Chapter 4

Measuring segment distances acoustically

In Chapter 3 we described three feature systems, namely the system of Hoppenbrouwers & Hoppenbrouwers (H & H), Vieregge & Cucchiarini (V & C) and Almeida & Braun (A & B). The systems can be used for the corpus frequency method (see Section 2.3.2), the frequency per word method (see Section 2.3.3), and the Levenshtein distance (see Section 5.1). The Levenshtein distance is the focus of this thesis. When using the Levenshtein distance the distances may be calculated on the basis of these feature systems.

While the use of features as linguists have developed them yields satisfactory results, one may nonetheless question the physical basis of the feature assignments, in fact seeking a more objective foundation. The advantage of the system of H & H is that vowels and consonants can be compared with each other. For vowels, the more consonant-specific features get default values, and for consonants the more vowel-specific features get default values. The disadvantage of Hoppenbrouwers’ system is that feature values are not based on physical measurements. The SPE feature system, which H & H use, was not developed to reflect physical, perceptual or articulatory differences directly, but rather to facilitate the coding of phonological rules. It is of course, to be expected that this coding reflects the physical properties of speech in some way. The advantage of the system of V & C is that it is partly based on real measurements, found by experiments. The system of the A & B is interesting because of its use of the well-known IPA system. However, just as for the H & H system, the IPA system is not based on real measurements.

Another inadequacy of the three feature systems concerns the definition of ‘silence’, which is needed in the Levenshtein algorithm (see Section 5.1). The
way in which ‘silent vowels’ and ‘silent consonants’ are defined was described for each system in Chapter 3. A definition of ‘silence’ in terms of features will always be somewhat artificial.

When acquiring language, children learn to pronounce sounds by listening to the pronunciation of their parents or other people. The acoustic signal seems to be sufficient to find the articulation which is needed to realize the sound. Acoustically, speech is just a series of changes in air pressure, quickly following each other. With a spectrograph or a computer a spectrogram can be made, representing an analysis of the speech sample. A spectrogram is a “graph with frequency on the vertical axis and time on the horizontal axis, with the darkness of the graph at any point representing the intensity of the sound” (Trask, 1996, p. 328). In a spectrogram the formant structure of a sound can be identified. A formant is “a concentration of acoustic energy within a particular frequency band, especially in speech” (Trask, 1996, p. 148). Especially for vowels, formants can be easily recognized in a spectrogram as thick dark bars.

In this chapter we present the use of spectrograms and formant tracks for finding sound distances. We will show that the disadvantages of feature systems as mentioned above do not apply to acoustic representations when plausible segment classifications may be obtained on the basis of acoustic representations. Both spectrograms and formant tracks are based on physical measurements. When using a spectrogram or formant definition instead of a feature definition, the distance between a vowel and a consonant can be measured in the same way as the distance between a vowel and a vowel, or between a consonant and a consonant. When using a spectrogram ‘silence’ is defined in a natural way: for all frequencies for all times the intensities are equal to 0. Something similar applies for the formant definition of ‘silence’: there are no vibrations, so the frequencies are set to 0.

We explored the use of the different acoustic representations for finding segment distances which are intended for use in the Levenshtein algorithm. In this section we show how distances between sounds can be found using spectrograms and formant tracks. In Section 4.1 it is shown that spectrograms can be regarded as pictures of sounds. In Section 4.2 we discuss the samples of the sounds we used. Section 4.3 discusses several spectrogram models, as well as the formant track model. Classifications on the basis of the different representations are also given in this section. The way we deal with diphthongs and affricates is explained in respectively Section 4.4 and Section 4.5, while Section 4.6 describes how suprasegmentals and diacritics are processed. The comparison of sounds is explained in Section 4.7. In Section 4.8 the different acoustic representations are compared with each other and with respect to discrete representations. Segment distances which are obtained from the different systems are correlated with each other. Finally, in Section 4.9 we draw some conclusions.
4.1 Visible speech

In Potter et al.’s (1947) *Visible Speech*, spectrograms are shown for all common English sounds (see pp. 54-56: *The ABC’S of Visible Speech*). Examining the spectrograms the formant structure of vowels and sonorants, the high frequency noise of certain fricatives and the periods of voicing and voicelessness can be clearly identified (Trask, 1996, p. 328). Looking at the spectrograms we can already see which sounds are similar and which are not. We expect that visible (dis)similarity between spectrograms reflects perceptual (dis)similarity between sounds to some extent. In Figure 4.1 the spectrograms of some sounds are shown as pronounced by John Wells on the audio tape *The Sounds of the International Phonetic Alphabet* (Wells and House, 1995). The x-axis gives the time and the y-axis the frequency. For each frequency at each time the intensity is visualized by the darkness. The spectrograms are made with the computer program PRAAT.\(^1\)

4.2 Samples

For finding spectrogram distances or formant track distances between all IPA sounds, for each sound we need samples from one or more speakers. We found these samples on the tape *The Sounds of the International Phonetic Alphabet* on which all IPA sounds are pronounced by John Wells and Jill House. On the tape the vowels are pronounced in isolation. The consonants are sometimes preceded, and always followed by an [a]. We cut the part preceding the [a], or the part between the [a]’s. We are aware of the fact that information on the F2 transition is lost. Rietveld and Van Heuven (1997) explain that the F2 transition in the transition zone from consonant to vowel gives information about the place of articulation. In our research we focus only on the sound itself. We also realize that the pronunciation of sounds depends on their context. For both vowels and consonants Stevens (1998) gives a discussion of some influences of context on speech sound production (pp. 557-581). Since we use samples of vowels pronounced in isolation, and samples of consonants selected from a limited context, our approach is a simplification of reality. However, Stevens (1998, p. 557) also observes that

“by limiting the context, it was possible to specify rather precisely the articulatory aspects of the utterances and to develop models for estimating the acoustic patterns from the articulation”.

The two speakers on the tape give us two sets of IPA samples. However, some sounds were missing or not properly pronounced. Therefore, the [?] , [r] and [u]

\(^1\)The program PRAAT is a free public-domain program developed by Paul Boersma and David Weenink at the Institute of Phonetic Sciences of the University of Amsterdam and available via [http://www.fon.hum.uva.nl/praat/](http://www.fon.hum.uva.nl/praat/).
Figure 4.2: Different acoustic representations of four sounds as pronounced by John Wells. Starting with the first row we see respectively spectrograms, Barkfilters, cochleagrams and formant tracks obtained on the basis of the *monotonized* samples.
of Jill House were substituted by the corresponding sounds of John Wells. The original \[v\] of Jill House is used as \[w\] for both John Wells and Jill House.

The burst in a plosive is always preceded by a period of silence (voiceless plosives) or a period of murmur (voiced plosives). When a voiceless plosive is not preceded by an \[a\], it is not clear how long the period of silence which really belongs to the sound lasts. Therefore, we have always cut each plosive in such a way that the time span from the beginning to the middle of the burst is equal to 90 ms. In a spectrogram the burst is recognized as a small dark vertical bar at the end of the period of silence or murmur (see e.g. the \[p\] in spectrograms in Figure 4.1). The middle of the burst was estimated by eye. Among the plosives which were preceded by an \[a\] or which are voiced (so that the real time of the start-up phase can be found), we found no sounds with a period of silence or murmur which was clearly shorter than 90 ms.

In voiceless plosives, the burst is followed by an \[h\]-like sound before the following vowel starts. When including this part in the samples, the consequence is that bursts will often not match when comparing two voiceless plosives. However, since aspiration is a characteristic property of voiceless sounds, we retained aspiration in the samples (see Figures 4.3 and 4.4 for both speakers). In the voiced sounds the burst is immediately followed by the vowel. In some cases it was not clear where the burst ended and the vowel started. For the voiced sounds it cannot be guaranteed that nothing from the following vowel is included, although any error here will be minimal. In general when comparing two voiced plosives, the bursts will match (see Figures 4.3 and 4.4 again). When comparing a voiceless plosive and a voiced plosive the bursts will not match.

To keep trills comparable to each other, we always cut three periods, even when the original samples contained more periods. When there were more periods the most regular looking sequence of three periods was cut.

To get a sample of ‘silence’ we cut a small silent part on the IPA tape. This assures that silence has about the same background noise that the other sounds have.

To make the samples as comparable as possible, all vowel and extracted consonant samples are monotonized on the mean pitch of the 28 concatenated vowels. The mean pitch of John Wells was 128 Hertz, the mean pitch of Jill House was 192 Hertz. In order to monotonize the samples the pitch contours were changed to flat lines. Figure 4.1 shows spectrograms of non-manipulated samples while Figure 4.2 shows spectrograms of the corresponding monotonized samples.

