ZASPiL Nr. 40 – July 2005

Speech production and perception: Experimental analyses and models

Editors: Susanne Fuchs, Pascal Perrier
and Bernd Pompino-Marschall

Preface

This special issue of the ZAS Papers in Linguistics contains a collection of papers of the French-German Thematic Summerschool on "Cognitive and physical models of speech production, and speech perception and of their interaction".
Organized by Susanne Fuchs (ZAS Berlin), Jonathan Harrington (IPdS Kiel), Pascal Perrier (ICP Grenoble) and Bernd Pompino-Marschall (HUB and ZAS Berlin) and funded by the German-French University in Saarbrücken this summerschool was held from September 19th till 24th 2004 at the coast of the Baltic Sea at the Heimvolkshochschule Lubmin (Germany) with 45 participants from Germany, France, Great Britain, Italy and Canada.

The scientific program of this summerschool that is reprinted at the end of this volume included 11 key-note presentations by invited speakers, 21 oral presentations and a poster session (8 presentations). The names and addresses of all participants are also given in the back matter of this volume.

All participants was offered the opportunity to publish an extended version of their presentation in the *ZAS Papers in Linguistics*. All submitted papers underwent a review and an editing procedure by external experts and the organizers of the summerschool. As it is the case in a summerschool, papers present either works in progress, or works at a more advanced stage, or tutorials. They are ordered alphabetically by their first author's name, fortunately resulting in the fact that this special issue starts out with the paper that won the award as best pre-doctoral presentation, i.e. Sophie Dupont, Jérôme Aubin and Lucie Ménard with "A study of the McGurk effect in 4 and 5-year-old French Canadian children".

Acknowledgements
We want to thank the German-French University in Saarbrücken for the funding of this summerschool, the staff of the Heimvolkshochschule Lubmin for their hospitality, the external reviewers for their helpful comments of the papers printed here, the invited speakers for their contribution, and all the participants who made this meeting a unique combination of scientific discussions at every time of the day, tango dances, songs, plays, a boat trip … You were all so brave for putting up with cold and rainy weather and for eating potatoes every day ☺.

Table of Contents

Sophie Dupont, Jérôme Aubin and Lucie Ménard
A study of the McGurk effect in4 and 5-year-old French Canadian children…. 1

Sascha Fagel
Merging methods of speech visualization 19

Alban Gebler and Roland Frey
Anatomical structures involved in non-human vocalization..... 33

Cédric Gendrot
Acoustic, kinematic and aerodynamic aspects of word-initial and word-final vowels in pre-boundary context in French..... 45

Ian S. Howard and Mark A. Huckvale
Learning to control an articulatory synthesizer by imitating natural speech ... 63

Bernd J. Kröger, Julia Gotto, Susanne Albert and Christiane Neuschaefer-Rube
A visual articulatory model and its application to therapy of speech disorders: a pilot study 79

Shinji Maeda
Face models based on a guided PCA of motion-capture data: Speaker dependent variability in /s/-/ʃ/ contrast production 95

Pascal Perrier
Control and representations in speech production..... 109

Hartmut R. Pfitzinger
Towards functional modelling of relationships between the acoustics and perception of vowels..... 133

Bernd Pompino-Marschall
Von Kempelen et al. –
Remarks on the history of articulatory-acoustic modelling..... 145

Nicolas Ruty, Annemie Van Hirtum, Xavier Pelorson, Ines Lopez and Avraham Hirschberg
A mechanical experimental setup to simulate vocal folds vibrations. Preliminary results *161*

Willy Serniclaes
On the invariance of speech percepts *177*

Antoine Serrurier and Pierre Badin
Towards a 3D articulatory model of nasals based on MRI and CT images..... *195*

Ralf Winkler and Walter Sendlmeier
Open quotient (EGG) measurements of young and elderly voices: Results of a production and perception study..... *213*

Appendix: Participants, invited speakers & organisers..... *227*

A study of the McGurk effect in 4 and 5-year-old French Canadian children

Sophie Dupont
Jérôme Aubin
Lucie Ménard
University of Quebec in Montreal, Montreal, Canada

It has been shown that visual cues play a crucial role in the perception of vowels and consonants. Conflicting consonantal stimuli presented in the visual and auditory modalities can even result in the emergence of a third perceptual unit (McGurk effect). From a developmental point of view, several studies report that newborns can associate the image of a face uttering a given vowel to the auditory signal corresponding to this vowel; visual cues are thus used by the newborns. Despite the large number of studies carried out with adult speakers and newborns, very little work has been conducted with preschool-aged children. This contribution is aimed at describing the use of auditory and visual cues by 4 and 5-year-old French Canadian speakers, compared to adult speakers, in the identification of voiced consonants. Audiovisual recordings of a French Canadian speaker uttering the sequences [aba], [ada], [aga], [ava], [ibi], [idi], [igi], [ivi] have been carried out. The acoustic and visual signals have been extracted and analysed so that conflicting and non-conflicting stimuli, between the two modalities, were obtained. The resulting stimuli were presented as a perceptual test to eight 4 and 5-year-old French Canadian speakers and ten adults in three conditions: visual-only, auditory-only, and audiovisual. Results show that, even though the visual cues have a significant effect on the identification of the stimuli for adults and children, children are less sensitive to visual cues in the audio-visual condition. Such results shed light on the role of multimodal perception in the emergence and the refinement of the phonological system in children.

1. Introduction

Studies have shown that speech perception not only relies on the decoding of the waveform, but also on visual information transmitted by the speaker's jaw and lips positions. Interactions between speakers under unimodal communication conditions are frequent and most of the time successful. Telephonic conversations, where only the auditory modality is used, are good examples of such unimodal interactions. Considering speech in natural face-to-face interactions does not discern between the contribution of both modalities and their integration in the audiovisual speech perception; that's why, since the first data relative to this kind of experiment were published in 1976, the McGurk effect has constituted an efficient research paradigm for the study of audiovisual speech perception.

The McGurk effect was first introduced in an article by McGurk and MacDonald (1976) in which results of conflicting audiovisual stimuli percepts were presented. The first manifestation of this effect occurred when an auditory stimulus /ba/ was dubbed to a visual stimulus /ga/. Most of the time, people perceived /da/, a percept which resulted from the fusion of the auditory and visual information. Using the McGurk effect, many studies have since been conducted on adults but only a few investigated the emergence of this effect in speech development.

When McGurk and MacDonald (1976) discovered this audiovisual integration phenomenon, they decided to study its robustness in English speaking children. They tested 54 adults (18-40 years old), 21 pre-school aged children (3-4 years old) and 28 children of school age (7-8 years old) and presented them stimuli in two conditions : audiovisual (for example, A (audio) /ba/ - V (visual) /ga/, A /ga/ - V /ba/, A /pa/ - V /ka/ and A /ka/ - V /pa/) and visual-only. The authors found that adult subjects were more influenced by the visual input than were the two other groups. The description of this phenomenon by the authors also showed that percepts reflecting the absence of any audiovisual integration were audio percepts for child subjects, and visual percepts for adult subjects.

In a 1978 study, MacDonald and McGurk sought to explain the phenomenon of the conflicting auditory and visual information by testing the manner-place hypothesis, according to which, in a face-to-face situation, manner of articulation in consonants is best recognized by ear, while articulation place is best integrated by eye. In their experiment, 44 English speaking adults participated in a perception test where they had to repeat the perceived utterance produced by a female speaker presented on a screen. Stimuli were conflicting

dubbings of the consonants /p, b, t, d, k, g, m, n/ in an /a/ vocalic context. The manner-place hypothesis was validated with percepts for which the audio part of the stimuli consisted of labial consonants and the visual part consisted of non-labial stimuli (for example, A /ma/ - V /ga/ stimuli were mostly identified as /na/), but was "less satisfactory with respect to nonlabial sound/labial lips combinations" (MacDonald et McGurk 1978: 256) such as A /ga/ - V /pa/ stimuli, which were identified as /ga/. This study showed the general effect of vision in speech perception in face-to-face interactions.

In 1984, Massaro conducted a study concerning the evaluation of the integration of information in speech perception, which aimed at studying the developmental aspects of speech perception. He compared results from 11 children to those from 11 adults in a phonemic identification task, and found that the children showed half the visual influence of the adults. He mentioned that children seemed sensitive to the correspondence of visual and auditory information, but auditory information had a greater influence on their categorical perception in language acquisition.

Rosenblum et al (1997) investigated the McGurk effect in 5-month-old babies. They wanted to show that speech representation of babies was amodal and that babies processed the audiovisual input by making a supramodal association with the audiovisual input instead of an association with the phonemic identity. By creating three types of audiovisual stimuli (conflicting, non-conflicting and audio-only) and by using an habituation technique, they measured fixation time and found a significant difference between the fixation time of the stimulus A /da/ - V /va/ and the stimulus A /va/ - V /va/, but not between A /ba/ - V /va/ and the stimulus A /va/ - V /va/. Since they ensured that a general preference of the babies for the A /da/ - V /va/ stimuli or an auditory similarity between the stimuli could not account for this significant difference, they concluded that they had "observed evidence for a McGurk-type effect in 5-month-old infants".

More recently, Robert-Ribes et al. (1998) identified the complementarity and synergy between vision and hearing as two factors that influence the effectiveness of audiovisual speech perception. Complementarity is related to the manner-place hypothesis of MacDonald and McGurk (1978), in that it explains that the manner of articulation is best transmitted by the auditory channel and that place of articulation is best transmitted by the visual channel. Synergy is a property related to the perception enhancement caused by the interaction between both the auditory and the visual modalities; audiovisual perception is thus always better than visual-only perception and auditory-only perception. The authors also used the notion of viseme, first introduced by

Fisher (1968) to refer to “visual phonemes” and this notion will be relevant to the present study.

Altogether, these studies suggest that audiovisual perception development is not equally weighted during speech development: babies and young children would rely less than adults on visual cues. But since we know that they can use them in specific perception tasks, like matching acoustical signals to visual patterns as early as 4 and 5-months old (Kuhl and Metzoff, 1984) and that attention can be monitored during the perception tests (Massaro, 1984; Tiipana et al., 2004), we could hypothesize that children use as many cues as available during speech development, but to an extent specific to each child. The goal of this research is thus to describe the role of auditory and visual cues during speech development in French Canadian speaking subjects.

2. Method

2.1. Stimuli Recordings

Audiovisual recordings of a French Canadian adult speaker producing symmetric VCV sequences were carried out, using a *Panasonic* Mini-DV camcorder. Vowels /a/ and /i/ and consonants /b/, /d/, /g/, and /v/ were used to construct the following sequences, repeated three times: [aba], [ada], [aga], [ava], [ibi], [idi], [igi], and [ivi]. Speech rate, intonation, and intensity were controlled by the experimenters in order to maintain them at constant levels. As can be observed in Figure 1, a close-up on the lower part of the speaker’s face was made. After the second vowel of each sequence, the speaker was told to return to a neutral position, lips closed.

Those sequences were digitized at a frame rate of 29.94 frames per second and at a sampling frequency of 44.1 KHz. Audiovisual data were imported using Imovie (Apple) and bisyllabic sequences were edited. From the three repetitions recorded for each sequence, we kept only the one with the highest acoustic quality and movement clarity; each sequence lasted about 2800 ms.

Figure 1: Speaker's image after producing a sequence.

2.2. Audio and Video Processing

Recorded stimuli were edited in order to create the four following conditions. Bimodal non-conflicting stimuli were composed of the original image and sound of the speaker, repeating the 8 VCV sequences ([aba], [ada], [aga], [ava], [ibi], [idi], [igi], [ivi]) and did not require any editing. The same stimuli were used to construct the 8 visual-only stimuli, in which the sound was turned off. Unimodal audio condition stimuli consisted of the audio track of the original audiovisual non-conflicting stimuli.

Bimodal conflicting stimuli were constructed to generate consonantal conflicts between the acoustic and visual signals. *Adobe Premiere Pro 7.0* was used to separate the signals and superimpose them. Prior to this dubbing, closing and opening times of the mouth and lips were measured to ensure optimal synchronization between the audio and the video tracks. The combination of each of the four consonants in the auditory modality to every different consonant in the visual modality into the two vocalic contexts gave rise to a total of 24 stimuli.

2.3. Perception Test

2.3.1. Subjects

Eight children (all female) from four years and three months to five years and nine months (mean of four years and seven months) participated in this study. They were recruited from a daycare center in Montreal. 10 women, from 22 to 31 years old (mean of 25 years old) also served as subjects. All adult participants were enrolled in a Linguistics degree at the University of Quebec in Montreal. Children received a certificate of participation as a gift for their

contribution and adults received no compensation. All subjects were native French Canadian speakers and presented no vision or hearing problems. Most of the subjects did not know the goal of the experiment and did not seem to be conscious of the conflicting nature of the presented stimuli. Three adult subjects noticed the conflicting nature of the stimuli and believed this was done in order to disturb their perception. Percepts of those subjects were kept since McGurk and MacDonald (1976) mentioned the fact that being aware of the way stimuli were constructed did not inhibit the manifestation of the McGurk effect. Furthermore, informal comparisons showed no difference between the results of the two groups (those who were aware of the conflict and those who were not).

For both groups of subjects (children and adults), conflicting and non-conflicting stimuli were grouped in the bimodal condition category. Unimodal visual and auditory conditions, as well as bimodal conditions, were presented in three separated sessions to participants. Unimodal conditions each consisted of 8 stimuli, while the bimodal condition had 32 stimuli. Stimuli were randomized and presented once, separated by a four-second silent black screen.

2.3.1.1. Adults

Using *Windows Media player 9.0*, stimuli were presented to adult subjects in a perception test where the task was to be attentive to the movie presented on the screen and to the sound presented in the earphones and to write down the perceived utterance. In addition to the four-second black screen between each stimulus, subjects could also press pause between sequences to allow them more time to write down their responses. In order to avoid a lack of attention due to the task of writing and watching the screen at the same time, we asked the last three subjects who participated in the test to repeat the perceived utterance, instead of writing it down. A random list of stimuli was presented in the following order: audiovisual (conflicting and non-conflicting mixed) stimuli, audio stimuli, and visual stimuli. This test took place in a quiet room and lasted about fifteen minutes.

2.3.1.2. Children

For children, this test was presented as a game in which they were invited to be attentive to what was presented on the screen and to what they were hearing in the speakers. Their task was to repeat what they thought the speaker had just said. Two experimenters were present to write down the children's percepts, providing an inter-judge agreement and ensuring that the children were paying attention to what they were being presented. Experimenters monitored the test

by pressing the pause button between each utterance. A random list of stimuli was presented in the following order: audio stimuli, visual stimuli and audiovisual stimuli.

2.4. Data processing

For each condition, classification criteria were established for the consonant parts of the percepts; vocalic parts were not analyzed. Results are based on the mean numbers of percepts for a given group at a given condition for a given classification criterion. Then, results obtained from the children's data to a given classification criterion were compared to those of the adults for the same criterion. One-way ANOVAs were computed with Statistica 6.1. Student's t-tests were used to measure the perception difference of non-conflicting stimuli among a same group of subjects, but for different conditions (audio-only, visual-only and audiovisual). Student's t-tests were also performed to measure the perception difference of conflicting stimuli among a given type of percept (audio, visual, fusion, combination, other) between the two groups of subjects. The difference observed between the children and the adults will be described in these statistical terms, but remarks on specific percepts will be allowed only if at least 2 responses are similar. Indeed, since 16 answers were collected with the child subjects to each stimulus (8 subjects * 2 vocalic contexts) and 20 answers with the adults subjects (10 subjects * 2 vocalic contexts), a percept has to have been proposed twice by a subject or once by two subjects to be considered; percentage of 13% for children (2 percepts/16 responses) and 10% for adults (2 percepts/20 responses) is thus required.

3. Results

The principal goal of the following analysis was to describe the use of auditory and visual cues in the perception of four French Canadian consonants (/b/, /d/, /g/, /v/) by preschool-aged children and adults. To do so, we examined the results of the auditory-only and visual-only conditions to assess the use of each modality by the listeners. The analysis of the results of the conflicting stimuli revealed an influence of the visual component of the stimuli for both preschool-aged child and adult subjects, but the influence in children was clearly lower.

3.1. Control conditions

The analysis of the responses given to the unimodal (audio-only and visual-only conditions) and non-conflicting bimodal stimuli aimed at measuring: 1) the listeners' ability to identify the sequences in audio only and visual only

conditions and 2) the extent to which visual cues presented together with audio information improves intelligibility. Results can also provide an assessment of the quality of the audio components of the consonants as good exemplars of their phonetic categories.

3.1.1. Correct responses

For the auditory-only and the non-conflicting auditory-visual stimuli, percepts for which place and manner of articulation were the same as the presented stimuli have been treated as correct responses. For the visual-only condition, the responses for which the percepts and the stimuli shared the same place of articulation have been considered as correct responses. Table 1 presents examples of correct responses given by the two groups of subjects for each condition. Percepts which did not meet the criteria mentioned above have been classified as "others".

Table 1: Some examples of the classification of the percepts in the control conditions.

Control conditions	**Stimuli**		**Percepts**	
	Audio component	**Visual component**	**Correct responses**	**Others**
Auditory-only	/b/	-	/b/	/d/
Visual-only	-	/d/	/d/, /t/	/b/
Auditory-visual	/g/	/g/	/g/	/v/

Figure 2 shows the average values of correct responses given by the adult subjects (top graph) and child subjects (bottom graph) for each consonant in the three control conditions. It is noteworthy that a ceiling effect is present for the /b/, /d/, and /g/ adults' results, which all reach a proportion of correct responses of at least 85.0%. The slightly lower performances observed for the /v/ stimuli in the audiovisual and audio-only conditions cannot be explained by specific acoustic characteristics since no effect was observed on the identification of the conflicting stimuli having /v/ as audio part. Results of ANOVA did not reveal any significant difference in adults between any of the three control conditions. It shows that the audio and the visual components of the stimuli were intelligible and that the adults were competent in the tasks of listening and lip-reading.

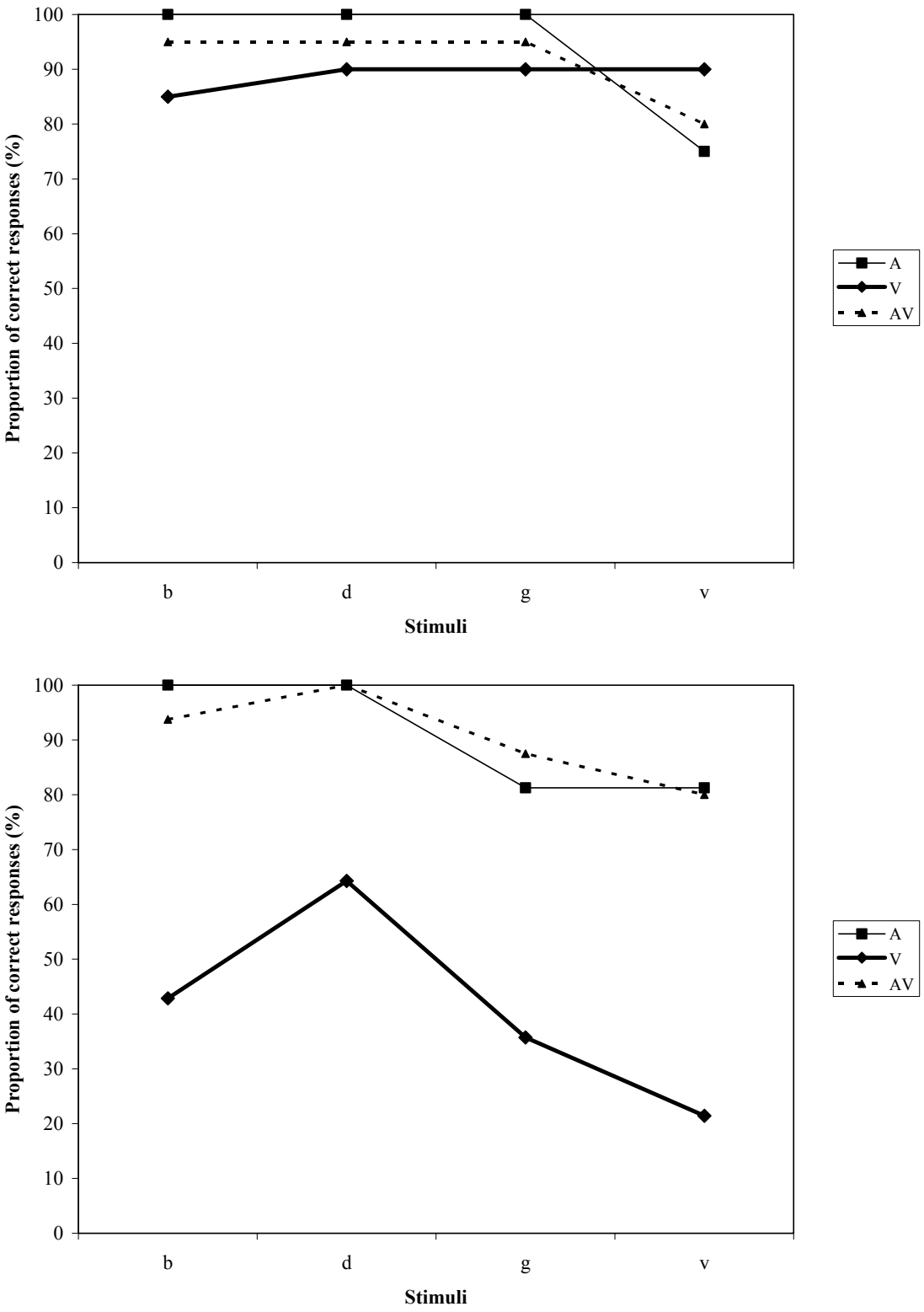


Figure 2: Proportion of correct answers for each consonant given by the adult subjects (top) and children subjects (bottom) in the three control conditions: visual-only (thick black line), auditory-only (thin black line) and non-conflicting auditory-visual conditions (dotted line).

In the bottom graph, we can observe the same kind of ceiling effect in the children results for the auditory-only and the auditory-visual conditions. However, a t-test revealed significant differences between the visual-only and the auditory-only conditions ($t=4.97$ $p<0.005$) and between the visual-only and the non-conflicting auditory-visual conditions ($t=4.75$ $p<0.01$). The performance of children in lip-reading was clearly poorer than the adults' and than their own performance in the two other control conditions. This visual-only condition was undoubtedly the most difficult one for children of this age since this task is rather unusual and infrequent in their day-to-day life. Since they were not given any choice of answer, we could not compare their results with a specific chance level; on the other hand, they gave correct answers between 21.4% and 64.2% of the time so we considered that they were sensitive to a certain extent to the speaker's articulatory movements.

3.2. Conflicting conditions

The analysis of the responses given to the conflicting auditory-visual stimuli aimed at studying the influence of the auditory and the visual modalities in multimodal speech perception. The identification of the conflicting stimuli gave rise to five classes of percepts: audio, visual, combination, fusion and other percepts. Data will be described and compared across the groups of child and adult subjects. Table 2 presents examples of each of the five classes of percepts given by the two groups of subjects to the conflicting stimuli.

Table 2: Some examples of the 5 classes of the percepts of the conflicting stimuli.

Stimuli		**Percepts**					
Audio component	**Visual component**	**Audio**	**Visual**	**Combinations**		**Fusions**	**Others**
				Strict	**Extended**		
/b/	/d/	/b/	/d/	-	-	-	/t/, /vg/
/b/	/g/	/b/	/g/	-	-	/d/, /θ/	-
/g/	/v/	/g/	-	/vg/, /gb/	/bg/	/d/	/f/
/v/	/d/	/v/	/d/	-	/bv/, /zd/	/θ/	/t/

3.2.1. Audio percepts

Percepts for which place and manner of articulation were similar to those of the audio component of the conflicting stimulus have been considered as audio percepts. They emphasize the absence of conflicting perception between the auditory and visual modalities. Figure 3 illustrates the means of the proportion of conflicting stimuli identified as audio percepts.

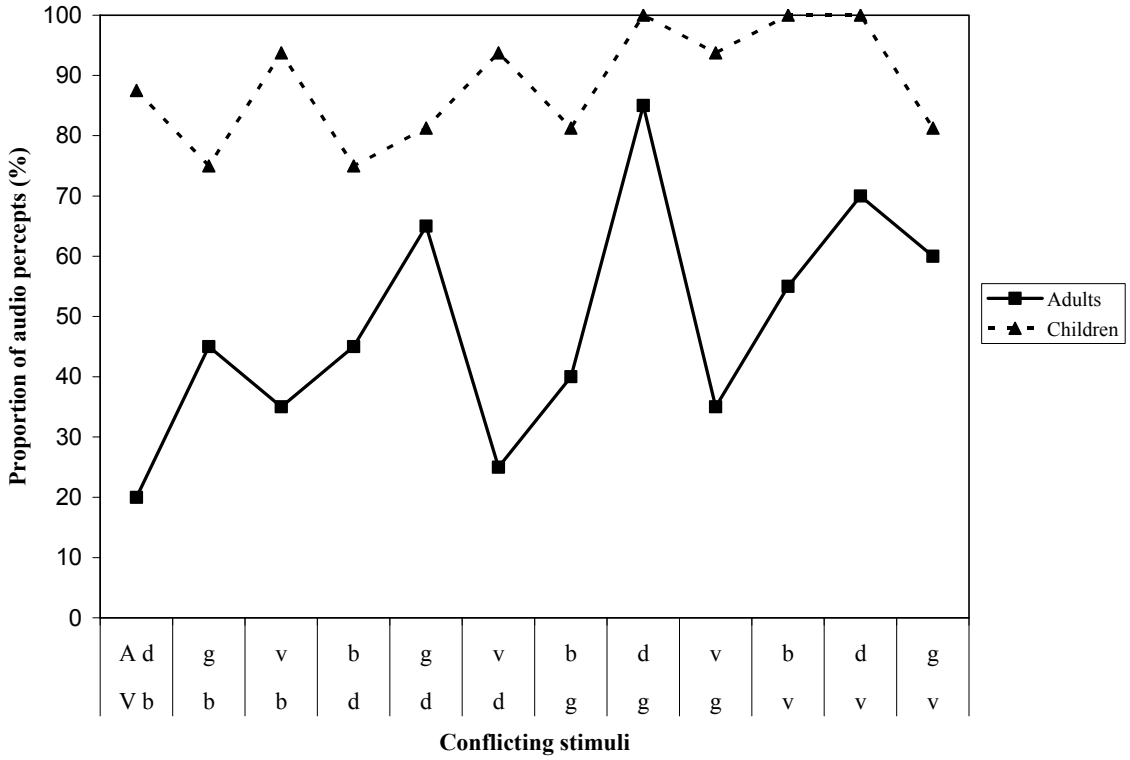


Figure 3: Proportion of audio percepts given by the adult subjects (black line) and child subjects (dotted line) for the 12 conflicting stimuli.

The proportion of audio percepts is significantly higher for children than for adults, as revealed by a one-way ANOVA ($F(1,22)=42.10$, $p<0.001$). These results are related to those found in the control conditions in children, since they already seemed to be more sensitive to auditory cues. We can also observe a tendency in the children's results: stimuli for which places of articulation of the auditory and visual components are the closest seem to give rise to more audio percepts. This tendency shows that children are more sensitive to auditory cues when less conflicting visual information is present. Even if the adults show a smaller proportion of audio percepts than children, the same tendency (for example for stimuli like A/g/-V/d/) is present.

3.2.2. Visual percepts

Visual percepts are those for which place and manner of articulation are similar to the visual component of the conflicting stimuli. Such percepts demonstrate a major dominance of the visual modality over the auditory modality. Figure 4 illustrates the means of the identification of the conflicting stimuli as visual percepts.

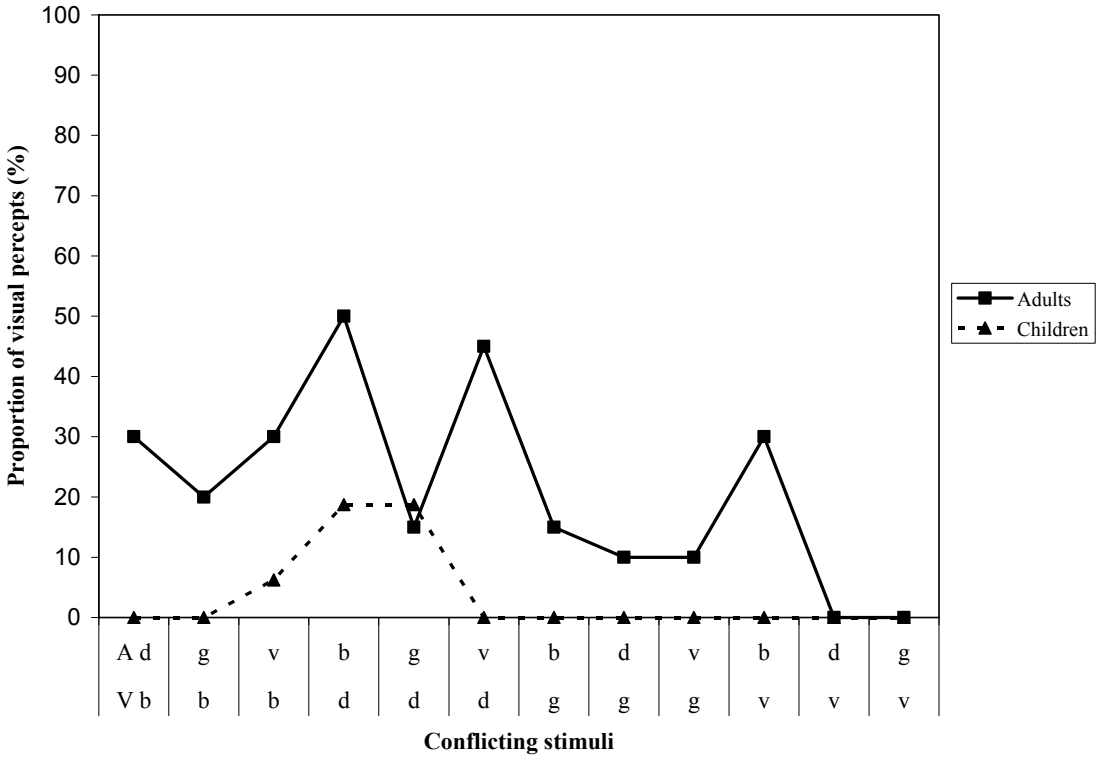


Figure 4: Proportion of visual percepts given by the adult subjects (black line) and children subjects (dotted line) for the twelve conflicting stimuli.

The proportion of visual percepts is significantly lower for children than for adults ($F(1,22)=11.90$, $p<0.01$). Since the results in the control conditions and the results of the audio percepts revealed the dominance of the auditory modality of children in speech perception, it is not surprising to observe such an effect. According to the criterion defined in section 2.4, children showed visual percepts in two contexts (A /b/ - V /d/ and A /g/ - V /d/), both of them including the dental /d/ visual consonant, which is surprising since dental place of articulation is not the most visible one. Note that the visual percept associated with A /v/ - V /b/ corresponds to only one response and could hardly be significant.

Visual percepts emphasize the important influence of the visual modality in speech perception; stimuli were presented in a quiet room and no surrounding noise could have distracted the subjects in their auditory perception. Thus, these visual percepts emphasize the dominance of the use of the visible articulatory gestures (visual cues) over the acoustical characteristics (auditory cues) of the conflicting stimuli.

3.2.3. Combination percepts

As can be seen in Table 2, combination percepts have been divided into two subclasses: strict combinations and extended combinations. Strict combination percepts include both the auditory and the visual consonants of the conflicting stimuli. Extended combination percepts are composed of two phonemes: one directly taken from the auditory or the visual component, and another one which is not directly taken from the components. Figure 5 shows the means of the proportion of conflicting stimuli perceived as combinations (strict and extended together).

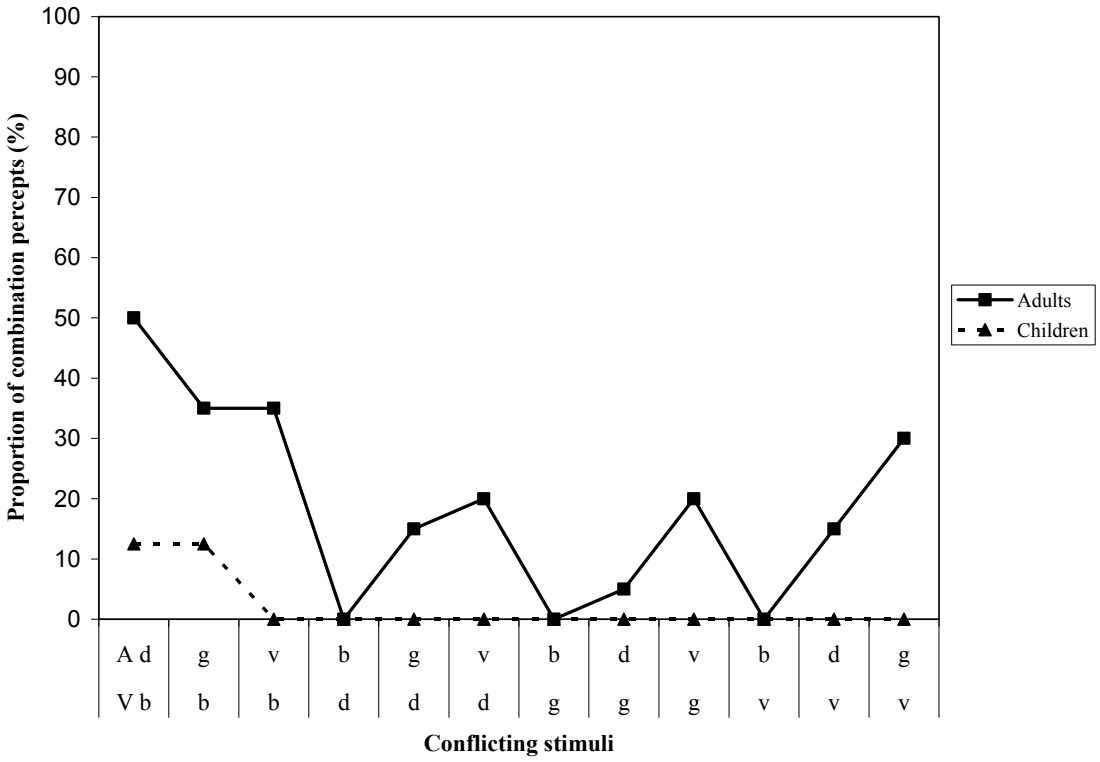


Figure 5: Proportion of combination percepts given by the adult subjects (black line) and child subjects (dotted line) for the twelve conflicting stimuli.

The proportion of combination percepts is significantly lower for children than for adults ($F(1,22)=11.58$, $p<0.01$). Since children showed only strict combinations, we compared the number of strict combination percepts of the children to the number of both strict and extended combinations percepts of the adults. Children showed combination percepts in two contexts: A /d/ - /V/ b/ and /A/ g/ - V /b/ and these contexts are those in which adults showed the highest proportion of combination percepts as well. This result suggests that the articulatory movements of the bilabial /b/ gave rise to more combination percepts when combined with a less visible visual consonant (dental /d/ and

velar /g/). On the other hand, some of the contexts in which adults gave the highest proportion of combination percepts include the audio velar /g/ consonant (A /g/ - V /b/ and A /g/ - V /v/). This result is not in agreement with McGurk and MacDonald (1976) nor MacDonald and McGurk (1978) who found that combination percepts were more frequent when dental and bilabial consonants were presented in the audio modality.

Combination percepts demonstrate the simultaneous influences of both the information provided by the auditory and the visual modality so it is interesting to note that children seem to be sensitive to the visual information, though to a smaller extent than the adults.

3.2.4. Fusion percepts

Fusion percepts are the most typical percepts of the McGurk Effect. When the subjects identified the conflicting stimuli by using only one phoneme and that phoneme was not that of the auditory part nor of the visual part, but instead, an "integration" percept, the responses were classified as fusions. Figure 6 shows the means of the identification of the conflicting stimuli as fusions.

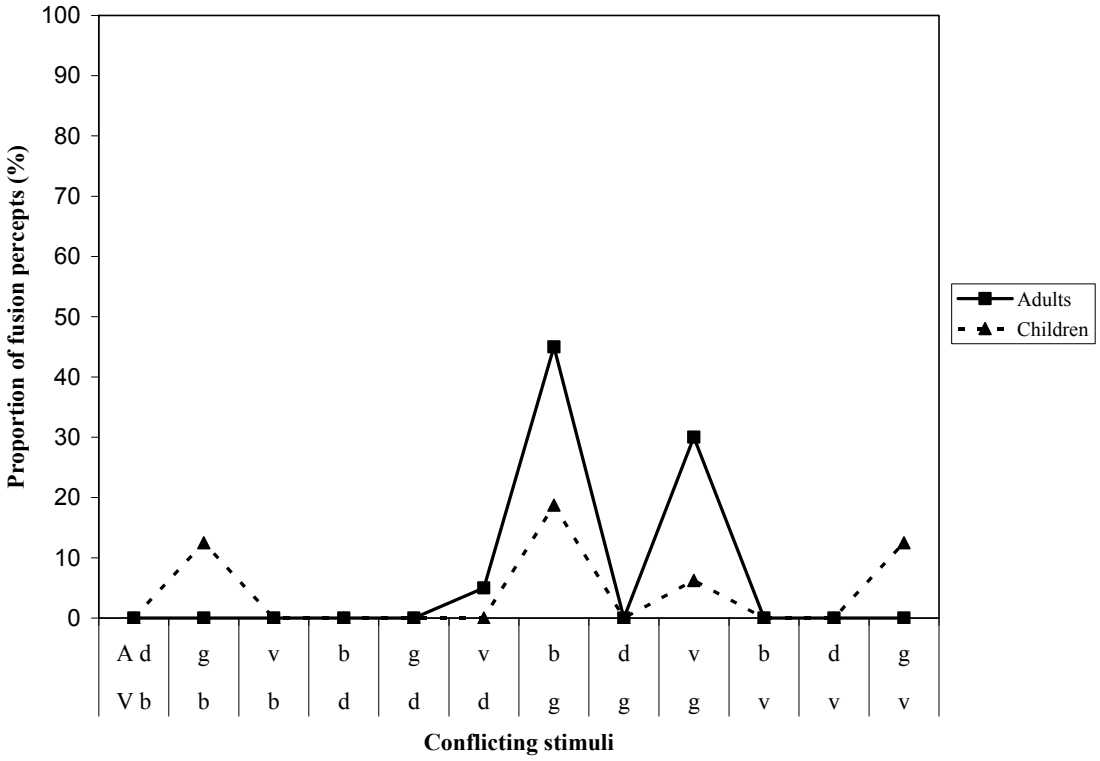


Figure 6: Proportion of fusion percepts given by the adult subjects (black line) and child subjects (dotted line) for the 12 conflicting stimuli.

The proportion of fusion percepts of children is not significantly different from that of adults. However, it is interesting to note that children showed fusions in three contexts (A /d/ - V /b/, A /b/ - V /g/ and A /g/ - V /v/) and adults in only two contexts (A /b/ - V /g/ and A /v/ - V /g/). The A /b/ - V /g/ stimuli are the ones which gave rise to the most /d/ or /θ/ fusion percepts. These percepts show that both children and adults were sensitive to visual cues and that they integrated them to the auditory cues giving rise to a fused perceived consonant.

3.2.5. Other percepts

All percepts given by the subjects which did not meet the conditions described previously have been classified as other percepts. Figure 7 shows the means of the identification of the conflicting stimuli as other percepts.

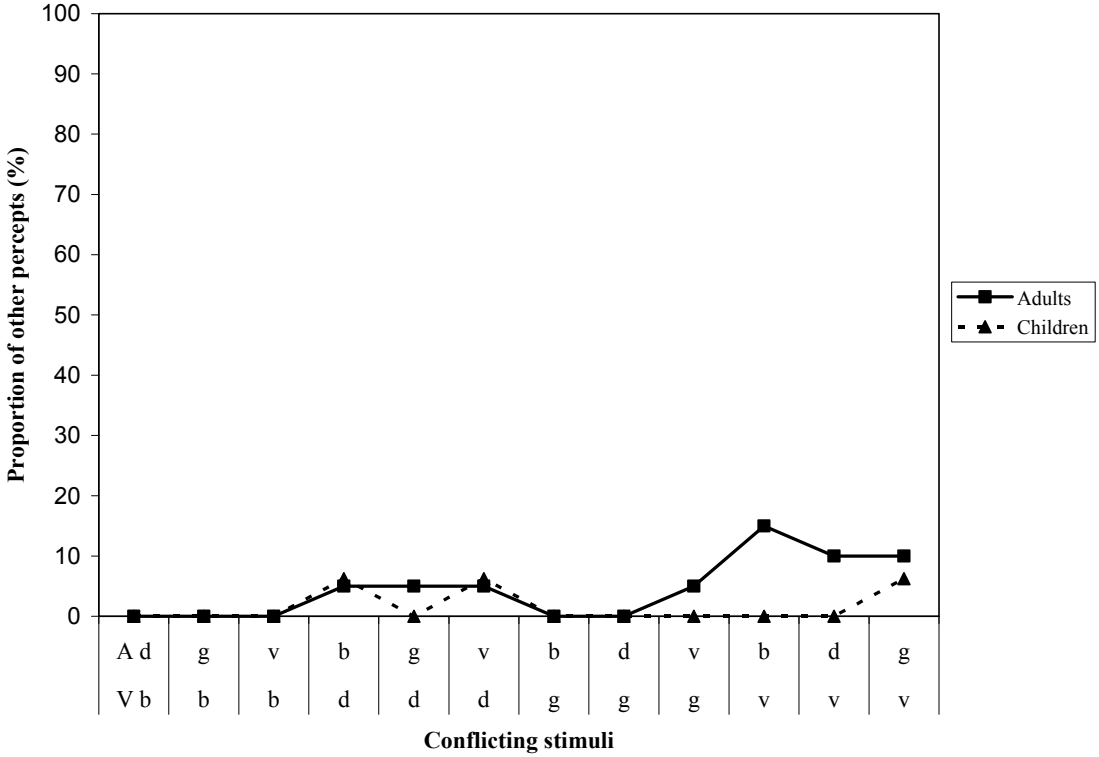


Figure 7: Proportion of other percepts given by the adult subjects (black line) and child subjects (dotted line) for the twelve conflicting stimuli.

The proportion of other percepts is not significantly different from that of adults. Some of the other percepts cannot be recovered by the auditory or the visual cues (ex: A/b/ - V/d/ identified as /vg/) and could be considered as random responses or errors. However, the interpretation of some other percepts can be related to the visual component. For example, 36.4% of the other percepts given

by the adults consisted in the devoiced visual component of the conflicting stimuli.

It is also interesting to note that the children did not give more 'other' percepts than adults; we could have hypothesized that they would have given more random responses due to a lack of attention or to the difficulty of the task, but it seems not to have happened.

4. Discussion and conclusion

Many differences between the child subjects' percepts and the adult subjects' percepts appeared to be significant. Overall, the children subjects were more sensitive to auditory cues. Their low performance in lip-reading and their greater number of audio percepts in response to the conflicting stimuli are closely akin to Massaro's postulate (1984) on the greater influence of the auditory information on the perception of the phonetic categories in language acquisition.

Even if the children subjects have been much more influenced by the auditory information, they have nonetheless showed the McGurk effect in some contexts, but the effect was lower than in adults and much more variable. These results have to be taken into account in a description of the role of the vision and audition in the phonological development in children.

But many questions still need to be investigated to provide a coherent description. The magnitude of the McGurk effect across development seems to be non-linear; Rosenblum's (1997) data collected by a head-turn procedure with 5-month-old infants suggest that they are sensitive to the McGurk effect and data collected in preschool-aged children like ours (McGurk and MacDonald, 1976; Massaro 1984) suggest it too but to a smaller extent. Finally, evidence has widely demonstrated the McGurk effect in adults. We could have hypothesized that since infants are sensitive to visual cues and since children are in the process of acquiring their phonological categories, they would use as many cues as available. However, this hypothesis is not supported by the present data.

Further work should be conducted with more subjects over a wider range of age to try to localize some kind of "McGurk effect critical period" and to study what would underlie such a period. Finally, the relationship between speech perception and production could be explored in order to compare the use of auditory cues and visual articulatory movements in children and adults. Bilingualism might be a way to explore this issue. It may be of interest to

attempt to quantify second language fluency in some way, to determine whether some correlation might exist with observed McGurk percepts.

Acknowledgment

We would like to thank our speaker, Raoul Bugueno, for his patience. This work greatly benefited from the discussions and suggestions of the participants at the French-German summerschool on "Cognitive and physical models of speech production, perception and perception-production interaction", and from the thorough review by Pascal Perrier, Mark Tiede and Susanne Fuchs. This work was supported by FQRSC and CRLMB grants.

References

Fisher, C. L. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11: 796-804.

Kuhl, P. K. & Meltzoff, A. N. (1984). The intermodal representation of speech in infants, *Infant Behavioral Development*, 7: 361-381.

MacDonald, J. M. & McGurk, H. (1978). Visual influences on speech perception process. *Perception and Psychophysics,* 24: 253-257.

Massaro, D. W. (1984). Children's Perception of Visual and Auditory Speech. *Child Development*, 55: 1777-1788.

McGurk H. & MacDonald, J. M. (1976). Hearing lips and seeing voices. *Nature*, 264: 746-748.

Robert-Ribes, J., Schwartz J-L, Lallouache T. & Escudier, P. (1998). Complementary and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise. *Journal of Acoustical Society of America*, 103 (6): 3677-3689.

Rosenblum, L. D., Schmuckler M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, 59(3): 347-357.

Tiipana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception . *European Journal of Cognitive Psychology*, 16: 457-472.

Merging methods of speech visualization

Sascha Fagel
Technical University Berlin

The author presents MASSY, the MODULAR AUDIOVISUAL SPEECH SYNTHESIZER. The system combines two approaches of visual speech synthesis. Two control models are implemented: a (data based) di-viseme model and a (rule based) dominance model where both produce control commands in a parameterized articulation space. Analogously two visualization methods are implemented: an image based (video-realistic) face model and a 3D synthetic head. Both face models can be driven by both the data based and the rule based articulation model.

The high-level visual speech synthesis generates a sequence of control commands for the visible articulation. For every virtual articulator (articulation parameter) the 3D synthetic face model defines a set of displacement vectors for the vertices of the 3D objects of the head. The vertices of the 3D synthetic head then are moved by linear combinations of these displacement vectors to visualize articulation movements. For the image based video synthesis a single reference image is deformed to fit the facial properties derived from the control commands. Facial feature points and facial displacements have to be defined for the reference image. The algorithm can also use an image database with appropriately annotated facial properties. An example database was built automatically from video recordings. Both the 3D synthetic face and the image based face generate visual speech that is capable to increase the intelligibility of audible speech.

Other well known image based audiovisual speech synthesis systems like MIKETALK and VIDEO REWRITE concatenate pre-recorded single images or video sequences, respectively. Parametric talking heads like BALDI control a parametric face with a parametric articulation model. The presented system demonstrates the compatibility of parametric and data based visual speech synthesis approaches.

1. Introduction

Speech communication usually consists of two coherent information streams, i.e. audition and vision. This is possible due to the fact that the movements of the speech organs that form the utterance become manifest in the acoustical and optical domain and hence are audible and visible. At least under acoustically bad conditions both information streams are used jointly to increase the robustness against transmission errors (Sumby & Pollack, 1954; Erber, 1969). This property of natural speech can also be helpful for speech synthesis (Benoît et al., 1995; Beskow, 2003). This is the case although there is not necessarily a single underlying process like in natural speech but – at least in current unlimited audiovisual speech synthesis systems ("talking heads") – the audio signal and the video signal are synthesized in most cases separately and are played back synchronously.

Although they borrow some techniques from one another, the visualization method of most audiovisual speech synthesis systems can be classified as either image based or parametric. The first class of systems concatenates parts of pre-recorded video speech material (comparable to concatenative audio synthesis systems). The second class models the speech production process by means of physiological, articulatory, or facial parameters. However, the present paper shows that both approaches are not mutually exclusive and that they can be combined in a single system.

2. Parametric and image based talking heads

Parametric visual speech synthesizers generate a sequence of values for a number of fixed parameters. These parameters can be e.g. virtual articulators like tongue tip, tongue back, lip opening, and so on. A synthetic face is then manipulated according to the parameter values. Simple spatial co-articulation, i.e. movements of an articulator caused by the movement of another one near to it, can be modeled in the facial animation. But the temporal co-articulation, i.e. articulators start to move towards their target position for one speech segment in preceding segments and partly carry over the target positions to subsequent segments, is modeled by generating appropriate parameter values.

In contrast, image based visual speech synthesizers use pre-recorded single images (e.g. MIKETALK: Ezzat & Poggio, 2000) or video sequences (VIDEO REWRITE: Bregler et al., 1997). These image databases are indexed by phonemes (or visemes) or phoneme (or viseme) sequences, respectively. Co-articulation can be taken into account by recording a database containing phonemes in possibly all needed contexts. Co-articulation differences as they occur between different languages (e.g. between lip rounding in Turkish and American English:

Boyce, 1990) cannot be realized with the same database. Some main properties of speech visualization systems are summarized in Table 1.

Table 1: General properties of talking heads.

Either	**Or**
synthetic face	natural images
one instance of the face	many instances of one face
(articulation) parameters	database indexed by (classes of) phonemes
co-articulation (mostly) outside the face model	co-articulation (hidden) inside the face model

Some recent developments do not completely fit in the parametric vs. image based distinction: MARY101 (Ezzat, 2002) defines a set of prototypic images and the optical flow (Horn & Schnuck, 1981) between them. The system generates appropriate video frames which do not necessarily have to lead from one prototype to another (which was the case for MARY101's predecessor MIKETALK). VOICE PUPPETRY (Brand, 1999) is trained by audiovisual recordings and then provides a sequence of facial motion vectors related to the audio track. These facial motion vectors can be applied to other prepared faces. The visual extension (Minnis & Breen, 2000) of the concatenative audio synthesis system LAUREATE (British Telecom) associates N-visemes with face deformations which can be applied to a 3D face model.

3. The MODULAR AUDIOVISUAL SPEECH SYNTHESIZER

The system that demonstrates the compatibility of parametric and image based approach is called MASSY (MODULAR AUDIOVISUAL SPEECH SYNTHESIZER). A plain text serves as system input. The phonetic articulation module creates the phonetic information, which consists of an appropriate phone chain on the one hand and - as prosodic information - phone and pause durations and a fundamental frequency curve on the other hand. From this data, the audio synthesis module generates the audio signal and the visual articulation module generates motion information. This motion information consists of control commands for virtual articulators given by an articulation model. Hence, the control part of the visualization follows in general the parametric approach. The face module interprets the motion information and adds the audio signal to create the complete audiovisual speech output. Figure 1 shows a system overview.

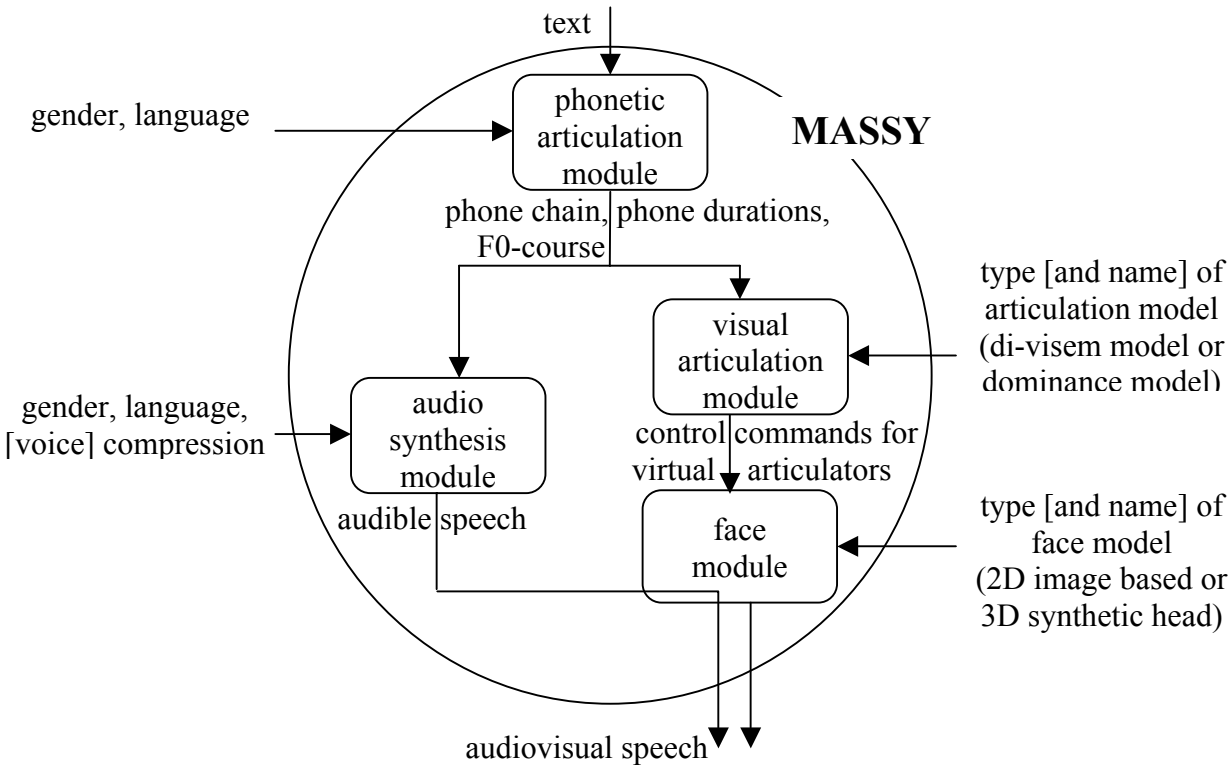


Figure 1: Schematic system overview of MASSY.

The system in its present state can be tested at http://avspeech.info. This website provides a user interface to a fully functional text-to-audiovisual-speech synthesis system built of the modules of MASSY. The phonetic transcription and the voice can be German or English, male or female. Both face models described below are available including a simple tool to mark facial feature points in any uploaded image file to make a new face talking. Among others, experimental settings of speaking rate, hyper/hypo-articulation, and phoneme/ viseme replacements to generate McGurk stimuli (McGurk & MacDonald, 1976) are possible.

3.1. Visual articulation module

The visual articulation module generates a sequence of values for articulation parameters to synthesize a phone chain given by the phonetic articulation module. These parameters control the face model implemented in the subsequent module. The articulation parameters of the articulation model currently are

- lip width
- jaw height
- lip height

- tongue tip height
- tongue back height
- lower lip retraction

The lip width is 0 at neutral position (relaxed state), 1 at maximum narrowing and -1 at maximum spreading. For the production of some vowels the real vocal tract is lengthened by lip protrusion, for some other vowels shortened by lip spreading. A negative correlation between lip protrusion and spreading is assumed. At least in German and English which are the languages MASSY currently can "speak" there is no acoustic-articulatory need to spread the lips while protruding them or to narrow them without protrusion. This is an appropriate simplification if the goal is to realize one plausible articulation and not to clone a specific speaker. Hence, lip rounding and narrowing are combined to one articulation parameter. The lower jaw height is 0 at closed jaw and 1 at maximum opening. The lip height is 0 at neutral position relative to the upper and lower teeth. It is 1 for the lips moved maximum towards each other on the upper and lower jaw and -1 if the lips are moved maximum apart. So the vertical lip opening depends on both the jaw height and the lip height and the lips can be closed only if the jaw is not wide open (see Figure 2).
The tongue tip height and tongue back height are 0 at relaxed tongue and 1 at tongue contact at the alveoli or the palate, respectively. For this, the absolute values of the displacement vectors for tongue tip height and tongue back height are scaled by the lower jaw height in order not to break through the palate. The retraction of the lower lip is 0 at neutral position and 1 at retracted position for labiodental constrictions. The set of motion parameters was chosen with respect to the visibility of German phones displayed by MASSY. Motion parameters for tongue advance and velum closure are currently under construction. The visual articulation module implements alternatively two different articulation models to generate the values for the articulation parameters: a di-viseme model and a dominance model (Löfqvist, 1990). Details of the algorithms can be found in Fagel & Sendlmeier (2003).

3.2. Face module

The face module visualizes the articulator movements described by the articulation parameters. The module creates an animation of a face and dubs the synthesized speech audio. One face model is a 3D synthetic head with a set of displacement vectors for each articulation parameter. A second alternative face model is image based. Figure 3 shows image sequences generated by the two face models.

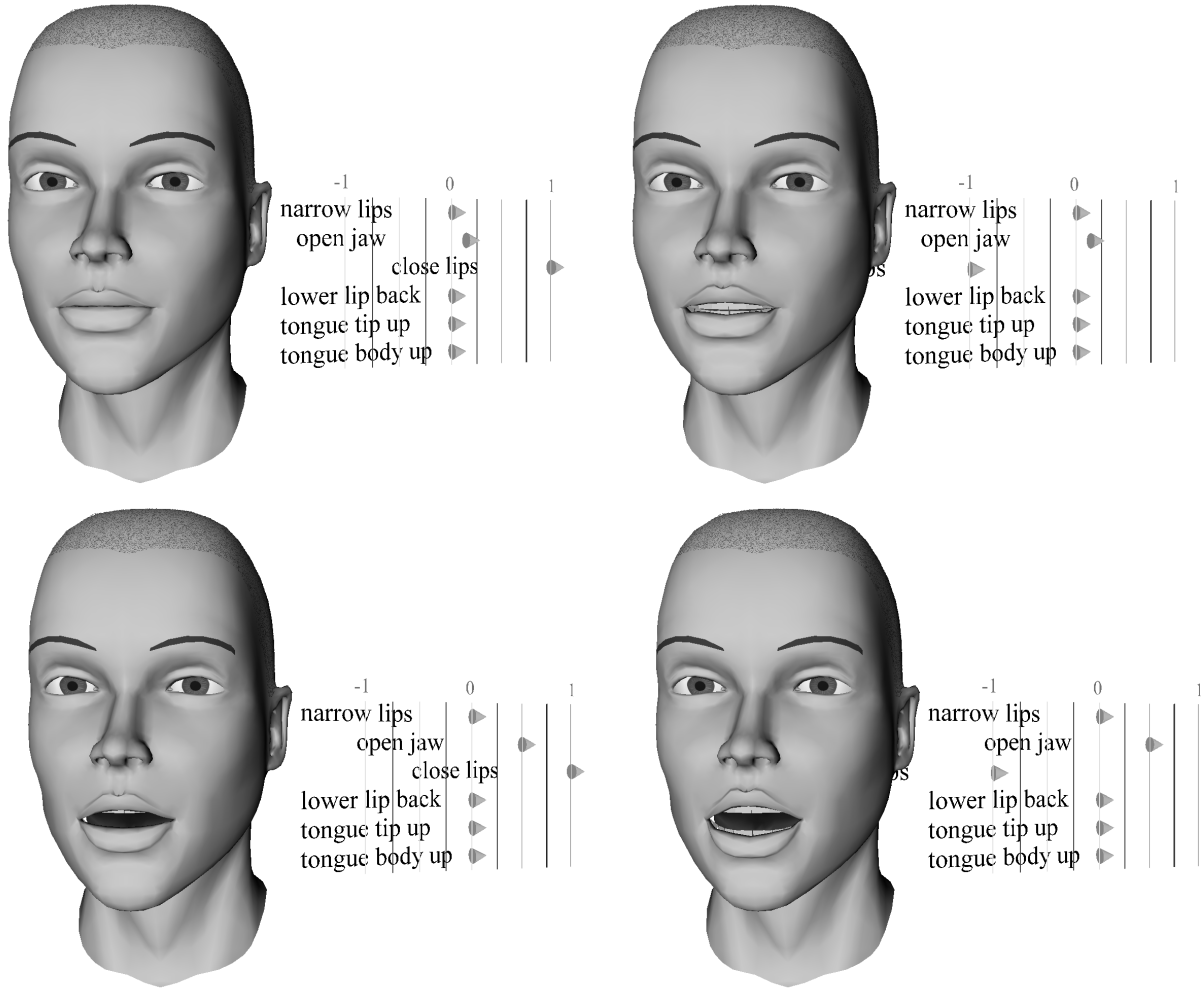


Figure 2: Minimum (both left images) and maximum (both right images) lip height at lower jaw nearly closed (both top images) and half opened (both bottom images) displayed with the 3D synthetic face.

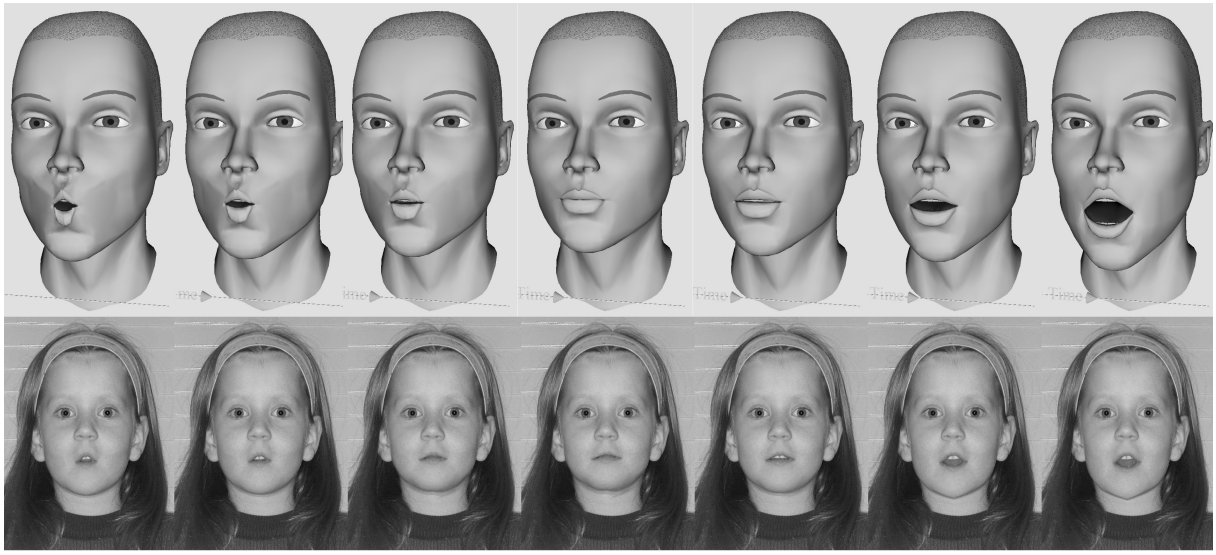


Figure 3: Image sequences of the utterance /oma/ generated by the 3D synthetic head (top) and the image based face model (bottom).

3.2.1. 3D synthetic head

MASSY's 3D face model is realized in VRML and uses the six motion parameters provided by the visual articulation module. The difference of the 3D model in neutral state to the model deformed to the maximum position of one articulator constitutes an articulator excursion. The difference vectors of all affected vertices besides the concerning vertex index are stored as a so called displacer. All possible articulator positions result from linear combinations of the vertex difference vectors contained in the displacers. A facial animation is generated from a sequence of motion parameter values.

3.2.2. Image based face model

The image based face model consists of an image database indexed by facial properties (instead of phonemes/visemes). These facial properties are a subset of the articulation parameters with respect to the visibility of articulators in the images. Currently these facial properties are implemented:

- the lip width and
- joint lip and jaw height and the lower lip retraction.

The image database can be built by deforming a reference image to fit the facial properties. For this procedure 37 feature points are defined in a reference image of a face. 27 of them correspond to feature points standardized in MPEG-4. Five additional feature points define a surrounding of the lower jaw area to prevent sharp edges when the lower jaw is displaced. Another five feature points mark the upper teeth to save them from being deformed or displaced.
Two displacement vectors (one per facial property) are assigned to each of the 37 feature points. These two displacement vectors are linearly combined – weighted with the magnitude of the facial property – before being applied to the feature point for deformation. The pixels of the face image are displaced using a bilinear interpolation between the combined displacement vectors of three feature points surrounding the pixel. Details of the algorithm can be found in Fagel (2004). Figure 4 shows the lower part of a reference image including the feature points, the triangle mesh built of them, and schematically the two displacement vectors for each feature point.

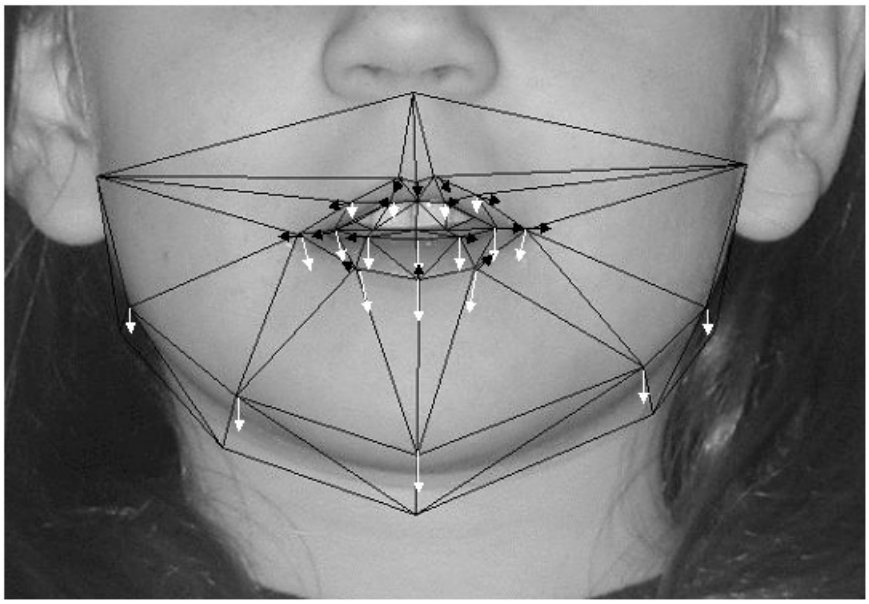


Figure 4: Lower part of a neutral face with the triangle mesh built of the feature points and – schematically – the two displacement vectors for each feature point. Black arrows show the displacements for lip spreading (which also leads to a lip slimming), white arrows show the displacements for vertical mouth opening.

A simple software tool for marking the feature points in an image by hand was developed. This software tool can also be used to mark displaced feature points in an image with one facial property different from neutral. The differences of these feature points and those in the reference image form the displacement vectors. Instead of defining displacement vectors for a new face, a predefined set of displacement vectors for each facial property can be used as a preset. These displacement vectors are scaled by the width of the lips in the reference image of the new face to fit the proportions.

Alternatively to the image deformation approach an example image database was created. A male speaker was videotaped and the frames were extracted. Outer lip height and width were annotated automatically by a lip feature extraction program. In case of duplicates images with upper lip position and vertical lip center near the average upper lip position and average vertical lip center were chosen. A more sophisticated criterion for similarity of images which constitute a database will follow. Figure 5 shows the frames selected from the database for the utterance /oma/ as well as the frames generated by deforming one image of the database with the method described above for the same utterance.

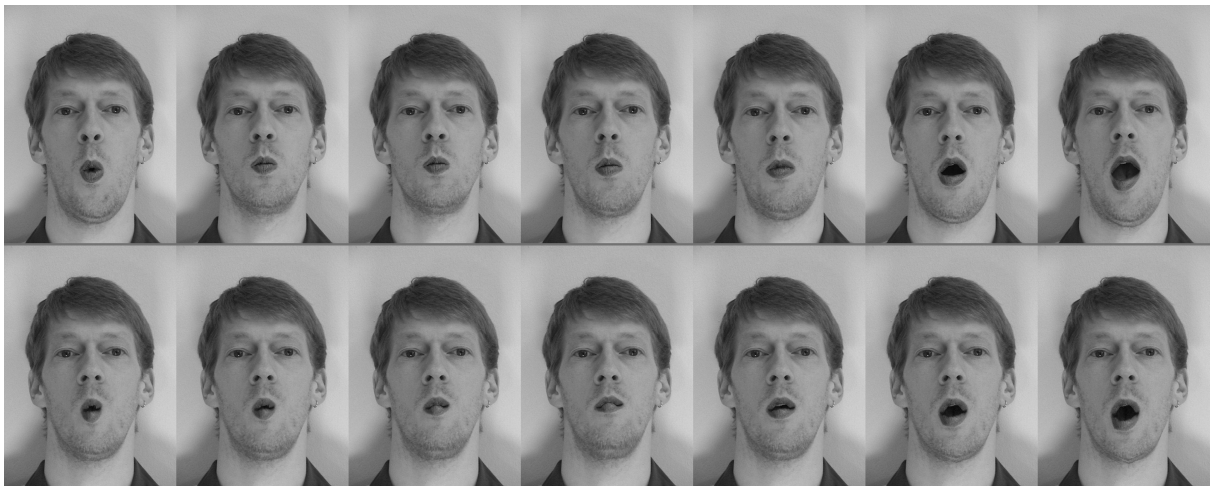


Figure 5: Frames of the utterance /oma/ selected from the database (top) and generated by deforming one image of the database (bottom).

