On variation and change in diphthongs and long vowels of spoken Dutch
Jacobi, I.

Citation for published version (APA):
Jacobi, I. (2009). On variation and change in diphthongs and long vowels of spoken Dutch

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
2. ON MEASURING AND ANALYZING VOWELS AND DIPHTHONGS

Abstract This chapter provides an overview of various aspects that need to be taken into account when measuring and comparing vowel qualities in spontaneous speech. The goal was to find an efficient and reliable method for measuring and comparing vowel qualities across speakers and sexes. In an optimal procedure for our vowel variation analysis, the variance caused by the speakers’ differing vocal tract properties should be reduced, whilst linguistic trends rather than artifacts should be maintained. First, acoustic vowel properties and the most important sources of variation within and between speakers are given. Second, two different methods to measure vowel quality acoustically, formant analysis and principal component analysis based on spectral filter output, are reconsidered, as well as procedures to normalize for unwanted speaker-effects in linguistic vowel research. Principal components derived from a principal component analysis built on the vowels /a/, /i/, /u/, which have been unaffected by linguistic trends and delimit the acoustic vowel space, are expected to yield the most objective results when measuring acoustic variation within other vowel phonemes.
2. On Measuring and Analyzing Vowels and Diphthongs

2.1 Introduction

In casual speech, the speaker can neglect the articulatory-acoustic quality of a vowel realization up to a certain degree without being misunderstood. This is tolerated by the listener due to the complex speech processing, which weighs the various layers of speech depending on meaning, context, predictability and redundancy, and supports a quick accommodation to the interlocutor’s sound inventory. So generally, variation hardly hinders communication, and perceptually, the accommodation to speech variability is an automatic process (Magnuson & Nusbaum, 2007 [93]).

When analyzing the quality of vowels and (social) variation, aspects of duration, speech mode, speech rate and context need to be considered. The following section will outline these effects, followed by sections on two methods of spectral analysis, common methods to normalize for unwanted speaker effects on vowel quality, and conclusions on how to objectively analyze the diphthongs and diphthongized long vowels for our research.

2.2 Aspects of Duration, Speech Mode, Speech Rate, and Context

In this section we will dwell on durational aspects and aspects of context that need to be considered when analyzing vowel quality in acoustics and perception.

Though for example duration generally adds to the identification of Dutch vowels, in spontaneous speech, vowel duration can be heavily influenced by speech style, speech rate, and context. Generally, stressed vowels are longer and they are articulated more accurately (more peripheral in the articulatory-acoustic vowel space) compared to unstressed vowels (for studies on Dutch see e.g. Koopmans-van Beinum, 1973 [76], van Bergem, 1993 [150]). This implies that they are less affected by coarticulation and more reliable in terms of acoustic regularity.

Differences in vowel quality are also apparent in isolated tokens versus read speech, versus spontaneous speech (see Koopmans-van Beinum, 1980 [77]). The (static) spectral values of many studies are taken from accurately read speech recorded in noiseless environments, or from synthesized stimuli, and thus are based on a vowel quality that is rarely reached in spontaneous speech. As an example, vowels taken from casual speech are often not identified as belonging to the phoneme category intended by the speaker when presented out of context. Contrary to vowel realizations of isolated tokens or read speech, vowel realizations from spontaneous speech are often more centralized in the articulatory-acoustic vowel space (Joos, 1948 [68]), and phoneme categories are more diffuse and overlap considerably. Next to effects of coarticulation, following Lindblom’s (1971 [91]) argumentation, the main reason for the quality differences is the varying speech rate, causing an ‘undershoot’ in reaching the articulatory-acoustic target position, with increases in tempo and decreases in vowel duration respectively. Other studies could not find effects of un-
dershoot with increasing tempo, and suggest that compensatory articulations such as an increase in articulatory velocity or an increase in coarticulatory overlap let the articulators nevertheless reach their target positions (van Son 1993 [158], see Harrington & Cassidy, 1999 on this topic [45]).

Nonetheless, besides the possibility of undershoot due to speech rate differences, the speaker’s awareness of his production, attention, and communicative intention probably differ for each speech condition. Differences in speech conditions can then result in differences in e.g. prosodic realization, or, more generally, differences in suprasegmental realization.

