

The /i/-/a/-/u/-ness of Spoken Vowels

Hartmut R. Pfitzinger ^{a,b}

^a Department of Phonetics and Speech Communication, University of Munich, Germany

^b JST/CREST at ATR Human Information Sciences Research Laboratories, Kyoto, Japan

hpt@phonetik.uni-muenchen.de, hrpfitz@atr.co.jp

Abstract

This paper investigates acoustic, phonetic, and phonological representations of spoken vowels. For this purpose four experiments have been conducted. First, by drawing the analogy between the spectral energy distribution of vowels and the vowel space concept of Dependency Phonology, we achieve a new phonologically motivated vowel quality representation of spoken vowels which we name the */i/-/a/-/u/-ness*. As a second step, it is shown that the extension of this approach is connected with the work of Pols, van der Kamp & Plomp 1969 [1] who, among other things, predicted formant frequencies from the spectral energy distribution of vowels. Third, the vowel quality relating to the IPA vowel diagram is derived directly from the spectral energy distribution. Finally, we compare this method with a formant and fundamental frequency based approach introduced by Pfitzinger 2003 [2]. While both the */i/-/a/-/u/-ness* of vowels as well as the perceived vowel quality prediction are quite robust and therefore useful for both signal pre-processing and vowel quality research, the formant prediction achieved the lowest accuracy for the mapping to the IPA vowel diagram.

0. Introduction

The acoustics of vowels and their phonetic as well as phonological representations still present problems for spoken language research. Studies on the vowel transformation between the acoustic and the phonetic domain and vice versa have a long tradition. Recent efforts (Ran, Millar, Macleod & Rose 1994–96 [3, 4], Pfitzinger 1995/2003 [5, 2]) still do not provide commonly accepted solutions but at least appear to be promising.

By contrast, the acoustic-to-phonological mapping is always considerably degraded by the huge amount of acoustic variation of spoken vowels. One of the outstanding studies on the relationship between the acoustics of vowels and their phonological categories has been conducted by Syrdal & Gopal in 1986 [6]. A three-Bark frequency distance criterion between two formants or between one formant and the fundamental frequency was suitable to generate a system of binary features that allowed to considerably reduce the errors when deciding on the corresponding phonological category. Nevertheless, the transition from the acoustic to the phonological layer is not well-understood. Thus, the present study addresses this problem.

0.1. Dependency Phonology

In contrast to phonological descriptions of vowels based on Distinctive Feature Theory (Jakobson, Fant & Halle 1952 [7]), Dependency Phonology (Anderson & Ewen 1987 [8]) makes use of “anchor” vowels *li*, *la*, *lu*, and *lɔ*. The vowel space is characterized by four components: *li* (*palatality* or *acuteness/sharpness*), *la* (*lowness* or *sonority*), *lu* (*roundness* or *gravity/flatness*), and *lɔ* (*centrality*) (Anderson 1986 [9, p.25]). Each vowel in the vowel space is either one of these primi-

tive components, or a ‘mixed’ vowel (e.g. the *ly* is a combination of *li* and *lu*, i.e. *li,u*). Thus, all vowels are described by defining their “distances” to these anchor vowels. Because of the still categorical description, maximally three intermediate vowel quality steps are describable between two anchor vowels, e.g. *le/=li;a* (*li* governs *la*), *lɛ/=li;a* (*li* and *la* mutually govern each other), and *læ/=la;i* (*la* governs *li*). The *lɔ* was introduced to enable the symbolic description of centralized vowels (e.g. *li/=li;ɔ*). Despite the fact that the *li*, *la*, and *lu* components are intended to characterize quantal vowels in the sense of Stevens 1972 [10] and therefore “are grounded in phonetic theory” [9, p.27], Anderson states that the anchor vowels “do not correspond to precise articulatory manoeuvres or acoustic values” [9, p.28]. Nonetheless, we searched for corresponding acoustic values as described in the following.

1. First Experiment

The relation between vowel acoustics and the vowel space concept of Dependency Phonology is investigated. It is not our concern here to train HMM- or ANN-based vowel classifiers to a given set of phonological vowel categories. Instead, we used a phonological theory as a starting point which does not completely ignore the continuous character of vowel quality, namely the Dependency Phonology.