4. Evaluation

4.1. 3D synthetic head

4.1.1. Method

A phonetically balanced rhyme test (Sendlmeier & v. Wedel, 1986) covering initial and final consonants and medial vowels was used in the first experiment for the evaluation of the 3D synthetic head. A trained female speaker uttered the items of the corpus and was videotaped. The recorded items were split and saved as single files. The audio channel was separated. All phones of the recorded items were labeled by hand and the fundamental frequency curves were extracted using the speech analysis software Praat. This information was handed to the face module as input. For each item the synthesizer generated an animated face and an audio file synchronous to the recorded natural utterances.
Both the synthetic head and the videotaped face were paired with both the synthesized and recorded voice. Synthetic and natural audio alone conditions were used as references. Several measures of audiovisual integration (Massaro, 1987, Braida, 1991, Grant & Seitz, 1998) also use visual alone conditions. These were included as well, resulting in a total of eight conditions: synthetic stimuli in audio alone, visual alone and audiovisual conditions (a_s, v_s, a_sv_s), natural stimuli in audio alone, visual alone and audiovisual conditions (a_n, v_n, a_nv_n) , and mixed audiovisual stimuli (a_nv_s, a_sv_n). All audio material was mixed with white noise at -6dB signal-to-noise ratio. The visual alone stimuli were presented without noise. 36 undergraduate students of communication science participated in the test voluntarily. Every subject was presented with the stimuli in a different pseudo-random order.

4.1.2. Results

An analysis of variance was carried out for all pairs of conditions. Table 2 shows the mean recognition scores where conditions producing non-significantly differing results are grouped ($p<0.05$). All audiovisual conditions resulted in higher recognition scores than all unimodal conditions (audio alone or visual alone respectively). Both natural unimodal conditions led to higher scores than the corresponding synthetic condition. In case of synthetic audio in bimodal condition the pairing with natural video showed better recognition scores than the pairing with synthetic video, but the scores reached with natural and synthetic video paired with natural audio did not differ significantly from each other.

Table 2: Mean recognition scores of the first experiment in %.

condition	subgroup				
	1	2	3	4	5
a_s	43.6				
v_s		48.3			
a_n			58.5		
v_n			60.8		
$a_s v_s$				67.2	
$a_n v_s$					75.3
$a_s v_n$					77.2
$a_n v_n$					79.6

4.2. Image based face model

4.2.1. Method

For the image based approach a simpler evaluation experiment was carried out. 12 of the most frequent German phones (Kohler, 1995) were chosen: /p, b, m, t, d, n, l, s, z, k, g, ŋ/. Additionally 3 of the most frequent vowels /a, ɪ, ʊ/ that nearly span the German vowel space were used to build all 36 items of the form VCV. These items were synthesized with identical phone durations and constant fundamental frequency. White noise at -4.5dB signal-to-noise ratio was added to the audio signals. Stimuli in the conditions audio alone (a), visual alone (v) and audiovisual (av) were generated resulting in 108 stimuli in total. All stimuli were presented in the same pseudo-random order to five students of communication science (two male, three female, 23-28 years, mean age 25.4, all

normal hearing and normal or corrected to normal vision). The subjects were asked to mark the answer for each stimulus among all possible 36 items.

4.2.2. Results

Vowel and consonant identification was analyzed separately, regarding segments as correctly identified if the chosen answer contained the right of the three vowels or the right of the 12 consonants, respectively. Table 3 shows the mean recognition scores for vowels and consonants in the conditions audio alone, visual alone, and audiovisual. An analysis of variance revealed that vowels and consonants in the audiovisual condition were significantly better recognized ($p<0.05$) than in the audio alone condition. Except for consonant identification in one subject (where the recognition was identical) all subjects reached higher recognition scores in audiovisual than in audio alone condition. The relative error reduction (audiovisual benefit, Sumby & Pollack, 1954) was 55% for vowels and 15% for consonants.

Table 3: Mean recognition scores for vowels and consonants in the second experiment in %.

condition	vowels	consonants
a	61.7	22.8
v	63.9	12.2
av	82.8	34.4

4.2.3. Discussion

The evaluation of the image based face model shows further interesting data which have to be confirmed in forthcoming experiments. The visual information is obviously integrated into the audiovisual perception although the visual alone identification of consonants (12%) is only slightly above chance level (8.3%). If the chance level is taken into account (Equation 1: chance level correction) a super-additive information usage can be seen (Table 4). The correct consonant identification above chance is 4.2% for video alone, 15.8% for audio alone, and 28.5% in the audiovisual condition which is more than the sum of audio+video. This super-additivity of speech perception was already observed by Saldaña & Pisoni (1996) in an audiovisual speech intelligibility test with sine-wave speech as audio signal. Furthermore it was implicitly described by Schwartz (2003) where a non-informative – if presented alone – video increased the distinction of voiced and unvoiced plosives when added to the audio signal.

$$R' = (R-C)/(1-C) \quad (1)$$

where R': chance level corrected recognition score,
R: recognition score, $0 \leq R \leq 1$
C: chance level, $0 \leq C \leq 1$,
here: $C=1/N$ with N: quantity of response alternatives.

Table 4: Chance level corrected recognition scores for vowels and consonants in the second experiment in %.

condition	vowels	consonants
a	42.6	15.8
v	45.9	4.2
av	74.3	28.5

5. Summary and future work

Both the 3D synthetic head and the image based speech synthesis enhance the intelligibility of audible speech. The visualization methods realize different levels of abstraction from natural static appearance and natural dynamics. With increasing abstraction the benefit of visual speech decreases (Benoit, 1996). There are some studies investigating the influence of spatial and temporal resolution (de Paula et al., 2000, Massaro, 1998) on the speech perception process. Knowledge in this area will help to design maximum intelligible talking heads at minimum system performance requirements and programming effort. Natural speech including shape and appearance (face topology and texture) and dynamics will be simulated with MASSY (methods for "speaker cloning" have been reported e.g. by Odisio & Bailly, 2003). Then the precision of the simulation will successively be reduced in order to determine the crucial synthesis properties.

The MODULAR AUDIOVISUAL SPEECH SYNTHESIZER combines parametric and image based approaches of speech visualization. In this way the visual speech output can be video-realistic but controlled by a specific articulation model not included in the image database. Co-articulation taking into account potentially all preceding and subsequent speech segments (instead of being limited to e.g. neighbors) becomes possible. The separation of articulation and visualization enables the synthesis of different speaking styles within one visual database. But an appropriate database is required for successful synthesis. Ideally all facial properties that are visible should be annotated to the images and not only speech

specific properties. So images with similar annotated facial properties look only marginally different from each other. This guarantees smooth transitions when images are concatenated. A database built of deformed versions of a reference image fulfils this requirement. When using databases consisting of naturally recorded material the recording conditions have to be constant (or an accordingly huge corpus has to be recorded) and more facial properties than the two described above have to be annotated to the material. An example database was automatically created by means of a lip feature extraction software. But similarities regarding facial features that are not yet annotated (e.g. eye closure) currently must be detected manually.

Acknowledgements

I thank my students – especially Åsa Wrange – for their active participation in my course on audiovisual speech perception. Parts of the present paper are based on their work.

References

Benoît, C., Abry, C., Cathiard, M., Guiard-Marigny, T. & Lallouache, T. (1995). Read my Lips: Where? How? When? And so ... What? In B. Bardy, R. Bootsma & Y. Guiard (eds.) *Poster Book of the 8th International Congress on Event Perception and Action*, Marseille.

Benoît, C. (1996). On the Production and the Perception of Audio-Visual Speech by Man and Machine. In H. Bertoni, Y. Wang & S. Panwar (eds.), *Multimedia and Video Coding*. Plenum Press, New York.

Beskow, J. (2003). Talking Heads – Models and Applications for Multimodal Speech Synthesis. PhD Thesis, Stockholm.

Boyce, S. E. (1990). Coarticulatory Organisation for Lip Rounding in Turkish and English. *Journal of the Acoustical Society of America*, 88: 2584-2595.

Braida, L. D. (1991). Crossmodal Integration in the Identification of Consonant Segments. *Quarterly Journal of Experimental Psychology*, 43, 647-677.

Brand, M. (1999). Voice Puppetry. *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, Los Angeles: 21-28.

Bregler, C., Covell, M. & Slaney, M. (1997). Video Rewrite: Driving Visual Speech with Audio. *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, Los Angeles: 353-360.

de Paula, H. B.,Yehia, H. C., Shiller, D., Jozan, G., Munhall, K. & Vatikiotis-Bateson, E. (2003). Linking Production and Perception Through Spatial and Temporal Filtering of Visible Speech Information. *Proceedings of the 6th International Seminar on Speech Production*, Sydney: 37-42.

Erber, N. P. (1969). Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli. *Journal of Speech and Hearing Research*, 12: 423-425.

Ezzat, T. & Poggio, T. (2000). Visual Speech Synthesis by Morphing Visemes. *International Journal of Computer Vision*, 38: 45-57.

Ezzat, T., Geiger, G. & Poggio, T. (2002). Trainable Videorealistic Speech Animation. *Proceedings of ACM SIGGRAPH*, San Antonio: 388-398.

Fagel, S. & Sendlmeier, W. F. (2003). An Expandable Web-based Audiovisual Text-to-Speech Synthesis System. *Proceedings of the 8th EUROSPEECH European Conference on Speech Communication and Technology*, Geneva: 2449-2452.

Fagel, S. (2004). Video-realistic synthetic speech with a parametric visual speech synthesizer. *Proceedings of the INTERSPEECH*, Korea.

Fagel, S. (2004a). Audiovisuelle Sprachsynthese – Systementwicklung und -bewertung. Logos Verlag, Berlin.

Fagel, S. & Clemens, C. (2004). An Articulation Model for Audiovisual Speech Synthesis – Determination, Adjustment, Evaluation. *Speech Communication*, 44: 141-154.

Grant, K. W. & Seitz, P. F. (1998). Measures of Auditory-Visual Integration in Nonsense Syllables and Sentences. *Journal of the Acoustical Society of America*, 104: 2438-2450.Löfqvist, A. (1990). Speech as Audible Gestures. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modeling*. Kluwer Academic Publishers, Dodrecht.

Kohler, K. J. (1995). Einführung in die Phonetik des Deutschen. Schmidt, Berlin, 1995.

Massaro, D. W. (1987). Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Erlbaum, London.

Massaro, D. W. (1998). Illusions and Issues in Bimodal Speech Perception. *Proceedings of Audiovisual Speech Processing*, Sydney: 21-26.

McGurk, H. & MacDonald, I. (1976). Hearing Lips and Seeing Voices. *Nature*, 264: 746-748.

Minnis, S. & Breen, A. (2000). Modeling Visual Coarticulation in Synthetic Talking Heads Using a Lip Motion Unit Inventory with Concatenative Synthesis. *International Conference on Spoken Language Processing*, Beijing: 759-762.

Odisio, M. & Bailly, G. (2003). Shape and appearance models of talking faces for model-based tracking. *Proceedings of Audiovisual Speech Processing*, St. Jorioz: 105-110.

Sumby, W. H. & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, 26: 212-215.

Saldaña, H. & Pisoni, D. (1996). Audio-Visual speech perception without speech cues. *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia: 2187-2190.

Schwartz, J.-L., Berthommier, F. & Savariaux, C. (2002). Audio-Visual Scene Analysis: Evidence for a "Very-Early" Integration Process in Audio-Visual Speech Perception. *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver: 1937-1940.

Sendlmeier, W. F. & v. Wedel, H. (1986). Ein Verfahren zur Messung von Fehlleistungen beim Sprachverstehen – Überlegungen und erste Ergebnisse. *Sprach, Stimme, Gehör*, 10: 164-169.

Anatomical structures involved in non-human vocalization

Alban Gebler
Roland Frey
Leibniz Institute for Zoo and Wildlife Research (IZW), Berlin, Germany

In order to understand the functional morphology of the human voice producing system, we are in need of data on the vocal tract anatomy of other mammalian species. The larynges and vocal tracts of four species of Artiodactyla were investigated in combination with acoustic analyses of their respective calls. Different evolutionary specializations of laryngeal characters may lead to similar effects on sound production. In the investigated species, such specializations are: the elongation and mass increase of the vocal folds, the volume increase of the laryngeal vestibulum by an enlarged thyroid cartilage and the formation of laryngeal ventricles. Both the elongation of the vocal folds and the increase of the oscillating masses lower the fundamental frequency. The influence of an increased volume of the laryngeal vestibulum on sound production remains unclear. The anatomical and acoustic results are presented together with considerations about the habitats and the mating systems of the respective species.

1. Introduction

Although the human ability for speech is the most prominent example, acoustic communication has also evolved in many other species. Apart from intellectual requirements, the anatomy of the human larynx and vocal tract seems to be highly adapted to produce a huge variety of sounds. But what about adaptations in other vocalizing species?

The debate on a lowered larynx as a precondition for modern human speech (Lieberman & Crelin 1971) is an example of how the knowledge of animal vocal tract morphology influences the understanding of the evolution of the human vocal apparatus. A constant or temporary lowered larynx position has been documented in an increasing number of species. Thus, it becomes more and more implausible that larynx descent itself can be used as a criterion for the

ability to articulated speech (cf. Fitch & Reby 2001, Boe et al. 2002, Nishimura et al. 2003).
In contrast to human speech, the acoustic communication system of many animal species comprises only few more or less fixed call types. The acoustic parameters of these calls, e.g. the amplitude of sound pressure, the fundamental frequency and the spectral distribution of energy within the frequency range, are adapted to the acoustical properties of the species' habitat. Additionally, the emission of specific calls as one component of mating behaviour can be crucial for the mating success in many animal species. Despite the basic similarity of the vocal apparatus in mammals, the great diversity of habitats and social systems has led to the evolution of an accordingly great variety of vocal tract specializations. Unfortunately, our knowledge of the head and neck anatomy of almost all wild living species is sparse. In particular, details of the vocal tract morphology are still lacking in many species.

Figure 1: Investigated species (all males):
Muntiacus reevesi (top left), *Procapra gutturosa* (top right),
Budorcas taxicolor (below left) and *Ovibos moschatus* (below right)
© Zoo Köln (*M. reevesi*, *O. moschatus*), H. Mix (*P. gutturosa*), A. Michailov (*B. taxicolor*)

This contribution presents morphological studies of the vocal tract anatomy of the Chinese muntjac (*Muntiacus reevesi*), the Mongolian gazelle (*Procapra gutturosa*), the muskox (*Ovibos moschatus*), and the takin (*Budorcas taxicolor*). All of these species (Fig. 1) belong to the Artiodactyla, i.e. even-toed hoofed

mammals. Vocalizations play an important role in the social systems of all four species. Practically, the investigation of the species was restricted by the difficulty of obtaining specimens and acoustic recordings of respective vocalizations.

Muntiacus reevesi is a small deer of approximately 15 kg body mass. *M. reevesi* is originally distributed in dense subtropical forests of Southern China and on the island of Taiwan. The individuals live mostly solitary and territorial. Normally, the home ranges of males exclude each other but include parts of the home ranges of several females. The mating system of *M. reevesi* is promiscuous and copulations can occur all over the year. Apart from acoustic communication, the olfactory communication plays an important role in the life of *M. reevesi*.

The body mass of *Procapra gutturosa* is approximately 30 kg in the male and 25 kg in the female. The habitats are open arid steppes and semi-desert regions in China, Mongolia and the Altai region of Russia. *P. gutturosa* is a nomadic species except during the rut. In the mating season dominant adult males occupy individual territories in which they retain several females, thus forming a harem. This behaviour is characteristic of a polygynous mating system.

The body mass of the arctic *Ovibos moschatus* is about 350 kg for adult males and one third less for females. Today, the animals inhabit the arctic tundras along the polar circle in the North of Canada and the North and Northeast of Greenland. The polygynous *O. moschatus* lives in mixed herds throughout most of the year. During the rut, a temporary harem-like association of females is formed and defended by a dominant male.

The male *Budorcas taxicolor* has a mean body mass of 300 kg, the female of 200 kg. Three subspecies are distributed in disjunctive mountain areas in Assam, Bhutan, Burma and China (including Tibet). The natural habitat of *B. taxicolor* ranges from grass covered alpine rocks to dense bamboo forests. In winter the animals live in small groups which unite to form larger herds during the summer months before the rut. Details of the mating system are unknown yet.

2. Material and Methods

The left halves of frozen heads and necks of *Muntiacus reevesi*, *Ovibos moschatus*, *Budorcas taxicolor*, and *Procapra gutturosa* were macroscopically dissected while the specimen were submerged in cold water. Consecutive dissection steps were documented by series of analogue photographic slides (Nikon F3) and also by digital images (Nikon CoolPix 950, Minolta DImage 7). Major stages were recorded in proportional drawings. The measurement of a string arranged along the vocal tract, i.e. from the vocal cleft to the upper lip, yielded the length of the vocal tract.

In addition, heads of *M. reevesi*, of *O. moschatus* and of *P. gutturosa* were scanned by means of computer tomography (GE Lightspeed 4-Slice Spiral Computertomograph). Slice thickness was 0.6 mm.
The vocalizations of *O. moschatus* and *P. gutturosa* were recorded be means of a DAT (digital audio tape) recorder (Sony TCD-D 100) and an appropriate directional monophonic microphone with wind cover (Sennheiser ME 80, including preamplifier K3U), whereas for the recordings of *B. taxicolor* the same recorder and a stereophonic microphone with overlapping kidney-shaped directionalities (Sony ECM-MS 907) were used. Additionally, we used earlier analogous audiotape recordings of *M. reevesi* made by unknown authors from a small herd formerly kept at the University of Bielefeld. Spectrograms were calculated by means of the program 'praat' (P. Boersma & D. Weenink, University of Amsterdam, Netherlands; analysis width: 100 ms, time steps: 10^3, frequency steps: 10^3, Gaussian window).

3. Results and Discussion

*3.1. The Chinese muntjac (*Muntiacus reevesi*)*

The larynx of *M. reevesi* may be taken as morphologically unspecialized (Fig. 2). It lacks any conspicuous enlargement of the laryngeal cartilages and any further specializations such as ventricles or air sacs. Significant differences between the larynges of both sexes of *M. reevesi* were not found. The membraneous vocal folds extend between the thyroid cartilage and the vocal process of the arytenoid cartilage. Thyroid and cricoid cartilage surround the laryngeal cavity.
The vocal repertoire of *M. reevesi* comprises miscellaneous types of calls from whimpers to barks with a broad overall frequency range. The best call type to demonstrate the filter impact of the vocal tract is the harmonic ‘cheep’ call (Fig. 3). Another call type is the noisy ‘barking’ which is used mainly as an alarm call. Only the latter can pass through the dense vegetation of *M. reevesi*’s natural habitat for longer distances.

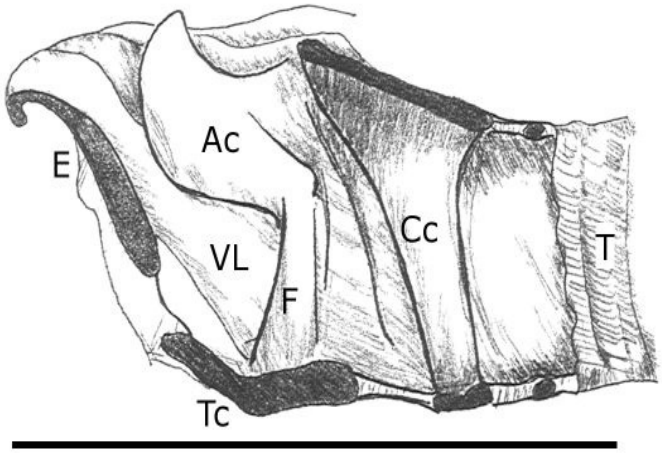


Figure 2: Larynx of *Muntiacus reevesi*
Ac arytenoid cartilage, Cc cricoid cartilage, E epiglottis,
F right vocal fold, T trachea, Tc thyroid cartilage,
VL laryngeal vestibulum. Scale bar = 5 cm.

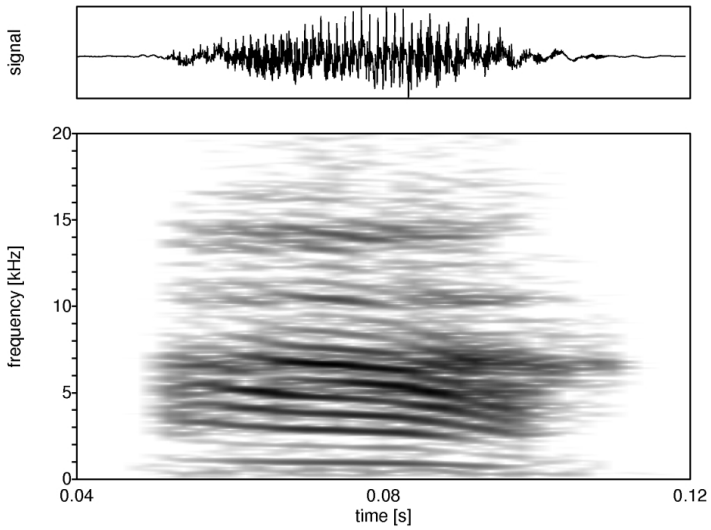


Figure 3: Narrow band spectrogram
of *Muntiacus reevesi*'s cheep call (adult, unknown sex)

*3.2. The Mongolian gazelle (***Procapra gutturosa***)*

In this species, sexual selection has led to a remarkable dimorphism of the larynx that is greater than expected by the differences of body masses. All cartilages of the male's larynx are enlarged. The thyroid bulges the ventral neck region (Fig. 4).

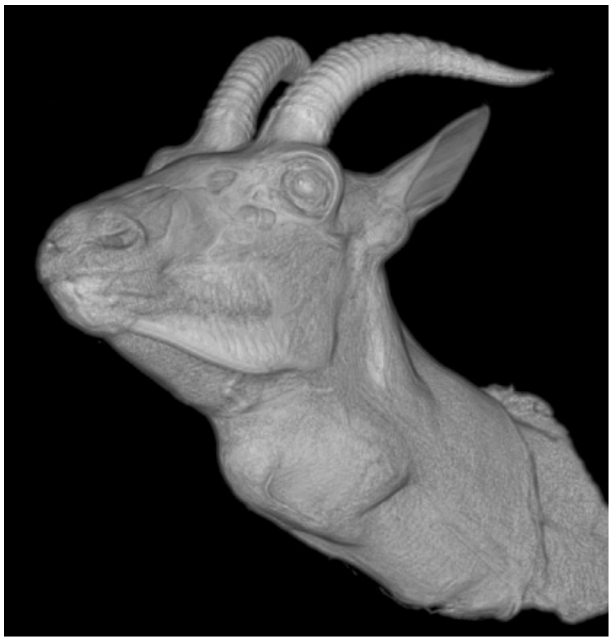


Figure 4: Head reconstruction of adult male *Procapra gutturosa* without its hair coat, based on data from computer tomography

The arytenoid cartilages, together with the epiglottis, form an exceptionally large entrance to the larynx. Paired lateral ventricles are located between the arytenoid cartilage and the thyroarytenoid muscle. The vocal folds of the male *P. gutturosa* are supported by large fibroelastic pads (Fig. 5).

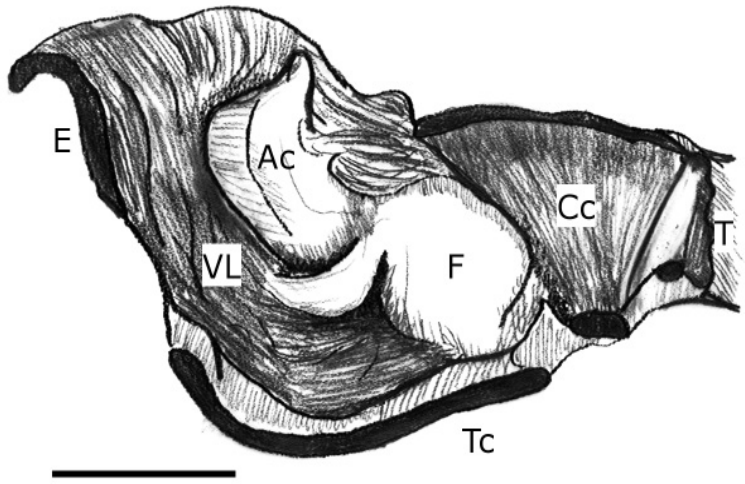


Figure 5: Larynx of *Procapra gutturosa*
Ac arytenoid cartilage, Cc cricoid cartilage, E epiglottis, F right vocal fold including fibroelastic pad, T trachea, Tc thyroid cartilage, VL laryngeal vestibulum.
Scale bar = 5 cm.

Compared to the call of the female which has a fundamental frequency of about 600 Hz, the specialized larynx of the male reduces the fundamental frequency

by about 100 Hz (Fig. 6). In the open habitat of *P. gutturosa* a call with a lower pitch has a more homogeneous directivity. This may be an adaptation suited for multidirectional advertisement calls during the rut.

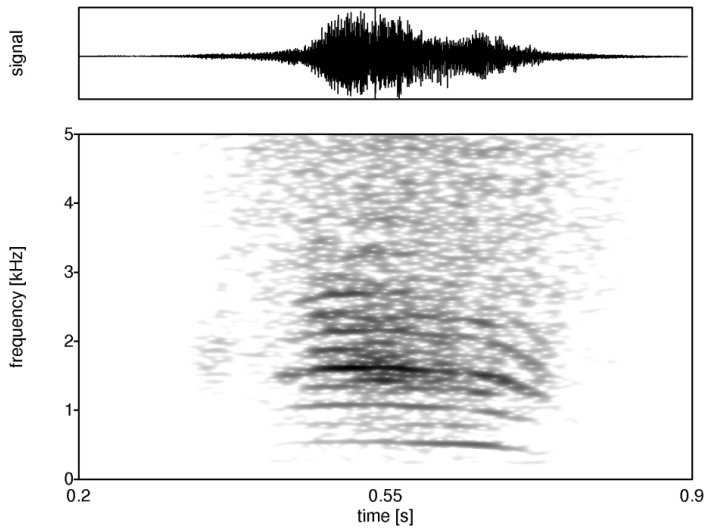


Figure 6: Narrow band spectrogram
of a call of *Procapra gutturosa* with harmonic and
subharmonic structures (adult male)

*3.3. The muskox (*Ovibos moschatus*)*

The male of *O. moschatus* has a small unpaired ventrorostral ventricle which extends into the extralaryngeal space (Fig. 7).

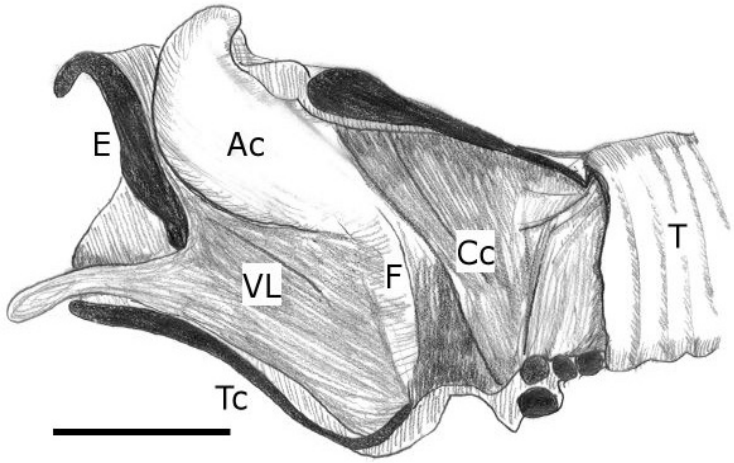


Figure 7: Larynx of a male *Ovibos moschatus*
Ac arytenoid cartilage, Cc cricoid cartilage, E epiglottis,
F right vocal fold, T trachea, Tc thyroid cartilage,
VL laryngeal vestibulum. Scale bar = 5 cm.

This specialization may contribute to producing high amplitude calls with low fundamental frequencies. Both sexes achieve low frequencies of around 120 Hz (Fig. 8).

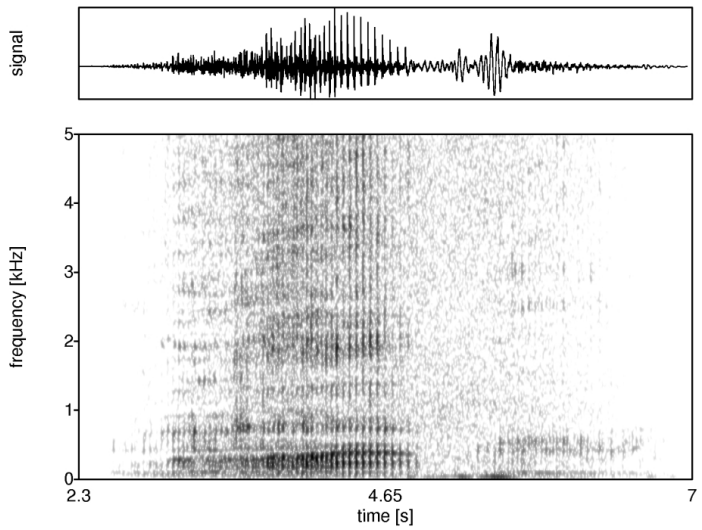


Figure 8: Narrow band spectrogram of *Ovibos moschatus*'s call with an additional pulse modulation (adult male)

Furthermore, the monotonous calls are characterized by their pulsed structure. The pulse rate averages 20 Hz.
Low frequency parts in the glottal signal together with the ability to produce lower formants by vocal tract filtering might be important for the male's mating success within a polygynous mating system. For both sexes, low frequencies may be advantageous for the intra-group short-range communication under windy weather conditions predominating in their natural habitat.

*3.4. The takin (*Budorcas taxicolor*)*

Caudoventrally, the thyroid cartilage of both sexes of *B. taxicolor* forms a voluminous hollow structure (Fig. 9). This thyroid bulla is partially filled by extensions of the conspicuously enlarged vocal folds. As a consequence, the frequencies of vocalization decrease because of the mass increase of the vocal folds. This may be understood as an adaptation to communicate with conspecifics in a habitat covered with dense vegetation that attenuates higher frequencies.

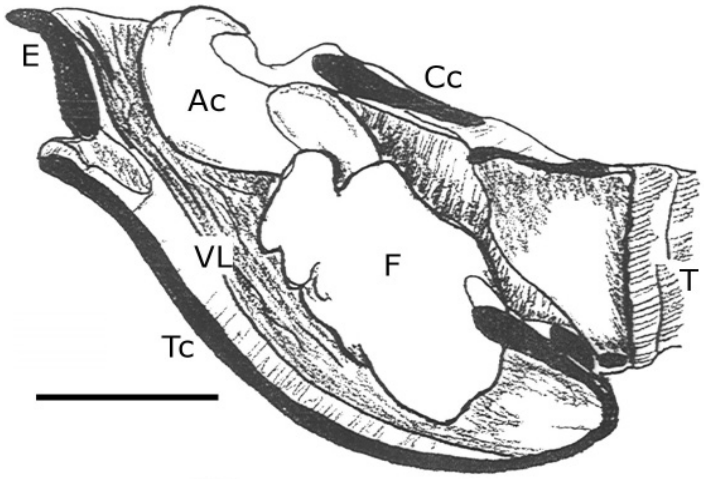


Figure 9: Larynx of *Budorcas taxicolor*
Ac arytenoid cartilage, Cc cricoid cartilage, E epiglottis, F enlarged right vocal fold, T trachea, Tc thyroid cartilage, VL laryngeal vestibulum. Scale bar = 5 cm.

As in *O. moschatus*, the vocalization of *B. taxicolor* is also pulsed with a rate of about 20 Hz (Fig. 10). Lowest prominent frequencies are around 210 Hz.

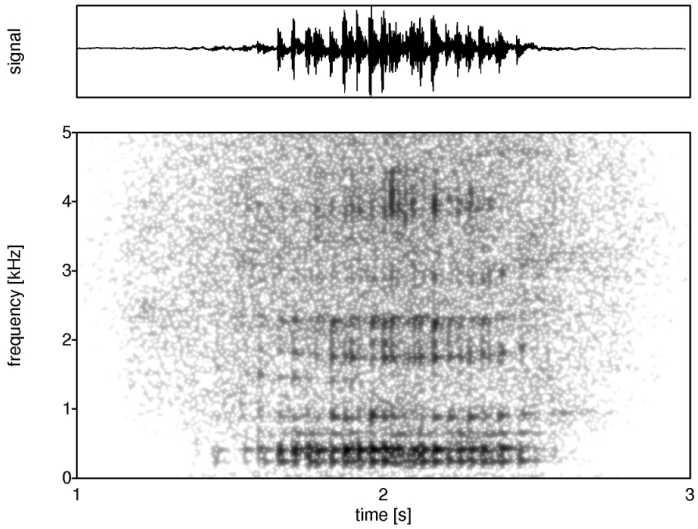


Figure 10: Narrow band spectrogram
of a pulse modulated call of *Budorcas taxicolor*
(adult male)

4. Conclusions

Tab. 1 summarizes the results. As the morphology of sound producing structures is a key to the interpretation of acoustic data, one species with less specialized and three species with more specialized larynges were investigated. In the latter three examples, the laryngeal characters vary in the following parameters:

1. the lengths of the vocal folds
2. the masses of the vocal folds
3. the volume of the laryngeal air spaces,

i.e. the combined volumes of the laryngeal vestibulum and additional laryngeal ventricles

Evolutionary elongation and mass increase of the vocal folds led to a lowering of the fundamental frequency. The low frequency calls of the investigated species evolved as a consequence of, mainly, sexual (*Procapra gutturosa*) or natural (*Budorcas taxicolor*) selection, or both (*Ovibos moschatus*).

Table 1: Comparison between investigated species demonstrating differences of laryngeal anatomy, of fundamental frequencies, of acoustic properties of the natural habitats and of the mating systems

Species	*Muntiacus reevesi*	*Procapra gutturosa*	*Ovibos moschatus*	*Budorcas taxicolor*
Sexual dimorphism of larynges	no	pronounced	size (proportional to body mass?)	size (proportional to body mass?)
Laryngeal anatomy	both sexes: unspecialised	male: enlarged vocal folds with fibroelastic pads, paired lateral ventricles	both sexes: ventrorostral ventricle, enlarged vestibulum	both sexes: large thyroid bulla, enlarged vocal folds
Vocal tract length (male/female) [mm]	130 / 140	370 / ?	470 / ?	470 / ?
Fundamental frequency f_0 (male/female) [Hz]	1000 / 1000	500 / 600	90 / 95	210 / 210
Acoustic properties of habitat	attenuation of most call frequencies	attenuation of higher call frequencies	unknown (long distance communication might be not important)	attenuation of higher call frequencies
Mating system	promiscuous	polygynous	polygynous	unknown

Although laryngeal ventricles were evolved independently in several taxonomic groups, their function is still unclear. Laryngeal ventricles might act as additional resonator devices which modify the acoustic filter function of the

upper respiratory tract. Supporting evidence comes from the investigation of the human piriform recess (Dang & Honda 1997).

References

Boe, L.J., Heim, J.L., Honda, K. and Maeda, S. (2002): The potential Neanderthal vowel space was as large as that of modern humans. *Journal of Phonetics* 30: 465-484.

Dang, J. & Honda, K. (1997): Acoustic characteristics of the piriform fossa in models and humans. *Journal of the Acoustic Society of America* 101 (1): 456-465.

Fitch, T. & Reby, D. (2001): The descended larynx is not uniquely human. *Proceedings of the Royal Society B* 268: 1669-1675.

Frey, R. & Gebler, A. (2003): The highly specialized vocal tract of the male Mongolian gazelle (*Procapra gutturosa* Pallas, 1777 - Mammalia, Bovidae). *Journal of Anatomy* 203: 451-471.

Frey, R., Gebler, A. & Fritsch, G. (2005): Arctic roars – laryngeal anatomy and vocalization of the muskox (*Ovibos moschatus* Zimmermann, 1780, Bovidae). *Journal of Zoology* (accepted)

Frey, R. & Hofmann, R.R. (2000): Larynx and vocalization of the Takin (*Budorcas taxicolor* Hodgson, 1850 - Mammalia, Bovidae). *Zoologischer Anzeiger* 239: 197-214.

Gebler, A. (2003): The Vocal Tract Anatomy of the Chinese Muntjac (*Muntiacus reevesi* Ogilby, 1839). *Proceedings of the First International Conference on Acoustic Communication by Animals*, College Park, Maryland: 93.

Lieberman, P. & Crelin, E. S. (1971): On the speech of Neanderthal man. *Linguistic Inquiry* 2: 203-222.

Nishimura, T., Mikami, A., Suzuki, J. & Matsuzawa, T. (2003): Descent of the larynx in chimpanzee infants. *Proceedings of the National Academy of Science of the United Nations of America* 100 (12): 6930-6933.

Acoustic, kinematic and aerodynamic aspects of word-initial and word-final vowels in pre-boundary context in French

Cédric Gendrot
LPP (Laboratoire de Phonétique et de Phonologie). Université Paris3 Sorbonne Nouvelle – ILPGA, France. CNRS UMR 7018.

Four speakers repeated 8 times 15 sentences containing 'pVp' syllables (V being /a/, /i/ and /u/). The 'pVp' syllables were located in final, penultimate and antepenultimate position relatively to the Intonational Phrase (IP) boundary. They were embedded in lexical words of 1-3 syllables and were either word-initial or word-final. Results show that the closer the vowel in word-final position is to the IP boundary, the longer the duration and the higher the fundamental frequency of the vowel; it is also characterised by larger lip opening gestures. The potential reduction or coarticulation of vowels in word-initial position compared to their counterparts in word-final position is discussed.

1. Introduction

1.1. Articulatory prosody

This study is located within the field of *articulatory prosody*, i.e. the influence of prosody on the articulation of speech sounds. In this framework the most obvious variations are linked to the boundaries (at the edges of prosodic domains, in the immediately preceding and following context) and to prominences in numerous languages. Segments around the boundaries and in prominent position are characterised by a strengthening of their articulatory properties (Fougeron 1998, Keating et al. 2003, Cho 2001). The term 'strengthening' may be defined here as an increase of spatial and temporal dimensions, such as for example consonants will be articulated with more extreme and longer constrictions; vowels will be articulated reaching their

respective targets and staying there for a longer duration. This prosodically conditioned strengthening is usually considered as linguistically significant, insofar as it can lead to increased linguistic contrast between segments. This is true on a syntagmatic scale (increased contrast between neighbouring segments) as well as on a paradigmatic scale (increased difference between contrastive phonemes in the language sound system). For example more jaw opening for enhancing the vocalic character of a vowel may generate a louder sound, thus making it more distinct from the surrounding consonants. This can also be true for closed vowels such as /i/ and /u/ (Erickson 1998, Harrington et al 2000). In the latter study, the authors reported a possible conflict and argued that a higher and fronter tongue position for /i/ gave support for a more peripheral vowel in its articulatory and acoustic aspects, while a lower jaw position gave support for the increased sonority hypothesis (cf. section 1.4 for a quick presentation of both hypotheses).

1.2. Prosody in French

In French, a word or a group of words which constitute a 'meaning unit' or "groupe de sens" (Grammont 1933) tend to form a single acoustic unit and the last syllable of the meaning unit carry a so-called final, primary or logical accent. This group accent has a demarcative function, namely it signals the end of a group and is thus correlated with the occurrence of a boundary.
The first prosodic position we will deal with here is the major boundary in French: the so-called "continuation majeure" in Delattre's terms (1966). In read isolated sentences, it is usually found at the juncture between the Noun Phrase (NP) and the Verb Phrase (VP) (cf. Fig. 1). The "continuation majeure" is acoustically cued by a continuation rise (H% in French ToBI notation by Jun & Fougeron (1997)), and by an extra lengthening of the last syllable before the juncture, as compared to the other sense-groups in the sentence.

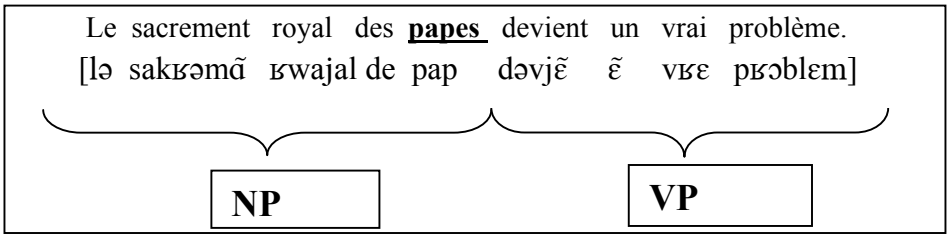


Figure 1: Example of major boundary occurring after the lexical word 'papes' [pap]. Translation: Royal sacrament of popes is becoming a real problem.

It is also hypothesised to be realised with greater articulatory gestures: consonants are being produced with a stronger closing gesture and vowels with a greater opening (Straka 1963).
The second position of interest in this work is the word initial position in French. It may be a specific position for what is called the "word initial accent". The word initial accent's most puzzling characteristic is its optional realisation as its presence and its exact position may be conditioned by different factors such as rhythm (Pasdeloup 1990), speaker identity (Vaissière 1975) or style (Vaissière 1975, Lucci 1983). It is acoustically signalled by a f0 rise, which is timed with the very beginning of the word, but may end on one of the first syllables, generally the first or the second. We will globally consider the word initial position in this study, for which no study has yet provided an articulatory description in French.

1.3. The π-gesture activity

Byrd and colleagues (Byrd et al. 2000, Saltzman & Byrd 2000b, Byrd & Saltzman 2003), after Beckman, Edwards and colleagues (1992), introduced the *π-gesture (prosodic-gesture)* into the Articulatory Phonology framework.

> *"We expand on this concept [Articulatory Phonology] by proposing that prosodic (non-constriction-based) gestures also exist, occuring at domain edges. We refer to such prosodic boundary gestures as π-gestures. Like articulatory gestures, π-gestures have an inherent temporal extent. These π-gestures are hypothesised to cause slowing at phrase edges by affecting the ongoing stiffness parameter values of all constriction gestures that are active within the π-gesture's activation interval. Stronger π-gestures slow the local speaking rate more (i.e. lower stiffness more) than weaker π-gestures. The prosodic boundary strength defines the activation level of a π-gesture."* (Byrd 2000: 13).

These π-gestures will alter constriction gestures during a temporal interval which varies according to the boundary strength. Its activation effect is also determined by boundary strength. According to this model articulatory gestures are slowed down - and possibly enlarged due to less overlapping according to their simulations (Byrd & Saltzman 2003) - when in close contact to a boundary, and this slowing effect is attenuated when going further from this boundary. The consequence of such a model, allowing to rely upon dynamic prosody and aspects of articulatory gestures, has not been tested for French as far as we

know, and the hypothesised slowing and enlarging of articulatory gestures will be tested in this work in pre-boundary context.

1.4. Hypothesis

Our hypothesis is that the closest vowels to the boundary in word-final position will be realised with longer and larger gestures compared to their counterparts further away from the boundaries in word-initial position. Thus we will consider the span of the activation of the π-gesture, or more generally if the results are coherent with the π-gesture hypothesis.

Our aim is to characterise the strengthening hypothesised to occur on the last syllable in immediate pre-boundary context (Tabain 2003a,b). Indeed the closest the vowel is to the boundary, the most strengthened it should be, compared to the same vowel further away from the boundary. Two hypotheses concerning the nature of the strengthening are found in the literature:
- Sonority Expansion Hypothesis: Edwards & Beckman (1988), Beckman & Edwards (1992) following Straka's hypothesis (1963) state that

> *" the lengthening associated with accented sequence, on the other hand, is a head effect. [...] That is, the purpose of the lengthening seems to be not to make the sequence longer per se, but rather to expand the portion of the syllable where maximum energy is radiating out of the mouth."* (Beckman & Edwards, 1992: 369).

They predict that the opening contrast between the accented vowel and the following consonant is maximised, thus participating in a greater syntagmatic highlighting of the syllabic nucleus under accent.

- De Jong (1995) proposes a slightly different hypothesis according to which the distinctive features of vowels should be increased, allowing a paradigmatic contrast relying on phonemic distinction between segments occurring on a same syntagmatic position or between lexical units. This hypothesis is inspired by the works of Lindblom (1990), relying more specifically on hyper-articulation of syllables.

2. Method

2.1. Measurements

A set of experimental methods simultaneously recorded (cf. figure 2) will be used in order to allow us to evaluate the activity of the vocal apparatus at different levels:
- ***The sub-glottic effort***. Subglottal Pressure (Ps) tends to be constant along the sentences, but Ps increase has been observed in sentence stress/emphasised syllables (Ladefoged 1958, Benguerel 1970). While direct Ps measurement is invasive, intra-oral pressure (PIO) is a non invasive method used to estimate the mean subglottal pressure for a vowel for adjacent unvoiced stop sounds alternating with vowel phonations (Hertegard 1995). The method is based on the fact that the pressure is equalised below and above the glottis during the closure period of unvoiced stop consonants when vocal folds are abducted. The closed syllable 'pVp' is used in our study with the assumption that there will not be much change in Ps between the vowel and the surrounding consonants. These data were collected with the FJ-electronics Aerophone II device at 1 KHz and then resampled at 25 KHz.

- ***The glottic activity***. f0 was measured from Electro-Glotto-Graphy (EGG) signal using a Laryngograph Ltd. In modal voice register, an increase of f0 mainly corresponds to an increase in crico-thyroid and thyro-arytenoid activity. The EGG as well as the acoustic signal was collected with DIANA physiological station at 25 KHz.

- ***The supraglottic articulators.*** From video-camera recordings (Digital Camera Canon XM2 at 34 frames/second). The lip aperture is measured manually using Matlab. **Lip aperture movement** is measured as the maximum amplitude of opening between the maximally closed position of the first /p/ and the maximally closed position of the second /p/ (e.g. „pap").
Data curves were obtained from these measurements. These curves were first synchronised to the acoustic signal using a second acoustic signal from the camera and then up-sampled at 25 KHz.

- ***Acoustic measurements*** (F1 and F2 formant analysis, duration) completed the different physiological values recorded for this corpus. A manual segmentation was realised so as to measure the duration of analysed vowels. Praat scripts were used to get these acoustic values (Gendrot & Adda 2004). Formant values were systematically checked however as closed vowels such as /u/ and /i/ are sometimes partly unvoiced and thus more difficult to analyse. Moreover the first

two formants of /u/ being close to each other can bring some difficulties for automatic formant detection systems.

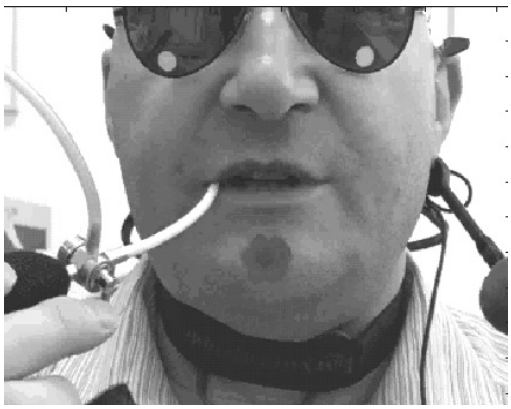


Figure 2: Example of speaker M1 with multiple device (EGG, intra-oral pressure, video, microphone)

2.2. Corpus and data collection

The vowels (/a/, /i/ and /u/) were embedded in a 'pVp' syllable context. Each syllable was tested in 5 different contexts (thus 5 different sentences) developed below. The heading we shall use afterwards to refer to these conditions is given between square brackets:

1. last syllable of a lexical word in a pre boundary context [word final]
   a. monosyllabic word [1syll_fin]
   b. bisyllabic word [2syll_fin]
2. first syllable of a lexical word in pre-boundary context [word initial]
   a. bisyllabic word [2syll_ini]
   b. trisyllabic word [3syll_ini]
3. last syllable of a bisyllabic lexical word followed by a monosyllabic word in a pre-boundary context. The first lexical word (soupapes neuves) should still be characterised by word lengthening.

In short there were altogether 3 vowels * 5 different sentences (conditions) = 15 sentences each repeated 8 times during the recording.

Table 1: Sentences used in the corpus for vowel /a/. Studied syllables are in bold and underlined ; σ: syllables ; [ ]: word ; #: boundary

1a	[**σ**] #	Le sacrement royal des **papes** [pap] devient un vrai problème. Royal sacrament of popes is becoming a real problem.
1b	[σ **σ**] #	La mise en rayon des sou**papes** [supap] devrait vite s'arrêter. The shelving of valves should stop soon.
2a	[**σ** σ] #	La dernière collection des **pap**erasses [papʁas] a lieu le 20 juin. The last collection of papers will take place on june the 20th.
2b	[**σ** σ σ] #	La mise en hypothèque des **pap**erasseries [papʁasʁi] m'énerve beaucoup. The mortgage of paperwork makes me really angry.
3	[σ **σ**] [σ] #	La mise en rayon des sou**papes** neuves [supapnœv] devrait vite finir. The shelving of new valves should stop soon.

The last condition (3) is considered as the reference. The examples were chosen so as to make sure the two words could be said in a straight way by the speakers. The syllable [pap] is then located before a word boundary and should undergo word final lengthening, although fairly small as compared to that of higher boundaries (Di Cristo 1998 and references within).
According to our hypothesis, the vowels in condition 1a and 1b should be more strengthened than in condition 2a and 2b respectively. Both vowels in condition 3 and condition 2a are one syllable away from the boundary and in this regard could show similar results. However the tendency for word final lengthening in French could explain longer vowel duration (together with other differences) in condition 3 compared to 2a.

2.3. Procedure

Four speakers were recorded in a quiet room of HEGP's hospital, 2 male speakers: M1 (39 years-old) and M2 (30 years-old); 2 female speakers: F1 (26 years-old) and F2 (25 years-old).
All of them have linguistic knowledge and have no identifiable accent. The author was conscious that utterances realised by our different speakers would not necessarily be representative of the desired prosodic conditions, i.e. Intonational Phrase (IP) boundaries. The author who guided the recording sessions made sure that the sentences in the corpus were not read too fast, but did not in any way induced the speaker to produce the expected realisations.
Following the same methodology, word initial positions could be realised as word initial accents according to the speaker's will (cf. section 1.2). The word initial accents were selected according to f0 values that were significantly higher than each speaker's mean values in the same condition. These outliers (23

altogether) were discarded from further analysis. The reference condition (e.g. “soupapes neuves”) could also be differently realised according to the speaker’s will. The author made sure that no explicit pause would occur between the two lexical words during the experiment. In that case the speakers were asked to repeat after the explanation that a pause in this location would be unnatural.

3. Data analysis

3.1. Duration

Syllables in final position were expected to be longer than syllables in non-final position. A longer duration of the monosyllabic (condition 1a) as compared to the disyllabic (condition 1b) was also expected: the longer the words, the shorter the syllables should be (Nooteboom 1997).

Table 2: ANOVAs'[1] results for f0 and duration values for each speaker and vowel according to the condition factor.

	A		I		u	
All Subjects						
duration	F(4,141)=7.422	**	F(4,148)=5.35	*	F(4,158)=4.67	*
f0	F(4,134)=2.69	*	F(4,145)=3.14	*	F(4,154)=5.95	*
Subject M1						
duration	F(4,33)=21.942	**	F(4,35)=13.59	**	F(4,34)=20.4	**
f0	F(4,26)=57.57	**	F(4,35)=50.52	**	F(4,34)=45.76	**
Subject M2						
duration	F(4,30)=7.26	*	F(4,29)=1.07	-	F(4,31)=3.3	*
f0	F(4,30)=13.85	**	F(4,29)=4.68	*	F(4,31)=47.52	**
Subject F1						
duration	F(4,32)=0.47	-	F(4,35)=3.2	*	F(4,45)=2.56	-
f0	F(4,32)=3.695	*	F(4,32)=10.82	**	F(4,41)=32.1	**
Subject F2						
duree	F(4,31)=4.4	*	F(4,34)=6.76	*	F(4,33)=1.81	-
f0	F(4,31)=15.6	**	F(4,34)=40.42	**	F(4,33)=26.86	**
Significance levels: ** p<0.0001; * p<0.05 ; - not significant						

[1] The statistical package used was StatView 5.0.

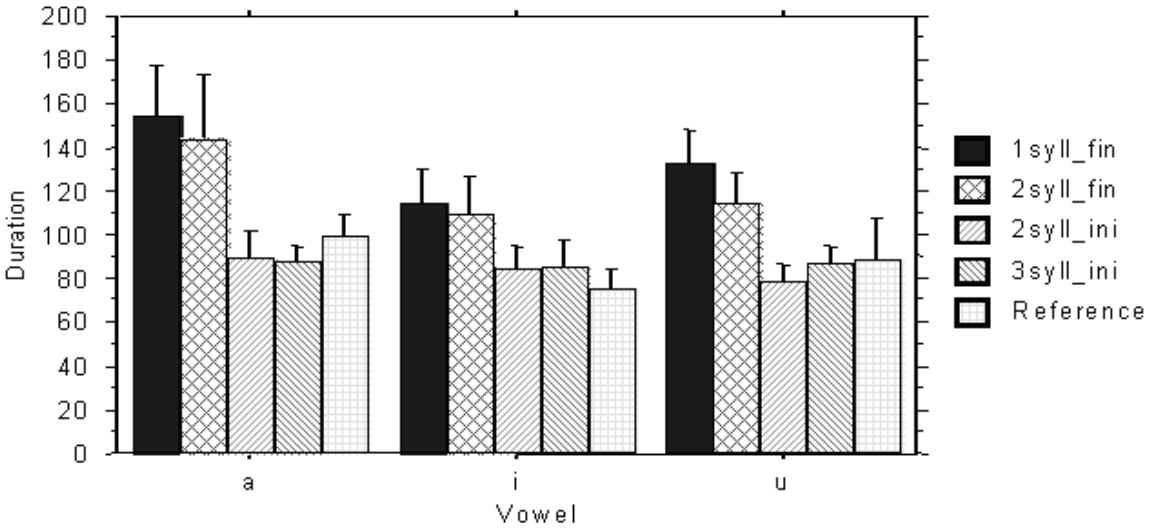


Figure 3: Duration values for /a/, /i/ and /u/ for speaker M1. The five columns of each set are from left to right: 1_syll_fin, 2_syll_fin, 2_syll_ini, 3_syll_ini, reference.

The results show that the effect of condition is significant on vowel duration in most cases (except for speaker F1 on /a/ and /u/ for example: cf. table 2). As can be seen on figure 3, the values can be separated in two main groups: vowels in word final syllables are significantly longer than vowels in word initial syllables and in the reference condition. This is consistent with the boundary adjacent vocalic lengthening that is often mentioned in the literature (Delattre 1966).

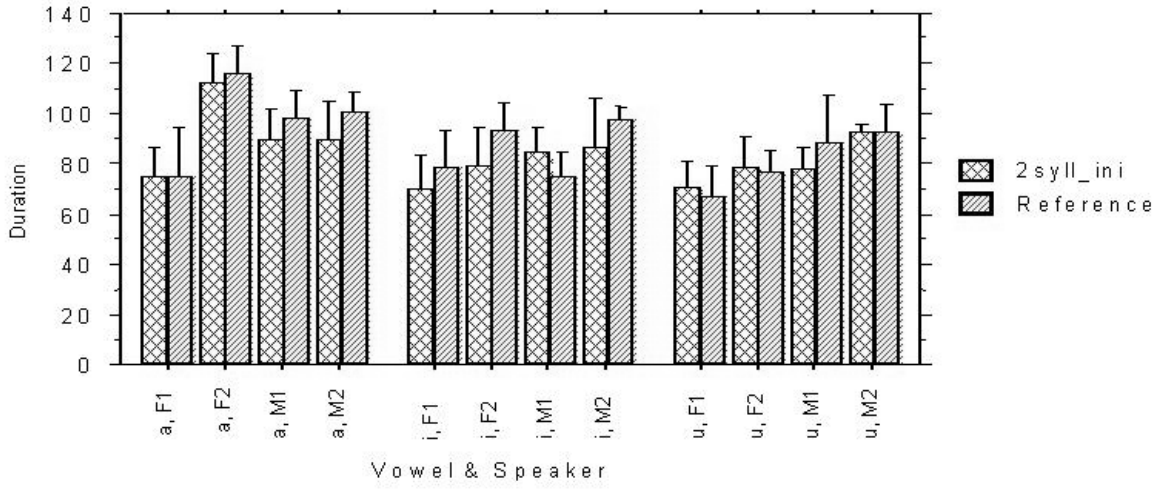


Figure 4: Duration values for /a/, /i/ and /u/ for each speaker, comparing 2syll_ini and the reference conditions.

The results for the reference condition (soupapes neuves) show no significantly different values compared to condition 2a (p=0.067 given by the Fisher's PLSD

Test), although the mean durations are usually slightly longer for the reference condition (the difference is significant for speaker M1 on /i/ and /u/ only, as can be seen on figure 4). This is coherent with the π-gesture hypothesis, as the presence of the word boundary (reference condition) generally does not significantly lengthen the vowel.

3.2. f0

Next we turn to a comparison of f0 values in word initial and word final position. The ANOVAs show that speakers significantly produce increased values of f0 for the vowels in final syllables (1syll_fin and 2syll_fin) compared to their counterparts in word initial syllables. This is consistent with the boundary adjacent f0 continuation rise (H%) that is often mentioned in the literature.
The only exception noticed is for vowel /a/ for speaker F1. Although the ANOVA shows a significant effect of condition on the f0 values, no tendency consistent with the other results can be found. This is similar to the results we found in the previous part for this speaker. As an increase in duration and f0 are useful indicators of boundary realisation, it suggests that speaker F1 did not realise the boundaries as intended (on sentences with /a/). As for sentences with vowel /u/ (cf. table 3: no significant effect of duration), the significant increase in f0 on the last syllable before the boundary suggests that speaker F1 realises boundaries by a continuation rise only (and not accompanied by a lengthening). We decided not to get rid of this speaker for further analyses as the kinematic analyses (lip aperture) could bring interesting results. Will the kinematic results be strictly coherent with the acoustic results observed (f0 and duration)?

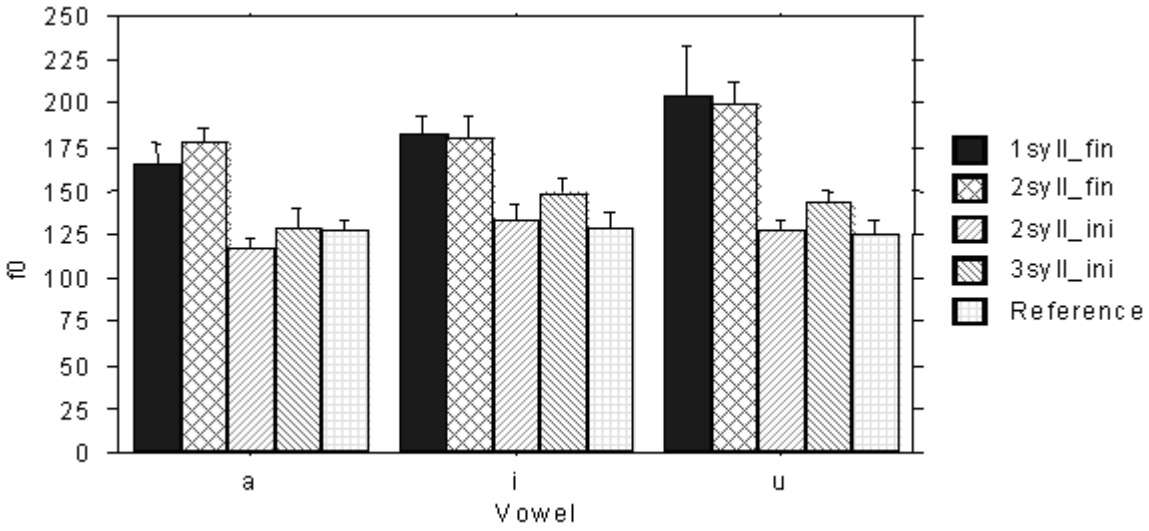


Figure 5: f0 values for /a/, /i/ and /u/ for speaker M1.

3.3. Formants

Vowels in word final syllables (1_syll_fin and 2_syll_fin) were hypothesised to be more extreme in terms of a higher F1 when located in pre-Intonational Phrase boundary (as found by Tabain 2003a). This may be due to hyper-articulation and/or to longer duration. Vowels extracted from a large database (2 hours for French and for German) have shown to have more extreme values in terms of higher F1 and F2 with increasing duration (Gendrot & Adda 2004). However, as for /a/, results for F2 are less systematic and obvious as noted by both studies. This result can be explained by the fact that) /a/ in French is rather central on the frontness-backness scale. If we consider the Sonority Expansion, F1 values would be systematically higher when close to a boundary, while no explicit hypothesis can be drawn for F2. If we refer to coarticulation, then vowels close to a boundary would be less coarticulated and thus have a higher F2 (a coarticulation effect would make F2 transition point towards a locus at approximately 600 Hz; Delattre et al. 1955).
Target vowels /i/ and /u/ produced by speaker F1 were often devoiced and we were not able to investigate any results for this speaker. F1 and F2 values for /i/ and /u/ were not significantly different according to 2 factor („condition“ and „vowel“) ANOVAs. We assume that this is due to a lack of variability of these extreme vowels as these are known to be less variable than /a/ (for example Recasens 1999). However, considering /a/, the effect of condition is significant for F1 and F2 values for all speakers. There is an overall tendency for vowels in word final syllables to have higher F1 values than vowels in word initial syllables. For speakers M1 and F1, vowels in condition 2b (3syll_ini) have significantly lower F1 values than vowels in final position (conditions 1a and 1b). For the 2 others speakers, vowels in condition 2a (2syll_ini) have significantly lower F1 values than vowels in final position. The 1st formant values for the reference condition are also significantly lower (except for speaker F2).
As for F2 values, the results follow the same trends as F1, but are significantly different for speaker M1 and M2 only. When considering that F1 is related to aperture and F2 to anteriority of the vowel, the results indicate that /a/ is more posterior and closed for vowels in word-initial syllables (cf. figure 7). This is interpreted here as a higher coarticulation, rather than more centralisation, as more centralisation would lead F2 to move towards 1500 Hz instead of lowering, as can be noticed here. The 2nd formant values for the reference condition are slightly higher than vowels in word initial position (for both speakers M1 and M2, not shown in the figure), which suggests that they are rather centralised than coarticulated.

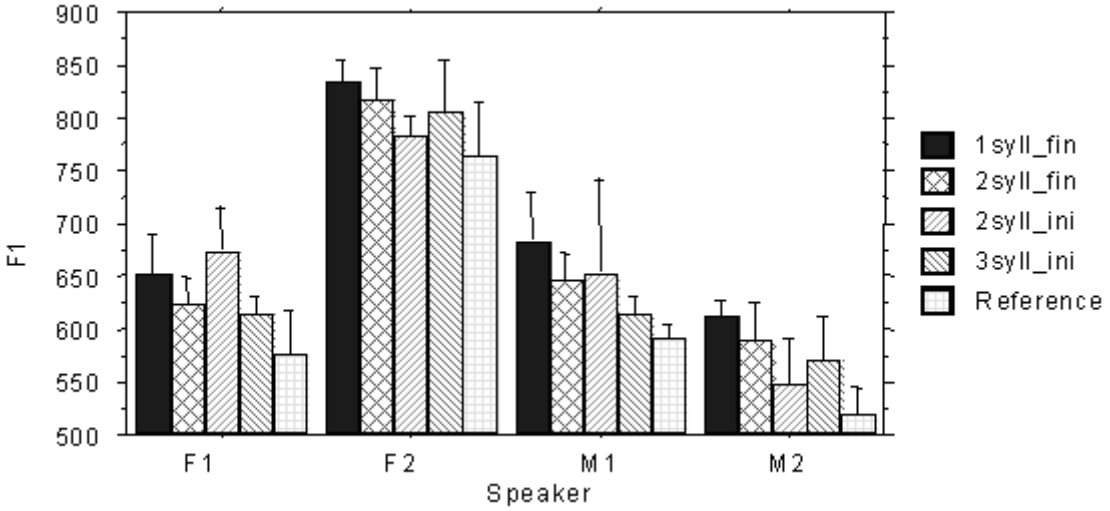


Figure 6: F1 values for /a/ for all speakers, comparing all conditions.

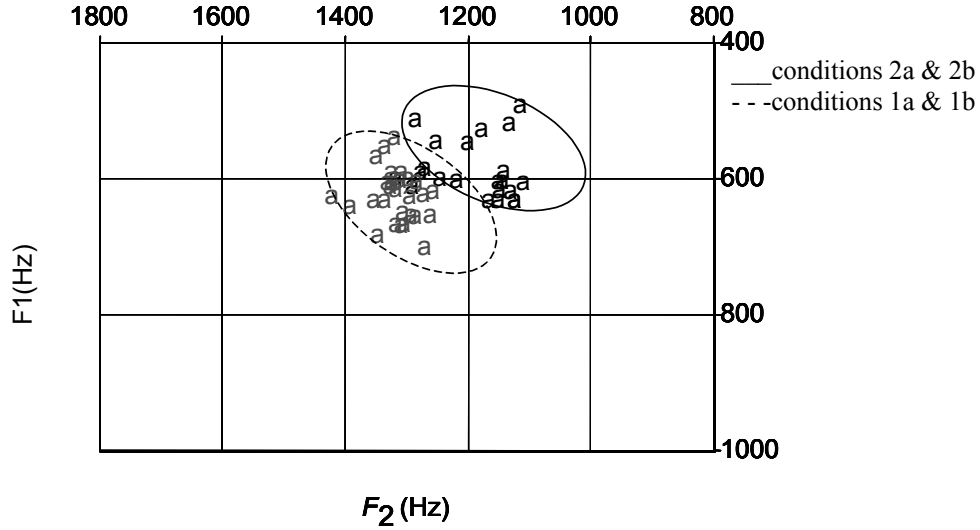


Figure 7: F1 and F2 values for /a/ for speakers M1 and M2, comparing vowels in word initial syllables (conditions 2a & 2b) and vowels in word final syllables (conditions 1a & 1b).

3.4. Intra-Oral Pressure (PIO) measurements

We expected no significant variations of PIO according to the different conditions as an increase in PIO would reflect an emphasis on the analysed vowel. The overall tendency is for vowels in word initial syllables to show a slightly higher PIO than their counterparts in word final syllables. However no effect of boundary surrounding on PIO could be observed for speaker M1

([F(4,59)=0.405, p=0.805]), speaker M2 ([F(4,61)=2.54, p=0.064]) and speaker F2 ([F(4,53)=1.34, p=0.27]). An exception can be noticed for speaker F1 [F(4,63)=5.28, p=0.0011] who produced vowels in word final syllables with greater values (of PIO) than vowels in word initial syllables. We cannot provide a clear interpretation of the results for speaker F1 here. This still suggests that word initial positions (when not emphasised) are in most of the cases not produced with a significant increase or decrease in intra-oral pressure compared to their counterparts in word-final positions.

3.5. Kinematic measurements

Word final vowels are generally characterised by a greater lip aperture movement as can be seen in figure 8 and 9. This is specifically consistent for /a/ and /i/. It is true for all speakers including speaker F1 whose duration and f0 values didn't show the same tendencies as the other speakers (cf. 3.1. and 3.2).
The reference condition (condition 3) for /a/ and /i/ shows larger opening movements than the initial syllables conditions in many cases; however these differences are rarely significant. Once again we cannot claim that the π-gesture's activity has been altered by the presence of the lexical word boundary (reference), as it rarely significantly conveys a supplementary lip aperture compared to the vowel in condition 2a (2_syll_ini).

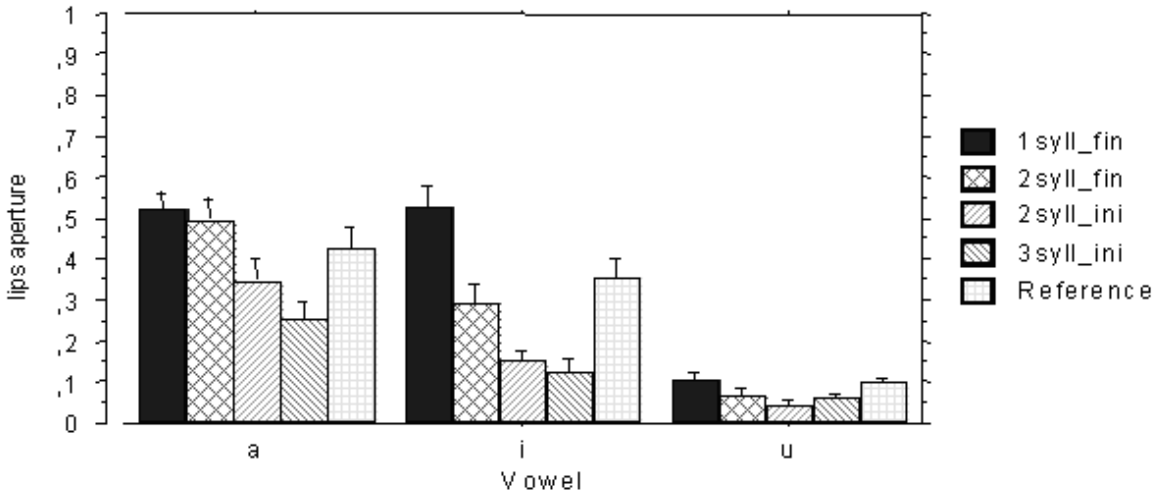


Figure 8: Lip aperture values for /a/, /i/ and /u/ for speaker F1.

Following the idea that these temporal and spatial (lips movements) prosodic changes may not be interdependent (specifically for speaker F1), it would be interesting to investigate whether a spatial enlargement is possible without a

temporal slowing down and vice versa: one way to evaluate this is to measure the velocity of the opening (and closing) gestures.

