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.

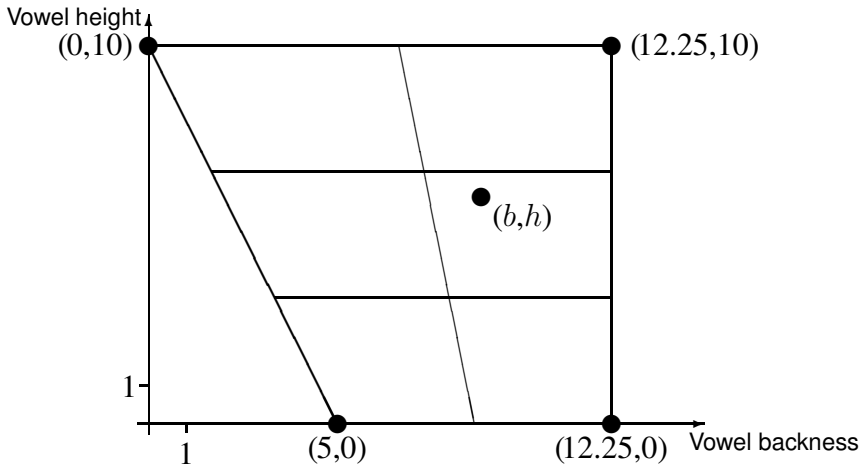

Figure 1: Dimensions of the Cardinal Vowel diagram used in our perception tests and in the vowel quality prediction formulae.

