Acoustic Correlates of the IPA Vowel Diagram

Hartmut R. Pfitzinger
Department of Phonetics and Speech Communication
University of Munich, Schellingstr. 3, 80799 München, Germany
hpt@phonetik.uni-muenchen.de

ABSTRACT

This study investigates the relationship between the acoustic features of short steady-state vowels and perceived vowel quality, and describes formulae that estimate perceived vowel quality from the speech signal and vice versa. A subsequent evaluation showed these formulae to be more exact than the individual judgements of eight out of ten trained phoneticians.

1 INTRODUCTION

Research on the relation between the acoustic features of vowels and perceived vowel quality has been conducted throughout more than a century (see e.g. Lloyd 1890 [1]). At least since the investigation of Peterson & Barney 1952 [2] F1/F2 formant charts are known to yield overlapping regions for phonologically distinct vowels. In addition, absolute formant frequencies in themselves do not sufficiently represent perceived vowel quality. Consequently, neither positions nor distances in this kind of chart allow us to reliably draw conclusions about the perception of vowels.

By contrast, a point within the familiar Cardinal Vowel diagram, introduced by Daniel Jones in 1917 [3], represents a particular vowel quality which a skilled phonetician should be able to recognize and to reproduce. This kind of vowel chart adequately represents perceived vowel quality but it provides little information on the underlying vocal tract configurations or on the acoustic features of vowels.

Undoubtedly, F1 roughly correlates with perceived vowel height and F2 with perceived vowel backness (Ladefoged 1993 [4]). Numerous studies have been conducted to discover and to characterize the details of this relationship: the inclusion of F0, of Bark-transformation, and of formant frequency distances partly succeeded in improving the prediction accuracy of vowel perception on the basis of acoustic vowel features alone [5, 6, 7, 8, 9].

All these approaches are aimed at solving the so-called vowel normalization problem which is summarized as follows: Men, women, and children are able to produce equivalent vowel qualities despite the fact that they possess individually shaped vocal tracts with various lengths and with individual formant frequencies. Only human perception is able to decode the intended vowel quality from the highly varied combinations of the acoustic features of vowels.

The Cardinal Vowel diagram implicitly solves this many-to-one problem. Reference points in the diagram for various vowels would enable formulae to be developed which estimate the point positions from the acoustic features (i.e. fundamental frequency as well as the first and second formant frequency). In this way, a solution might become explicit.

One of today's challenges in the phonetic sciences is to find a mathematical description for the relationship between vowel acoustics and perception which is sufficiently accurate and straightforward to be used in phonetic research and teaching. This is the goal of the present study which comprises four experiments: i) a vowel quality assessment task to get reference data, ii) a reliability test, iii) the development of prediction formulae, and iv) an evaluation of these formulae.

2 VOWEL QUALITY ASSESSMENT

100 vowel stimuli each having a duration of 80 ms came from the *PhonDatII*-corpus of continuously read German speech. They were randomly selected and then cut from the quasi-steady portion of different vowels from 6 male and 6 female speakers considering the criterion of filling up the F1/F2 space as evenly as possible to ensure the presence of reduced vowel realizations and of gender- and speaker-specific variation. 40 volunteers (9 phoneticians, 24 students of phonetics, and 7 skilled phonetics teachers) from Munich who were intensively trained in narrow phonetic transcription for at least one year participated in this experiment.

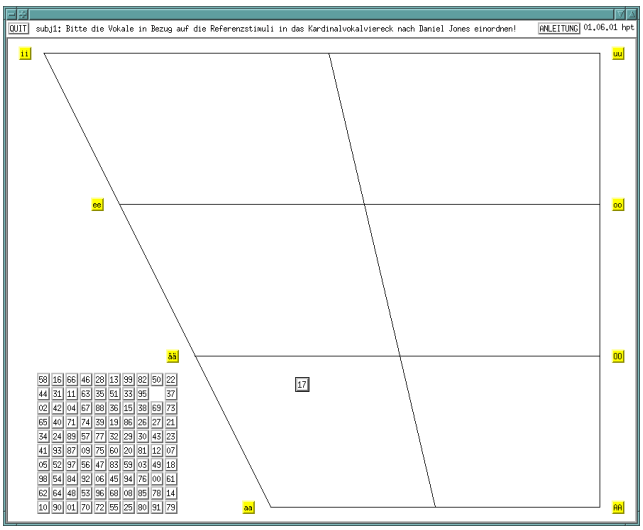

Figure 1: User interface of the interactive perception test.