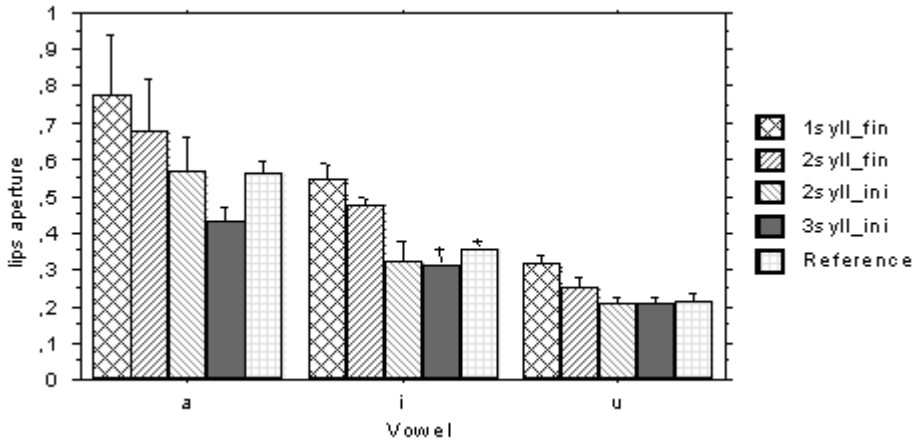


Figure 9: Lips aperture values for /a/, /i/ and /u/ for speaker M1.

Maximum velocity was measured as the maximum value on the derivative of the lips' opening movements as found in the literature (Beckman & Edwards 1992).

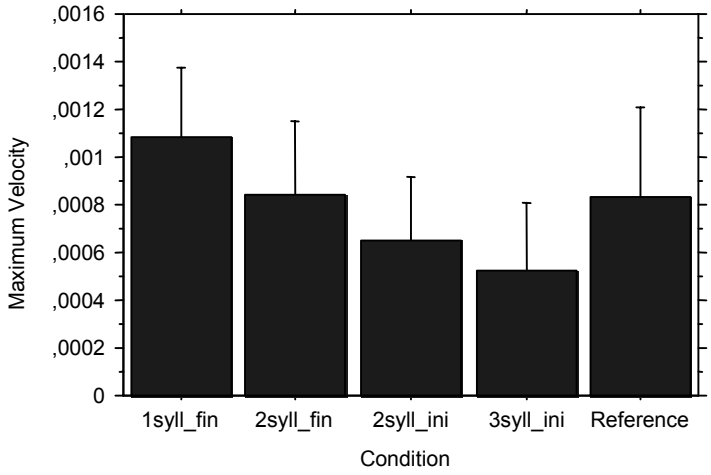


Figure 10: Maximum velocity values for /a/ for speaker F1.

The results for speaker F1 (cf. figure 10) indicate the same tendency as was found for duration for the other speakers. This suggests that the vowel /a/ in word final syllables is realised with a larger opening thanks to a higher maximum velocity while the duration of the vowel is not significantly higher. This configuration is different from other speakers for which temporal (longer vowels) and spatial (larger lip movements) changes seem to be interdependent.

4. Conclusion

The π-gesture hypothesis seems to be confirmed in this experiment, i.e. vowels (in word final position) that are closer to the boundary are longer, have a higher f0, larger opening lips gestures. As shown for speaker F1, a higher maximum velocity during the vowel, still allowing the same gestural extent, may also compensate for a short duration, but this will have to be further investigated.
The results for the reference condition are coherent with the π-gesture hypothesis. The presence of the lexical word boundary (when compared to the condition 2a) generally conveys no significant extra lengthening and no larger lip opening gestures.
In our corpus we tested vowels in word initial syllables versus vowels in word final syllables; the word initial position is quite strong in French (as in several languages) and we know that it is a location for possible supra-glottic tension - as mentioned by Vaissière (2004). The results for PIO indicate that vowels in word initial positions are not realised with a lower pressure compared to vowels in word final positions. However we found evidence for temporal reduction as well as for reduced lip aperture movements. The analysis of formant values suggested that vowels in word initial syllables could be considered as more coarticulated, rather than centralised. A complete investigation of the correlations between duration and gestural extent is needed to check whether a higher velocity in word-initial position may compensate for the temporal reduction.

Acknowledgements

My thanks to all at the Laboratoire de Phonétique et Phonologie, especially Jacqueline Vaissière, Shinji Maeda, Cécile Fougeron, Alexis Michaud. Preliminary findings were presented in 2004 at the German-French summerschool on Cognitive and physical models of speech production and perception and the production-perception interaction in Lubmin. Recordings were made at the "Hôpital Européen Georges Pompidou" with the help of Stéphane Hans. Many thanks for his useful hints and his obligingness. My deepest thanks to Jean Léo Léonard, Olivier Corbin, Céline Raynal and Frédérique Bénard for their patience during the recording sessions. Recording equipment was lent by the Laboratoire Phonétique et Phonologie-CNRS UMR7018, the École Doctorale 268 and the "Hôpital Européen Georges Pompidou". Finally the author would like to thank reviewers Christine Mooshammer and Bernd Pompino-Marschall for many helpful comments on this paper, as well as all the organisers of the German-French summerschool (...) for their tremendous work.

References

Beckman, M. & Edwards, J. (1992) Intonational categories and the articulatory control of duration. In: Y. Tohkura, E. Vatikiotis-Bateson, Y. Sagisaka (eds.), *Speech Perception, Production, and Linguistic Structure*. Tokyo: Ohmsha, Ltd. 359-375.

Benguerel, A. (1970) *Some physiological aspects of stress in French*. Ann Arbor: The university of Michigan, Phonetics Laboratory.

Boersma, P. & Weenink, D. (1999) *Praat, a system for doing phonetics by computer*. Institute of Phonetic Sciences of the University of Amsterdam. 132-182.

Browman, C. & Goldstein, L. (1992) Articulatory phonology: An overview. *Phonetica*, 49: 155-180.

Byrd, D. (2000) Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica,* 57: 3-16.

Byrd, D., Kaun, A., Narayanan, S. & Saltzman, E. (2000) Phrasal signatures in articulation. In Michael B. Broe & Janet B. Pierrehumbert (eds.) *Papers in Laboratory Phonology V*. Cambridge: Cambridge University Press. 70-87.

Byrd, D. & Saltzman E. (2003) The elastic phrase: Modelling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31: 149-180.

Cho,T. (2001) *Effects of prosody on articulation in English PhD dissertation*. University of California, Los Angeles. (Distributed as UCLA Dissertations in Linguistics, number 22).

de Jong, K.J. (1995) The supraglottal articulation of prominence in English. *Journal of the Acoustical Society of America*, 97: 491-504.

Di Cristo, A. (1998) *Intonation in French*, in Intonation Systems: A Survey of Twenty Languages, edited by D. Hirst and A. di Cristo (Cambridge University Press, Cambridge), pp. 195–212.

Delattre, P., Liberman, A., and Cooper, F. (1955) Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27(4).

Delattre, P. (1966) *Studies in French and Comparative Phonetics*, Mouton, La Haye.

Edwards, J., Beckman, M. and Fletcher, J. (1991) The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, 89(1): 369-382.

Erickson, D. (1998) Effects of contrastive emphasis on jaw opening. *Phonetica*, 55: 147-169.

Gendrot, C. & Adda, M. (2004) Analyses formantiques automatiques de voyelles orales : évidence de la réduction vocalique en langues française et allemande. *MIDL*,Paris: 29-30.

Grammont, M. (1933) *Traité de phonétique*. Paris: Librairie Delagrave.

Harrington, J.; Fletcher, J.; Beckman, M. (2000). Manner and place conflicts in the articulation of accent in Australian English. In Beckman & Kingston (eds.): Papers in Laboratory Phonology V: Acquisition and the Lexicon. Cambridge: University Press, 40-51.

Hertegard, S., Gauffin, J. & Lindestad, P. A. (1995) A comparison of subglottal and intraoral pressure measurements during phonation. *Journal of Voice*, 9(2):149-55.

Jun, S.-A., & C. Fougeron. (1997) A Phonological model of French intonation. *Poster presented in the ESCA Workshop on Intonation*, Athenes.

Keating, P., Cho, T., Fougeron, C., & Hsu, C.S. (2003) Domain-initial strengthening in four languages. In J. Local, R. Ogden & R. Temple (eds). *Papers in Laboratory Phonology VI.* Cambridge University Press. Cambridge, U.K.: 143-161.

Ladefoged, P. & Draper, M. (1958) Syllables and Stress. *Miscellanea Phonetica*, 3: 1-14.

Lindblom, B. (1990) Explaining phonetic variation: a sketch of the H & H theory. In Hardcastle, W. & Marchal, A., eds., *Speech production and speech modelling*. Dordrecht. Kluwer: 403-440.

Lucci, V. (1983) *Etude phonétique du français contemporain à travers la variation situationnelle*. Grenoble: Publications de l'Université des Langues et Lettres de Grenoble.

Nooteboom, S.(1997) The prosody of speech: melody and rhythm. In: W. J. Hardcastle & J. Laver (eds*.), The Handbook of Phonetic Sciences*, Basil Blackwell Limited, Oxford: 640-673.

Pasdeloup, V. (1990*) Modèles de règles rythmiques du fraçais appliqué à la synthèse de la parole.* Thèse de Doctorat Nouveau régime, Université de Provence.

Recasens, D. (1999) Lingual coarticulation. In W. Hardcastle 1 N. Hewlett (eds), *Coarticulation: theory, data and techniques*, Cambridge University Press. Cambridge, U.K: 80-104.

Saltzman, E. & Byrd, D. (2000a) Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, 19(4): 499-526.

Saltzman, E & Byrd, D (2000b) Demonstrating effects of parameter dynamics on gestural timing. Meeting of the Acoustical Society of America, *Journal of the Acoustical Society of America*, 107(5,2): 2904.

Straka, G. (1963) La division des sons du langage en voyelles et consonnes peut-elle être justifiée ? *Travaux de Ling. et de Littérature*, U. de Strasbourg, 2: 17-98.

Tabain, M. (2003a) Effects of prosodic boundary on /aC/ sequences: acoustic results. *Journal of the Acoustical Society of America*, 113(1): 516-531.

Tabain, M. (2003b) Effects of prosodic boundary on /aC/ sequences: articulatory results. *Journal of the Acoustical Society of America*, 113(5): 2834-2849.

Vaissiere, J. (1975) Une procédure de segmentation automatique de la parole en mots prosodiques, en Français. *VIIiè mes Journées d'Etudes sur la parole*, Nancy : 193-208

Vaissière, J. (2004) Perception of Intonation. In *Handbook of Speech Perception*, Blackwell, Oxford.

Learning to control an articulatory synthesizer by imitating real speech

Ian S. Howard
Sobell Department, Institute of Neurology, UCL, England

Mark A. Huckvale
Phonetics & Linguistics, UCL, England

The goal of our current project is to build a system that can learn to imitate a version of a spoken utterance using an articulatory speech synthesiser. The approach is informed and inspired by knowledge of early infant speech development. Thus we expect our system to reproduce and exploit the utility of infant behaviours such as listening, vocal play, babbling and word imitation. We expect our system to develop a relationship between the sound-making capabilities of its vocal tract and the phonetic/phonological structure of imitated utterances. At the heart of our approach is the learning of an inverse model that relates acoustic and motor representations of speech. The acoustic to auditory mappings uses an auditory filter bank and a self-organizing phase of learning. The inverse model from auditory to vocal tract control parameters is estimated using a babbling phase, in which the vocal tract is essentially driven in a random manner, much like the babbling phase of speech acquisition in infants. The complete system can be used to imitate simple utterances through a direct mapping from sound to control parameters. Our initial results show that this procedure works well for sounds generated by its own voice. Further work is needed to build a phonological control level and achieve better performance with real speech.

1. Introduction

Several different approaches have been adopted in order to get a machine to speak. One approach is to record human speech, chop it up into pieces and then reassemble them in a new desired order. Another approach is to program phonological-to-synthesizer control mapping rules by hand. The approach we

take here is to try to discover an acoustic-to-synthesizer control mapping using machine learning techniques. We gain inspiration from infant speech acquisition and in this vain also use an articulator-based synthesiser for our work, to make our system's speech production apparatus more like that of a human. Our work here is obviously only a first step towards producing a useful system, which would also need to learn other associations such as phonological-to-articulatory mapping to be a useful system. Several other authors have made similar investigations. Bailly et al. (1997) modelled the generation of formant trajectories. Guenther (1994, 1995) has also carried out similar work.

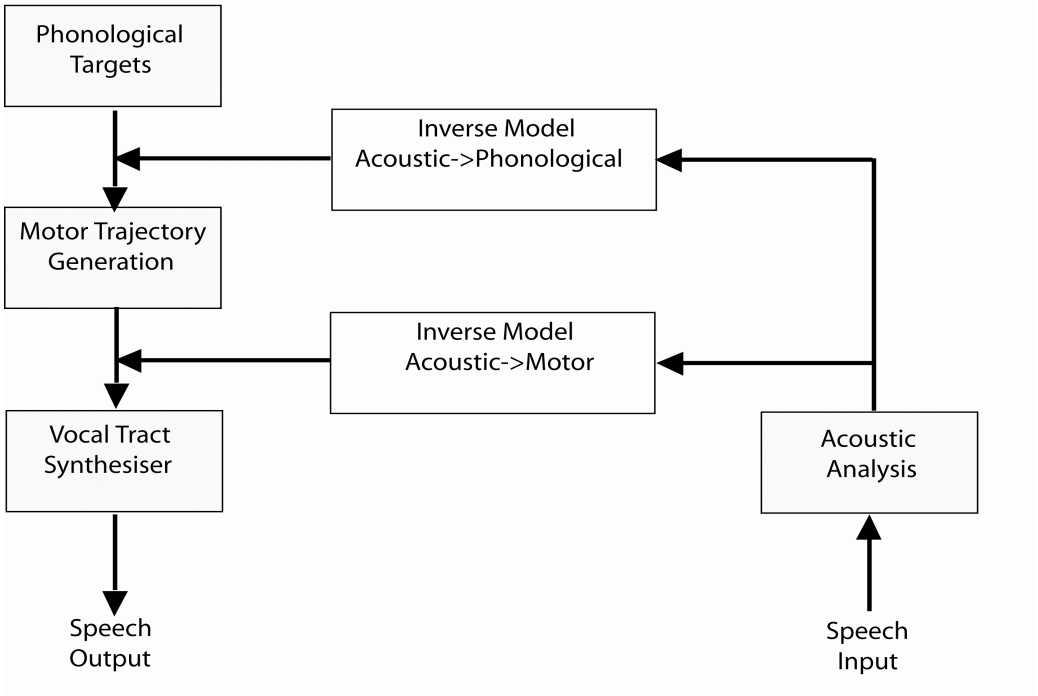


Figure 1: Inverse models. Mappings between acoustic, phonological, and articulator representations of speech.

2. Inverse models

Let us consider the speech production and analysis system shown in figure 1. Here we have a system that can both generate artificial speech and also perform a basic acoustic speech analysis.

In this model, the vocal tract synthesiser is controlled by a vector of articulatory parameters which change as a function of time, as specified by a motor

trajectory generation stage. It is the task of the latter to move the articulators of the model in such a fashion that the desired speech utterance is generated by the synthesiser.

If we feed back the speech signal generated using this process into the acoustic analysis pathway, our system then has an explicit representation of the motor commands it uses to generate speech, as well as their acoustic consequences. Since we can have access to both the input to our synthesiser, and also its acoustic consequences, we can use this information to define an inverse transformation that will map an acoustic representation of speech back to the motor commands needed to generate it. This inverse transformation is marked on figure 1 as an acoustic-to-motor inverse model. Clearly, for such an inverse model to be useful in practice, it must perform well over a representative range of conditions, corresponding to the kinds of inputs the synthesiser would experience during normal use. It must also account for any time delay between the motor commands and their sensor consequences. In the work described here, this alignment is also performed by the inverse model. A discussion regarding the training of the inverse model is given in the next sections. The concept of inverse models is well established in the field of motor control; see (Wolpert, 1997) for a further discussion of the issues involved.

If our speech production system contained a higher hierarchical level of control, it would also be possible to define an inverse model to a more abstract level of representation. For example, if our motor trajectory generator was controlled by a phonetic input representation, we could define an inverse model pathway mapping between acoustic and phonological representations of speech (also shown in figure 1). In the work described here, we only investigate the low-level acoustic to vocal tract parameter inverse model using an articulator synthesiser based on the work of Maeda (Maeda, 1990) and a simple acoustic analysis based on the JSRU channel vocoder (Holmes, 1982), together with a simple autocorrelation estimate for fundamental frequency. The Maeda parameters are shown in table 1. In our implementation, they are specified at a sampling rate of 8kHz to generate speech signal output at the same rate. This gives the synthesiser an acceptable speech quality without requiring excessive computational resources to run it.

Assuming that we can find the inverse model for our system, it can be used to provide a basic mechanism to control a synthesiser. All we must do is to process input speech with the acoustic analysis and then map this representation to the vocal tract control parameters using the inverse model.

Table 1: Maeda's articulator model parameters.

Parameter	Description
P1	Jaw position
P2	Tongue dorsum position
P3	Tongue dorsum shape
P4	Tongue apex position
P5	Lip height (aperture)
P6	Lip protrusion
P7	Larynx height
P8	Voicing (glottal area)
P9	Fundamental frequency

3. Learning the inverse model by babbling

The main task in the approach we describe here is the estimation of the inverse model that will map between an acoustic representation of speech and the required vocal tract parameters. If we use some kind of motor trajectory generator to vary the vocal tract parameters as a function of time, and then use them to generate synthesised speech (which is then subsequently acoustically analysed), we can create a data set with which to define the input and output relationships of the inverse model. This is shown in figure 2. Training the inverse model then constitutes a classical supervised learning task.

As is the case with regression analysis and pattern recognition in general, we want the inverse model to operate well over a wide range of representative inputs and also to generalize well on previously unseen ones. In order to generate an inverse model that meets these requirements, appropriate training data is required, as well as a suitable pattern recognition/regression technique capable of learning the required transformation. With regard to the generation of suitable training data, the ideal input to the synthesiser would correspond to motor trajectories that resulted in high quality speech output over a wide range of utterances. This input would then sample the vocal tract parameter space in a fashion consistent with its intended use. Training the inverse model on relevant rather than irrelevant data will also result in higher performance, because it will be optimised to implement the transformations that are needed, rather than ones that are not.

At the beginning of the estimation process, the generation of ideal motor trajectories is clearly not possible, because we do not know what they are ***a priori***. We

must therefore resort to a method of synthesiser control that will sample its input space in a fashion consistent with our knowledge of speech generation. In the field of control theory, such system identification of a complex system is often carried out using some kind of random input excitation. However, this excitation should be matched to the task in hand. Since we are using an articulatory synthesiser, we know that there are limitations on the maximum rate of change of articulator positions due to biophysical constraints. To sample vocal tract parameter space, we thus adopt an approach that randomly investigates this space, but only does so in a slowly varying fashion consistent with the changes in articulator positions appropriate for the generation of speech. For example, there would be no point in instantaneously moving the jaw position from up to down, because we know that such a change could not occur in a real vocal tract.

We have implemented this slowly varying random signal generation scheme using a Hidden Markov Model (HMM) with output interpolation, as explained below. We describe this as a babble generator, because it generates sounds that have similarities with the babble produced by infants, although the exact form of the sound sequences generated by this approach does not exactly correspond to baby babble. Its task is only to explore the input space of the synthesiser in a way that is useful for the purposes of training the inverse model.

For an inverse transformation to exist, it is necessary for the forward transformation due to the plant (represented here by the synthesizer and acoustic analysis) to be unique. If this is not the case, it will not be possible to find an inverse. That is, if many different vocal tract configurations give rise to the same acoustic output, it will be impossible to know, on the basis of an acoustic measurement, which vocal tract configuration was responsible. One approach that can be adopted in this case is based on distal supervised learning (Jordan & Rumelhart, 1992), which involves first training a simpler forward model and then using this to find the inverse. Although the instantaneous mapping between vocal tract configuration and acoustic output is not in general unique, this problem can also be overcome using a wide acoustic context at the input to the inverse mapping. The latter approach was used here and an acoustic window of 50ms was used as an input to the inverse model.

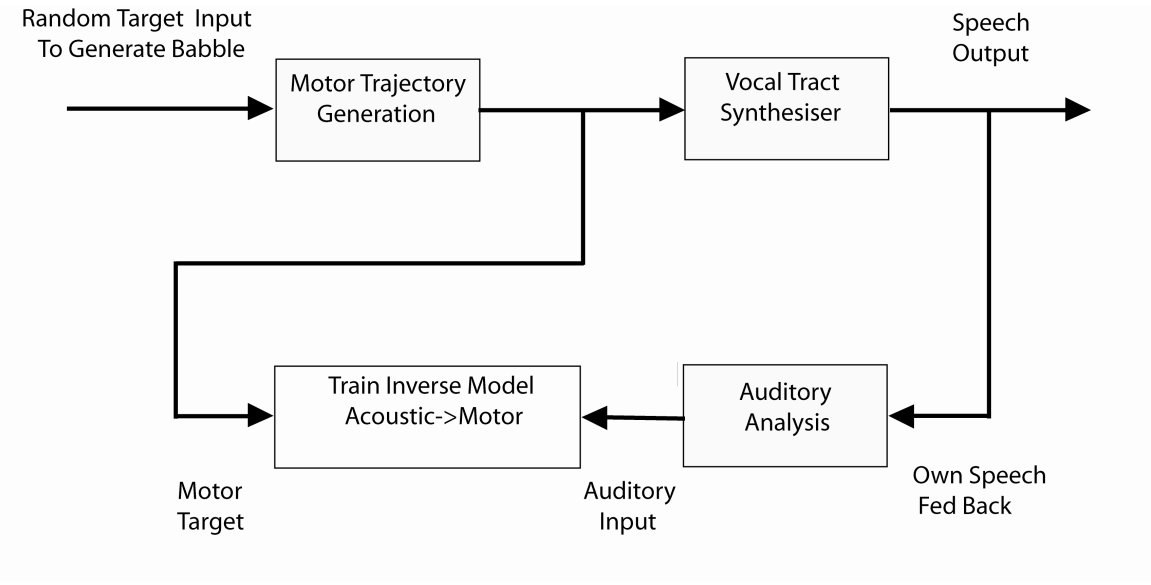


Figure 2: Learning an inverse model by babbling. Motor command space is sampled using 'random babbling'. Direct training of the inverse model then becomes a classical supervised learning task.

4. Babble generator

The task of the babble generator is to generate sequences of parameters that explore the input space of the synthesiser in a fashion relevant for speech production. The basic idea is to use an HMM to sample phonetically significant regions of synthesiser space and then to interpolate the output to generate a slowly varying time signal vector. By restricting the states of the HMM to phonetics targets (such as vowels and consonant targets) we can incorporate ***a priori*** speech knowledge into the babble generator. Consider the HMM shown in figure 3. Using this model structure, we can sample three points in vocal tract parameter space, represented by the states V1, V2 and V3. Each of these states has associated transition probabilities to the other states and also an associated output parameter vector. We can use this approach to sample part of vowel and consonant space (which forms a subset of the much larger total parameter space) by associating each of the states with the parameters relating to a particular vowel/consonant generated by the articulatory synthesiser. Using a duration parameter in our output vector, we can interpolate between successively sampled vectors to generate continuous trajectories needed to drive the Maeda vocal tract synthesizer.

We have also investigated a more naive scheme by directly sampling randomly from the entire parameter space, rather than from vowel and consonant space. In this case, the output from our babble generator did not sound as much like baby babble. Configurations of the articulators arose that were not relevant in the

production of speech. Comparisons of the output speech generated by various babbling schemes can be found online (see section 5).

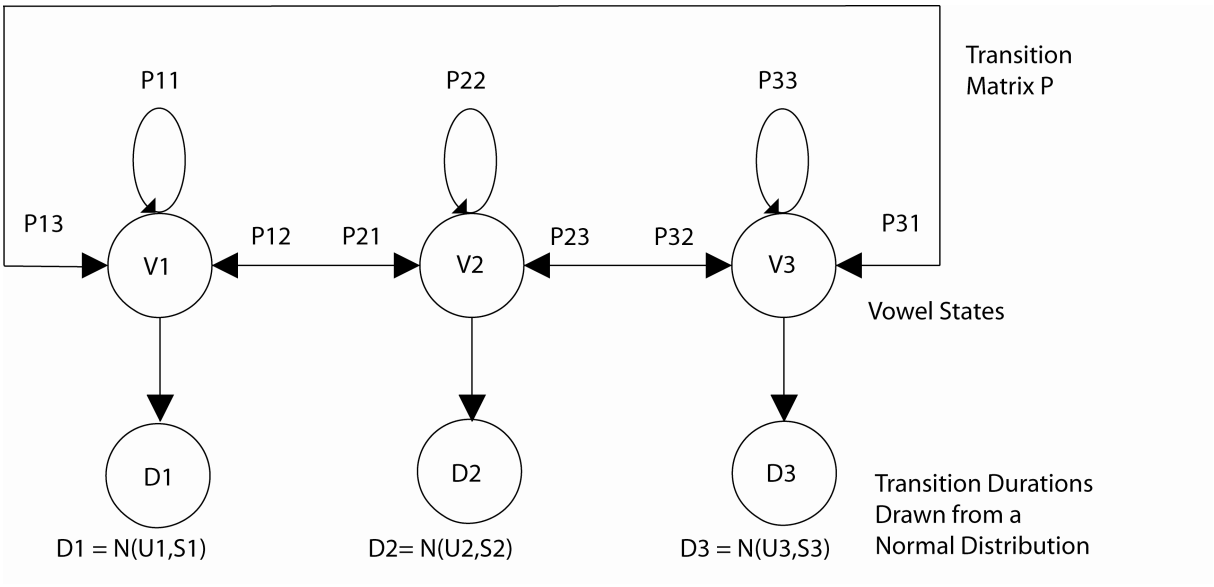


Figure 3: HMM babble generator. Each state represents a possible vocal tract configuration. In this case, only 3 vowels are shown for clarity. The transition matrix determines vowel sequences. The generation matrix specifies related vocal tract parameters which are then interpolated to provide smooth trajectories.

In this initial work, the HMM states were limited to five pure vowels and the consonants /b/ and /g/. By manipulating the transition probabilities between the states it was possible to generate babble either over this entire vowel space and consonant space, or only over a small part of it. An example of the latter was generation of the sequence /babababab/.

To generate smooth parameter trajectories in time at the sampling rate needed by the synthesiser, cosine interpolation was performed on the vectors generated by the state sequences according to the relation

$$\mathbf{VTpar(X) = VTparStart + (Cos(X.PI\ /duration)\ -1)\ .\ (VTparStart\ -\ VTparEnd)/2}$$

$$\mathbf{for\ X = 0,\ 1,\ 2,\ \ldots\ ,\ duration\ samples}$$

Figure 4 shows the parameter trajectory outputs from the babble generator, as well as the corresponding vocoder analysis of the resulting synthesised speech.

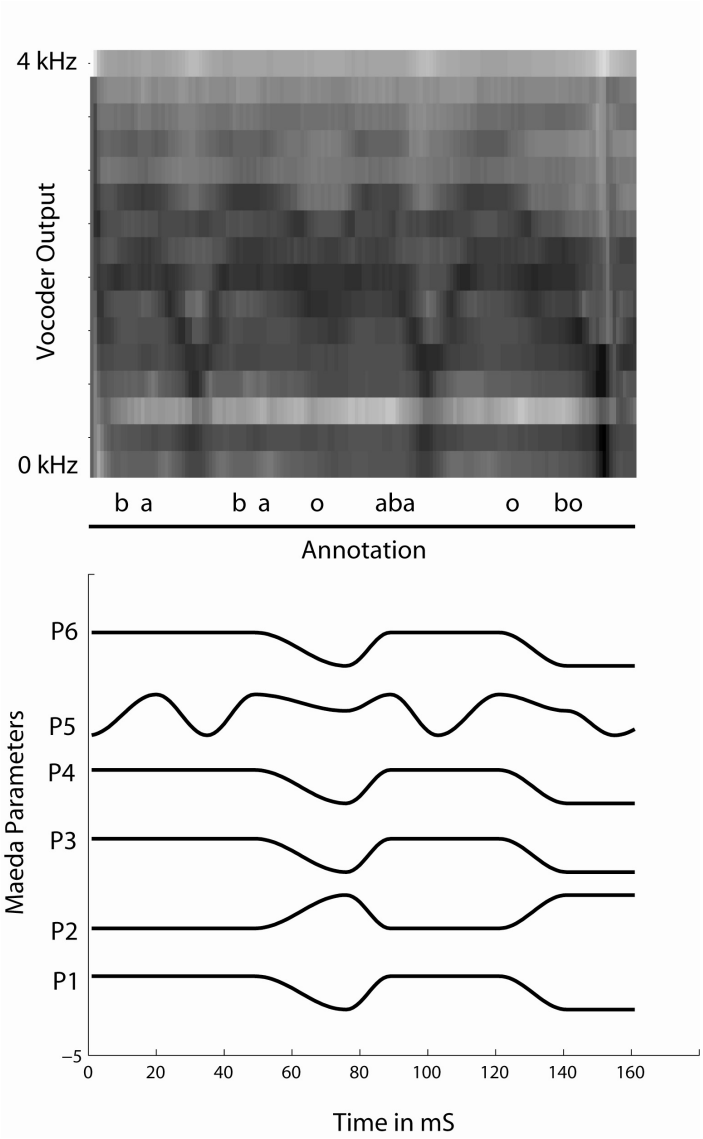


Figure 4: Babble generator output trajectories for parameters 1 to 6 and the corresponding vocoder analysis of the resulting speech signal. In this case the transition matrix was set up to babble over vowel space and also include the consonants /b/ and /g/

5. Experimental details

The babble generator was run on vowel and consonant targets to generate 180 seconds of articulator trajectory data. In our current implementation this data length was limited by computational considerations. This data was then used to drive the Maeda synthesiser and generate a time waveform representation of the output speech signal. The latter was then subjected to acoustic analysis with the channel vocoder, which generated 17 frequency outputs, as well as a fundamental frequency estimate, every 10ms.

For our experiments, a Matlab implementation of the multilayer perceptron (MLP) with linear output units was used to implement the inverse model (Nabney & Bishop, 1995), although in principle many other regression techniques could have been used. Input and output data patterns were normalized by subtracting their mean value and dividing by their standard deviation (after recognition, the inverse procedure was used on the output to reconstruct the estimated data's range). It was trained using back-propagation (Rumelhart et al., 1987). As mentioned previously, the input vector spanned 50ms in time and consisted of 5 adjacent vocoder frames. This time window also provided an automatic means for the inverse model to compensate for any time delay between the acoustic data and the synthesiser control parameters. Since the input to the inverse model spanned 25ms forward and 25ms backwards in time from the current synthesiser control vector, any information in the acoustic data that lay within these limits could be related to the current motor command. If a more time localized representation of the acoustic input had been used (for example a single spectral input frame) it would have been necessary to explicitly account for any time shift (although this could also easily be achieved by using delay lines and optimising the MLP error over delay). The number of hidden units in the MLP was determined by experimentation and a final value of 20 hidden units was used.

The output of the network consisted of 9 linear units, and these were mapped to the 9 Maeda synthesiser parameters. Training the inverse model involved 1000 passes over the data set (which comprised around 18000 different patterns). This value was again arrived at by experimentation and doubling the cycles to 2000 did not significantly improved performance. The output of the inverse model was smoothed using a median filter to remove undesirable spikes. Figure 5 shows the operation of the inverse model on testing data for the jaw position parameter.

Apart from the articulator synthesiser, which was implemented in the C language, all analysis was carried out in Matlab on a PC running under Windows XP. A supplement to this paper is available on the web at www.ianhoward.de/ZASPIL2005.htm and contains .wav files of all the input and output utterances described in this paper.

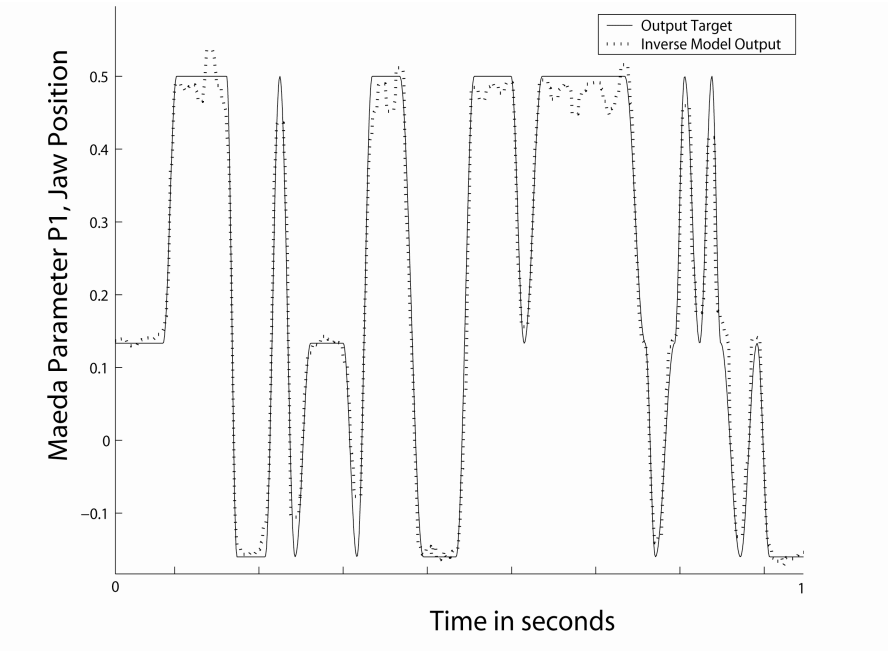


Figure 5: Training target and output generated by the inverse model for jaw parameter P1. A median filter was used to smooth the output.

6. Re-synthesising its own speech

After an inverse model had been trained during a babbling stage, it was clearly interesting to test its performance by re-synthesising some input speech. Evaluations were carried by listening to the speech generated by the system and also by observation of the corresponding wideband spectrograms. At this early stage of the work, this was considered to be the most appropriate form of assessment. In the first instance, we were concerned with whether any useful results could be generated by our system. At some time in the future, more rigorous quantitative evaluations will no doubt become useful. For example, the phonetic analysis of confusion matrices from listening tests on real and re-synthesised utterances could shed light on the deficiencies of the system.

Re-synthesising input speech involved passing an externally generated speech signal (that is, one from another synthesizer or a human subject) through the acoustic analysis and then through the inverse model. This produced a time-varying estimate of the vocal tract control parameters needed to regenerate the original speech utterance. It is clear that there would only be a perfect reconstruction of the input speech if the inverse model were perfect and vocal tract used to generate the speech was identical that of our vocal tract synthesiser. If the two vocal tracts had different physical dimensions, it may not be possible to get an exact reconstruction of the input speech. Such a mismatch arises when

children mimic the speech of adults, where there are clearly differences in the relative size of the speech production apparatus. This also manifests itself at an acoustic level. However, at a more abstract phonetic level, similarity between the two can still be achieved. This naturally raises the question of speaker normalization and speech matching criterion. In this initial work we do not focus on these two issues, although they will be investigated in more detail in future work.

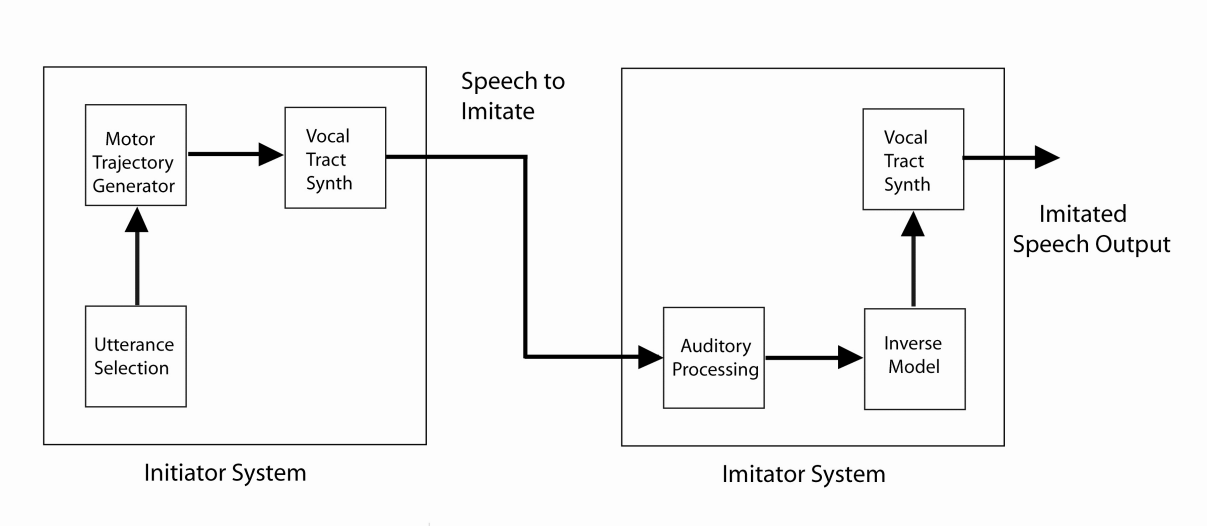


Figure 6: Imitating speech generated by an identical system. This is the simplest case and the issue of speaker normalization does not arise.

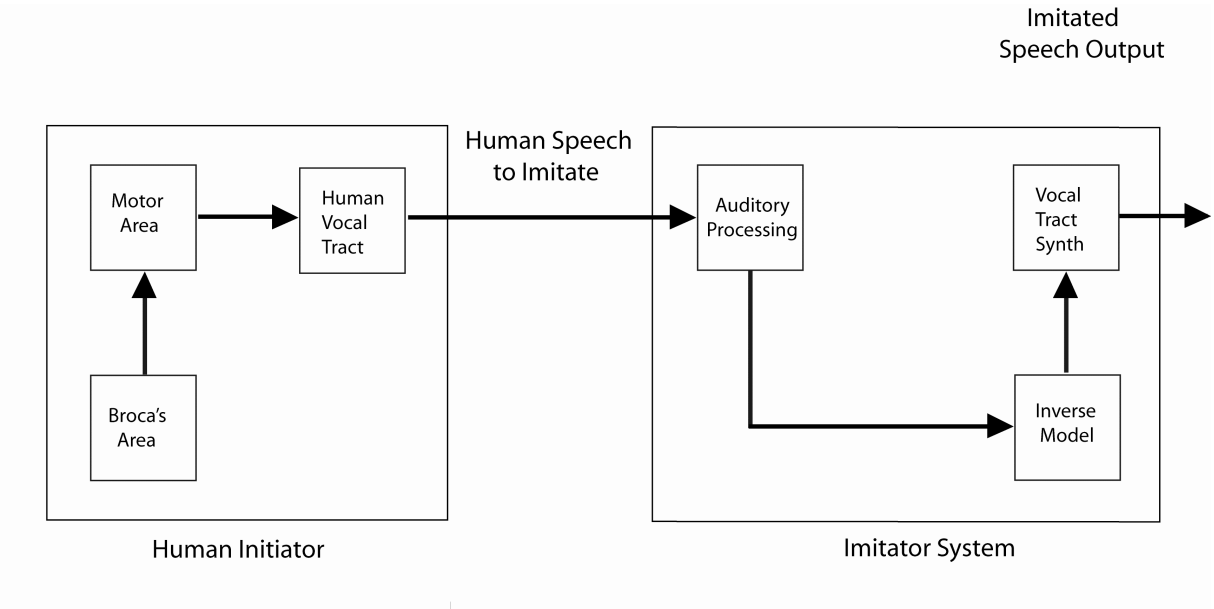


Figure 7: Imitating speech generated by a different (and human) system. In this case speaker normalization is needed.

The simplest task for the inverse mapping between acoustic and vocal tract parameters is the case when both generator and mimic vocal tracts are identical. This is shown in figure 6. We simulated this situation by using the articulator

model itself to generate babbled speech utterances (20 seconds of speech was used) and then used these as input to our imitation system. On its own speech, performance was very good (examples on the web), although this could probably still be improved further by using more training data.

7. Re-synthesising speech from human subject

A much more difficult case arises when speech generated by a different vocal tract must be imitated by the system. This is the case when real human speech is used as an input, as shown in figure 7. Simple utterances from one male speaker were used for the evaluation. In this case, the performance was lower, although simple utterances such as /babababababa/ and /bugi bugi babababa/ were still intelligible after re-synthesis (examples on the web). Wideband spectrograms for the input utterance (A) and the synthesiser output (B) are shown in figure 8. It can be seen that the imitated speech has voicing continuing into the silent parts of the utterance. Formant F1 corresponds well to that of the original speech, although F2 tends to be too low during the /i/ vowel section. Formants F3, F4 also appear somewhat too low in their values.

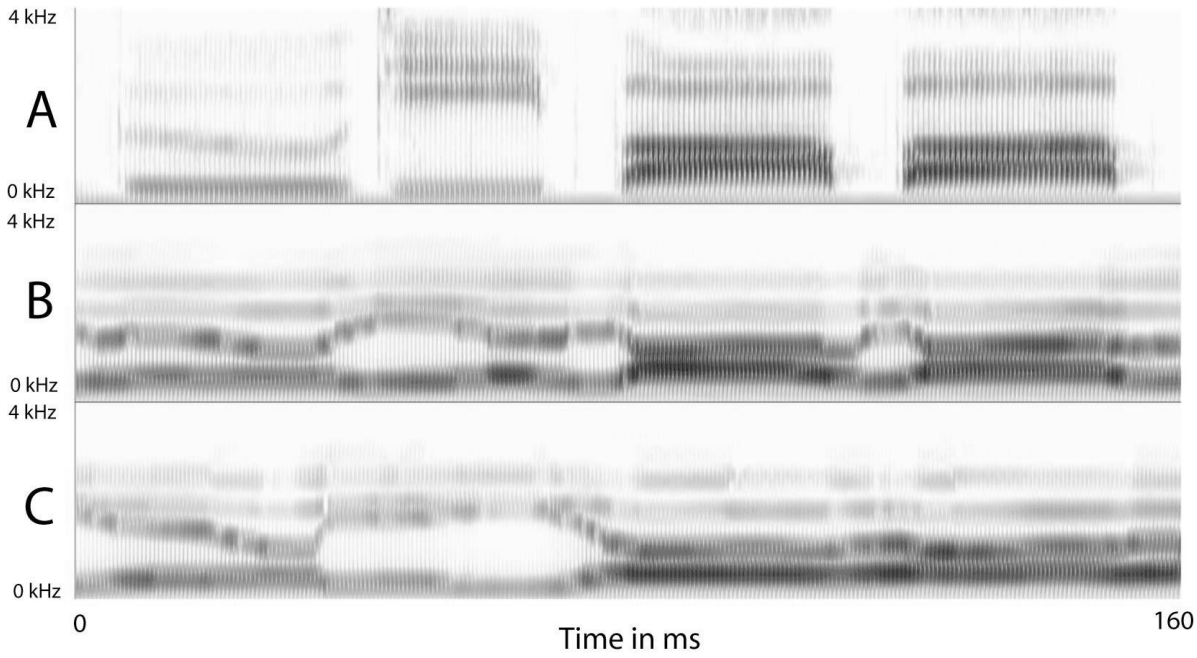


Figure 8: Wideband spectrograms for input utterance /boo gie ba ba/ from male speaker and re-synthesised outputs generated by imitation system. A shows real speech, B re-synthesis using vocoder analysis and C re-synthesis using an additional sparse coding stage.

8. Improved acoustic analysis

In order to improve the performance of the current system, we looked for inspiration from what is known about front end sensory processing in the human nervous system.

It has been known for quite a long time that the lower levels of sensory processing are matched to the statistics of the stimuli that they represent (Barlow, 1962). Recent work in the visual system has been quite successful in modelling receptive fields on neurons in the primary visual cortex by using sparse coding strategies (Olshausen & Field, 1996). Indeed it has been recently shown that such strategies appear relevant in auditory processing (Lewicki, 2002). The interpretation of this processing is still the subject of much debate, but it appears that the early sensory system is involved in both efficient coding as well as the extraction of features in the input modalities that relate to useful aspects of the input. A simple spectral vocoder vector provides quite a general representation of the acoustic input, whereas sparse coding may start to code in terms of features that operate over time and frequency and are specifically relevant for speech. We maintained the vocoder analysis and investigated adding a sparse coding to its output. The input to the latter consisted of 5 adjacent vocoder frames as before. It was implemented in two stages consisting of a whitening filter followed by a sparse filtering stage. The former consisted of linear Sobel edge detection filter and it effectively removed all short-term temporal and spatial correlations in the vocoder data (it effectively differentiated the vocoder data in 2-dimensions). The sparse filter itself was implemented by a 2-dimensional linear filter with as many outputs as inputs (90 in all). Its coefficients were optimized to minimize a cost function that requires the outputs to be both independent and also rarely active. Details of this coding scheme can be found in (Olshausen & Field, 1996). The sparse coding stage also was trained on 360 seconds of speech from one speaker.
After training, the auditory sparse coding stage was run on the babble data and used to train the inverse model as before. Figure 9 shows this modified system. Performance evaluation was once again carried out using real input speech, which was once again run through the inverse model pathway and re-synthesised as before. It was noted that the effect of the sparse coding was to improve performance on transitions (example on the web). Figure 8 C shows a wideband spectrogram of the output.

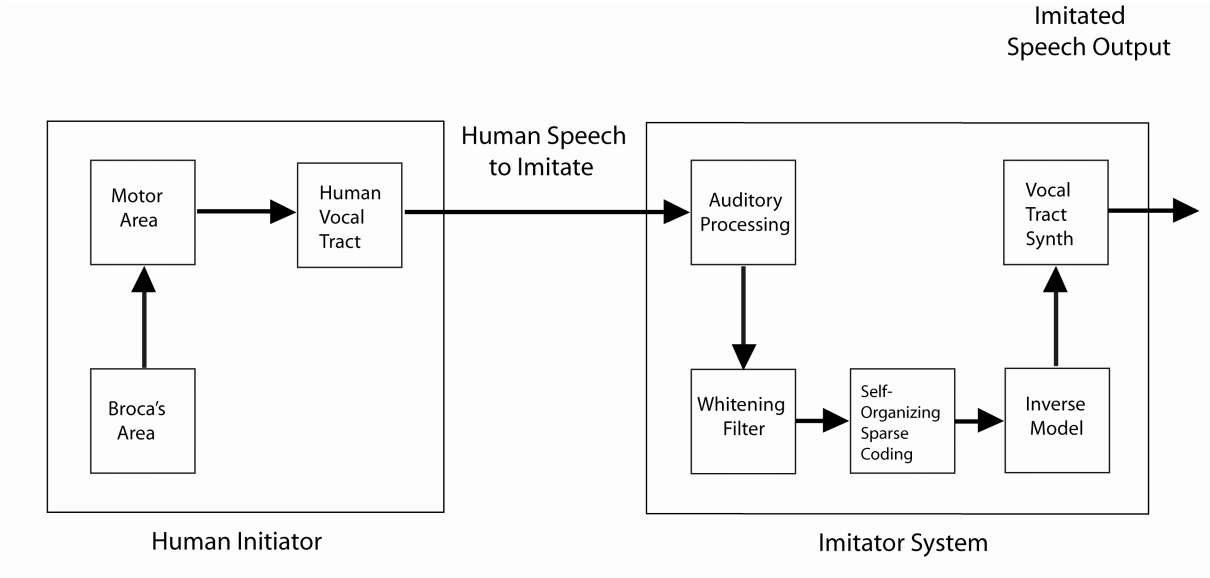


Figure 9: Imitating speech using sparse coding of the auditory representation.

9. Future work

Using our current system architecture, there are several technical improvements that could be made. The feed forward network used to implement the inverse model only makes use of a relatively short time window on the acoustic data. In addition, this network performed a memory-less mapping from input to output and did not take the continuity in the output trajectories into account. That is, it made an estimation of the articulator trajectories without using any prior knowledge regarding their dynamics. A consequence of this was that spikes were sometimes generated in its output signal, and these had to be smoothed out using post-processing. Such issues are elegantly addressed by Kalman filtering techniques, which is also worth of investigation for this kind of task and has been used for similar decoding tasks in many fields (e.g., Wu et al., 2004).

The simple inverse model used here to map between speech and the control of its articulator synthesiser obviously differs quite considerably on what is going on inside an infant's brain during speech acquisition. Firstly, in our scheme, the babble generation is totally separate from the imitation process. In a developing child, it is likely that the mechanisms that cause initial speech production (i.e. cooing and then babbling) are adapted over time to match its linguistic environment (i.e. to the speech to which the infant is exposed) and that these mechanisms develop into ones that are later used for the production of linguistically significant utterances. A more biologically relevant modification to our current scheme would thus be to include an abstract speech generator

stage. Initially this would be configured to generate only the simplest of speech sounds, such as those involved in babbling. It should then develop, due to the system's exposure to external speech and feedback of its own speech, to produce more complex speech utterances. As mentioned previously, the best source of internally generated articulator movements to train the inverse model would be those used for the generation of real speech. It would therefore seem advantageous if the training of the inverse model should be an ongoing process (and actually never stop). One may expect its performance to improve as the speech generation process also improved. At the same time there also should be ongoing adaptation sparse coding stages. These issues are currently being investigated and are depicted graphically in figure 10.

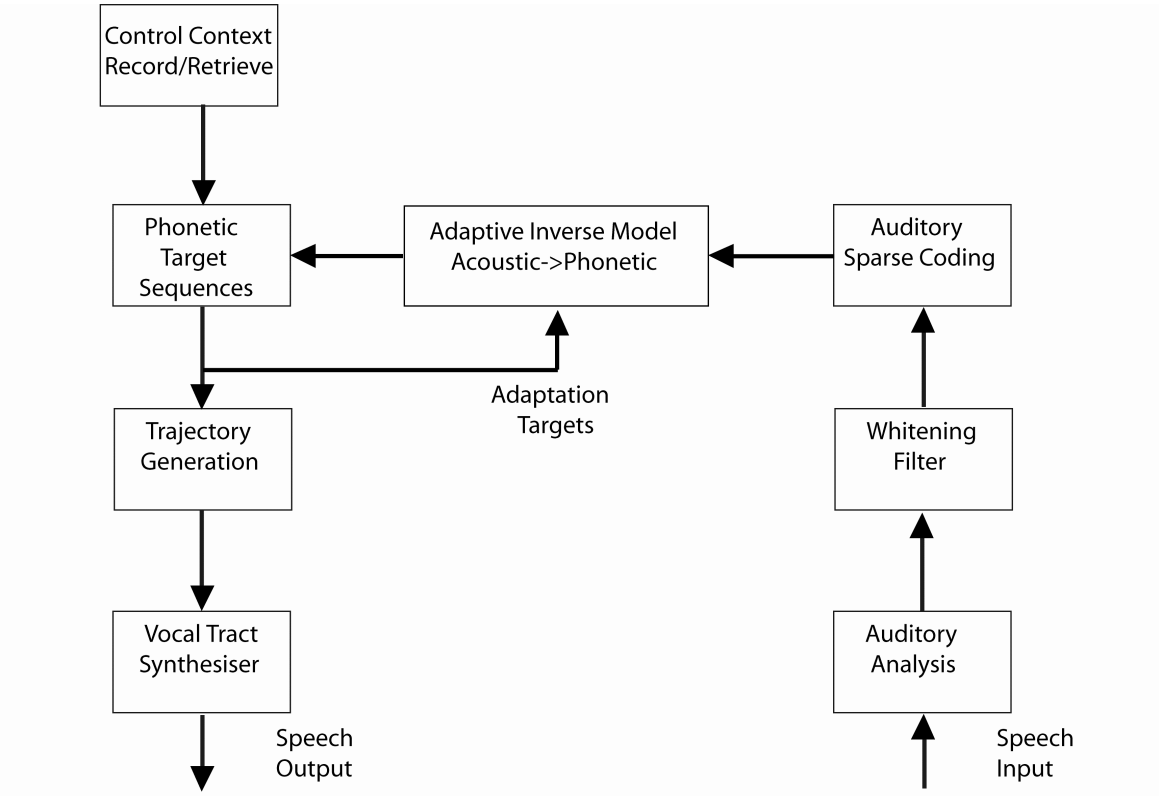


Figure 10: Improved system structure. The same generation process produces babble and then develops into a linguistic speech production mechanism. The inverse model and acoustic features detection are continually adapted and improved, not only during a separate babbling phase.

Acknowledgements

We wish to thank Daniel Wolpert for supporting this work. The implementation of the articulator synthesizer was based on an implementation by Shinji Maeda

within the DOS program VTCALCS. We wish to thank Pascal Perrier and an unknown reviewer for commenting on the manuscript.

References

Bailly, G. (1997). Learning to speak. Sensori-motor control of speech movements,'' *Speech Commun.* 22: 251–267.

Barlow H.B. (1961). Possible principles underlying the transformation of sensory messages. In Rosenblith, W. (ed.) *Sensory Communication*. M.I.T. Press, Cambridge MA.

Guenther, F. H. (1994.) A neural-network model of speech acquisition and motor equivalent speech production. *Biol. Cybern.* 72: 43–53.

Guenther, F. H. (1995). 'Speech sound acquisition, coarticulation, and rate effects in a neural-network model of speech production. *Psychol. Rev.* 102: 594–621.

Holmes, J.N. (1980). The JSRU Channel Vocoder. *Proc. IEEE*, 127, Pt. F, 53-60.

Jordan, M.I., and Rumelhart, D.E. (1992). Forward models—Supervised learning with a distal teacher. *Cogn. Sci*. 16:307–354.

Lewicki, M.S. (2002). Efficient coding of natural sounds. *Nature Neurosci.* 5(4):356-363.

Maeda, S. (1990). Compensatory articulation during speech: evidence from the

analysis and synthesis of vocal tract shapes using an articulatory model. In Hardcastle W.J. and A. Marchal (eds.) *Speech production and speech modelling*. Kluwer Academic Publishers, Boston. p.131-149.

Nabney, I. and Bishop, C. (1995). Netlab: Netlab neural network software. http://www.ncrg.aston.ac.uk/netlab/.

Olshausen B.A. and Field D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*. 381(6583): 607-9.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986) Learning representations by back-propagating errors. *Nature* 323: 533-536.

Wolpert DM. (1997) Computational approaches to motor control. *Trends in Cognitive Sciences*. 1(6): 209-216.

Wu, W., Shaikhouni, A., Donoghue, J. P., and Black, M.J. (2004). Closed-loop neural control of cursor motion using a Kalman filter. *Proc. IEEE Engineering in Medicine and Biology Society*: 4126-4129.

A visual articulatory model and its application to therapy of speech disorders: a pilot study

Bernd J. Kröger
Department of Phoniatrics, Pedaudiology and Communication Disorders, University Hospital of RWTH Aachen, Germany

Julia Gotto
Department of Neurology, Neurolinguistics, University Hospital of RWTH Aachen, Germany

Susanne Albert
Christiane Neuschaefer-Rube
Department of Phoniatrics, Pedaudiology and Communication Disorders, University Hospital of RWTH Aachen, Germany

A visual articulatory model based on static MRI-data of isolated sounds and its application in therapy of speech disorders is described. The model is capable of generating video sequences of articulatory movements or still images of articulatory target positions within the midsagittal plane. On the basis of this model (1) a visual stimulation technique for the therapy of patients suffering from speech disorders and (2) a rating test for visual recognition of speech movements was developed. Results indicate that patients produce recognition rates above level of chance already without any training and that patients are capable of increasing their recognition rate over the time course of therapy significantly.

1. Introduction

Applications for phonetic models of speech production are rare. Basically these models are used as research tools in phonetics and phonology. However more practical applications like high quality speech synthesis (see for example Birkholz et al. 2003 and this volume) may be aimed for. The visual articulatory model described here serves as a generator for visual stimuli used in therapy of speech disorders as suggested by Heike et al. (1986).

2. Visual articulatory model

The visual articulatory model developed in Cologne and Aachen (Kröger 1998 and 2003) is a geometrical model based on a set of static MRI-data. It generates still images for single sounds and articulatory movement sequences (i.e. animations, video sequences) for syllables, words and sentences. The basic concept of this model is the separation of vocalic and consonantal articulation which allows a straightforward modelling of coarticulation.

2.1. Static MRI-data: the basis of the model

A corpus of midsagittal articulatory MRI-data of one speaker (JD) was collected for isolated German long vowels, nasals, fricatives, and the lateral (Kröger et al. 2000). *Edge contours* were extracted manually for all articulators (i.e. lips, tongue, jaw, palate, velum, pharyngeal wall, larynx) and all sounds from these MRI data using a predefined number of contour points per articulator (e.g. 23 contour points for tongue body, see fig. 1).

Figure 1: MRI-data for edge contours of vowels [i:], [a:], and [u:]. Edge contour points are indicated by white dots for different articulators. Contour points of tongue body are indicated by grey points.

2.2. Modelling vocalic and consonantal articulation

Functional control parameters (table 1) are predefined with respect to vocalic and consonantal articulation. Articulator positions (figure 2) can be described by sets of control parameter values and thus articulatory movements by time variation of control parameter values (figure 3). Model contours corresponding to definite sets of control parameter values are calculated by interpolation from edge contours.

Three functional vocalic parameters are defined: close-open, front-back, and unrounded-rounded (table 1, parameters 1 to 3). Interpolation of vocalic midsagittal contours is done for each contour point on the basis of three edge vowels [i:], [a:], and [u:]. It can be seen from the data (figure 1), that the variation of vocalic parameters is *global*, i.e. affects all articulators (lips, tongue, jaw, velum, larynx). Furthermore, vocalic parameters are *absolute*, since we can relate vocalic parameter values to vocalic midsagittal contours in an one-to-one relation.

Table 1: Parameters of the model (name and abbreviation), correlated basic articulatory gestures, and examples for resulting sounds for Standard German.

name	**abbr.**	**correlated basic articulatory gestures**	**examples (German)**
(1) close-open	voc ↕	vocalic raising / lowering of tongue (with lower jaw)	[a]↔ [i] [a]↔ [u]
(2) back-front	voc ↔	fronting / backing of tongue	[u]↔ [i] (and [x] ↔[ç])
(3) unrounded-rounded	lips = O	unrounding / rounding of lips	[i]↔ [y], [a]↔ [u]
(4) lips: vocalic-closure	lab ↕	consonantal closure / opening of mouth (lips)	[b, p, m], [f, v] („Abi, ab, am, auf, wahr“)
(5) tongue body: vocalic - closure	dors ↕	consonantal raising / lowering of tongue body	[g, k, ŋ, x], [ç] („Egge, Acker, Enge, ach, ich“)
(6) tongue tip: vocalic - closure	apic ↕	consonantal raising / lowering of tongue tip	[d, t, n, l] („da, Hut, an, All“)
(7) tongue tip: back - front	apic ↔	fronting / backing of tongue tip (place of artic.: dental / alveolar / postalveolar)	[s, z], [ʃ, ʒ] („Ass, sah, Asche, Genie“)
(8) velum: lowered - raised - strongly raised	vel ↕	lowering (for nasals) / raising (for non-nasal sonorants, e.g. vowels) / strong raising (for obstruents)	[m, n, ŋ] ↔ vowels vowels ↔ [b, d, g]
(9) glottis: tightly closed - closed - opened - opened widely	glott ↔	tight closure for [ʔ], closed for phonation, opened for voicelessness, opened widely for breathing	[ʔ] („ Aa“) ↔ phonation (e.g. vowels) ↔ voiceless sounds (e.g. [p, t, k, f, s, ʃ, ç, x, h]) ↔ breathing

In contrast, consonantal articulation affects *local* functional regions (e.g. lips, tongue tip, tongue body including lower jaw) and is modelled in this approach by defining three parameters for oral closure controlled by lips, tongue tip and

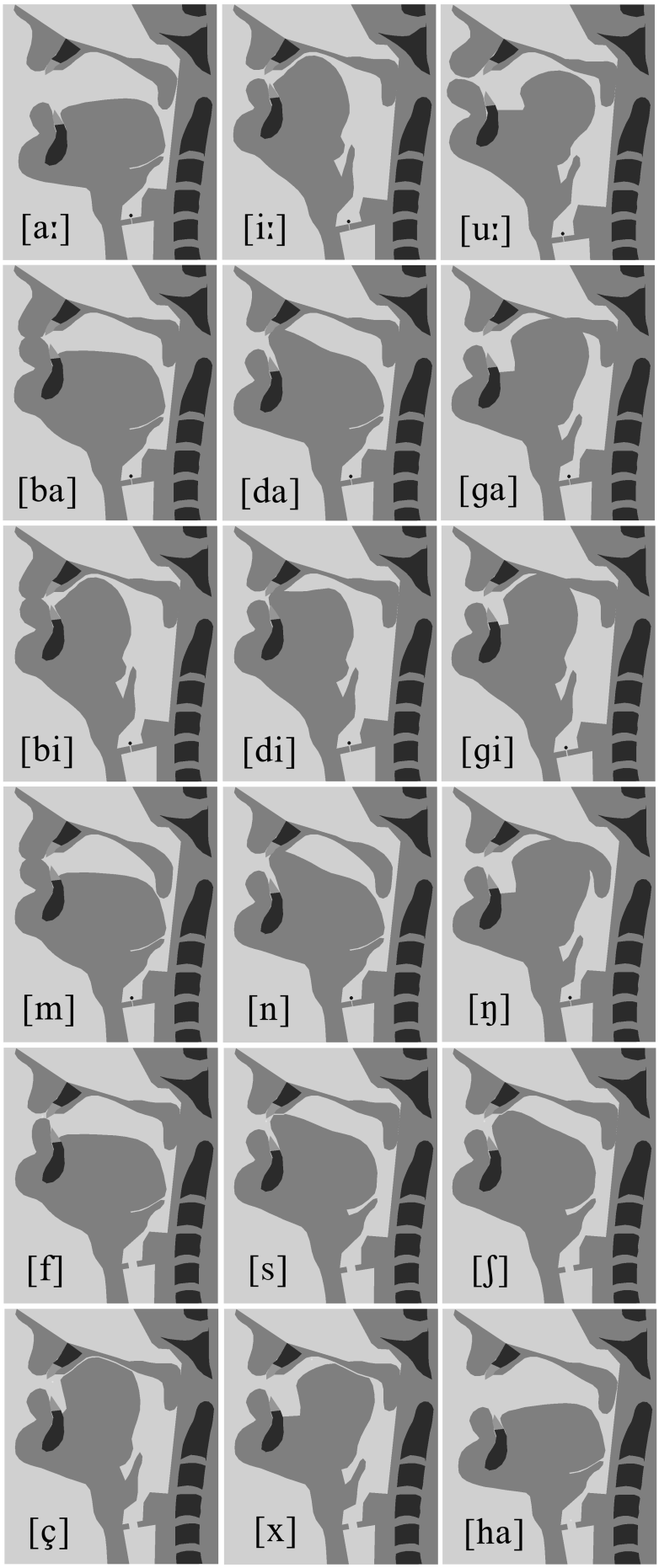


Figure 2: Midsagittal views of edge vowels [i:], [a:], and [u:] and consonants (plosives, nasals and fricatives). The plosives are given in the context of [a:] and [i:]. Other consonants are given in context [a:].

tongue body respectively (table 1, parameters 4 to 6). Interpolation of consonantal midsagittal contours is done for each contour point of the active consonantal articulator on the basis of the actual vocalic contour on the one hand and a consonantal edge contour for lips (bilabial closure), tongue tip (alveolar closure), or tongue body (velar closure) on the other hand. This is sufficient for modelling plosives.

In the case of fricatives a labio-dental obstruction contour is used instead of the bilabial stop contour (figure 2). In all other cases the consonantal constriction is approximated by the consonantal closure contours. This approximation should be replaced by using dynamic MRI-data for fricatives collected recently (Kröger et al. 2004). The control of place of articulation in the case of dorsal fricatives is done by the back-front-parameter (table 1, parameter 2). In the case of apical and laminal fricatives a further control parameter is introduced (table 1, parameter 7).
The control of velum and glottis is accomplished by separate parameters (table 1, parameter 8 and 9). For the velum in addition to edge values a mid-range value (i.e. “raised”) is defined. The corresponding velum contour is interpolated with respect to the vocalic edge contours (see figure 1 and 2). The appertaining velum position represents the non-obstruent raising, which leads to different positions of the velum, e.g. for high vs. low vowels. For glottis articulation the mid-range value (i.e. “closed”) represents a light glottal closure as needed for normal phonation.
Consonantal parameters are local as described above and *relative* since the interpolation of contour points for consonant articulation depends not exclusively on consonantal edge contours but also on an actual underlying interpolated vocalic contour. Thus consonantal closure results in smaller closing gestures for example in [i]- vs. [a]-context (figure 2, row 2 vs. row 3).
According to this separation of vocalic and consonantal articulation, modelling of a time course of a word (or sentence) is relatively simple. At first a score of independent vocalic and consonantal articulatory features is defined in terms of the model parameters (figure 3, upper panel). Subsequently these features are used for specification of spatio-temporal targets for each control parameter and each sound (figure 3, lower panel, see rectangles). Then the vocalic targets are simply connected by monotonic trajectories, thus producing slowly varying tongue body (and lip rounding) movements (figure 3, lower panel). Consonantal parameters reset to zero if a target position needs no longer to be hold by an articulator.

	[k]	[ɔ]	[m]	[p]	[a]	[s]
length		[-long]			[-long]	
voc ↕		[-close] [-open]			[+open]	
voc ↔		[+back]			[+back]	
lips =O		[+rnd]			[-rnd]	
lab ↕	[/]		[bilab] [-cont]	[bilab] [-cont]		[/]
dors ↕	[velar] [-cont]		[/]	[/]		[/]
apic ↕	[/]		[/]	[/]		[postalv] [+cont]
apic ↔	[/]		[/]	[/]		[+ant]
vel ↕	[-son]	[-nas]	[+nas]	[-son]	[-nas]	[-son]
glott ↔	[+open]	[+voice]	[+voice]	[+open]	[+voice]	[+open]

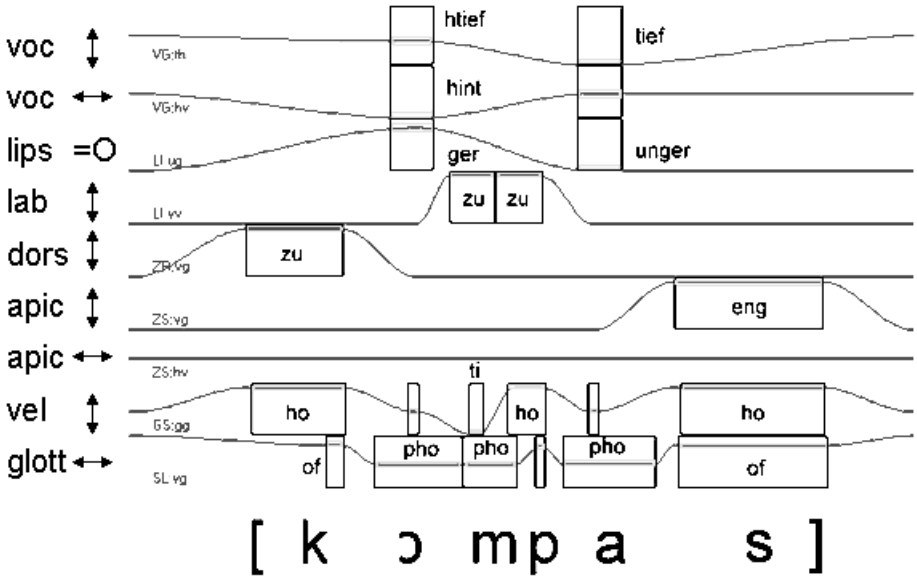


Figure 3: Score of articulatory features (upper panel) and articulatory plan including phonetic transcription (lower panel) for the German word "Kompass". For definition of parameters see table 1.

2.3. Modelling coarticulation

A huge amount of coarticulatory movements results from articulatory underspecification in our approach and is referred to as *consonant-vowel-coarticulation*. Since vocalic and consonantal features are defined by different parameters (i.e. vocalic or consonantal parameters), free slots occur in the table

of feature specification for each word (figure 3 upper panel, empty slots). In addition consonantal features are solely specified for the actual constriction forming oral articulator i.e. lips, tongue body, or tongue tip (see free slots marked by [/] in figure 3, upper panel). Thus (i) the basic vocalic movements are relatively uninfluenced by consonants and (ii) all articulators not primarily involved in producing oral closure or constriction are free for coarticulatory variation according to the underlying vocalic movements. This can be exemplified by comparing the German words "Kompass" vs. "Kampus" (figure 4). For these words only the vocalic parameter trajectories are different (figure 4 top) leading to different coarticulatory forms of tongue body and lip rounding during the consonantal closure of [m] and [p] in both words (figure 4 bottom).
Furthermore the articulatory plans displayed here clarify the parallelism of the terms "coarticulation" and "temporal gestural overlap" (e.g. Browman and Goldstein 1992). Figure 4 indicates that vocalic gestures are activated during the preceding consonantal closure or constriction periods: Vocalic movements mainly occur during consonantal production periods and many consonantal movements occur during the target period of vowels.