The volume was not normalized because volume contains too much segment specific information. For example, it is specific for the \[v\] that its volume is greater than that of the \[f\].
Figure 4.3: Spectrograms of voiceless (left) and voiced (right) plosives as pronounced by John Wells. Starting with the first row spectrograms are given for [p] and [b], [t] and [d], [c] and [ç], [k] and [g], [q] and [ç]. When comparing voiceless plosives, the aspiration parts will match, and when comparing voiced plosives the bursts will match. When comparing a voiceless plosive with voiced plosive, the burst of the voiced plosive will partly match the aspiration part of the voiceless plosive.
Figure 4.4: Spectrograms of voiceless (left) and voiced (right) plosives as pronounced by Jill House. Starting with the first row spectrograms are given for [p] and [b], [t] and [d], [ɾ] and [l], [ʃ] and [ʒ], [k] and [g], [ɡ] and [c]. When comparing voiceless plosives, the aspiration parts will match, and when comparing voiced plosives the bursts will match. When comparing a voiceless plosive with voiced plosive, the burst of the voiced plosive will partly match the aspiration part of the voiceless plosive.
4.3 Representation of segments

On the basis of the samples, manipulated spectrograms can be made or formant tracks can be found. We do not use the most common type of spectrogram with a Hertz-scale, but instead use more perceptually oriented models. In Section 4.3.1 we describe the Barkfilter. A Barkfilter has a frequency scale which is roughly linear below 1000 Hz, and roughly logarithmic above 1000 Hz. The logarithms of the intensities are mapped. In Section 4.3.2 we explain the cochleagram which is based on the Barkfilter, but may be more similar to human perception. The cochleagram uses the same frequency scale as the Barkfilter, but in the cochleagram, rather than the intensities themselves, the loudnesses as perceived by the ear are given. Besides two spectrogram like representations we also consider the formant track representation. Essential for perceiving vowels is that spectral peaks are recognized by the ear. The same applies for the sonorant consonants. These peaks are called formants. The formant track representation is discussed in Section 4.3.3. Our brief explanation of the three different representations is based on Rietveld and Van Heuven (1997).

4.3.1 Barkfilters

4.3.1.1 Representation

In the most commonly used type of spectrogram the linear Hertz frequency scale is used. The difference between 100 Hz and 200 Hz is the same as the difference between 1000 Hz and 1100 Hz. However, our perception of frequency is non-linear. We hear the difference between 100 and 200 Hz as an octave interval, but the difference between 1000 to 2000 Hz is perceived as an octave as well. Our ear evaluates frequency differences not absolutely, but relatively, namely in a logarithmic manner. Therefore, in the Barkfilter, the Bark-scale is used which is roughly linear below 1000 Hz and roughly logarithmic above 1000 Hz (Zwicker and Feldtkeller, 1967). In the program PRAAT, for a given frequency in Hertz, the corresponding frequencies in Bark are found with a formula of Schroeder et al. (1979):

\[ \text{Bark} = 7 \times \ln \left( \frac{\text{Hertz}}{650} + \sqrt{1 + \left( \frac{\text{Hertz}}{650} \right)^2} \right) \]

(4.1)

Hertz frequencies are plotted against Bark frequencies in Figure 4.11. The graph shows two curves. The upper curve shows Bark values calculated with the formula of Schroeder et al. (1979) (applied when using Barkfilters and cochleagrams, see also Section 4.3.2), the lower curve shows Bark values calculated with the formula of Traunmüller (1990) (applied when using formant tracks, see Section 4.3.3). In the plot the Hertz-scale runs from 20 to 20,000 Hertz, the frequency range which
can be perceived by a human being. This corresponds with a frequency range of 0.22 to 28.84 Bark for the Schroeder et al. curve.

In the commonly used type of spectrogram the power spectral density is represented per frequency per time. The power spectral density is the power per unit of frequency as a function of the frequency. In the Barkfilter the power spectral density is expressed in decibels (dB’s). “The decibel scale is a way of expressing sound amplitude that is better correlated with perceived loudness” (Johnson, 1997, p. 53). The decibel-scale is a logarithmic scale. Multiplying the sound pressure ten times corresponds with an increase of 20 dB. On a decibel scale intensities are expressed relative to the auditory threshold. The auditory threshold of 0.00002 Pa corresponds with 0 dB (Rietveld and Van Heuven, 1997, p. 199).

A Barkfilter is created from a sound by band filtering in the frequency domain with a bank of filters. In PRAAT the lowest band has a central frequency of 1 Bark per default, and each band has a width of 1 Bark. There are 24 bands, corresponding with the first 24 critical bands of hearing as found along the basilar membrane (Zwicker and Fastl, 1990). A critical band is an area within which two tones influence each other’s perceptibility (Rietveld and Van Heuven, 1997, pp. 204–205). Due to the Bark-scale the higher bands summarize a wider frequency range than the lower bands.

In the Figures 4.1 and 4.2 Barkfilters are shown, obtained on respectively non-manipulated and monotonized samples. In this type of spectrogram the Bark-scale is used as frequency scale, while intensities are given in Decibels. The frequencies range from 0 to 24.67 Bark. They are divided in 24 equal intervals, where for each interval the mean intensity is given. The sound signal is probed each 0.005 seconds with an analysis window of 0.015 seconds. Here we used the standard settings in the program PRAAT. Other settings may give different results, but since it is not a priori obvious which results are optimal, we restricted ourselves to the default settings.

### 4.3.1.2 Classification

In Section 4.7 we describe how the distance between two spectrograms is measured in our research. The measure described in that section enables us to calculate the distances between all sounds. For the RND we have 18 vowels and 27 consonants. Also ‘silence’ is added. This gives a total of 46 sounds. In the NOS data the modern IPA system is used. In the IPA system 28 vowels and 58 pulmonic consonants are given. We added the [w] so we get 59 consonants. ‘Silence’ is added as well. So we get in total 88 sounds. Because we have two sets of samples (namely of John Wells and Jill House), two distance matrices are obtained for the RND and the NOS. Next the matrices of the two speakers are averaged, resulting in distances which are more general and less speaker dependent. We are aware of
the fact that the number of speakers – two – is minimal. Useful future research would be to repeat this experiment on the basis of many more speakers.

Besides averaged matrices for all sounds, we also made matrices containing the averaged distances between vowels only and consonants only. Since it is difficult to appreciate all distances individually we have chosen to visualize the results to give an impression of the results.\(^2\) On the basis of the IPA versions of these matrices we performed multidimensional scaling. Multidimensional scaling was not performed on the RND matrices because the RND sounds are just a subset of the IPA sounds. The multidimensional scaling technique is described in more detail in Section 6.2 and shows us the relations between the sounds in two-dimensional space. This allows us to compare the ordering of the sounds with the way in which they are ordered in the IPA system. The multidimensional scaling plots also show differences between the classifications of the different representations.

Note that for dialect comparison the real sound distances are used, not the multidimensional scaling distances. The multidimensional scaling plots are used here only to visualize the distances and suggest that the spectrogram or formant track distances yield a reasonable measure of pronunciation.

**Vowels** When using multidimensional scaling on the basis of the vowel distances, one dimension explains already 85% of the variance, two dimensions 98% and three dimensions 98% as well. In Figure 4.5 a two-dimensional multidimensional scaling plot is shown. The first dimension (the vertical dimension in the plot) represents the height, and the second dimension (horizontal) the advancement. The positions of the [i], [u], [o] and [a] resemble those in the IPA quadrilateral clearly. We see a clear division between high and low vowels. Note that the [o] belongs to the higher vowels of the lower group, while in the IPA quadrilateral this sound is located exactly in the center. This may be explained by the fact that in our calculations information additional to the F1 and F2 were used. When scaling to three dimensions, it appears that the third dimension does not distinguish between spread and rounded vowels as in the IPA quadrilateral, but distinguishes between central vowels on the one hand, and front and back vowels on the other hand.

