Measuring and Comparing Vowel Qualities in a Dutch Spontaneous Speech Corpus

Irene Jacobi, Louis C.W. Pols, Jan Stroop

Amsterdam Center for Language and Communication,
University of Amsterdam, Netherlands
i.jacobi@uva.nl

Abstract

Recent studies of spoken Standard Dutch support an ongoing change in the phonetic quality of the diphthong /ɛɪ/ [1, 2]. However, there is a need for broader analyses and larger data sets. Here, we took Dutch vowel variants of 44 speakers from a spoken Dutch speech corpus, the CGN [3]. The vowels were measured and compared on the basis of 15.000 vowel segments, consisting of productions of /ɛɪ/, /ɑu/, /ʌy/, /oː/, and /eː/, as well as the anchor vowels /a/, /i/, /u/. It was our aim to analyze changes in vowel quality dependent on the speakers' sociological backgrounds and ages, and to deal with the variable recording qualities of the corpus. All vowels were taken from spontaneously uttered sentences and were analyzed automatically by means of a principal component analysis (PCA) on the vowels' bark-filtered spectra, as well as by formant analysis.

Recalculating spectral positions in the principal components (pc's) plane displayed the spectral interaction of the first formants in the pc1-pc2 plane, and explained the better separability of the vowels compared to the F1-F2 plane, as well as the high correlation of the first three formants with pc1 and pc2. The first pc's turned out to be rather insensitive to sex-differences, but they were sensitive to the signal-to-noise ratio of the speech data. Variable recording qualities manifested themselves in speaker-specific location and size of the vowel spaces. Good signal-to-noise ratios could be transformed to poorer signals by increasing the lowest possible dB values per filter. Having analyzed the influence of noise on our data, we could normalize the data by taking each speaker's /a-i-u/ positions and the focal point as references for better inter-speaker comparison.

The results clearly show different vowel quality patterns dependent on the speakers' education and age, and indicate a progress of quality changes with as parameters the lowering and the degree of diphthongization of the long vowels and diphthongs.

Index Terms: vowel variation, speech quality, social background, Dutch.

1. Introduction

In our previous study, the acoustic properties of vowels from Dutch spontaneous speech of twelve speakers were compared by means of formants and a principal component analysis (PCA) on their bark-filtered spectra [2]. The latter analysis is more robust since it needs no hand correction and can be fully automatized. To be able to interpret the individual variation, the speakers' anchor vowels /a/, /i/, and /u/ were used as references on which the PCA was calculated. The resulting first two components (pc1 and pc2) of the PCA on the bark-filtered spectra of the sound segments were comparable to F1 and F2 in bark of the same sound segments. When it came to find acoustic cues to the perceived diphthong variety of /ɛɪ/, the pc1-pc2 plane was more meaningful.

These initial results led to further investigations including the other Dutch diphthongs, as well as investigations on the dynamics of the Dutch so-called 'pseudo' diphthongs, and possible dynamic changes within such long monophthongs from words with <ee>, <oo> and <eu>; /eː/, /oː/ and /øː/. A larger sample of speakers might then reveal the temporal order of change within the whole vowel system over the last decades, taking into account the aspects of age and social background.

In this paper we will concentrate on the Dutch genuine diphthongs /ɛɪ/, /ɑu/, and /ʌy/ of 44 speakers, as well as on the long and slightly diphthongized monophthongs /oː/ and /eː/. Compared to the other vowels, the third Dutch long and slightly diphthongized monophthong /øː/ is less frequent in the data. Due to the small amount of data for /øː/, it will be neglected in this study.

2. Data

The spontaneous speech of 44 adult speakers of different age groups and with different sociological backgrounds was taken from the Spoken Dutch Corpus[1] (CGN). At the time of recording (around 2000), the 22 female and 22 male speakers were between 20 and 74 years old. All speakers had been acknowledged as speakers of Standard Dutch concerning their first, home, and work language. Their speech had been recorded during interviews, gatherings, discussions and private conversations.

From the spontaneous utterances we selected all stressed realizations of the vowels /a/, /i/, /u/, /ɛɪ/, /ɑu/, /ʌy/, /eː/ and /oː/, in a variety of phonetic contexts. The extraction criterion was based on lexical stress and a minimum duration of the vowel. Segments with overlapping speakers or strong accidental signal distortions were excluded. Due to possible strong retroflexal or velar coarticulatory influences, vowel segments from special environments were excluded also, e.g. vowels followed by final /r/ or /l/.

For the segment boundaries and vowel classes we relied on the corpus segmentations and annotations that fitted our research in terms of a broad transcription: the phonemic representation of the corpus is based on the orthographic transcriptions of the corpus and was generated fully automatically by TreeTalk [4]. The symbols were derived from SAMPA in such a way, that the produced sounds were related to the phonemes of Dutch [5], thus giving the same symbol to all variants of a phoneme: "E+" to all /ɛɪ/, "A+" to /ɑu/, "Y+" to /ʌy/, "e" to /eː/, and "o" to /oː/. For one million words, the automatic transcriptions of the corpus had been checked manually [3]. We checked all our data manually and excluded (the minute amount of) suspect transcriptions and segmentations respectively.

The frequency of occurance for words and vowels differed among

[1] http://lands.let.kun.nl/cgn/ehome.htm

speakers, also depending on topics. Most frequent within the long vowels were segments of /eː/, /ɛɪ/, and /oː/, less frequent were /ɑu/ and /ʌy/. The longest mean vowel durations were found for /ɑu/ (118ms) and /ʌy/ (116ms). The shortest, and together with /ɛɪ/ (112ms) also the most homogenous durations were found for /eː/ (102ms). All measurements were done using the Praat program [6].

3. Method

All vowel segments were bandfiltered and formant tracked automatically at the same points in time. For the anchor monophthongs /a/, /i/, /u/, the analysis was performed at the middle of the vowel. The diphthongs and the other long vowels were analyzed at one tenth and nine tenth of their duration. For temporal analyses, the long vowels were also barkfiltered at every ten milliseconds of the total vowel duration. We used 20 barkfilters up to 21 bark, and took the mean of the first two filters to bar variance in these filters caused by the speakers' varying F0 (see [2]). For the analysis, the barkfiltered segments were level normalized to 80 dB. We calculated a PCA on the mean barkfilter values of each speakers' /a/, /i/, /u/, altogether 132 means from 7575 vowel segments (Fig.1), and used the resulting dimensions for further analysis of all vowel segments. A PCA on the 572 means of all measured begin and end values of the long vowels and diphthongs of the 44 speakers, plus the anchor vowel means, resulted in barely different eigenvectors and fractional variances, and so we continued using the PCA based on merely the anchor vowels.

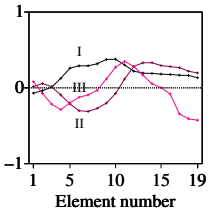


Figure 1: *Eigenvectors 1 to 3 of the PCA on all mean barkfiltered /a/, /i/, /u/ of the 44 speakers. The first three dimensions explained 93% of the variance: pc1 explained 65%, pc2 23%, and pc3 4%.*

	$F1_{bark}$	$F2_{bark}$	$F3_{bark}$
pc1	+0.837	+0.129	-0.187
pc2	-0.204	+0.885	+0.312
pc3	-0.129	+0.313	-0.350

Table 1: *Correlations of the first three pc's with the first three formants (bark), based on 572 means (/a/, /i/, /u/, as well as the long vowel on- and offsets) of the 44 speakers.*

4. Recording quality

Various recording qualities are a characteristic of the spontaneous speech part of the CGN, and so we investigated the implications of this variability on our vowel analyses. We compared the most extreme speakers in as far as their vowel space size and location in the pc1-pc2 plane was concerned. The one extreme recording had also been perceived as being of rather low quality, a radio recording with music in the background. The influence of background noise on the vowel space size was furthermore tested by degrading speech of good quality. Every filtered value that was below 20/30/40/60 dB was set equal to 20/30/40/60 dB, all other values were kept as they were. The increase of the minimum dB in the filters resulted ultimately in a mere point in the plot for more than 60 dB. Figure 2 shows an example of two cases with different recording qualities. Results display that different locations of the speakers' vowel spaces in the pc1-pc2 plane can at least partly be led back to the signal-to-noise ratio. In this regard, we furtheron normalized the spectral data by setting the speakers' /a/-/i/-/u/ focal points to 0, and therewith abated most noise differences.

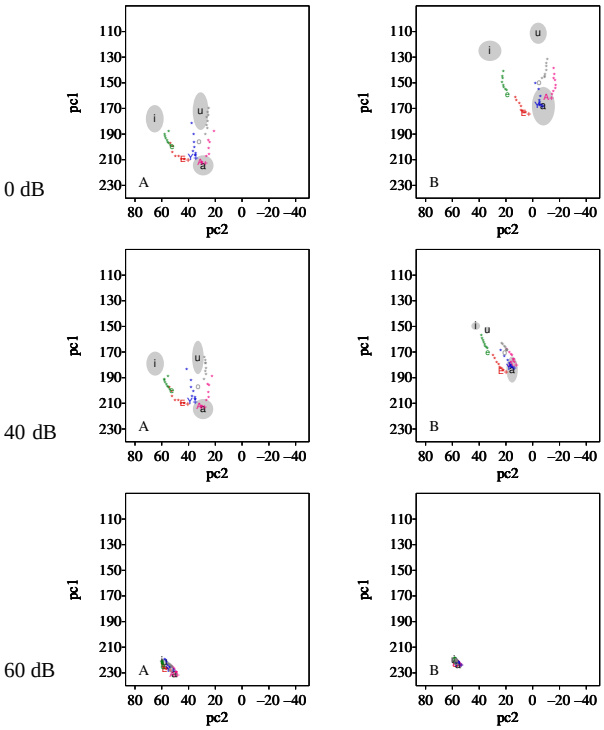


Figure 2: *Speaker A (rather poor recording quality) and speaker B (good recording quality) before (top) and after (bottom) increasing filter minimum values. Increasing the minimal filter values results in decreasing sizes and shifting positions of the vowel space. One sigma ellipses represent /a/,/i/,/u/. The points indicate 10ms steps in time of the mean vowel movement for /oː/, /eː/, /ɛɪ/, /ʌy/ and /ɑu/.*

5. The interaction of formants

As can be seen in Table 1, the first formant (in bark) highly correlates with pc1, and the second formant with pc2. F3 seemed to steadily correlate higher with pc2 and pc3, the more speakers we put in the PCA, and the more diverse the recording qualities of the data respectively. Considering that F2 and the higher formants merge and split, a representation by merely F1 and F2 has earlier been reported as being inadequate for the multi-dimensional nature of vowel qualities [7].

When comparing recalculated spectra in the pc1-pc2 plane (cf. Figure 3), the interaction with the first two formants, and the second and higher formants alternatively, became obvious, as well as the sensitivity to noise. Only in relation to each speaker's other (anchor) vowels do the speaker-specific spectra of the phoneme classes make sense. To make the speaker-specific data comparable between speakers, we had to put the vowel positions in the pc1-pc2 plane in relation to each other by measuring the relative distances within the vowel sets.

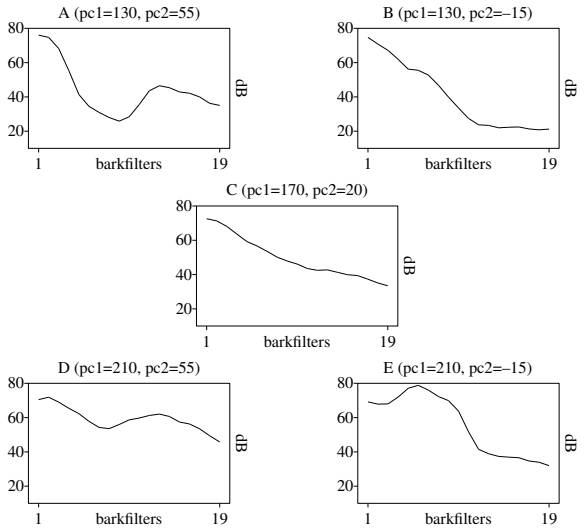
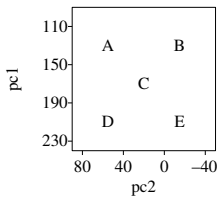


Figure 3: *Recalculated spectra (above) from the corners and from the center of the pc1-pc2 plane, indicated by A, B, C, D, E. For all pc3 to pc19 values, the 44 speakers' /a/-/i/-/u/ focal point values of pc3 to pc19 were taken.*



6. Comparing speaker vowel sets

To detect certain patterns within the vowel data, we compared the pc1-pc2 values of the speakers. Aim was to find out if a speaker's vowel set would highlight his or her educational or occupational level. The level of occupation (high or low) and the level of education (high or low) turned out to be the same for all except one speaker, and so we concentrated on only one level, the level of education. All speakers had been acknowledged as speakers of Standard Dutch. However, it has to be mentioned that the data of the elder low educated (male) speakers had to a certain amount perceptible dialectal characteristics, which was less obvious for the rest of the speakers.

Since the speakers' vowel spaces had diverse measures, we put all vowels into perspective of the location of their anchor vowels /a/, /i/, and /u/. Comparable to the method of measuring the Euclidean distance in van Heuven et al. [1], we started with measuring the distance of each diphthong and long vowel onset value to /a/. This distance was then related in percentage to the distance between /a/ and /i/ of the same speaker, with the distance of /a/ and /i/ always being normalized to 100%. This relation of onset position and /a/, compared to the distance of /a/-/i/, resulted in two percentage fractions for pc1 and pc2, representing the postition of each long vowel and diphthong onset.

To compare the degree of diphthongization of the long vowel and diphthong segments, the distances between on- and offset were related to, again, the speaker-specific /a/-/i/ distance. For the back vowels /oː/ and /ɑu/, we chose the /a/-/u/ distance as relation. The pc2 /a/-/u/ distances related to the other vowel onsets turned out to be highly diverse for some speakers. As the low energy level in the higher barkfilters is one main characteristic of /u/, the /u/-quality is the first to be affected by noise. Baring this in mind we had to be cautious with the pc2 values of the back vowels.

7. Speaker group patterns

For males and females, all correlations of the pc1 onset values with each other had been positive, apart from /ʌy/ for the females, and thus already indicated, that the onsets of the long vowels and diphthongs interdepend in the first dimension. Ploting all 44 speakers displayed educational group patterns.

7.1. Degree of diphthongization

A MANOVA on the 44 speakers' means of their relative on- and offset distances for /oː/, /eː/, /ɛɪ/, /ʌy/, and /ɑu/, with factors sex and level of education displayed significant effects for the vowels ($F(4,37)=22.88$, $p=.000$), for all vowels and their pc1-pc2 distance values ($F(4,37)=4.84$, $p=.003$), and for the vowels, their pc1-pc2 values and the level of education ($F(4,37)=2.69$, $p=.046$). Since pc1 is the most important parameter, we did a t-test on the means of the relative pc1-distances between on- and offset positions for the high vs. low educated within the group of males and females. The test displayed significant differences within the group of males for /ɛɪ/, /ʌy/, /eː/ (all $p<.05$), and /oː/ ($p<.005$) (see Fig.4, top): higher educated males diphthongized to a larger extent than lower educated males. A t-test on the group mean pc1-differences for the females concerning the degree of diphthongization showed significant differences for high vs. low educated speakers for /ʌy/, /oː/, and /eː/ ($p<.005$), and for /ɛɪ/ ($p<.05$).

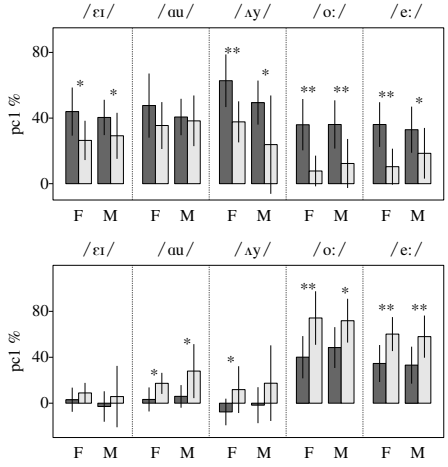


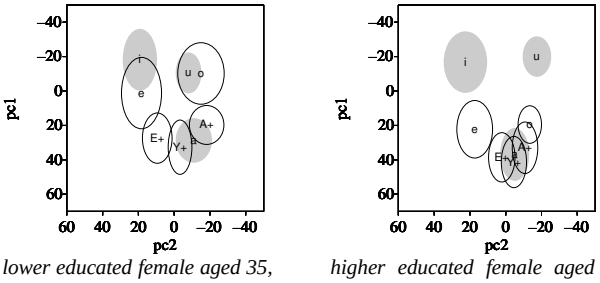
Figure 4: *T-tests on the group means of the relative degree of diphthongization (top), and relative onset values (bottom): F for females, M for males; dark bars represent higher educated, light bars lower educated speakers. The stars display significance.*

7.2. Long vowel and diphthong onset positions

A MANOVA performed on the means of the relative onset positions of the long vowels and diphthongs with factors level of education and sex revealed significant effects for the vowels ($F(4,37)=713.52$, $p=.000$), the vowels and level of education ($F(4,37)=3.26$, $p=.022$), the vowels and sex ($F(4,37)=4.78$, $p=.003$), the vowels and their onset values ($F(4,37)=20.39$, $p=.000$), and for the vowels, their onset values and the level of education ($F(4,37)=4.79$, $p=.003$). When split into males and females, the group of females indicated a significant effect of interplay for the vowels, their pc1-pc2 values, and the level of education, whereas the males indicated a marginal effect for the vowels and the level of education. With pc1 as the most important indi-

cator, a t-test on the pc1 means of the higher vs. lower educated group of males showed significant differences for /ɑu/, /oː/ (for both $p<.05$), and /eː/ ($p<.005$). Higher educated males displayed lower vowel onsets than lower educated males (Fig.4, bottom). For the females, a t-test on the means of the onsets for the higher vs. lower educated group showed significance for the relative onset of /ɑu/ and /ʌy/ (both $p<.05$), and a significant pc1 difference between the groups and their /oː/ and /eː/ values (both $p<.005$) (Fig.4, bottom).

When comparing the correlations within the higher educated group of males to females, all male onset values (for /oː/ significantly) correlate conversely to the female onsets with the year of birth (apart from /ɛi/) . In other words, the younger the high educated females are, the lower their vowel onsets appeared to be, contrary to the high educated males. More data to split age groups with representative speaker numbers will hopefully clarify the correlations and connection between vowel lowering, education and the year of birth. A clear pattern was found for the larger group of "middle aged" females, where the contrast between speakers of different educational background is rather steady (compare Figure 5, top).



lower educated female aged 35, postal worker

higher educated female aged 37, teacher secondary school

lower educated male aged 31, mechanic

higher educated male aged 33, job unspecified

Figure 5: *Example of 4 speakers of roughly the same age and different backgrounds. Pc1-pc2 planes with one sigma ellipses of anchor vowels /a/, /i/, /u/ and begin values of /ɛɪ/, /ɑu/, /ʌy/, /oː/, /eː/ (E+, A+, Y+, o, e). All speakers' focal points were set to 0. The speakers on the right show lowered diphthongs as opposed to those on the left. The arrows in the lower panel display the distance between mean on- and offset of the long vowels/diphthongs.*

8. Discussion and conclusion

Whithin a large corpus of spontaneous speech that was recorded under various circumstances, comparing vowel variants is a difficult task since the recording quality, as well as speaker-dependent physical attributes probably confuse measurements. We tried to make the speaker data comparable by using a reliable automatic method for analyzing the vowels, which reduced speaker-dependent physical attributes in its building process, and which was based on all speakers' anchor vowels /a/, /i/, /u/. The resulting pc1-pc2 dimensions of our spectral analysis highly correlated with the first two formants (in bark), which are used to traditionally represent vowel qualities. The second and third dimensions show also correlations with the third formant, and confirmed the role of F3 in vowel variant analysis, that is often neglected. Normalizing the speakers' focal points reduced the artefacts of variable recording qualities while keeping the vowel variation.

When analyzing the data, each speaker's unique vowel array was taken into account by refering the long vowels and diphthongs to the anchor vowels, which represent the extreme vowel qualities, and are supposed not take part in quality changes. Comparing the relative distances within each of the 44 speakers' vowel set revealed speaker spanning behaviours within the first two dimensions. Generally, the strong correlations of the relative vowel onsets with each other point out, that the vowel locations in the first two dimensions differ hinging on each other. The year of birth seemed to have more impact on high educated (female) speakers when it came to the process of lowering, where females and males showed opposite behaviours, though not significantly.

The results showed, that, although there might be a continuum of diphthong variants, there are definite trends. Speakers lowering the genuine diphthongs /ɛɪ/, /ɑu/ and /ʌy/, also lowered /oː/ and /eː/. These speakers (Fig.5, plots on the right side) also diphthongized them to a larger extent than speakers who did not lower diphthongs and long vowels (compare Fig.5, plots on the left side). Looking at the metadata, the group of speakers who do not lower the long vowels and diphthongs, differs in education and, for females, in age from the group of speakers who do lower the long vowels and diphthongs. As already mentioned, more data will be needed to specify the age groups that differ in behaviour, though sharing the same level of education. All in all, the results indicate sound changes in progress with as most salient parameters the lowering and the differing distances between on- and offsets of the genuine diphthongs, as well as the 'pseudo' diphthongs /oː/ and /eː/.

9. References

[1] van Heuven, V.J., Edelman, L. & van Bezooijen, R., "The pronunciation of /ɛɪ/ by male and female speakers of avant-garde Dutch", Linguistics in the Netherlands, 2002, p.62-72.

[2] Jacobi, I., Pols, L.C.W. & Stroop, J., "Polder Dutch: Aspects of the /Ei/-lowering in Standard Dutch", Proc. Interspeech, 2005, p.2877-2880.

[3] Oostdijk, N., Goertier, W., van Eynde, F., Boves, L., Martens, J.P., Moortgat, M. & Baayen, H., "Experiences form the Spoken Dutch Corpus project", Proceedings 3rd LREC, 2002, p.340-347.

[4] Daelemans, W. & van den Bosch, A. "TreeTalk: Memory-Based Word Phonemisation", in: Data-Driven Techniques in Speech Synthesis. editor: Damper, R.I., Klewer Academic Publishers, 2001, p.149-172.

[5] Gillis, S. "Protocol voor de Brede Fonetische Transcriptie", http://lands.let.kun.nl/CGN/home.htm, 2001.

[6] Boersma, P. & Weenink, D. "Praat: doing phonetics by computer" (Version 4.4.13) [Computer program]. Retrieved March 8, 2006, from http://www.praat.org/.

[7] Bladon, A., "Two-formant models of vowel perception: shortcomings and enhancements", Speech Communication, Vol.2, 1983, p.305-313.