ISBN 1-876346-48-5

2.1 PROCEDURE

These subjects carried out a computer-aided interactive combined discrimination/identification test using the Cardinal Vowel diagram in which they could place and reorganize the labels of the vowel stimuli and compare their acoustics as often as they wished (see Fig. 1). Each session took about one hour. Denoised and decrackled versions of the first eight Cardinal Vowels taken from the second record originally spoken by Daniel Jones [3] served as the anchor stimuli.

2.2 RESULTS AND DISCUSSION

Averaging the judgements of the 40 subjects yields 100 reference points in the Cardinal Vowel diagram. Fig. 2 shows the positions as well as the 90% confidence intervals. The mean correlation of individual perception results with the group mean is r_h=0.917 for tongue height and r_b=0.919 for tongue backness. Randomly generated responses typically are uncorrelated and contain a four to five times greater deviation from the reference positions than individual subject responses. Obviously, subjects are able to accomplish the perception task. Nevertheless, individual results deviate remarkably from the group mean as can be seen when comparing Fig. 2 and 4. The same amount of deviation has been observed in previous studies [9, 10] and seems to be near to the limit of human precision.

Two separate two-way ANOVAs, both with the factors '*subject*' and '*stimulus*', were applied to perceived tongue height and backness since height is not clearly correlated with backness (r=–0.174). As expected, the factor '*stimulus*' had a significant influence on perceived vowel height (F(99,3861)=192.75, p<0.001) as well as on backness (F(99,3861)=208.49, p<0.001) and explained 83.0% and 84.1% of the observed variance, respectively. Although the factor '*subject*' was also significant (F(39,3861)=1.883, p=0.001) for perceived vowel height, it had no effect on backness (F(39,3861)=0.698, p=0.921) and explained only 1.8% and 0.7% of the observed variance, respectively. These

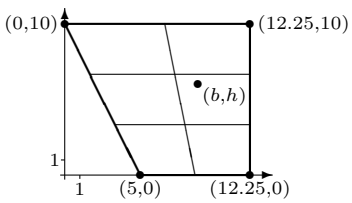

Figure 3: Dimensions of the Cardinal Vowel diagram used in the perception tests and in the prediction formulae.

results provide convincing evidence that the variance of the perception results is mainly determined by the stimuli, supporting the validity of the reference data.

3 RELIABILITY TEST

To evaluate the reliability of the preceding experiment it was repeated after a period of one year. So far, only five subjects have taken part in the repeated-measures test. Therefore final inferential statistics will be postponed until at least ten listeners have been recorded. Instead, for each of the five subjects, the mean deviation between the results of the two tests was calculated. On average, the deviations were 12.7% smaller than the deviations obtained from comparing individual results with the mean group results. These preliminary results strongly support the reliability of the experimental procedure.

Fig. 4 shows the judgements of one subject in the former as well as the latter experiment. Preferences for e.g. the [ɔ] and the [o] region in the Cardinal Vowel diagram, which were discernible in the former test, disappeared in the latter test and other preferences e.g. for the [e] and [ɑ] region appeared. The remaining subjects behaved similarly. Subjects seem to change their preferences when plotting vowel positions in the vowel space. Therefore, the individuality of preferences is not the main source of deviations. A random dispersion seems to be most influential, which supports the hypothesis that this experimental procedure reveals judgements near to the limit of human precision.

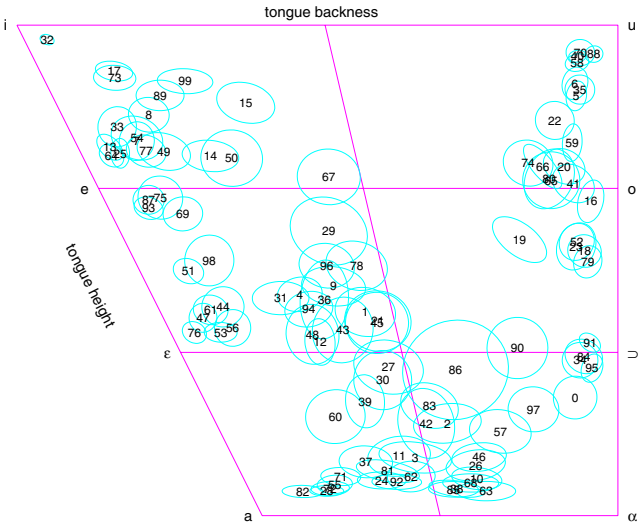

Figure 2: Mean perception results and 90% confidence ellipses estimated from individual judgements of 40 subjects.

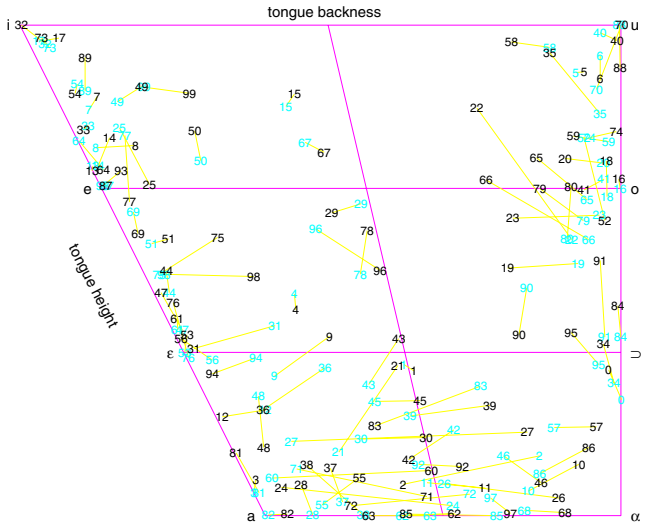

Figure 4: Former *(light color)* and later results *(black)* of one of the subjects who repeated the test after one year.

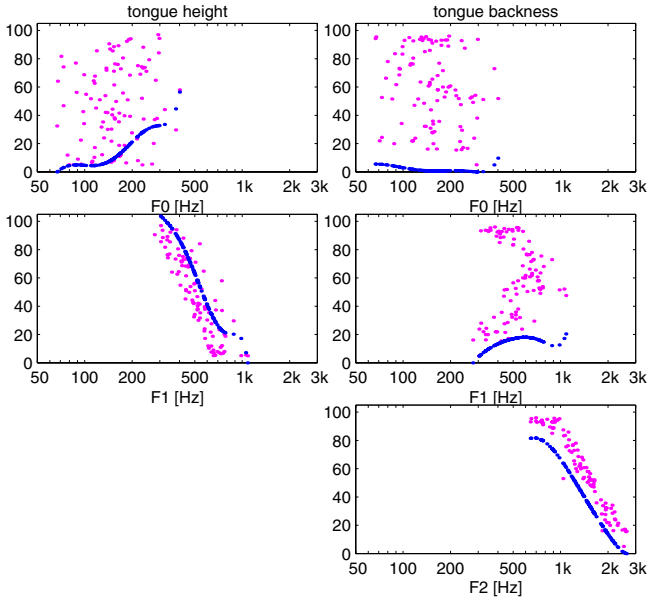

Figure 5: Model A. *Light points:* Scatter plots of perceived tongue height (0=low, 100=high) or backness (0=front, 100=back) versus pitch, first or second formant. *Dark curves:* Effect of frequency values on position displacement.

4 VOWEL QUALITY PREDICTION

Acoustic analysis of the vowel stimuli yielded F0, F1, and F2 which, additionally, were transformed to logarithmic, Bark, and ERB scales. Then, Linear Regression Analysis was applied to the perception and acoustic data. The effect of the acoustic features on predicted vowel height and backness is presented in Fig. 5: e.g. the diagram of F2 vs. tongue backness shows a dark curve which means that an F2 of 800 Hz or less shifts the perceived vowel by 80% to the back (the bend of the curve represents displacement clipping). Fig. 6 shows the predicted vowel qualities. Estimated versus perceived height (r_h=0.972) and backness (r_b=0.983) are highly correlated. None of the 40 subjects exceeded these correlation coefficients.

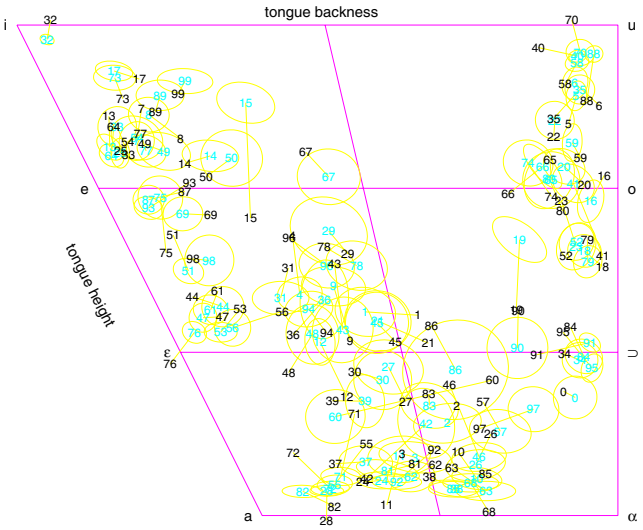

Figure 6: Model A. Mean perception results *(light numbers)* and predicted vowel qualities *(black numbers)*.