1.1. Method

To find distinctive frequency ranges, the mean logarithmic frequency spectra of the four “anchor” vowels were estimated from 128 *li*, *la*, *lu*, and *lɔ* vowels giving in total 512 vowels which were randomly selected from 270 *li*:/, 1727 *la*:/, 400 *lu*:/, and 1195 *lɔ*/ vowels from the *PhonDatII*-corpus [11] of continu-

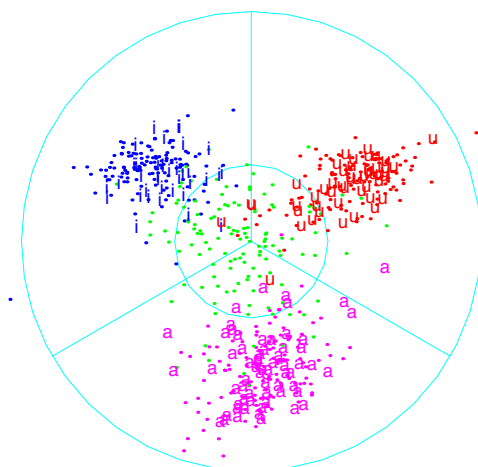


Fig. 1: */i/-/a/-/u/-ness* of training (dots) and evaluation vowels.

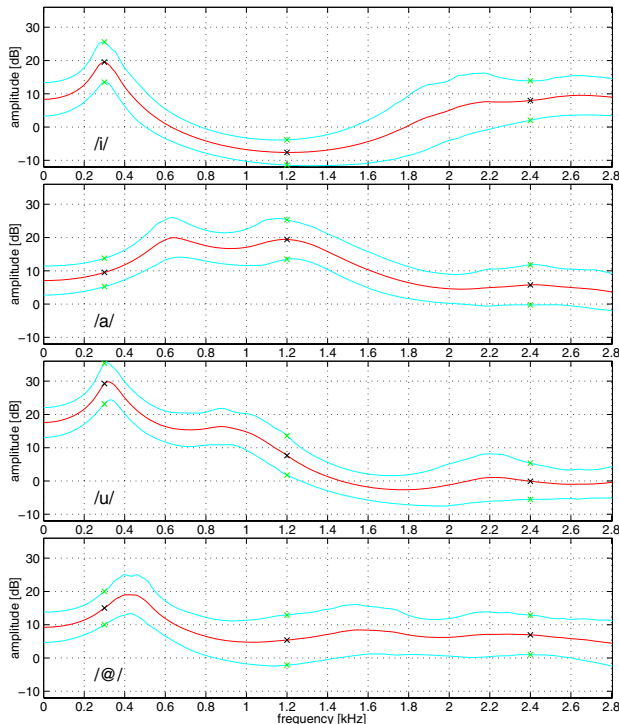


Fig. 2: Mean frequency spectra and standard deviations over the 128 training vowels (16 speakers \times 8 repetitions) for each class.

ously read German speech containing 10 male and 6 female speakers (16 speakers \times 4 vowel types \times 8 repetitions = 512 vowel tokens).

Fig. 2 shows the resulting LP-based and amplitude normalized frequency spectra. Visual comparison of the spectral differences encouraged us to build, as a preliminary approach, a linear regression model based on only three frequency bands: *i*) at 300 Hz the energy is highest for /u/ and might represent *gravity* (see Sec. 0.1), *ii*) at 1.2 kHz the energy is highest for /a/ thus possibly representing *sonority*, and *iii*) 2.4 kHz is near to the prototypical F2 of an /i/ and thus might represent *acuteness*.

The phonologically motivated target vowel space is shown in Fig. 1 and 3: The centers of each third of the circle represent the prototypical vowel categories /i/, /a/, and /u/. Their Cartesian coordinates were used as the training targets. The /@/ target is located in the center of the circle diagram.

Because the preliminary approach yielded strongly overlapping vowel positions the procedure was refined as follows: Two linear regression models (for the x- and the y-dimension) based on 19 logarithmic spectral amplitudes measured between 250 Hz and 2500 Hz and equally spaced by 125 Hz were estimated from the 512 vowel spectra of the training corpus.

1.2. Evaluation and Discussion

For the evaluation of the refined prediction method, we used another set of 48 /i/, 48 /a/, and 48 /u/ vowels. The positions of both the training and evaluation vowels are plotted in Fig. 1. Additionally, we estimated the positions of 530 /e/, 361 /ɛ/, 371 /ɔ/, and 323 /o/ vowels (see Fig. 3) which were also not included during the training procedure.