**Consonants** When using multidimensional scaling on the basis of the consonant distances, one dimension explains 59% of the variance, two dimensions 94% and three dimensions 99%. In Figure 4.6 a two-dimensional multidimensional scaling plot is shown. The first dimension (the vertical dimension in the plot) makes a distinction between voiceless (upper) and voiced (lower) sounds. The second dimension (horizontal) distinguishes between continuous (left) and non-continuous consonants (right). Comparing these results with the IPA table, the

\(^2\)For 28 vowels we get \(\binom{28}{2} = 378\) distances, for 59 consonants we get \(\binom{59}{2} = 1711\) distances, and for vowels plus consonants plus ‘silence’ we get \(\binom{88}{2} = 3828\) distances.
Figure 4.5: Two-dimensional multidimensional scaling plot obtained from the Barkfilter distances between all pairs formed by the 28 vowels. Two dimensions explain 98% of the variance. The first dimension (y-axis) corresponds with height and the second dimension (x-axis) with advancement. We might have expected the schwa [ə] to be placed more highly. A third dimension would distinguish between central vowels on the one hand, and front and back vowels on the other hand.
place of articulation hardly plays any role, in contrast with the manner of articulation which is important. Striking is the position of the [j] between the approximants and the voiced plosives. With respect to the voiced plosives the [j] is most similar to the [g]. We illustrate this relation by two examples. In German, *Morgen* ‘morning’ is usually pronounced as [mɔʁən], but in Berlin the same word is also pronounced as [məjɔn]. The German word *gemacht* ‘made’ is mostly pronounced as [gəmaxt]. In Berlin, however, this word is also pronounced as [jamaxt]. Apart from the [g], among the voiced plosives the [d] is most similar to the [j]. The relation between the [j] and the [d] can easily be illustrated as well. E.g., the Dutch words *goede* ‘good’ [xudə] and *rode* ‘red’ [roːdə] are also pronounced as [xuːja] and [roːja]. Also striking is the position of the [f] rather close to the voiceless plosives. This relation can be found e.g. in the word *father* [faðə], where the [f] arose from Indo-European [p] (cf. Latin [pater]). Unfortunately, a similar close relation between the [x] and the [k] cannot be found here. Looking at the approximants it is striking that the retroflex variants do not cluster with the other approximants, but are located in the neighborhood of the trills. When scaling to three dimensions the third dimension distinguishes between voiceless plosives, retroflex, velar, uvular, pharyngeal and glottal voiceless fricatives and r-like consonants on the one hand, and (lateral) alveolar, postalveolar and palatal fricatives, the palatal voiced plosive and palatal (lateral) approximants on the other hand. We cannot explain what this distinction is based on.

**All sounds** When using multidimensional scaling on the basis of all sound distances, one dimension explains 76% of the variance, two dimensions 96% and three dimensions 98%. In Figure 4.7 a two-dimensional multidimensional scaling plot is shown. The first dimension (the vertical dimension in the plot) corresponds with intensity. The [a] is loudest and ‘silence’ is most silent (of course). The second dimension (horizontal) represents clearness. The [ʃ] is the clearest and the [u] is the darkest sound. However, one might expect some consonants (e.g. the [f]) to be located in the darker area. In this context, a sound is clear when it has many harmonic tones, and it is dark when harmonic tones are lacking. The voiceless plosives and the voiced plosives can clearly be identified as different groups. However, the other sounds form a continuum. Drawing a line from [i] to [a], from [a] to [o], from [ɔ] to [u], and from [u] to [i] we recover the IPA vowel quadrilateral. Here the nasals appear as high vowels. Also, note the position of the [r], [ɾ] and [r] in the IPA vowel quadrilateral. A close relation between these liquids and (central) vowels can be illustrated by the fact that e.g. the Dutch word *vier* ‘four’ is sometimes pronounced as [fiːr] and sometimes as [fiːə]. Here we see that the [r] can correspond with the [ə]. Less easy to explain is the appearance of the voiced fricative [ɣ] on the border of the quadrilateral, close to the [i]. We note that the [u] and [w] are very close, but the [i] and the [j] are not close. Maybe this is due to the fact that the [j] has a lower intensity than the
Figure 4.6: Two-dimensional multidimensional scaling plot obtained from the Barkfilter distances between all pairs formed by the 59 consonants. Two dimensions explain 94% of the variance. The first dimension (y-axis) distinguishes between voiceless and voiced consonants, the second dimension (x-axis) between non-continuous and continuous consonants. A third dimension would less easily be interpreted.
Further the position of ‘silence’ is very near the glottal stop. When scaling to three dimensions the third dimension distinguishes between low vowels, voiceless plosives, retroflex, velar, uvular, pharyngeal and glottal voiceless fricatives and r-like consonants on the one hand, and high vowels, (lateral) alveolar, postalveolar and palatal fricatives, the palatal voiced plosive and (lateral) approximants on the other hand.

Since the vowel classification is like the IPA quadrilateral, and the consonant classification reflects the different manners of articulation we conclude that the Barkfilter representation is useful for finding segment distances.

### 4.3.2 Cochleagrams

#### 4.3.2.1 Representation

The cochleagram represents the behavior of the basilar membrane in the cochlea. The cochlea is the inner part of the ear. Just as in the Barkfilter, in a cochleagram the Bark-frequency scale is used. In the computer program PRAAT Bark values in a Barkfilter are found using the formula of Schroeder et al. (1979) (see Section 4.3.1). In PRAAT the same formula is used for the frequency scale of cochleagrams.

In a Barkfilter for each time and for each frequency the intensity is given. In a cochleagram for each time for each frequency the loudness is given. When two sounds have the same intensity but different frequencies, they will probably be perceived as differing in loudness. Human aural sensitivity varies with frequency. Loudness is the perceived intensity. Loudnesses are expressed in reference intensities. In a cochleagram the reference intensities are the intensities of a frequency of 1000 Hz. This is the basis for the measurement of loudness in phon. If a given sound is perceived to be as loud as a 60 dB sound at 1000 Hz, then it is said to have a loudness of 60 phon. The relation between the reference loudness and the loudness of another given intensity at a specific frequency is determined experimentally.

Since only one pure tone leads to activation of hair cells over a large surface on the basilar membrane, the ear is not able to perceive other neighboring frequencies. One tone is masked by the other. There are two types of masking: lateral masking and forward masking. Lateral masking occurs when at the same time different but neighboring frequencies are recorded. One tone may make other nearby tones (nearly) inaudible. In general a low tone will mask a high tone rather than the opposite. Forward masking appears when tones occur after each other. E.g., after hearing a strong sound our ears may be stunned for a short time. The more successive sounds resemble each other, the stronger the masking will be. In a cochleagram both the lateral and the forward masking is modeled.
Figure 4.7: Two-dimensional multidimensional scaling plot on the basis of the Barkfilter distances between all pairs formed by the 28 vowels, the 59 consonants and ‘silence’. In the plot, the ! is used for ‘silence’. Two dimensions explain 96% of the variance. The first dimension (y-axis) represents intensity (lower sounds are louder) and the second dimension (x-axis) clearness (sounds on the right are darker). The shaded area represents a quadrilateral formed on the basis of the vertices of the IPA quadrilateral.
In our research we did not consider forward-masking. In our case all sounds were pronounced in isolation (vowels) or cut from their context (consonants). The effect of forward-masking would mainly be found at the begin of a segment and models the phenomenon that a sound has a gradual onset. For long sounds the effect is relatively greater than for short sounds. Thus the relative influence depends on the absolute length of a sample. However, not all samples have reliable lengths. On the IPA tape the vowels are pronounced in isolation. Therefore, our vowel durations will be longer than the durations of vowels which are pronounced in words. Consonant durations reflect the property of the segment to some extent. However, for trills no more than three periods were cut even if there were more periods (see Section 4.2). Besides cutting both vowels and consonants always involves inaccuracies. Therefore, we did not apply forward-masking.

In the Figures 4.1 and 4.2 cochleagrams are shown, obtained on respectively non-manipulated and monotonized samples. In this type of spectrogram loudnesses are given instead of intensities, expressed in phon. Further lateral and forward frequency masking is modeled. The darker lines in the pictures represent formant tracks (see Section 4.3.3.1). The frequencies range from 0 to 25.6 Bark. They are divided in 256 equal intervals, where for each interval the mean loudness is given. The sound signal is probed each 0.01 seconds with an analysis window of 0.03 seconds. The forward-masking time is set at 0.00 seconds. This means that the effect of forward-masking is not regarded in our results. Except for the forward-masking time, here we used the standard settings in the program PRAAT. Other settings may give different results, but just as for the Barkfilter it is not clear which results will be optimal beforehand. Therefore, we restricted ourselves to the default settings.

4.3.2.2 Classification

Just as for the Barkfilter representation, the distances between the IPA sounds are calculated (see Section 4.7). Multidimensional scaling is performed on the basis of vowel distances, the consonant distances and the distances between both vowels and consonants including ‘silence’. Since we followed the same procedure as for the Barkfilter representation, the reader is referred to Section 4.3.1 for more details.