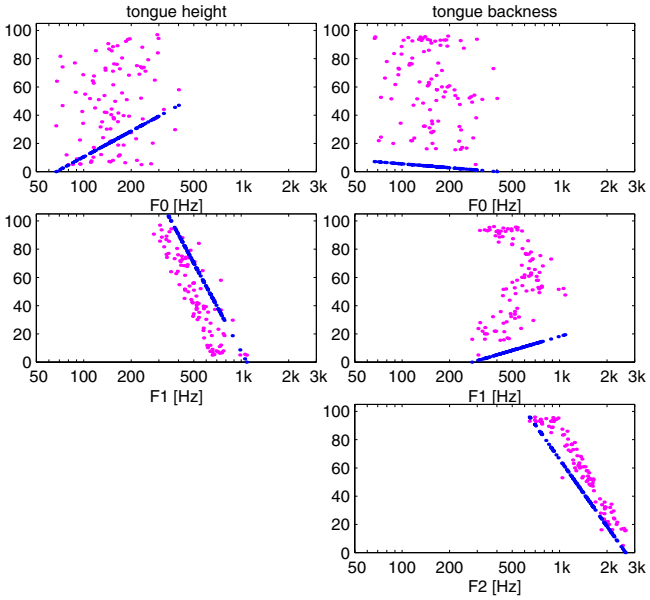

Figure 7: Model B. Compare with Fig. 5.

Then for practical applications and for avoiding overadaptation, a second model (B) was developed which is based on only logarithmic F0, F1, and F2. Fig. 7 and 8 show the results. The correlation coefficients degrade slightly to r_h=0.953 and r_b=0.974, respectively. ERB- as well as Bark-based models are slightly better only at the height estimation (r_h=0.961). The underlying formulae of the logarithmic model are as follows:

$$h = 2.621\log(\mathrm{F_0}) - 9.031\log(\mathrm{F_1}) + 47.873$$

$$b = -0.486\log(\mathrm{F_0}) + 1.743\log(\mathrm{F_1}) - 8.385\log(\mathrm{F_2}) + 59.214$$

Fundamental and formant frequencies are in Hz. The estimated values for height (h) and backness (b) refer to Fig. 3. The inverse formulae enable F1 and F2 to be estimated from a given F0 and a given vowel quality:

$$\overline{\mathrm{F_1}} = e^{0.2902\log(\mathrm{F_0}) - 0.1107h + 5.3013}$$

$$\overline{\mathrm{F_2}} = e^{0.0024\log(\mathrm{F_0}) - 0.0230h - 0.1193b + 8.1637}$$

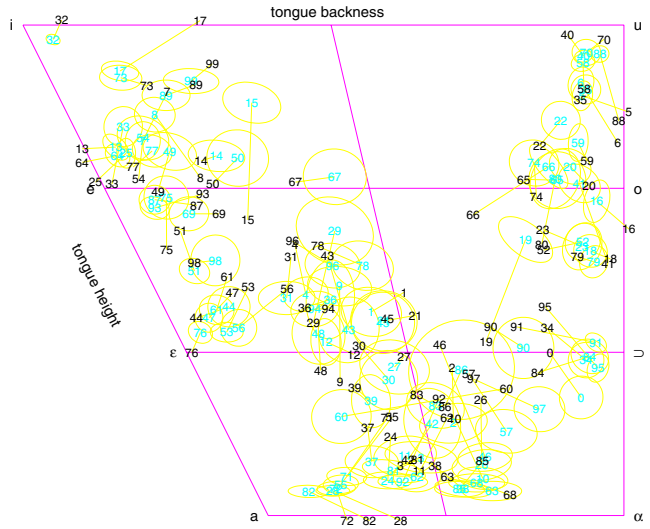

Figure 8: Model B. Mean perception results *(light numbers)* and predicted vowel qualities *(black numbers)*.

ISBN 1-876346-48-5

Since the model B omits displacement clipping, the predicted vowel qualities (h, b) finally have to be clipped to the borders of the Cardinal Vowel diagram in Fig. 3. As for F3, the inverse formulae become ambiguous when including it. In addition, F3 only marginally increases the prediction quality. Therefore, the prediction formulae disregard it.