To control for durational and coarticulatory effects on vowel realizations, only stressed vowels will be considered in our analysis. This will also reduce the strongest effects of coarticulation (influence of neighboring sounds) on the target sounds.

Considering diphthongs, results of experiments with mostly synthesized stimuli showed that the temporal pattern of diphthong movement varies depending on dialect or language, and it was not duration itself that differentiated the quality of the vowels. In a cross-language perception test, Peeters (1991 [113, 114]) studied subjects’ preferences for synthesized possible productions of diphthongs. The stimuli were continua of long vowels and diphthongs (/e/, /o/, /ai/, and /au/) with manipulated onsets, offsets and transitions, and subjects were asked to choose the best match of two. The languages included were Dutch, English, and German. The results suggested that not duration but spectral transitions were relevant for the perceived quality.

Gay (1968 [39]) investigated American English diphthongs in three conditions of speaking rate. The results indicated that the offset target positions were variable across different diphthong durations, whereas the onset target position and the rate of change of the second formant were constant. Bladon (1985 [9]), studying fast speech, stated that the integrity of diphthongs is not compromised by offset undershoot since the second targets have little competition.

Following the latter studies, the most promising acoustic values for the analysis of our diphthongal vowels seem to be the spectral composition at the vowel onset position and spectral change. The following section will focus on how spectral information reflects the articulatory-auditory quality of a vowel and in what ways this acoustic information is commonly represented.

### 2.3 Acoustic Cues to Vowel Quality

The usual acoustic parameter to characterize and differentiate between and within vowels is the spectral energy distribution. When measuring and representing the distribution of energy of vowel spectra, the question arises what forms a better sketch of the acoustic and perceptual vowel cues: whole spectrum representations or formant representations. In the
2. On Measuring and Analyzing Vowels and Diphthongs

following sections, these two different ways to acoustically cover vowel quality, and their advantages and disadvantages with respect to the present variation research, are described.

2.3.1 Formants

The importance of the first two formants for vowel quality distinctions has been reported by Helmholtz and others from as early as the middle of the 19th century on (Helmholtz, 1862 [161]). Formants represent concentrations of energy around particular frequencies in the vocal tract. During vowel production, these resonance frequencies occur while the vocal tract is excited by the chain of air pulses that passes through the folds during the open phases of the vibratory cycle. Here, the vocal tract resembles a one-side-closed tube with minimum pressure and maximum velocity at the open end (mouth opening), and maximum pressure and minimum velocity at the closed end (the glottis). Based on Webster’s horn equation to describe pressure waves in a duct, natural frequencies that correspond to the vocal tract area functions can be calculated (compare the early investigations on vowels by Chiba and Kajiyama, 1942 [15]).

Formants are defined by the frequency at which air vibration is maximal, the center frequency, and by the bandwidth, which is defined as the range of neighbouring frequencies falling within 3 dB below the peak amplitude. Corresponding to the characteristics of the vocal tract filter, the resonance frequencies differ with the tract size, and shift with its shape (compare fig. 2.1, p. 21). Accordingly, they indicate the articulatory pattern. A detailed description can be found in Stevens (1998 [137]).

The most commonly used acoustic cues for vowels are the first two formants. From the fourth formant on, the formant attributes carry mainly speaker-specific and little vowel-specific information; they contribute to natural sound perception and are of importance for speaker recognition. The first formant (F1) is associated with the degree of constriction, and indicates the vertical tongue position and mouth opening. A narrowing of the cross-sectional front part of the vocal tract is accompanied by a widening in the back part, and results in a decrease of F1 (Stevens, 1998 [137]). The more the jaw is lowered, and the more open the vocal tract, the higher F1 (Lindblom, 1971 [91]). The second formant (F2) is dependent upon the length of the front cavity (Fant, 1970 [36]), and indicates the articulatory front-back dimension. Lip rounding, which increases vocal tract length, was found to lower all formants (Lindblom, 1971 [91]).

An aggravating factor when trying to detect articulatory patterns based on formant tracking is the variety of human vocal tract sizes and shapes, causing their characteristic tract resonances to differ accordingly. Peterson & Barney (1952 [115]) tried to relate formant patterns to vowel qualities in a seminal experiment. A plotted F1-F2 plane, created from the vowel values of uttered /b/vowel/ /d/-words could be divided into vowel areas. However, the areas overlapped considerably for males, females and children, and absolute
2.3. Acoustic Cues to Vowel Quality

F1 and F2 values as the only parameters were not sufficient to define the produced and perceived vowel categories.

Usually, transformations based on the cochlea’s tonotopic bank of filters are applied on formant values (see e.g. Traunmüller, 1981 [147] and Syrdal & Gopal, 1986 [142]). Frequency scales that have been used for vowel analysis such as the Koenig scale, the cochlear position scale, the mel scale, or the Bark scale are all based on auditory findings. Generally, all these scales share a linear Hertz-scale in the low-frequency region and a logarithmic scaling in the high frequencies (see Miller, 1989 [97]). By mapping the differences in Hertz-values onto the perceptual vowel quality or timbre domain, it is taken into account that the same distance (of say 100 Hz) between low tones is experienced as greater than the same distance between high tones.

The perceived vowel quality is also affected by several other acoustic properties. Proposed on the basis of psychophysical considerations, the relevance of spectral relations (feature interaction) as opposed to formant peak extraction was suggested (Miller, 1989 [97]). Crucial for auditory perception are thus not only the positions of the formant peaks but also the distance between them. Several studies have pointed out the role of f0 in rela-

\[ z = 26.81 \cdot \left( \frac{1}{f + 1960} \right) - 0.53 \]

\[ f = 1960 \cdot \left( \frac{z + 0.58}{26.28 - z} \right) \]

---

Figure 2.1: Top row from left to right: oscillograms of the vowels /a/, /i/, and /u/, three periods each. Middle row from left to right: spectra of /a/, /i/, /u/, with superimposed smoothed spectra (grey lines). Bottom row from left to right: barkfiltered spectra of /a/, /i/, /u/.
tion to F1 (for instance Miller, 1953 [98], Traunmüller, 1981 [147], Hoemke & Diehl, 1994 [51]). Syrdal and Gopal (1986 [142]) found the F3-F2 difference to be a more accurate cue for the perceived front-back dimension than the F2-F1 difference. An improvement of judgements by taking F3 into account had been reported earlier, as well as a combined influence of the higher formants on vowel perception, as in Delattre et al. (1952 [26]), or Bladon (1983 [8]). In variation studies, however, F3 is hardly ever mentioned.

Although formants are important acoustic vowel features, extracting reliable formant values from a harmonic spectrum is notoriously difficult. To draw any conclusions from the speech signal about the momentary vocal tract characteristics, the spectral envelope has to be separated from the source signal. Measuring formant frequencies by spectrographic analysis was one of the first attempts to accomplish this, but this method is not very accurate (c.f. Monsen & Engebretson, 1983 [102]). Usually, formants are calculated from the mean frequency spectrum of the target segment. For vowels of monophthongal quality, the mean of its mid section is taken, whereas for vowels of diphthongal quality the mean of a section in the beginning of the vowel is taken, as well as the mean of the end section of the vowel.

Linear Predictive Coding (LPC) is used to separate the effects of source and filter. With Linear Predictive Coding (Markel & Gray, 1976 [94]), the signal is broken down into a signal source and LPC coefficients. Knowing that neighbouring consecutive samples are statistically not independent, a subsequent value can be approximately predicted by the weighted sum of preceding values. The signal is encoded as a set of coefficients (usually N=10), representing the vocal tract filter, plus an error signal that represents the difference between predicted and real value (Tempelaars, 1991 [144]). With a flat spectrum the minimized error signal looks like an impulse train (a glottis signal) or noise, and when using only the LPC coefficients a spectrum can be calculated without the interference of the source/error signal. Formants are computed from the LPC coefficients. However, the model is not perfect and the coefficients also contain information about less variable unwanted filters of the vocal tract. These can be factored out by pre-emphasis.