Vowels When using multidimensional scaling on the basis of the vowel distances, one dimension already explains 86% of the variance, two and three dimensions 98%. In Figure 4.8 a two-dimensional multidimensional scaling plot is shown. The plot is very similar to the Barkfilter plot (see Figure 4.5), so the conclusion is that it does not matter whether the Barkfilter or the cochleagram representation is used when finding distances between vowels. For the cochleagram vowel plot the same remarks apply as for the Barkfilter vowel plot (see Section 4.3.1).
Figure 4.8: Two-dimensional multidimensional scaling plot obtained from the cochleagram distances between all pairs formed by the 28 vowels. Two dimensions explain 98% of the variance. The first dimension (y-axis) corresponds with height and the second dimension (x-axis) with advancement. We might have expected the schwa [ə] to be placed more highly. A third dimension would distinguish between central vowels on the one hand, and front and back vowels on the other hand.
4.3. REPRESENTATION OF SEGMENTS

Consonants When using multidimensional scaling on the basis of the consonant distances, one dimension explains 81% of the variance, two dimensions 96% and three dimensions 98%. In Figure 4.9 a two-dimensional multidimensional scaling plot is shown. Just as the plot based on the Barkfilter distances (see Figure 4.6) the first dimension (the vertical dimension in the plot) makes a distinction between voiceless (upper) and voiced sounds (lower). The second dimension (horizontal) distinguishes between continuous (left) and non-continuous consonants (right). The plot is very similar to the Barkfilter consonant plot. The only important difference is that the division between voiceless and voiced sounds in the cochleagram plot is sharper than in the Barkfilter plot. Probably this is explained by that fact that the cochlear model uses loudness instead of intensity. Perceptually, the distinction between voiceless and voiced sounds is greater than pure intensities indicate. For further comments see the explanation of the Barkfilter consonant plot (see Section 4.3.1).

All sounds When using multidimensional scaling on the basis of all sound distances, one dimension explains 88% of the variance, two dimensions 98% and three dimensions 99%. In Figure 4.10 a two-dimensional multidimensional scaling plot is shown. The first dimension (the vertical dimension in the plot) corresponds with intensity. The [a] is loudest and ‘silence’ is most silent (of course). The second dimension (horizontal) represents clearness. The [s] is the clearest and the [b] is the darkest sound. However, one might expect some consonants (e.g. the [f]) to be located in the darker area, just as in the Barkfilter plot. In the plot the distinction between voiceless and voiced sounds is sharper than in the Barkfilter plot, just as we saw for the consonants. Drawing a line from [i] to [a], from [a] to [n], from [n] to [u], and from [u] to [i] we recover the IPA vowel quadrilateral. Just as in the Barkfilter plot the nasals appear as high vowels. In contrast to the Barkfilter most r-like sounds are outside the vowel quadrilateral now. Only the retroflex flap is still in the quadrilateral. Both retroflex approximants are moved to the center area of the quadrilateral. The [u] and the [w] are still closer than the [i] and the [j]. Additionally the position of ‘silence’ is very near the glottal stop. For the explanation of the third dimension see the Barkfilter representation (Section 4.3.1).

Because the vowel classification is like the IPA quadrilateral, and the consonant classification reflects the different manners of articulation we conclude that the cochleagram representation is useful for finding segment distances. Cochleagrams differ from the Barkfilter representation in virtue of the sharper distinction between voiceless and voiced sounds, and between vowels and r-like sounds.
Figure 4.9: Two-dimensional multidimensional scaling plot obtained from the cochleagram distances between all pairs formed by the 59 consonants. Two dimensions explain 96% of the variance. The first dimension (y-axis) distinguishes between voiceless and voiced consonants, the second dimension (x-axis) between non-continuous and continuous consonants. A third dimension would be less easily interpreted.
Figure 4.10: Two-dimensional multidimensional scaling plot on the basis of the cochleagram distances between all pairs formed by the 28 vowels, the 59 consonants and ‘silence’. In the plot, the ! is used for ‘silence’. Two dimensions explain 98% of the variance. The first dimension (y-axis) represents intensity (lower sounds are louder) and the second dimension (x-axis) clearness (sounds on the right are darker). The shaded area represents a quadrilateral formed on the basis of the vertices of the IPA quadrilateral.
4.3.3 Formant tracks

4.3.3.1 Representation

Another way to analyse the acoustic signal is to investigate formants. When using a spectrogram with a large analysis window (about 20 ms) the frequency resolution will be high. Individual harmonics will show up as horizontal lines through the spectrogram (see the spectrograms, Barkfilters and cochleagrams in Figures 4.1 and 4.2). The lowest line represents the fundamental frequency or pitch (F0). However, when using a small analysis window (about 3 ms) the frequency resolution will be lower. Individual harmonics get blended together. Instead of lines, bands will show up through the spectrogram. The center frequency at one time in a band is called a formant, the range of center frequencies in the course of time forms a formant track. A formant in the lowest band is called F1, a formant in the next band F2, etc. Formants represent a frequency region that is enhanced by the resonances of the vocal tract.\(^3\)

In the Figures 4.1 and 4.2 formant tracks are shown, obtained on respectively non-manipulated and monotonized samples. When finding formants in the computer program PRAAT, the time step was set to 0.01 seconds with an analysis window of 0.025 seconds. The ceiling of the formant search range should be set to 5000 Hz for males, and to 5500 Hz for females. So for the samples of John Wells the ceiling was set to 5000 Hz, and for Jill House to 5500 Hz. Pre-emphasis starts at 50 Hz. In the manual which can be found in the PRAAT program pre-emphasis is explained as follows:

“This means that frequencies below 50 Hz are not enhanced, frequencies around 100 Hz are amplified by 6 dB, frequencies around 200 Hz are amplified by 12 dB, and so forth. The point of this is that vowel spectra tend to fall by 6 dB per octave; the pre-emphasis creates a flatter spectrum, which is better for formant analysis because we want our formants to match the local peaks, not the global spectral slope.”

In PRAAT several algorithms can be chosen for finding the Linear Predictive Coding (LPC) coefficients. We chose the algorithm of Burg. This algorithm may initially find formants at very low or high frequencies. However we used in PRAAT the version which removes formants below 50 Hz and formants above 5000 Hz (males) or 5500 Hz (females) minus 50 Hz. In this way the algorithm will identify the traditional F1 and F2. The algorithm of Burg is much more reliable than the Split Levinson algorithm which always finds the requested number of formants in every frame, even if they do not exist. Since we found at least two formants for every frame in every sample when using the more reliable Burg

algorithm we do not use the Split Levinson algorithm. More about the algorithms can be found in the manual in PRAAT program.

When using formant tracks we had to decide how many formant tracks should be taken into account. It is a well-known fact that in the IPA vowel quadrilateral the height corresponds with the F1 (the lower the F1, the closer the vowel) and that the advancement corresponds with the F2 (the higher the F2, the further sounds are fronted, see Rietveld and Van Heuven (1997, p. 133)). The F2 of rounded vowels is a little lower than the F2 of unrounded vowels. The meaning of higher formants is less clear. At the risk of ignoring information important to dialect recognition we therefore decided to compare sounds only on the basis of the F1 and the F2. Before comparing formant frequencies in the comparison of words the frequencies in Hertz are converted to Bark, which is, as mentioned above, a more faithful scale perceptually. For this purpose we used the formula of Traunmüller (1990) as suggested in standard works about phonetics (Rietveld and Van Heuven, 1997):

\[ \text{Bark} = \frac{26.81 \times \text{Hertz}}{1960 + \text{Hertz}} - 0.53 \]  

The relation between Traunmüller’s formula and Schroeder et al.’s (see the formula in (4.1) in Section 4.3.1.1) is shown in Figure 4.11, in which the Hertz frequencies are plotted against the Bark frequencies. The graph shows two curves, the upper one based on the formula of Schroeder et al. (1979), the lower one found by using the formula of Traunmüller (1990). As mentioned in the Sections 4.3.1 and 4.3.2 the Schroeder et al. formula is used for Barkfilters and cochleagrams. In the plot the Hertz-scale runs from 20 to 20000 Hertz, the frequency range which can be perceived by a human being. This corresponds with a frequency range of -0.26 to 23.89 for the Traunmüller curve.

4.3.3.2 Classification

Just as for the Barkfilter representation and the cochleagram representation the distances between the IPA sounds are calculated (see Section 4.7). Multidimensional scaling is performed on the basis of vowel distances, the consonant distances and the distances between both vowels and consonants including ‘silence’. Since we followed the same procedure as for the Barkfilter representation, the reader is referred to Section 4.3.1 for more details.