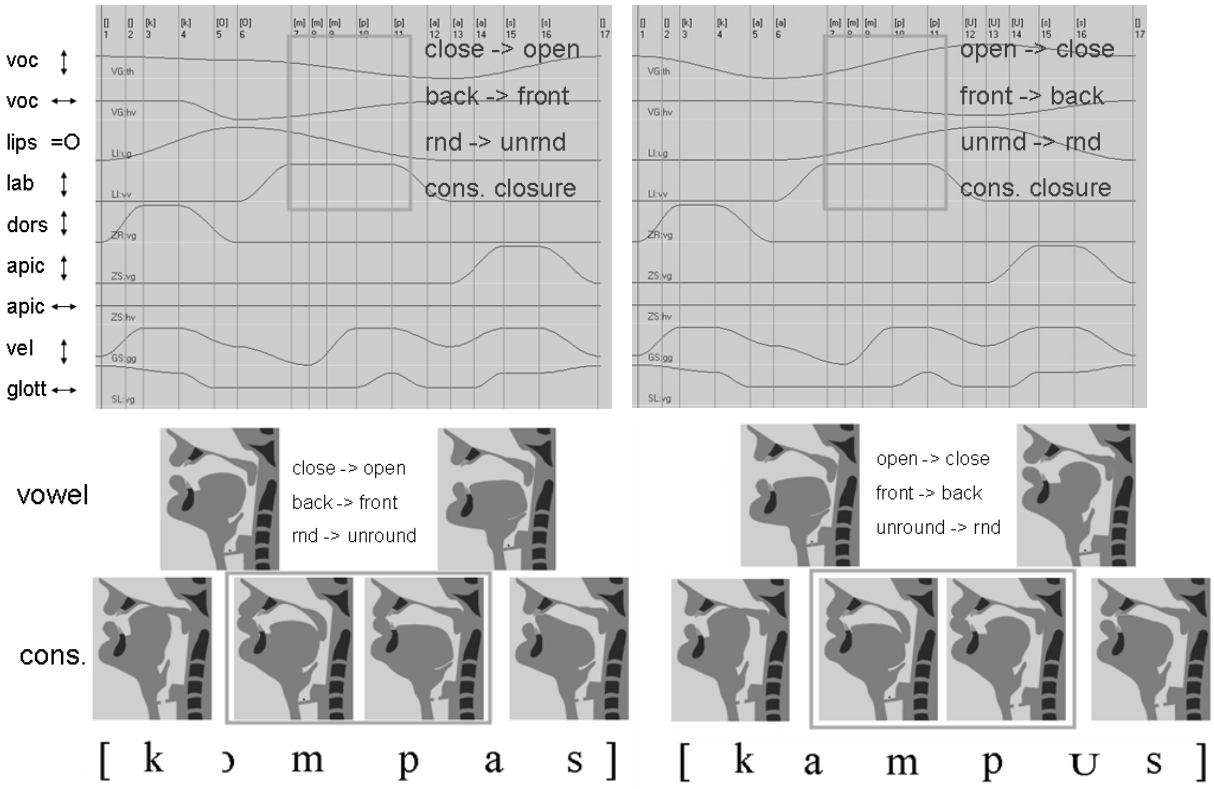


Figure 4: Articulatory plan (top) and midsagittal views of the model articulator positions in the temporal centre of the sounds for the German words "Kompass" (left) and "Kampus" (right). The underlying vocalic articulatory movements during the bilabial closure of [m] and [p] are marked by rectangles within the articulatory plans.

Further coarticulatory movements result from *synergistic movement effects* (e.g. Lindblom 1983) also modelled in this approach. Focussing again on the example of a lip closing gesture in [a]-context, the complete movement needed for lip closure can be divided into cooperative partial movements of upper lips, lower lips and lower jaw. Thus the lower jaw participates in the lower lip raising movement as can be seen in figure 2 for [a:] vs. [ba]. This lower jaw raising caused by the lip closing gesture leads also to a passive raising of the front part of the tongue body and thus to tongue body coarticulation in the case of consonantal lip closure.

2.4. User-friendly interface "SpeechTrainer"

The main idea for application of this model is its use in therapy of speech disorders as a visual stimulation technique. Thus it is very important to integrate the visual articulatory model in a user-friendly interface in order to allow the model to be controlled without effort by therapists and patients.
The model can be controlled by prompting broad phonetic transcriptions for the generation of syllables, words or sentences, by prompting orthographic text (converted via a text-to-allophone-module), or by clicking sound symbols in a phonetic symbol chart. Furthermore word lists can be created by users in order to apply specific word lists to various disorders or to various patients with different kinds and different degrees of dysfunctions. The articulatory movement patterns generated by the model can be displayed in different speech rates (fast, normal, and various degrees of slow motion). Furthermore the animations generated by the model can be synchronised with natural acoustic signals recorded by users. This programm ("SpeechTrainer") can be used as freeware by therapists and researchers (see: http:\\www.phoniatrie.ukaachen.de > research > speechtrainer).

3. Application in therapy of speech disorders

The main idea for application of this model is its use in therapy of speech disorders as a visual stimulation technique. At the moment the visual articulatory model is applied in therapy of developmental speech disorders (i.e. articulation disorders, Albert 2005) and in therapy of neurogenic speech disorders (i.e. apraxia of speech, Gotto 2004 and dysarthria, O'Neill 2004). Patients with articulation disorders show deficits in producing definite single sounds (e.g. velar plosives). Patients with apraxia of speech exhibit deficits in syllable or word production while the production of isolated sounds is nearly unproblematic. The latter patients have deficits at the level of motor planning of articulatory movements, e.g. coproduction of articulatory gestures (McNeil et al.

1997). In therapy of both types of disorders, the visual information (i.e. the mediosagittal still images for single sounds or video clips of speech movements for syllables and words) are used in addition to the auditive and visual information naturally produced by the therapist (i.e. acoustic speech output and synchronous view of the face). But this visual stimulation technique using the articulatory model is not meant as an independent therapy method; rather this technique should be integrated as one part in comprehensive therapy concepts.

In the case of *velar consonants*, patients profit from mediosagittal visual information of consonantal target positions, since the acoustic cues for place of articulation are very complex and eventually not prominent enough. Furthermore the natural facial visual information does not contain prominent cues for the rear places of articulation. Thus the mediosagittal visual information of the model is a non-redundant perceptual information.

In the case of apraxia of speech patients profit from the visual perception of the *dynamics* of articulatory movements instead of viewing still images (i.e. consonantal target positions). In the case of the visual articulatory model described here, patients get an impression of coarticulatory overlap of vocalic and consonantal speech movements. Thus the movement sequences (i.e. animations, video sequences) generated by the visual articulatory model help to rebuild patients articulatory plans for syllables as well as for words or sentences.

At the moment it is still unclear how the visual stimulation technique works. It can be argued that these visual signals are too complex for the patient. Speech movements are mainly acquired by auditory perception using auditive sensorimotor circuits and thus our perception system is not trained in processing these complex visual stimuli in the case of the speech production and speech perception mode. However, due to the existence of multimodal perception-production links (Liberman and Mattingly 1985) a profitable use of visual perception of speech gestures seems to be of particular importance for the treatment of various pathological speech disorders.

Two case studies (Gotto 2004 and Albert 2005) were carried out in order to evaluate the benefit of this stimulation technique. Unfortunately according to the huge variety of factors influencing therapy effects we were not able to prove within these pilot studies that patient - using this visual stimulation technique as a part of their therapy - profit significantly more by this therapy than patients using conventional therapy techniques. As a first step we created a rating method for visual recognition of speech movements measuring the increase in recognition rate over the time period of therapy. An increase in patients recognition rate indicates interaction between patient and model and makes feasible that patients benefit from the visual stimulation technique.

3.1. Rating test

A rating test was performed by presenting mute visual stimuli produced by the model (i.e. 4 corner vowels [i], [a], [u], [y], 11 consonants [p], [t], [k], [m], [n], [ŋ], [f], [s], [ʃ], [ç], [x] in the case of articulation disorders and 4 vowels, 6 consonants [p], [t], [k], [f], [s], [x] and 10 syllables [pa:], [ti:], [ke:], [fu:], [sa:], [u:x], [ksa:], [u:st], [pfi:], [u:xt] in the case of apraxia of speech) in randomised order. Patients were asked to mime each visual stimulus acoustically (verbal realisation). Patients' responses were recorded (DAT, 44,1 kHz, 16 bits, mono) and phonetic transcriptions were accomplished afterwards by the examiner.
In the case of apraxia of speech the distance between patients' dislocated realisations and target realisations is quantified with respect to a four-level scale following Huber et al. (1983) (table 2). In the case of patients with articulation disorders, patients' responses were differentiated with respect to a system of articulatory-visual features (table 3). In comparison to an undifferentiated quantification (i.e. target sound correctly reproduced or not), this kind of quantification for realisation-target-distance leads to more detailed results, since patients often recognise some articulatory-visual features correctly albeit the target segment as a whole is not recognised. For example if the patient produces a [t] as reaction on the visual item [n], 4 of 6 articulatory-visual features were identified correctly (i.e. narrowness: closed, articulator: tongue tip, rounding: not rounded, place of obstruction: alveolar ridge, see table 3).

Table 2: Quantification of patients' responses with respect to a four-level scale (case: apraxia of speech).

patients performance	**score**
no reaction or evasive answer	0
dissimilar target-sound/-syllable: consonants: wrong place *and* manner of articulation vowels: vowel quality is *strongly* aberrant	1
similar target-sound/-syllable: consonants: wrong place *or* manner of articulation vowels: vowel quality is *slightly* aberrant	2
target-sound/-syllable is produced correctly	3

Table 3: System of articulatory-visual features for quantification of patients' responses (case: articulation disorders). From left: Area of appearance of the articulatory-visual feature, name of feature, possible specification values for each feature.

area	**feature**	**specification**		
		(1)	**(2)**	**(3)**
oral region	narrowness of obstruction	open	narrow	closed
oral region	articulator	lips	tongue body	tongue tip
lips	rounding	rounded	not rounded	-
tongue	place of obstruction	hard palate or alveolar ridge	soft palate	pharyngeal wall
velum	nasality	nasal	non-nasal	-
larynx	voice	voiced	voiceless	-

3.2. Results

For both types of speech disorders the rating test was carried out at begin, middle, and end of the therapy period, and at a follow-up test a few months after therapy was completed. All patients had no experience with midsagittal views at the beginning of the therapy. Therefore in the case of the initial rating test, patients were introduced to midsagittal views by explaining the articulators (i.e. lips, tongue, velum, pharyngeal wall, larynx) and by explaining the basic articulatory movements (i.e. closure or constriction produced by lips, tongue tip, tongue body; fronting, backing, lowering, and raising the tongue; spreading and rounding of lips; raising and lowering of velum; opening and closing of glottis) produced by our model.

Two patients with articulation disorders (child 1: female 4;6 years old, child 2: male 4;7 years old) used the visual articulatory model during therapy (Albert 2005). The therapy period lasted three months with two one-hour therapy sessions per week. The visual stimulation technique was applied in the first half of the therapy period for child 1 and in the second half of the therapy period for child 2. The rating test was performed at begin, middle, and end of the therapy period and a follow-up testing was performed 6 weeks after end of therapy. The results are displayed in figure 5.

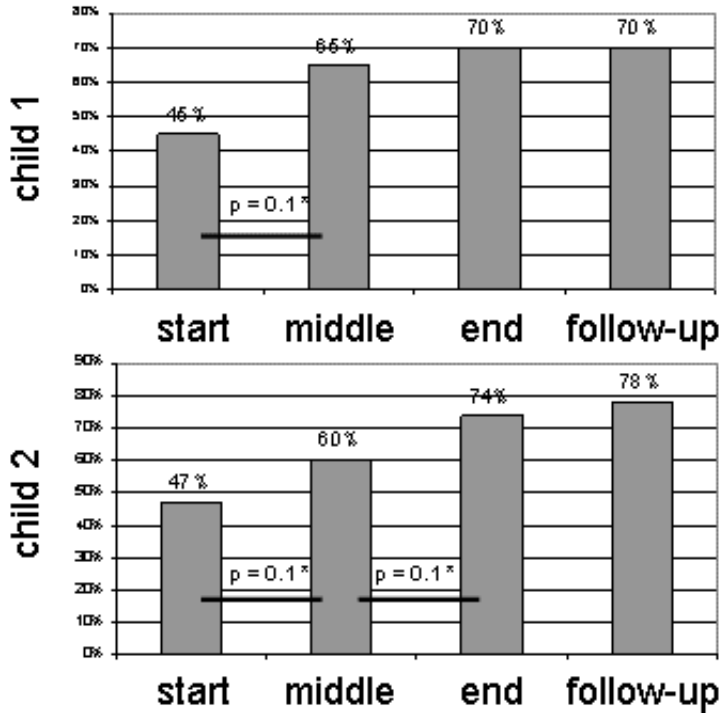


Figure 5: Recognition score in % at begin, middle, end of therapy and at a follow-up testing (from left to right) for two children. Differences reaching significance are indicated by horizontal lines (Willcoxon-sign-rank-test two-tailed).

In addition 9 other children (3 male, 6 female, age from 4;7 to 8;3) were tested at begin of the therapy period exclusively (Albert 2005). Thus, in total 11 children were tested: 7 children suffered from articulation disorders without further deficits, while 4 children exhibited articulation disorders as a part of a specific language impairment. Results of this initial rating test is given in table 4. Recognition scores are round about 40% to 70% in the case of differentiated quantification (mean value) and about 10% to 30% in the case of undifferentiated quantification. Mean values of recognition rate of the 11 children separated for each visual-articulatory feature are given in figure 6. It can be seen, that the feature "lip rounding" is recognised best. This may be related to the fact, that this lip feature is well known from normal (i.e. facial) visual speech perception. The articulatory-visual feature "place of obstruction" refers to vocalic and consonantal tongue articulation (i.e. palatal or /i/-like, velar or /u/-like, pharyngeal or /a/-like articulation). This feature is identified second best. This may be related to the fact that the change of tongue contour is visually prominent in the midsagittal view (see figure 1, row 1).

Table 4: Recognition rate in % for visual articulatory features (see table 3), for mean value of differentiated rating, and for non-differentiated rating in the case of 11 children suffering from articulations disorders. The rating test was done before starting the therapy. Children with specific language impairment are indicated by asterisk. Child 1 and 2 are also described in figure 5.

child	age	narrowness	articulator	rounding	place	nasality	voice	differentiated (mean)	undifferentiated
1	4;6	40	33	64	42	44	47	45	11
2	4;7	40	53	60	44	47	40	47	13
3	4;7 (*)	51	64	78	71	78	44	64	20
4	4;11 (*)	38	31	58	40	38	31	39	11
5	5;5	47	69	76	80	62	44	63	24
6	5;9 (*)	60	73	87	78	71	49	70	22
7	5;9	49	62	76	64	44	49	57	29
8	6;5 (*)	49	62	73	64	49	51	58	22
9	6;6	44	62	68	64	51	36	54	13
10	7;9	58	77	73	82	67	44	67	27
11	8;3	56	64	73	64	69	58	64	27

The features identified next best are “articulator” and “nasality” followed by the features “narrowness of obstruction” and “voice”. Especially the specification of the last two features corresponds to very small changes only within the medio-sagittal view of the visual articulatory model and thus may be not noticed by patients.

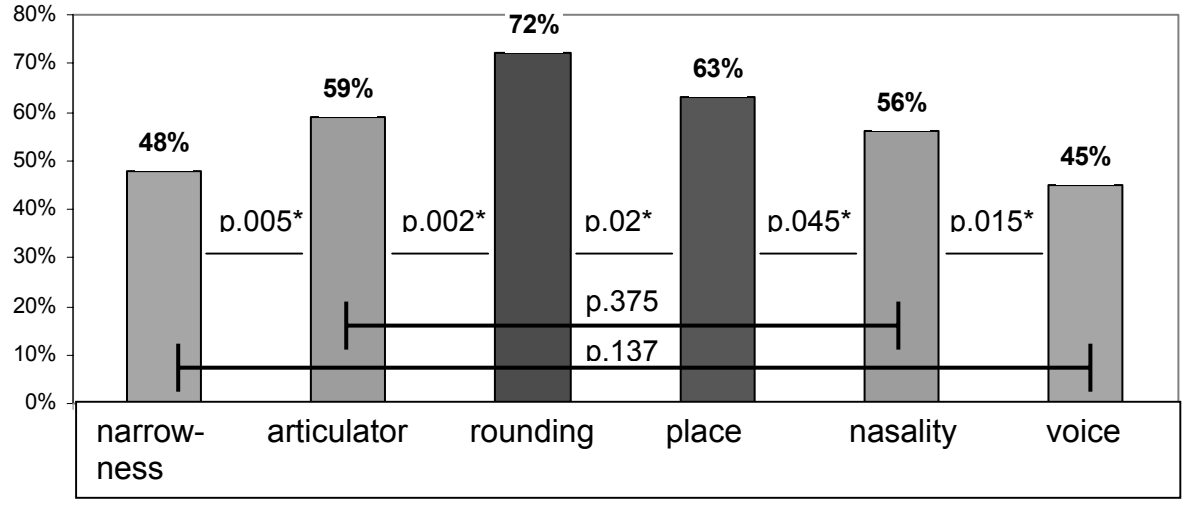


Figure 6: Mean values of recognition rate for visual articulatory features for all 11 children described in table 4. The rating test was executed before starting the therapy. Differences reaching significance were indicated by horizontal lines (Wilcoxon-sign-rank-test two-tailed).

Furthermore these two features are relatively abstract and patients may be unable to interpret these features in the articulatory domains. But it is an important result of this study, that features like "nasality" or "place of obstruction" - which are not easily understandable in the visual domain - are specified correctly above level of chance without any training.

In the case of apraxia of speech our visual stimulation technique was used in therapy of a 47-years-old female patient (Gotto 2004). This patient suffered from a severe Broca-aphasia in combination with a severe apraxia of speech, resulting from a left ischemic MCA infarction, occuring 5;7 years before beginning of the therapy described here. The patient exhibited no neuropsychological deficits, especially no deficits in visual perception and visual processing. The period of therapy lasted two months with five one-hour sessions per week. The rating test was performed at the begin, middle, and end of the therapy period and at a follow-up testing three months after end of therapy. The results are displayed in figure 7. The recognition score significantly increased over the time period of therapy from 33% to 67% and remained stable afterwards (follow-up-rate: 58%). If results were separated with respect to recognition of segments (single sounds) and recognition of syllables, it can be seen clearly that the patient mainly increased the recognition rate with respect to syllables. This result is not surprising, since patients suffering from apraxia of speech often show deficits with respect to production of sound sequences, while single (i.e. isolated) sounds are produced with less problems.

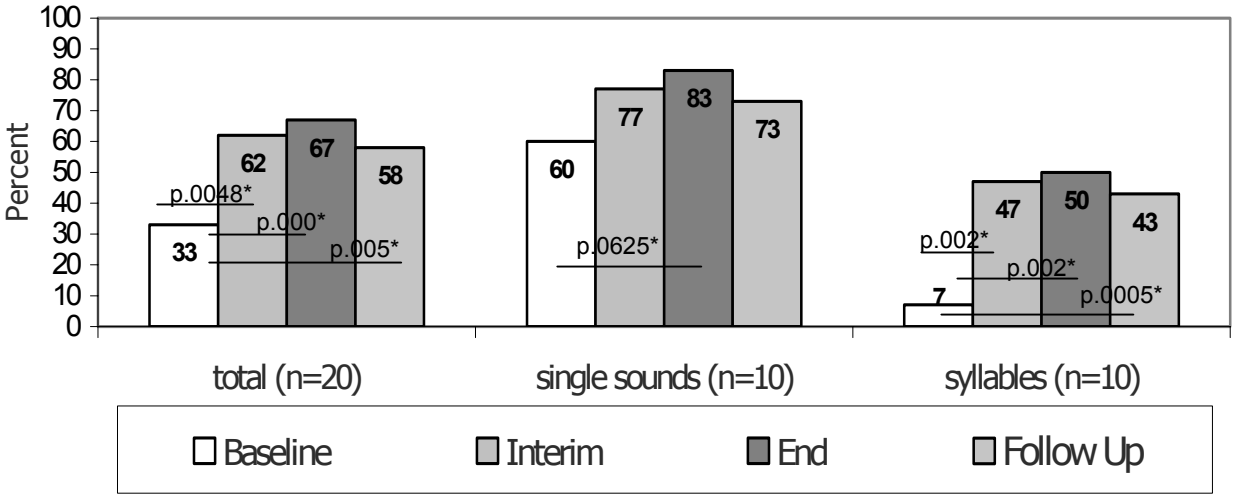


Figure 7: Recognition score in % at the begin, middle, end of the therapy period and at a follow-up testing (from left to right by different tones). The recognition score is displayed in total (left side) and separated with respect to segments (middle) and syllables (right side). Differences reaching significance were indicated by horizontal lines (Wilcoxon-sign-rank-test one-tailed).

4. Discussion and perspective

A visual articulatory model comprising basic coarticulatory mechanisms has been introduced. Articulator positions are calculated by interpolating edge contours extracted from MRI-data. The model displays segmental articulatory targets as still images and articulatory movement sequences as video sequences. The model is used as a visual stimulation technique in therapy of articulation disorders and apraxia of speech. A rating test was developed in order to evaluate the increase in visual recognition of sounds and syllables over the time course of therapy. In both kinds of speech disorders a significant increase in recognition rate is found. Even in the case of young children we were able to demonstrate that recognition of articulatory-visual sound features is unproblematic at their first exposure to the visual articulatory model (i.e. without training).

A major disadvantage of the rating test described here, is that the visual recognition scores are also influenced by the production abilities of the patients. In order to get a better separation of production and perception abilities, it is planned to design acoustic-visual matching experiments. Since the results of this study have promise, we are going to use this visual stimulation technique in therapy with a larger amount of patients. The goal is to evaluate the effectiveness of this therapeutic aid as well as an estimation of therapy duration effects with and without using this visual stimulation technique in the therapy of speech disorders.

References

Albert, S. (2005). Einsatz des SpeechTrainers in der Artikulationstherpie bei Kindern. Diplomarbeit, Studiengang Lehr- und Forschungslogopädie, RWTH Aachen.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49: 155-180.

Birkholz, P., & Jackèl, D. (2003). A three-dimensional model of the vocal tract for speech synthesis. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain. 2597-2600.

Gotto, J. (2004). Therapie der Sprechapraxie: Eine Einzelfallstudie zum PC-Programm SpeechTrainer. Diplomarbeit, Studiengang Lehr- und Forschungslogopädie, RWTH Aachen.

Heike, G., Philip, J. & Hilger, S. (1986). Computergrafikdarstellung von Artikulationsbewegungen zur Unterstützung des Artikulationstrainings. *Sprache - Stimme - Gehör* 10: 4-6.

Huber, W., Poeck, K., Weniger, D. & Willmes, K. (1983) *Aachener Aphasie Test (AAT).* Göttingen: Hogrefe.

Kröger, B.J. (1998). *Ein phonetisches Modell der Sprachproduktion.* Tübingen: Niemeyer.

Kröger, B.J. (2003). Ein visuelles Modell der Artikulation. *Laryngo-Rhino-Otologie*, 82: 402-407.

Kröger B.J., Winkler, R., Mooshammer, C. & Pompino-Marschall, B. (2000). Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results. *Proceedings of 5th Seminar on Speech Production: Models and Data*. Kloster Seeon, Bavaria. 333-336.

Kröger, B.J., Hoole, P., Sader, R., Geng, C., Pompino-Marschall, B., & Neuschaefer-Rube, C. (2004). MRT-Sequenzen als Datenbasis eines visuellen Artikulationsmodells. *HNO*, 52: 837-843.

Liberman, A. M. & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition,* 21: 1-36.

Lindblom, B. (1983). Economy of speech gestures. In: MacNeilage P.F. (Eds.) *The production of speech*. New York: Springer. 217-246.

McNeil, M. R., Donald, A. R. & Schmidt, R.A. (1997). Apraxia of speech: Definition, differentiation, and treatment. In M.R. McNeil (ed.) *Clinical management of sensorimotor speech disorders*. New York: Thieme. 311-344.

O'Neill, G. (2004) The use of a new computer programme in the treatment of dysarthria: A single case studie. Bachelor Thesis, Trinity College Dublin.

Face models based on a guided PCA of motion-capture data: Speaker dependent variability in /s/-/ʃ/ contrast production

Shinji Maeda
ENST/TSI & CNRS/UMR5141, Paris, France

We measure face deformations during speech production using a motion capture system, which provides 3D coordinate data of about 60 markers glued on the speaker's face. An arbitrary orthogonal factor analysis followed by a principal component analysis (together called a guided PCA) of the data has showed that the first 6 factors explain about 90% of the variance, for each of our 3 speakers. The 6 derived factors, therefore, allow us to efficiently analyze or to reconstruct with a reasonable accuracy the observed face deformations. Since these factors can be interpreted in articulatory terms, they can reveal underlying articulatory organizations. The comparison of lip gestures in terms of data derived factors suggests that these speakers differently maneuver the lips to achieve contrast between /s/ and /ʃ/. Such inter-speaker variability can occur because the acoustic contrast of these fricatives is shaped not only by the lip tube but also by cavities inside the mouth such as the sublingual cavity. In other words, these tube and cavity can acoustically compensate each other to produce their required acoustic properties.

1. Introduction

When data consist of a large number of variables having correlation structures between them, a factor analysis becomes effective. Motion capture data on the face deformations during speech production is such a case. In our experiment, a motion capture system measures 3D coordinates of individual markers glued on the speaker's face. Movements of markers are necessarily linked. The position of markers is affected by jaw gestures, by lip gestures like rounding and protrusion, and so on. Each of these gestures would deform the face in a particular way, creating a particular correlation pattern among markers' coordinates. Since the

measured face deformations present the sum of the effects of those gestures involved, markers' coordinates should have the correlation structure as the sum of individual correlation patterns.

Formally, let Y be a data matrix, which consists of observations of a set of variables. Y is centered and then normalized to obtain its Z score as

$$Z=(Y-m)/\sigma, \qquad (1)$$

where m and σ are, respectively, mean and standard deviation vector. A factor analysis only describes variations around the means of individual variables. Then, Z is assumed to be a weighted sum of factors as

$$Z=AX, \qquad (2)$$

where A is a matrix consisting of rows of weights, called factor patterns, specifying degrees of influence of corresponding factors upon individual variables. In other words, the sum of weighted factors presents the observed variations of individual variables. A factor analysis determines the factor pattern A and then values of factors X, *i.e.*, factor scores from A and Z.

The most basic factor analysis method is the principal component analysis (PCA) that determines factors so as to extract the maximum of variances. The derived factors, however, are not always interpretable. In comparison, an arbitrary orthogonal factor analysis (Overall, 1962) followed by PCA helps us to extract a set of interpretable factors from observed data (Badine *et al.*, 2002; Gabioud, 1994; Maeda, 1990). As in the case of PCA, the guided PCA derives an uncorrelated factor set. The total variance, therefore, becomes equal to the sum of variances explained by individual factors. If each of factor patterns can be interpreted in articulatory terms, the linear equation Eq. 2 can be considered as an articulatory model. In this report, we shall describe some examples for demonstrating the usefulness of such a data-derived functional model in analysis of face deformations during speech.

2. A face model based on a guided PCA of motion capture data

One American English male speaker (S1) and 2 French, male and female, speakers (respectively, S2 and S3) read a corpus consisting of a sequence of nonsense VCV syllables and a short text in the corresponding language. These 3 speakers were instructed to read the VCV syllables in a clear and hyper articulated way, and a text with 3 different speaking rates, slow, normal, and

rapid. For English, VCV sequences where V = /i/, /a/, or /u/ and C = one of 24 consonants are used. These 3 vowels and plus the high front rounded vowel /y/ are combined with 20 consonants in the French VCV sequences. The short English text consists of 28 syllables and the French text of 62 syllables.

Maeda et al. (2002) have reported, for S1, the details of the data acquisition and of the guided PCA analysis to extract uncorrelated articulatory factors that efficiently describe the measured face data. The same method was used for the data from the 2 French speakers. Briefly, a Vicon Motion Capture system with 6 infrared video cameras tracked 3D coordinates of 61 markers glued on the S1's face. For the 2 French speakers, 8 cameras tracked 63 face markers. For all the 3 speakers, common 61 face markers were approximately placed at the same relative locations. The camera speed was 120 frames/s for all the 3 speakers.

Before applying the guided PCA, effects of head movements on the position of markers are eliminated by head alignment. Y in Eq.1 becomes a matrix of head-aligned motion data of a speaker. For example, Y of S3 consists of 171 data variables, interlaced 3 coordinates of 57 markers in columns and 23164 frames of observations in rows.

First, we calculate the correlations between data variables (in Z), *i.e.*, a correlation matrix C. A factor analysis determines factor weights A. In the guided PCA, we specify factors to extract particular correlation structures. The first factor (f1), therefore, is determined so that it extracts the correlations between the z-coordinate of a marker on the chin, remaining 2 coordinates of the same marker, and 3 coordinates of all other face markers. Since this z-coordinate can be considered as a measure of the vertical jaw position, f1 represents the effects of vertical close/open jaw motions upon the face including the lips. Then the extracted correlation structure by f1 is subtracted from C. In this way, we determine, step-by-step the second (f2) and then the third factor (f3) representing, respectively, the effects of horizontal front/back jaw motions (along the y-axis) and those of horizontal left/right motions (along x-axis).

It may be interjected here that the skin, on which the chin marker is glued, can slide with up and down jaw motions, possibly causing a discrepancy between the measured marker movements and those of the lower jaw. As long as the marker movements are proportional to those of the jaw, the discrepancy would not affect our linear modeling. Since there is no guaranty about the proportional relationships, it would be more assuring to use a jaw splint (Badin *et al.*, 2002) fixed on the lower front incisors. With the Motion Capture system however, the use of such a device was not recommended, because it would disturb the real-

time automatic detection of markers' coordinates. We here assume therefore that measured chin marker motions are the reasonable representative of the jaw movements at least as a first-order approximation.

Second, the principal factors are determined from the residual of C after the subtractions of those first 3 jaw-related correlation structures. Table 1 summarizes variances explained by each of the first six factors determined for the 3 speakers.

Table 1: Explained variances (%) of the first 6 factors in a guided PCA. Factors in the columns are organized by functions and the corresponding factor numbers are indicated in the parentheses.

	Arbitrary orthogonal factors (Jaw gestures)			PCA (Intrinsic lip gestures)		
Speaker	high/low (f1)	front/back (f2)	left/right (f3)	round/ spread	protrude/ retract	(cheeks) lower/raise
S1 (English)	31	13	11	26 (f4)	7 (f5)	3 (f6)
S2 (French)	34	9	9	17 (f4)	3 (f6)	15 (f5)
S3 (French)	21	23	8	28 (f4)	3 (f6)	7 (f5)

In Table 1, the factors with number f5 and f6 are organized by functions, which we shall discus later. The first 6 factors explain about 90% of the variance for every speaker. Moreover, it is interesting to note that the cumulative variance over 3 jaw-related factors and that over the 3 lip-related factors vary little across speakers, respectively, about 53% and 37%. Nevertheless, there exist fairly important differences in the variance of individual factors across speakers. Note that S2 primarily uses the vertical jaw motion (f1 with 34% of variance) and much less front/back (f2 with 9%) and left/right motions (f3 with 9%). S3 uses more front/back (f2 with 23%) than high/low jaw motion (f1 with 21%). Moreover, S2 uses the cheek lowering/raising (f5 with 15%) to control the upper lip position, as explained latter. These findings suggest speakers employ different strategies in speech production. To make this point clearer, let us describe the effects of each factor upon the face deformation.

The effects of a factor can be visualized by varying the value of each factor from one extreme to the other while that of other factors is kept at zero. Figure 1 shows such visualization for selected factors for S1. The first 2 factors represent the effects of high/low (f1) and of front/back (f2) lower jaw motion in the order of their extractions, respectively in Figure 1a and 1b. The contribution of the vertical jaw motion is primarily lowering and elevation of the lower lip and to a lesser extent of the lip comissures. The center of the upper lip is hardly affected by the jaw motion. The other 2 speakers show similar patterns as the

consequence of jaw lowering and raising. The front/back jaw motion also primarily has an effect on the lower lip, which advances and retracts following jaw displacements. The Speaker S2 also exhibits this pattern, but not S3 as shown in Figure 2.

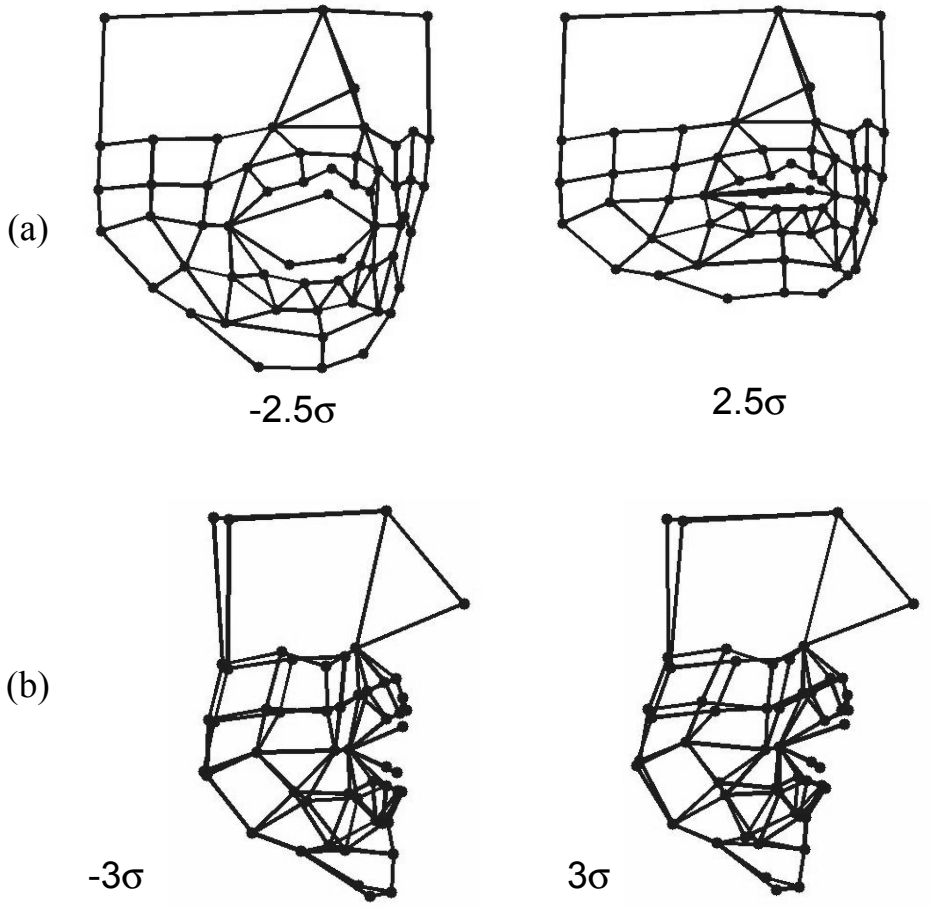


Figure 1: Model face of Speaker S1, where the value of each of 2 factors, f1 in (a) and f2 in (b), is varied from one extreme to the other as indicated while that of other factors is kept at zero. Faces in (a) and in (b) therefore indicate the effects of, respectively, jaw lowering/raising and front/back movements.

As mentioned before the speaker S3 is characterized by the high value of variance explaining the effects of the jaw front/back factor, f2 (23%), which is in fact greater than the variance of f1 (21%). Figure 2 shows that not only the lower lip deforms from front to back along the jaw movement, but also the upper lip appears to open up, resulting in a larger lip opening area in the back jaw position than in the front position. Presumably, the upper lip actively coordinates with front/back jaw movements, which is interpreted as a correlation structure in the factor analysis. This explains the high value of f2 variance observed in this speaker S3.

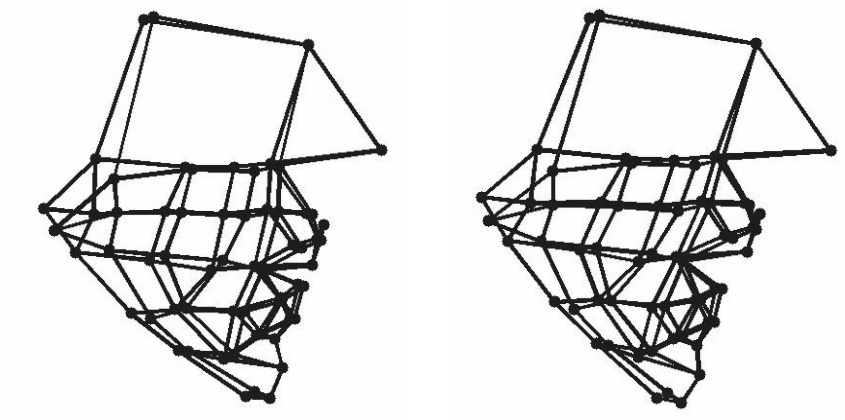


Figure 2: Model face of Speaker S3 where the value of the front/back jaw factor, f2, is varied from one extreme to the other as in Figure 1.

Badin et al. (2002) report a similar correlation between lip protrusion and jaw advance in their speaker. These authors consider the jaw advance as a part of lip related gesture and the lip-protrusion factor was extracted in the guided way, which resulted in a very high value of the explained variance. Here we assume a hierarchy in articulators. For example, the jaw gestures can be considered higher than the intrinsic lip gestures, since the influences of intrinsic lip gestures are localized in the lips themselves whereas that of the jaw extends over not only the lips but also the tongue and to an extent the larynx. In our analysis, therefore, the effects of the jaw positions were extracted before those of the lips following the hierarchical order.

The remaining 3 factors were determined by PCA and numbered in the order from high to low explained variance, as f4, f5, and f6. They must represent face deformations related not to the jaw gestures but to the intrinsic lip gestures, because they are determined on the residual of C after the subtractions of the correlation structures related to jaw motions in the 3 dimensions. The functions of these PCA derived factors must be visually identified by observing synthesized faces, as shown in Figure 3 for the speaker S1. It appears that the first PCA factor (f4) represents rounding/spreading in which the lip opening mainly deforms horizontally as seen in Figure 3a. The factor f4 of the French speakers, S2 and S4, also exhibits this kind of horizontal deformation of the lips. We interpret the factor f5 of S1 in Figure 3b as protrusion/retraction gestures involving both the lower and upper lips for this particular speaker. Although it may not be so visible in Figure 3b, the protrusion appears to be accompanied with a rotation of each lip to open up the aperture, which is visible in the video-clip. The factor f6 seems to represent a lowering/raising of the upper lip that is a direct consequence of the corresponding cheek movements, which is clearer in the f5 of S2, as seen in Figure 4.

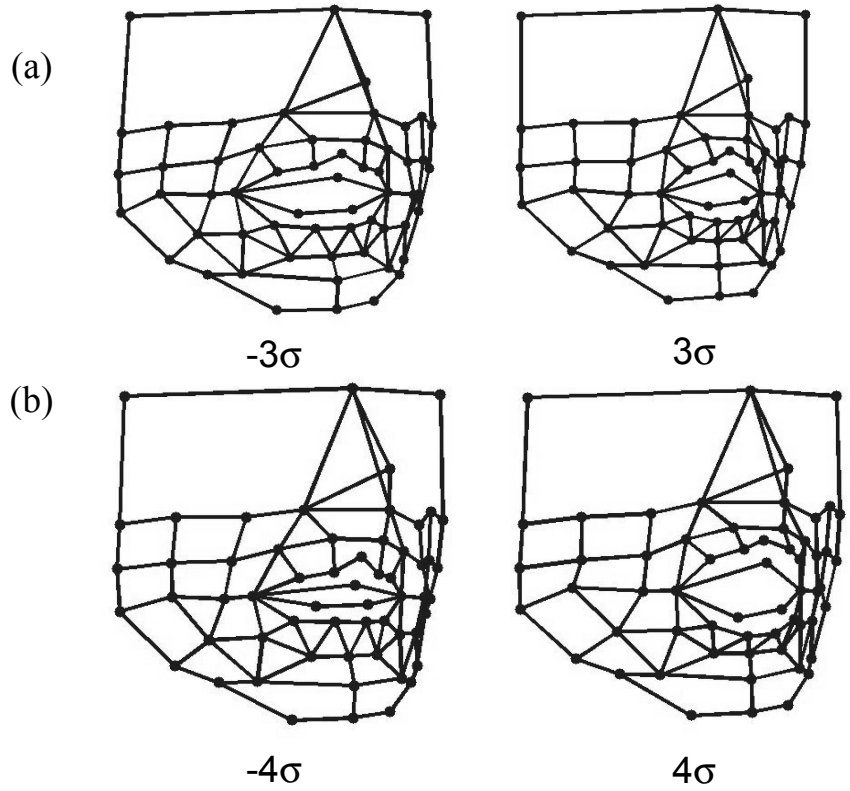


Figure 3: Model face of Speaker S1, where the value of each of 2 factors, f4 in (a) and f5 in (b), is varied from one extreme to the other as in Figure 1.

For the speaker S2, and also for S3, f5 seems to deform the face by lowering/rising of the cheeks as shown in Figure 4. The apparent larger lip opening is primarily due to a rising of the upper lip. In detail, the magnitude of cheek rise from the low position (with -4σ) to the high position (with 4σ) is greatest at the cheeks and then the upper lip, whereas the position of the lower lip is hardly affected. From this observation, we conclude that it's the cheeks which pull up the upper lip, and not the upper lip pushes up the cheeks. It may be obvious that the lips cannot push the cheeks up, if one considers the arrangements of the facial muscles (e.g., Gomi *et al*., 2002).

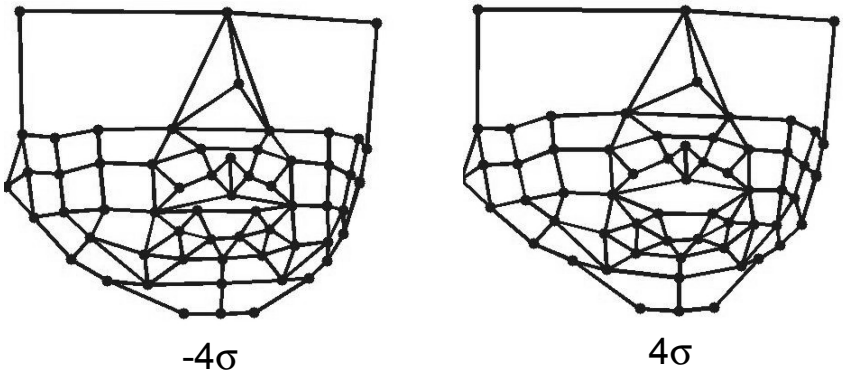


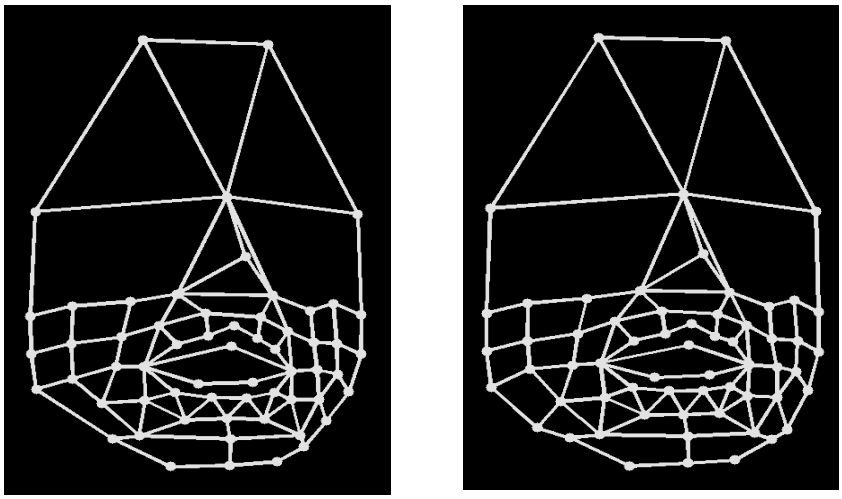
Figure 4: Model face of Speaker S2 where the value of the lower/rising cheek factors, f5, is varied from one extreme to the other as in Figure 1.

It may be noted here that functions of the 2 higher order factors, f5 and f6, are interchanged between S1 and French speakers S2 and S3, as already indicated in Table 1. For speaker S1, f5 represents the intrinsic lip gesture, retraction and protrusion and f6 having the smallest variance represents the cheek lowering and rise. For speakers S2 and S3, f5 accounts for the lowering/rising of the cheeks with the concomitant upper lip displacements and f6 that accounts for the lip protrusion has the smallest variance. It is safe to state therefore that S1 uses lip protrusion to control the length and aperture of the lip tube, while S1 and S2 employ the cheek lowering and rising to control the lip aperture. These observations suggest speakers use different articulatory maneuvers to produce speech sounds. In the next section, we shall describe in detail how differently speakers make the contrast between /s/ and /ʃ/ in terms of lip gestures.

Since each of those 6 factors can be considered as a functional elementary articulator, they tell us about how speakers articulate the lower jaw and the lips during speech production. As mentioned before, those first 6 factors explain about 90% of the variance for every speaker. Since the variance explained by any higher factor is less than 1.5 %, we discard factors higher than f6. In fact, it is not so evident to identify the functions of the higher factors because of their small individual influence on the face deformation. As a face model therefore, we use Eq. 2, but the full weight matrix A is replaced by its truncated version, A_6, for the first 6 factors, X_6, as follow:

$$Z=A_6X_6. \qquad (3)$$

Now, a synthetic factor model, Eq. 3, can be interpreted as an articulatory model as follows. The deformation Z from the mean face marker position is the sum of uncorrelated 6 linear components. Each weight, actually a set of coefficients of which number equals to that of variables (for example, 171 for S2 and S3), determines a particular face deformation pattern and the value of the factor, or the articulatory parameter, specifies the magnitude of that deformation. To obtain markers' positions in the original 3D coordinates, the deformation Z must be de-normalized using the inverse of Eq. 1. As described before, one of the interesting features of the face model is that the values of factors are calculated from the observed deformation Z and the truncated factor pattern A_6 in Eq. 3. Figure 5 shows the measured position of the markers during [i] and its reconstructed version with the first 6 factors. Note that markers' positions illustrated by dots are connected by lines so as to obtain a face like object. At present, we use this rather rudimentary face representation that was used already in Figures 1-5. The shapes of these 2 faces are hardly indistinguishable by eyes.



(a) (b)

Figure 5: The measured marker positions, indicated by the dots, of the speaker S1 during [i] in (a) and its reconstructed version using the first 6 factors shown in (b).

3. How speakers make [s] vs. [ʃ] contrast?

As described before, the values of the 6 factors can be directly calculated from the data and they should indicate an articulatory organization underlying the speech production. Let us show, as an example of a data analysis, how speakers make /s/-/ʃ/ contrast. It is known that /ʃ/ is produced with somewhat protruded open lips in English and French, whereas /s/ with unprotruded lips, although the lip shapes are phonologically unspecified (Gentil, 1980).

Figure 6 compares observed faces at about the center of [s] and of [ʃ] in /iCi/ syllables produced by those 3 speakers. All the lips appear spread due to the coarticulation of the vowel [i]; no matter the consonant is /s/ or /ʃ/. An obvious difference is that the lips are more open in /ʃ/ than in /s/ for all the speakers. Somewhat less obvious, but the lips appear to be protruded in /ʃ/, at least, in the speaker S1 and S2. These raw data clearly show the systematic geometrical differences in lip shapes, which are common to all the speakers. However, this doesn't necessarily mean that the same articulatory organizations underlie to produce the common geometrical differences, as describing in the next.

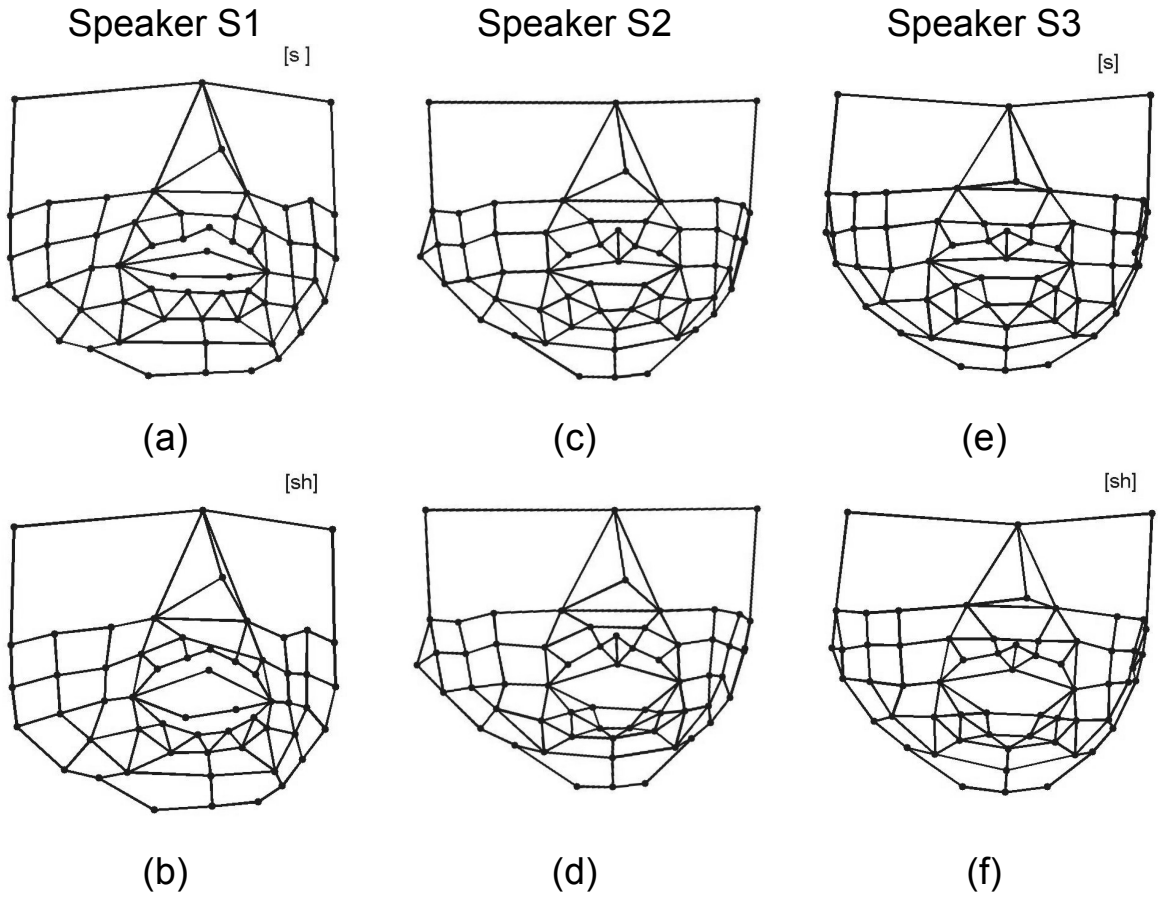
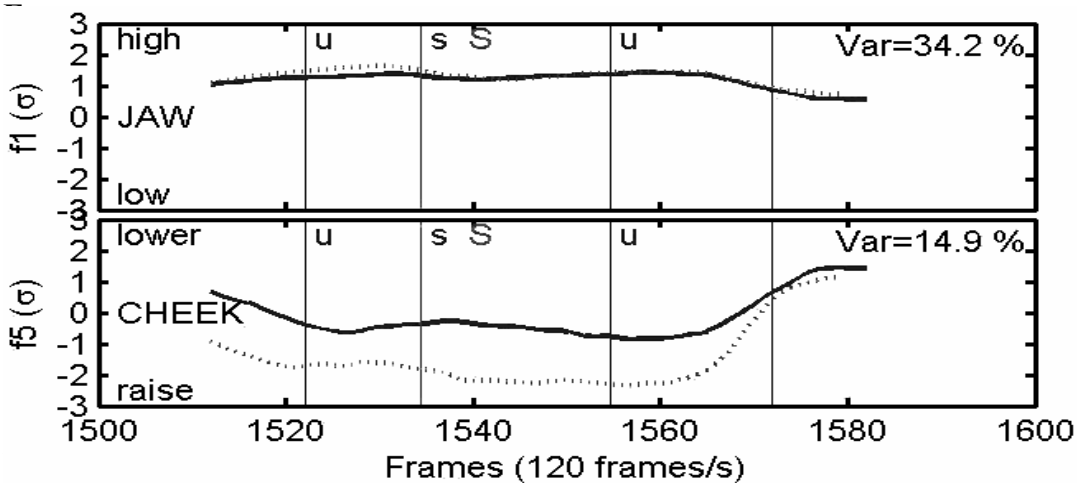


Figure 6: The measured faces at the center of [s] on the first row and of [ʃ] (denoted [sh] in the figures) on the second row. These plots are extracted from /iCi/ tokens uttered by the 3 speakers.

Figure 7 illustrates temporal variations of 2 selected factors, f1 and f6, along 2 syllables, [usu] in solid curves and [uʃu] in dotted curves, uttered by S2. Recall that f1 accounts for high/low jaw motions and f5 for the raising/lowering of the upper lip due to the cheek motions. In the case of f1, there is hardly any difference between the 2 curves, suggesting that f1 doesn't contribute to the [s] vs. [ʃ] distinction. The remaining factors also exhibit no important difference between [s] and [ʃ], except f5. In f5, shown in Figure 7, those 2 curves differ in an important way, suggesting the speaker S2 uses only f5 to differentiate [s] and [ʃ]. In other words, this factor (or this elementary articulator) is active for making the contrast. The speaker S2 raises the upper lip, which makes in turn the lip opening larger as seen in Figure 6d. The table 2 summarizes subjective judgments on which factors are active for speakers to make the contrast.



p

oral traces of 2 selected factors, f1 and f5, along syllables [usu] in solid curves and [uʃu] (denoted 'S' in the figures) in dotted curves, uttered by the speaker S2.

The "+" symbol in **Table 2** indicates the occurrence of marked difference between the values of a factor during [s] and the corresponding [ʃ] segment in a given vowel context. It is evident that 3 speakers use quite different articulatory organizations to produce the contrast. For example, S3 creates the contrast almost exclusively with the high/low jaw motion (f1). The use to the cheek raising/lowering, presumably to control the lip aperture, is unique to S2 among the speakers. Note that effects of vowel context are less pertinent than those of speaker difference. It may be noteworthy that the participation of cheek rise in S2 excludes with the vowel context /a/, suggesting an incompatibility of the upper lip rise in the open vowel context. Note also that no factor is active for making the contrast with the vowel context /u/ in S3. As reported by Gentil (1980), the anticipatory coarticulation of the labial attribute of the second vowel /u/ during /s/ could be the cause of the absence of the contrast. We shall give an additional discussion on this absence of difference in the lip gestures below.

Table 2: Active elementary articulators (factors) making /s/-/ʃ/ contrast

Factors		Speaker S1			Speaker S2				Speaker S3			
		aCa	iCi	uCu	aCa	iCi	uCu	yCy	aCa	iCi	uCu	yCy
JAW	high/low								+	+		+
	front/back			+	+	+						
LIPS	round/spread	+	+	+	+	+		+				
	protrude/retract	+	+	+								+
CHKS	lower/raise					+	+	+				

It becomes clear that speakers control the lip geometry to distinguish those 2 fricatives with quite different articulatory organizations. What acoustically accounts is the shape of the lip opening tube, which is roughly represented by its aperture (lip cross-sectional area) and its length. Although, our data using markers (i.e., flesh points) don't give us the exact geometry of the lip tube, it is safe to assume that the geometry is related to that of the flesh point representation, as seen in Figure 6, roughly in a proportional way. Then Figure 6 suggests that the lip aperture would be systematically greater in /ʃ/ than in /s/. The larger aperture of /ʃ/ than that of /s/ is created differently depending on speakers, by the combination of lip protrusion and spread in S1, by rising of the upper lip in S2, and merely by lowering of the jaw in S3. It isn't so systematic, however, in the case of the lip-tube length control. Speaker S1 and, a lesser extent, S2 seem to lengthen the lip tube with protrusion in /ʃ/ in comparison with /s/, which is neutral or slightly rounded. Note that protrusion/retraction in S2 is not marked in Table 2. This is because S2 protrudes the lips, to a certain degree, in both /s/ and /ʃ/. S3 does not lengthen at all in /ʃ/ relative to /s/.

Why is such an inter-speaker variation in lip geometry, especially lip tube length possible? In fact, Toda et al. (2002) have shown that differences in the observed lip configurations for /s/ and /ʃ/ alone cannot explain fairly important and distinctive differences in the spectral shape of these 2 classes of fricatives. The lower cutoff frequency in the noise spectrum of /s/ is much higher than that of /ʃ/ in the same vocalic contexts. Those authors have suggested that not only the differences in the lip geometry, but also those in tongue position and shape must contribute to the formation of the spectral characteristics of /s/ and of /ʃ/. In fact, Stevens (1993, 1999) has pointed out the acoustic influence of a relatively long sublingual cavity in /ʃ/ and the absence of such a cavity in /s/. Toda and Honda (2003) have confirmed Stevens' assertion by an MR imaging study. Then, the observed inter-speaker variations in the lip tube length can be explained as the consequence of an adjustment of the total length of the sublingual cavity plus the lip tube. As an extreme case, the lip gestures don’t differ much in the production of /s/ and /ʃ/ in the /uCu/ context (see Table 2) by the speaker S3. Presumably, this speaker makes the distinction only by tongue position. Extended studies on the labial and sublingual geometries in relation to their acoustic consequences are underway by those 2 authors and will be reported elsewhere.

4. Concluding remarks

We have shown in Section 2 that the guided PCA allows us to derive a compact and rational model of face deformations during speech. It is compact, because

the model consists of 6 uncorrelated (orthogonal) linear components. In other words, observed face deformations with apparent complexity have functionally, say, only 6 degrees of freedom. It is rational, because its 6 components can be interpreted by articulatory terms and thus, as shown in Section 3, the data analysis by factor values provides us with some insights about the underlying articulatory organizations.

Acknowledgements

The author is grateful to an anonymous reviewer and the editors Susanne Fuchs and Pascal Perrier for helpful feedback on an earlier draft. This work was supported, in part, by the project FEEDAT/ARC of the INRIA/LORRAINE, France.

References

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., and Savariaux, C., (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, 30, 533-553.

Gabioud, B., (1994). Articulatory models in speech synthesis. In *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art and Feature Challenges*, ed. E. Keller, John Wiley & Sons, 215-230.

Gentil, M. (1980). Sibilation et labialité en français – Coarticulation vocalique et valeur consonantique cible. *Labialité et Phonétique, Publications de l'Université des Langues et Lettres* de Grenoble, 181-201.

Gomi, H., Honda, M., Ito, T., and Murano, E., (2002). Compensatory articulation during bilabial fricative production by regulating muscle stiffness. *Journal of Phonetics*, 30, 261-279.

Maeda S., Toda M., Carlen A. J. and Meftahi L., (2002). Functional modeling of face movements during speech. *Proc. International Congress on Speech and Language Processing*, 17-20 September 2002, Denver (Colorado), 1529-1532.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech production and modeling. (NATO Advanced Study Institute series)*, eds. W. J. Hardcastle and A. Marchal, Kluwer Academic Publishers, 131-149.

Overall, J. E., (1962). Orthogonal factors and uncorrelated factor scores. *Psychological Reports*, 10, 651-662.

Stevens K. N., (1993). Modelling affricate consonants. *Speech Communications*, 13, 33-43.

Stevens K. N., (1999). *Acoustic Phonetics*. The MIT Press.

Toda M., Maeda S., Carlen A. J. and Meftahi L., (2002). Lip gestures in English sibilants: Articulatory-acoustic relationship. *Proc. International Congress of Speech and Language Processing*, 17-20 September 2002, Denver (Colorado), 2165-2186.

Toda, M. and Honda, K., (2003). An MRI-based cross-linguistic study of sibilant fricatives. *Proceedings of the 6th International Seminar on Speech Production*, Sidney, December 7 to 10, 2003. 290-295.

Control and representations in speech production

Pascal Perrier
Institut de la Communication Parlée, UMR CNRS 5009, Institut National Polytechnique de Grenoble & Université Stendhal, Grenoble, France

In this paper the issue of the nature of the representations of the speech production task in the speaker's brain is addressed in a production-perception interaction framework. Since speech is produced to be perceived, it is hypothesized that its production is associated for the speaker with the generation of specific physical characteristics that are for the listeners the objects of speech perception. Hence, in the first part of the paper, four reference theories of speech perception are presented, in order to guide and to constrain the search for possible correlates of the speech production task in the physical space: the *Acoustic Invariance Theory*, the *Adaptive Variability Theory*, the *Motor Theory* and the *Direct-Realist Theory*. Possible interpretations of these theories in terms of representations of the speech production task are proposed and analyzed. In a second part, a few selected experimental studies are presented, which shed some light on this issue. In the conclusion, on the basis of the joint analysis of theoretical and experimental aspects presented in the paper, it is proposed that representations of the speech production task are multimodal, and that a hierarchy exists among the different modalities, the acoustic modality having the highest level of priority. It is also suggested that these representations are not associated with invariant characteristics, but with regions of the acoustic, orosensory and motor control spaces.

1. Introduction

The concept of representation is a key concept in language and speech research. However, it can have different meanings, according to the specific field of interest that is considered. Representations in language can be *lexical*, when they refer to the verbal characterization of the world, or *phonological*, if they aim at describing speech sequences in an invariant and segmental way. Representations in speech communication can also be *motor*, if motor control processes underlying speech production are under investigation, *spectro-temporal*, if the

characteristics of the acoustic speech signal are analyzed, or *auditory*, if the focus of the research is the perceptual processing of speech. In this paper, we will not address all the different kinds of meaning of this concept. The focus of the paper concerns the representations of the speech production task which could be elaborated in the human brain, with the purpose of setting up the motor commands and generating the movements of the vocal source and of the vocal tract articulators that will permit the production of intelligible speech sequences.

In cognitive sciences (see Jeannerod, 1994, for a challenging tutorial), the term *representation* relates to the mental imagery process underlying brain functions of human beings in their interactions with the external world surrounding them. When the brain function of interest is the production of speech, the external world includes the peripheral speech apparatus (vocal folds, jaw, tongue, velum, lips and the vocal tract as a whole) as well as the acoustic speech signal. Hence, from this perspective, the terms *representations in speech production* refer both to the mental imagery of the measurable physical characteristics associated with the articulation of intelligible speech sounds (i.e. muscle activities, articulators' positions, geometrical shape of the vocal tract, together with the spectro-temporal characteristics of the acoustic signal), and to the mental imagery of the peripheral apparatus itself (i.e. of the muscle anatomy, of the relations between muscle activations, articulatory positions and characteristics of the speech signal, of the dynamical and biomechanical articulators' properties …). In other words, studying representations in speech production means trying to find answers to the two following questions:

1. How does the human brain characterize what a speaker wants (probably unconsciously) to generate in the physical world when he/she produces speech? Articulatory positions? Spectral properties? Temporal properties?
2. How does the human brain characterize the relations between motor commands and the expected objectives of speech production in the physical world?

The focus of this paper will be limited to the question of the nature of the *representations of the speech production task* in the speaker's brain (i.e. question 1). In the first part of the paper, fundamental theoretical aspects will be presented via a summary of the four main speech perception theories. In the second part, a few selected experimental studies published in the literature will be described in order to show how these theories can be questioned and used to understand the representations of the speech production task. In these two parts of the paper, the different theoretical hypotheses and the interpretations of the experimental works will be presented as objectively as possible. In the conclusion, I will set out my own interpretation of this whole theoretical and experimental material in terms of representations.

As concerns the second question, the one of the nature of the *representation of the peripheral speech apparatus* in the speaker's brain, readers could refer to Guenther (1995) and Guenther *et al.* (1998) for theoretical studies about the use of these representations in speech motor control, to Perkell *et al.* (1997, 2000) for experimental evidences supporting the hypothesis of the existence of such representations in the brain, and to Laboissière *et al.* (1996), Perrier *et al.* (2003) and Perrier (2005) for theoretical and modeling studies of the complexity of these representations. More generally, theoretical foundations of the concept of *internal representation of the motor system* in the brain can be found in Kawato *et al.* (1987), and Jordan & Rumelhart (1992), while the controversy between Gomi & Kawato (1996) and Gribble *et al.* (1998) illustrates well the debates about the nature and the complexity of these representations.

2. What do models of speech perception tell us about the representation of the speech production task?

The complexity of the issue of the speech production task's representation arises essentially from the combination of three factors: the truly perceptual nature of the ultimate objective of speech production; the multimodality of the physical correlates of this perceptual objective; the many-to-one relations between the articulatory and the acoustic domains of speech production.

The ultimate objective of speech production is not defined in the directly accessible and measurable physical world, but in the brain of the listeners. Indeed, speech signals have no meaning by themselves. They only make sense in relation with the perception that a listener can have of them. In other words, speech production does not ultimately aim at producing specific phenomena in the physical environment of the speakers, such as movements or spectral properties of the acoustic signal, but at transmitting a code that can be interpreted by the listeners. To do so, speech production does use physical carriers, which are both articulatory movements and acoustic signals. In natural speech, articulatory movements and the acoustic signal are obviously strongly coupled, since articulatory movements are the source of the acoustics and determine its spectro-temporal characteristics. However, there is some experimental evidence suggesting that, when both modalities are available, they are both taken into account and processed in speech perception, even if they don't match with each other as it is the case in audiovisual illusions (McGurk & MacDonald, 1976). In summary, there is no way to measure the characteristics of the ultimate objective of speech production, because it is perceptual and, then, related to the listeners, but multimodal correlates of this objective (i.e. articulatory, gestural, acoustic or/and aerodynamical) can be found in the physical world.

This multimodal nature constitutes the second factor of complexity in attempts to characterize the speech production task's representations, because the characteristics in the different modalities are not simply different faces of the same object. Indeed, the many-to-one nature of the relations between the different physical domains of speech production has been shown many times (see for example Atal *et al.*, 1971). Thus, numerous muscle activations can underlie the same articulatory configuration, various articulatory positions can produce similar vocal tract shapes, and a number of vocal tract shapes can be associated with very similar characteristics of the acoustic speech signal. This is the third major source of complexity.

The search for physical correlates of the perception of speech categories (for example of phonemes) has been one of the crucial issues of speech production and speech perception research for the last 3 decades. Four major theories have been proposed in the literature; they have at the same time served as rationales for a very large amount of experimental and modeling work, and been at the core of numerous controversial debates (Perkell & Klatt, 1986; McGowan & Faber; 1996): the *Acoustic Invariance Theory* (Stevens & Blumstein, 1978; Blumstein & Stevens, 1979), the *Motor Theory* (Liberman et al., 1967; Liberman & Mattingly, 1982), the *Direct Realist Theory* (Fowler, 1986; 1991), and the *Adaptive Variability Theory* (Lindblom, 1988; 1990). In the two following subsections, the main hypotheses of these reference theories will be summarized.

2.1. Acoustic Invariance Theory and Adaptive Variability Theory: the object of speech perception is acoustic

Stevens (Stevens & Blumstein, 1978; Blumstein & Stevens, 1979) proposed that speech perception would be based on invariant properties of the acoustic signal:

> *"* [...] *there is an acoustic invariance in the speech signal corresponding to the phonetic features of natural language. That is, it is hypothesized that the speech signal is highly structured in that it contains invariant acoustic patterns for phonetic features, and these patterns remain invariant across speakers, phonetic contexts, languages.* [...] *the perceptual system is sensitive to these invariant properties. That is, it is hypothesized that the perceptual system can use these invariant patterns* [...] *to process the sounds of speech in ongoing perception"* (Blumstein, 1986, p. 178).

Typically, these acoustic invariants could be formant patterns for vowels or the spectral shape of the burst for plosives. Stevens does not deny a possible role of

articulatory information in speech perception, but not with a primary status, and only in addition to the basic process based on the processing of acoustic events:

> *"In fact the occurrence of acoustic events arising from implementation of* [phonological] *features could provide landmarks that guide the search for other features that are more directly the result of manipulation of particular articulators"* (Stevens, 1996, p. 1693).

For example, knowledge about the articulatory-acoustic relations could be helpful when the acoustic information is ambiguous or not complete:

> *"The listener must also know how to access items in the lexicon based on partial information and must also know which kinds of modifications are permitted in the sound* [...] *and which are not. Speech production clearly can play an important role in acquiring these sources of knowledge* [...] *"* (Stevens, 1996, p. 1693).

In coherence with the acoustic invariance theory, and also in strong support of it, the *quantal theory of speech* (Stevens, 1972; 1989) proposes that structure of phonological systems in the world languages would have been determined by the non-linearities of the articulatory-acoustic relation, in order to associate phonemes with the most stable acoustic patterns. Thus, the articulatory configurations would have been selected in order to minimize the acoustic variability associated with the articulatory inaccuracy existing in ongoing production of speech, and to ensure the best achievement of the acoustic features characterizing the phoneme.

The Adaptive Variability theory of Lindblom (Lindblom, 1988; 1990) defends also the primacy of acoustics over articulation for the characterization of the physical correlates of speech perception.

> *There is at present no evidence suggesting that gestures have* [any] *particular advantage over acoustic patterns.* [...] *articulatory recovery* [...] *does not seem like a compelling alternative to exploiting acoustic/auditory systematicities in efficiently precompiled form"* (Lindblom, 1996, p. 1690).

However, as indicated by its name, the Adaptive Variability theory rejects the hypothesis of the existence of any physical invariant whether in the acoustic space or in the articulatory one.

> *"Looking for invariance cannot be seen as a phonetic problem. It is not a signal analysis problem at all. The invariance of linguistic categories is ultimately to be defined only at the level of listener comprehension."* (Lindblom, 1988, p. 160).

Thus, the physical realizations of phonetic units would be fundamentally variable, depending on the phonetic context, on the speaking style, on the speaking rate, and, more generally, on the speaking condition. However, this variability would also be controlled and adapted by the speaker who evaluates what it is necessary to generate, in order ensure a good perception of the message.

> *"Intraspeaker phonemic variation is genuine and arises as a consequence of the speaker's adaptation to his judgment of the need of the situation. In the sense of the biologist's term speech is an adaptive process"* (Lindblom, 1988, p. 163).

In spite of this physical variability, a correct perception of invariant phonetic categories should be possible:

> *Any sets of intelligible pronunciations are, by definition, articulatorily, acoustically, and auditorily equivalent with respect to the goal of perceiving the given lexical item correctly, but that does not logically entail assuming articulatory, acoustic or auditory invariance in the phonetic behavior. The common point of these examples is that an invariant (non signal) end is reached by variable (signal) means* (Lindblom, 1996, p. 1685).