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

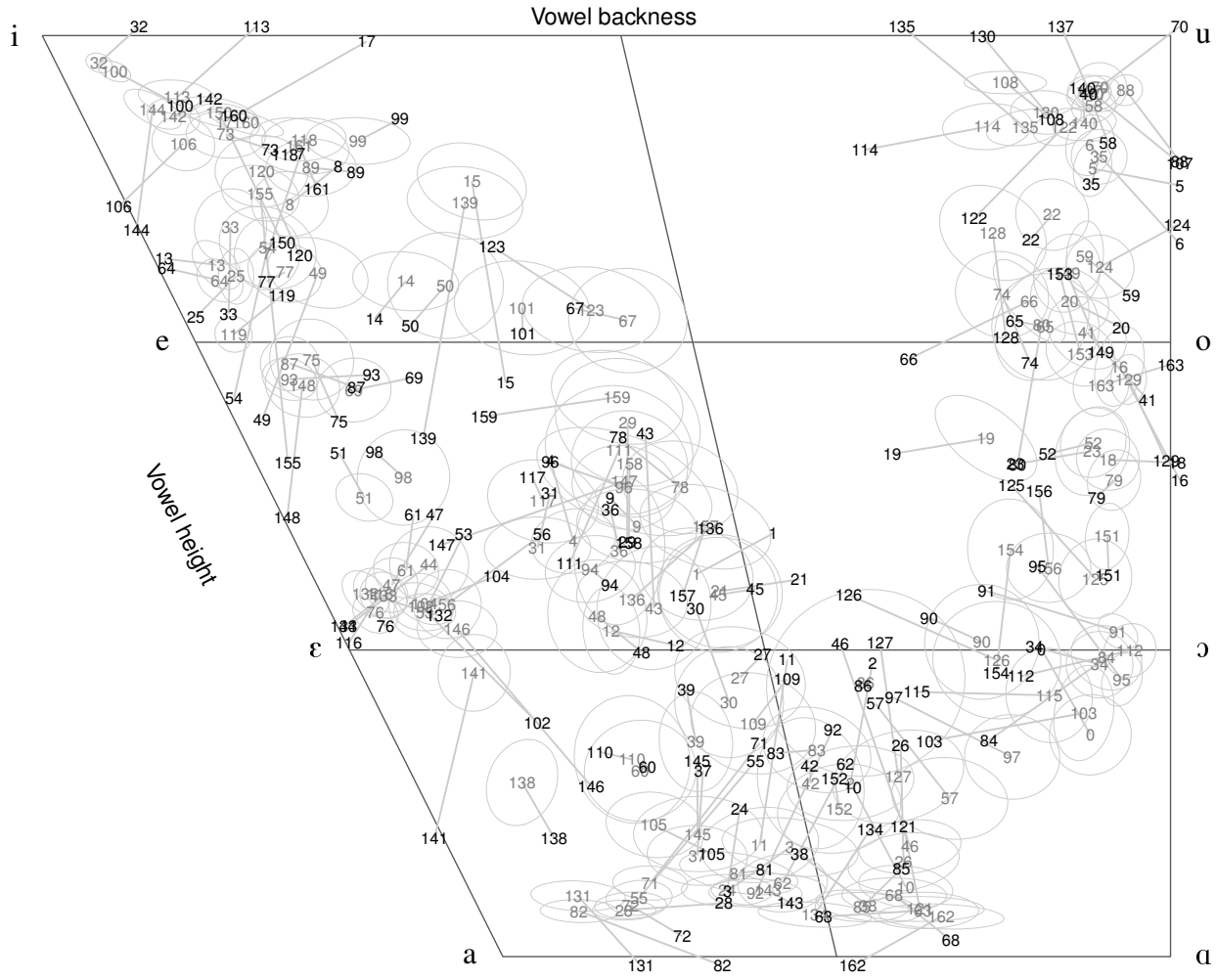

Figure 2: Mean perception results and 90% confidence ellipses (*light colors*) of 164 German monophthong stimuli and their predicted positions (*dark numbers*).

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.

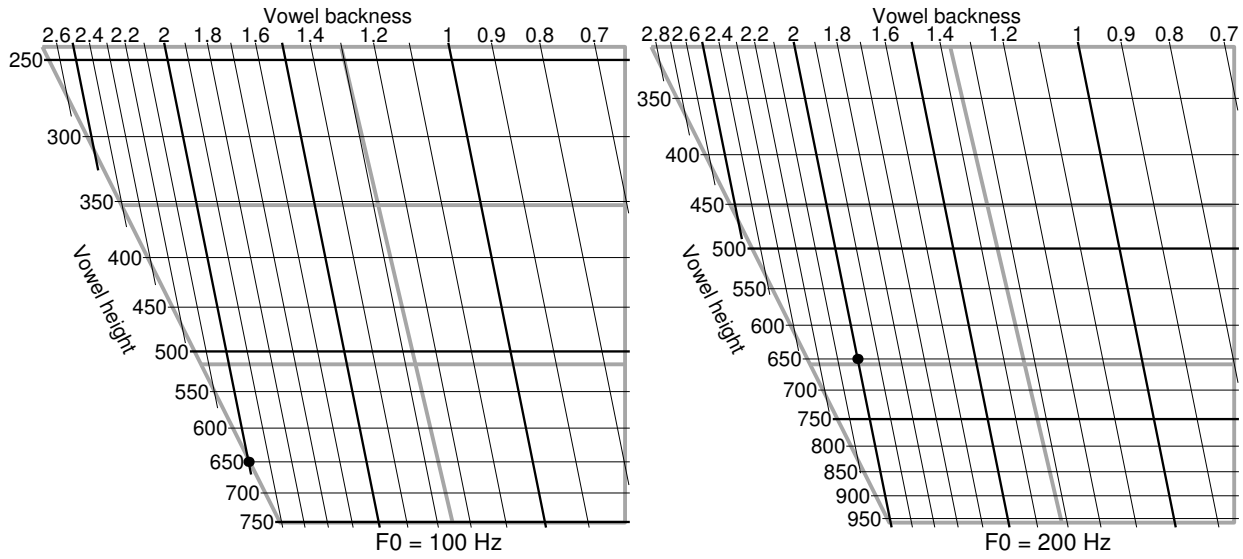

Figure 3: Relationship between formant frequencies (vertical: F1 in Hz, horizontal: F2 in kHz) and the Cardinal Vowel diagram estimated by the refined model for two selected F0 frequencies. Both marked vowels have the same F1 and F2 frequencies.