Vowels When using multidimensional scaling on the basis of the vowel distances, one dimension already explains 67% of the variance, two dimensions 99% and three dimensions 100%. In Figure 4.12 a two-dimensional multidimensional scaling plot is shown. The first dimension (the horizontal dimension in the plot) represents the advancement, the second dimension (vertical) height. The plot is rather similar to the Barkfilter plot (see Figure 4.5) and the cochleagram plot.
Figure 4.11: Frequencies in Hertz versus frequencies in Bark. The Hertz scale runs from 20 Hz to 20,000 Hz. The upper line shows Bark values calculated with the formula of Schroeder et al. (1979) (applied when using Bark filters or cochleagrams), the lower line shows Bark values calculated with the formula of Traunmüller (1990) (applied when using formant tracks). Below 1000 Hz both curves are roughly linear, above 1000 Hz they are roughly logarithmic.
4.3. REPRESENTATION OF SEGMENTS

(see Figure 4.8). However, when drawing a line from [i] to [a], from [a] to [u], from [o] to [u], and from [u] to [i] we get a triangle rather than a quadrilateral, where the [o] is on the line between the [a] and the [u]. This agrees with results found in Rietveld and Van Heuven (1997, p. 133). They show a triangle based on mean formant values of male speakers derived from Pols (1977). In our plot front vowels have a high F2 and back vowels a low F2. The second dimension represents the height, corresponding with F1. High vowels have a low F1 and low vowels have a high F1. In the plot the [a] is nearly in the center, while in the Barkfilter plot and the cochleagram plot the sound belonged to the higher vowels of the lower group. The division between high vowels and low vowels is not as sharp as in the Barkfilter plot and the cochleagram plot. When scaling to three dimensions, the third dimension makes a distinction between the [u], [i], [e], [e] and [æ] on the one hand, and the other vowels on the other hand. This may be interpreted as a distinction of high and front vowels versus other vowels, although this is not consistently true.

Consonants When using multidimensional scaling on the basis of the consonant distances, one dimension explains 79% of the variance, two dimensions 92% and three dimensions 96%. In Figure 4.13 a two-dimensional multidimensional scaling plot is shown which was obtained on the basis of the consonant distances. The first dimension (the vertical dimension in the plot) distinguishes between voiced (upper) and voiceless sounds (lower) and the second dimension (horizontal) represents the place of articulation very vaguely, albeit in a different way than that used in the IPA pulmonic consonant table. The palatals appear in Figure 4.13 as front consonants (left), and the velar, uvular, pharyngeal and glottal consonants appear as back consonants (right). The [w] which is specified as a voiced labial-velar approximant in the IPA system is found here as a back consonant as well as the [v]. Other consonants are more central. Drawing a line from the [j] to the [w], from the [w] to the [h] and from the [h] to the [j] a similar triangle is found as in the vowel plot. Voiced sounds have a low F1 and voiceless sounds have a high F1. Front consonants have a high F2 and back consonants a low F2. However, this is a simplified sketch, many exceptions can be found if one examines the plot more precisely. The role of manner of articulation is found in the plot. The nasals (upper right), voiced obstruents, (upper central), liquids (central, below the nasals and the voiced obstruents), and voiceless obstruents (low) can be identified. The plot is different from the Barkfilter plot (see Figure 4.6) and the cochleagram plot (see Figure 4.9). Most striking is the fact that there is no sharp separation between plosives and fricatives. We examined the third dimension as well but found no obvious interpretation for this.

All sounds When using multidimensional scaling on the basis of all sound distances, one dimension explains 72% of the variance, two dimensions 96% and
Figure 4.12: Two-dimensional multidimensional scaling plot obtained from the formant track distances between all pairs formed by the 28 vowels. Two dimensions explain 98% of the variance. The first dimension (x-axis) corresponds with advancement and the second dimension (y-axis) with height. Note that the schwa [ə] is located about in the middle. A third dimension would be less easily interpreted. In this plot some symbols are shifted a little bit.
Figure 4.13: Two-dimensional multidimensional scaling plot obtained from the formant track distances between all pairs formed by the 59 consonants. Two dimensions explain 92% of the variance. The first dimension (y-axis) distinguished between voiceless and voiced sounds. For the second dimension (x-axis) we found no obvious interpretation. For a third dimension we found no interpretation as well.
three dimensions 98%. In Figure 4.14 a two-dimensional multidimensional scaling plot is shown on the basis of all sounds. The first dimension (the vertical dimension in the plot) distinguishes between high vowels (high) and low vowels (lower), and between voiced consonants (high) and voiceless consonants (low). The second dimension (horizontal) distinguishes between front vowels (left) and back vowels (right), and between ‘front consonants’ (left) and ‘back consonants’ (right). See the separate plots of vowels and consonants for more explanation.

When drawing a line from [i] to [a], from [a] to [u], from [u] to [i], and from [u] to [i], we get a triangle again. High vowels and voiced consonants have a lower F1, and low vowels and voiceless consonants a higher F1. Front vowels and ‘front consonants’ have a higher F2, and back vowels and ‘back consonants’ a lower F2.

In the middle of the triangle we find the r-like sounds, similar to the Barkfilter plot (see Figure 4.7). The nasals are located around the line between the [i] and the [u], however, closer to the [u] than to the [i]. The laterals are located in the corner of the [i]. Most voiced plosives are located above the line between the [u] and the [i], closer to the [i] than to the [u]. The voiceless velar, uvular, and pharyngeal fricatives, and both the voiceless and voiced glottal fricatives are located between the [a] and the [u]. The [i] and the [j] are much closer than the [u] and the [w]. This is the opposite of what we saw in the Barkfilter plot (see Figure 4.7) and the cochleagram plot (see Figure 4.10). The correct closeness of the [i] and the [j] may be explained by the fact the formants are not sensitive to differences in intensity. However, the relatively large distance between the [u] and the [w] can be explained from the fact that we used a sample of Jill House for the [w] in the set of samples of John House (this is justified in Section 4.2). Formants do not neutralize differences in gender. Furthermore, ‘silence’ is not as close to the glottal stop as when using the Barkfilter representation or the cochleagram representation (see Figure 4.10). We also examined the third dimension but found no clear interpretation for this.

Summarizing, we found that the vowel classification is like the IPA quadrilateral, and the consonant classification reflects the different manners of articulation. Therefore, we conclude that the formant track representation is useful for finding segment distances. For both the vowels and the consonants we found striking differences with respect to the Barkfilter representations and cochleagram representation. For the vowels we found no quadrilateral, but a triangle when connecting the corner points [i], [u], [o] and [a]. This is in accordance with results found in literature (data of Pols (1977) visualized by Rietveld and Van Heuven (1997, p. 133)). For the consonants there is no clear separation between plosives and fricatives. This may be explained from the fact that only formant tracks are used which gives less information than the Barkfilter and cochleagram representation. Therefore, the formant track representation is more accurate for finding vowel distances than for finding consonant distances.
Figure 4.14: Two-dimensional multidimensional scaling plot on the basis of the formant track distances between all pairs formed by the 28 vowels, the 59 consonants and ‘silence’. In the plot, the ! is used for ‘silence’. Two dimensions explain 92% of the variance. The first dimension (y-axis) represents the F1 (higher sounds have lower F1 values) and the second dimension (x-axis) the F2 (sounds on the left have higher F2 values). The shaded area represents a vowel triangle.
4.4 Diphthongs

As mentioned in Section 3.2, there are two possibilities for processing diphthongs. When using the acoustic representations we want to be able to consider both approaches again. In the first approach, a diphthong is considered as nothing more than a sequence of two monophthongs. In the second approach, a diphthong may also be regarded as one sound with a changing color. In the optimal case we should have samples of all diphthongs that may occur in our data. However, on the tape *The Sounds of the International Phonetic Alphabet* no samples of diphthongs can be found. To be able to process diphthongs as one sound nonetheless, we modify the calculations in ways which should have the desired effect.

The distance between a monophthong and a diphthong is calculated as the mean of the distance between the monophthong and the first element of the diphthong and the distance between the monophthong and the second element of the diphthong. So the distance between e.g. [ai] and [ɛ] is calculated as the mean of the distance between [a] and [ɛ] and the distance between [i] and [ɛ]. We will discuss this in more detail. Assume the front vowels [a], [æ], [ɛ], [e] and [i] are on a straight line, just as in the IPA quadrilateral. Assume the distance from [a] to [æ] is 0.5, from [æ] to [ɛ] is 0.5, from [ɛ] to [e] is 1, and from [e] to [i] is 1. What are now the distances from the [ai] to the starting point, the intermediate points and the end point? For each point we calculate the distance to the [a] and to the [i] and take the average of both segments:

<table>
<thead>
<tr>
<th></th>
<th>[a]</th>
<th>[i]</th>
<th>[ai]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[i]</td>
<td>3.0</td>
<td>0.0</td>
<td>1.5</td>
</tr>
<tr>
<td>[ɛ]</td>
<td>2.0</td>
<td>1.0</td>
<td>1.5</td>
</tr>
<tr>
<td>[ɛ]</td>
<td>1.0</td>
<td>2.0</td>
<td>1.5</td>
</tr>
<tr>
<td>[æ]</td>
<td>0.5</td>
<td>2.5</td>
<td>1.5</td>
</tr>
<tr>
<td>[a]</td>
<td>0.0</td>
<td>3.0</td>
<td>1.5</td>
</tr>
</tbody>
</table>

We see that for each of the points the distance to the [ai] is the same. In fact the distance is simply equal to \(d([a],[i])/2\). This is in accordance with the idea that the color of a [ai] gradually changes from [a] to [æ], from [æ] to [ɛ], from [ɛ] to [e], from [e] to [i]. All (intermediate) points are heard in the diphthong during an infinitesimally small moment in time. However, in our acoustic results the [æ], [ɛ] and [i] do not lay exactly on the line from [a] to [i] (see Figures 4.5, 4.8 and 4.12). The distances of these ‘intermediate points’ will be greater than \(d([a],[i])/2\).

The distance between two diphthongs is calculated as the mean of the distance between the first elements and the distance between the second elements. So the distance between e.g. [au] and [ci] is calculated as the mean of the distance between [a] and [c] and the distance between [u] and [i].
4.5 Affricates

In the RND data no affricates are used. However, in the NOS data they do appear. On the IPA tape *The Sounds of the International Phonetic Alphabet* two affricates can be found, namely the [kp] and the [ts]. However, to be able to process many more affricates, we did not use these sample but applied the more general approach as given in Section 3.3. When processing affricates both elements are processed as extra-short, separated elements.

4.6 Suprasegmentals and diacritics

The sounds on the tape *The Sounds of the International Phonetic Alphabet* are pronounced without suprasegmentals and diacritics. However, a restricted set of suprasegmentals and diacritics can be processed in our system. Since no features can be changed, only those suprasegmentals and diacritics are taken into account which can be processed by changing the weighting of segments or by averaging sound distances. Suprasegmentals and diacritics which are processed by the first approach are discussed in Section 4.6.1, those which are processed by the second approach in Section 4.6.2.

4.6.1 Weighting segments

In Section 3.4.2 we stated that the weighting of extra-short sounds should be halved with respect to the weighting of short sounds. Just as we did with the discrete representations, for the acoustic representations we realize this by changing the transcription beforehand. We retain the extra-short sounds as they are and double all other sounds.

When using discrete representations, we consider two approaches for the processing of *half-long* and *long*. When using acoustic representations, both approaches are regarded again. In fact, for the acoustic representations exactly the same applies as for the discrete phone representations. In the first approach, *half-long* and *long* are not processed since they may sometimes be redundant to some extent. In the second approach, the two length marks are processed by changing the transcription. Half-long sounds are trebled and long sounds are quadrupled. For more details see Section 3.4.2.

In the RND, consonants may also be vocalized. We process vocalized sounds as syllabic sounds. Vocalized (RND) or syllabic sounds (NOS) are marked with the diacritic *syllabic*. Using the acoustic representation, syllabic sounds are processed in the same way as when using the discrete phone representation. We consider two approaches for processing syllabic sounds which corresponds with the two approaches that are regarded when processing *half-long* and *long*. In the first approach *syllabic* is not processed since it may be redundant. In the second
approach this diacritic is processed by changing the transcription beforehand. Syllabic sounds are processed as long sounds, i.e. they are quadrupled in the transcription. For more details see Section 3.4.2 again.

For the RND *aspirated* is not processed. However, for the NOS data it is processed. An [h] is inserted after the phone which was noted to be aspirated. This [h] is noted as extra-short, so the significance is halved. For more details see Section 3.4.5.

### 4.6.2 Averaging segments

When using acoustic representations, the diacritics *voiceless*, *voiced*, *apical* and *nasalized* can be processed. When comparing sound $x$ and sound $y$, one or more diacritics may be noted after one or both sounds. To process them, first the distance between $x$ and $y$ is calculated as it is without any diacritics. This is a basic distance and mentioned in the first part of Table 4.1. A counter is set to 1. Next, we check whether the diacritics *voiceless* or *voiced* are used. The possible combinations are listed in the second part of Table 4.1 in the column ‘condition’. If one of the conditions applies, the corresponding distance increase as given in the column ‘distance increase’ is calculated and added to the basic distance. The counter is increased by 1. Subsequently we check whether the diacritic *apical* is used. The possible combinations are listed in the third part of Table 4.1 under ‘condition’. If one of the conditions apply, the corresponding distance increase as suggested under ‘distance increase’ is added to the basic distance, and the counter is increased by 1. Finally it is checked whether the feature *nasal* is used. If one of the conditions in the fourth part of Table 4.1 apply, the corresponding distance increase is added to the basic distance and the counter is increased by 1.

If no diacritics were noted, the total distance is equal to the basic distance, and the counter is equal to 1. If all diacritics were found, the largest distance is obtained, and the counter is equal to 4. Now the final distance is equal to the total distance (basic distance plus optionally one or more diacritic increases) divided by the counter.

The idea behind the calculation of the distance increase of the diacritics *voiceless* and/or *voiced* is that a voiced voiceless sound or a voiceless voiced sound is exactly intermediate between a voiceless sound and a voiced sound. In the table $X$ is the voiced counterpart of a voiceless $x$, or the voiceless counterpart of a voiced $x$. The $Y$ is defined analogously to the $X$. For voiced sounds which have no voiceless counterpart (the sonorants), or for voiceless sounds which have no voiced counterpart (the glottal stop) the sound itself is used. Since the RND sounds are a subset of the IPA sounds, there are no voiced counterparts for the [c] and the [h], so respectively the [c] and the [h] are returned. When the diacritic
voiceless is noted under a voiceless sound, it is most likely that this an error. Instead of this diacritic the diacritic voiced is processed. Conversely the diacritic voiced under a voiced sound is processed as the diacritic voiceless.

The diacritic apical is implemented in the NOS system only. The implementation of this diacritic was made to be able to process Romanesque languages as well. Only the [s] and [z] are allowed to be apical in our system. The /s/ “of standard Spanish is an apical-alveolar sound” (Pountain, 2001, p. 299). The tip of the tongue is often “retroflexed or turned back as it touches the alveolar ridge” (Dalbor, 1969, p. 91). In some Sardinian dialects the same apical-alveolar sound is found. When comparing a non-apical sound with an apical sound, the distance increase is equal to the distance between the non-apical sound and the [s] (if the apical sound was a [s]) or the [z] (if the apical sound was a [z]).

The thought behind the way in which the diacritic nasal is processed is that a nasal sound is about intermediate between its non-nasal version and the [n]. So when comparing a non-nasal sound with a nasal sound, we quantify the effect of the diacritic nasal by calculating the distance between the non-nasal sound and the [n].

4.7 Comparison of segments

In this section, we explain the comparison of segments in order to get distances between segments that will be used in the Levenshtein distance (see Section 5.1). In a Barkfilter or cochleagram, the intensities or loudnesses of frequencies are given for a range of times. A spectrum contains the intensities or loudnesses of frequencies at one point in time. In a formant track representation, the formants are given for a range of times. A formant bundle contains the formants for one point in time. The smaller the time step, the more spectra or formant bundles in the acoustic representation. Per acoustic representation we consistently used the same time step for all samples.

It appears that the duration of the segment samples varies. This may be explained by variation in speech rate. Duration is also a sound-specific property. E.g., a plosive is shorter than a vowel. The result is that the number of spectra of formant bundles per segment may vary, although for each segment sample the same time step was used. Since we want to normalize the speech rate and regard segments as linguistic units, we see to it that two segments get the same number of spectra or formant bundles when they are compared to each other.

When comparing one segment of m spectra or formant bundles with another segment of n spectra or formant bundles, each of the m elements is duplicated n times, and each of the n elements is duplicated m times. So both segments get a length of m × n. Below two segments are schematically visualized, one with 3 elements (black bars) and one with 2 elements (grey bars). Now both get a
Table 4.1: When $x$ and $y$ are sounds with the diacritics, first the basis distance is calculated (1) and a counter is set to 1. Next for each valid condition under ‘Condition’ (2,3,4) the basis distance is increased with the corresponding distance under ‘Distance increase’ and the counter is increased by 1. The final distance is the total distance divided by the counter. $X$ and $Y$ are the voiceless or voiced counterparts of the voiced or voiceless $x$ and $y$. 

<table>
<thead>
<tr>
<th>Diacritic</th>
<th>Condition</th>
<th>Distance increase</th>
<th>Counter increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 basis</td>
<td>$x$ vs. $y$</td>
<td>$d(x,y)$</td>
<td>1</td>
</tr>
<tr>
<td>2 voice</td>
<td>$x$ vs. $y$</td>
<td>$d(X,Y)$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$x$ vs. $y$</td>
<td>$d(x,y)$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$x$ vs. $y$</td>
<td>$d(X,Y)$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$x$ vs. $y$</td>
<td>$d(X,Y)$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$x$ vs. $y$</td>
<td>$d(X,Y)$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$x$ vs. $y$</td>
<td>$d(X,Y)$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$x$ vs. $y$</td>
<td>$d(X,Y)$</td>
<td>1</td>
</tr>
<tr>
<td>3 apical</td>
<td>[s] vs. $y$</td>
<td>$d([s],y)$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>[z] vs. $y$</td>
<td>$d([z],y)$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$x$ vs. [z]</td>
<td>$d([s],[z])$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$x$ vs. [s]</td>
<td>$d([z],[s])$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>[s] vs. [z]</td>
<td>$d([s],[z])$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>[z] vs. [s]</td>
<td>$d([z],[s])$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>[s] vs. [s]</td>
<td>$d([s],[s])$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>[z] vs. [z]</td>
<td>$d([z],[z])$</td>
<td>1</td>
</tr>
<tr>
<td>4 nasal</td>
<td>$x$ vs. $y$</td>
<td>$d([n],y)$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$x$ vs. $y$</td>
<td>$d([n],[n])$</td>
<td>1</td>
</tr>
</tbody>
</table>
length of 6 when each of the 3 elements are duplicated 2 times, and each of the 2 elements are duplicated 3 times.

For finding the distance between two sounds the Euclidean distance is calculated between each pair of corresponding spectra or formant bundles, one from the first, and one from the second sound. Assume a spectrum or formant bundle \( e_1 \) and \( e_2 \) with respectively \( n \) frequencies or formants, then:

\[
d(e_1, e_2) = \sqrt{\sum_{i=1}^{n} (e_{1i} - e_{2i})^2}
\]

(4.3)

The distance between two segments is equal to the sum of the spectrum or formant bundle distances divided by the number of spectra or formant bundles. In this way we found that the greatest distance occurs between the [a] and ‘silence’ (Barkfilter, Cochleagram) or between the [u] and the [j] (formants). We regard the maximum distance as 100%. Other segment distances are divided by this maximum and multiplied by 100. This gives segment distances expressed in percentages. Word distances and dialect distances which are based on them may also be given in terms of percentages.

In Section 3.7 we explained that segment distances obtained on the basis of feature definitions may be used in two ways. First, the distances can be used directly, i.e. linearly. Second the logarithms of the distances can be taken. The latter reflects the fact that in perception small differences in pronunciation may play a relatively strong role in comparison with larger differences. For the acoustic segment distances we consider both the linear and logarithmic approach again. Since the logarithm of 0 is not defined, and the logarithm of 1 is 0, distances are increased by 1 before the logarithm is calculated. To obtain percentages, we calculate \( \frac{\ln(distance + 1)}{\ln(maximum\ distance + 1)} \).

4.8 Correlation between systems

In this section, we compare the different acoustic systems in order to check for striking differences between them. In Section 4.8.1 we compare the different
representations by calculating the correlation coefficient between the distances, just as we did for the different feature representations (see Section 3.8).

We also compare the acoustic systems with the discrete systems. In Section 4.8.2, we correlate the acoustic based distances with the distances obtained on the basis of feature representations (see Section 3.1) in order to examine to what extent acoustic distances differ from feature based systems.

We calculate correlations on the basis of both the set of RND sounds and the set of IPA sounds. Again we have 18 vowels and 27 consonants for the RND, and 28 vowels and 59 consonants for the IPA. When we correlate on the basis of both vowels and consonants, we get respectively 46 sounds for the RND and 88 sounds for the IPA.

For finding the significance of a correlation coefficient we used the Mantel test, just as in Chapter 3. The Mantel test was also used to determine whether one correlation coefficient is significantly higher than another. The Mantel test is explained in Section 3.8.2. As significance level we again choose $\alpha = 0.05$.

### 4.8.1 Acoustic vs. acoustic

The Tables 4.2 and 4.3 show the correlation coefficients between matrices of segment distances obtained on the basis of the Barkfilter, the cochleagram and the formant track representation. Correlations are given for both the RND and the IPA segment distances. Results are given for vowels, consonants and all sounds. When all segments are used, ‘silence’ is also included. All correlations are significant for $\alpha = 0.05$.

For both the RND and the IPA the Barkfilter and the cochleagram distances correlate significantly more strongly than the other pairs of representations for vowels, consonants and all segments. This outcome is not surprising since these representations are most similar. Looking at the vowels, the correlation between the formant-track distances and the Barkfilter distances is not significantly weaker or stronger than the correlation between the formant track distances and the cochleagram distances. However, when looking at the consonants and all sounds, the formant-track distances yield correlations significantly stronger with the cochleagram distances than with the Barkfilter distances. We observed this but cannot explain it.

Since the correlations between the Barkfilter distances and cochleagram distances are rather high ($0.94 \leq r \leq 1.00$), dialect distances based on Barkfilter distances and cochleagram distances are expected to be similar. The correlations between the formant track distances and the Barkfilter distances are lower ($0.45 \leq r \leq 0.78$) just as the correlations between the formant track distances and the cochleagram distances ($0.51 \leq r \leq 0.78$). So we expect that the use of formant track segment distances in dialect comparison will give significantly different results than when using Barkfilter distances or cochleagram distances.
4.8. CORRELATION BETWEEN SYSTEMS

Table 4.2: Correlation coefficients among RND segment distances between vowels (vow.), consonants (cons.) and all segments obtained on the basis of the Barkfilter (Bark.), the cochleagram (coch.) and the formant track (form.) representation. When all segments are used, ‘silence’ is also included.

<table>
<thead>
<tr>
<th></th>
<th>vow.</th>
<th>cons.</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bark. vs.</td>
<td>1.00</td>
<td>0.94</td>
<td>0.96</td>
</tr>
<tr>
<td>Coch.</td>
<td>0.77</td>
<td>0.64</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table 4.3: Correlation coefficients among IPA segment distances between vowels (vow.), consonants (cons.) and all segments obtained on the basis of the Barkfilter (Bark.), the cochleagram (coch.) and the formant track (form.) representation. When all segments are used, ‘silence’ is also included.

<table>
<thead>
<tr>
<th></th>
<th>vow.</th>
<th>cons.</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bark. vs.</td>
<td>1.00</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>Coch.</td>
<td>0.78</td>
<td>0.59</td>
<td>0.51</td>
</tr>
</tbody>
</table>

4.8.2 Acoustic vs. features

We also examined the correlations between the acoustic distances and the feature-based distances. Results are given in Tables 4.4 and 4.5. Almost all correlations between the acoustic representations and the feature representations are significant for $\alpha = 0.05$. Only the lowest correlations which are found between the A & B representations (Manhattan and Euclidean distance) and the formant track representation for the RND consonants are not significant.

Looking at the vowel correlations, we observe that the Barkfilter distances correlate strongest with the A & B distances, regardless which feature bundle metric is used. The three A & B correlations are not significantly higher than the corresponding ones of the two other feature systems, however. Just as for the Barkfilter distances, the cochleagram distances correlate strongest with the A & B distances. However, the correlations for the three feature bundle metrics do not differ significantly among the different feature systems. The formant track distances correlate strongest with the V & C distances for most metrics. However the correlations for all three feature bundle metrics are for the most part not significantly higher than the same metrics applied to different feature systems. In the feature system of A & B height has a greater weight than advancement (see Table 3.12). Examining Figures 4.5 and 4.8, we observe that for the Barkfilter and the cochleagram representation height is weighted more strongly than advancement as well. In the feature system of V & C advancement has a greater
Table 4.4: Correlations among segment distances as specified by three feature systems and three acoustic systems on the basis of the distances between the RND segments. Feature bundle distances are found by calculating the (M)anhattan distance, (E)uclidean distance or (P)earson correlation coefficient.

<table>
<thead>
<tr>
<th></th>
<th>Bark.</th>
<th>Coch.</th>
<th>Form.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>vow.</td>
<td>cons.</td>
<td>vow.</td>
</tr>
<tr>
<td>H &amp; H</td>
<td>0.39</td>
<td>0.29</td>
<td>0.38</td>
</tr>
<tr>
<td>E.</td>
<td>0.41</td>
<td>0.29</td>
<td>0.39</td>
</tr>
<tr>
<td>P.</td>
<td>0.31</td>
<td>0.35</td>
<td>0.30</td>
</tr>
<tr>
<td>V &amp; C</td>
<td>0.54</td>
<td>0.35</td>
<td>0.52</td>
</tr>
<tr>
<td>E.</td>
<td>0.55</td>
<td>0.36</td>
<td>0.54</td>
</tr>
<tr>
<td>P.</td>
<td>0.36</td>
<td>0.40</td>
<td>0.34</td>
</tr>
<tr>
<td>A &amp; B</td>
<td>0.67</td>
<td>0.19</td>
<td>0.65</td>
</tr>
<tr>
<td>E.</td>
<td>0.67</td>
<td>0.21</td>
<td>0.66</td>
</tr>
<tr>
<td>P.</td>
<td>0.71</td>
<td>0.21</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 4.5: Correlations among segment distances as specified by three feature systems and three acoustic systems on the basis of the distances between the IPA segments. Feature bundle distances are found by calculating the (M)anhattan distance, (E)uclidean distance or (P)earson correlation coefficient.

<table>
<thead>
<tr>
<th></th>
<th>Bark.</th>
<th>Coch.</th>
<th>Form.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>vow.</td>
<td>cons.</td>
<td>vow.</td>
</tr>
<tr>
<td>H &amp; H</td>
<td>0.40</td>
<td>0.31</td>
<td>0.39</td>
</tr>
<tr>
<td>E.</td>
<td>0.40</td>
<td>0.30</td>
<td>0.39</td>
</tr>
<tr>
<td>P.</td>
<td>0.24</td>
<td>0.38</td>
<td>0.24</td>
</tr>
<tr>
<td>V &amp; C</td>
<td>0.52</td>
<td>0.36</td>
<td>0.51</td>
</tr>
<tr>
<td>E.</td>
<td>0.51</td>
<td>0.36</td>
<td>0.50</td>
</tr>
<tr>
<td>P.</td>
<td>0.36</td>
<td>0.42</td>
<td>0.34</td>
</tr>
<tr>
<td>A &amp; B</td>
<td>0.68</td>
<td>0.23</td>
<td>0.67</td>
</tr>
<tr>
<td>E.</td>
<td>0.68</td>
<td>0.25</td>
<td>0.66</td>
</tr>
<tr>
<td>P.</td>
<td>0.72</td>
<td>0.27</td>
<td>0.71</td>
</tr>
</tbody>
</table>
4.8. CORRELATION BETWEEN SYSTEMS

weight than height (see Table 3.6). Examining Figure 4.12 we also find for the
formant track representation that advancement has a greater weight than height.
All acoustic representations correlate worst with the H & H system, although
not significantly lower than with other systems. The lower correlations may be
explained by the unnatural way in which height is defined (see Section 3.1.2.1).

Looking at the consonant correlations we observe that the H & H distances in
most cases and the V & C distances in all cases correlate significantly better with
the Barkfilter distances than the A & B distances do. The V & C correlations
are higher than the H & H correlations, but they are not significantly higher.
The H & H distances and the V & C distances correlate in all cases significantly
better with the cochleagram distances than the A & B distances do. Just as for
the Barkfilter the V & C correlations are higher than the H & H correlations,
but again they are not significantly higher. The higher correlations of both the
Barkfilter distances and the cochleagram distances with the V & C distances may
be explained by the categorical way in which manner of articulation is defined
in the system of V & C. The worse correlation with the A & B system may be
explained by the fact that in this feature system, manner of articulation is defined
as a scale. The H & H distances and the V & C distances correlate in all cases
significantly better with the formant track distances than the A & B distances
do. Using the RND consonants, the formant track distances correlate strongest
with the V & C distances, but not significantly more strongly than with the
H & H distances. When using the IPA consonants, the formant track distances
 correlate strongest with the H & H distances, but not significantly more strongly
than with the V & C distances. The higher correlations of the V & C distances
and the H & H distances may again be explained by the fact that manner of
articulation is defined as a scale in the A & B system. The difference between
the RND and the IPA may be explained by the fact that in the system of H & H
all RND consonants are uniquely defined, but all IPA consonants are not.

As explained in Section 3.1.3, the V & C feature system is perceptually based.
We expect that the V & C distances will correlate more strongly with the cochleagram
distances than with the Barkfilter distances since the cochlear model
is a more exact model of the cochlea than the Barkfilter model. For the vowels
we see exactly the opposite: the Barkfilter distances correlate more strongly
with the V & C distances than the cochleagram distances, although no significant
differences between correlation coefficients were found. For the consonants
we find what we expected: the cochleagram distances correlate more strongly
with the V & C distances than the Barkfilter distances. However, the differences
between the correlations coefficients are not significant. It should be interesting
to correlate the complete set of vowels and consonants with perceptually based
distances. However, vowels and consonants are separated in the V & C system,
so unfortunately, this was not possible.

Although almost all correlation coefficients are significant, they are not extremenly high. For the vowels the highest correlation for the Barkfilter distances
is 0.72 (with respect to A & B, Pearson, IPA), for the cochleagram distances 0.71 (A & B, Pearson, IPA) and for the formant track distances 0.74 (V & C, Euclidean, IPA). For the consonants the highest correlation for the Barkfilter distances is 0.42 (with respect to V & C, Pearson, IPA), for the cochleagram distances 0.47 (V & C, Pearson, IPA) and for the formant track distances 0.35 (H & H, Manhattan and Pearson, RND). The vowel distances correlate more strongly than the corresponding consonant distances in most cases. Especially when regarding the lower consonant correlations, we expect that the use of acoustic segment distances in dialect comparison will give results that are different with respect to feature-based results.

4.9 Conclusions

In this chapter we presented the use of acoustic representations for finding distances between segments. In contrast to most feature systems acoustic representations are based on physical measurements. We examined the Barkfilter, cochleagram and formant track representation, which are more perceptually oriented models. We performed multidimensional scaling on the acoustic distances and scaled them to two dimensions. For all representations we obtained a vowel classification which is like the IPA quadrilateral, and the consonant classification reflects the different manners of articulation. Therefore, we conclude that the three representations are useful for finding segment distances. The Barkfilter and the cochleagram representations correlate significantly more strongly than any other pairs of representations. The results obtained on the basis of the formant track representation are more different. With the formant track representation a vowel triangle is obtained, and for the consonants no clear separation between plosives and fricatives was found.

When correlating distances obtained by the feature representations with distances obtained by the acoustic representations, it appears that, for the vowels, both the Barkfilter distances and the cochleagram distances correlate strongest with the A & B distances. The formant track distances correlate strongest with the V & C distances for most metrics. The correlation coefficients were for the most part not significantly higher than other comparable correlation coefficients. All acoustic representations correlate worst with the H & H system, but not significantly lower than with other systems. The lower correlations can be explained by the unnatural way by which height is defined in the system of H & H. Therefore, for vowels we prefer A & B and V & C to H & H. However we made no choice between A & B and V & C. On the one hand, the Barkfilter and cochleagram representations contain more information, on the other hand, the formant track representation may be limited to information which is relevant in perception.

For the consonants, the Barkfilter distances and the cochleagram distances correlate strongest with the V & C distances, but only significantly better than
with the A & B distances. The formant track distances correlate strongest with the V & C distances (RND) or H & H distances (IPA). The correlation coefficients were only significantly higher than the comparable ones of A & B. This suggests that manner of articulation should not be represented as a scale (as in A & B), but as different categories (as in V & C and H & H). Therefore, for consonants we find V & C and H & H preferable to A & B, but cannot make a choice between V & C and H & H.

When correlating feature-based segment distances with acoustically-based distances all correlation coefficients are significant. For both vowels and consonants they are not extremely high, although for vowels higher correlation coefficients were found than for consonants. Therefore, in Chapter 7 the use of both feature-based and acoustically-based segment distances will be validated.
Figure 4.1: Different acoustic representations of four sounds as pronounced by John Wells. Starting with the first row we see respectively spectrograms, Bark-filters, cochleagrams and formant tracks obtained on the basis of the original samples.