To achieve a correct perception, the speech perception system would not only take into account the information carried by physical speech signals, but also information about the conditions under which speech is produced. Thus, the Adaptive Variability Theory assumes

> *"that, in all instances, speech perception is the product of both signal-driven and signal independent information, that the contribution made by the signal-independent processes show short-term fluctuations, and that speakers adapt to those fluctuations. It says that* [...] *adaptive behavior is the reason for the alleged lack of invariance in the speech signal"* (Lindblom, 1990, p. 431).

Thus, according to this theory, the physical correlates of speech perception would be variable acoustic properties that would have a "*sufficient discriminative power*" (Lindblom, 1990, p. 431) to allow the identification of the different phonetic classes when contextual information is taken into account. For example, for vowels, the physical correlates could be the (F1, F2) formant patterns that would not be interpreted independently, but relatively to each other by taking into account the limits of the maximal acoustic (F1, F2) vowel space that can be produced by the speaker for the considered language and under the considered speaking conditions. The question of the mechanisms permitting the integration of contextual information in order to predict the size of the maximal vowel space, is still an unsolved question.

2.2. Motor Theory and Direct-Realist Theory: the object of speech perception is articulatory

The Motor Theory and the Direct-Realist Theory defend the idea that the objects of speech perception would be in the articulatory domain. Thus, along the lines of Stetson, they both suggest that "*Speech is rather a set of movements made audible than a set of sounds produced by movements.*" (Stetson, 1928, p.29). However, these two theories strongly differ about two important points. First, the Motor Theory does not assume the existence of a measurable articulatory invariant (i.e. of a physical invariant), while the Direct-Realist Perception Theory does. Second, and it is a consequence of the first point, the Motor Theory assumes the existence of a speech specific perceptual processing (a hypothesis classically summarized with the sentence "*Speech is special*"), while the Direct-Realist Theory is based on general human perception principles proposed by Gibson (1966). These points will be further developed below.

The Motor Theory rejects the idea that the object of speech perception would be in the acoustic domain, because

> *"*[...] *there is typically a lack of correspondence between acoustic cue and perceived phoneme, and in all cases it appears that perception mirrors articulation more closely than sound"* (Liberman *et al*., 1967, p. 453).

According to this theory, the acoustic signal would rather be for the listener

> *"a basis for finding his way back to the articulatory gestures that produced it, and thence, as it were, to the speaker's intent"* (Liberman *et al*., 1967, p. 453).

However, it should be noted that this theory does not pretend that the perception of a phoneme is associated with the existence of a measurable invariant, such as a specific set of articulatory positions or a specific vocal tract shape. Its authors rather suggest that the invariant would be at the level of the motor commands, and not in the physical external world:

> *"The invariant is found far down in the neuromotor system, at the level of the commands to the muscles"* (Liberman *et al*., 1967, p. 454).

This hypothesis was refined almost 20 years later in the "*revised*" version of the Motor Theory (Liberman & Mattingly, 1985) by introducing a clear link between the speaker's intent and physical phonetic characteristics:

> *"The objects of speech perception are the intended phonetic gestures of the speakers, represented in the brain as invariant motor commands that call for movements of the articulators through certain linguistically significant configurations. These gestural commands are the physical reality underlying tradi-*

> *tional phonetic motions – for example " tongue raising", "tongue backing", "lip rounding" and "jaw raising" – that provide the basis for phonetic categories"* (Liberman & Mattingly, 1985, p. 64).

According to these authors, in speech production the existence of invariant motor commands associated with an intended phonetic gesture does not imply that any invariance exists at the articulatory or at the acoustic level. Indeed, the successive gestures necessary for the production of a sequence of phonemes are not produced purely sequentially. They partly overlap each other in time, in such a way that an invariant intended gesture will generate various movements and various acoustic signals because of two main factors: first, the nature of the preceding and following phonemes, and, second, the speaking rate determining the time overlap between successive gestures. This is the consequence of coarticulation. A strong feature of the Motor Theory consists in the fact that, thanks to the concept of overlap between invariant intended gestures, the observed physical variability of speech signals is fully compatible with the concept of phoneme related invariance. However, at the same time, the absence of congruence between the motor commands underlying an intended phonetic gesture and the associated measurable articulatory or acoustic properties, raises the question of how an intended gesture can be recovered by a listener. The solution proposed by the Motor Theory is that the perception of speech is special, and based on the use of a specialized "*phonetic module*" in the brain. This module would describe the very complex acoustic consequences of gestural overlaps in speech production, in order to infer the sequence of intended gestures from the acoustics.

> *"Incorporating a biologically based link between perception and production, this specialization prevents listeners from hearing the signal as an ordinary sound, but enables them to use the systematic, yet special, relation between signal and gesture to perceive the gesture. The relation is systematic because it results from lawful dependencies among gestures, articulator movements, vocal-tract shapes, and signal. It is special because it occurs only in speech"* (Liberman & Mattingly, 1985, p. 67).

Thanks to the "*phonetic module*",

> *"Speech perception is immediate, no cognitive translation from patterns of pitch, loudness, and timbre is required"* (Liberman & Mattingly, 1989, p. 489).

As said above, this last point is in strong disagreement with the Direct-Realist Theory of speech perception elaborated by Fowler (1982, 1986) who suggests

that the basic mechanisms of speech perception are just the same as the ones underlying visual or tactile perception:

> *"An informational medium, including reflected light, acoustic signals and the perceiver's own skin, acquires structure from an environmental event specific to certain properties of the event; because it acquires structure in this way the medium can provide information about the event properties to a sensitive perceiver. A second crucial characteristic of an informational medium is that it can convey its information to perceivers by stimulating their sense organs and imparting some of its structure to them"* (Fowler, 1986, p. 5).

In speech perception the informational medium is the acoustic signal, and the event, the source of information, is the articulating vocal tract. Fowler considers the vocal tract itself and not the motor commands that are at the origin of its shaping:

> *"* […] *studies of the activities of individual muscles or even individual articulators will not in themselves reveal the systems that constitute articulated phonetic segments"* (Fowler, 1986, p. 5).

Hence, from this perspective also the Motor Theory and the Direct-Realist Theory strongly differ. The Motor Theory suggests that listeners perceive the speaker's intent, even if this intent is hidden by the gestures associated with the surrounded phonemes, while, according to the Direct-Realist Theory, the object of perception is a set of actual characteristics of the vocal tract. What are these characteristics? Fowler does not give an answer, but she assumes that they are produced anyway, whatever the context and in spite of the observed variability of the articulatory patterns associated with the production of a phonetic segment:

> *"* [...] *from an event perspective, the primary reality of the phonetic segment is its public realization as vocal-tract activity"* (Fowler, 1986, p. 10).

This "*public*" vocal-tract activity could be perceived directly in the acoustic signal without any complex cognitive or non-cognitive processing, and it would be the direct image of the mental intention of the speakers:

> *"* [...] *the idea that speech production involves a translation from a mental domain into a physical, non-mental domain such as the vocal tract must be discarded.* [...] *we may think of the talker's intended message as it is planned, uttered, specified acoustically, and perceived as being replicated intact across different physical media from the body of the talker to that of the listener"* (Fowler, 1986, p. 10-11).

A consequence of this "*direct*" conception of the speech production-speech perception system lies in the fact that the invariant at the phonological level not only appears as a vocal tract activity, but also in the acoustic signal itself as "*specifiers or invariants*" (Fowler, 1996, p. 1731). Hence, from this perspective, the Direct-Realist Theory does agree with the Invariance Acoustic Theory, and this statement logically raises the following question: Why does Fowler defend the hypothesis of the perception of invariant vocal tract properties via the acoustic signal, rather than Stevens' hypothesis of the perception of invariants in the acoustic signal? There are two main reasons that justify this theoretical approach. First, speech perception should not obey different rules than other animal perception systems

> *"Perceptual systems have a universal function. They constitute the sole mean by which animals can know their niches.* [...] *even though it is the structure of the media (light for vision, skin for touch, air for hearing) that sense organs transduce, it is not the structure of those media that animals perceive. Rather, essentially for their survival, they perceive the components of their niche that caused the structure"* (Fowler, 1996, p. 1732).

Second, theoretical models of the different levels of human communication with language have to be as congruent and compatible with each other as possible, in order to offer a coherent theoretical framework, in which general models integrating interactions between these different levels can be developed (Fowler, 1996).

2.3. Conclusions for the representation of the speech production task in the speaker's brain

It is common sense to say that speech is produced to be perceived and that the relevance of physical characteristics of speech should only be assessed from this perspective. However, from a speech motor control perspective this common sense tells us also that the task of the speaker should be to generate in the physical world information that listeners will be able (1) to perceive and (2) to interpret in terms of phonetic categories and/or in lexical and semantic terms. Consequently, depending on the speech perception model, representations of different natures can be proposed for the speech production task in the speaker's brain.

The Acoustic Invariance Theory suggests that representations should be associated with absolute invariant temporal and/or spectral characteristics of the acoustic signal. Thus, the production of French rounded vowel /u/ could be represented as a low frequency (300Hz, 800Hz) point in the (F1, F2) space,

while producing the stop consonant /k/ could mean generating a short burst with a maximum of energy around 2.5 kHz.

The Adaptive Variability Theory suggests something more complex associating, on the one hand, some kind of acoustical characteristics, such as formant patterns or burst spectrum, that could vary within certain limits as the result of a permanent negotiation between speaker and listener, and, on the other hand, some kind of extra-linguistic information about the speaker, the speaking style or the speaking rate. From this perspective, thinking about the representations of the speech production task in the speaker's brain implies thinking about the terms of the speaker-listener negotiation and about the implementation of this negotiation in the brain.

If the Motor Theory is right, the speaker should have in mind the production of a sequence of overlapping phonetic gestures. Thus, producing a rounded vowel followed by a nasal labial stop would imply, for example, to generate with a certain time overlap, a combination of a lip movement toward protruded lips and a lowering of the larynx, for the rounded vowel, and a movement toward closed lips associated with a lowering of the velum, for the stop. There is no requirement for the speaker to actually achieve these articulatory goals. It is just necessary for them to send the appropriate commands to the motor system, the final movements depending on the gestural overlap in time.

If we follow the Direct-Realist Theory, the speaker should have the objectives to achieve a number of specific characteristics of the vocal tract. For example, producing the vowel /u/ should mean achieving a constriction in the velar region of the vocal tract together with rounded lips.

Why is it important, in terms of speech motor control, to be able to make a choice among these different hypotheses? Indeed, after all, when a French /u/ is produced, we do actually observe at the same time, rounded lips, a constriction in the velar region and a low frequency (F1, F2) pattern. Hence, why do we care whether the speaker's objective was to produce the articulatory or the acoustic characteristics? It is because of the non-linear and non bi-univocal characteristics of the relations between the articulatory and the acoustic domains of speech production.

Indeed, the non-linearity generates a warping of the relations between configurations within a space, when one moves from a space to the other. Thus, configurations that are very close in one domain, could be far from each other in another domain. This is well illustrated by the French vowels /i/ and /y/, which are quite close in the space of the first three formants, and are very well separated in the articulatory domain along the lip rounding dimension. As a consequence, if we suppose that control strategies underlying the production of speech sequences could involve a minimization of distance in a given space, the resulting optimal strategy could be different according to which space is

considered. Similarly, requirements in terms of accuracy and stability of the control could be very different according to which physical space is taken into account to measure accuracy and stability.
The non bi-univocal characteristic of the articulatory-acoustic relations could also largely influence speech motor control strategies. Indeed, a specific configuration in one domain can be associated with a number of configurations in the other domain. Thus, a given formant pattern can be produced by several combinations of jaw and tongue positioning. Similarly, a given jaw aperture can be generated by different recruitments of the jaw muscles. In summary, the number of degrees of freedom in the achievement of an objective is not the same depending on the space where the objective is defined. This implies, in particular, that different compensation strategies could be involved, and this would generate different coarticulation mechanisms…
It should be also noted that the issue of the nature of the representation of the task is an important issue not only for speech production but for motor control in general, and it has been shown to be crucial to understand the motor control strategies underlying human movements. For example, as concerns target pointing tasks with a finger, many research works have studied whether movements are controlled in the geometrical space of the finger position, in the space of the joint angles (wrist, elbow, shoulder) or in the space of the torques at the joints. The challenges of this research were well illustrated by the studies carried out by Soechting & Lacquaniti (1981), Wolpert *et al.* (1995) or Sabes & Jordan (1997).

3. Some insights into the nature of speech task representations from recent experimental studies

The four speech perception theories described in Section 2, and the corresponding hypotheses about possible speech production task's representations in the speaker's brain, are very controversial and still at the center of numerous debates. It is not the purpose of this paper to present an exhaustive review of the arguments in favor of or against each of them. Numerous, very interesting discussions were published in the literature about this topic, in particular in the book *Invariance and Variability in Speech Processes* (Perkell & Klatt, 1986), in two special issues of the *Journal of Phonetics,* the one centered on the Direct-Realist Theory for speech perception and on the Task Dynamics for speech production (*Journal of Phonetics, 14*, Vol. 1, 1986) and the one devoted to Stevens' Quantal Theory of Speech (*Journal of Phonetics, 17*, Vol. 1/2, 1989), and in the group of papers published in 1996 in the *Journal of the Acoustical Society of America* after a special session entitled *Speech Recognition and Perception from an articulatory point of view* held during spring 1994 in the

ASA meeting (*Journal of the Acoustical Society of America, 99(3),* 1680-1741). Two books respectively published by Alvin Liberman (Liberman, 1996) and by Ken Stevens (Stevens, 1998) offer numerous details about the theories defended by their authors. Two critical tutorial papers in the field of speech perception, Schwartz *et al.* (2002) and Hawkins (2004), should also be recommended. Finally, in order to close this list of publications related to the notion of representations, it is necessary to mention Sock's contribution, which totally rejects the concept of representation in speech (Sock, 2001).

To illustrate the content of the debates and the kind of studies that aim at clarifying the nature of the representations of the speech production task in the speaker's brain, a few selected experimental works will be presented in this section. First, three experimental studies supporting the hypothesis of acoustic representations will be presented. Then, an experiment suggesting the existence of strong articulatory specifications for the speech production task will be described. Finally, in a more speculative approach, we will see how the recent discovery of *mirror neurons* in monkeys' brain could offer new perspectives.

3.1. The Lip Tube experiment

In the *Lip Tube experiment*, Savariaux *et al.* (1995, 1999) perturbed the production of the French [u], pronounced in isolation, by introducing a 2 cm diameter tube between the lips of 11 subjects speakers of French. This induced a strong increase of the lip area, and limited the range of variation of jaw position. After the insertion of the tube, the subjects could train 19 times, in order to find out how to compensate for the perturbation, if they felt that compensation should be made. Then, by the twentieth repetition, they were asked to pronounce /u/ with the lip tube one more time but with the strategy that they considered to be the best among the 19 preceding training trials. Compensatory strategies were observed in the articulatory domain with sagittal cineradiography.

The production of /u/ in French is normally achieved with very rounded and protruded lips associated with a back and high tongue position generating a vocal tract constriction in the palato-velar region. However, the classical acoustic theory of vowel production (Fant, 1960) predicts that the same (F1, F2) pattern could be also produced with open lips and with a back tongue position generating a vocal tract constriction in the velo-pharyngeal region. By inserting the tube between the lips, Savariaux *et al.* wanted to test how the subjects reacted to the perturbation, with the following hypothesis in mind: if the subjects move their tongue back in order to generate a velo-pharyngeal articulation, it would support the hypothesis of an *acoustic* representation of the speech task. On the contrary, if the subjects do not change anything to their tongue position, or if they produce changes that are not compatible with an enhancement of the

(F1, F2) pattern, it would rather support an *articulatory* representation of the speech production task.
A systematic analysis of the acoustic signal in terms of pitch and (F1, F2, F3) formant patterns was performed, and perceptual tests were run to evaluate the perceptual quality of the /u/ produced under perturbed conditions. Results were as follows. First, none of the subjects could compensate for the perturbation in the first trial, and, in case of compensation, a number of trials were always necessary to achieve it. Second, only one subject actually produced a velo-pharyngeal constriction and could compensate for the perturbation in the (F1, F2) plane. Three other subjects, still keeping their tongue in the palato-velar region, moved it back sufficiently to generate an improvement of the (F1, F2) formant pattern, which, in combination with the pitch frequency, permitted the production of a well perceived /u/: the backward tongue movement permitted to limit the F2 variation, while an increase of the pitch maintained (F1-F0) sufficiently low. Third, all the remaining subjects tested a number of new articulatory strategies during the 19 trials of the training phase. Some of them provided a small improvement of the perceptual quality of their /u/ and actually moved their tongue slightly backward. Some of them did not, but none produced a forward movement of the tongue which would have led to a worse (F1, F2) pattern and to a decrease of the perceptual quality of their /u/.
These observations support largely the hypothesis that, in all cases, the compensatory maneuvers were elaborated in order to generate (F1, F2) formant patterns as close as possible to the normal patterns. Hence, they speak in favor of an acoustic nature of the representation of the speech production task in the speaker's brain. The perceptual tests also permitted to suggest that acoustic representations of vowels could be associated with regions of a space combining the formant patterns and the pitch frequency. Last, it was observed that, at the end of the training phase, 3 subjects finally had selected the original palato-velar articulation as the best one of their 19 trials, after they had stated that they could not enhance the bad (F1, F2) pattern and the bad perceptual quality of their /u/. This suggests that these canonical articulations could also be part of the representation of the speech production task.

3.2. The Dental Prosthesis experiment

Jones & Munhall (2003) investigated for six native speakers of Canadian English the contribution of the auditory feedback to the process of adapting for a geometrical perturbation of their vocal tract during the production of the fricative /s/ in the context of the word /tɑs/. The perturbation consisted in a dental prosthesis that lengthened the upper incisor teeth between 5 and 6 millimeters, without affecting the subjects' bite. As compared to the lip tube, this

perturbation has the noticeable advantage of not modifying at all the normal articulation and the normal proprioceptive and tactile information within the vocal tract. In other words, with this perturbation the natural tongue and jaw positions underlying the production of /s/ are still possible and the corresponding proprioceptive feedback is not altered.

On the contrary, in the acoustic domain this perturbation has a noticeable impact. In the production of /s/, the noise source arises from a jet of air, generated by the vocal tract constriction, and hitting the surface of the front teeth (Shadle, 1989). This source of noise excites essentially the small front cavity of the vocal tract located between the constriction and the lips. This is at the origin of a maximum of energy in the high frequency domain of the speech spectrum. According to Jones & Munhall (2003) the lengthening of the upper incisor teeth essentially induces an enlargement of the front cavity, and, thus, a lowering of the frequency of the maximum of energy in the spectrum. As a consequence, in the absence of compensation, /s/ pronounced with the dental prosthesis is expected to sound more like /ʃ/.

Each experimental session consisted of two sub-sessions, and each of these sub-sessions consisted of 15 blocks of 10 repetitions of /tɑs/ under 4 different conditions: (C1) normal condition; (C2) without the dental prosthesis in the mouth and with a masking of the auditory feedback with a white noise; (C3) with the prosthesis in the mouth and with masked auditory feedback; (C4) with the prosthesis in the mouth and with normal auditory feedback. The ordering of the 15 blocks was as follows: C1, C2, C3, C4, C3, 4 alternations (C4-C3), C2, C1. The acoustic production of /s/ was evaluated in the spectral domain by measuring the ratio between the slope of the spectral envelope below 2.5 kHz and the slope between 2.5 kHz and 8 kHz, and its perceptual quality was rated by perceptual tests carried out by 16 listeners.

Results are as follows. In the first block in condition C3, the acoustic production of /s/ was altered by the dental prosthesis, and the spectral impact conformed to the theoretical predictions: /s/ sounded more like /ʃ/. No improvement was observed during the 10 repetitions in this block. Compensation started only during the first block in condition C4, when auditory feedback became available. Afterwards, in the sequence of (C4-C3) alternations an improvement was generally noted within each block with or without auditory feedback, but the improvement was larger in the presence of auditory feedback. In addition, a learning effect was observed, since the improvement increased continuously across the 5 repetitions of the alternations (C4-C3), and since the improvement obtained at the end of sub-session 1 was maintained during sub-session 2.

The observation of the major role of auditory feedback in the compensation process supports the hypothesis of an acoustic representation of the speech production task in the speaker's brain. The results also suggest that, for

fricatives, the spectral steepness difference could be a good physical correlate of the acoustic representation. In addition, the learning effect observed within each session and across them, together with the fact that, once compensation was initiated with auditory feedback, improvement was also possible in the absence of it, suggest that speakers were immediately able to transpose requirements in the acoustic domain into articulatory terms. This suggests that the primary acoustic representation of the speech production task could be immediately associated with a secondary representation in the articulatory domain, which is more proximal for the speaker.

3.3. Velocity dependent perturbations of jaw movements

With a robotic device connected to the mandibular teeth and controlled by computer, Tremblay *et al.* (2002) delivered velocity dependent mechanical perturbations to the jaw of subjects in speech and non speech conditions. Perturbing forces were applied in the sagittal plane along an axis parallel to the occlusal plane, in the direction of jaw protrusion. Analyses of kinematic data of these perturbations induced a change of the motion path of the jaw, and thus a change of the somatosensory feedback. The larger the velocity the stronger the perturbation force applied to the jaw, and, then, the change provided to the motion path. Three different conditions were tested: production of the utterance [siat] slowly and clearly with vocalization; articulation of the same utterance without vocalization (silent speech) and still slowly and clearly; non-speech jaw movement that matched the amplitude and duration of the two speech conditions. For each condition the session started with 20 repetitions of the task without perturbation; it continued with 20 repetitions of the task with perturbation, and finally the perturbation was removed and the task was again repeated 20 times.

Results were as follows. In the first trials following the introduction of the perturbing force field, a noticeable modification of the motion path of the jaw was observed for all subjects and for the three conditions. For the two speech conditions (vocalized and silent), after training, an adaptation to the perturbation was observed: after a few trials, the motion path of the jaw became similar to the one produced without the perturbation. In addition, still for the two speech conditions, an after-effect was noticed, since, once the perturbation was removed, a few trials were again necessary for the subjects to go back to their normal jaw movements. In the non-speech condition no adaptation was observed. In order to understand why the speech conditions induced a specific behavior of the subjects, the authors assessed whether the perturbation of the jaw path did provide changes to the speech acoustics. For that, they measured and compared for vocalized speech the frequencies of the first two formants

during the transition from [i] to [a] under 4 conditions: (1) in normal condition, before the introduction of the perturbation; (2) at the beginning of the perturbed condition; (3) at the end of the perturbed condition; (4) at the end of the normal condition after the perturbation was removed. No significant differences were observed between the different conditions. Perceptual tests were also carried out and no systematic distinction could be made by the listeners between the stimuli produced in the different conditions. The last two results speak against the hypothesis of an adaptation process guided by the non-achievement of specific acoustic or perceptual goals.
In this experiment, since both speech and non speech movements used the same articulators in very similar ranges of displacement and duration, the differences observed between these two categories of movement in the impact of the perturbation cannot be attributed to any peripheral phenomenon, such as muscle mechanics or jaw dynamics. They have to be associated with differences in the motor control at the origin of the movements. According to Tremblay *et al.* (2002), they reflect differences in the specification of the goals in the articulatory domain: the time variation of the somatosensory information during the movement is part of the goal for speech production, while it is not the case for non speech movements.
These observations support the hypothesis of an articulatory representation of the speech production task in the speaker's brain.

3.4. Mirror neurons

Before concluding, it seems important to mention an extremely interesting finding that was recently made in neurophysiology for monkeys, namely the *mirror neurons* (Rizzolatti *et al.*, 2001). Indeed, this finding could become a strong support for the Motor Theory of speech perception, if similar findings could be made for human subjects in the future.
Rizzolatti and colleagues have discovered, in area F5 of the premotor cortex of macaque monkeys, neurons that are activated when a monkey grasps food with its hand, and also when the monkey does not move but observes an experimenter grasping the food with his hand. In other words, the discovery of the mirror neurons shows that the observed action leads to resonance in the internal neural circuit of the observer, which is normally activated during execution of a similar action. It should be noted that these neurons are not activated in visual perception if the observed movement does not belong to a category of movement that the monkey is able to produce, and to produce for an identified purpose. Thus, mirror neurons constitute a neural system matching action observation with action execution. It was suggested that this matching system could be at the basis of action understanding.

More recently, mirror neurons associated with the production and the observation of orofacial movements, such as lip-smacking or lip protrusion to take food, were observed in a brain area of macaque monkeys which is close to the Broca area of human brain (Ferrari *et al,* 2003). According to Rizzolatti & Craighero (2004),

> "*There are no studies in which single neurons were recorded from the putative mirror-neuron areas in humans.*[…] *There is, however, a rich amount of data proving, indirectly, that a mirror-neuron system does exist in humans. Evidence of this comes from neurophysiological and brain-imaging experiments*" (p. 174).

These experiments have shown in particular that when humans observe another human who is achieving a motor task, their motor cortex becomes active, even if no movement is actually produced. This suggests the existence of

> *"a neurophysiological mechanism* [in humans] *that creates a common* […], *non arbitrary, semantic link between communicating individuals"* (Rizzolatti & Craighero, 2004, p. 183; see also Rizzolatti & Arbib, 1998, for more details related to this hypothesis for speech).

Such a mirror-neuron system could be the basis of sensorimotor representations of speech. This is why Rizzolatti and colleagues' discovery is often seen as a potential support for the Motor Theory of speech perception.

4. Conclusions

In a theoretical approach assuming that the characteristics of the speech production and speech perception systems are the results of a strong mutual interaction, it was proposed to link the representations of the speech production task in the speaker's brain with the potential objects of speech perception. In pursuit of this aim, four reference speech perception theories were analyzed. In doing so, we ended up with three main questions:

1. Are the speech production task's representations acoustic or articulatory?
2. Do they correspond to invariant or to variable characteristics?
3. Is it necessary to actually achieve these characteristics in ongoing speech production?

To illustrate how experimental studies could help us finding (partial) answers to these questions, the results of a few recent perturbation experiments and of a neurophysiological study of animals were presented. What did we learn?

First of all, the representation of the speech production task in the speakers' brain is probably not purely acoustic and not purely articulatory. Evidence was

found in the lip tube experiment and in the dental prosthesis experiment that these representations have an acoustic component. At the same time, the velocity dependent jaw perturbations demonstrated the existence of an articulatory component in these representations. Hence, our proposal is that speech production task's representations are multimodal, i.e. both acoustic and articulatory.

However, it is important to emphasize that the acoustic and the articulatory modalities do not seem to be equally important, and that a hierarchy seems to exist among them. Indeed, some experimental works that were proposed in this paper (the lip tube and dental prosthesis experiments) show clearly that when both articulatory and acoustic characteristics of normal speech are modified by external perturbations, the speakers elaborate new strategies, in order to correct the acoustical output as their main priority. Compensation did only correct the articulatory configurations when the perturbations did not endanger the achievement of the acoustical goal (see the velocity dependent jaw perturbation experiment). We are not aware of experiments where the speakers accepted changes in the acoustical output, in order to preserve specific articulatory properties. Hence, it can be logically hypothesized that the acoustic component of the speech production task's representation is essential, primary, and that the articulatory component are of secondary importance.

The observation of learning in the alternations of the C4 and C3 blocks in the dental prosthesis experiment, and in particular the fact that improvements were also observed during the C3 blocks without auditory feedback, suggests that the articulatory component of speech production's representations could emerge from a learning guided by the acoustic representation, and that this learning could be very fast for adult speakers. The emergence of this secondary component of the task's representation as a correlate of the primary acoustic component could be, for the speakers, a way to simplify the control by projecting a distal objective (the acoustic product in the external world) into a more proximal one (in the orosensory domain). For speech this could be a particularly efficient way to simplify the control, because the transformations from the orosensory domain to the acoustic one are non-linear and non bi-univocal. In the continuity of this hypothesis, it can assumed that, within the articulatory domain, the speaker could also learn a representation in terms of motor commands associated with the orosensory goals. In the light of the role of mirror neurons in visual perception in monkeys, this projection in the motor domain could provide an efficient framework for identification and classification of phonetic units. Thus, after speech learning, the representation of the task could consist of components in the motor control domain, in the orosensory domain and in the acoustic domain, with an increasing importance from the motor control component to the acoustic one. In normal speech

production, these three levels of representation are equivalent. Planning and monitoring of speech production could thus be made in either of these domains, or in a hybrid domain based on complex, possibly phoneme-dependent combinations of the three components. However, when perturbations modify the speech conditions, so that the goals of different nature cannot be matched simultaneously, priority will be given to the achievement of the acoustic goals.

Are these representations associated with invariant characteristics? The velocity dependent jaw perturbations experiment suggests the existence of an absolute invariant in the orosensory domain, since the motion path of the jaw in its whole is quite perfectly reproduced from repetition to repetition and across experimental conditions. The dental prosthesis experiment suggests that acoustic representations could be associated with a relative invariant, which describes relations between physical characteristics of speech (in this case the steepness ratio between the slopes of the spectral envelop in the low and in the high frequency domain of the speech spectrum). The lip tube experiment suggests that speech goals would be regions of the acoustic space combining F0, F1 and F2, rather than relative or absolute invariants. This last hypothesis is more compatible with the well-known articulatory and acoustic variability of natural speech than the invariant hypothesis. Hence, in agreement with Guenther *et al.* (1998) (see also Keating, 1988, for a first proposal along these lines), our suggestion is that representations of the speech production task associate phonetic units with specific regions in the motor, orosensory and acoustics domains. The size of these regions could be variable according to the phonetic unit and, also, to the speaking style. This last hypothesis could explain intrinsically a part of the observed variability of speech signals.

Is it part of the specification of the speech task for the speaker to generate motor, orosensory and auditory characteristics that are in these regions? According to the hypothesis that we proposed about a hierarchy among the different levels of representation of the speech production task, the answer should be negative for what concerns the motor and the orosensory domains. For the acoustic domain, things are less clear. The different experiments that were presented in this paper do not permit an answer, since they did not involve changes in speaking rate, speaking style or clarity. The answer is strongly dependent on whether and to what extent the human speech perception system could be able to recover intentions in motor tasks that are not achieved. The mirror neurons, if they exist in the human brain, could participate to such an intention recovering process, since the gesture identification and classification based on these neurons seem to be related to the gesture intentionality. From this perspective the projection of the primary acoustic component of speech production task's representation into

the motor control domain would be particularly helpful. Other proposals have also been made involving target recovery in case of target undershoot based on internal models of the peripheral speech apparatus in the brain (Lœvenbruck & Perrier, 1997). In the line of the Adaptive Variability Theory we could also imagine that non-linguistic contextual information could be integrated to deal with cases where the acoustic regions defining the speech goals are not reached. This is still an unsolved question.

An important aspect of the representations of the speech production task was not treated in this paper: the representation of time. This is obviously an important drawback, since time is an essential component of speech. It carries phonemic and prosodic information that is at least as important as the configurational aspects that were considered in this paper, in the spatial and in the frequency domains. The time issue is even more complex since it addresses at the same time the cognitive issue of the representation and the perception of time in human beings, and the physical issue of the relation between time and dynamics in physical systems. Time in speech is the complex combination of both aspects. Addressing this issue, together with those of the internal representations of the peripheral speech apparatus in the brain, of the intentionality recovering, and of the potential role of mirror neurons in speech perception, could be a nice challenge for a tutorial during the next Lubmin Summerschool on *Cognitive and physical models of speech production and speech perception and of their interaction*…..

References

Atal, B.S., Chang, J.J., Mathews, M.V. & Tukey, J.W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique. *Journal of the Acoustical Society of America, 63*, 1535-1555.

Blumstein, S.E. (1986). On acoustic invariance in speech. In J.S. Perkell & D.H. Klatt (Eds.), *Invariance and Variability in Speech Processes* (pp. 178-193). Hillsdale N.J.: Lawrence Erlbaum.

Blumstein, S.E. & Stevens, K.N. (1979). Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America, 66 (4)*, 1001-1017.

Fant, G. (1960). *Acoustic Theory of Speech Production*. he Hague: Mouton.

Ferrari, P. F., Gallese, V., Rizzolatti, G. & Fogassi, L. (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *European Journal of Neuroscience, 17,*1703-1714.

Fowler, C.A (1979). "Perceptual centers in speech production and speech perception. *Perception and Psychophysics, 25*, 375-388.

Fowler, C.A. (1986). An event approach of the study of speech perception from a direct-realist perspective. *J. Phonetics, 14,* 3-28.

Fowler, C.A. (1991). Auditory perception is not special: We see the world, we feel the world, we hear the world. *Journal of the Acoustical Society of America,89*, 2910-2915.

Fowler, C.A (1996). Listener do hear sounds no tongues. *Journal of the Acoustical Society of America, 99(3)*, 1730-1741.

Gibson, J.J. (1966). *The sense considered as perceptual systems*. Boston: Houghton Mifflin. (cited by Fowler, 1986)

Gomi, H. & Kawato, M. (1996). Equilibrium-point control hypothesis examined by measured arm stiffness during multijoint movement. *Science, 272,*117-120.

Gribble, P.L., Ostry, D.J., Sanguineti, V. & Laboissière, R. (1998). Are complex control signals required for human arm movement? *Journal of Neurophysiology, 79*, 1409-1424.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation and rate effects in a neural network model of speech production. *Psychological Review, 102,* 594–62.

Guenther, F. H., Hampson, M. & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review, 105,* 611–633.

Hawkins, S. (2004). Puzzles and patterns in 50 years of research on speech perception. *Proceedings of the Conference "From Sound to Sense"* (CDROM) (pp. B-223 – B246). Cambridge, Massachusetts: Research Laboratory of Electronics, Massachusetts Institute of Technology.

Jones, J.A. & Munhall, K.G. (2005). Learning to produce speech with an altered vocal tract: The role of auditory feedback. *Journal of the acoustical Society of America, 113(1)*, 532-543

Jordan, M.I. & Rumelhart, D.E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science, 16,* 316-354.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences, 17(2)*, 187-245.

Kawato, M., Furukawa, K. & Suzuki, R. (1987). A hierarchical neural network model for control and learning of voluntary movement. *Biological Cybernetics, 57,* 169-185.

Keating P.A. (1988). The window model of coarticulation: articulatory evidence. *UCLA Working Papers in Phonetics, 69*, 3-29. Los Angeles : University of California.

Laboissière, R., Ostry, D.J. & Feldman, A.G. (1996). Control of multi-muscle systems: human jaw and hyoid movements. *Biological Cybernetics, 74,* 373–384.

Liberman, A.M; (1996). *Speech: A special code*. Cambridge Massachusetts: MIT Press.

Liberman, A.M. & Mattingly, I.G. (1985). The motor theory of speech production revised. *Cognition, 21*, 1-36. (Note: The page numbers referenced in the text correspond to the reproduction of the *Cognition* paper published in *Haskins Laboratories Status Report on Speech Research, SR-82/83*, pp. 63-93)

Liberman, A.M. & Mattingly, I.G. (1989). A specialization for speech perception. *Science, 243,* 489-494.

Liberman, A.M., Cooper, F., Shankweiler, D. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431-461.

Lindblom, B. (1988). Phonetic invariance and the adaptive nature of speech. In Ben A.G. Elsendoom & H. Bouma (Eds.), *Working Models of Human Perception* (pp. 139-173). London, UK: Academic Press.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W.J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 403-439). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America, 99(3)*, 1683-792.

Lœvenbruck, H. & Perrier, P. (1997). Motor control information recovering from the dynamics with the EP Hypothesis. *Proceedings ofEUROSPEECH'97* (Vol. 4, pp. 2035-2038). International Speech Communication Association.

McGowan R. S. & Faber A. (1996). Introduction to papers on speech recognition and perception from an articulatory point of view. *Journal of the Acoustical Society of America, 99(3)*, 1680-1682.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices, *Nature, 264*, 746-748.

Perkell, J.S. & Klatt, D.H. (1986). *Invariance & Variability in speech processes.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Perkell, J., Matthies, M.L., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J. & Guiod, P. (1997). Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech Communication*, *22*, 227-250.

Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Perrier, P., Vick, J., Wilhelms-Tricarico, R. & Zandipour, M. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics, 28,* 233-272.

Perrier, P. (2005). About speech motor control complexity. In J. Harrington & M. Tabain (eds), *Speech Production: Models, Phonetic Processes, and Techniques.* Psychology Press: Sydney, Australia (To appear)

Perrier, P., Payan, Y., Zandipour, M., & Perkell, J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *Journal of the Acoustical Society of America, 114,* 1582-1599.

Rizzolatti, G., Fogassi, L. & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Review Neuroscience, 2* , 661-670

Rizzolatti, G. & Arbib, M.A. (1998). Language within our grasp. *Trends in Neuroscience, 21,* 188-194.

Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27,*169-192.

Sabes, P.N. & Jordan, M.I. (1997). Obstacle avoidance and a perturbation sensitivity model for motor planning. *The Journal of Neurosciences, 17(18).,* 7119-7128

Savariaux, C., Perrier P. & Orliaguet, J.-P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: a study of the control space in speech production. *Journal of the Acoustical Society of America, 98*, 2428-2442.

Savariaux, C., Perrier, P., Orliaguet, J.P. & Schwartz, J.L. (1999). Compensation strategies for the perturbation of French [u] using a lip tube. II. Perceptual analysis. *Journal of the Acoustical Society of America, 106*, 381-393.

Shadle, C.H. (1989). Articulatory-acoustic relationships in fricative consonants. In W.J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 211-240). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Schwartz, J.L., Abry, C., Boë, L.J. & Cathiard, M. (2002). Phonology in a theory of perception-for-action-control. In J. Durand & B. Laks (eds.) *Phonology: from Phonetics to Cognition* (pp. 255-280). Oxford: Oxford University Press.

Sock, R. (2001). La théorie de la viabilité en production-perception de la parole. In D. Keller, J.-P. Durafour, J.-F. Bonnot & R. Sock (Eds.), *Percevoir : monde et langage* (pp. 285-316). Liège, Belgium: Mardaga (Psychologie et Sciences Humaines)

Soechting, J.F & Lacquaniti, F. (1981). Invariant characteristics of a pointing movement in man. *The Journal of Neurosciences, 1,* 710-720.

Stetson, R.H. (1928). Motor Phonetics: a study of speech movements in action. *Archives néerlandaises de phonétique expérimentale, 3*, 1-216. (2nd edition., 1951, North Holland: Amsterdam. 1988, by J.A.S. Kelso & K.G. Munhall, Boston).

Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In Jr. E.E., David & P.B., Denes (Eds.) *Human Communication: A unified view* (pp. 51-66). New York: Mc Graw Hill.

Stevens, K.N. (1998). *Acoustic phonetics*. Cambridge, Massachusetts: MIT Press.

Stevens, K.N. & Blumstein, S.E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64,* 1358-1368.

Stevens, K.N. (1989). On the quantal nature of speech. *J. Phonetics, 17*, 3-45.

Stevens, K.N. (1996). Critique: Articulatory-acoustic relations en their role in speech perception. *Journal of the Acoustical Society of America, 99(3),* 1693-1694

Tremblay, S., Shiller, D.M. & Ostry, D.J. (2003). Somatosensory basis of speech production. *Nature, 423,* 866-869.

Wolpert, D.M., Ghahramani, Z. & Jordan, M.I. (1995). Are arm trajectories planned in the kinematic or dynamic coordinates? An adaptation study. *Experimental Brain Research, 103*, 460-470.

Towards functional modelling of relationships between the acoustics and perception of vowels

Hartmut R. Pfitzinger
Inst. of Phonetics and Speech Communication, Munich University, Germany

This paper summarizes our research efforts in functional modelling of the relationship between the acoustic properties of vowels and perceived vowel quality. Our model is trained on 164 short steady-state stimuli. We measured F1, F2, and additionally F0 since the effect of F0 on perceptual vowel height is evident. 40 phonetically skilled subjects judged vowel quality using the Cardinal Vowel diagram. The main focus is on refining the model and describing its transformation properties between the F1/F2 formant chart and the Cardinal Vowel diagram. An evaluation of the model based on 48 additional vowels showed the generalizability of the model and confirmed that it predicts perceived vowel quality with sufficient accuracy.

1. Introduction

As early as 1890 Lloyd claimed that vowels with similar qualities have similar formant frequency relations. During the following 62 years almost none of the phonetic investigations on vowel quality contradicted this claim, which was at the time remarkable. And even until recently, most vowel quality studies (e.g. Fricker 2004) take into account only F1 and F2 and persistently ignore the knowledge acquired during the second half of the 20th century. Therefore, it appears to be useful to recall some of the relevant studies which led to decisive conclusions and moved the vowel quality research forward.
Peterson and Barney (1952) recorded 76 American subjects (33 male, 28 female, and 15 children) producing 10 isolated monosyllabic words[1] two times. The resulting 1520 words ($= 76 \cdot 10 \cdot 2$) were presented to 70 listeners who had to judge which of the 10 words they perceived, leading to 106400 judgements. Formant charts with all 1520 vowels in the F1/F2 space revealed strongly overlapping regions even if non-uniformly judged vowels as well as all vowels of

[1]These words were *heed, hid, head, had, hod, hawed, hood, who'd, hud,* and *heard.*

female speakers were excluded. Obviously, the F1/F2 space failed to represent the vowel quality sufficiently, either in absolute or in relative terms.
Miller (1953) systematically varied the fundamental frequency of synthetic monophthongs while keeping their spectral envelopes constant. He found a shift of perceived vowel quality, i.e. if F0 was doubled the perceived vowel height was raised. Many subsequent studies have confirmed these results (Traunmüller 1981; Syrdal and Gopal 1986; Di Benedetto 1987; Rooney, Vaughan, Hiller, Carraro, and Laver 1993). They suggest that the distance between Bark-transformed F1 and F0 corresponds to perceptual vowel height, and the distance between Bark-transformed F2 and F1 represents perceptual vowel backness.
Inspired by the vowel perception experiments of Ladefoged (1967) who presented single-syllable word stimuli to a group of skilled phoneticians thus achieving reliable vowel quality assessments, in Pfitzinger (1995) we investigated the perception of isolated monophthong stimuli produced by a single speaker and judged by 20 skilled phoneticians. Based on the perception results we developed our first functional model for speaker-dependent prediction of perceptual vowel quality from acoustic measurements of F0, F1, and F2.
In Pfitzinger (2003a, 2003b) we developed and improved a functional model based on Multiple Linear Regression analysis of acoustic and perception data of 100 monophthongs cut from German read speech produced by 12 speakers. Again, F0, F1, and F2 were measured to represent the acoustic properties. Judgements of 40 phonetically trained subjects measured as x- and y-coordinates in the Cardinal Vowel diagram (Jones 1962) served as perception data. The resulting model appropriately and speaker-independently predicts perceptual vowel quality from acoustic measurements. The inverse formulae of the model enable the frequencies of F1 and F2 to be estimated from a perceptually specified vowel quality (b,h) and a given target fundamental frequency. (b,h) refer to *perceptual vowel <u>b</u>ackness* and *<u>h</u>eight* in an arbitrarily defined coordinate system superimposed on the Cardinal Vowel diagram as shown in Figure 1.

1.1. Functional Modelling

Functional modelling is central to most of our investigations. It involves not only understanding the function of a component and its impact on other components of the speech chain. It also provides a formal description (usually in the form of a computer program) which allows the accuracy of the model to be evaluated. The evaluation step is obligatory since all functional models are in some sense only simplified imitations of samples of natural real-world processes and thus always show a more or less imperfect behaviour.

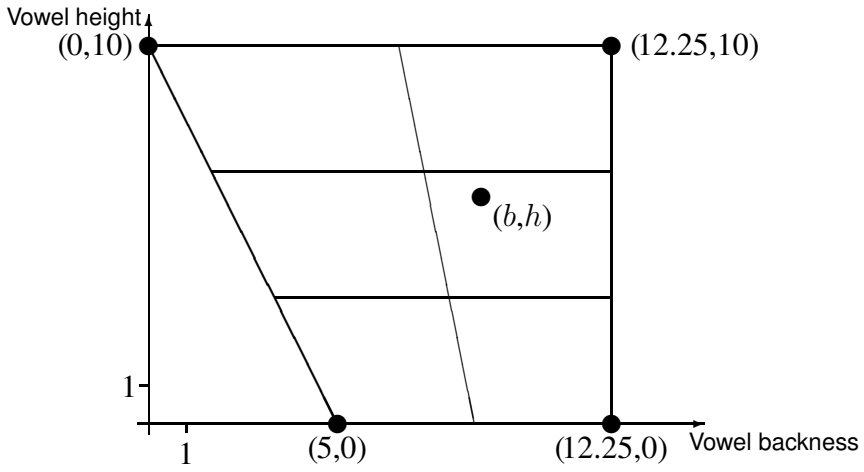


Figure 1: Dimensions of the Cardinal Vowel diagram used in our perception tests and in the vowel quality prediction formulae.

2. Refinement of the Vowel Quality Prediction Model

In the present study we refined the reliability of the model developed in Pfitzinger (2003a) by estimating the model coefficients from extended acoustic and perception data: F0, F1, and F2 frequencies together with the coordinates (b,h) of 164 vowel tokens were submitted to Multiple Linear Regression analysis. They consist of the original 100 monophthongs cut from German read sentences produced by 6 female and 6 male speakers, and of 64 new monophthongs cut from German spontaneous speech of further 4 female and 4 male speakers. The new vowels were also judged by 40 phonetically trained subjects.

3. Results

The increased number of stimuli changed the vowel quality prediction formulae presented in Pfitzinger (2003a) slightly: while the former model predicted vowel backness more accurate when including F0, backness prediction of the refined model did not benefit from F0 information. Presumably, the inclusion of F0 in backness prediction of the former model was due to over-adaption to the training data. Therefore, the corresponding formula was reduced to only three coefficients. The refined formulae are:

$$\begin{aligned} \hat{h} &= 3.122\log(\mathrm{F}_0) - 8.841\log(\mathrm{F}_1) + 44.16 \\ \hat{b} &= 1.782\log(\mathrm{F}_1) - 8.617\log(\mathrm{F}_2) + 58.29 \end{aligned} \tag{1}$$

where F0, F1, and F2 are in Hz. The estimated values for perceptual vowel height $\hat{h}$ and backness $\hat{b}$ refer to the dimensions displayed in Figure 1. The *end-of-scale effect* (Traunmüller 1981, p. 1469) poses a problem to any vowel

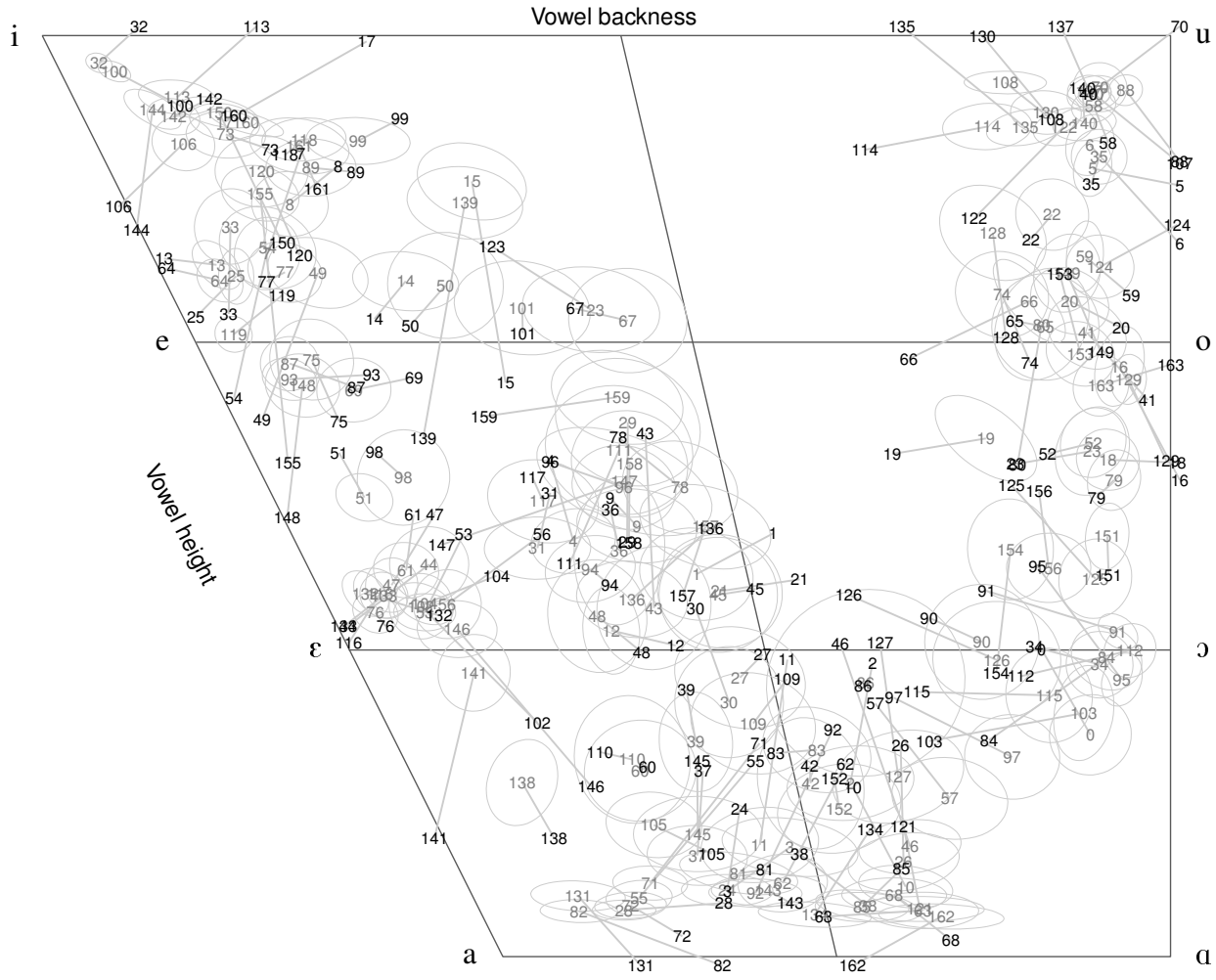


Figure 2: Mean perception results and 90% confidence ellipses (*light colors*) of 164 German monophthong stimuli and their predicted positions (*dark numbers*).

quality prediction model. Generally, it is a perceptual saturation effect limiting acoustic values that exceed the end of a perceptual scale to the perceptual limit. Accordingly, whatever fundamental and formant frequencies vowels have, they lie within the Cardinal Vowel diagram boundaries.

In favour of its simplicity our model ignores these effects and therefore transforms vowels with end-of-scale F0, F1, or F2 into positions outside the Cardinal Vowel diagram. Consequently, it is necessary to graphically move vowel tokens from outside the diagram boundaries to the boundary coordinate values.

The correlation coefficients of this refined model are $r = 0.98$ between perceptual and predicted vowel backness and $r = 0.96$ between perceptual and predicted height of the 164 vowels used in the training of the model. This corresponds to a mean deviation of $\pm\frac{1}{18}$ of the mean Cardinal Vowel diagram width and $\pm\frac{1}{15}$ of its height. The training vowels and their predicted positions are shown in Figure 2.

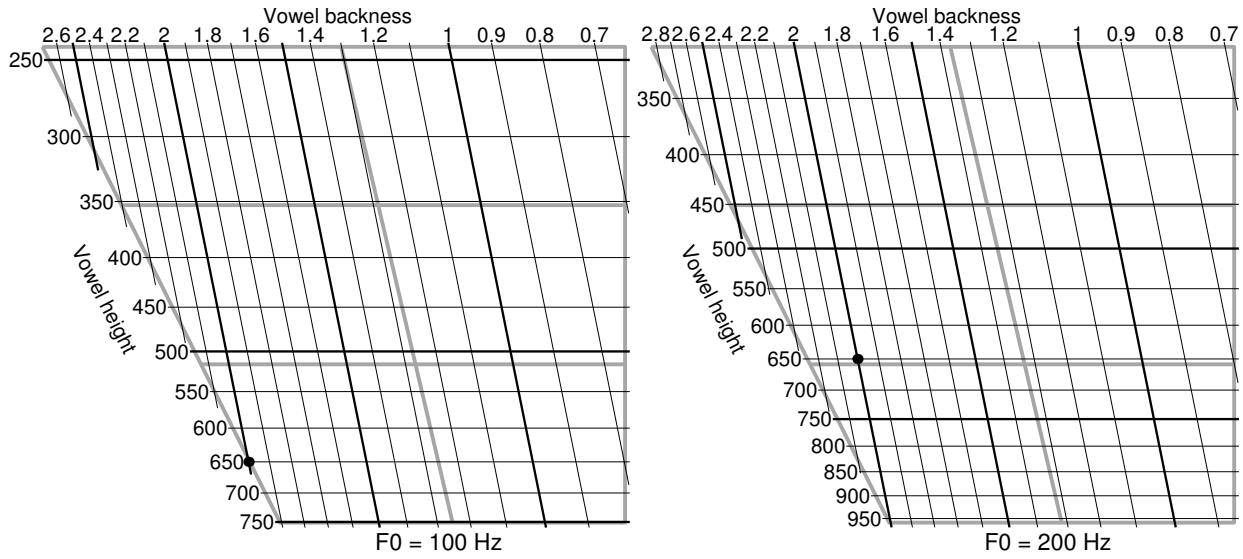


Figure 3: Relationship between formant frequencies (vertical: F1 in Hz, horizontal: F2 in kHz) and the Cardinal Vowel diagram estimated by the refined model for two selected F0 frequencies. Both marked vowels have the same F1 and F2 frequencies.

4. Analysis of the Transformation Properties

The formulae of the model also allow to systematically project all F1/F2 formant combinations onto the Cardinal Vowel diagram for a given F0. As an illustrative example, this is done for two different fundamental frequencies (100 Hz and 200 Hz) and displayed in Figure 3. It clearly shows the warping of the two-dimensional formant space caused by the refined model. In particular, the effect of F0 on perceptual vowel height is evident: while at a fundamental frequency of 100 Hz a first formant frequency of 750 Hz is sufficient for the perception of an open vowel, an F1 of about 950 Hz is necessary if F0 is 200 Hz.
F0 also influences perceived vowel backness: e.g. a vowel with F1/F2 frequencies of 650 Hz/2 kHz and an F0 of 100 Hz is perceived as a front vowel. But with an F0 of 200 Hz it is perceived more retracted (and raised) (see Figure 3).
The coefficients of the inverse formulae of the refined model, which predict the frequencies of F1 and F2 given the coordinates of a vowel quality in the Cardinal Vowel diagram (b,h) (see Figure 1) as well as a fundamental frequency, also changed only slightly compared with Pfitzinger (2003a):

$$\begin{aligned}\widehat{\mathrm{F}_1} &= e^{0.3532\log(\mathrm{F}_0)-0.1131h+4.9951}\\ \widehat{\mathrm{F}_2} &= e^{0.0730\log(\mathrm{F}_0)-0.0234h-0.1160b+7.7974}\end{aligned} \tag{2}$$

It is not surprising that while the refined model in formula (1) predicts perceptual vowel backness b independent of F0, the inverse model in formula (2) requires F0 in both equations. The reason is that in the equation for estimating

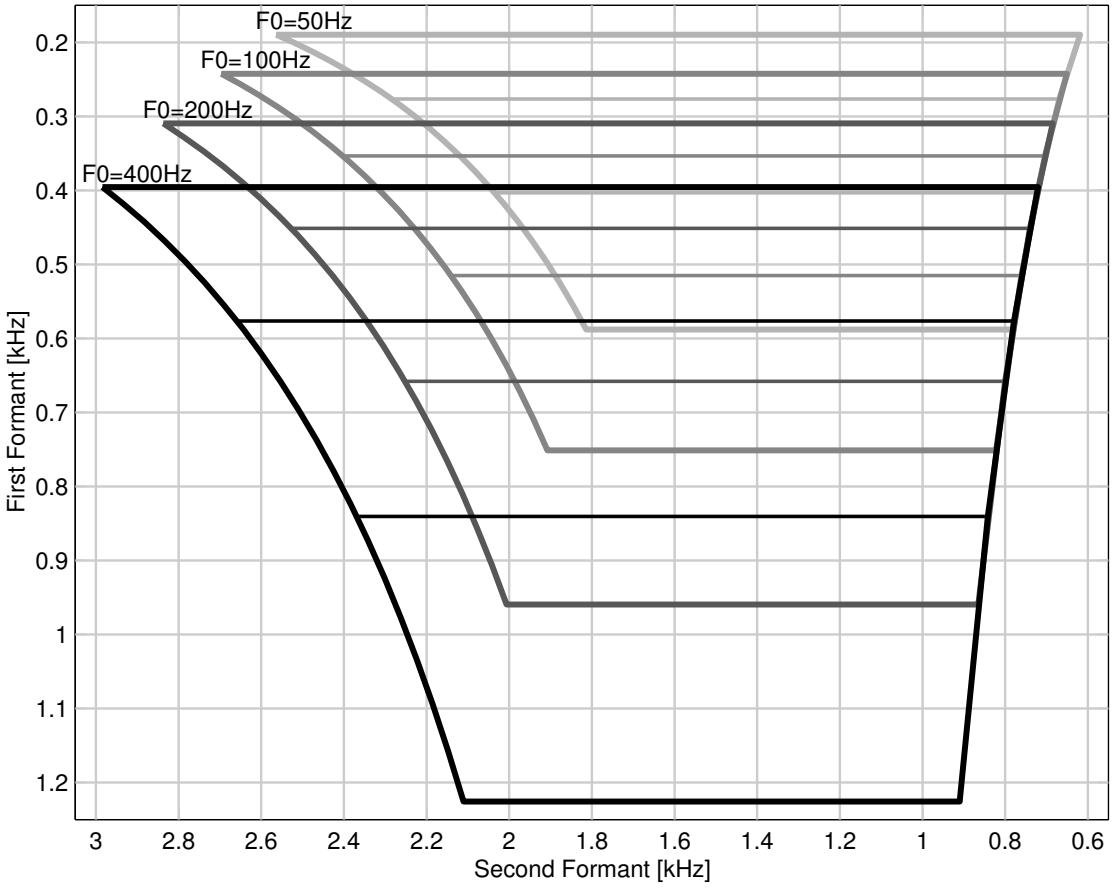


Figure 4: Using the inverse refined model with four different F0 frequencies to project the Cardinal Vowel diagram onto the linear F1/F2 frequency space.

$\widehat{\mathrm{F}_2}$ the original term 'F_1' has been substituted by the equation for estimating $\widehat{\mathrm{F}_1}$. In Figure 4 the boundaries of the Cardinal Vowel diagram are projected onto the linear F1/F2 space using formula (2) and four different F0 frequencies.

It is important that the combination of the trapezoid shape of the Cardinal Vowel diagram and the Cartesian coordinate system being used in this study (see Figure 1) lead to vowel backness values b between 0 and 12.25 for high vowels, while low vowels are transformed into values only between 5 and 12.25. Thus, this vowel quality measurement method per se introduces a small amount of correlation between backness and height.

If the goal is to analyse perception results statistically, the amount of correlation which is technically introduced by the Cartesian coordinate system should be removed from (b,h) measurements by transforming them into a square space:

$$B = 10\,\frac{b + 0.5h - 5}{0.5h + 7.25} \tag{3}$$

The resulting coordinates (B, h) are also useful if the objective is to modify the vowel quality in equally-spaced steps within the Cardinal Vowel diagram.

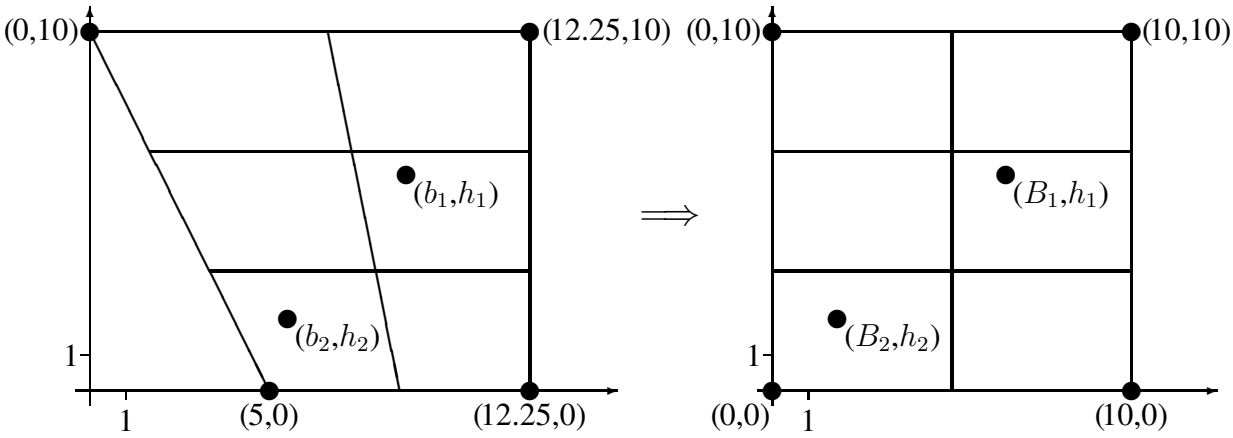


Figure 5: Transformation of perceptual vowel qualities (b_i,h_i) within the Cardinal Vowel diagram (*left*) into a square space (*right*) via formula (3).

Then, the inverse transformation of modified positions within the square space into the Cardinal Vowel diagram is also needed:

$$b = \frac{B\,(0.5h + 7.25)}{10} - 0.5h + 5 \tag{4}$$

Figure 5 portrays the effect of formula (3) on the shape of the Cardinal Vowel diagram. It should be emphasized that when discarding vowel height information the remaining values of the vowel backness dimension b are meaningless except for (i) the case of back vowels or (ii) when referring to B.

5. Evaluation

Acoustic and perception data of 48 monophthongs taken from Pfitzinger (1995) were used to evaluate the refined model. These vowel stimuli corresponded to the following 11 vowel phonemes of the German vowel system: /i/, /e/, /ɛ/, /a/, /ɔ/, /o/, /u/, /ɪ/, /ə/, /ɐ/, and /ʊ/. Additionally, an allophonic realization of /a/ was also recorded which is used in some dialects of southern Germany. Since it is more retracted than the standard German /a/ it is denoted by the symbol /ɑ/.
A native German speaker produced the 12 vowels in isolation with two different fundamental frequencies: 105 Hz and 230 Hz ($\pm$1 semitone). The mean duration of the resulting 24 vowels was 208 ms ($\pm$18%). Inappropriate articulation or too large variation of F0 or duration has been immediately rejected during the vowel stimulus recordings. A second stimulus set was created by carefully shortening the 24 vowels to 50 ms in order to investigate the effect of vowel length on vowel perception (Weiss 1972).
20 skilled German phoneticians perceptually judged each of these 48 vowels 5 times. Thus, each perceptual reference position in the Cardinal Vowel diagram

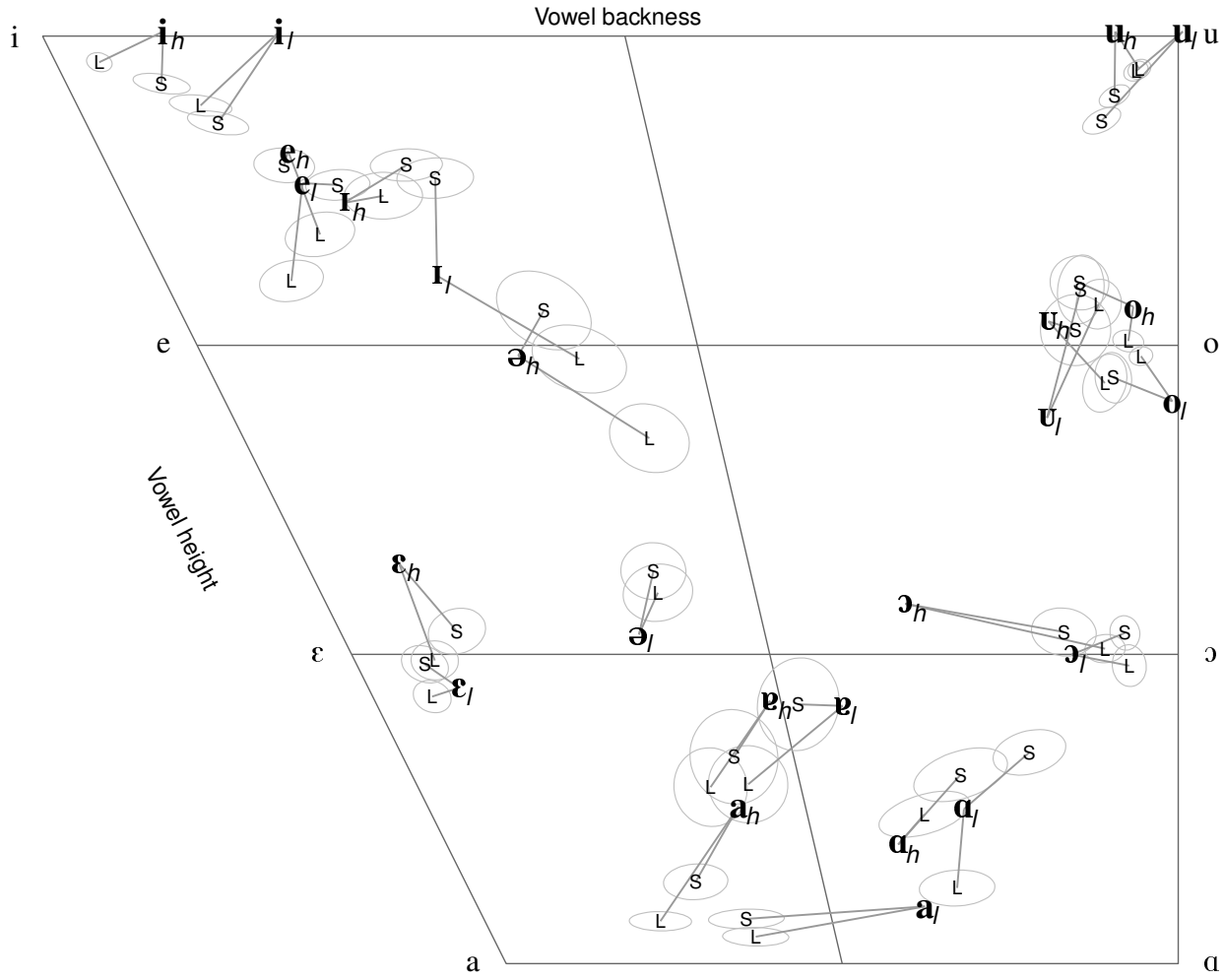


Figure 6: Evaluation results: Perceptual and predicted positions of 24 vowels. Half of them have an F0 of 108 Hz (*l*) and the other half 230 Hz (*h*). Lines join each predicted position (*bold*) with two perceptual reference positions for long (*L*) and shortened (*S*) versions of the vowels.

is derived from 100 judgements.

Figure 6 shows the application of the refined vowel quality prediction model to acoustic measurements of F0, F1, and F2 of the underlying 24 vowels and the 48 perceptual reference positions. The mean deviation of the prediction results from both reference positions (= the average length of all joining lines in Figure 6) is $\pm\frac{1}{17}$ of the mean Cardinal Vowel diagram width and $\pm\frac{1}{17}$ of its height. These deviations are similar to those achieved with the 164 training vowels.

A significant amount of error is due to the vowels /ɔ$_h$/, /a$_l$/, /a$_h$/, /ɪ$_l$/, and /ə$_h$/. Both perceptual reference positions for the vowels /ɔ$_h$/, /a$_l$/, and /a$_h$/ are very close which means that the judgements of the phoneticians were not biased by the length of these vowels. Thus the error is caused by the model.

But for /ɪ$_l$/ and /ə$_h$/ the influence of vowel length on perception is obvious and in accordance with the literature (Weiss 1972): The shortening of these vowels

leads to a raised vowel height perception. The predicted positions are closer to the reference positions for the short vowels which might be due to the fact that all training vowels had a comparatively short duration of 80 ms.

6. Discussion

The evaluation of the refined perceptual vowel quality prediction model revealed that (i) the model generally achieved a reasonable prediction accuracy, and that (ii) even a jury of skilled phoneticians is not able to completely ignore the phonological system of their native language in case of very few vowels. Since the mean duration of the long vowels was 208 ms and regarded as phonologically long, we do not expect our prediction model, which was developed on the basis of short vowels (80 ms), to be able to predict phonologically biased vowel quality judgements with high accuracy.

In Pfitzinger (1995) we have already shown that the shortening of isolated monophthong vowels leads to a significant raising of vowel height judgements of skilled German phoneticians ($\hat{t} \approx 2.639 > t_{0.01;2398} \approx 2.581$, **). And in Dioubina and Pfitzinger (2002) we found that phonetically trained subjects with different native languages do not perfectly agree when judging vowel quality by means of the Cardinal Vowel diagram. Finally, in Pfitzinger (2003a) we reported that skilled phoneticians are not able to exactly repeat their judgements after a period of one year.

Obviously, the experimental method of judging vowel quality by plotting its position in the Cardinal Vowel diagram yields perception data near to the limit of human precision. This method is also suited to evaluate and compare the different levels of experience of phoneticians since in all our perception experiments some phoneticians steadily produced small deviations from the mean group results and from their former individual judgements.

However, we still conclude that mean perception results of a group of phoneticians are the most reliable source for the assessment of vowel quality (Pfitzinger 2003a). The reason for this is that in a group of subjects random deviations of individual subjects from a target position compensate each other so that only systematic effects remain. By increasing the number of participants the reliability of the mean results also increases.