Both Figures show clearly, that although the centers of the distributions of vowels belonging to the same class are arranged in the circular vowel space as expected, the acoustic variation of vowels is also reflected. An informal perceptual inspection of 10% of the vowels, which deviate most from the centers of

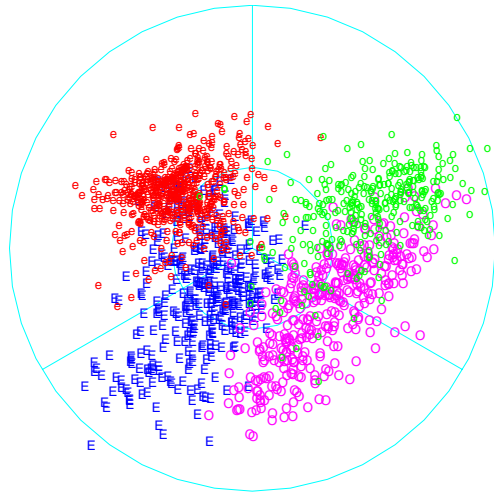


Fig. 3: /i/-/a/-/u/-ness of 530 /e/, 361 /ɛ/, 371 /ɔ/, and 323 /o/ taken from the evaluation corpus.

their corresponding distributions, revealed for 37% of the vowels strong deviations between their phonological labels and their phonetic vowel qualities.

But despite their strong deviations in the circular vowel space, 63% of the inspected vowels perceptually match their phonological labels. Thus this approach is, not surprisingly, also confronted with the vowel normalization problem (i.e. phonetically equivalent vowels show strong acoustic variations) which occurs in every acoustic-to-phonetic mapping. Obviously, the acoustic-to-phonological transformation suffers from both the vowel normalization problem and the problem of phonetically different vowels belonging to the same phonological class. In the following sections the vowel normalization problem is addressed.

2. Second Experiment

Formant frequencies are considered to be, on the one hand, the most important acoustic features of vowels but, on the other hand, the most difficult to reliably extract from the speech signal. The aim of this experiment is to estimate the frequencies of the formants 1, 2, and 3 directly from linear combinations of the logarithmic amplitudes of successive frequency bands. This approach could be interpreted as an extension of the method introduced in the previous section.

While improved versions of the cepstrum-to-formant mapping introduced by Broad & Clermont 1989 [12] are applied in speech research (Mokhtari & Campbell 2003 [13]), direct mapping from the frequency spectrum to formant frequencies has to our knowledge not reached this state. Pols, van der Kamp & Plomp 1969 [1, p.464] achieved correlation coefficients of $r_{F1} = 0.985$ and $r_{F2} = 0.981$ between predicted and measured F1 and F2 but they averaged over 50 speakers. Their analysis was based on a method introduced by Plomp, Pols & van de Geer 1967 [14] to predict vowel categories from a linear combination of 6 principal components over 18 bandpass filters.

2.1. Method

100 vowel signals cut from 6 male and 6 female speakers of the *PhonDatII*-corpus (see Sec. 1.1) together with their manually measured and repeatedly verified F1, F2, and F3 values served as the training data. A careful selection of the underlying frequency range for each formant has been conducted by reduc-

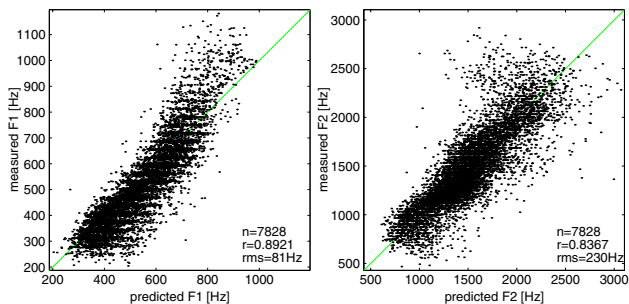


Fig. 4: Scatter plots of measured versus predicted F1 and F2.

ing the number of bandpass filters as much as possible without considerably degrading both r and the root-mean-square error (rms). As a consequence, F1 is predicted by a linear combination of the logarithmic amplitudes at the five frequencies 250, 500, 750, 1000, and 1250 Hz while F2 uses 14 frequencies in the range of 250–3500 Hz equally spaced by 250 Hz. F3 is predicted from 10 frequencies between 2000 and 4250 Hz.

2.2. Results and Discussion

Fig. 4 presents scatter plots, correlation coefficients, and rms -values for F1 and F2 of 7828 vowels having a duration of more than 40 ms and coming from the formant database built by Heid, Wesenick & Draxler 1995 [15]. Predicted F3 showed only a correlation coefficient of $r=0.6966$ and an rms of 229 Hz.