The Praat program (Boersma & Weenink, 1992 [12]) is frequently used in vowel variation studies, and with its standard procedure to analyze formants using LPC. In the standard setting for formant analysis, after resampling and pre-emphasis, the coefficients are computed with the Burg algorithm and five formants are assigned to candidates in the frequency range from 0 to 5500Hz (for an adult female speaker).

Though formants are clearly indispensable in human vowel perception, formant tracking algorithms produce errors which are not comparable to listeners' errors (Bladon, 1982 [7]). Using LPC it is assumed that there are no prominent antiformants, which could cause problems e.g. when vowels are nasalized and spectral valleys are of significance (c.f. Johnson, 2003 [64]). Also, for formant tracking, the specification of the number of formants is important to anticipate the right peaks: For e.g. Praat [12], the given standard value for
analyzing a human female voice assumes five formants within the first 5500Hz, for a male voice five formants within 5000Hz. Difficulties occur when f0 and F1 interact, formant peaks lie close, or when formants are low (e.g. in high back vowels).

The spectral interaction of frequency bands lays open one of the problems of depicting and comparing formants. Especially with high fundamental frequencies formants are hard to define. The splitting and merging of formants and antiformants entails the assignment of spurious formant peaks or ‘missing’ peaks, which causes errors in the serial numbering of the formants, and thereby problems in the further processing of the results. There is hardly a formant-based study that does not mention a hand correction of data.

The mentioned integration or adjoining effects on the acoustic level are comparable to some perceived formant integration effects. Following Chistovich et al. (1979 [16]), perceived formant averaging or integration occurs within a critical distance of 3 to 3.5 Bark for two-formant signals. Also, spectral amplitude relations and spectral density seem to determine vowel quality (Chistovich et al. 1979, Ito et al., 2001 [16, 57]). Similarly to the results of Bladon & Lindblom (1981 [10]) and Bedder & Hawkins (1990 [6]), Kieft & Khender (2005 [70]) conclude that gross spectral properties (tilt) at least contribute to more detailed spectral cues (formants peaks) in vowel perception. However, after their experiments with synthesized monophthongs and diphthongs, they also found that the role of spectral tilt is less important in signals with changing spectral characteristics, and they argue that in this case, it is change over time that dominates the perception of speech.

2.3.2 Whole-Spectrum Representations

Alternatively to formant tracking, from the 1960’s on, Dutch researchers used principal component analyses (PCAs) on bandfiltered spectra to analyze and compare vowels (starting with the studies of Plomp et al., 1967 [117], Pols et al., 1969 [123], followed by others [73, 157, 122, 138, 129, 160, 124, 150]). Using this method, it is assumed that there is a finite amount of independent variation that appears in the spectral data. Instead of using raw spectral data, the ensemble of spectral variation is used to arrange the data: The original n-dimensional feature space is rotated in such a way that in the new n-dimensional space most of the variability is placed in the first dimension; the smallest (noisy) variance will end up in the highest dimension.

By analyzing the whole spectrum in frequency bands, one band represents one dimension and the different levels of energy within the band can be described as a coordinate in the single dimension. Changes in the spectrum then reflect different concentrations of constituents, and each spectral sample becomes a point in a multidimensional Euclidean space (compare Pols, 1971 [120]). Based on the principle of combining two or more (correlated) variables into a single factor, the principal component analysis breaks down the information in the bandfilters into its most basic variations. Physical or psychophysical
properties of the human listener can be included by choosing filters of the same properties as the bandfilters of the human ear, frequency-dependent excitation levels, or other neurophysiological or psychophysical scales.