5 EVALUATION TEST

To evaluate the precision of the vowel quality prediction formulae, additional perception data were required. The stimuli were cut from a large corpus of German spontaneous speech *(VerbMobil)* so that speaking style, vocabulary and speakers differed from the reference data. The number of stimuli was reduced from 100 to 50, thus decreasing the total duration and possibly the granularity of the listening test. Ten trained phoneticians participated in the perception test under the conditions as described in Section 2.1.

Two separate two-way ANOVAs were used for statistical analysis of the raw perception results. The factors '*stimulus*' as well as '*subject*' had significant influence on perceived vowel quality:

	height	backness
stimulus	F(49,441)=47.003 p<0.001	F(49,441)=59.981 p<0.001
subject	F(9,441)=8.206 p<0.001	F(9,441)=2.190 p=0.022

The variance explained by the factor '*stimulus*' slightly decreased to 80.4% for height (from 83.0% in Section 2.2) and 80.7% for backness (from 84.1%). The factor '*subject*' explained 2.8% and 2.3% of the variance, respectively. Therefore, the results confirm the validity of the reference data.

Then, model B was applied to F0, F1, and F2 of the evaluation stimuli. The correlation coefficients of predicted vowel quality with perceived height (r_h=0.949) and backness (r_b=0.967) were again higher than the mean correlation coefficients of the ten individual results with the group mean (r_h=0.924, r_b=0.931). The prediction accuracy of model B exceeded the results of eight out of the ten trained phoneticians.

6 CONCLUSIONS

Dioubina & Pfitzinger 2002 [10] found out that phonetically trained subjects do not perfectly agree when judging vowel quality by means of the Cardinal Vowel diagram. The second experiment of the present study supports these results: Even skilled phoneticians are not able to exactly repeat their judgements after a period of one year. All results of this study contradict the hypothesis that "there is a high degree of agreement among the judgements of skilled phoneticians" (Ladefoged 1967 [11, p.52]). We assume that Ladefoged received quite uniform vowel quality judgements in his studies because he did not present isolated vowel stimuli but monosyllabic words which enable the listener to make use of additional acoustic features like formant transitions and other coarticulation effects.

Nevertheless, the mean result of a group of phoneticians is the most reliable source for the assessment of vowel quality. Only a few skilled phoneticians are able to determine vowel qualities more precisely than the prediction formulae developed here. Consequently, the automatic prediction can be used instead of individual results of the majority of trained phoneticians. A vowel transcription training evaluation is planned.

Many investigations remain to be done: i) the perception of the vowels in the Secondary Cardinal Vowel diagram, ii) the importance of dynamic acoustic features (e.g. formant transitions) for vowel perception, and iii) a perceptual evaluation of the inverse formulae.

REFERENCES

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

[2] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[3] D. Jones, *An outline of English phonetics*, W. Heffer & Sons Ltd., Cambridge, 9. edition, 1962.

[4] P. Ladefoged, *A course in phonetics*, Harcourt Brace College Publishers, Fort Worth, Philadelphia, San Diego, New York, 3. edition, 1993, (1. edition: 1975).

[5] R. L. Miller, "Auditory tests with synthetic vowels," *J. of the Acoustical Society of America*, vol. 25, no. 1, pp. 114–121, 1953.

[6] H. Traunmüller, "Perceptual dimension of openness in vowels," *J. of the Acoustical Society of America*, vol. 69, no. 5, pp. 1465–1475, 1981.

[7] J. N. Holmes, "Normalization in vowel perception," in *Invariance and variability in speech processes*, Joseph S. Perkell and Dennis H. Klatt, Eds., chapter 16, pp. 346–357. Lawrence Erlbaum Associates, Hillsdale, 1986.

[8] A. K. Syrdal and H. S. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *J. of the Acoustical Society of America*, vol. 79, no. 4, pp. 1086–1100, 1986.

[9] H. R. Pfitzinger, "Dynamic vowel quality: A new determination formalism based on perceptual experiments," in *Proc. of EUROSPEECH '95*, Madrid, 1995, vol. 1, pp. 417–420.

[10] O. I. Dioubina and H. R. Pfitzinger, "An IPA vowel diagram approach to analysing L1 effects on vowel production and perception," in *Proc. of ICSLP '02*, Denver, 2002, vol. 4, pp. 2265–2268.

[11] P. Ladefoged, *Three areas of experimental phonetics*, Oxford University Press, London, 1967.