If in a study on vowel quality a group of phoneticians is not available the prediction model could be applied to approximately imitate their mean judgements since only a few skilled phoneticians are able to determine vowel qualities more precisely than the prediction model.

Phonological bias is a top-down process, that means a listener interprets the

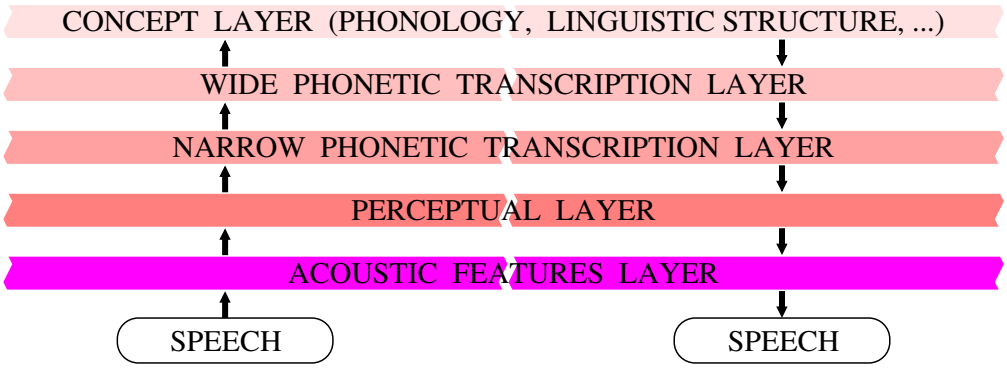


Figure 7: Model of speech analysis/synthesis along layers which have different degrees of abstraction.

sounds of speech with reference to knowledge of the phoneme system and co-occurrences of phonemes of a specific language. Listeners with different native languages interpret the same acoustic stimuli in different ways. E.g. shortening of the vowels /e/ and /o/ leads to a tense/lax category change and the perception of /ɪ/ and /ʊ/ for German listeners (Weiss 1972) but not for Danish listeners (Fischer-Jørgensen 1975). And Dutch listeners perceive an /ɔ/ when shortening an /o/ (van Son 1993).

Because of the presence of effects like these a functional model which makes no use of phonological knowledge and which uses only vowel intrinsic features can not sufficiently predict the phonological vowel category. Nevertheless, many studies (e.g. Syrdal and Gopal 1986) try to solve the problem of acoustic-to-phonological mapping by taking into account only vowel intrinsic features. It seems that this problem is underestimated.

In Figure 7 we try to illustrate the outline of our theory on vowel identification: between the acoustic layer and the concept layer (which contains the phonological layer and other higher-level knowledge bases) are at least three additional layers with different degrees of abstraction. The "wide phonetic transcription layer" is e.g. used in spoken language databases to enable access to the speech signal by means of a very limited set of labels. These coarse labels are phonologically motivated but denote real speech segments which appear in various allophonic realizations. In contrast, the "narrow phonetic transcription layer" additionally provides all phonetic symbols and diacritics to symbolically describe the segmental features of the speech sounds as precisely as possible. Finally, the "perceptual layer" is a continuous layer closely related to the acoustic features of speech but with parameters in a meaningful and easy-to-modify domain (such as the Cardinal Vowel diagram). Note that only the "acoustic features layer" contains — highly encoded — information about the gender or age of a speaker.

In this paper we only solved the problem of transition from the acoustic features of vowels to the perceptual layer. In Pfitzinger (2003b) we investigated several ways to further abstract from vowel quality information but with only limited success since we did not include contextual or dynamic information. This remains to be done.

Since only F0, F1, and F2 are taken into account the generation of synthetic two-formant stimuli via the inverse model could lead to mean deviations greater than the investigated transformation accuracy from the acoustic to the perceptual vowel quality representation. Therefore experiments with synthetic vowels are subject to our future research. The vowels of the Secondary Cardinal Vowel diagram are conspicuously excluded from this paper since their investigation is not finished yet.

Acknowledgements

I would like to thank Hansjörg Mixdorff, Parham Mokhtari, Uwe Reichel, and two anonymous reviewers for their helpful comments on first drafts of this paper, and BMW Group Research and Technology Pty Ltd, Munich for their financial support.

References

Di Benedetto, M.-G. (1987). On vowel height: Acoustic and perceptual representation by the fundamental and the first formant frequency. In: *Proc. of the XIth Int. Congress of Phonetic Sciences*, vol. 5. Tallinn, 198–201.

Dioubina, O. I. and H. R. Pfitzinger (2002). An IPA vowel diagram approach to analysing L1 effects on vowel production and perception. In: *Proc. of ICSLP '02*, vol. 4. Denver, 2265–2268.

Fischer-Jørgensen, E. (1975). Perception of German and Danish vowels with special reference to the German lax vowels. In: G. Fant and M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech*. London, New York, San Francisco: Academic Press, 153–176.

Fricker, A. B. (2004). The change in Australian English vowels over three generations. In: *Proc. of the 10th Australian Int. Conf. on Speech Science and Technology (SST 2004)*. Sydney, 189–194.

Jones, D. (1962). *An outline of English phonetics* (9. ed.). Cambridge: W. Heffer & Sons Ltd.

Ladefoged, P. (1967). *Three areas of experimental phonetics*. London: Oxford University Press.

Lloyd, R. J. (1890). Speech sounds: Their nature and causation. *Phonetische Studien 3*, 251–278. (1891, vol. 4: 37–67, 183–214, 275–306).

Miller, R. L. (1953). Auditory tests with synthetic vowels. *J. of the Acoustical Society of America 25*(1), 114–121.

Peterson, G. E. and H. L. Barney (1952). Control methods used in a study of the vowels. *J. of the Acoustical Society of America 24*(2), 175–184.

Pfitzinger, H. R. (1995). Dynamic vowel quality: A new determination formalism based on perceptual experiments. In: *Proc. of EUROSPEECH '95*, vol. 1. Madrid, 417–420.

Pfitzinger, H. R. (2003a). Acoustic correlates of the IPA vowel diagram. In: *Proc. of the XVth Int. Congress of Phonetic Sciences*, vol. 2. Barcelona, 1441–1444.

Pfitzinger, H. R. (2003b). The /i/-/a/-/u/–ness of spoken vowels. In: *Proc. of EUROSPEECH '03*, vol. 1. Geneva, 809–812.

Rooney, E., R. Vaughan, S. Hiller, F. Carraro, and J. Laver (1993). Training vowel pronunciation using a computer-aided teaching system. In: *Proc. of EUROSPEECH '93*, vol. 2. Technische Universität Berlin, 1347–1350.

Syrdal, A. K. and H. S. Gopal (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *J. of the Acoustical Society of America 79*(4), 1086–1100.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *J. of the Acoustical Society of America 69*(5), 1465–1475.

van Son, R. J. J. H. (1993). *Spectro-temporal features of vowel segments*. Studies in Language and Language Use, 3. Amsterdam: IFOTT.

Weiss, R. (1972). Perceptual parameters of vowel duration and quality in German. In: *Proc. of the VIIth Int. Congress of Phonetic Sciences (Montréal 1971)*. The Hague, Niederlande, Paris: Mouton & Co., 633–636.

Von Kempelen et al. – Remarks on the history of articulatory-acoustic modelling

Bernd Pompino-Marschall

Humboldt-Universität zu Berlin
and
ZAS Berlin

The contribution of von Kempelen's "Mechanism of Speech" to the 'phonetic sciences' will be analyzed with respect to his theoretical reasoning on speech and speech production on the one hand and on the other in connection with his practical insights during his struggle in constructing a speaking machine.
Whereas in his theoretical considerations von Kempelen's view is focussed on the natural functioning of the speech organs – cf. his membraneous glottis model – in constructing his speaking machine he clearly orientates himself towards the auditory result – cf. the bag pipe model for the sound generator used for the speaking machine instead. Concerning vowel production his theoretical description remains questionable, but his practical insight that vowels and speech sounds in general are only perceived correctly in connection with their surrounding sounds – i.e. the discovery of coarticulation – is clearly a milestone in the development of the phonetic sciences: He therefore dispenses with the Kratzenstein tubes, although they might have been based on more thorough acoustic modelling.
Finally, von Kempelen's model of speech production will be discussed in relation to the discussion of the acoustic nature of vowels afterwards [Willis and Wheatstone as well as von Helmholtz and Hermann in the 19th century and Stumpf, Chiba & Kajiyama as well as Fant and Ungeheuer in the 20th century].

1. The person

Wolfgang von Kempelen (1734-1804), civil servant – in later years in the rank of a privy councillor – at the Royal Hungarian Court at Preßburg (today's Bratislava), protégé of Maria Theresa, is present in public memory foremost because of his geniously constructed chess playing 'Turk' (although it was based on deception), an 'automaton' that defeated – among others – the Russian

empress, Catherine the Great, at this royal game (cf. Figure 2). Napoleon's stepson, Prince Eugène de Beauharnais, later bought this 'machine' (but, alas, without the chess champion hidden inside).

Figure 1: Self portrait of Wolfang von Kempelen
(charcoal drawing; Szépmüvészeti Múzeum, Budapest) and signature

But, typical son of his times, Wolfgang von Kempelen was a multitalented person, experimenting in quite different fields of science and engineering.

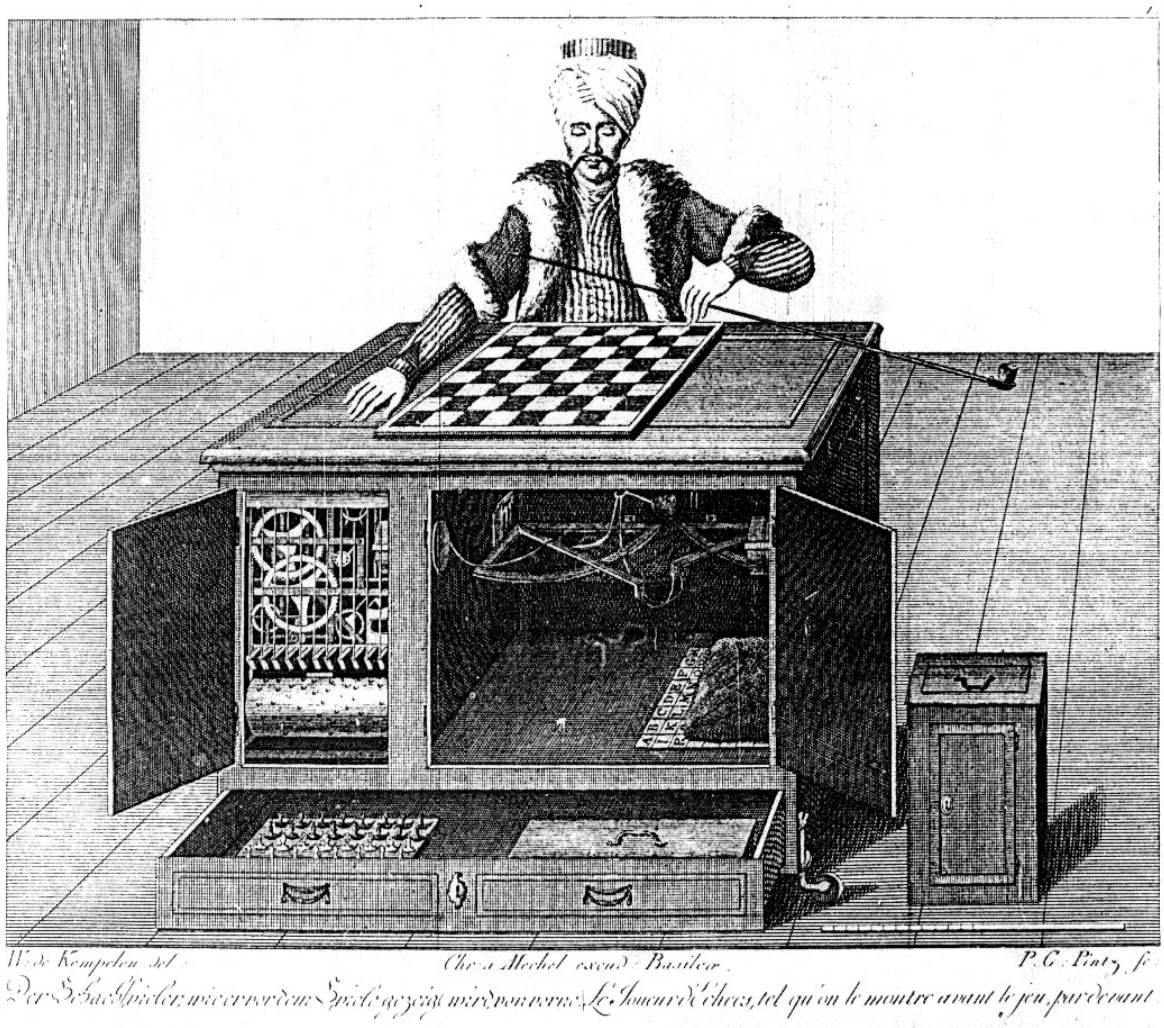


Figure 2: The chess playing 'Turk' as shown from front before the game after the engravings accompanying the "Letters …" of Windisch (1783b)

In focus here is his interest in the mechanism of human speech to which Kempelen dedicated a whole book, "The Mechanism of Human Speech Including the Description of His Speaking Machine" published on demand[1] in a German-French parallel edition of together 195 copies[2] in 1791 (cf. Figure 3). Brücke (1856: 6) – German 'Lautphysiologe' (speech physiologist) and one of the founders of modern phonetics – clearly recommends this book of Kempelen

[1] Cf. the bookseller's 1789 announcement of the publication of the "Mechanism ..." in Figure 3.

[2] At least according to the list of subscribers in the German edition. The German edition is set in black letters, the French edition in Roman type letters.

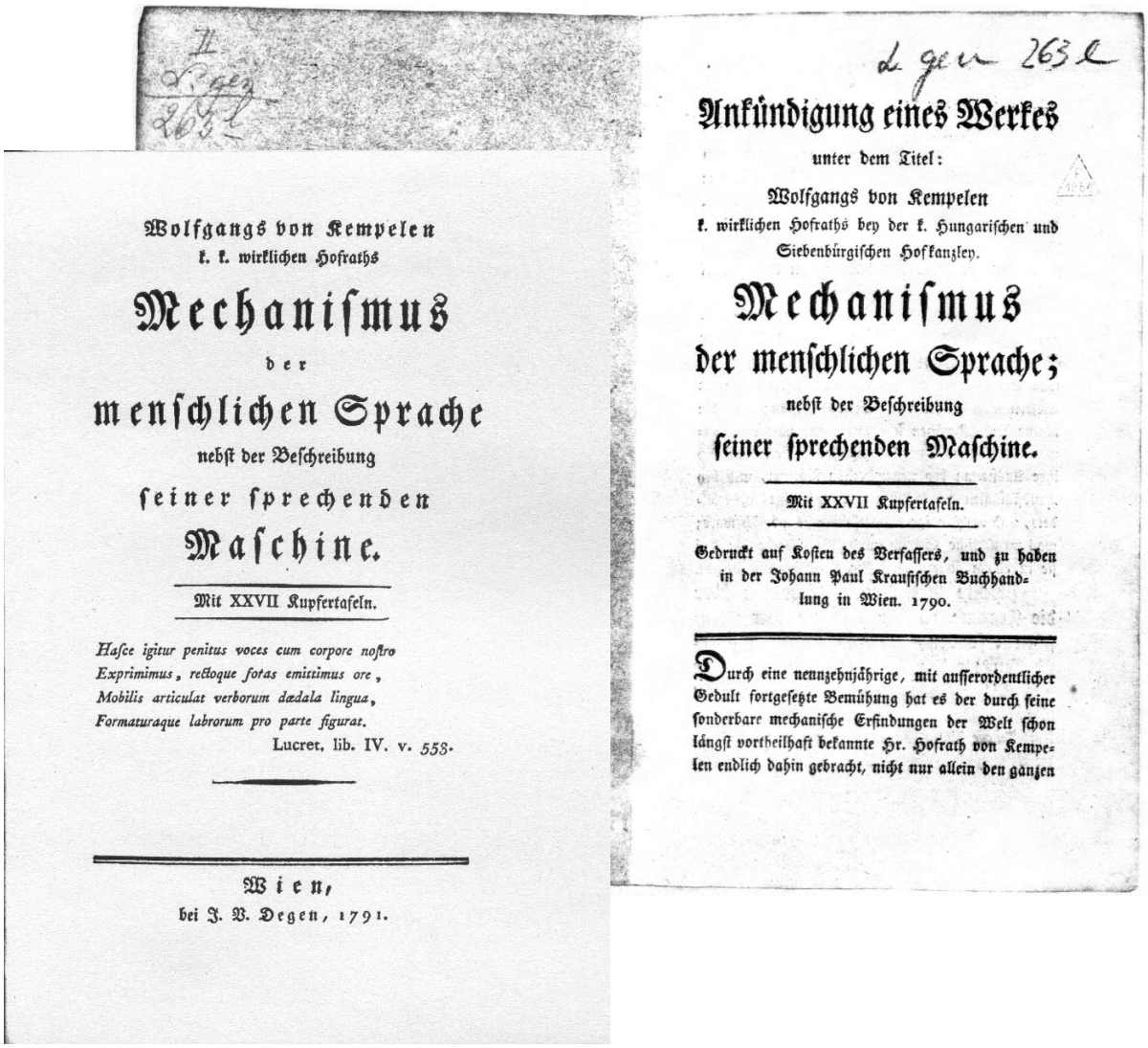
"to all linguists interested in the purely mechanical part of the theory of speech sounds."

2. The construction of the speaking machine

In his "Mechanism …" Kempelen himself tells us about the long time he needed to construct his speaking machine: "I can't tell exactly what forced me to imitate human speech. But I remember that already during my work on the chess player [cf. Figure 2] in 1769 I was eager to find musical instruments resembling the human voice." (Kempelen 1791: 389f.; my translation). His starting point thus was that human speech can be nothing but vibrating air since it is obvious that we breathe for speaking and while exhaling the air is set in motion by the voice membrane.

In his book he then continues to describe how by chance he got hold of the mouthpiece of a shepherd's bagpipe (cf. Figure 4) that sounded to him like a singing child. This kind of mouthpiece as a first step was used by him as a sound generator in an unfinished 'vox humana' organ he bought. For this kind of machine he went on to construct different variable resonators that could be controlled by pressing the keys of a keyboard (cf. Figure 5). He notes some difficulties with the vowel /i/, but since he had then already reached the conclusion that although it would be possible to construct a 'vox humana' for single speech sounds it wouldn't be possible to concatenate these sounds into syllables he was no longer interested in learning more about the Kratzenstein tubes (cf. Figure 6).
The leading ideas behind his approach at a speaking machine at this times can be summarized as following:

- Since speech sounds are only discernable in relation to one another you have to use a *single glottis* and a *single mouth*.
- The mouth and tongue are in *continuous* motion producing obstacles for the *sounding (!)* air.
- And since it is almost mathematically proven that
  speech = voice passing through openings
  it follows that
  for a speaking machine you need nothing else but
  - a lung
  - a glottis
  - and a mouth.



Wolfgangs von Kempelen
k. k. wirklichen Hofraths

Mechanismus
der
menschlichen Sprache
nebst der Beschreibung
seiner sprechenden
Maschine.

Mit XXVII Kupfertafeln.

Hasce igitur penitus voces cum corpore nostro
Exprimimus, rectoque fotas emittimus ore,
Mobilis articulat verborum dædala lingua,
Formaturaque labrorum pro parte figurat.
Lucret. lib. IV. v. 558.

Wien,
bei J. V. Degen, 1791.

Ankündigung eines Werkes
unter dem Titel:
Wolfgangs von Kempelen
k. wirklichen Hofraths bey der k. Hungarischen und Siebenbürgischen Hofkanzley.

Mechanismus
der menschlichen Sprache;
nebst der Beschreibung
seiner sprechenden Maschine.

Mit XXVII Kupfertafeln.

Gedruckt auf Kosten des Verfassers, und zu haben in der Johann Paul Kraussischen Buchhandlung in Wien. 1790.

Durch eine neunzehnjährige, mit ausserordentlicher Geduld fortgesetzte Bemühung hat es der durch seine sonderbare mechanische Erfindungen der Welt schon längst vortheilhaft bekannte Hr. Hofrath von Kempelen endlich dahin gebracht, nicht nur allein den ganzen

Figure 3: Title page (left) and bookseller's announcement (1789) of Kempelen's "Mechanism …" (1791a)

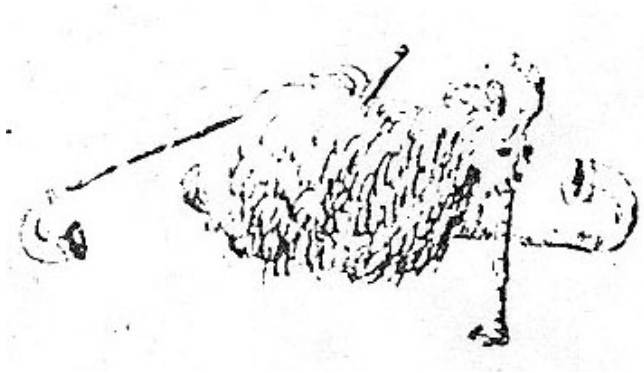


Figure 4: Kempelen's drawing of a Hungarian bagpipe (epigraph to an occasional poem dedicated to Magdalena von Wiesenthal in his family book "Gedichte. von W. v. K." [Lyrics. of W. v. K.]; 1757 ff.; National Hungarian Library)

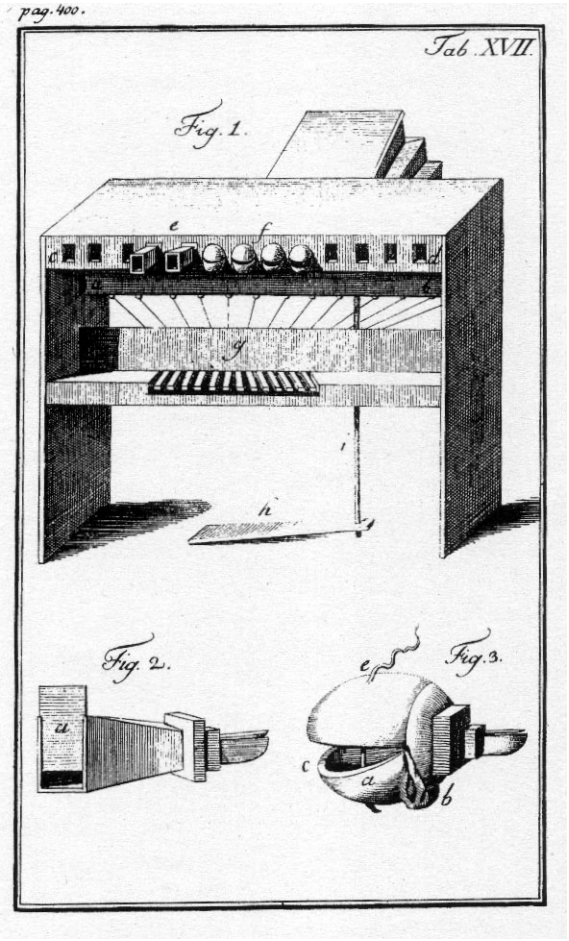


Figure 5: Kempelen's 'vox humana' trial

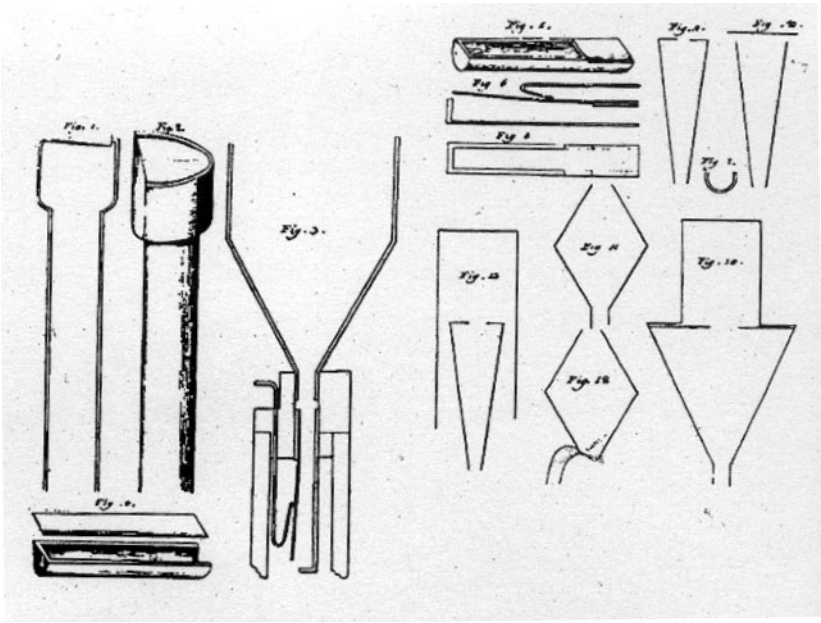


Figure 6: Kratzenstein's vowel tubes (after Panconcelli-Calzia, 1940)

In 1778, according to Bois-Reymond (1862: 129), Kempelen (partially) successfully finished the construction of his speaking machine. Clearly documented in the newspaper literature of that time, 1782 till 1784 Wolfgang von Kempelen was granted a sabbatical by Joseph II during which he undertook a European journey exhibiting both of his 'automata'. He went through Switzerland, stayed in Paris, went on to London and visiting the German fairs at Frankfurt, Dresden and Leipzig on his way back to Hungary, always accompanied by the letters of his friend Karl Gottlieb von Windisch (1783a, b, c, 1784).
The first picture of the machine – more complicated than the one of the "Mechanism ..." (cf. Figure 8) – is given by Hindenburg (1784; cf. Figure 7).

3. Kempelen: Observer vs. engineer

Taking a closer look to his "Mechanism ..." one can see Kempelen's twofold interest in language and speech production as a natural process on the one hand and the engineering task of building a speaking machine whose output sounds like human speech on the other hand.

In describing the phonatory functions of the larynx e.g. he developed a far more realistic membranous glottis model (cf. Figure 10) in contrast to the bagpipe mouthpiece that he used as sound generator in his speaking machine (cf. Figure 8, above left). Comparing the intermediate machine of Figure 7 – and the one at the "Deutsches Museum", Munich (cf. Figure 9) – with the one of the "Mechanism ..." one can also see that Kempelen discards additions that he could not handle correctly: One of these pieces is the small wire at the mouthpiece's tongue that eventually should control pitch variation.
Kempelen also makes suggestions how to construct a mechanical tongue (cf. Figure 11) instead of only changing the resonance characteristics by (partly) closing the rubber mouth or putting the fingers of his left hand inside. But he leaves it at this since he has problems with the audible burst for plosives then.

4. Kempelen and the theory of acoustic articulation

Kempelen didn't construct his speaking machine on the base of acoustic theories but went the engineering way of analysis-by-synthesis – or trial and error. He was mainly interested in the audible result that should be reached by a simple mechanism as close as possible to our articulatory apparatus on the one hand and playable like a musical instrument on the other.

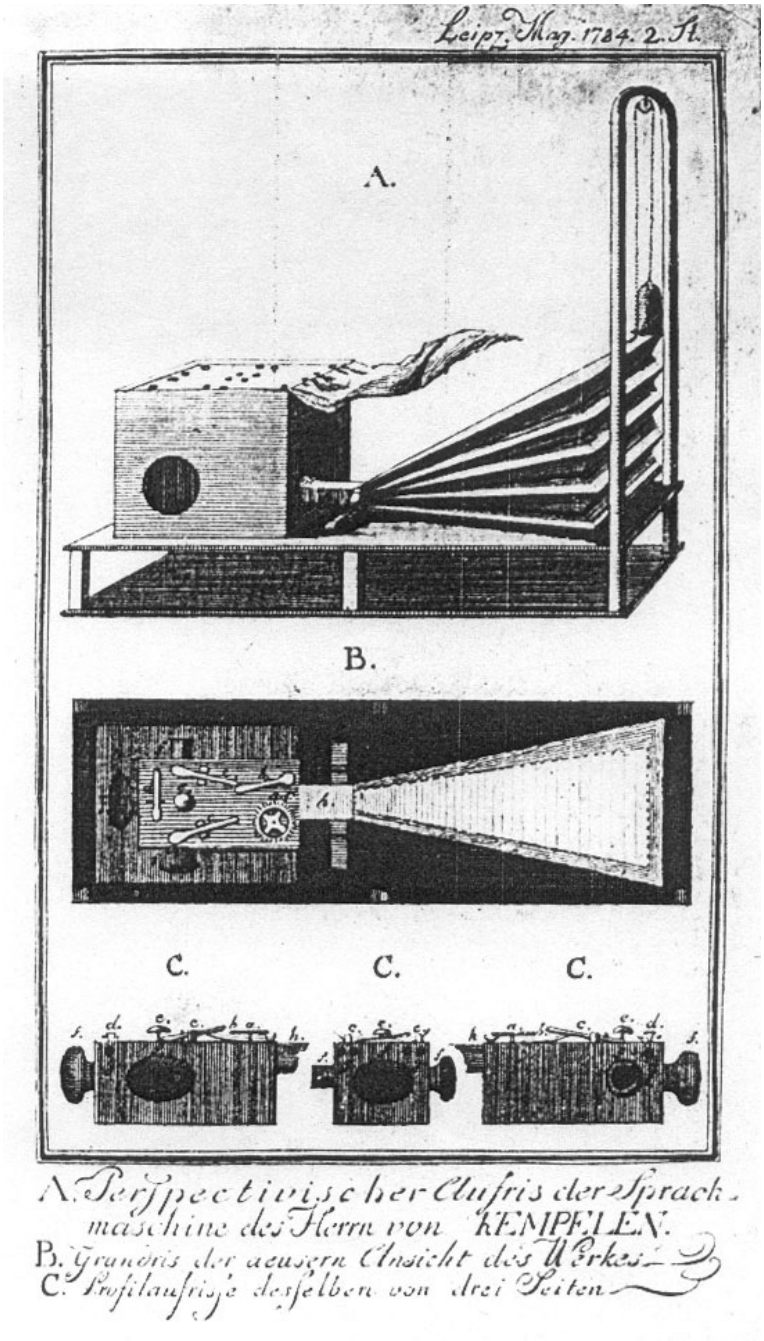


Figure 7: The first picture of the speaking machine (Hindenburg, 1784)

Kratzenstein, inspired by Euler on the other hand, tried to find his way into the nature of vowels through geometric-acoustic considerations based on reflections within elliptical cones (cf. Figure 12) although these were wrong and the tubes he finally used didn't resemble these constructions very much (cf. Figure 6).

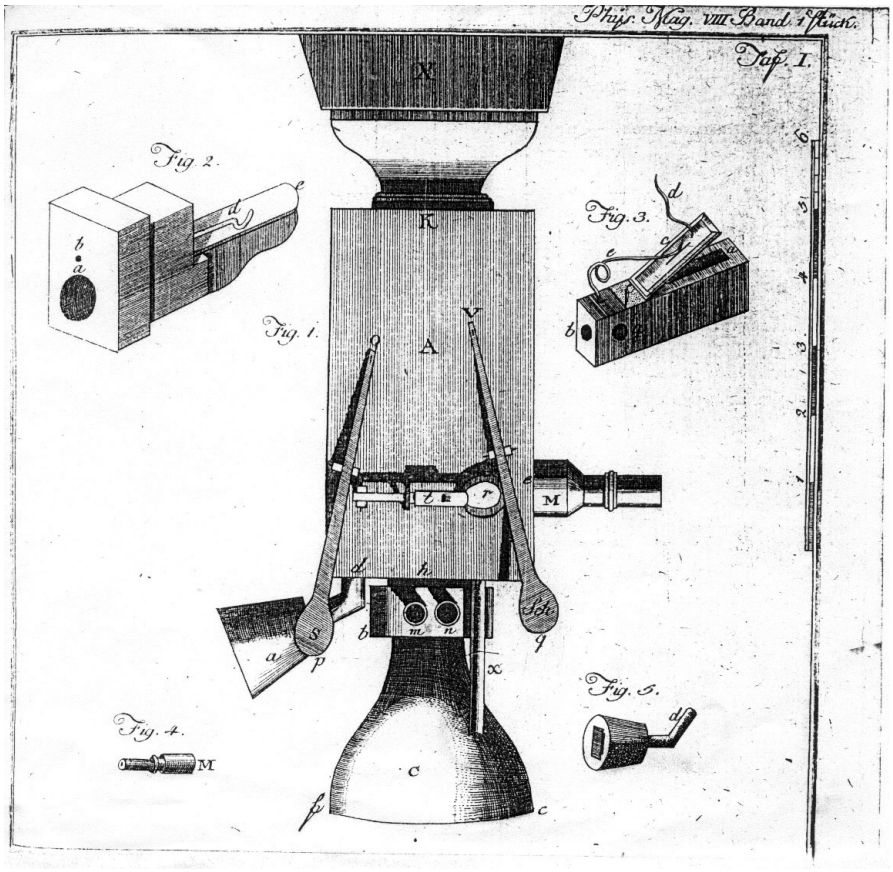


Figure 8: The machine of the "Mechanism ..."
(anonymous review of 1792)

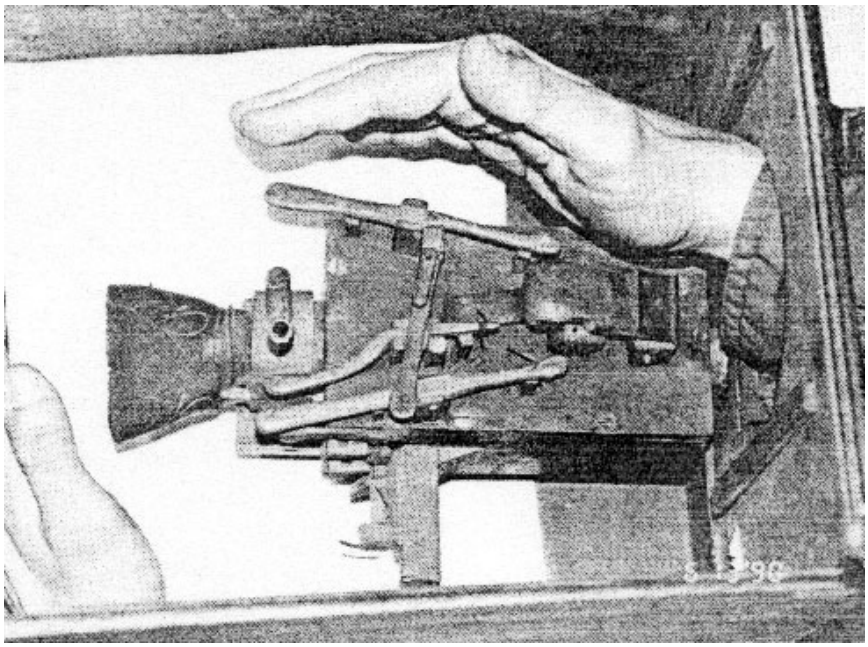


Figure 9: Kempelen's speaking machine at the "Deutsches Museum", Munich

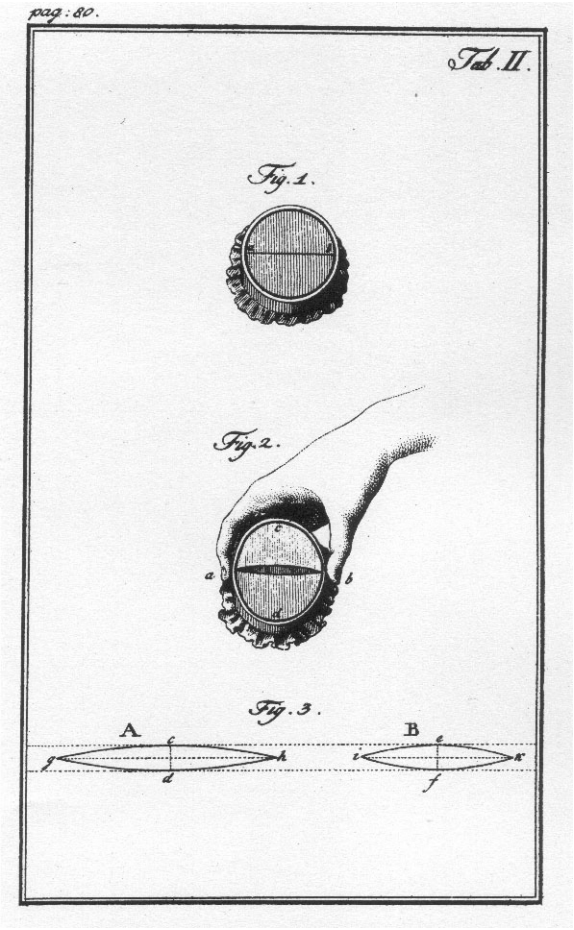


Figure 10: Kempelen's membranous glottis model

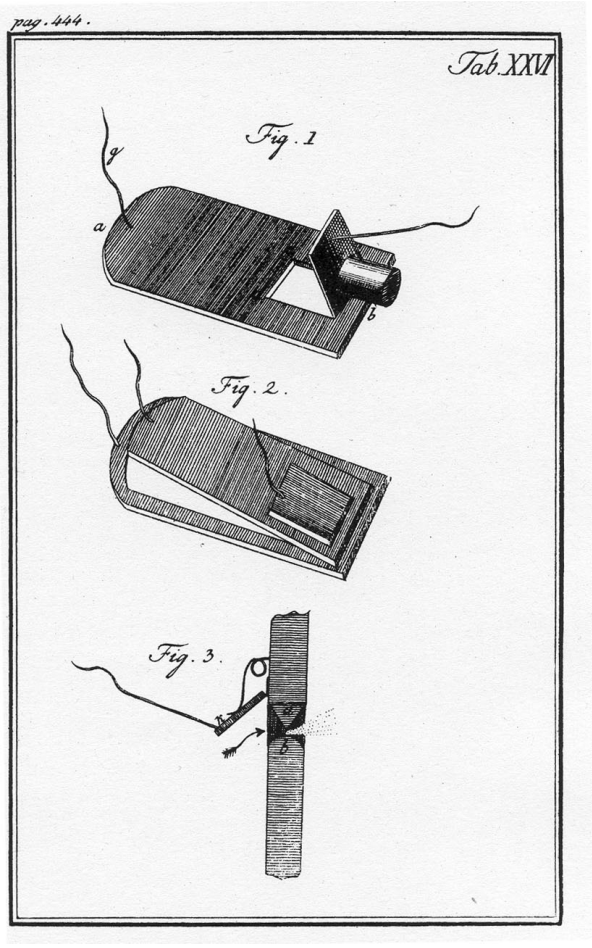


Figure 11: Kempelen's possible solution for a mechanical tongue for lingual stop production

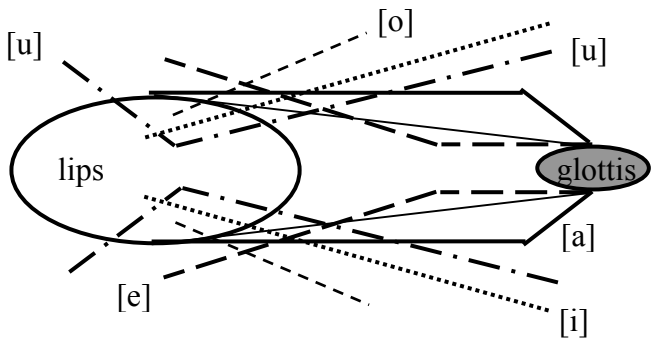


Figure 12: Kratzenstein's geometric-acoustic considerations based on reflections within elliptical cones (after Gessinger 1994)

Kempelen only once in his "Mechanism ..." gets deeper into vowel acoustics (cf. Figure 13).
He classifies the vowels according to the width of the lip channel giving a ranking of A > E > I > O > U and the width of the so called tongue channel that can be interpreted as horizontal tongue position. Kempelen goes on to remark that although he tried to produce the different vowels at the same pitch the vowel with a smaller tongue channel seemed to be higher in pitch. Although Kempelen isn't very explicit here, the observation clearly resembles the perceptual analysis of the second formant in whispered vowels described a century before by Reyher (1679; as cited in Kohler, 2000; cf. Figure 14) and the vowel tunes of von Helmholtz (1862; cf. Figure 15)

In 1830 it was Willis, starting from the ideas of Kratzenstein and von Kempelen, who first gained reasonable insight in the resonating properties of neutral tubes that would be able to give the illusion of different vowels (cf. Figure 16). In 1838 Wheatstone who also rebuilt Kempelen's machine added the theory of multiple resonance. During the 19th and part of the 20th century there existed allegedly contradictory theories on the nature of vowel sounds: On the one hand there was the harmonic theory stating that vowel frequencies have to be simple multiples of the fundamental frequency (Wheatstone, Helmholtz; Stumpf, Fant) and cavity tone theories (Willis, Hermann; Chiba & Kajiyama, Ungeheuer) that denied this. Today we know that harmonic analysis and resonance analysis are not real contradictions to one another but are merely two sides of the same coin. But a thorough theory of acoustic articulation (without simplifications) is still missing.
Kempelen's "Mechanism ..." is therefore a milestone in the history of phonetics, incorporating many insightful observations on articulatory mechanisms, whereas the speaking machine clearly a milestone in audio engineering.

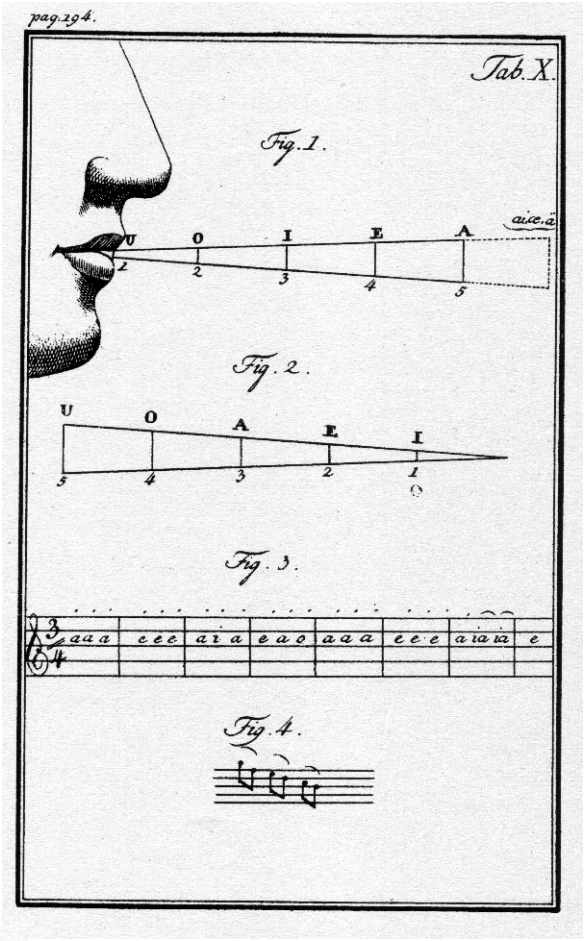


Figure 13: Kempelen's vowel categorisation according to the width of the tongue channel and the lip channel

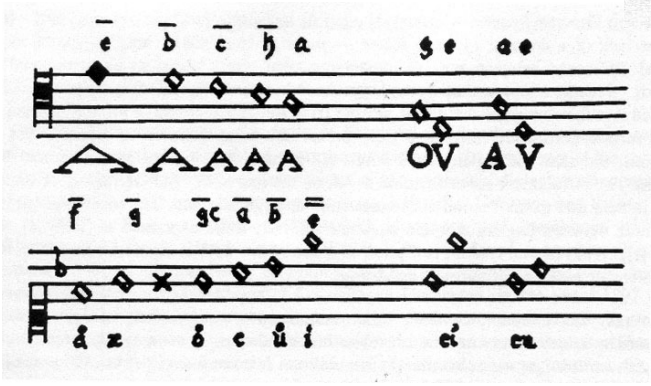


Figure 14: Whispered vowel tunes of Reyer (1679; after Kohler 2000)

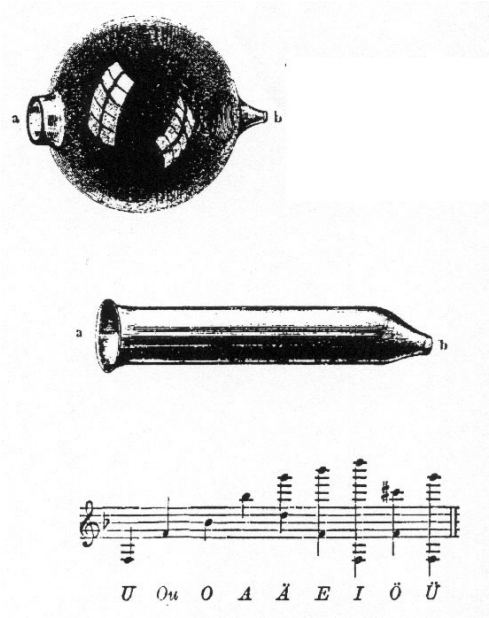


Figure 15: Vowel resonances after Helmholtz (1862)

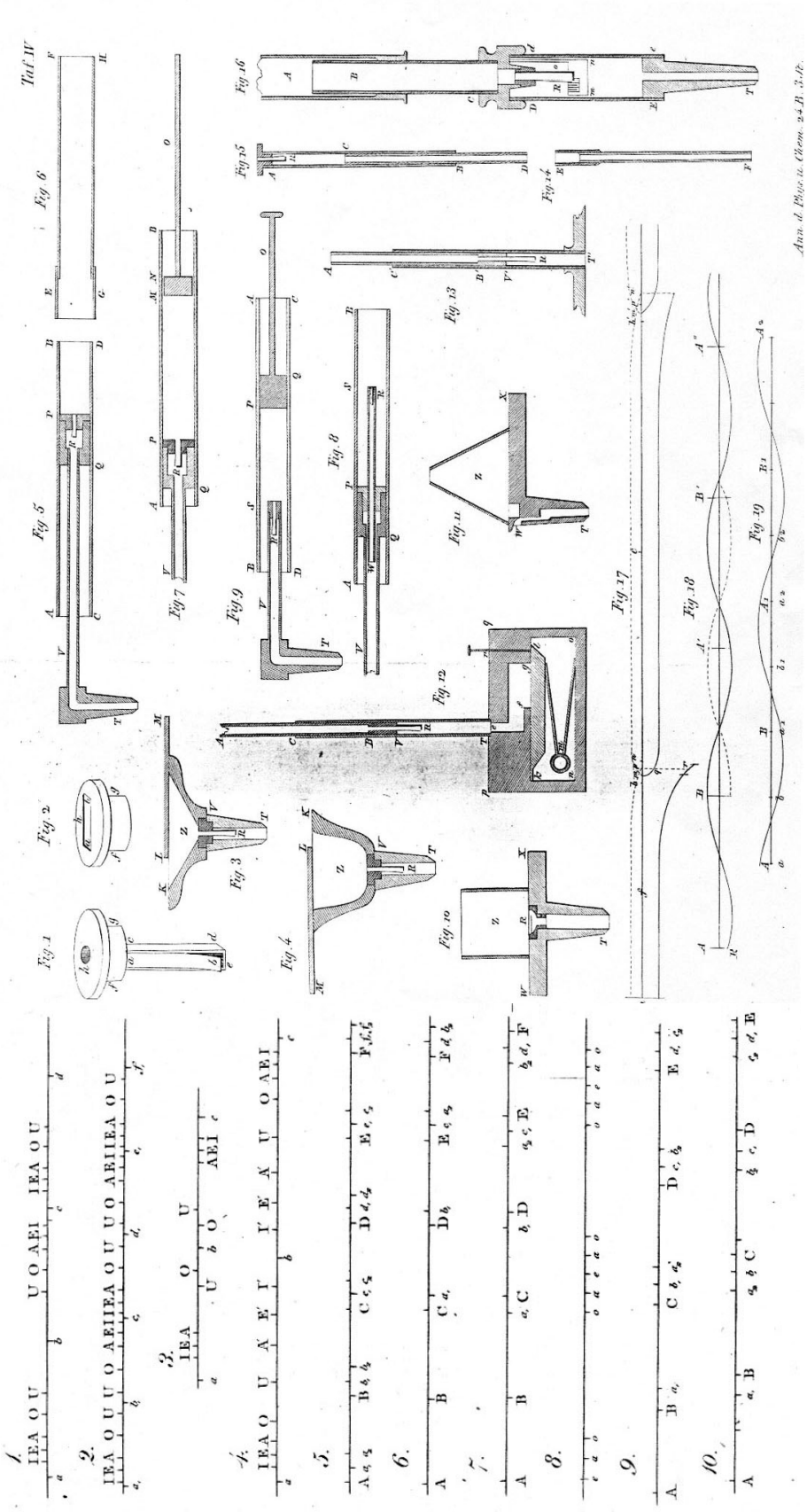


Figure 16: The experimental set-up of Willis (1832)

Acknowlegements

I wish to thank Gordon Ramsay for his considerable help on an earlier version of this paper.

References

[Anonymous] (1792). Nachricht von der Sprachmaschine des Herrn Hofr. von Kempelen. *Magazin für das Neueste aus der Physik und Naturgeschichte*. Bd. 8, 1. St., 84-102.

Brücke, E. (1856). *Grundzüge der Physiologie und Systematik der Sprachlaute für Linguisten und Taubstummenlehrer*. Wien.

Chiba, T. & Kajiyama, M. (1941). *The Vowel, its Nature and Structure*. Tokyo.

du Bois-Reymond, F. H. (1862). *Kadmus oder allgemeine Alphabetik vom physikalischen, physiologischen und graphischen Standpunkt*. Berlin.

Fant, C. G. M. (1961). *Acoustic Theory of Speech Production*. The Hague.

Gessinger, J. (1994). *Auge und Ohr. Studien zur Erforschung der Sprache am Menschen 1700-1850*. Berlin.

Helmholtz, H. v. (1862). *Die Lehre von den Tonempfindungen, als physiologische Grundlage für die Theorie der Musik*. Braunschweig.

Hermann, L. (1889 ff.). Phonophotographische Untersuchungen I-IV. *Archiv für die gesamte Physiologie des Menschen und der Thiere* 45 (1889) 582-592, 47 (1890) 44-53, 48 (1890) 347-391, 53 (1893) 1-51.

Hindenburg, C. F. (1784). *Ueber den Schachspieler des Herrn von Kempelen. Nebst einer Abbildung und Beschreibung seiner Sprachmaschine*. Leipzig.

Kempelen, W. v. (1791a). *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*. Wien.

Kempelen, W. v. (1791b). *Le Mécanisme de la parole, suivi de la déscription d'une machine parlante et enrichi de XXVII planches*. Vienne.

Kohler, K.J. (2000). The future of phonetics. *JIPA* 30, 1-24.

Kratzenstein, Ch. G. (1782). Sur la formation et la naissance des voyelles. *Journal de Physique* 21, 358-380.

Panconcelli-Calzia, G. (1940). *Quellenatlas zur Geschichte der Phonetik*. Hamburg.

Pompino-Marschall, B. (1991). Wolfgang von Kempelen und seine Sprechmaschine. *FIPKM* 29, 181-252.

Reyher, S. (1679). *Mathesis Mosaica, sive Loca Pentateuchi Mathematica Mathematice Explicata*. Kiel.

Stumpf, C. (1926). *Die Sprachlaute. Experimentell-phonetische Untersuchungen*. Berlin.

Ungeheuer, G. (1962). *Elemente einer akustischen Theorie der Vokalartikulation*. Berlin.

Wheatstone, C. (1838). Art. II. – 1. On the vowel sounds, and on Reed Organ Pipes. By Robert Willis [...] 2. Le Méchanisme de la Parole, suivi de la Description d'une

Machine Parlante. Par M. de Kempelen [...] 3. C.G. Kratzenstein. Tentamen Coronatum de Voce. [...] *The London and Westminster Review*, 27-41.

Willis, R. (1830). On vowel sounds, and on reed organ pipes. *Transactions of the Cambridge Philosophical Society* III, 231-276.

Willis, R. (1832). Ueber Vocaltöne und Zungenpfeifen. *Poggendorfs Annalen der Physik und Chemie*, 24(3), 397-437 + 1 plate.

Windisch, K. G. (1783a). *Briefe über den Schachspieler des Herrn von Kempelen*. Preßburg.

Windisch, K. G. (1783b). *Karl Gottlieb Windisch's Briefe über den Schachspieler von Kempelen nebst drey Kupferstichen die diese berühmte Maschine vorstellen*. Basel.

Windisch, K. G. (1783c). *Lettres de M. Charles Gottlieb Windisch sur le joueur d'echec de M. Kempelen*. Basel.

Windisch, K. G. (1784). Inanimate reason; or circumstantial account of that astonishing piece of work, M. de Kempelen's Chess-Player ... London.

A mechanical experimental setup to simulate vocal folds vibrations. Preliminary results

Nicolas Ruty
Annemie Van Hirtum
Xavier Pelorson
ICP, UMR 5009 CNRS/INPG/Université Stendhal Grenoble, France

Ines Lopez
Avraham Hirschberg
TU/e, Eindhoven, The Netherlands

This paper contributes to the understanding of vocal folds oscillation during phonation. In order to test theoretical models of phonation, a new experimental set-up using a deformable vocal folds replica is presented. The replica is shown to be able to produce self sustained oscillations under controlled experimental conditions. Therefore different parameters, such as those related to elasticity, to acoustical coupling or to the subglottal pressure can be quantitatively studied. In this work we focused on the oscillation fundamental frequency and the upstream pressure in order to start (on-set threshold) either end (off-set threshold) oscillations in presence of a downstream acoustical resonator. As an example, it is shown how this data can be used in order to test the theoretical predictions of a simple one-mass model.

1. Introduction

The interaction of expiratory airflow with the vocal folds tissues is known to be the primary source of human voiced sound production (Titze, 1977; Fant, 1980). The air-flow through the larynx induces instability of the vocal folds. The resulting vocal fold vibrations modulate the airflow giving rise to a periodic sequence of pressure pulses which propagates through the vocal tract and is radiated as voiced sound. Modelling of the ongoing fluid-structure interaction and the vocal fold oscillations is important in the understanding of phonation (Fant, 1982a; Fant, 1982b; Ikeda et al., 2001), the synthesizing of voiced sound

(Ishizaka et al., 1972; Koizumi, 1987; Story, 1995), and the study of voice disorders. Physical modelling of the vocal folds and the 3D fluid-structure interaction between the living tissues and the airflow has a long and rich history. Exact solutions for the flow through the glottis, as well as the flow-induced vocal fold deformation, are impossible to derive analytically. Full numerical simulation is still limited to over-simplified glottal configurations or flow conditions and is therefore not of practical use for many applications such as speech synthesis. Simplified models, such as distributed models (Ishizaka et al, 1972; Liljencrants, 1991; Kob, 2001), sometimes described as "caricatures", are therefore of interest for many practical applications. Concerning these simple models, two major problems can be formulated:

(1) how accurate are these simplified theories ?
(2) how can they be related to human physiology ?

One way to answer the first question (1) is to test the theoretical models against physical measurements performed on mechanical models of the vocal folds. Since the pioneer work of van den Berg (1957) many research groups in the world have developed mechanical replicas for this purpose and with an increasing complexity (and expected realism) Scherer et al., 1983; Pelorson et al., 1996; Gauffin, 1988. At present time, the most realistic set-ups are certainly those simulating the vibrations of the vocal folds (Barney, 1995; Zhao et al., 2002; Zhang et al., 2002). In this paper we present a new set-up capable of generating self-sustained oscillations of an elastic replica of the vocal folds. This mechanical model is interesting because it allows to generate and to measure unsteady flows at a rate comparable to those encountered during speech (50Hz <F0<1kHz). Of great interest for many applications, parameters associated to the onset and offset of vibration can also be measured. In addition, since the replica is producing sound, the acoustics can also be addressed using this setup as it will be shown in this paper.
Lastly, we will show how the second problem (2) can be addressed and tested using this experimental setup. It will be shown how the mechanical parameters of a simple one-mass model can be determined by direct observations performed on the replica.

2. Experimental setup

2.1. Description

The experimental set up is illustrated in figure 1. Up-scaled by a factor 3, it represents the human phonatory system composed of several elements. A pressure reservoir (approximate volume 0.75m³) represents the lungs and forces an airflow through a deformable vocal fold replica. Airflow is supplied by a

compressor, which controls the pressure in the reservoir up to 50000Pa. The air pressure at the entrance of the reservoir is controlled thanks to a Norgren precision pressure regulator. The walls of the reservoir are covered with absorbing foam, in order to reduce its acoustical resonance.

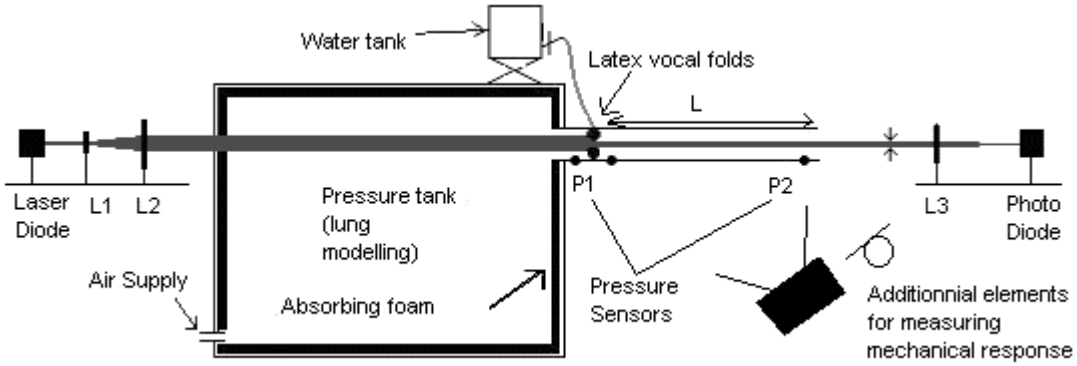


Figure 1: Schematic overview of the experimental set up used to produce self sustained oscillations and to detect the upstream pressure threshold necessary to obtain these oscillations. The water tank is used to control the internal pressure of the vocal folds replica.

The replica shown in Figure 2, is made of two latex tubes (Piercan Ltd) of 1.1cm diameter and 0.2mm latex thickness. Precision is on 10 percent on thickness and 1 percent on diameter. The tubes are mounted on a metal cylinder of 1cm diameter of which half of the diameter has been removed over a length of 3cm.

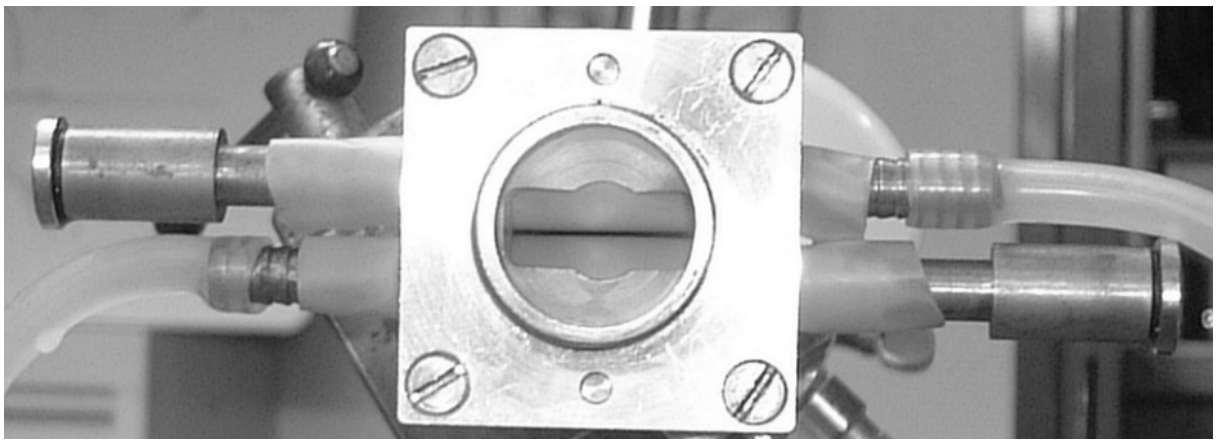


Figure 2: Vocal fold replica, made of two latex tubes, filled with water.

A central duct of 2mm diameter along the axis of the cylinder allows to fill the replica with water of which pressure (henceforth called the internal pressure Pc) is controlled by a water column. Changing the internal pressure Pc involves different types of changes. Firstly, because increasing the internal pressure has

the consequence to fill the replica with more water, the mass of the replica increases. Secondly, the initial deformation of the replica is changed too. When internal pressure Pc is low (e.g. Pc=4000Pa), the replica remains well rounded as suggested in figure 3 whereas when the internal pressure is increased, the vocal fold replica are contacting which involves a more complex glottal geometry. Lastly, of course, increasing the internal pressure decreases the elasticity of the vocal fold replica. This vocal fold replica is connected with the reservoir directly.

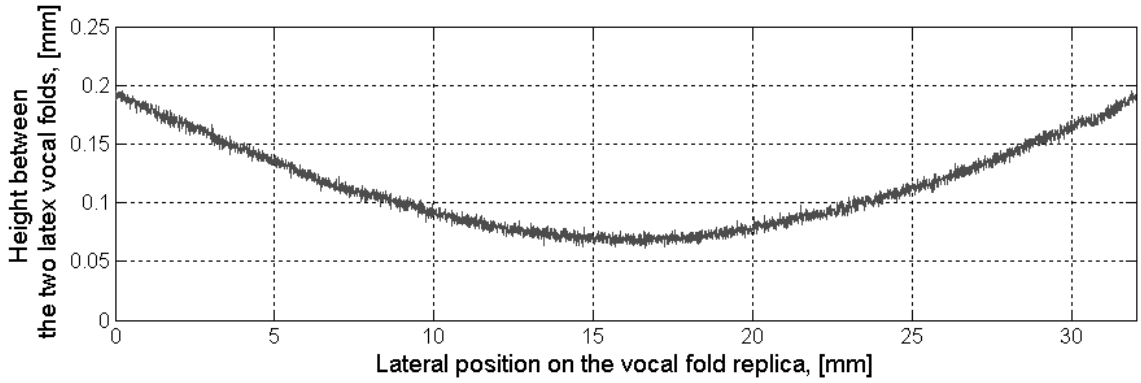


Figure 3: Frontal view of the deformable vocal fold replica, realised displacing laterally the laser beam and the photo sensor. Internal pressure Pc = 4000Pa

A downstream resonance pipe, simulating a vocal tract, of varying length L is also connected downstream of the glottal replica. Its section is constant, of 2.5cm diameter. Two Kulite pressure sensors XCS-093-0.35-Bar-G (P1, P2), supplied by a Labor-Netzgerät power supply EA-3005S, enable us to measure upstream pressure and downstream pressure respectively. The pressure sensors were calibrated against a watermeter with typical accuracy of +/- 5Pa.
Thanks to an optical laser system, the replica opening h, i.e. the vertical 1D movement of the deformable replica, can be recorded. This optical system consists in a laser diode (supplied by a P. Fontaine Dc amplifier FTN2515) which laser beam is increased by two convergent lenses (L1, f=50mm; L2, f=100mm), at a distance of 125mm. Then, the laser beam goes through the reservoir, and further through the vocal folds replica. When oscillating, the glottal replica is therefore interrupting the laser beam. The corresponding intensity fluctuations are recorded by a photo sensor (supplied by a Solartron DC Power Supply).

2.2. Data acquisition

This system is linked to an acquisition chain which enables us to detect the presence of oscillation as well as the oscillation frequency from the registered changes in replica aperture.
This data acquisition system consists in a filtering preamplifier (National Instruments SXCI-1121) in which all sensor signals are filtered and amplified between +10V and -10V. The outputs of this filter are plugged onto a National Instruments BNC2080 card linked to an acquisition card National Instruments PCI-MIO-16XE. Lastly the acquired data are processed in Labview7.

2.3. Calibration

Pressure sensors, optical system and measurement microphone have to be calibrated.
Pressure sensors have a linear response. We have to determine their gain. Airflow is forced in a rigid tube. Pressure is simultaneously measured with a Kimo pressure gauge and with a pressure sensor. For each value of pressure ranging between 50 and 1000Pa, the real pressure values measured from the pressure gauge are related to the electrical voltage measured by the pressure sensors. From the linear experimental curve the gain of each pressure sensor is estimated.
The optical system is used to measure the vocal folds replica opening. The intensity of the light receives by the photo sensors have to be related to the opening of the obstacle the laser beam crosses. For the calibration the replica is replaced by a rigid rectangular aperture of a height ranging between 0.05mm and 3mm. Each value of the aperture is related to the voltage measured by the photo sensors.
Lastly the Bruël and Kjaer measurement microphone is calibrated with a Bruël and Kjaer Sound Level Calibrator Type 4230 which produce a pure sinusoidal audio signal of 1000Hz frequency and 94dB amplitude.

3. Results

3.1. Oscillation threshold

The set-up illustrated in Figure 1 and 2 is able to produce self-sustained oscillations. In the experiments detailed here, a downstream pipe of length L=49cm was used. The internal pressure of the vocal folds was ranging between 3500 and 6500Pa. In order to get this range of variation, the height of the water column that fills the replica was ranging between 35cm and 65cm.

For a given internal pressure, the upstream pressure is increased until oscillations appear. Oscillations are detected by spectral analysis of the pressure signal recorded by the sensor P2. The presence of oscillations is validated as soon as the spectrum is composed of at least a fundamental frequency and two superior order harmonics. Hence we obtain the mechanical experimental phonation on-set threshold.
When stable oscillations are achieved, the fundamental frequency of oscillation is recorded and the upstream pressure is decreased. The same procedure is followed to retrieve the transition from self-sustained oscillation to no-oscillation state. The upper upstream pressure for which oscillations disappear corresponds to the phonation off-set threshold.

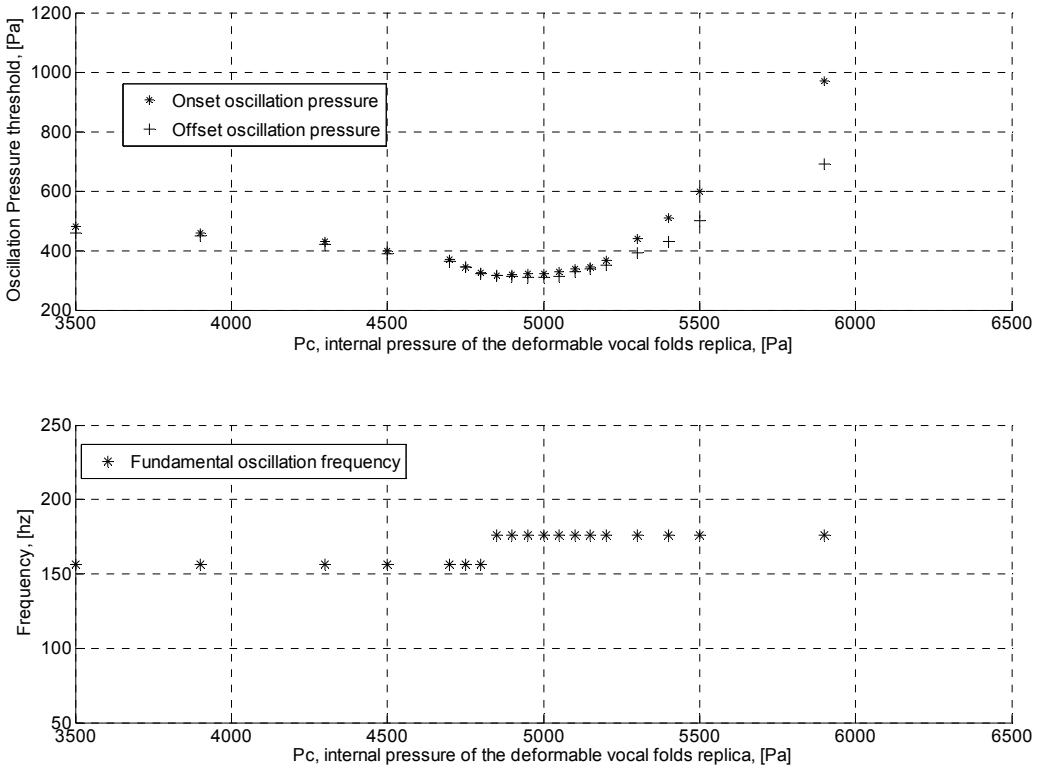


Figure 4: (a), Upstream pressure oscillation threshold, as a function of the internal deformable replica pressure. (b), fundamental frequency of the oscillation.

Figure 4 shows an example of the results obtained using this procedure. A lot of observations can be made. Firstly, the range of values we obtain for the threshold is comparable to the human phonation threshold, with an average value of 400Pa (Baken, 1987). Secondly, we notice that a hysteresis phenomenon is present when we compare the on-set and the off-set pressure as noted by Lucero, 1999. Furthermore a minimum threshold pressure is reached at Pc=5000Pa. That confirms further studies considering the oscillation threshold (Chan et al, 1997; Titze et al., 1995). More precisely, we observed that this

minimum is reached when the vocal folds replica is almost closed at rest position (Figure 5). This behaviour can be related to the optimum configuration for the ease of phonation discussed by (Lucero, 1998).

3.2. Oscillation frequency

As illustrated in Figure 4, the fundamental frequency of oscillations observed experimentally is close to the downstream pipe acoustical resonance. Indeed, the resonance frequency of the pipe is equal to 173.5Hz and what we observe in Figure 4 is fundamental frequency ranging between 155Hz and 180Hz. Fundamental frequencies seem to be constrained between a short range of values. We can note that further measurements have to be done. Indeed, we observe only two values for fundamental frequencies. It is in fact not the case, but for real time measurements, we have to reduce the FFT scale and loose in precision for the measurement of fundamental frequency. What we obtain is an order of range for the fundamental frequency of the oscillations (with a precision of 10Hz approximately).
Then, concerning this range, we observe that such values are between male and female phonation fundamental frequencies.

3.3. Influence of initial parameters

In this section, we examine the influence of the internal pressure, Pc as a control parameter for the deformable vocal folds replica. Of particular interest is its effect on the equilibrium position of the vocal folds and on their mechanical characteristics. The consequences of these changes can be observed in terms of oscillation thresholds and oscillation frequencies since changing the internal pressure, Pc affects the mechanical characteristics. Therefore the equilibrium aperture of the replica in absence of upstream and downstream resonators as well as the mechanical response of the deformable replica is measured. Measurement results are detailed below.

3.3.1. Equilibrium positions

The procedure used to determine the equilibrium positions is as follows. For each given internal pressure, Pc (ranging between 1500Pa and 6500Pa), the upstream pressure is increased from 50Pa to 1000Pa by steps of 50Pa. Thanks to the optical laser system, the aperture between the two latex tubes can be measured. The results are presented in Figure 5.

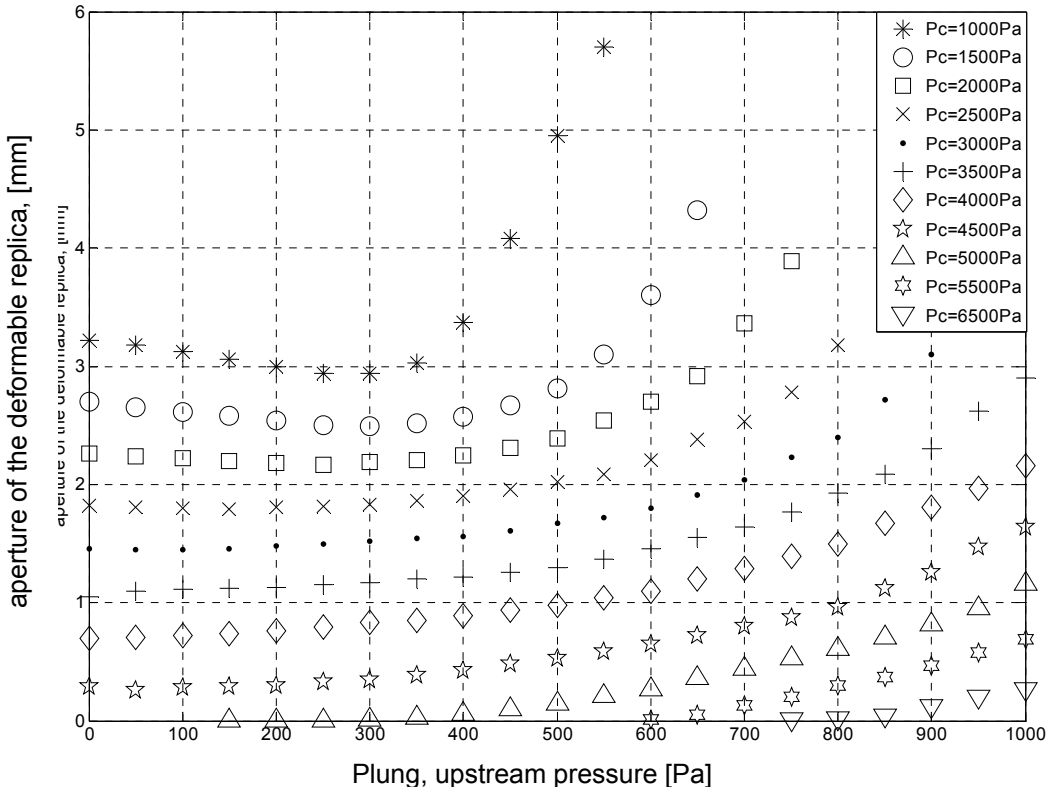


Figure 5: Equilibrium position of the deformable vocal fold replica measured as a function of the upstream pressure (Plung), for different values of the internal pressure (Pc) in the deformable replica

From this figure some interesting behaviour can be observed. For low internal pressures, namely lower than 3000Pa, increasing the upstream pressure at first decreases the replicas aperture until a minimum opening is reached; further pressure raise increases the equilibrium aperture. When the internal pressure Pc is higher than 3000Pa, increasing the upstream pressure results in forcing the vocal folds replica to open.

Concerning the position at rest, without upstream pressure, we observe that increasing internal pressure involves a decrease of the aperture until the tubes are in contact for internal pressures up to 5000Pa. 5000Pa appears to be a critical value for the internal pressure. This corresponds to the frontier between an opened rest position and a condition where the vocal folds are contacting at rest. It also corresponds to the point where oscillation pressure threshold is minimal, as described in section 3.1.

We can conclude from this measurement that there is a strong correlation between the oscillations threshold and the equilibrium position.

3.3.2. Mechanical response

In addition to that, we use a measurement microphone Bruël and Kjaer type 4192 with its preamplification unit Bruël and Kjaer and a loud speaker driven by an Onkyo Integra Stereo Amp.
This second measurement consists in the determination of the mechanical response of the deformable vocal folds replica from which we extract the resonance frequency and a quality factor.
The experimental protocol is the following. In addition to the experimental set-up depicted in Figure 1, a loud speaker and a Bruël and Kjaer (type 4192) measurement microphone were added in the vicinity of the vocal folds replica. Using the loudspeaker the replica was excited using a sinusoidal signal with a frequency varying between 100Hz and 400Hz. The varying maximal opening h is captured during the vertical movements of the replica by way of the photo-diode through the acquisition chain. Hence, after filtering, the needed parameters (resonance frequencies, quality factor and equilibrium position) are derived. This filtering consists in a deconvolution of the displacement, by the microphone signal.
An example of mechanical response measured for internal pressures Pc varying from 1500 to 6500Pa is presented in Figure 6. Resonance frequencies are detected if a peak is more important (higher than 10dB) in terms of its amplitude, than the mean signal. We focus our attention on the peak close to the one observed during the oscillations (see 2.1). The quality factor is calculated using the following formula:

$$Q = \frac{f_{res}}{\Delta f_{-3dB}}, \quad (1)$$

where f_{res} is the resonance frequency and Δf_{-3dB} is the associated 3dB bandwidth.

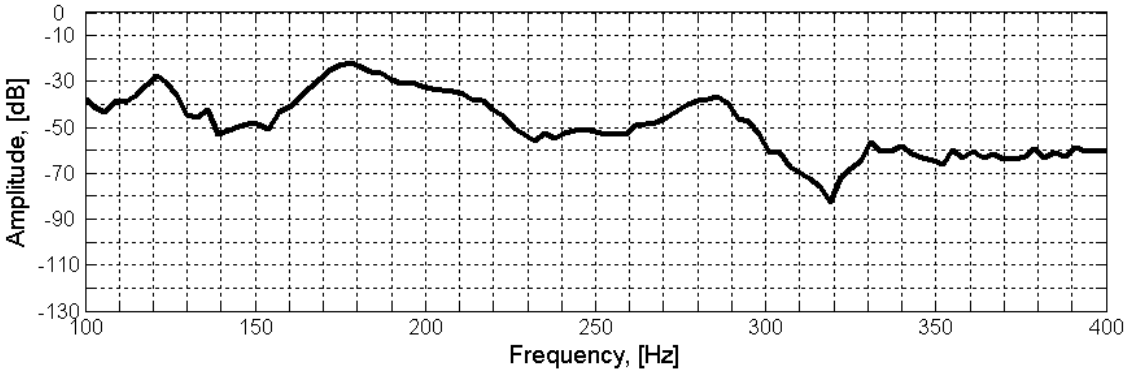


Figure 6: Mechanical response of the deformable vocal folds replica, for an internal pressure of 6000 Pa.

The experimental parameters obtained from these experiments are tabulated in Table 1. We observe that resonance frequency is close to the one measured when the replica self-oscillates. However, increasing internal pressure involves an increase in resonance frequency.

Table 1: Recapitulative table of the different experimental parameter values used for the numerical simulation.

Internal Pressure Pc(Pa)	**3500**	**4000**	**4500**	**5000**	**5500**	**6000**	**6500**
Resonance frequency (Hz)	123	124	130	140	166	169	175
Quality Factor	8.2	8.267	8.3	10	5.5	14.08	17.5

This effect is to be related again to a variation of the stiffness of the replica. High internal pressure involves a higher vocal folds replica stiffness and thus higher resonance frequency as observed in real life on human speakers.
Another observation is that the resonance frequencies are varying between 123Hz and 175Hz while when self-sustained oscillations where present frequencies were observed in a smaller range (between 155Hz and 180Hz). This tends to illustrate the effect of the downstream resonator on the oscillation frequencies of the replica.
In the absence of extensive data concerning the effect of the subglottal system on human phonation, the experimental study of the influence of an upstream pipe in the mechanical setup is currently omitted since it is difficult to conclude about the relevance.

4. Using the measured data in physical models

As an example of application we present in the following an attempt to use the set-up for testing a physical model of human phonation. For the sake of simplicity, the model chosen will be based on a simple one mass model of the vocal folds. This model is depicted in figure 7.

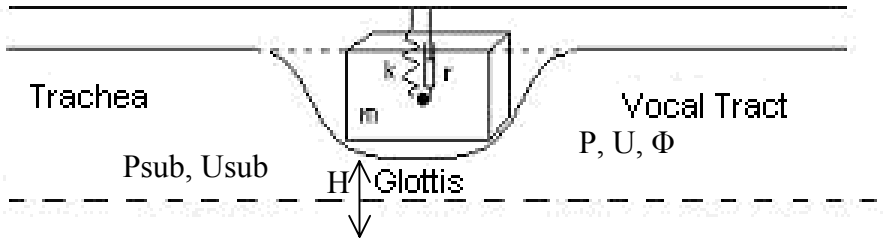


Figure 7: One spring mass model of the vocal folds, with k, r and m respectively the stiffness, the damping and the mass of a vocal fold. H is the displacement, between the two vocal folds, Psub and Usub respectively the upstream pressure and flow velocity. Φ is the volume flow, P and U respectively the downstream pressure and flow velocity.

Assuming small variations around the equilibrium position, this system, coupled with a downstream resonator (a uniform section pipe), can be described with the following set of equations:

$$\frac{d^2h}{dt^2} + \frac{\omega_l}{Q_l}\frac{dh}{dt} + \omega_l^2 h = -\frac{1}{\mu}p \qquad (1)$$

$$\phi = b\left(-\frac{\bar{h}\,p}{\rho\,\bar{u}} + \bar{u}\,\bar{h}\right), \bar{u} = \sqrt{\frac{2\,\bar{p}_{sub}}{\rho}} \qquad (2)$$

$$\frac{d^2\psi(t)}{dt^2} + \frac{\omega_a}{Q_a}\frac{d\psi(t)}{dt} + \omega_a^2\psi = \frac{Z_a\omega_a}{Q_a}\phi \qquad (3)$$

(1) is the mechanical equation describing the elasticity behaviour of the vocal folds. h is the variation of the glottal opening around the equilibrium position $\bar{h}$. p is the variation of transglottal pressure. ω_l is the resonance pulsation. $\omega_l = \sqrt{k/m}$, k is the stiffness of the spring, m the mass. Q_l is the quality factor associated to this resonance. $Q_l = \omega_l / r$, r is the damping of the one mass model.

(2) is the fluid mechanical equation describing the behaviour of the airflow through the glottis. Φ is the volume flow. b is the glottal width, ρ is air density, $\bar{u}$ is the mean airflow velocity, $\bar{p}_{sub}$ is the mean subglottal (or upstream pressure)

(3) is the acoustical equation describing the propagation of acoustic waves inside the vocal tract, approximated here as a uniform pipe. $p = d\psi / dt$, where ψ is the acoustical eigenfunction. ω_a, Q_a, Z_a are the resonance

pulsation, the quality factor and the static impedance of the pipe respectively.

While geometrical or acoustical parameters in the above equations can be determined directly, the same cannot be said about the mechanical characteristics in equation (1). Therefore, for each internal pressure, Pc, the values attributed to the parameters Q_l and ω_l were obtained in order to fit the experimental mechanical response (see Table 1).
Then for subglottal pressure $\overline{p}_{sub}$ ranging between 20 and 1000 Pa a linear stability analysis method, as proposed by Cullen et al. (2000), was performed.
Hence the results of simulation can be compared to the experimental ones as shown in figure 8.

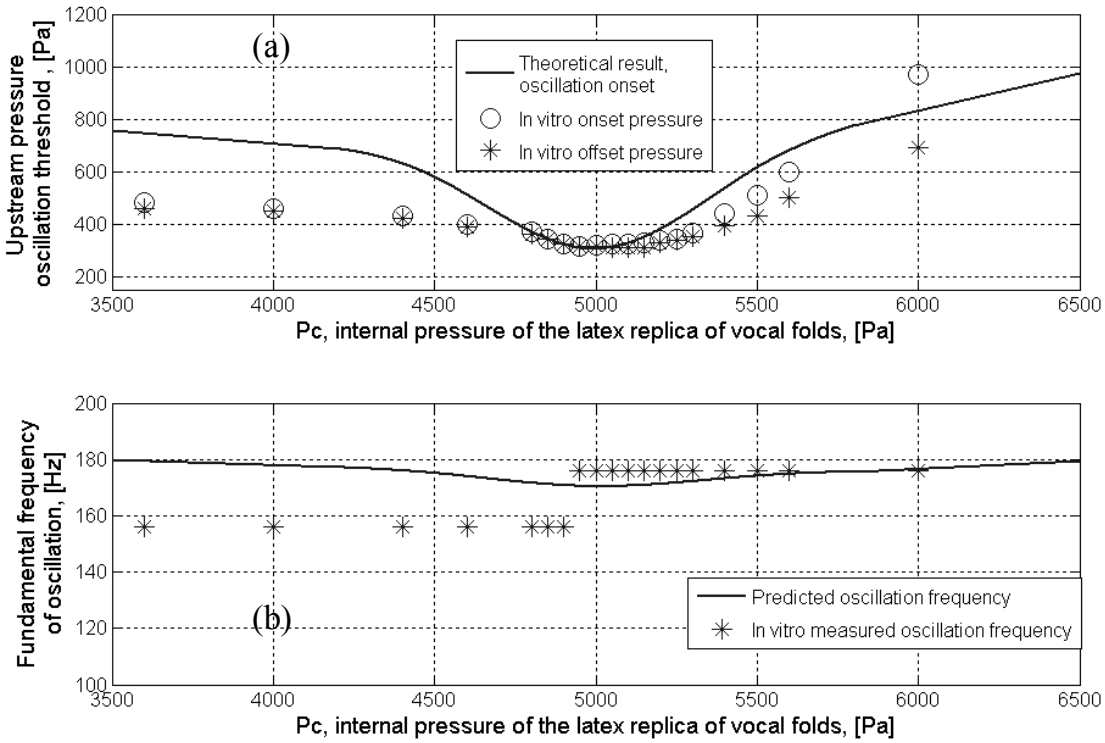


Figure 8: (a), Upstream pressure oscillation threshold, comparison of experimental results and theoretical prediction. (b), Comparison of experimental results and prediction for the fundamental frequency of the oscillation.

On the same graph (figure 8a), theoretical and experimental threshold are presented showing that even such a crude model can explain qualitatively the experimental results. Moreover, fundamental frequencies are also theoretically predicted and compared to the experimentally measured ones as shown in figure 8b. Once again, theory appears to be close to experiments. The same phenomena are observed on the two curves, i.e. the presence of a minimum threshold for an

internal pressure of 5000Pa and a fundamental oscillation frequency close to the downstream pipe (vocal tract) resonance frequency.

5. Conclusion

In this paper we have presented a new experimental set-up designed in order to test physical models for human voice production. This set-up presents a self-oscillating vocal folds replica made of latex filled with water which is coupled to acoustical resonators (representing the subglottal or the supraglottal tract). This allows to measure aerodynamical (the coupling between the elastic structure and the flow) and aeroacoustical (the coupling between the acoustic field and the flow) phenomena at the same time, which, to the best of our knowledge, has never been done using a mechanical replica. Further, compared with existing set-ups this one has the advantage to generate the self-sustained oscillations of a replica whose mechanical characteristics (controlled using the internal pressure, Pc) can be controlled quantitatively. This is particularly important when willing to compare to low order mechanical models such as the one- or the two-mass model.
The first results obtained with this replica tend to be encouraging. Oscillation threshold pressures were found quite comparable to those observed in human speech and with comparable frequencies, at least for healthy male speakers. Even more interesting is the experimental evidence for a minimum threshold pressure which was already theoretically predicted by Lucero (1998).
Of course, the greatest care should be taken in the interpretation of the measured data. This set-up has been built in order to test the theoretical models, as shown for the one-mass model example here, rather than to mimic reality.
Further study is therefore ongoing, testing systematically different approximations for the mechanical behaviour of the vocal folds as well as the effect of acoustical coupling. The set-up itself is also under study, in particular in order to generate a broader quantity of configurations such as the presence of asymmetrical vocal folds, high frequency fundamental frequency of oscillation, etc …

Acknowledgements

This work has been supported by the Region Rhône-Alpes (project Emmergence) a PhD grant from the French Minister of Research and Education. We would like to acknowledge the skills of Pierre Chardon who helped us to design and to build the set-up. Lastly, we wish to thank Susanne Fuchs and an anonymous reviewer for useful remarks.

References

Avanzini F., Van Walstijn M. (2004). Modelling the mechanical response of the reed-mouthpiece-lip system of a clarinet. Part 1. A one-dimensional distributed model. *Acta Acustica United with Acustica, 90:* 537-47.

Baken R.J. (1987). Clinical Measurement of Speech and Voice. *Ed. Allyn and Bacon.*

Barney A. (1995). Fluid Flow in a dynamic mechanical model of the larynx. Thesis submitted for the degree of Doctor of Philosophy. *Department of Electronics and Computer Science. Faculty of Engineering an Applied Science*. University of Southampton.

Chan R.W., Titze I.R., Titze M.R. (1997). Further studies of phonation threshold pressure in a physical model of vocal fold mucosa. *J. Acoust. Soc. Am. 101 (6)*: 3722-27.

Cullen J.S., Gibert J., Campbell D.M. (2000). Brass instruments: linear stability analysis and experiments with an artificial mouth. *Acta Acustica, 86:* 704-24.

Fant G. (1982)a. Preliminaries to analysis of the human voice source. *STL-QPSR 4:* 1-27.

Fant G. (1980). Speech production. A voice source dynamics. *STL-QPSR 2-3:* 17-37.

Fant G. (1982)b. The voice source – Acoustic modelling. *STL-QPSR 4.*: 28-48.

Gauffin J., Liljencrants J. (1988). Modelling the air flow in the glottis. *Annual Bulletin RILP 22*: 41-52.

Ikeda T., Matsuzaki Y., Aomatsu T. (2001). A numerical analysis of phonation using a two dimensional flexible channel model of the vocal folds. *Journal of Biomechanical Engineering, Vol. 123*: 571-79.

Ishizaka K. (1985). Air resistance and intra-glottal pressure in a model of the larynx. *Vocal Fold Physiology: Biomechanics, Acoustic and Phonatory Control. Ed.: Titze I.R, Scherer R.C.:* 414-24.

Ishizaka K., Flanagan J.L. (1972). Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords. *Bell Syst. Tech. Journal 51*. 6, pp 1233-67.

Ishizaka K., Flanagan J.L. (1977). Acoustic properties of longitudinal displacement in vocal cord vibration. *The Bell Syst. Technical Journal, Vol. 56, No. 6*: 889-918.

Liljencrants J. (1991). A translating and rotating mass model of the vocal folds. *STL-QPSR 1*. pp 1-18.

Kob M. (2002). Physical modeling of the singing voice. *Berlin : Logos-Verl.*

Koizumi T., Taniguchi S., Hiromitsu S. (1987). Two-mass model of the vocal cords for natural sounding voice synthesis. *J. Acoust. Soc. Am. 82 (4)*: 1179-92.

Lopez I., Schellekens M.H., Driessen N.M., Hirschberg A., Van Hirtum A., Pelorson X. (2004). Buzzing lips and vocal folds: the effect of acoustical feedback. *Flow Induced Vibration, Ed.: de Langre, Axisa.*

Lucero J.C. (1999). A theoretical study of the hysteresis phenomenon at vocal fold oscillation onset-offset. *J. Acoust. Soc. Am. 105 (1)*: 423-31.

Lucero J.C. (1998). Optimal glottal configuration for ease of phonation. *J. Voice 12(2):* 151-58.

Pelorson X., Hirschberg A., Van Hassel R.R., Wijnands A.P.J., Auregan Y. (1996). Theoretical and experimental study of quasisteady-flow separation within the glottis during the phonation. Application to a modified two-mass model. *J. Acoust. Soc. Am. 96 (6)*: 3416-31.

Pelorson X., Hirschberg A., Wijnands A.P.J., Baillet H. (1995). Description of the flow through in-vitro models of the glottis during phonation. *Acta acustica 3*: 191-202.

Scherer R.C. (1991). Physiology of Phonation: A review of basic mechanics. *Phonosurgery: Assessment and surgical management of voice disorders. Raven Press:* 77-93.

Scherer R.C. (1983). Pressure-flow relationships in a laryngeal model having a diverging glottal duct. *Acoustical Society of America meeting, Cincinnati, Ohio.*

Scherer R.C., Titze I.R., Curtis J.F. (1983). Pressure-flow relationships in two models of the larynx having rectangular glottal shapes. *J. Acoust. Soc. Am. 73 (2)*: 668-76.

Story B.H., Titze I.R. (1995). Voice simulation with body cover model of the vocal folds. *J. Acoust. Soc. Am. 97 (2)*: 1249-60.

Titze I.R. (1977). On the mechanics of vocal-fold vibration. *J. Acoust. Soc. Am. 60 (6):* 1366-80.

Titze I.R. Schmidt S.S., Titze M.R. (1995). Phonation Threshold pressure in a physical model of the vocal fold mucosa. . *J. Acoust. Soc. Am. 97 (5)*: 3080-84.

Van Den Berg Jw., Zantema J.T., Doornenbal P. (1957). On the air resistance and the Bernoulli effect of the human larynx. *J. Acoust. Soc. Am. 29 (5):* 625-31.

Vilain C.E., Pelorson X., Hirschberg A., Le Marrec L., Op't Root W., Willems J. (2003). Contribution to the physical modeling if the lips. Influence of the mechanical boundary conditions. *Acta Acustica United with Acustica, Vol.89*: 882-87.

Warren D.W. (1976). Aerodynamics of speech production. *Contemporary Issues in Experimental Sciences*, *Academic Press, New-York*: 105-137.

Zhao, W., Zhang, C., Frankel, S.H., and Mongeau, L. (2002). "Computational Aeroacoustics of Phonation, Part I: Numerical Methods, Acoustic Analogy Validation, and Effects of Glottal Geometry," *J. Acoust. Soc. Am.*, Vol. 112, No. 5, pp. 2134-2146.

Zhang, C., Zhao, W., Frankel, S.H., and Mongeau, L. (2002). "Computational Aerocoustics of Phonation, Part II: Effects of Subglottal Pressure, Glottal Oscillation Frequency, and Ventricular Folds," *J. Acoust. Soc. Am.*, Vol. 112, No. 5, pp. 2147-2154.

On the invariance of speech percepts

Willy Serniclaes
CNRS & Université René Descartes, Paris 5

A fundamental question in the study of speech is about the invariance of the ultimate percepts, or features. The present paper gives an overview of the non-invariance problem and offers some hints towards a solution. Examination of various data on place and voicing perception suggests the following points. Features correspond to natural boundaries between sounds, which are included in the infant's predispositions for speech perception. Adult percepts arise from couplings and contextual interactions between features. Both couplings and interactions contribute to invariance. But this is at the expense of profound qualitative changes in perceptual boundaries implying that features are neither independently nor invariantly perceived. The question then is to understand the principles which guide feature couplings and interactions during perceptual development. The answer might reside in the fact that: (1) adult boundaries converge to a single point of the perceptual space, suggesting a context-free central reference; (2) this point corresponds to the neutral vocoïd, suggesting the reference is related to production; (3) at this point perceptual boundaries correspond to the natural ones, suggesting the reference is anchored in predispositions for feature perception. In sum, perceptual invariance seems to be grounded on a radial representation of the vocal tract around a singular point at which boundaries are context-fee, natural and coincide with the neutral vocoïd.

1. Introduction

You never bath twice into the same river (Heraclitus). While everything always changes what remains invariant? A fairly classical solution to the non-invariance problem is to look for constant relationship. According to Everitt (1998), invariance is "A property of a set of variables or a statistic that is left unchanged by a transformation" (p. 168). The purpose of this paper is to give some hints for handling the non-invariance problem in speech communication.

Features, the ultimate units of language (Jakobson, 1973), are the best candidates as building blocks for speech perception (Jakobson, Fant & Halle, 1952). Features were first defined on phonological grounds, as a function of their distinctive function in the language, hence "distinctive" features. They were later defined on articulatory grounds in the framework of Generative phonology; hence "phonetic" features (Chomsky & Halle, 1968). Though features are key concepts in empirical investigations, their perceptual invariance has been repeatedly questioned (Fromkin, 1979). How can we pretend that features are perceptually constant when there is massive evidence (Repp, 1982) to show that the *perception* of a given feature (e.g. stop place of articulation) depends on the phonetic context (e.g.: the following vowel, Schatz, 1953)? Simply by looking at contextual variations in feature *production*. Features are invariant to the extent that perceptual variations parallel those in production. Whenever this is true, the relationship between perception and production does not change across contextual transformations, conforming to the very definition of invariance.

Practically, invariance can be tested by comparing perceptual boundaries with productive categories, i.e. those present in speech production and which can be specified with acoustic measurements. With two different categories (e.g. /b/ and /p/) separated by a single feature (e.g. voicing), the perceptual boundary is the point along some acoustic continuum at which the categories are equally perceptible. Boundaries are usually measured by collecting labeling responses to stimuli generated by modifying an acoustic cue known to play a major role in the perception of the feature (e.g. for voicing: Voice Onset Time, VOT; Lisker & Abramson, 1964), and the boundary value corresponds to the point at which the two labeling responses are equi-probable (e.g. 50 % /b/ and /p/ labeling). Perceptual boundaries can then be matched with the distributions of the major cue in the production of the categories (Figure 1). Results on voicing perception (in English: Lisker & Abramson, 1976; in French: Serniclaes, 1987) show that both the perceptual boundary and the productive categories change with the context (e.g. voicing boundary and related productive categories change from /ba-pa/ to /gi-ki/ in Figure 1). However, as the relationship between voicing boundaries and categories remains fairly constant across contexts (as in Fig.1), feature perception can be considered to be nearly invariant. Studies on place of articulation also suggest parallel contextual shifts in perception and production (Dorman et al., 1977).

The fact that contextual variations do not grossly affect the relationship between perceptual boundaries and productive categories suggests that featural percepts are invariant. However, as we will see, this is at the expense of cross-

dependencies in the perception of different phonetic features: the perception of a given feature (e.g. voicing) depends on other features (e.g. place or vowel), and vice-versa.

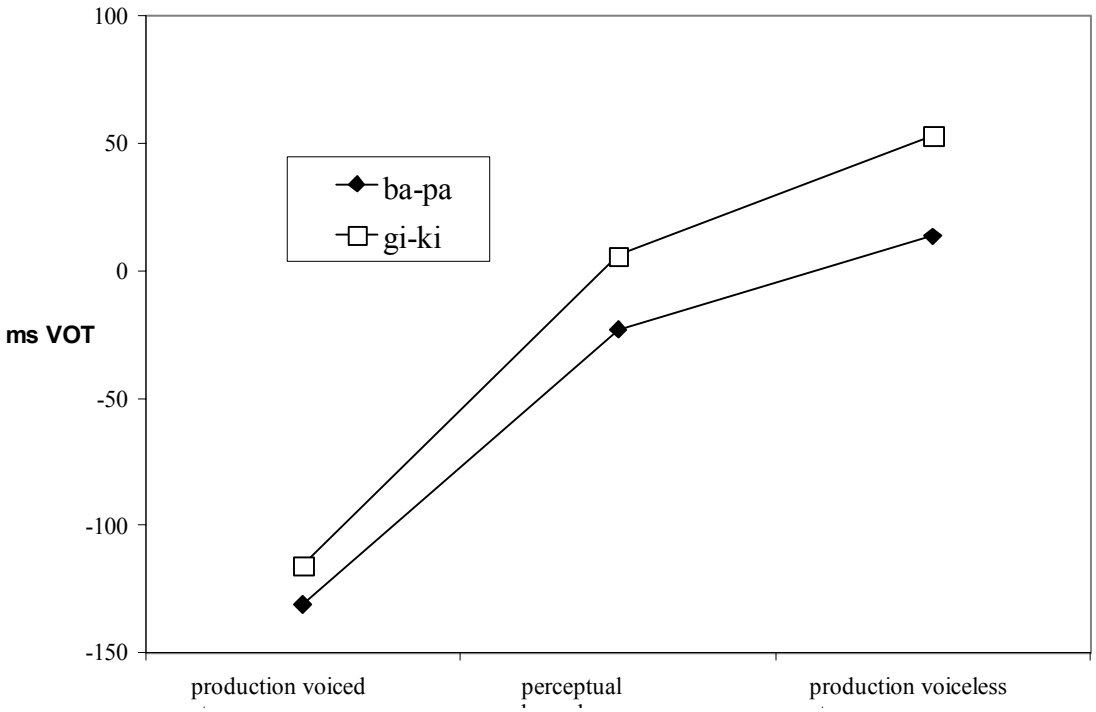


Figure 1. Relationship between voicing perception and production in French (adapted from Serniclaes, 1987). Mean acoustic measurements of VOT in voiced and voiceless stops as well as the mean perceptual boundaries along a synthetic VOT continuum are given in two phonetic contexts, i.e. /labial stop + a/ (/ba-pa/) and velar /stop + i/ (/gi-ki/). The contextual shift in perception (29 ms VOT) is about half-way between those in production (15 and 39 ms VOT for voiced and voiceless stops respectively; geometric mean = 25 ms). Perceptual boundaries follow the productive variations, resulting in a fairly stable relationship across contexts.

The present paper gives an overview of the non-invariance problem and offers some hints towards a solution. First, the empirical evidence for cross-dependencies in feature perception is reviewed. Then, data which suggest that adult percepts arise from couplings between perceptual predispositions for the perception of phonetic features will be presented. Perceptual couplings are combinations between phonetic features giving rise to language-specific features, hence "phonological" in nature. A further question will be to understand the nature of the representation which guides the development of feature couplings during language acquisition. At this point I will consider two basically different models of speech perception, the one based on auditory properties (Stevens, 1989), the other on motor ones (Liberman & Mattingly,

1985). I will argue that feature couplings are driven by a specific version of the speech-specific model, based on a radial representation of the vocal tract. Further, I will argue that this representation is based on a central reference corresponding to the neutral vocoïd (the "schwa") and that the distinction between language-specific and auditory-like processing disappears around that central reference. The latter is therefore not only central but also singular.

2. Perceptual dependencies between features

There are numerous examples to suggest that the perception of a given feature is affected by the phonetic context (for a review: see Repp, 1982). As a rule, contextual variations in feature production are paralleled by contextual adjustments in feature perception. Several models of contextual adjustment are possible. According to the "Auditory-acoustic" model, contextual effects in perception are due to simultaneous changes of acoustic cues which affect both the target feature and the contextual features. For example, the duration of formant transitions affects both the perception of voicing and place of articulation in stop consonants: longer transitions indicate both back vs. front place of articulation (/g-k/ vs. /b-p/) and voiced vs. voiceless category (/k-p/ vs. /g-b/). The inclusion of transition duration in the repertoire of voicing cues therefore contributes to the shift of the VOT boundary towards longer values (i.e. more voiceless) in a /b-p/ vs. /g-k/ context (Figure 1), transitions being longer (i.e. more voiced) in the latter. More generally, the multiple cueing of phonetic features might open the way for solving the non-invariance problem since the acoustic cues contributing to the perception of the same feature vary in a complementary way across contexts (Serniclaes, 1975; Dorman et al., 1977): when one cue is weaker (e.g. VOT is short in /p/, long in /k/), another is stronger (e.g. transitions are short in /p/, long in /k/). As the contextual variations of the cues compensate for each other, cue integration might give the key for solving the non-invariance problem.

While acoustic cue integration undoubtly contributes to perceptual invariance, this is not the whole story. According to the "Phonetic" model – to take back the classical Haskins' terminology (Carden et al., 1981) - contextual effects also truly arise from cross-dependencies in the perception of different features and, as we will see, this model is supported by the results of fairly sophisticated experiments. Two different "Phonetic" models are in turn possible. Perception of a given feature might simply bias the phonetic categorization of another feature. This is the "Additive" model. Alternatively, perception of a given feature might affect the processing of the acoustic cues involved in the perception of another feature. This is the "Interactive" model.

2.1. Auditory invariance: Locus model

The Locus model of place perception is undoubtedly the most elaborated form of Auditory-acoustic model of feature perception. According to this model, first settled by Delattre (Delattre, Liberman, & Cooper, 1955) the perceptual invariance of stop place of articulation in CV syllables is based on the virtual onset of F2 transition, extrapolated from its acoustic onset and offset. According to Delattre, the invariant for each place category is the frequency *value*, or Locus, towards which F2 transitions point in different vocalic contexts. As further research demonstrated that the Locus was not constant across vocalic contexts, the model has since been reformulated by Sussman (Sussman, McCaffrey, & Matthews, 1991; Sussman, Fruchter, Hilbert & Sirosh, 1998). Instead of a single value, now it is the *linear relationship* between the onset and offset of F2 transition which is supposed to be invariant for each place category (Equation 1).

Equation 1. $(F2)_{onset} = I + B_{Place} * (F2)_{offset}$

where Place $\in$ {labial, coronal, dorsal, velar}
and $(F2)_{onset}$, $(F2)_{offset}$ correspond to acoustic measurements of the second formant in CV syllables
where I is the intercept

The invariants were originally formulated in terms of categories because they were primarily intended to be tested with production data, but they can be easily transposed into boundary invariants in order to cope with perceptual data (Equation 2).

Equation 2. $(F2)_{onset} = I + B_{labial\text{-}coronal} * (F2)_{offset}$

where $B_{labial\text{-}coronal}$ is a linear transform of B_{labial} and $B_{coronal}$
and $(F2)_{onset}$, $(F2)_{offset}$ correspond to the acoustic values of the second formant at the perceptual boundary.

This model is motivated by both ecological and phylogenetic considerations. According to Sussman et al. (1998): (1) linear relationships are quite common in the acoustic environments of species which are able to operate complex auditory processes; (2) vertebrates are endowed with pre-adapted mechanisms for processing linear relationships; (3) the human vocal system would result from an evolutive pressure leading to the production of stimuli which conform to these relationships. With this linear conception, the Locus is not fixed for each place but depends on the vocalic context. However, the invariant remains acoustic in

nature because contextual adjustments operate through acoustic cue integration and do not depend on the perception of the adjacent vowel. According to the Locus equations, the percept does not depend on variations in the vocalic percept as long as the acoustic stimulus remains unchanged. This implies that fluctuations in vowel perception occurring with ambiguous stimuli should not affect consonant perception.

Among the widespread criticisms which have been addressed to the Linear model (cf. the comments to Sussman et al., 1998), the most important ones for our concern here are those related to the non-invariance problem. To sum up these criticisms, invariance has to rely on several different acoustic cues – not only F2 transition but also F3 transition and the burst- and the relative weightings of these cues should depend on the speaker and context (in the comments to Sussman et al., 1998: Carré p.262; Blumstein, p.260; Diehl, pp.264; Nearey, p.277). While this meshes neatly with the abundant evidence on cue multiplicity (Delattre, 1968) and contextual changes in the contribution of the different cues (such as those of formant transitions and burst: Dorman et al., 1977), the question is to know whether the contextual effects are indeed entirely acoustic in nature.

2.2. Perceptual dependencies between features

2.2.1. The phonetic vs. acoustic model

Although there is an acoustic component in contextual adjustments, the acoustic model cannot account for different data which suggest that identification of a given feature depends on the perceived identity of the surrounding features. These data show that in conditions where all the possible effects of acoustic cues were controlled, including those arising from random fluctuations in cue extraction with the same stimulus, contextual effects were still present and could then only arise from perceptual dependencies. Carden et al. (1981) demonstrated that place perception in consonants depended on whether exactly the same stimuli were presented either as stops or as fricatives. Similarly, using /Stop+ Vowel/ stimuli in which both voicing and place cues were fixed at ambiguous values, we showed that fluctuations in voicing categorisation depended on those in place categorization (Serniclaes & Wajskop, 1992). Further, the inclusion of vowel identification responses is necessary to account for consonant place identification as evidenced by the analysis of perceptual data with Logistic Regression models (Nearey, 1990).

2.2.2. The phonetic interactive vs. additive model

While these experiments suggest that the Auditory-acoustic model is too simple, different speech specific models are in turn possible. Perception of a given feature might simply bias the phonetic categorization of another feature. This is the "additive" model (Equation 3). Alternatively, perception of a given feature might affect the processing of the acoustic cues involved in the perception of another feature. This is the "interactive" model (Equation 4).

Equation 3. $(F2, F3)_{onset} = I + Vowel + B_{labial\text{-}coronal} * (F2, F3)_{offset}$

Equation 4. $(F2, F3)_{onset} = I + Vowel + B_{(labial\text{-}coronal)} * (F2, F3)_{offset} * Vowel$

where 'Vowel' represent the perceived identity of the vowel.

Examination of previous data on the perception of English synthetic /si, ʃi, su, ʃu/ syllables by Nearey (1990) led to the conclusion that effects of vowels on consonant identification were additive. Logistic Regression functions were used by Nearey for testing the additive vs. interactive perceptual models. Vowel and consonant bias terms were significant but interactive terms were not significant, which supported the additive model. However, in a more recent study on Dutch fricative-vowel syllables, Smits (2001a) found evidence supporting perceptual interactions using a Hierarchical Categorization model (HICAT: Smits, 2001b). HICAT allows to separate tests of the effects of vowel on consonant perception from those of consonant on vowel perception, a distinction which was not addressed in Nearey's work.

2.2.3. A specific phonetic interactive model: the Radial Model

We provided a further test of the perceptual dependencies between features in an experiment on the perception of synthetic /fricative+vowel/ syllables generated by factorial modification of formant transitions onset-offset, with F2 and F3 covarying (Serniclaes & Carré, 2002). The data also supported an interactive model of phoneme perception and further showed that the additive component was not necessary (Equation 5).

Equation 5. $(F2, F3)_{onset} = I + B_{(labial\text{-}coronal)} * (F2, F3)_{offset} * Vowel$

Geometrically, the absence of an additive component means that the boundaries converge to a single point in the space of formant transitions onset-offset,

(Figure 1). This means there is a point in the perceptual space at which place perception is context free. Interestingly, the convergence point corresponds to a stimulus with flat F2-F3 formant transitions with values corresponding to the neutral vocoïd (1500 Hz F2-2500 Hz F3), corresponding to the uniform vocal tract. Further, flat transitions constitute a natural auditory boundary between rising transitions and falling transitions (Cutting & Rosner, 1974). It thus seems that place perception is organized around a central reference characterized by both natural and context free boundaries, and corresponding to the neutral productive category. With the vocal tract in a fairly neutral position, place perception does not strongly depend on the perception of the vocalic context and is derived from natural auditory sensitivities. However, outside the neutral context, the interaction between place and vowel perception generates speech specific boundaries which become increasingly different with the distance from the neutral vocoïd measured on directions which depend on the perceived identity of the vowel. This suggests that place perception is based on a "radial" representation anchored on the neutral vocoïd. This representation is suggested by the fact that perceptual boundary for place of articulation executes a radial movement from the front vowel contexts (on the right-hand in Figure 2) to back vowel contexts (on the left-hand in Figure 2).

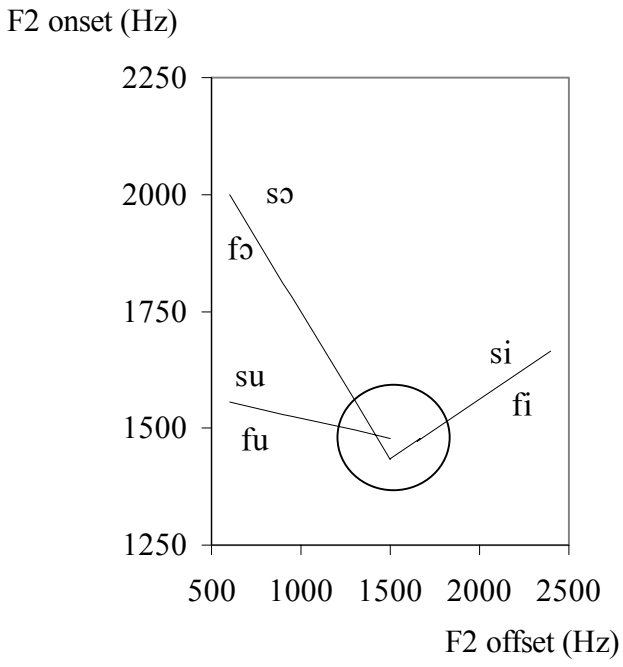


Figure 2. (adapted from Serniclaes et al., 2002): Perceptual boundaries in the F2 onset-offest plane. F3 onset-offset values covaried with those of F2 in this experiment and F3 was close to 2500 HZ for F2 of 1500 Hz, which corresponds to the neutral vocoïd values. In agreement with the radial model, the obtained boundary lines converge to the F2 flat transition when the offset value is close to 1500 Hz. For stimuli with offset values close to 1500 Hz F2 (circled region), place perception is fairly independent of the vocalic context.

3. Couplings between perceptual predispositions for speech

3.1 Models of speech development

Human infants are born with predispositions for perceiving all the possible phonetic contrasts, which are then activated or not as a function of the presence versus absence of the corresponding contrast in the linguistic environment. This fairly classical view on speech development (Werker & Tees, 1999) is grounded on a considerable amount of empirical evidence. Neonates can already discriminate between a range of phonetic categories (Eimas, Siqueland, Jusczyk & Vigorito, 1971), even between those which are not present in their ambient language (e.g. Lasky, Syrdal-Lasky & Klein, 1975). The initial ability to discriminate the universal set of phonetic contrasts however declines within the first year of life (Werker & Tees, 1984a) but the decline involves a change in processing strategies rather than a sensorineural loss (Werker & Tees, 1984b).

Infant studies not only show that the discrimination between phonetic categories is already present at birth, they also indicate that the location of phonetic boundaries already depends on several acoustic cues. Thus, the discrimination between voiced and voiceless stops by infants below six months of age depends both on voice onset time (VOT) and F1 transition duration, just as for adult speakers of English (Miller & Eimas, 1983). Innate mechanisms might thus also explain the integration of multiple cues for the perception of the same phonetic feature.

Perceptual development would be fairly simple if it was restricted to selecting the percepts in a stock of innate predispositions, as in Werker's model. Phillips (2001) calls this a "structure-adding" approach, all features being processed at a universal "phonetic" level processing and only those specific to the language at an upper-stage "phonological" level. Alternatively, the adult perceptual space might not be straightforwardly related to the universal predispositions (Kuhl, 1994; 2000), what Phillips considers as a "structure-changing" approach. A third possibility is that language specific features are generated by couplings between phonetic features (Serniclaes, 1987; 2000), which implies both structure-adding and structure-changing.

Couplings are combinations between features. Couplings create new functional entities inside which features are integrated. The term "coupling" is common-place in the study of visual perception, e.g. for describing perceptuo-motor integration in depth perception (Hochberg, 1981).

3.2 Test of a mixed model: coupling between predispositions

In support of the coupling model, previous research already suggested that voicing perception in several languages is based on a VOT boundary which is not precluded in the infant's predispositions. Up to about 6 months of age, infants discriminate three voicing categories, separated by two VOT boundaries (see Figure 2; Lasky, Syrdal-Lasky, & Klein, 1975; Aslin, Pisoni, Hennessy, & Perrey, 1981). After 6 months of age, only the positive VOT boundary remains active in languages with a single distinction between short vs. long positive VOT categories (e.g. English; Figure 3; Eilers, Wilson & Moore, 1979). Languages such as Spanish and French use a single distinction between negative VOT and moderately long positive VOT categories (Caramazza & Yeni-Komshian, 1974; Williams, 1977), and the perceptual boundary is located around 0 ms (Serniclaes, 1987). The fact that the boundary is located around 0 ms means that negative and positive VOT are equally important for voicing identification and hence that the categorical predispositions for the perception of negative and positive VOT are both activated and coupled in the course of perceptual development. It might be argued that the 0 ms VOT boundary simply emerges in the course of development, while the positive and negative boundaries are deactivated. While this is of course possible, the inclusion of predispositions combinations in the predispositions would seriously entail the parsimony of the model.

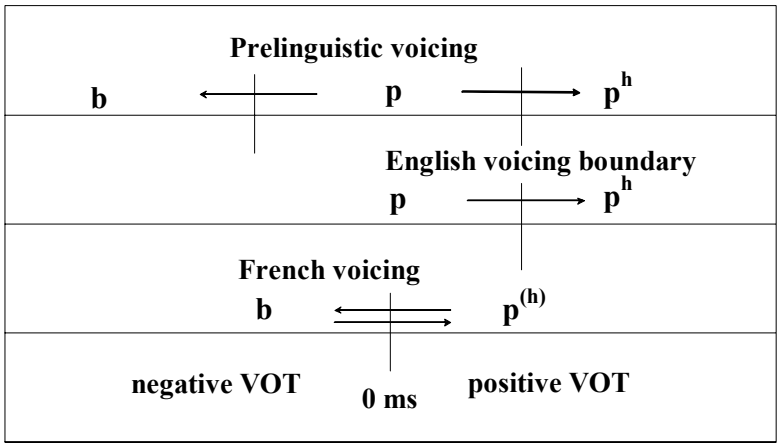


Figure 3 (from Serniclaes et al., 2004): Perceptual boundaries between voicing categories in prelinguistic children, in English and in French for stops in syllable-initial position . Prelinguistic boundaries correspond to predispositions (indicated by arrows) for the perception of all potential categories (voiced as for /b/, voiceless as for /p/ and voiceless aspirated as for /p^h/) in the world's languages. In English, a single predisposition is activated for performing the distinction between voiceless unaspirated and voiceless aspirated stops. In French, two predispositions are coupled in order to perform the distinction between voiced and slightly aspirated voiceless stops.

In support of the coupling hypothesis, examination of studies on children raised in Spanish-speaking environments showed that the 0 ms VOT boundary is not predicted by the infant's predispositions (Lasky et al., 1975), although it appears fairly early in the course of language development (Eilers et al., 1979). Recently, data collected on children raised in French-speaking environments suggested those around 4 months of age discriminated the negative and positive boundaries (located at -30 and +30 ms VOT respectively) whereas those around 8 months of age discriminated the 0 ms VOT boundary (Hoonhorst, 2004). Further evidence on couplings between predispositions has been obtained for the perception of place of articulation. F2 and F3 transitions allow separating the three place categories usually found in languages, i.e. labial, coronal and velar. In the neutral vocalic context, stimuli with raising F2-F3 transitions correspond to /b/ percepts, those with falling F2-F3 transitions correspond to /d/ percepts and those with falling F2 and rising F3 transitions to /g/ percepts (Carré, Liénard, Marsico & Serniclaes, 2002). However, a fourth category characterized by raising F2 and falling F3 transitions is also possible and it might correspond to the palatal consonants found in Czech (Jakobson et al., 1952) and also in Hungarian (Geng et al., 2005). As the perception of rising vs. falling transitions is grounded on natural boundaries –flat transitions, see above- the discrimination of F2 and F3 transitions is probably present in the newborn, although there is no direct evidence on this point. The predispositions for perceiving F2 and F3 transitions might straightforwardly be used in four-category languages, two binary features allowing to discriminate four place categories.

However, the natural F2 and F3 boundaries are not optimal for perceiving consonants in three-category languages. The F2-F3 perceptual space should be divided into three equally sized regions for optimal use, which would require new boundaries (Figure 4). These boundaries can only be obtained by trade-off between F2 and F2 transitions, e.g. a strongly falling F3 compensating for a slightly raising F2 for perceiving /d/ instead of /b/. Notice that if F2 and F3 transitions are not simply two different acoustic cues but are instead precluded into different perceptual predispositions, the very existence of a perceptual trade-off between F2 and F3 transitions means coupling between predispositions.

We have recently found evidence in support of this conjecture by collecting both identification and discrimination responses to /stop + neutral vocoïd/ synthetic syllables generated by either factorial or combined modification of F2 and F3 transition onsets. Preliminary results (Serniclaes, Bogliotti & Carré, 2003; see Figure 4) showed that French adult speakers discriminated natural F2 and F3 boundaries -i.e. those corresponding to flat transitions- though their labelling boundaries reflected trade-offs between F2 and F3. The fact that perceptual boundaries for place of articulation are built on trade-offs between of the

coupling hypothesis. These results have since been confirmed with a larger sample of subjects (Bogliotti, 2005) two acoustic cues, which are each endowed with natural boundaries, provides further support.

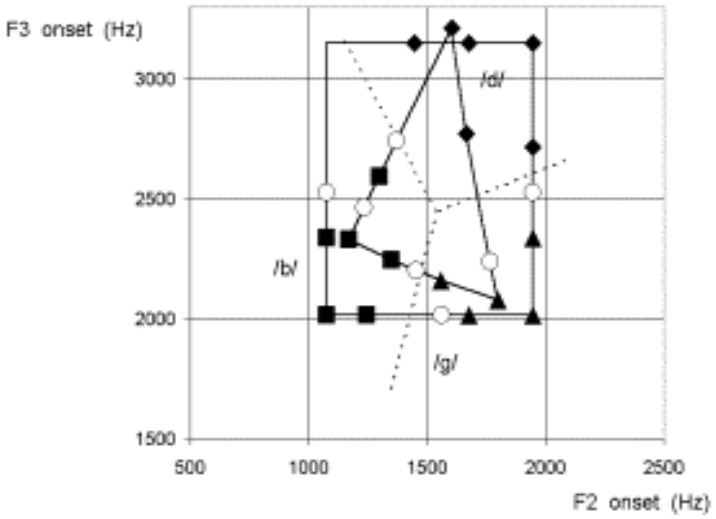


Figure 4. (from Serniclaes et al., 2003): Labeling and discrimination results for two places of articulation continua. Stimuli collecting at least 75% /b/, /d/ or /g/ responses are indicated by squares, diamonds and triangles respectively. Broken lines represent category boundaries. White circles indicate above chance discrimination peaks.

3.3 Speech perception in dyslexic children: a coupling deficit?

Phonological couplings between features imply considerable qualitative changes. It would therefore not be surprising to find coupling deficits in some part of the population. Our research on speech perception dyslexia lends support to the existence of coupling deficits and this paved the way to a new explanation of dyslexia. Our work in this domain is in the framework of the phonological explanation of reading deficits (for an overview see: Sprenger-Charolles, 2003). Previous investigations showed that dyslexics are affected by deficient grapheme-phoneme correspondences, deficits in phonological awareness, phonological short-term memory, phoneme discrimination and in categorical perception. But where is the core deficit? A first investigation showed that the categorical perception deficit in dyslexia is characterized by a better discrimination of within-category differences (Serniclaes, Sprenger-Charolles, Carré & Démonet, 2001). This result suggested a new hypothesis as to the origin of dyslexia, namely that it comes from a deficit in the coupling between predispositions in the course of perceptual development which gives rise to an allophonic, rather than phonemic, mode of speech perception. Allophonic perception offers a possible explanation to dyslexia. The child who perceives speech in allophones has an evident handicap for discovering the relationship

between the speech sounds and alphabetic symbols, knowing that the opacity of the writing system adds a further difficulty, of cultural order. Allophonic perception has several testable consequences as to the difference between dyslexics and controls. The main one is that dyslexics should be less categorical than controls for phonemic distinctions and should be more categorical for the allophonic ones. This prediction was recently confirmed by the results of several different investigations (Bogliotti, 2003; Serniclaes, Van Heghe, Mousty, Carré & Sprenger-Charolles, 2004; Burnham, 2003).

4. Discussion and conclusions

To summarize the evidence contemplated in the previous sections, two basic findings emerge. Firstly, while phonetic features were initially conceived as autonomous dimensions of speech, it now appears that they are not independently perceived. Secondly, while phonetic features are language-independent dimensions of speech, they are not always directly used for speech perception in a given language. Rather, speech perception is also based on language-specific couplings between phonetic features.

Phonetic features were conceived as language-independent autonomous dimensions of speech production. The discovery of infant's predispositions for feature discrimination and the traces they leave in the adults make it clear that features remain the best candidates as building blocks of speech perception. However, couplings between predispositions show that phonetic features do not constitute independent units for language perception in the adult. If features are independent units at the start, why are they interactively processed? Presumably because when features are put into action in a linguistic frame, they do not have invariant acoustic correlates. Couplings between features constitute an obvious remedy to contextual effects in production. For example, voicing contrasts are easier to produce in labial stops before open vowels (e.g. /ba/-/pa/), while aspiration contrasts are easier to produce in velars stops before closed vowels (e.g. /gi/-/ki/). Coupling voicing and aspiration then gives rise to a more stable and possibly invariant compound. However, the fact that voicing and place perception are not independent indicates that couplings are not sufficient for attaining invariance and that contextual interactions further contribute to it.
An important question is to understand the principles which guide the development of feature couplings and interactions during language acquisition. How is the search for invariance implemented in development? Invariance requires that perceptual boundaries fit into productive categories. There are basically two different ways by which feature compounds might be invariant: invariance might either be driven by motor representations in perception or by auditory

representations in production. Speech perception theories can be subdivided into four classes, depending on whether invariance is conceived with or without major contribution from learning and whether it is based on auditory or speech specific representations (Serniclaes, 2000). Both the Quantal Theory (Stevens, 1989), and the one based on "natural" psychoacoustic boundaries (Kuhl & Padden, 1983) are basically inneist as they consider that invariance is the by-product of auditory integration and that learning only plays a marginal role. While there are predispositions for feature perception, we have seen that acquisition plays a crucial role in speech perception. No wonder then if other auditory theories are centered on acquisition. Among them the 'Perceptual Magnet' theory (Kuhl, 1994; 2000) considers that adult percepts are shaped by linguistic experience. Though the perceptual magnets are not clearly related to innate dimensions, they might easily be accommodated with couplings between predispositions. However, while magnets are quite interesting concepts for understanding the genesis of linguistic categories, they should be conceived in speech specific rather than auditory terms.

Motor theories suppose that direct links exist between perception and motor commands (Liberman & Mattingly, 1985), a contention which received recent support by the existence of mirror neurons (Fadiga, Fogassi, Paresi & Rizzolatti, 1995; Rizzolatti, Fadiga, Gallese & Fogassi, 1996; Studdert-Kennedy, in press). In further support of this conception, fRMI results show that, with exactly the same acoustical stimuli, a change in perceptual mode from nonspeech to speech affects the localization of the brain activity (Dehaene-Lambertz, Pallier, Serniclaes, Sprenger-Charolles, Jobert & Dehaene, 2005). These results strongly suggest that speech perception is achieved through specific pathways different from those used in auditory perception because the change in the neural site of processing in the brain was obtained with exactly the same stimuli, thereby excluding possible confounding effects arising from differences in stimulus complexity.

While there is recent strong neuro-imagery evidence in support to the Motor theory, the latter is basically inneist, a view which is difficult to conciliate with couplings between predispositions. Articulatory theories, notably the “direct-realist” one (Studdert-Kennedy, 1985; Fowler, 1986), rely on the learning of invariants from environmental regularities and are therefore better suited for explaining the complexities of perceptual development.

While it seems fairly clear that speech perception is related to articulatory representations, the precise nature of these representations remains unknown. However, some hints might be found in our results on place of articulation perception (see above, Serniclaes et al., 2002; 2003). Place perception seems to be built up around a singularity of the perceptual space characterized by boundaries which are both context-free and natural. Further, this singularity

coincides with the neutral vocoïd, which corresponds to the uniform vocal tract. This provides a straightforward link between perception and production (Carré, Liénard, Marsico & Serniclaes, 2002). As contextual adjustments in perception correspond to radial movements of boundary lines around the neutral point, it would seem that perception occurs in a spatial representation of the vocal tract with radial lines as contextual variants of a central, and context-free, reference.
This "radial" model of speech perception needs to be refined and tested with appropriate means. But it can already find some support by the fact that the only neural site with is specifically dedicated to the categorization of speech features is located in the left supra-marginal gyrus (Dehaene-Lambertz et al., 2005), a region which is linked to the sensory representation of the mouth and might correspond to part of the auditory cortex devoted to the processing of spatial information.

Acknowlegments

I would like to thank Bernd Pompino-Marschall and Noel Nguyen for his insightful comments on a previous version of this manuscript.

References

Aslin, R.N., Pisoni, D.B., Hennessy, B.L., & Perrey, A.V. (1981). Discrimination of voice onset time by human infants: New findings and implications for the effect of early experience. *Child Development, 52,* 1135-1145.

Bogliotti, C. (2003). Relation between categorical perception of speech and reading acquisition. In M.J.Solé, D.Recaesens & J.Romero (Eds.). *Proc. 15th International Congress on Phonetic Sciences*, 885-888.

Bogliotti, C. (2005). Perception categorielle et perception allophonique: Incidences de l'âge, du niveau de lecture, et des couplages entre predispositions phonétiques. *PhD. Thesis, Université Paris 7 - Denis Diderot*.

Burnham, D. (2003). Language specific speech perception and the onset of reading. *Reading and Writing, 16,* 573-609.

Caramazza,A. and Yeni-Komshian,G.H. (1974). Voice onset time in two French dialects. *Journal of Phonetics, 2,* 239-245.

Carden, G., Levitt, A., Jusczyk, P.W., & Walley, A. (1981). Evidence for phonetic processing of cues to place of articulation: Perceived manner affects perceived place. *Perception and Psychophysics, 29,* 26-36.

Carré, R., Liénard, S., Marsico, E., & Serniclaes, W. (2002). On the role of the "schwa" in the perception of plosive consonants. In J.L.H. Hansen & B. Pellom (Eds.). *7th International Conference on Spoken Language Processing*, 1681-1684.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York Harper and Row.

Cutting, J.E., & Rosner, B.S. (1974). Categories and boundaries in speech and in music. *Perception and Psychophysics*, *16,* 564-570.

Dehaene-Lambertz, G., Pallier, Chr., Serniclaes, W., Sprenger-Charolles, L., Jobert, & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *NeuroImage*, *24,* 21-33.

Delattre, P.C., Liberman, A.M., & Cooper, F.S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27,* 769-773.

Delattre P. (1968). "From acoustic cues to distinctive features" *Phonetica* 18, 198-230.

Dorman, M.F., Studdert-Kennedy, M., & Raphaël, L.S. (1977). Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception and Psychophysics, 22,* 109-122.

Eilers, R., Wilson, W. and Moore, J. (1979). Speech discrimination in the language-innocent and the language-wise: a study in the perception of voice onset time. *Journal of Child Language, 6,* 1-18.

Eimas, P.D., Siqueland, E.R., Jusczyk, P. & Vigorito, J. (1971). Speech perception in infants. *Science, 171,* 303-306.

Everitt, B.S. (1998). *The Cambridge dictionary of statistics*. Cambridge University Press.

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology*, 73, 2608-2611.

Fowler, C.A. (1986). An event approach to the study of speech perception. *Journal of Phonetics, 14,* 2-28.

Fromkin, V.A. (1979). Persistent questions concerning distinctive features. In B.Lindblom and S.Öhman (Eds.) *Frontiers of Speech Communication Research.* London: Academic Press. 323-334.

Geng, C., Mády, K., Bogliotti, C., Messaoud-Galusi, S., Medina,V. & Serniclaes, W. (2005). Do palatal consonants correspond to the fourth category in the perceptual F2-F3 space? *ISCA Workshop on Plasticity in Speech Perception, London, June 15-17 2005:* 219-222.

Hochberg J. (1981). On cognition in perception: perceptual coupling and unconscious inference. *Cognition, 10,* 127-34.

Hoonhorst (2004). L'évolution de la discrimination phonologique des jeunes enfants entre 4 et 8 mois et ses implications sur la compréhension de l'étiologie de la dyslexie. Mémoire de Licence Spéciale en Logopédie, ULB-UCL.

Jakobson, R. (1973). *Essais de Linguistique Générale*. Paris: Editions de Minuit.

Jakobson, R., Fant, G. and Halle, M. (1952). *Preliminaries to speech analysis. The distinctive features and their correlates*. Cambridge Mass.: M.I.T. Press.

Kuhl, P.K. (1994). Learning and representation in speech and language. *Current Opinion in Neurobiology, 4,* 812-822.

Kuhl, P.K. (2000). Language, mind, and brain: Experience alters perception. In M. Gazzaniga (Ed.), *The cognitive neurosciences 2nd ed.* (pp. 99-115). Cambridge, MA: MIT Press.

Kuhl, P.K., & Padden, D.M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Journal of the Acoustical Society of America, 73,* 1003-1010.

Lasky, R.E., Syrdal-Lasky, A., & Klein, R.E. (1975). VOT discrimination by four to six and a half months old infants from Spanish environments. *Journal of Experimental Child Psychology, 20,* 215-225.

Liberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1-36.

Miller, J.L. and Eimas, P.D. (1983). Studies on the categorisation of speech by infants. *Cognition, 13,* 135-165.

Nearey, T.M. (1990). The segment as a unit of speech perception. *Journal of Phonetics 18*, 347-373.

Phillips, C. (2001). Levels of representation in the electrophysiology of speech perception. *Cognitive Science, 25,* 711-731.

Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 92,* 81-110.

Rizzolatti, G., Fadiga, L., Gallese,V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3,* 131-141.

Serniclaes, W. (1975). Perceptual processing of acoustic correlates of the voicing feature. In G.Fant ed. *Speech Communication, Proc. of the SCS, Stockholm 1974* (pp. 87-94). New York: J.Wiley.

Serniclaes, W. (1987). *Etude expérimentale de la perception du trait de voisement des occlusives du français*. Unpublished doctoral thesis. Université Libre de Bruxelles. Téléchargeable à http://www.vjf.cnrs.fr/umr8606

Serniclaes, W. (2000). La perception de la parole. In *La parole, des modèles cognitifs aux machines communicantes*. P. Escudier, G. Feng, P. Perrier, J.-L. Schwartz, Eds.; Paris: Hermès, 159-190.

Serniclaes, W., Bogliotti, C. & Carré, R. (2003). Perception of consonant place of articulation: phonological categories meet natural boundaries. In M.J. Solé, D. Recaesens & J. Romero (Eds.). *Proc. 15th International Congress on Phonetic Sciences*, 391-394.

Serniclaes, W. & Carré, R. (2002). Contextual effects in the perception of place of articulation: a rotational hypothesis. In J.L.H. Hansen & B. Pellom (Eds.). *7th International Conference on Spoken language Processing*, 1673-1676.

Serniclaes, W. & Sprenger-Charolles, L. (2003). Categorical perception of speech sounds and dyslexia. *Current Psychology Letters: Behaviour, Brain & Cognition, 10,* http://cpl.revues.org/documents379.html

Serniclaes , W., Sprenger-Charolles , L., Carré, R., & Démonet, J.-F. (2001). Perceptual discrimination of speech sounds in dyslexics. *Journal of Speech Language and Hearing Research, 44,* 384- 399.

Serniclaes, W., Van Heghe, S., Mousty, Ph., Carré, R. & Sprenger-Charolles, L. (2004). Allophonic mode of speech perception in dyslexia. *Journal of Experimental Child Psychology, 87,* 336-361.

Serniclaes, W., & Wajskop, M. (1992). Phonetic versus acoustic account of feature interaction in speech perception. in *Analytic Approaches to Human Cognition, Proc. of the Conference in Honour of Paul Bertelson*, Brussels June 1991. J. Alegria, D. Holender, J. Junça de Morais, and M. Radeau eds.(pp.77-91). Amsterdam: North-Holland.

Smits, R. (2001a). Evidence for hierarchical categorization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 1145-1162.

Smits, R. (2001b). Hierarchical categorization of coarticulated phonemes: A theoretical analysis. *Perception & Psychophysics, 63,* 1109-1139.

Sprenger-Charolles, L. (2003). Linguistic processes in reading and spelling: The case of alphabetic writing systems: English, French, German and Spanish. In T. Nunes and P. Bryant (Eds.). *Handbook of children's literacy (pp.43-65)*. Dordrecht: Kluwer Academic Publishers.

Stevens, K.N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17,* 3-45.

Studdert-Kennedy, M. (1985). Perceiving phonetic events. In W.H. Warren, Jr. and R.E. Shaw (Eds.), *Persistence and Change : Proceedings of the First International Conference on Event Perception* (pp. 139-156). Hillsdale, NJ : Erlbaum.

Studdert-Kennedy, M. (in press). How did language go discrete? In *Evolutionary Prerequisites of Language* (provisional title). M. Tallerman ed. Oxford: Oxford University Press.

Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: The orderly output constraints. *Behavioral and Brain Sciences, 21,* 241-259.

Sussman, H.M., McCaffrey, H.A., & Matthews, S.A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, *90,* 1309-1325.

Werker, J.F., & Tees, R.C. (1984a). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7,* 49-63.

Werker, J.F., & Tees, R.C. (1984b). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America, 75,* 1866-1878.

Werker, J.F. & Tees, R.C. (1999). Influences on infant speech processing: Towards a new synthesis. *Annual Review of Psychology, 50,* 509-535.

Williams, L. (1977).The voicing contrast in Spanish. *Journal of Phonetics, 5,* 169-184.

Towards a 3D articulatory model of velum based on MRI and CT images

Antoine Serrurier
Pierre Badin
Institut de la Communication Parlée, UMR CNRS 5009 – INPG – Université Stendhal, Grenoble, France

This paper describes the processing of *MRI* and *CT images* needed for developing a 3D linear articulatory model of velum. The 3D surface that defines each organ constitutive of the vocal and nasal tracts is extracted from *MRI* and *CT images* recorded on a subject uttering a corpus of artificially sustained French vowels and consonants. First, the 2D contours of the organs have been manually extracted from the corresponding images, expanded into 3D contours, and aligned in a common 3D coordinate system. Then, for each organ, a generic mesh has been chosen and fitted by elastic deformation to each of the 46 3D shapes of the corpus. This has finally resulted in a set of organ surfaces sampled with the same number of 3D vertices for each articulation, which is appropriate for *Principal Component Analysis* or *linear decomposition*. The analysis of these data has uncovered two main uncorrelated articulatory degrees of freedom for the velum's movement. The associated parameters are used to control the model. We have in particular investigated the question of a possible correlation between jaw / tongue and velum's movement and have not find more correlation than the one found in the corpus.

1. Introduction

The problem of nasality is complex and has given rise to a large number of studies, from both perception and production points of view (Rossato *et al.*, 2003; Dang *et al.*, 1994; Feng *et al.,* 1996; Teixeira *et al.*, 2000; Huffman *et al.*, 1993). The nasality feature is related to the velum position: lowering the velum, and thus opening the velopharyngeal port, is a simple gesture that induces strong and complex changes in the vocal tract acoustical behaviour. The realisation of nasality involves (1) an articulatory level that deals with the shape of the

articulators and their articulatory degrees of freedom, and (2) a control level that deals with the coordination of these articulators.
The present article describes our first attempts to acquire 3D data of the complex geometry of the articulators and cavities involved in nasality (velum, nasal passages, velopharyngeal port, paranasal sinuses) for developing a three-dimensional articulatory model based on one specific subject. A number of reasons motivate this work:

- The uvula – an appendix of the velum in the midsagittal region – is often in contact with both the back of the tongue and the pharyngeal wall, creating an occlusion in the midsagittal plane: lateral channels can however remain open and should thus be taken into account by a 3D geometry description.
- The complex muscular structure of this region, in particular the interspersion between muscles from the velum, the nasopharynx and the tongue, and the *Passavant's pad* made of the fusion of the fibres of the *palatopharyngeus* muscle with those of the *pterygopharyngeal* portion of the superior constrictor (Zemlin, 1968) leads to a sphincter like behaviour (Amelot *et al.*, 2003) that controls the velopharyngeal port opening and should be considered in three dimensions.
- A 3D model offers the possibility to provide accurate area functions of the complex nasal passages and nasopharyngeal port, with the associated articulatory control parameters, that are needed for acoustical models and thus for speech synthesis
- In the framework of the development of virtual audiovisual talking heads (cf. e.g. Badin *et al.*, 2003), the 3D visualisation of the velum constitutes an interesting addition.

2. Modelling approach

2.1. Duct vs. organs

Our aim is to develop a 3D model of the various structures and organs involved in the nasopharyngeal tract. A similar approach has been successfully applied to the tongue and to the lips (Badin *et al.*, 2002). The present model will thus constitute a complement to these models, based on the same French subject.

The final result of an articulatory model is the shape of the complete vocal and nasal tracts, needed for the aerodynamic / acoustic stage of the speech production process. We could have developed a model of *tract* or *duct*, as in Badin *et al.* (1998). However, this approach is not well suited to take precisely into account the complex geometry of the various speech articulators. This is

why we have decided to develop an *organ-based* model, i.e. to model each organ separately, and to reconstruct the oral and nasal tracts subsequently.

2.2. A subject-oriented linear modelling approach

Following a method already proven for orofacial articulatory modelling (Badin *et al.*, 2002), the 3D geometry of the various non-rigid organs will be modelled as the weighted sum of a small number of linear components. These components will be extracted by linear analysis from a set of vocal and nasal tract shapes representative of the speech production capabilities of the subject. The analysis is based on both *Principal Component Analysis (PCA)* and *linear regression*. The weights of the sum constitute the *articulatory control parameters* associated with the components: a given set of values of these parameters produces a given single shape of the organs.
Two stages are necessary to develop the model: the construction of the database of shapes from images (described in section 3.), and then the analysis of the corresponding 3D coordinates (an example is detailed in section 4.)

Note that our approach is *subject-oriented*, i.e. that the model will be based on a single (French) subject. This avoids to merge the physiological characteristics and control strategy of different subjects, which can be quite different. For example, the closure of the nasopharyngeal port can be partial for some speakers while total for others. As well, compensation strategies can be different between speakers. Finally, a practical reason is the large amount of data to process. This approach can be naturally criticized as it is speaker dependent and may not generalize extensively. However, in this first approach, the model does not intend to cover all the possible aspects of nasal articulation, but to explain a possible mechanism of articulatory movements.

The corpus consisted of a set of artificially sustained articulations designed as to cover the maximal range of articulations: the French oral and nasal vowels [a ɛ e i y u o ø ɔ œ ɑ̃ ɛ̃ œ̃ ɔ̃], and the consonants [p t k f s ʃ m n ʁ l] in three symmetrical contexts [a i u]. This corpus was supplemented by two specific articulations frequent in speech: the *rest* articulation (in a rest position) and the *prephonatory* articulation (in the preparatory phase preceding phonation). Finally we have a corpus of 46 French phonemes.

As described above, the corpus contains only 46 static articulations with many more oral articulations than nasal ones. The use of a limited number of target articulations to build a model is justified by the investigation of Badin *et al.* (1998) who showed that a limited number of target articulations leads to an

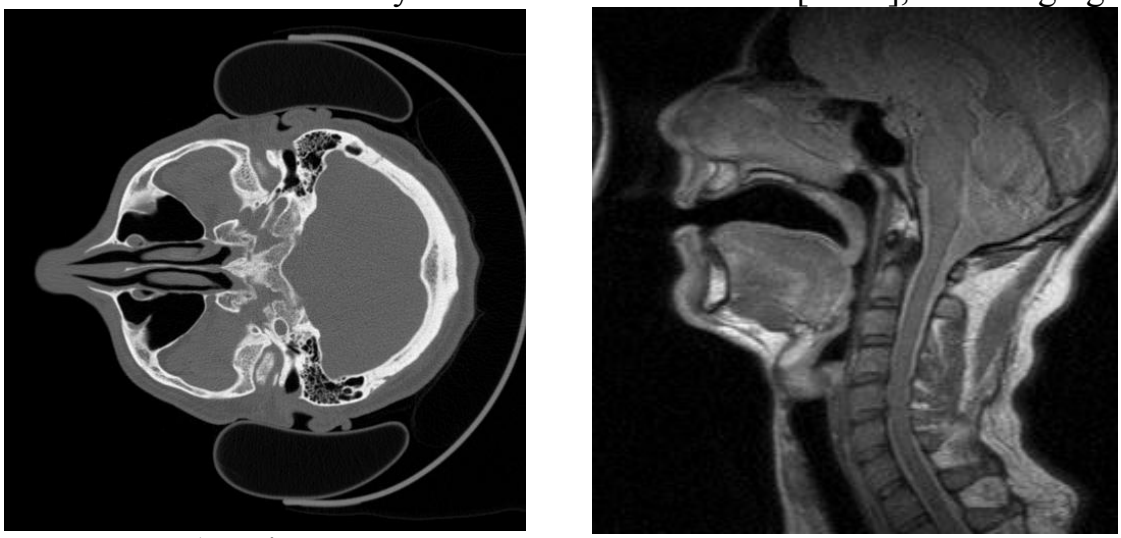
accurate articulatory model. The corpus intends to cover the maximal range of the articulators' positions, and the lack of balance between orals and nasals will not change the ability of the model to reach all possible articulations.

3. Determination of the organ shapes from MR and CT images

In order to be able to develop an articulatory model following the approach described above, it is needed to obtain a 3D surface representation of the organs of the vocal and nasal tracts (jaw, tongue, nasopharynx, oropharynx, velum, paranasal sinuses, nasal passages, nostrils, lips, epiglottis and hard palate) for each articulation of the corpus, based on stacks of MR and CT images.

3.1. Acquisition and pre-processing of the CT and MR images

A Computer Tomography scan of the head of the subject was made, to serve as a reference. A stack of 149 axial images with a size of 512×512 pixels, a resolution of 20 pixels/cm, and an inter slice space of 0.13 cm, spanning from the neck to the top of the head, was recorded for the subject at rest (see one example image in Figure 1a). These images allow making the distinction between bones, soft tissues and air, but do not allow discriminating different soft tissues. They will be used to locate bony structures and to determine accurately their shapes for reference (see section 4.2).
Stacks of sagittal MR images were recorded for the French subject sustaining artificially during about 45 sec. each of the 46 articulations of the corpus (Figure 1b illustrates vowel [a]). The subject was instructed to artificially sustain the articulation throughout the whole acquisition time. The consonants were produced in three different symmetrical vocal contexts [VCV], V belonging to

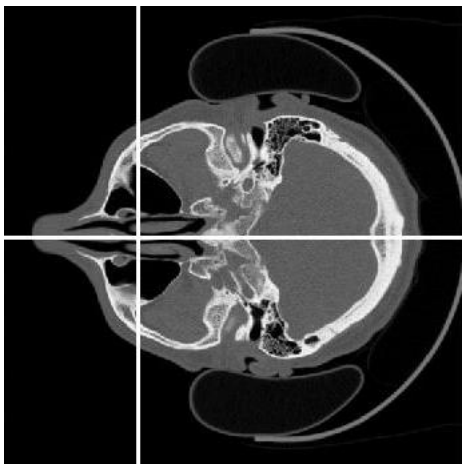


a) CT image b) MRI

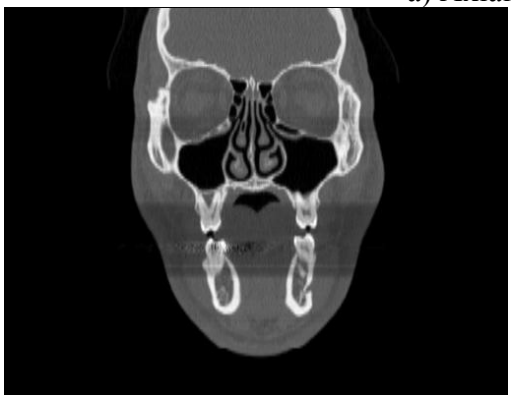
Figure 1: Original images: (a) an axial *CT image* and (b) an *MRI image* for vowel [a].

[a i u]. A set of 25 sagittal images with a size of 256×256 pixels, a resolution of 10 pixels/cm and an inter slice distance of 0.4 cm was obtained for each articulation. These images allow to make the distinction between soft tissues and air, and to discriminate soft tissues, but not to distinguish clearly the bones. Note that for both imaging techniques, the subject was in a supine position, which may alter somehow the natural shape of articulators.

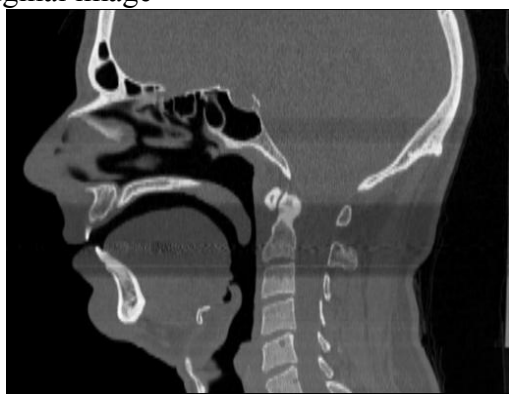
Due to the complexity of the contours of the various organs, to the relatively low resolution of the images, and to the need of an accurate reconstruction of the organs, the extraction of contours has been performed manually, plane by plane. This is a rather accurate process, except for regions where the surface of a structure is tangent to the plane, and thus the tracing difficult and not accurate: this happens for instance when tracing the tongue in sagittal planes far from the midsagittal one, and nearly tangent to the tongue sides. This is why we have supplemented the initial original stacks of images in one single orientation (*axial* for CT images and *sagittal* for MR images) by extra sets of images reconstructed by intersection of the initial stack with planes having a more useful orientation, i.e. being more perpendicular to the organ surface.



a) Axial – original image



b) Reconstructed coronal image

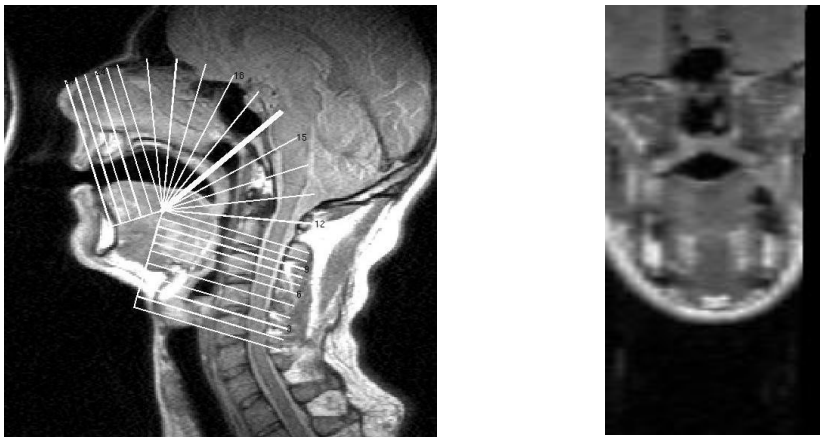


c) Reconstructed sagittal image

Figure 2: Example of CT images resclicing

The CT images have been thus resliced / interpolated in two stacks of 512 coronal images and 512 sagittal images (each new image has a size of 512×395 pixels, and a resolution of 20 pixels/cm), leading altogether to three stacks of perpendicular CT images that give a high resolution in the three orientations (see Figure 2 for an illustration).

For the MR images, the initial sagittal stack was resliced in images perpendicular to the vocal tract, considering that they will be used to extract organs shapes around the vocal tract (e.g. velum, tongue, etc.). They were thus resliced in 27 planes orthogonal to the midsagittal plane and intersecting it along a semipolar grid, as illustrated in Figure 3. Each new image is arbitrarily given a size of 200×100 pixels and a resolution of 10 pixels/cm. Finally we dispose of two redundant stacks of MR images for each articulation.



a) Sagittal – original image b) Reslice

Figure 3: Example of MR images reslicing: one perpendicular image for an [a] articulation

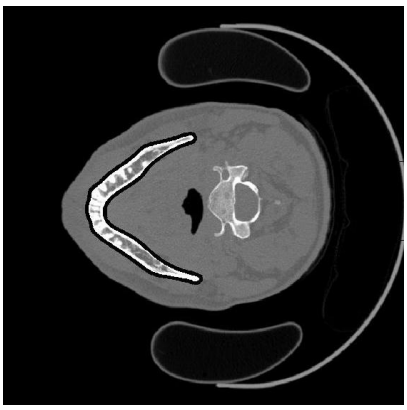
3.2. Determination of the rigid bony structures

A number of structures that makes up the vocal tract can be considered as rigid: jaw, hyoid bone, hard palate, nasal passages, nostrils and various paranasal sinuses. The shape of these structures has thus been reconstructed in the following way, from the CT images:

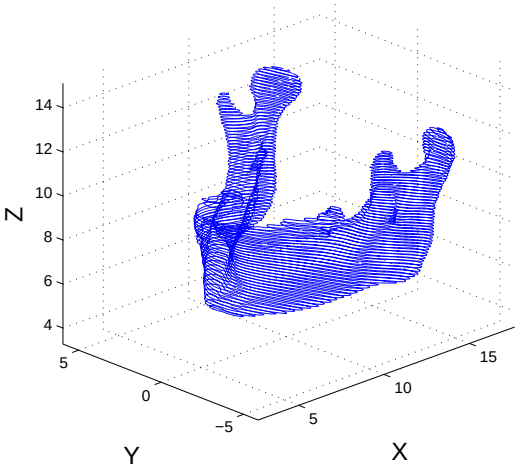
1. Manual edition of each organ, plane by plane, in one of the three stacks, or a combination of them, depending on the form and orientation of the organ, in such a way as to maximise the accuracy for complex organs (e.g. the nasal passages were hand-edited in coronal and axial stacks). Figure 4a shows, in black, the contour of the jaw manually edited from an axial CT image.

2. Expansion of the set of all 2D plane contours into a 3D coordinate system (see Figure 4b for the jaw). These 3D points are then processed through a 3D meshing reconstruction software service provided by the Prisme Research Group at INRIA (http://cgal.inria.fr/Reconstruction) to form a 3D surface meshing based on triangles (Figure 4c).

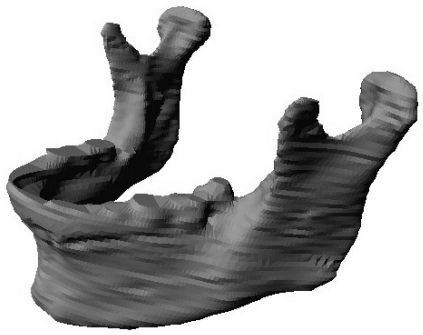
Figure 4 illustrates the 3D reconstruction process of the jaw: manual edition on plane images, and 3D meshing.



a) Manually edited contour in an axial image



b) Set of 2D contours



c) 3D surface reconstruction

Figure 4: Illustration of the 3D reconstruction of the jaw

3.3. Alignment of the images on a common reference

Before attempting to determine the shape of the soft structures, an important step is the alignment of the stacks of images for each of the articulations with a common reference. The process is composed of four steps:

1. We define an arbitrary common reference: the absolute 3D reference coordinate system is attached to the skull of the subject. In that way, some organs are *always* fixed in this system (hard palate, nasal passages,

paranasal sinuses, etc.). The reference coordinate system is arbitrary defined as follows (inherited from 2D reference coordinate system of Beautemps *et al.*, 2001):

- The x-axis is oriented from anterior to posterior in the midsagittal plane and approximately in the occlusal plane, the y-axis from left to right, and the z-axis from feet to head
- The point of coordinates (5, 0, 10) is arbitrarily set at the lower edge of the *Upper Incisors in the midsagittal plane*

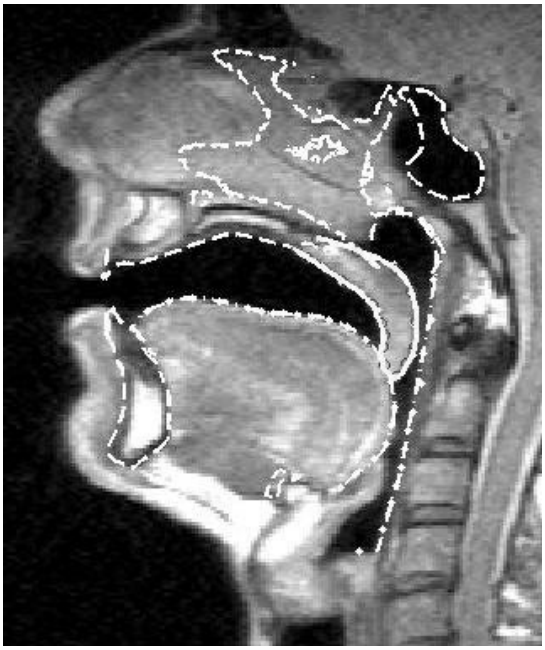
2. We align the stacks of CT images with this reference by manually placing the hard palate shape in the reference system defined above. The geometrical alignment transformation gives the position of the three stacks of CT images in the common reference. This transformation corresponds to the six degrees of freedom of a solid object and is thus defined by 6 parameters: 3 parameters for the 3D rotation and 3 parameters for the 3D translation; it will be referred to as a (*3D*) *rototranslation*.
3. Considering that the subject may have changed its position between two MR images stacks recording, each stack must be aligned with the common reference by using an appropriate 3D rototranslation. This rototranslation is obtained by aligning the rigid structures (hard palate, nasal passages, paranasal sinuses), extracted from CT images, with each of the MR images stack. The alignment of these rigid structures with a given MR images stack is a semi-automatic process: (1) anchor points of the rigid structures are manually marked with care on some of the MR images of the stack, (2) the minimization of the cumulated distance between these 3D points and the corresponding nearest points on the 3D rigid structures provides the 3D rototranslation needed. The minimization is carried out with the MATLAB function *fminunc* (minimization without constraint). A similar approach was proposed by Takemoto *et al.* (2004), the main differences being that their minimization error was the value of the volume overlap between the reference to align and the target data.
4. The previous procedure is also applied to the jaw and the hyoid bone for each articulation. This allows to determine the relative position of these bones in relation to the fixed rigid structures: by combining this relative 3D rototranslation and the absolute one corresponding to the given stack, the positions of these two structures are known in the reference common to each articulation.

It is then possible to project the intersection of the 3D surface boundaries of the rigid structures with the plane corresponding to the MR images on these images, in order to provide some useful anchor points for the interpretation of the images

and for the tracing of the soft structure contours (Figures 5a and 5b illustrate these images).

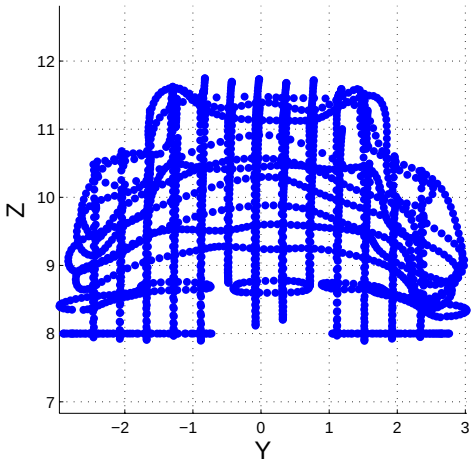
3.4. Determination of the soft structures

The determination of the soft structures (velum, nasopharynx, oropharynx, tongue, lips) is achieved in much the same way as for the rigid structures, but

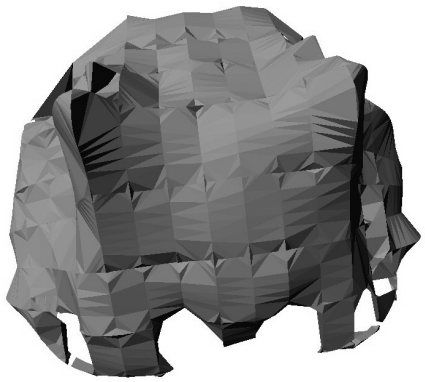


a) Edition in the sagittal plane

b) Edition in a perpendicular plane



c) Set of all 2D contours seen from front



d) 3D surface reconstruction

Figure 5: Example of manually edited contours (a and b), and of 3D reconstruction (c and d) of the velum for an [a] articulation. In a and b, the dashed lines correspond to bony structures superimposed on the images while dotted and dash-dotted lines correspond to soft structures other than the velum previously hand-edited. The solid line correspond to the velum being edited.

from the MR images of each articulation. As explained above, in order to maximise the accuracy, plane contours were edited in both stacks. The contours extracted from the rigid organs were superposed on the MR images in order to provide reference information. Figure 5 illustrates the edition of the velum in both MRI stacks for the vowel [i]. Bony structures and previously hand-edited soft structures other than the velum are superimposed on the images to help the detection of the velum (figures 5a and 5b). The velum contour (solid line on images 5a and 5b) is established as a 2D spline controlled by a limited number of points in the corresponding image plane. The set of all 2D plane contours expanded into the 3D coordinate system forms a 3D description of the given soft organ (figure 5c for a front view). The alignment with the common reference is then possible through the 3D rototranslation of the stack. An example of surface reconstruction through the 3D meshing reconstruction software service previously cited is given for the articulation [a] (figure 5d).

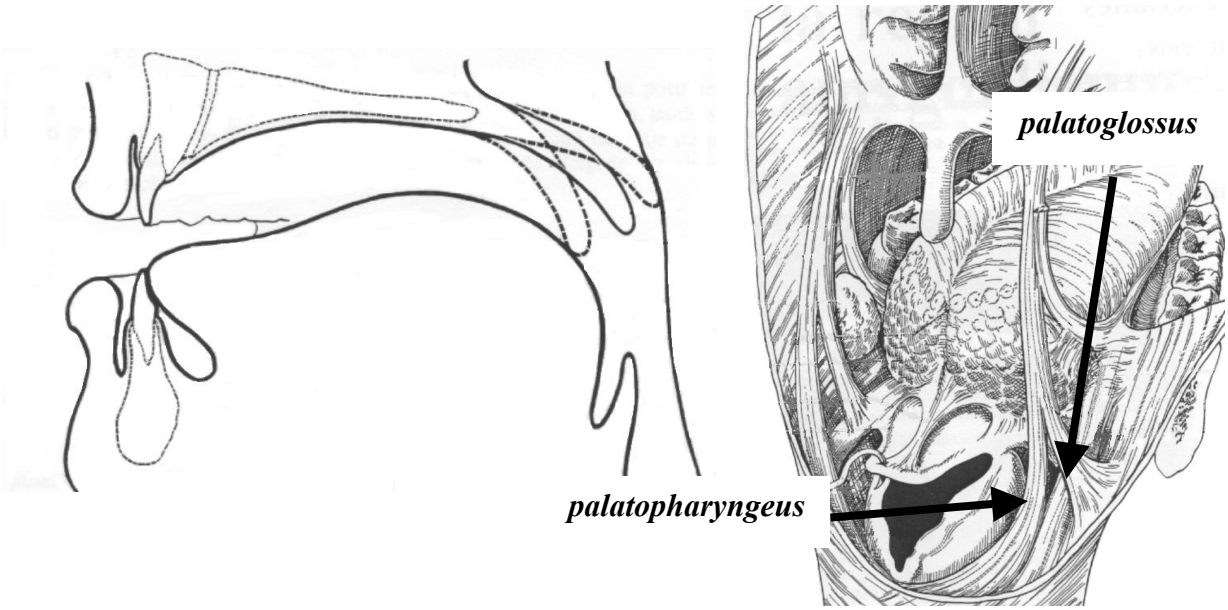
In order to ensure a common geometric representation of all the articulations for a given soft organ, a unique generic 3D surface mesh, made of triangles, is defined for each organ, and is fitted by elastic deformation to each of the 46 3D shapes of the corpus. This procedure provides thus a 3D sampling of each organ surface with the same 3D vertices for each of the 46 articulations of the corpus, which is appropriate for the statistical study and the linear modelling of the organs. The elastic deformation of the generic mesh to fit to each 3D shape extracted for the 46 articulations is computed through the matching software *TestRigid* developed at the TIMC laboratory in Grenoble (Couteau *et al.*, 2000). This reconstruction process finally provides a set of soft and rigid organ surfaces described in terms of triangular meshes having the same number of vertices, in a common reference coordinate system, for each of the 46 articulations of the corpus. This set of data forms the basis for the articulatory modelling of the subject, as will be illustrated in the next section.

4. Modelling of the velum

4.1. Brief anatomic description of the nasopharyngeal tract

The nasopharyngeal tract is made up of several organs: the velum, the nasopharyngeal walls, the nasal passages, the various paranasal sinuses and the nostrils. These organs can be classified into two types: rigid (or quasi rigid) structures (nasal passages, paranasal sinuses, nostrils) and non-rigid structures (velum, nasopharynx).

Rigid structures are geometrically complex. The septum separates the nasal passages. Each cavity is made up of a lot of thin and complex passages whose surface is very large. It is limited at one side by the nasopharyngeal port and by nostrils at the other one. Several paranasal sinuses (maxillary, sphenoid, etc.) are connected to these cavities through very narrow passages and thin membranes.
The velum and the nasopharynx constitute the non-rigid structures. The velum is mainly made of five muscles. It ensures the principal closure gesture of the velopharyngeal port through his major muscle, the *levator veli palatini* muscle, which stretches symmetrically from the medial region of the velum to the right and left Eustachian tubes. In addition, the velum is connected to the neighbouring organs: with the pharynx through the *palatopharyngeus* muscle, and with the tongue through the *palatoglossus* muscle (see Figure 6). The pharynx is principally active through the *superior*, *middle* and *inferior constrictors* muscles in sphincter action.



a) midsagittal view for three velum positions b) posterior view

Figure 6: Anatomy of the velum: midsagittal view (a) and posterior view (b). Muscles linked to the tongue and to the nasopharynx (from Bouchet *et al.*, 1980)

4.2. A first articulatory model of velum

We present in this section the results of a first attempt to establish an articulatory model of the velum based on the data obtained by the procedure described in the previous sections. The available observed data were the 3D coordinates of the 9794 vertices of the velum, of the 3369 vertices of the pharyngeal wall, and of the 4167 vertices for the tongue surface in the vicinity of the nasopharynx, for a

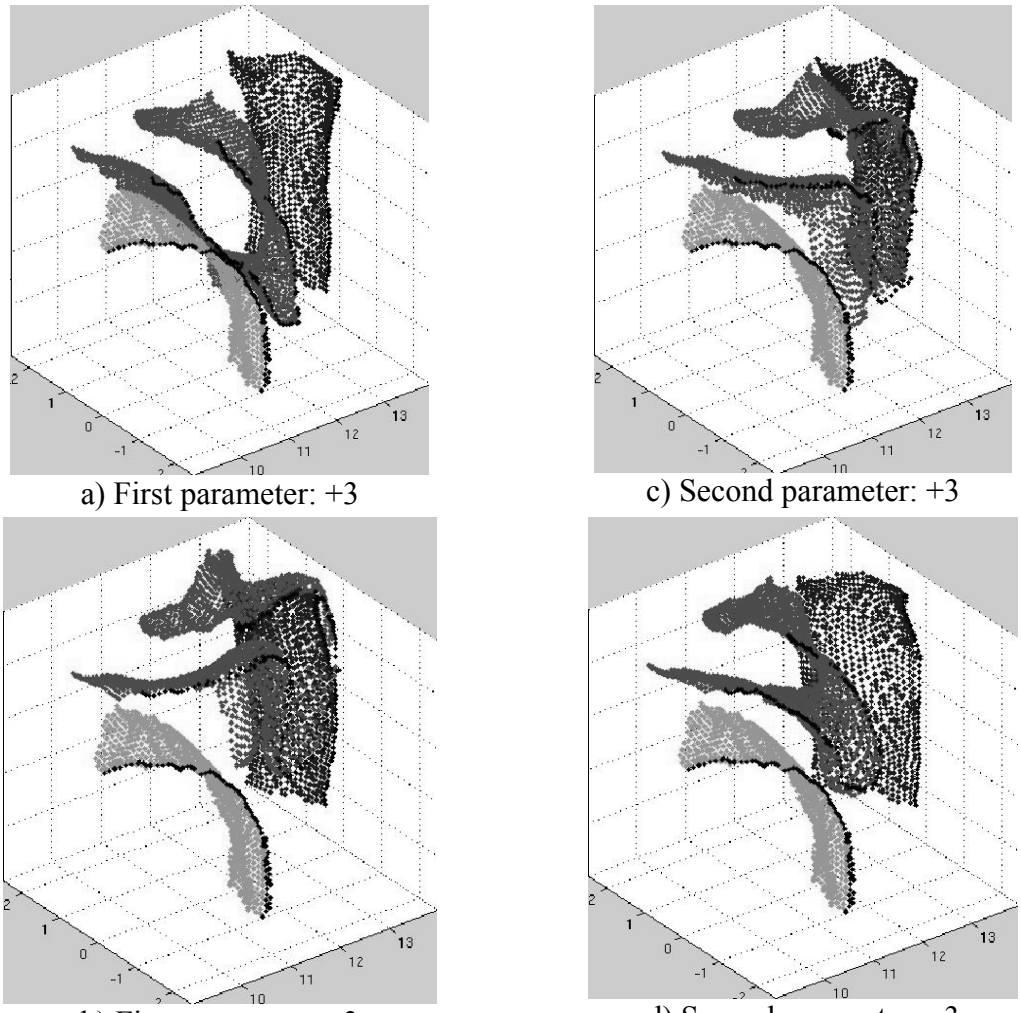
subset of 22 articulations taken among the 46 articulations of the full corpus: [a ɛ œ ɔ ɑ̃ ɛ̃ œ̃ ɔ̃ m^a m^i m^u p^a p^i p^u n^a n^i n^u t^a t^i t^u rest phrephonation].
Following the method described above, these generic meshes have been matched to the 22 targets extracted from the MRI data with a root mean square error of 0.12 cm for the part of the mesh located between the sagittal planes distant from ±1.5 cm from the midsagittal plane.

Our first attempt was to apply a *direct PCA* to the 22 observed velum shapes. The aim was to explain the maximum of the variance of the data by a minimum number of articulatory control parameters. A PCA was thus applied to these 29382 variables in order to extract a few articulatory control parameters by exploiting the correlations between neighbouring points and due to the physical continuity of the organs. Table 1 shows that almost 88 % of the cumulated variance of all the velum points is explained by only two articulatory control parameters, the first parameter explaining almost 75 % by itself. As explained below, the second parameter complements the closure mechanism of the velum and explains almost 13% of the full-cumulated variance of velum points. Table 1 gives the Root Mean Square error between the velum reconstructed from the parameters determined by PCA and the 22 shapes of velum: the error of reconstruction is less than 1 millimetre when two parameters are used.

Table 1: Explained variance and Root Mean Square error of reconstruction for the direct PCA model approach.

Parameter	Explained variance	Cumulated explained variance	RMS
First PCA parameter	74.9 %	74.9 %	0.11 cm
Second PCA parameter	12.9 %	87.8 %	0.07 cm

The figure 7 illustrates the shape of the velum (centre of the figure), of the tongue (partially represented in the bottom part of the figure) and of the nasopharynx (in the background) for two opposites values of each of the two parameters (±3). The action of the first parameter corresponds to a motion of the velum simultaneously along the vertical and horizontal directions. Considering its orientation and its prime importance for speech (Bell-Berti, 1993), the levator veli palatini muscle can be thought to be much involved in this movement. The second parameter is more related to an horizontal motion which complements the closure of the nasopharyngeal port by a back to front effect.



a) First parameter: +3 c) Second parameter: +3

b) First parameter: –3 d) Second parameter: –3

Figure 7: Representation of the velum, nasopharynx and tongue for opposite articulatory control parameters values for the first two components obtained by PCA. For clearness reasons we have represented only the right side of the three organs; the dark points are located in the midsagittal section.

In order to assess the possible correlation between velum and jaw / tongue through the palatoglossus muscle suggested in the literature (Wrench, 1999), we have used jaw height as the first articulatory control parameter of the velum, through a linear regression: we found that jaw height explains 37 % of the full cumulated variance of all the velum points (cf. Table 2). In comparison, jaw height explains 16,7% of the full variance of the tongue for the same subject and a similar corpus (Badin *et al.*, 2002).

Table 2: Variance of the velum points explained by the jaw height parameter.

Parameter	Explained variance
Jaw Height parameter	37.0 %

This explanation of the variance confirms the correlation between the complete shape of the velum and the jaw height in our corpus. The plot of velum height *vs.* jaw height for the 22 articulations of the corpus in figure 8 illustrates this correlation. The correlation coefficient between jaw and uvula heights is 0.37 for this corpus.

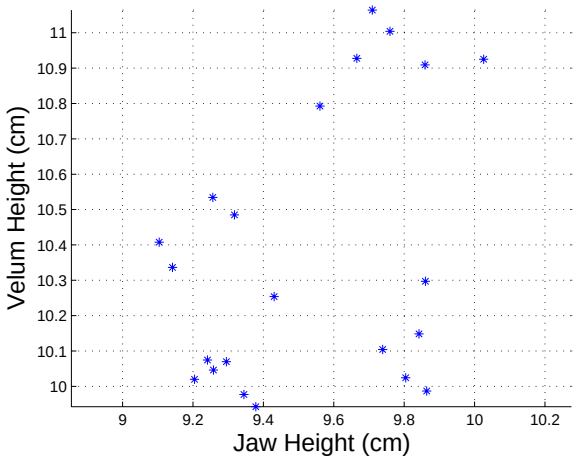
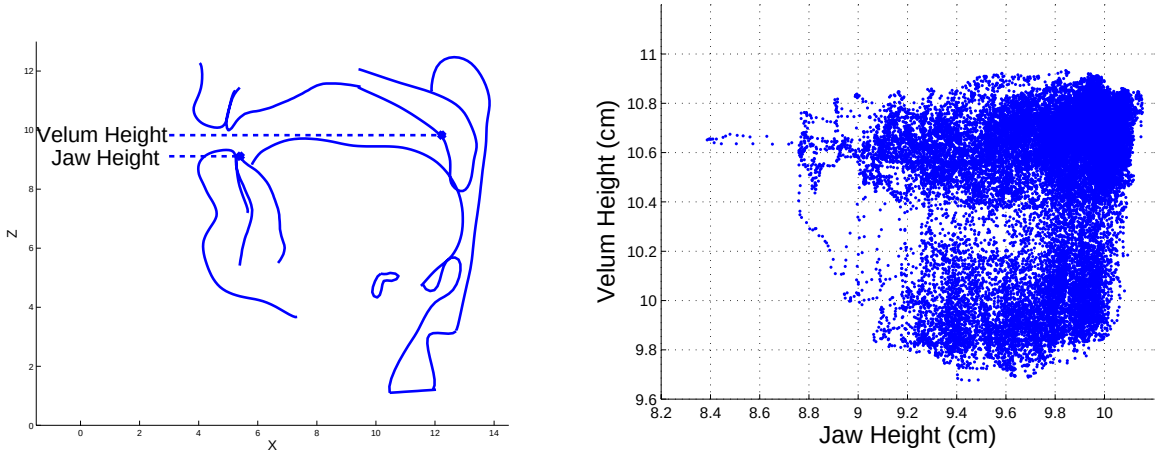


Figure 8: Velum height *vs.* jaw height for the 22 articulations. The associated correlation coefficient is 0.37.

However this correlation may be due to the content of our specific corpus of sustained articulations. In order to detect a possible corpus effect, a further investigation was carried out by means of an ElectroMagnetic midsagittal Articulograph. One of the coils of the articulograph was attached to the velum at about half way between the junction of the hard palate and of the velum and the extremity of the uvula of the subject, in the most mobile region of the velum, so as to provide an estimation of the velum movements, while another coil was attached to the lower incisors, in order to provide an estimation of jaw vertical movement (see figure 9a). A corpus of non-sense words [pVCVp] (V being one of the 14 French vowels and C one of the 16 French consonants) was recorded on the same subject. The figure 9b represents velum coil height *vs.* jaw coil height for the whole corpus (note that the complete trajectories are used). It appears clearly that the jaw and velum height parameters are not correlated. We found however a fairly strong correlation (0.88) between these parameters for the subset of the corpus constituted by the non-nasal vocalic targets; the correlation observed on the MRI corpus could be partly explained by a similar effect. Therefore, there is no reason to build a model with jaw height as first parameter.



a) Positions of the EMA coils b) Velum height *vs.* jaw height

Figure 9: Positions of the EMA coils and velum height *vs.* jaw height for the whole EMA corpus

5. Conclusion

The three-dimensional description of the velum shape for a set of 22 sustained articulations has allowed us to develop a first 3D linear articulatory model of the velum. The variance explained by the first two PCA factors for the 3D velum is not much below what we have found on the same corpus for the midsagittal contour of the velum sampled by means of a grid system similar to what Beautemps *et al.* (2001) used for the vocal tract, i.e. respectively 83.6% and 7.8% for the first and second factors. Badin *et al.* (2002) have shown that the whole 3D shape of the lips and of the tongue can be fairly well predicted from their midsagittal contours; this seems the case also for the velum and will be further checked. A 3D velum model driven by parameters measured in the midsagittal plane such as velum height for instance, could thus be developed. The next modelling step will be the determination of the area functions for both oral and nasal tracts as a function of the shapes of velum, tongue and pharynx wall, and thus the determination of the acoustical characteristics of the complete tract. A longer-term objective is finally to extend all this process to the complete 3D vocal and nasal tracts in order to build a complete 3D articulatory model of speech.

Acknowledgements

The medical images have been acquired at the Radiology Department of the Grenoble Regional University Hospital, in collaboration with the research unit INSERM/UJF 594 (Christoph Segebarth). We acknowledge the help of the

GMCAO team at TIMC for the use of the matching software *TestRigid* (Yohan Payan, Franz Chouly, Maxime Bérar). Finally we would like to thank Véronique Delvaux and Pascal Perrier for their helpful comments on the first version of the manuscript.

References

Amelot A., Crevier-Buchman L. & Maeda S. (2003) Observations of the velopharyngeal closure mechanism in horizontal and lateral directions from fiberscopic data. *Proc. of the 15th ICPhS*, Barcelona, Spain: 3021-3024.

Badin, P., Bailly, G., Raybaudi, M. & Segebarth, C. (1998). A three-dimensional linear articulatory model based on MRI data. In *Proceedings of the Third ESCA / COCOSDA International Workshop on Speech Synthesis*, Jenolan Caves, Australia, December 1998, 249-254.

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C. & Savariaux, C. (2002) Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics,* 30(3): 533-553.

Badin, P., Bailly, G., Elisei, F. & Odisio, M. (2003) Virtual Talking Heads and audiovisual articulatory synthesis (Invited talk at the symposium "*Articulatory synthesis. Advances and prospects"*). In *Proceedings of the 15th International Congress of Phonetic Sciences* (M.-J. Solé, D. Recasens & J. Romero, editors), Barcelona, Spain, 1: 193-197

Beautemps, D., Badin, P. & Bailly, G. (2001). Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. *Journal of the Acoustical Society of America,* 109(5): 2165-2180.

Bell-Berti, F. (1993). Understanding velic motor control: studies of segmental cotext. In Huffman & Krakow (Eds). *Phonetics and phonology. Nasals, Nazalisation and the Velum*. Vol 5. Academic Press Inc, 63-85.

Bouchet, A., & Cuilleret, J. (1980). Anatomie topographique, descriptive et fonctionnelle (la face, la tête et les organes des sens). *Edition Simep*.

Couteau, B., Payan, Y. & Lavallée, S. (2000). The mesh-matching algorithm: an automatic 3D mesh generator for finite element structures. *Journal of Biomechanics,* 33: 1005-1009.

Dang, J., Honda, K. & Suzuki, H. (1994) Morphological and acoustical analysis of the nasal and the paranasal cavities. *Journal of the Acoustical Society of America,* 96: 2088-2100.

Feng, G. & Castelli, E. (1996) Some acoustic features of nasal and nasalized vowels: a target for vowel nasalisation. *Journal of the Acoustical Society of America,* 99(6): 3694-3706.

Huffman, M. K. & Krakow, R. A. (1993). Phonetics and phonology. Nasals, Nazalisation and the Velum. Vol 5. Academic Press Inc.

Rossato, S., Badin, P. & Bouaouni, F. (2003) Velar movements in French: an articulatory and acoustical analysis of coarticulation. In *Proceedings of the 15th International Congress of Phonetic Sciences* (M.-J. Solé, D. Recasens & J. Romero, editors), Barcelona, Spain: 3141-3144.

Takemoto, H., Kitamura, T., Nishimoto, H. & Honda, K (2004). A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions. In *Acoustical Science and Technology*, 25(6):468-474

Teixeira, A., Vaz, F. & Príncipe, J.C. (2000) Nasal vowels following a nasal consonant. In *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Kloster Seeon, Germany: 285-288.

Wrench, A.A. (1999). An investigation of sagittal velar movement and its correlation with lip, tongue and jaw movement. In *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, United States: 435-438

Zemlin W.R. (1968). Speech and Hearing Sciences. Anatomy and Physiology. *Prentice-Hall, Inc.*

Open Quotient (EGG) measurements of young and elderly voices: Results of a production and perception study

Ralf Winkler
Walter Sendlmeier
Dept. of Speech Communication and Phonetics, Technical University, Berlin, Germany

This paper presents the results of Open Quotient measurements in EGG signals of young (18 to 30 year old) and elderly (59 to 82 year old) male and female speakers. The paper further presents quantitative results on the relation between the OQ and the perception of a speaker's age. Higgins & Saxman (1991) found a decreased OQ_{EGG} with increasing age for females, whereas the OQ_{EGG} in sustained vowel material increased for males as the speakers age increased. In Linville (2002), however, the spectral amplitudes in the region of F_0 (obtained by LTAS-measurements of read speech material) increased with increasing age independent of gender; this could be interpreted indirectly as an increasing OQ. We measured the OQ_{EGG} not only for sustained vowels, but also in vowels taken from isolated words. In order to analyse the relation between breathiness in terms of an increased OQ and the mean perceived age per stimulus a perception test was carried out in which listeners were asked to estimate speaker's age based on sustained /a/-vowel stimuli varying in vocal effort (soft - normal - loud) during production. The results indicated the following: (i) The decreased OQ for elderly females originally found by Higgins & Saxman is not apparent in our data for sustained /a/-vowels. For our female speakers no significant difference between the OQ of young and old speakers was found; for elderly males, however, we also found an increasing OQ with increasing age.(ii) In addition, a statistically significant increased OQ_{EGG} occurs for the group of the elderly males for the vowels from the word material. (iii) Our results show a strong positive relation between perceived age and OQ in male voices. Regarding (i) and (ii), at least the male speaker's voice becomes more breathy as age increases. Considering (iii), increased breathiness may contribute to the listener's perception of increased age.

1. Introduction

Age recognition from voice is possible with reasonable accuracy (e.g. Ptacek & Sander, 1966). Shipp & Hollien (1969) suggest that there is an identifiable set of measurable parameters that contribute to the perception of age from voice. Several features such as F_0 (pitch), jitter, shimmer and spectral tilt (voice quality) as well as temporal features like segment durations and pauses (speech tempo/timing) have been identified as markers of chronological and perceived age (for details see Linville, 2000).

Asked to describe the most salient features of aged voices, listeners reported increased breathiness among other features, at least for male voices (Hartman & Danhauer, 1976; Hartman, 1979). Winkler et al. (2003) qualitatively analysed the EGG signals and the Long-Term Average Spectra (LTAS) of sustained vowels and found that in males increased perceived age was linked to more sinusoidal EGG signals and considerable less spectral energy between two and four kHz. For female speakers this tendency did not occur. The increased spectral tilt values in males resulting from the energy drop around 2 kHz are another characteristic of breathy voice.

A stronger perceived breathiness is linked physiologically to an increased Open Quotient (e.g. Klatt & Klatt, 1990), as was indirectly measured by Higgins & Saxman (1991) for EGG signals of sustained /a/-vowels. They found decreased OQ_{EGG} values with increasing chronological age for females, whereas the OQ_{EGG} increased for males as the speakers' chronological age increased. The increase of OQ values in male voices could be interpreted as a consequence of laryngeal changes with increasing age. Although laryngeal degeneration due to increased age in females seems to occur to a lesser extent (e.g. Honjo & Isshiki, 1980), the decrease of OQ values in elderly female voices could not be explained.

According to Hanson et al. (2001) it is expected that as OQ increases, the glottal waveform more closely approximates a sinusoid of the fundamental frequency. Therefore in the frequency domain the amplitude of the first harmonic increases relative to the amplitudes of the higher harmonics. In Linville (2002) the spectral amplitudes in the region of F_0 (obtained by LTAS-measurements) increased with increasing age, independent of gender, which could be interpreted indirectly as increasing OQ values for both females and males in their read speech material.

Although results were derived from sustained vowels and read speech by means of analysing EGG signals and LTAS respectively, results from the literature regarding Open Quotient and the voices of chronologigal old speakers are contradictory.

In a study of Debruyne & Decoster (1999) possible acoustic correlates of age judgements based on sustained vowel material were analysed. Vowels unani-

mously judged as young or old were acoustically analysed. Among other distinctions between the vowels judged as young or old they found a smaller difference between the first and the second harmonic in the spectrum (H1-H2) for the vowels of the speakers judged unanimously as old. This difference was statistically significant for female voices only. Their findings could be indirectly interpreted as decreased OQ as a marker of perceived age. Although Debruyne & Decoster did not find a single acoustic feature that could explain the age judgements of their listeners, the decreased OQ especially for the elderly male voices do not match the results of Hartman & Danhauer (1976) and Hartman (1979).

The aim of the first part of this study was to measure OQ_{EGG} values not only for sustained vowels, but also for vowels taken from isolated words. With the extended stimulus material we are able not only to replicate the results of Higgins & Saxman but also to analyse whether the OQ_{EGG} values depend on stimulus type. The aim of the second part of this study was to quantify the relation between OQ_{EGG} values and the perception of a speaker's age. For this purpose a perception test was carried out in which listeners were asked to estimate the speaker's age based on sustained vowel material varying in vocal effort.

2. Open Quotient (EGG) and Chronological Age

2.1. Recordings

Our data consist of the sustained German vowel /a/ and /a/-vowels taken from words spoken one by one from a list of words. For the sustained vowel material the speakers were instructed to produce the vowel as stable as possible with self-selected comfortable pitch and loudness. For the words from the wordlist the speakers were instructed to speak the word as naturally as possible. The whole list contained 23 different words; the /a/-vowels used in this study were taken from two words. To account for the possible influence of word stress on voice quality two vowel versions were used: One vowel originated from the unstressed syllable of "l[a]wine" (avalanche), the second version from the stressed syllable in "kohlr[a]bi" (kohlrabi). From each sustained vowel, an interval of 500 ms centred on the midpoint of the vowel was selected for the analysis of the Open Quotient. In the word material the whole duration of the vowel segment was used to calculate the mean OQ_{EGG}. The recordings consist of an audio channel and a channel with the EGG signal; in this study only the laryngographic signals were used.

Table 1: Descriptive statistics (number of speakers, mean chronological age, standard deviation, minimum and maximum chronological age in years) for the speakers used in this study.

	Female					Male				
Group	**N**	**Mean**	**SD**	**Min**	**Max**	**N**	**Mean**	**SD**	**Min**	**Max**
					/a/ (sustained)					
Young	9	26.00	1.80	24	28	13	25.62	4.21	18	33
Elderly	9	69.33	6.67	61	82	9	63.56	3.32	59	70
					/a/ (words)					
Young	11	26.27	1.85	24	29	12	25.50	4.38	18	33
Elderly	16	69.56	5.85	61	82	12	66.75	7.44	59	82

We recorded 52 speakers: 27 female speakers and 25 males. The corpus of female data consists of 11 young (24-29 years old) and 16 elderly speakers (61-82 years of age). Regarding male speakers, our corpus includes recordings of 13 young (18-33 years old) and 12 elderly males (59-82 years of age). The details of the age distribution of the speakers can be found in Table 1. The number of speech items per stimulus type differs because not all speech items were available for all stimulus types and speakers. In addition abnormal EGG waveforms displaying any signs of voice disorders were excluded from the analysis. Because of the possible relationship between age and general health particular attention was taken to record healthy elderly speakers judged by a physician.

2.2. Methods

Derived Measures: Two alternative measurements of the Open Quotient (EGG) were taken. The time instant of glottal closure is well detectable as a positive peak in the time derivative (DEGG) of the EGG signal, but there is no agreement on defining the time instant of glottal opening. In the past, the minimum of DEGG has been used as marker for the glottal opening (Henrich et al., 2004), but often there does not exist any clear minimum during a glottal cycle. Even if there is a clear minimum, there is no agreement on whether this is actually the instant of glottal opening or not. An alternative approach to analyse the vibrational patterns of a glottal cycle in an EGG signal is to define a time instant corresponding to the point of intersection between the (falling edge of the) EGG signal and a threshold line. With the threshold intersection criterion different values have been placed at points representing 25%, 30%,40%, 50% and 75% of the signal peak-to-peak amplitude (Higgins & Saxman, 1991;

Orlikoff et al., 1997; Sapienza et al., 1998). There is an agreement on the fact that the results within a study do not strongly depend on the threshold value as long as a constant value is consistently used.
Generally the OQ is defined as the ratio between the duration of the phase where the glottis is open and the whole duration of the glottal cycle, multiplied by 100 to express the values in percent. In this study two OQ values were calculated: OQ1 uses the minimum of DEGG as the time instant of glottal opening, whereas OQ2 uses the threshold intersection criterion (30%) to define the time instant of glottal opening.

Statistical Analysis: One key question of this study was whether the OQ_{EGG} rises or falls with increasing chronological age of the speakers. Because the number of speakers is rather small we decided to use a non-parametric test. In order to compare the means of OQ1 and OQ2 between the group of young and old speakers, the Kolmogorov-Smirnov test for two independent samples was applied for the sustained vowel material as well as for the stressed and unstressed vowels from words. Because the process of aging as well as the voice characteristics generally differ between males and females, statistical analyses were arranged for each sex separately.

*2. 3. Result*s

The means of OQ1 and OQ2 were computed for every speaker and speaking condition. The mean value of one token represents the average value over every glottal cycle of the EGG signal under investigation. After computation of the means of the single tokens these means were again averaged to obtain group means for young and elderly speakers. The group means broken according to sex and age group can be found in Table 2.

Female speakers: For the sustained vowels the OQ_{EGG} values were greater for the older than the younger female speakers. The OQ1 group mean, where values originated from the time derivation of the EGG, rises from 52.12% (young) to 54.13% (elderly). The OQ2 group mean increases from 44.07% to 48.12%. Both measurements of Open Quotient show a tendency of increased breathiness with advanced age. Both differences were not statistically significant. Therefore, the increase can only be interpreted as a trend.
For the unstressed vowels from the word material the OQ1 group mean decreases from the young (56.34%) to the elderly (54.05%) speakers. The group means of OQ2 - 49.56% for the young and 49.36% for the elderly speakers - are nearly identical. The decrease of OQ1 as well as the very small decrease of OQ2 were not statistically significant.

Table 2: Means and standard deviations for OQ1 and OQ2 separated by sex, and age group.

	Females				Males			
	Young		Elderly		Young		Elderly	
Condition	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
OQ1 (Minimum in DEGG) [%]								
/a/ (sustained)	52.12	8.10	54.13	6.21	49.09	6.81	53.97	4.54
/a/ (unstressed)	56.34	9.55	54.05	7.74	48.17	6.58	59.49	8.47
/a/ (stressed)	54.06	6.96	53.01	6.44	47.79	6.69	56.75	6.65
OQ2 (Threshold criterion 30%) [%]								
/a/ (sustained)	44.07	8.36	48.12	8.25	48.91	5.43	50.92	5.41
/a/ (unstressed)	49.56	6.99	49.36	7.66	47.91	4.87	55.03	8.90
/a/ (stressed)	47.69	7.39	49.27	7.67	47.17	3.70	51.40	7.31

For the stressed /a/-vowels the OQ1 group mean of the young females (54.06%) decreases to 53.01% for the group of the elderly females. In contrast, the OQ2 group mean increases from 47.69% for the young group to 49.27% for the group of elderly females. Because the decrease of OQ1 as well as the increase of OQ2 was not statistically significant, the contradiction should not be overestimated. There is neither a trend for an increasing nor a trend for a decreasing OQ in stressed vowels.

Male speakers: For the sustained vowels the OQ1 group mean increases from the group of the young males (49.09%) to the group of elderly speakers (53.97%). The increase is apparent in the OQ2 measurement as well. The OQ2 group mean increases from 48.91% for the young to 50.92% for the elderly males. Although the group mean difference of OQ1 is slightly higher, this difference as well as the difference of OQ2 is not statistically significant.

The OQ1 group mean increases in the case of unstressed vowels from 48.17% for the young to 59.49% for the elderly males. A somewhat smaller difference is apparent between the OQ2 values. The group mean increases from 47.91% for the young to 55.03% for the elderly male speakers. Both differences are remarkably high, but only the group mean difference of OQ1 is statistically significant ($p<0.05$).

Similar in appearance are the OQ values for the stressed vowels from words. The OQ1 group mean increases from 47.79% for the young to 56.75% for the elderly males. The OQ2 group mean also increases from 47.17% for the young to 51.40% for the elderly males. Similar to the unstressed vowel versions, both

differences are high, but only the group mean difference of OQ1 is statistically highly significant (p=0.01).
For the sustained vowels results from both kinds of OQ measurements lead to the same trend. Regardless of the sex of the speakers, although not statistically significant, there is a weak trend for a higher OQ for the elderly compared to the younger speakers. The difference between the group of the younger and the older speakers is around 2% and 5%. Results from the vowels of the word material of the female speakers do not lead to any trend. In contrast, results from male speakers are clear. Both OQ measurements show an increase from the young to the elderly males. For the OQ1 measurement the difference for the vowels in unstressed as well as in stressed syllables is statistically significant. Although the differences for the OQ2 measurement are not statistically significant, the difference of approximately 7% for the unstressed vowels is remarkably high.

3. Open Quotient (EGG) and Perceived Age

Results from Hodge (2001) as well as from the first part of this study suggest that the Open Quotient consistently increases in the voice of elderly speakers, at least for males. Therefore a perception test was carried out to analyse whether the Open Quotient (derived by means of EGG) is a potential correlate of the perception of a speaker's age, at least for the distinction between the young and the elderly speakers. In order to expand the range of the acoustic correlates of perceived voice qualities of the stimulus set, special versions of the sustained /a/-vowel were used. According to e.g. Dromey et al. (1992), Hodge et al. (2001) and Sundberg et al. (2005) measures of vocal function strongly depend on vocal loudness. To capture a wide varity of different vocal functions, in addition to the sustained /a/-vowels from the first part of the study, /a/-vowels with systematic variation in vocal effort during production were included in the stimulus set. For the perception test all stimuli were normalized with respect to amplitude so that the task of the speaker was masked and the listener's focus was shifted to variations in perceived voice quality only.

3.1. Stimuli and Listeners

For the purpose of this perception experiment speakers were asked to produce two versions of the German vowel /a/. One instruction was to sustain the vowel as softly as possible but without whispering; the second instruction was to sustain the vowel as loudly as possible but without shouting. Three to five trials dependent on success of the speaker were recorded. For the perception experiment a 500 ms interval of the most stable token centred on the midpoint of

the vowel were selected. Furthermore 500 ms intervals of the vowels used in the first part of the study were added to the stimulus material.
The stimuli were produced essentially by the speakers described in subsection 2.1. Four female and two male speakers, neither chronological young nor old, were added to the corpus to maximize the number of speakers. Thus the full stimulus set consists of 60 stimuli from 22 female, and 72 stimuli from 24 male speakers, three stimuli each. For the female speakers four stimuli of the loud, and two stimuli of the soft production were excluded because of serious deviations regarding naturalness judged by the experimenter. The intensity of all stimuli was (peak-)normalized to suggest to the listeners a variation in voice quality features instead of an intensity variation.
Twenty subjects (six male, 14 female) listened to the female voice samples and 40 subjects (20 male/ 20 female) took part in the perception test with the male voices. For the experiment with the female voices the mean age of the listeners was 35.35 years (SD = 7.22 years). The mean age of all 40 listeners in the experiment with the male voices was 26.23 years (SD = 4.39 years). All listeners claimed not to have or know of any hearing problems either at present or in the past.
Listeners seated in front of a computer screen wearing earphones. After the presentation of one item the task of the listener was to rate the chronological age of the speaker immediately. In case of female voice stimuli the listeners rated on a five years scale from 20 to 95 years of age. Listeners rated the chronological age of the speaker in years directly for the male voices. One whole presentation consisted of three repetitions of each vowel signal (500 ms each) separated by a silent interval (500 ms). Subjects were able to listen to a stimulus only once. They were instructed to judge immediately after listening. The test started with a pre-presentation of five items to enable the listeners to adapt to the speed of the perception test.

3.2. Methods

The values of the mean perceived age per stimulus resulting from the perception experiment, that were computed by averaging the single values across all listeners, are quasi-continuous. These values were correlated with the values of the Open Quotient measured in the first part of this study. The single correlations for the condition of soft, normal or loud production level were not analysed in the context of this study because the task was given only to verify whether the OQ_{EGG} could be used as a predictor for perceived age.

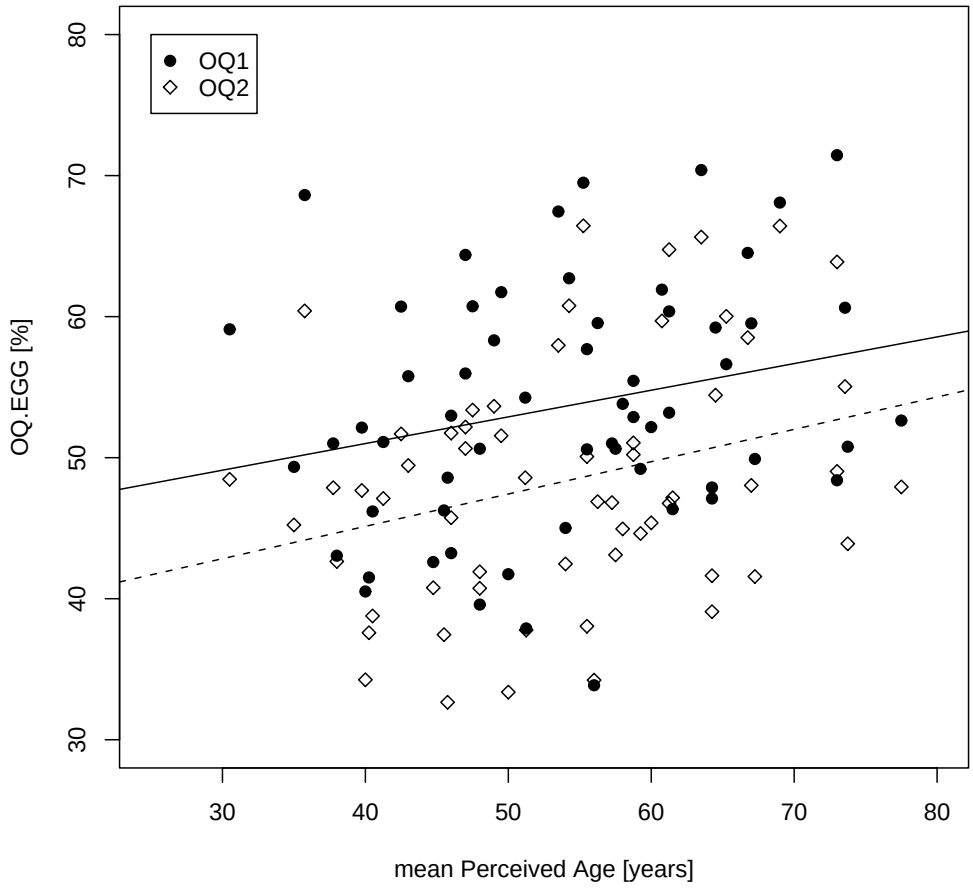


Figure 1: Relation and straight line fit between mean perceived age per stimulus and two versions of the Open Quotient (EGG) for female speakers.

3.3. Results

Female speakers: The mean perceived age per vowel stimulus slightly increases with increasing OQ1 as well as increasing OQ2 values. The relation between the mean perceived age per stimulus and the OQ_{EGG} values is depicted in Figure 1. The Pearson correlation coefficient between the mean perceived age and OQ1 is 0.24 (p=0.06, two sided). Between the mean perceived age and OQ2 the Pearson correlation coefficient is 0.30 and statistically significant (p=0.02).
From Figure 1 it appears that the OQ_{EGG} values scatter remarkably. There are quite a number of vowel stimuli with OQ_{EGG} values above 65% which (with one exception) were judged between 52 and 75 years old by the listeners. The stimuli with an OQ_{EGG} below 40% were judged between 40 and 55 years. The stimuli judged older than 40 years use nearly the whole range of OQ values

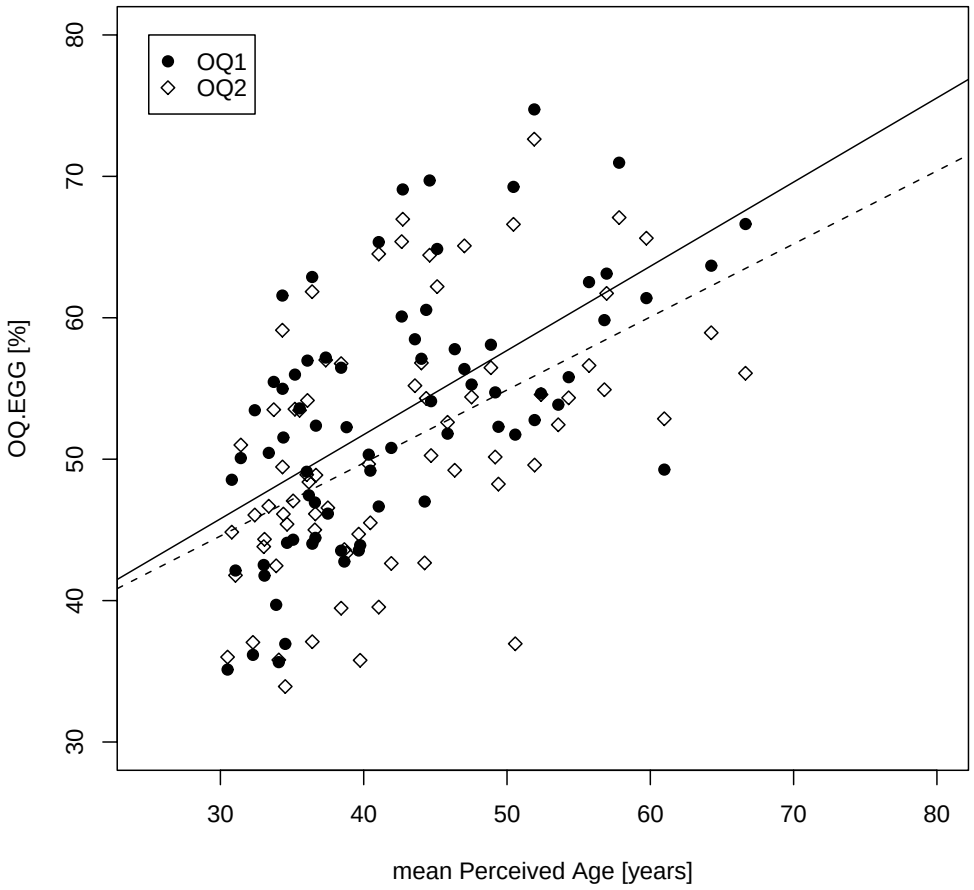


Figure 2: Relation and straight line fit between mean perceived age per stimulus and two versions of the Open Quotient (EGG) for male speakers.

between 30% and 70%. It seems that other features or combination of features guide the listener in judging a female speaker's age.

Male speakers: Similar to the pattern found for female voices the judged age increases as OQ values increase. The Pearson correlation coefficient between the mean perceived age per vowel stimulus and the OQ1 values is 0.59 (p=0.00), and 0.51 (p=0.00) between the mean perceived age and the OQ2 values. The relation between mean perceived age and the OQ values per stimulus is given in Figure 2. Both Pearson correlation coefficients are statistically highly significant.

From Figure 2 it appears that there is a strong relation between the mean perceived age and the OQ values. This strong relation is represented in the high correlation coefficients. For male voices there is a more systematic pattern in contrast to the female voices. Vowel stimuli judged as young (approximately 30

to 40 years old) have OQ values smaller than 65%, whereas stimuli judged older than approximately 50 years have OQ values between 50% and 75%.
Results from the second part of the study suggest that the relation between the Open Quotient and the mean perceived age per vowel stimulus is stronger for male than for female speakers. Although at least one of the correlations between OQ and mean perceived age in female voices is statistically significant, the two coefficients are rather small. Neither a consistent decrease nor a consistent increase of OQ_{EGG} is apparent in our data of female voices regarding the chronological age of the speakers; the relation between OQ and mean perceived age is only weak. The OQ_{EGG} for the group of elderly men is consistently higher in comparison to the group of the young speakers.

4. Summary & Conclusions

The increase of OQ_{EGG} for male elderly voices is apparent in our data; however, we did not find the decrease in OQ_{EGG} values for elderly female speakers originally found by Higgins & Saxman (1991) for sustained vowels. We further found a statistically significant increase in OQ_{EGG} values for the group of elderly male speakers in comparison to younger males for the vowels originated from words. These results are in line with findings of Linville and are plausible in terms of laryngeal changes which occur with advanced chronological age. The finding of significant differences between young and old males in stressed and unstressed vowels from words suggests that sustained vowels could be more affected by compensatory phonatory behaviour in terms of adjusted vocal effort or strain than the word material. Furthermore, our results for the female voices match the results of Sapienza & Dutka (1996), who did not find any significant changes in their amplitude-based airflow measurements related to the chronological age of their female speakers.
Regarding the perception of age, in case of male voices our results agree with the results of Hartman & Danhauer (1976) and Hartman (1979); and correlations give quantitative support to the impressions the listeners supplied in the two cited studies. For the female voices our results do not match the results of Debruyne & Decoster (1999) who found a significantly decreased OQ for vowels unanimously judged as old. In our material the OQ_{EGG} seems to have only minor importance in characterising the voices of chronological old female speakers as well as perceptually old female voices. Our results do not suggest any dependency between the stimulus type and the mean OQ_{EGG} values.
Perceptual judgments most likely depend on multiple acoustic cues. Furthermore human listeners seem to differ in the relative importance they give to different aspects of vocal quality. This is suggested by Kreiman et al. (1994) for the description of pathological voices by means of breathiness and hoarseness

ratings of expert listeners. Because the recognition of age by a group of listeners is possible to some extent, Shipp & Hollien (1969) stated that there is an identifiable set of acoustic parameters that contribute to the perception of age from voice. In our study listeners seem to be influenced in their judgements about the male elderly voice by increased OQ values. This could explain the relatively strong relation between the OQ_{EGG} and the mean perceived age in the second part of this study.
Finally it should be noted, however, that the maximum variation in voice quality features contained in the stimuli of the perception test usually does not appear in everyday speech. Other vocal qualities such as fundamental frequency which contribute to the perception of age, have not been analysed in the speech material of this study so far. In a next step it will be explored to what extent this feature is of perceptual relevance compared to other voice quality features.

Acknowledgements

The authors would like to thank Laura Koenig and Susanne Fuchs for many helpful comments and suggestions on an earlier version of this paper. We are further grateful for the financial support provided by the Deutsche Forschungsgemeinschaft (German Research Fund, SE 462/5-1).

References

Debruyne, F. & Decoster, W. (1999): Acoustic differences between sustained vowels perceived as young or old. *Logopedics Phoniatrics Vocology*, 24(1), 1-5.

Dromey, C., Stathopoulos, E.T. & Sapienza, C.M. (1992): Glottal airflow and electroglottographic measures of vocal function at multiple intensities. *Journal of Voice*, 6(1), 44-54.

Hanson, H.M., Stevens, K.N., Kuo, H.J., Chen, M.Y. & Slifka, J. (2001): Towards models of phonation. *Journal of Phonetics*, 29, 451-480.

Hartman, D.E. & Danhauer, J.L. (1976): Perceptual features of speech for males in four perceived age decades. *Journal of the Acoustical Society of America*, 59(3), 713-715.

Hartman, D.E. (1979): The perceptual identity and characteristics of aging in normal male adult speakers. *Journal of Communication Disorders*, 12, 53-61.

Henrich, N., d'Alessandro, C., Doval, B. & Castellengo, M. (2004): On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *Journal of the Acoustical Society of America,* 115(3), 1321-1332.

Higgins, M. & Saxman, J. (1991): A comparison of selected phonatory behaviours of healthy aged and young adults. *Journal of Speech and Hearing Research*, 34, 1000-1010.

Hodge, F.S., Colton, R.H. & Kelley, R.T. (2001): Vocal intensity characteristics in normal and elderly speakers. *Journal of Voice*, 15(4), 503-511.

Honjo, I. & Isshiki, N. (1980): Laryngoscopic and voice characteristics of aged persons. *Archives of Otolaryngology*, 106, 149-150.

Klatt, D.H. & Klatt, L.C. (1990): Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America,* 87(2), 820-857.

Kreiman, J., Gerratt, B.R. & Berke, G. S. (1994): The multidimensional nature of pathologic voice quality., *Journal of the Acoustical Society of America,* 96(3), 1291-1302.

Linville, S.E. (2000): The Aging Voice. In: *Voice Quality Measurement*, R. Kent and M. Ball, Eds., 359-376, Singular Thomson Learning, San Diego.

Linville, S.E. (2002): Source characteristics of aged voice assessed from long-term average spectra. *Journal of Voice*, 16, 472-479.

Orlikoff, R.F., Baken, R.J. & Kraus, D. H. (1997): Acoustic and physiologic characteristics of inspiratory phonation. *Journal of the Acoustical Society of America,* 102(3), 1838-1844.

Ptacek, P. & Sander, E. (1966): Age recognition from voice. *Journal of Speech and Hearing Research*, 9, 273-277.

Sapienza, C.M. & Dutka, J. (1996): Glottal airflow characteristics of woman's voice production along an aging continuum. *Journal of Speech and Hearing Research*, 39, 322-328.

Sapienza, C.M., Stathopoulos, E.T. & Dromey, C. (1998): Approximations of open quotient and speed quotient from glottal air flow and EGG wave forms: Effects of measurement criteria and sound pressure level. *Journal of Voice*, 12, 31-43.

Shipp, T. & Hollien, H. (1969): Perception of the aging male voice. *Journal of Speech and Hearing Research*, 12, 703-711.

Sundberg, J., Fahlstedt, E. & Morell, A. (2005): Effects on the glottal voice source of vocal loudness variation in untrained female and male voices. *Journal of the Acoustical Society of America,* 117(2), 879-885.

Winkler, R., Brückl, M. & Sendlmeier, W. (2003): The aging voice: an acoustic, electroglottographic and perceptive analysis of male and female voices. In: *Proceedings of the 15th ICPhS*, Barcelona, Spain.