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

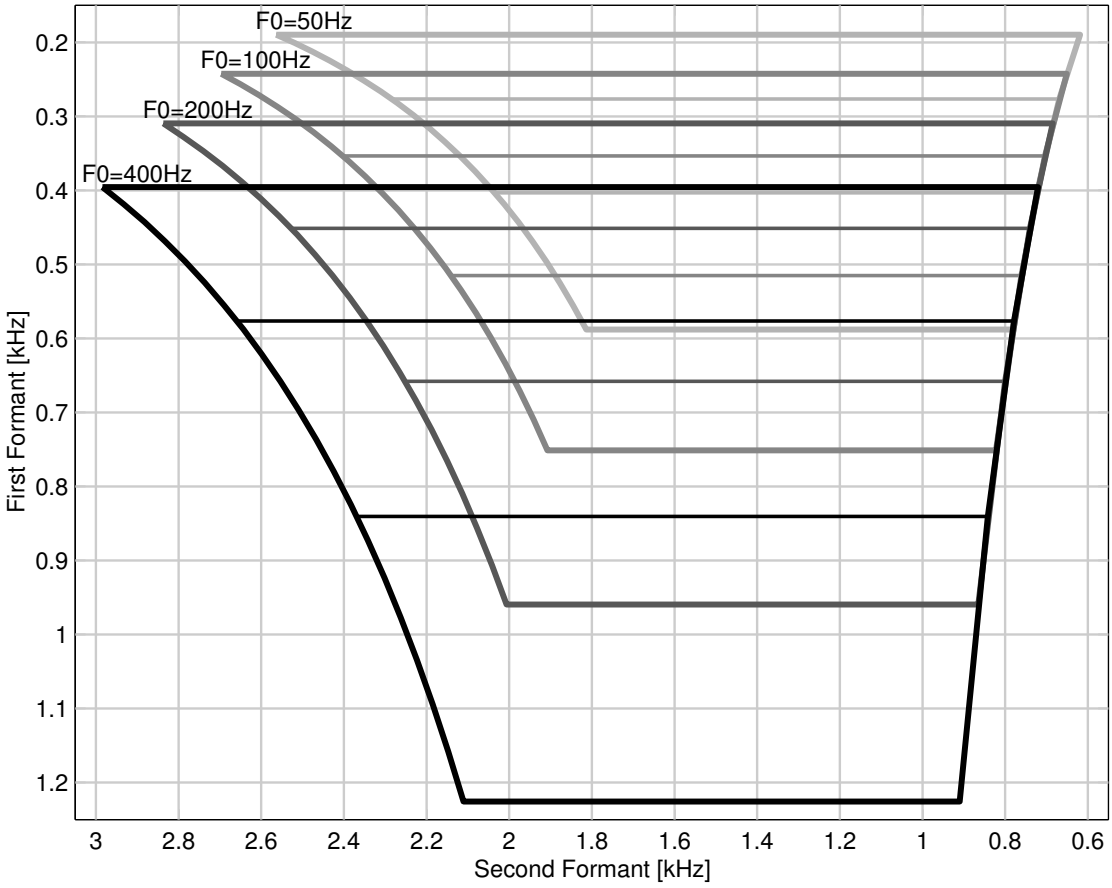

Figure 4: Using the inverse refined model with four different F0 frequencies to project the Cardinal Vowel diagram onto the linear F1/F2 frequency space.

$\widehat{\mathrm{F}_2}$ 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.

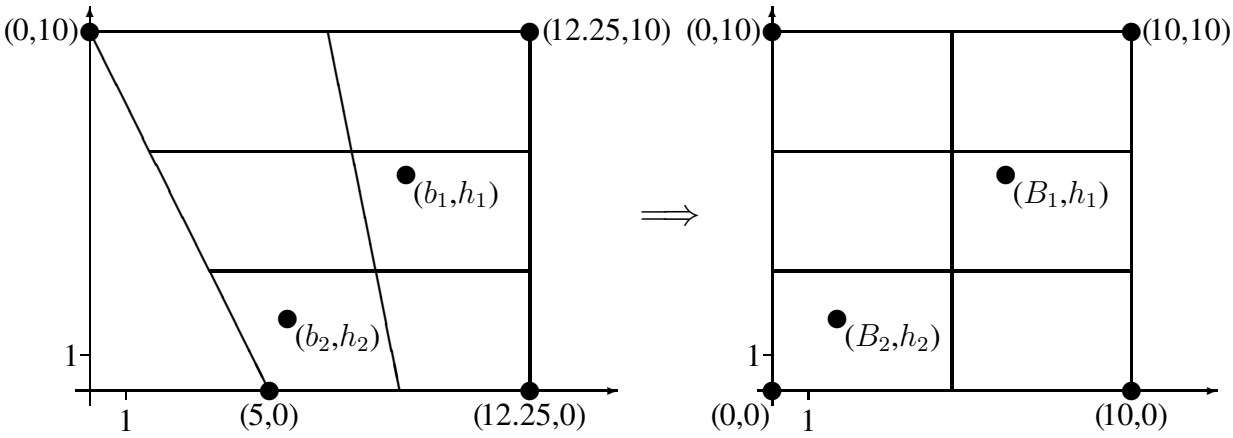

Figure 5: Transformation of perceptual vowel qualities (b_i,h_i) within the Cardinal Vowel diagram (*left*) into a square space (*right*) via formula (3).

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

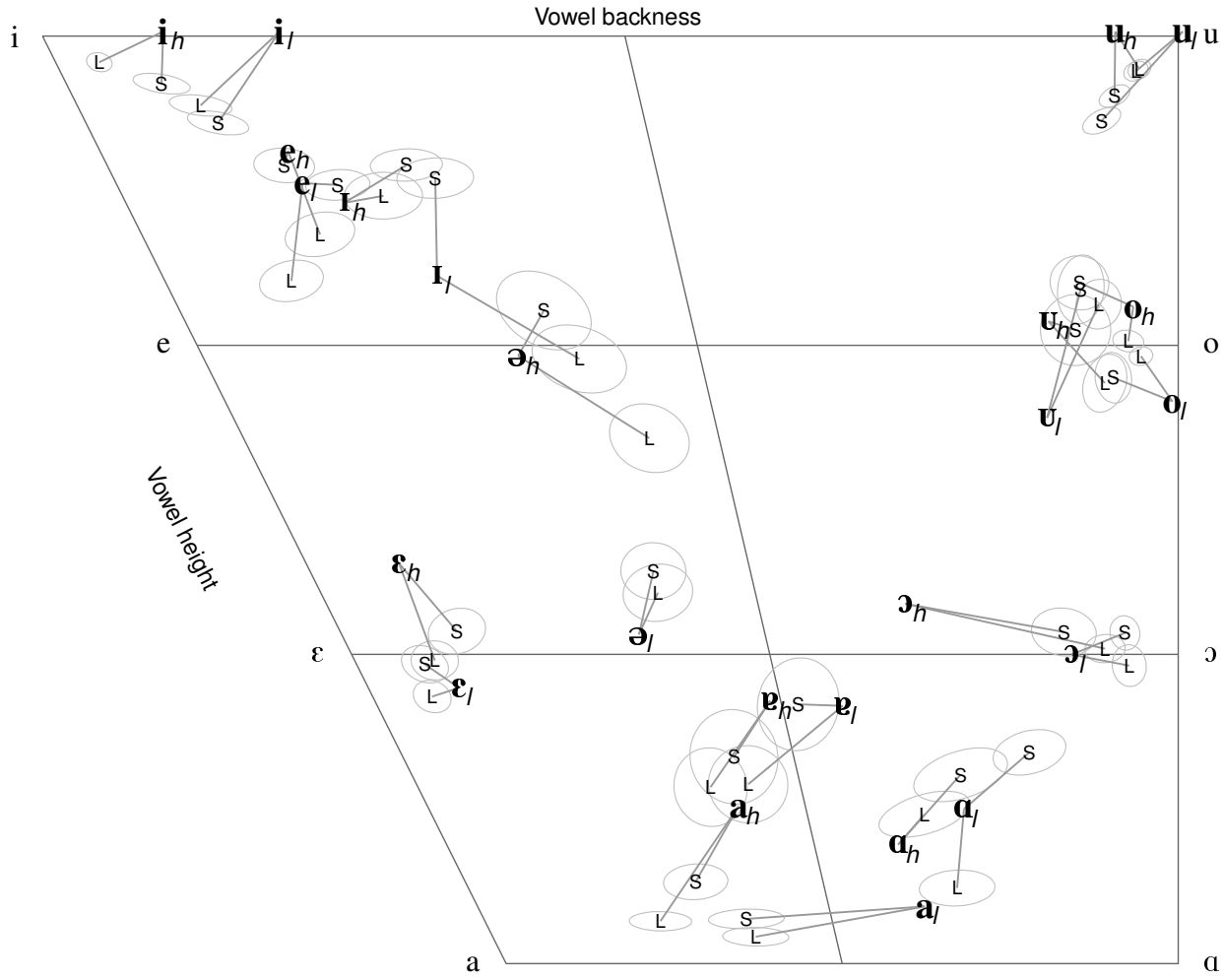

Figure 6: Evaluation results: Perceptual and predicted positions of 24 vowels. Half of them have an F0 of 108 Hz (*l*) and the other half 230 Hz (*h*). Lines join each predicted position (*bold*) with two perceptual reference positions for long (*L*) and shortened (*S*) versions of the vowels.

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

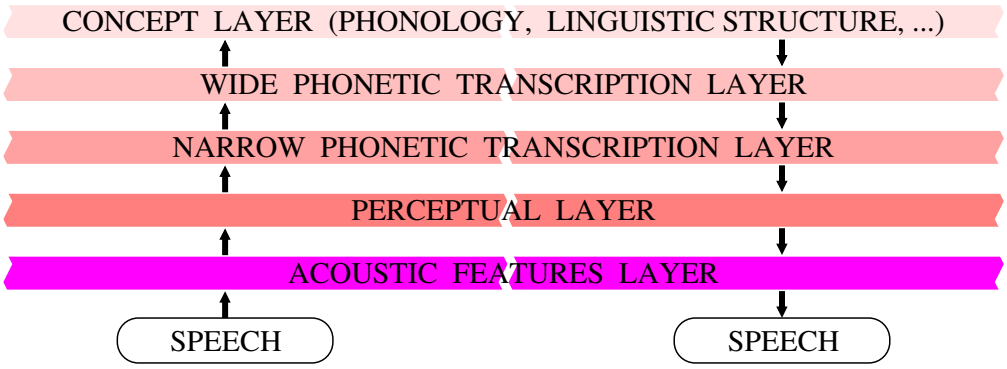

Figure 7: Model of speech analysis/synthesis along layers which have different degrees of abstraction.

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.