Numerous studies have shown that using multi-dimensional scaling, the first two principal components (pc) dimensions correspond to the frequencies of F1 and F2, though the acoustic properties they have been calculated from do differ from formant tracking algorithms: In the 70’s, Klein, Plomp, Pols and Tromp compared a principal component representation of 12 different bandfiltered /h/ vowel /h/-vowels, produced by 50 Dutch male speakers, to the frequency and level data of the first three formants from the same vowel segments (Klein et al., 1970 [73], Pols et al., 1973 [73, 122]). At the time, the formants were derived by drawing the spectral envelope by eye. The results were then compared to the results of the PCA. Corresponding to the ear’s critical bandwidths, the sound spectra had been filtered in 21 1/3-octave bands from 10 to 10000 Hz with the sound levels as dimensions. To reduce the influence of the fundamental frequency, the energy in the first three filters was added, and the energy in the fourth and fifth filter, resulting in 18 filters altogether. The overall sound pressure output levels were normalized. An analysis of the principal components yielded a reduction to four factors explaining 77% of the total variance in the 600 vowel spectra. The average vowel configuration resembled the logarithmically plotted F1-F2 formant plane. The largest part of variance caused by different vowel classes could be explained by logF1 and logF2, confirming F1 and F2 as the most characteristic vowel features. Adding logF3 further improved the identification score (Klein et al., 1970 [73], Pols et al., 1973 [122]).

2.3.3 Concluding Remarks

As reported, formants are the most important cues to vowel quality, and many experiments with (synthetic) speech have proven that changes in formant frequencies affect perceived vowel categories more than formant bandwidth or spectral bends do (e.g. Klatt, 1982 [72]).

However, finding formant peaks remains a problem when energy is distributed over a range of frequencies, or when formants come close together. Whenever formant tracking is used for vowel analysis, a considerable amount of hand correction of the formant data is reported. This problem does not occur when measuring the spectral distribution by principal components derived from spectral filters.

Next to the finding that no errors need to be corrected by hand, an important advantage using PCA on the whole spectra is that, contrary to formant analysis, no previous knowledge about vowel categories is needed. Thus, the analysis can be reliably automated. In vowel variation research with large amounts of vowel data, a reliable automation of the measuring procedure is highly desirable. Least influenced by expectations, this is likely to be the more objective way to analyze vowel variation, and hence we will apply this method
in the present research. Tracing back the articulatory patterns from the PCA coefficients' could be accomplished by building the PCA on only certain vowels, and by including and referring to vowels with clear or steady articulatory properties. Ideally, the measured vowel quality differences should correspond with perceived differences as well. Research showed that the first pc's are comparable to the first formants, the most important cues to perceived vowel quality. Direct evidence for the correspondence of pc’s and perceived differences was given in Klein et al. (1970 [73]).

The basis of our study will be vowels in spontaneous speech. Though we will only use stressed vowels, considering the artificial nature of the measured sounds in the cited studies on formants or PCA on barkfilters, and taking into account the reported effects of speech condition, we expect the vowels of spontaneous speech to be more centralized and/or coarticulated than vowels of semi-spontaneous or read speech. Also, with various speakers, we will have to deal with differences in vocal tracts shaping the acoustic output. The following section will consider a common problem in vowel studies: dealing with inter-speaker differences. An overview will be given of popular methods used in vowel formant analyses to normalize for speaker-specific effects, especially dealing with speaker-effects due to sex, in order to make a speaker-independent comparison of vowel quality possible.

2.4 Normalization Procedures

Numerous methods have been developed to reduce the impact of specific speaker effects and make a representation of e.g. vowels of a speaker community possible. Some procedures use only vowel-intrinsic information and categorize the vowel e.g. by transforming f0 and the formant patterns. Other extrinsic normalization procedures take into account information distributed over several vowels. A more detailed discussion of the different approaches and classification procedures can be found in Nearey, 1989 [105]. Joos (1948 [68]) was one of the first to suggest a speaker-specific normalization procedure. He suggested that listeners relate the phonetic quality to the speaker’s point vowels /i, a, u/. A decade later in 1957, Ladefoged and Broadbent confirmed his theory [86]. They shifted the complete vowel system in a synthesized sentence except for the test vowel. If, as a result, the vowel was placed within the acoustic category of another phonologically distinct vowel, the listeners reliably normalized: The same vowel was perceived as belonging to different phonological classes, dependent on the context (embedded in a sentence or separately)².

Considering vowels, most research has been focused on methods to generalize vari-

² By changing the formants’ range of the carrier sentence, e.g. the vowel of <head> could be made heard as the vowel of <hid>. A partial reproduction (by Malcolm Slaney) of the original experiment can be found on the web: http://cobweb.ecn.purdue.edu/~malcolm/interval/1997-056/VowelQuality3.html
2. On Measuring and Analyzing Vowels and Diphthongs

ation within vowel classes, for example to enhance robust speech recognition (see e.g. Weenink, 2006 [164]). Contrary to the more general aim of these procedures to minimize variation, for phonetic variation research, the organic variation (variation caused by physical attributes of the individual vocal tract) has to be disentangled from learned and/or acquired variation, the latter being the object of our interest. This implies that the rather complex or abstract relationship of the acoustic regularities within and between speakers versus the irregularities within and between speakers have to be defined and categorized.

On the basis of these findings, a normalization procedure can be built. Hence, for the present variation study, the phonemic speaker characteristics had to be further divided into physiological (anatomical) variation versus intra-phonemic variation, and the physiological/anatomical variation will have to be factored out.

Disner (1980 [30]) evaluated several formant normalization procedures to find out about their overall ability to reduce variance while yielding truly linguistic trends and not artifacts. She compared Gerstman’s (1968 [40]), Lobanov’s (1971 [92]), Neary’s (1977 [104]) and the PARAFAC procedure (Harshman, 1970 [46]). All procedures use a mean or standard deviation of the speaker’s whole vowel system. She came to the conclusion that for cross-linguistic studies, or for comparisons across dialects, the application of normalization procedures which use the mean or standard deviation of the vowel system are too effective in reducing interspeaker variance, and might result in procedural artifacts (Disner, 1980 [30]). In 2003, for the analysis of vowel variation in Dutch read speech, Adank favored the Lobanov procedure after evaluating formant normalization procedures (Adank, 2003 [1]).

A method of normalizing vowel data for variation analysis derived from Gerstman’s method was introduced by van Heuven et al. (2002, 2003 [156]). They compared formant values of the onset and the offset of Dutch /ɛi/ diphthongs, and concentrated on the height of the diphthong onset and the extension of the glide of the diphthongs. As reference they took a speaker’s most extreme high front vowel /i/ (in terms of F2), and the most extreme open front vowel /a/ (in terms of F1), to which the /ɛi/ onset was related. After calculating the Euclidean distance of F1 and F2 in Bark between diphthong onset and /a/, they related it to the distance between /a/ and /i/. The resulting values showed differences between the pronunciation of /ɛi/ by males and females in terms of relative onsets and diphthongization (both related to each speaker’s /a/ and /i/).

2.5 Conclusion

For our vowel analysis, we are searching for a reliable method to (automatically) measure variation in vowel realizations across speakers and sexes. As already mentioned, every sound is acoustically unique, even if uttered by the same speaker in a sequence. Considering vowel dispersion, there are considerable effects in terms of speech mode, accent,
stress, and speaker sex. Our speech mode will be spontaneous speech, and the conclusion is to consider only stressed vowels in our variation analysis.

The spectral composition at the target onset position and the rate of change, rather than duration, were found to determine the quality of diphthongal sounds, and when measuring our spontaneous vowel data, we will start our investigations with these acoustic properties.

For (socio-)phonetic variation research and between-speaker comparison, speaker-specific physical variation and ‘externally’ (environmentally) caused variation need to be separated. The variance caused by the physical differences between males and females needs special attention: Females are supposed to have more dispersed vowels, and for variation analysis, a normalization procedure should account for these sex differences, so that linguistic effects in terms of gender differences can still be differentiated from biological sex.

Our preferred method of analysis is a PCA on bandfilters. The vowels that are used for a PCA should also mark the size of the speaker’s vowel space (limited by the individual’s most extreme articulatory-acoustic realizations) to be able to capture all of the speakers’ vowel qualities, as for example the anchor vowels /ə/, /ɪ/, and /ʊ/. Ideally, these anchor vowels should be stable vowels with a minimal probability of carrying gender effects. Then, the variance within these vowel classes is merely due to sex differences and presumably smaller than the variation between the vowel classes, so that, when running a PCA, the variation of the total size of the vowel space due to speaker sex could be filtered out, whereas gender differences of paths within the vowel space should not be affected.

In the following chapter we will apply, and try to test this in more detail in a small variation study on the Dutch vowel /ɛi/, to prepare for the automatic analysis of vowel phoneme realizations in a larger corpus.