The results of Broad & Clermont [12, p.2015] who reported rms -values of 52 Hz for F1, 164 Hz for F2, and 218 Hz for F3, are based on vowels spoken by only four male speakers. In a similar but speaker-dependent and MFCC-based approach Högberg 1998 [16] reports rms -values of 52 Hz for F1, 97 Hz for F2, and 155 Hz for F3.

Our results seem to be considerably degraded by two facts: *i*) the multiple linear regression coefficients were trained on only 100 vowels and *ii*) both male as well as female speakers were used. Taking this into consideration, the degradation of our results was to be expected. But at least for F1 and F2 the prediction method is accurate enough to be used as an alternative formant detection in the next sections.

3. Third Experiment

The goal of this experiment was to develop a method for estimating the perceptual vowel quality as defined by the Cardinal Vowel diagram of Jones 1962 [17] and used in the IPA vowel diagram. For this purpose, reliable positions of vowels in the diagram were needed. This we obtained in an earlier perception experiment [2]: 100 vowel stimuli each having a duration of 80 ms were randomly selected and then cut from the quasi-steady portion of different vowels from 6 male and 6 female speakers of the *PhonDatII*-corpus mentioned in Sec. 1.1.

40 volunteers (24 students of phonetics who were intensively trained in narrow phonetic transcription for at least one year, 9 phoneticians, and 7 skilled teachers of narrow phonetic transcription) from Munich carried out a computer-aided interactive combined discrimination/identification test using the primary 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.

3.1. Method

We used the same bandpass filters as in the formant frequency prediction experiment (see Sec. 2) since it is well-known that F1 roughly correlates with vowel height and F2 with vowel back-

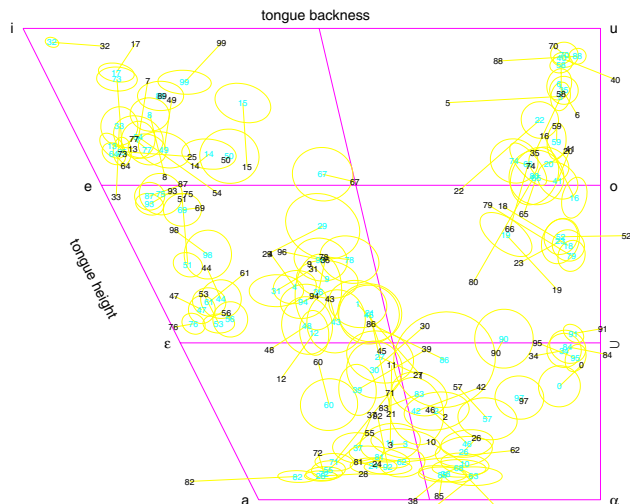


Fig. 5: Cardinal Vowel diagram with mean perception results (light color) and 90% confidence ellipses estimated from individual judgements of 40 subjects. Black numbers represent the positions predicted from spectral properties of the vowels.

ness. Two linear regression models were trained to predict the Cartesian coordinates of the mean perception results for the 100 reference vowels in the primary Cardinal Vowel diagram. Additionally, we included F0 values measured from the reference vowels in a second set of two linear regression models to enable the analysis of the influence of F0 on the accuracy of predicted perceived vowel quality, because F0 was found to be useful in speaker and gender normalization of vowel height [18, 19].

3.2. Results

The prediction accuracy is summarized in the following table:

	vowel height		vowel backness	
	r	mean dev.	r	mean dev.
spectrum→IPA (without F0)	.903	9.90%	.964	6.88%
spectrum→IPA (using F0)	.960	6.01%	.965	6.77%

r are the correlation coefficients between the perceived and predicted vowel height or backness, *mean dev.* are the mean absolute errors divided by the height or the mean width of the Cardinal Vowel diagram. Fig. 5 shows the positions of the perceived and predicted vowel qualities. An evaluation corpus of 50 vowels judged by 10 trained phoneticians caused the correlation coefficients to degrade only slightly by 0.005 and 0.007, respectively. Statistical analysis revealed that the influence of F0 on the vowel height prediction accuracy is highly significant.

4. Fourth Experiment

The last experiment of this study is concerned with the prediction of vowel quality using predicted formants (see Sec. 2). Two sets of two linear models, one based on only F1 (or F0 and F1, respectively) for perceived vowel height and the other on F1 and F2 (or F0, F1, and F2, respectively) for perceived vowel backness, were trained on the same 100 reference vowels as in the previous experiment. Also the influence of the presence or absence of F0 information on the prediction quality was tested.

4.1. Results

The following table summarizes the accuracy results of the two vowel quality prediction procedures which were trained on predicted formant frequencies (see Sec. 2) excluding and including F0 measurements:

