Automatic Selection of Speech Segments for Concatenative Speech Synthesis

presented by
THOMAS LUDWIG EWENDER
Diplom-Informatiker
born February 12, 1978
citizen of Germany

2012

Contents

List of Abbreviations

ANN	artificial neural network
F_0	fundamental frequency
GCI	glottal closure instant
GMM	Gaussian mixture model
MFCC	Mel-frequency cepstral coefficients
SVM	support vector machine
TD-PSOLA	time-domain pitch synchronous overlap add method
TTS	text-to-speech

Abstract

In concatenative speech synthesis, corpus generation so far has required tedious manual or semi-automatic work in the post-processing step and, most notably, in the selection of units from speech recordings. The reason for that is the lack of a quality measure to decide which phone segments are appropriate to be selected. This study presents such a measure that considers the following aspects of phone quality: spectrum, phase, fundamental frequency, duration, voicing, intensity and plosive quality. This quality measure is designed to favour phone instances with properties that are in any of these aspects desirable for concatenation-based speech synthesis.

To describe these aspects quantitatively with features, several novel signal analysis methods were developed, in particular to describe fundamental frequency, pitch marking and voicing. To weight and combine these features, two approaches are discussed. First, as a heuristic method a linear sum of penalty functions is proposed. This simple method is effective and practicable as it does not require training data and gives reproducible and transparent results. As a second method, a machine learning approach is presented. This approach was trained with few and unbalanced training data, that partly exhibited undefined feature values. To allow such a training, we devised a neural network variant based on back-propagation and tuned it to the training data with an extensive number of experiments on synthetic data. This machine learning approach is free from any heuristic penalty functions and we believe that the non-linear combination of features reflects the perceptual impression better than a linear sum.

The phone quality measure based on the machine learning approach was applied to create four diphone corpora from four different voices in a fully automatic way. The intelligibility of these corpora was evaluated with rhyme tests, one of the corpora was additionally evaluated for segmental quality. The results of these tests confirm that the phone quality measure can be applied to select a high-quality diphone set from a speech database. As a consequence of this study, tedious manual work in the creation of such diphone sets can now largely be eliminated.

Kurzfassung

Um einen Korpus für Konkatenationssynthese zu erstellen, mussten die Grundelemente bisher mühsam manuell oder halbautomatisch aus dem aufgenommenen Sprachmaterial ausgewählt sowie die Signale nachbearbeitet werden. Der Grund dafür ist, dass bisher kein Gütemass existiert, das entscheidet, ob ein Laut für einen Korpus geeignet ist oder nicht. Die vorliegende Arbeit stellt ein neues, geeignetes Gütemass vor, das die folgenden Aspekte von Lautqualität berücksichtigt: Spektrum, Phase, Grundfrequenz, Dauer, Leistung, Stimmhaftigkeit und Plosivqualität. Das Gütemass bevorzugt jene Laute, die in all diesen Aspekten die für Konkatenationssynthese erwünschten Eigenschaften aufweisen.

Um diese Aspekte quantitativ zu beschreiben, werden mehrere neue Methoden zur Signalanalyse vorgestellt, mit dem Fokus auf der Bestimmung von Grundfrequenz, Zeitpunkt der Glottisschläge und Art der Stimmhaftigkeit. Zur Gewichtung und Kombination der in der Signalanalyse gewonnen Merkmale werden zwei Ansätze erörtert: Als erstes wird ein heuristisches Verfahren gezeigt, das den Ausgang von Gewichtungsfunktionen aufsummiert. Dieses Verfahren ist effektiv und praktikabel, da es keine Trainingsdaten benötigt und reproduzierbare und transparente Ergebnisse liefert. Als zweites Verfahren wird ein statistisches Lernverfahren vorgestellt. Dieses Verfahren wurde mit wenigen und noch dazu ungleich verteilten Trainingsdaten, die teilweise undefinierte Werte aufwiesen, trainiert. Um diese Training zu ermöglichen, wurde ein neuronales Netz auf der Basis von Backpropagation abgewandelt. Die optimale Konfiguration und Parametrisierung dieses neuronalen Netzes wurden mit umfangreichen Versuchen anhand von synthetischen Daten bestimmt. Dieses statistische Lernverfahren benutzt keinerlei heuristisch motivierte Gewichtungsfunktionen und gibt, nach unserer Meinung, durch die nicht-lineare Kombination von Merkmalen den kombinierten Wahrnehmungseindruck besser wieder als eine einfache Summe.

Das auf dem statistischen Verfahren basierende Gütemass wurde benutzt, um vier verschiedene Diphonkorpora aus dem Sprachmaterial vier verschiedener Stimmen automatisch zu erstellen. Die Verständlichkeit dieser Korpora wurde mit Hilfe von Reimtests gemessen, für einen der Korpora wurde zusätzlich die segmentale Qualität bewertet. Die Auswertungen bestätigen, dass das vorgeschlagene Gütemass verwendet werden kann, um einen qualitativ hochwertigen Diphonkorpus aus aufgenommenem Sprachmaterial automatisch zu erzeugen. Dank dieser Arbeit ist es nun möglich, bei der Erstellung eines solchen Korpus auf mühevolle Handarbeit weitgehend zu verzichten.

Chapter 1

Introduction

1.1 Concatenative Speech Synthesis

Voice user interfaces have increasingly become part of new products, most noticeable in mobile and automotive industry, where we are on a cusp of a paradigm change in human-machine interaction. In many situations, voice output is making communication with the machine easier and less demanding for humans, especially in situations where either hands-free and eyes-free communication is necessary, or devices are too small for convenient direct manipulation. The former is the case in the cockpit of a car while driving, when traffic directions, messages or news articles are read to the driver while he can stay focused on the surrounding traffic. The latter is the case for pocket-size devices, such as PDAs or mobile phones, where direct manipulation interfaces like touchscreens or keyboards are demanding to use. A novel field, where speech synthesis applications are increasingly employed, is the speech synthesis of ebooks. In a similar way, eyes-free listening is often convenient if not necessary, and, in addition, speech synthesis is opening the way for illiterates and visually impaired persons to enjoy texts they so far did not have access to.

As a result of this considerable increase in speech synthesis applications, there is a growing demand for higher quality and a larger choice

of synthetic voices. However, voice production for any kind of text-to-speech (TTS) system is still very time-consuming and dependent on individual know-how. Today, most TTS technology in industry that creates high quality speech is based on *concatenative synthesis*.

Concatenative synthesis is based on the concept of generating arbitrary speech signals with the concatenation of appropriate signal segments, that are taken from natural human speech. This approach has the advantage that on the segmental level, which is on the level of signal segments, only natural signals are employed. These segments, which are used in the concatenation step, are assembled in a collection of speech signals, a so-called *corpus*. The basic elements of such a corpus are selected from a generally large number of speech recordings, spoken by a single voice talent. This compilation of speech recordings is called a *speech database*. In order to generate highly natural sounding speech, the segments that constitute a corpus have to be chosen in an appropriate way from such a speech database. Phone segments, for example, cannot be used, as their concatenation would produce abrupt changes in the speech signal, which would be perceived as highly unnatural. To avoid any abrupt changes, segments have to be concatenated where they are most stationary: in the centre of the phone. Accordingly, the segments must contain all possible phone transitions. Typically, polyphones are used, which are segments which start in the middle of a phone, may stretch over several phones and end in the middle of a phone. A special polyphone case is the diphone, which ranges from the middle of one phone to the middle of the next.

Two strategies are commonly used to generate speech corpora:

- For compact corpora, the basic elements are contained only once and are designed to be as short as possible. This approach is called *diphone synthesis*. In a speech signal that is supposed to sound natural, phones must adopt different fundamental frequency (F_0), duration and intensity, depending on their word and sentence position. As every diphone is only present exactly once in the corpus, the diphones must be prosodically modified before they are concatenated. This modification, however, may impair speech quality. The most common approach for duration and frequency modification is TD-PSOLA, which stands for time domain pitch synchronous overlap add method (see [CM89]).

- To avoid modification of the segments before concatenation, a corpus with more varied prosodic and spectral characteristics can be aimed for. To this end, a large single-speaker corpus is compiled. From this corpus polyphone segments, called *units*, are selected. This approach is known as *unit selection*. The units to be selected should be as long as possible, should contain the phone sequence that is required, and fit prosodically to the sentence to be synthesised. With this approach, more natural-sounding speech can be produced than with diphone synthesis as, on the one hand, the units do not have to be prosodically modified and, on the other hand, the signal contains fewer concatenation points. Nevertheless, the number of concatenation points depends on the corpus, so the quality and intelligibility of the synthesised sentences can vary considerably. In practise, unit selection does not go entirely without any prosodic modification of units as the F_0 has to be adapted around concatenation points to prevent sudden F_0 jumps, that generally sound disturbing.

The production of corpora, notwithstanding for diphone synthesis or unit selection, encompasses an elaborate process, which consists of various complex steps. First, a speaker has to be selected that is suitable in terms of voice pleasantness, accent and signal processability. Then, the speech database is recorded, which takes up to several weeks. These ongoing recordings are constantly monitored by experts from different fields. After that, the database is segmented and the signals are post-processed by a sound engineer to identify low quality segments, labelling errors or pronunciation variants. Finally, language experts generate a corpus from the speech database, by evaluating and selecting segments in a manual or semi-automatic way.

1.2 Problem Statement

This demand for tedious manual work in the selection of segments and partly also in the post-processing step is due to the lack of a quality measure that could help to decide which phone segments are appropriate to be selected. Such a quality measure should, first of all, be able to decide if phones are of good phonetic quality, in other words if they are clearly articulated and unambiguously identifiable instances of these phones. Furthermore, the measure must determine whether the phone signal can be prosodically modified without impairing the perceived speech quality and if it can be concatenated without producing audible artifacts at the concatenation points. Such a quality measure could then be used to determine the most suitable speech segments from a speech database. From these speech segments speech corpora could then be automatically generated.

Automatic phone quality judgement for corpus creation was only considered to a small extent so far. In [TH99], the best diphone variant is selected using the cepstral distance between the two semi-diphones and the corresponding phone centroids as the only automatic measure. Unit selection does not directly consider the quality of the selected units in their target costs during synthesis, because no acoustic properties for the target units are known. Phone quality is considered only indirectly through the concatenation costs, which only take into account spectral discontinuities (see [CB96]). However, phone quality, what concatenative speech synthesis is concerned, not only depends on spectral quality, but has several, partly orthogonal aspects.

In this thesis, for the first time a comprehensive phone quality measure is proposed, which can be applied for the automatic selection of speech segments for concatenative speech synthesis. This measure takes all necessary aspects into account, based on detailed signal analysis. These aspects include various characteristics like spectral characteristics, fundamental frequency, duration, pitch marks and voicing characteristics. In this context of signal analysis, novel methods for the analysis of fundamental frequency, voicing characteristics and pitch marking were developed and applied to determine the phone characteristics. These phone characteristics are described with features and combined to obtain a single measure for the overall quality of a phone. For this combination, we used a machine learning approach that weights and combines the features in a non-linear way. To evaluate the quality of the proposed approach, we created four diphone corpora from different voices and evaluated intelligibility and segmental quality with a listening test.

1.3 Scientific Contributions

The following contributions result from the present thesis:

1. We propose a novel approach to extract the fundamental frequency from a speech signal. This approach is based on a clear mathematical model and produces virtually error-free F_0 contours. In addition, it generates smooth interpolations of the F_0 contour in unvoiced speech segments that are located between voiced segments.

2. We suggest a frame classification approach that not only distinguishes voiced and unvoiced speech segments but also mixed excitation, irregularly glottalized and silence segments.

3. We propose a new approach on pitch marking that takes into account the signal properties and applies different features according to some heuristic. The proposed pitch marking algorithm clearly improves the quality of synthesised speech generated by a concatenative text-to-speech system that uses TD-PSOLA for prosodic modifications.

4. We used the information that can be gathered from the signal with the methods described above to augment the TD-PSOLA algorithm. Contributions are related to diverse problems, like treatment of mixed excitation and irregularly glottalized speech, treatment of plosives and compensation of energy variations.

5. We investigated phone quality aspects that play a role in the selection of phone segments for concatenative speech synthesis and quantified these aspects with various features. A modified neural network was trained with an educational learning method. This neural network allows the features to be combined in a single measure that assesses phone quality. This measure was finally used to select elements for the creation of diphone corpora.

1.4 Structure of the Thesis

The remainder of this thesis is structured as follows:

Chapter 2 presents a new method for the estimation of a continuous F_0 contour.

Chapter 3 is dedicated to a frame classification method that not only distinguishes voiced and unvoiced frames but also mixed, irregularly glottalized and silence frames.

Chapter 4 describes a new approach to pitch marking. Unlike other approaches that use the same combination of features for the whole signal, we take into account the signal properties and apply different features according to some heuristic.

Chapter 5 is concerned with an extended version of TD-PSOLA, which is based on the information that can be extracted from the signal with the methods presented in the previous chapters. This information allows for a more appropriate F_0 and duration modification, which is dependent on the local voicing characteristics of the signal.

Chapter 6 introduces the various aspects of phone quality and describes how these aspects can be quantified with features.

Chapter 7 describes a simple method to combine these phone quality features. We used penalty functions on these features and summed up these function values to create a single score for each phone.

Chapter 8 elaborates on an alternative approach to the one presented in Chapter 7. This approach replaced the penalty functions with a machine learning approach to weight the features and combine them in a non-linear way.

Chapter 9 presents the evaluation of diphone corpora by means of subjective listening tests. These corpora were created based on the machine learning approach described in Chapter 8.

Chapter 10 concludes the thesis with a final discussion.

Chapter 2

F_0 Extraction

This chapter describes a new method for the estimation of a continuous fundamental frequency (F_0) contour. First, the purpose of continuous F_0 contours will be motivated as a basis for the continuous fundamental wave, which is a main constituent for the frame classification method presented in Chapter 3 and for the pitch marking approach presented in Chapter 4. Subsequently, we will present the new method in detail by introducing high-resolution cepstrograms and describing the optimisation algorithm to determine the optimal T_0 contour in these cepstrograms. From the T_0 contour, the F_0 contour can then be easily derived.

2.1 Introduction

In voiced speech, airflow from the lungs via the trachea causes the vocal folds to vibrate in a quasi-periodic way. In literature, the rate of vocal fold vibration is referred to as *pitch* or *fundamental frequency*. The exact terminology depends on the point of view on the problem (see [Tal95, Hes08]): The term pitch is associated with the *perception* point of view and describes the auditory perception of a tone by the listener. The term fundamental frequency is associated with the *signal processing* point of view and characterises an inherent property of a quasi-periodic signal. The fundamental frequency correlates well with the perceived pitch, and in literature the term pitch is often used in a wider sense as some kind of common denominator for both terms. In the following, we will use the term fundamental frequency (F_0) as we focus on the signal processing point of view.

Fundamental frequency extraction has a long and extensive history with the most important developments being made in the 1960s and 1970s. A bibliography dating from 1983 already includes some 2000 entries [Hes83], at least another 1000 entries would have to be added to consider more recent developments [Hes08]. Therefore only a selection of approaches can be presented here.

A commonly used method for F_0 extraction is RAPT [Tal95], which is part of the well-known ESPS software package[1] for speech analysis. This approach first applies a normalised cross-correlation function, which corresponds to a modified autocorrelation function with some energy normalisation applied. Subsequently, dynamic programming is used to search for an optimal path through F_0 candidates from consecutive frames. YIN [dCK02b] is another well-known algorithm based on the autocorrelation method. The algorithm consists of an number of processing steps, one building upon the other to prevent errors which are typical for the autocorrelation method.

Numerous applications in speech processing depend on F_0 extraction, including speech synthesis, coding, recognition and segmentation. The quality of concatenative speech synthesis systems can benefit significantly from a good estimation of the F_0 contour of speech signals. It allows on the one hand an optimal prediction of the target prosody of the speech to be synthesised and on the other hand the selection and prosodic modification of the segments to be concatenated.

In terms of prediction of target prosody in general and F_0 contours in particular, a most successful approach is based on a recurrent multilayer perceptron as described in [Tra95] and [Rom09]. These studies show that much better models can be achieved with continuous F_0 contours rather than with only piecewise defined ones. Hence, we need a possibility to estimate for a given speech signal an F_0 contour that is

[1] `http://www.speech.kth.se/software/#esps`

accurate in voiced parts and reasonably smooth in unvoiced parts.

When segments are concatenated they usually have to undergo some prosodic modifications, in particular if the size of the speech corpus is minimal as in diphone synthesis. Such modifications can for example be performed with time or frequency domain PSOLA (see [CM89]). A prerequisite is again an accurate estimate of the F_0 contour. But prosodic modification also depends on an accurate frame classification method as the one presented in Chapter 3. From the continuous F_0 contour (see Section 3.2.1) a continuous fundamental wave can be computed, which is a basis for the features used in this frame classification method (see Chapter 3). In the same way, accurate pitch marking is crucial to achieve high quality results. The pitch marking approach presented in Chapter 4 also relies on the continuous fundamental wave as one of the two main features to determine pitch mark positions.

2.2 Estimation of the T_0 Contour

Our approach to estimate a continuous F_0 contour is motivated by the fact that humans can easily and reliably "see" the T_0 contour of a speech signal from a suitably drawn cepstrogram such as the one shown in Fig. 2.3: the contour has to follow the strong cepstral peaks (i.e. the bright tracks), has to be somewhat smooth and should not make unreasonable detours in completely unvoiced regions.

This task can be considered as an optimisation problem, namely to find the optimal curve along the cepstral peaks while the curve has to meet some constraints at the same time. As constraints we use the probability distribution of the local declination and curvature that was estimated from natural F_0 contours, as shown in Section 2.2.2. The optimisation is detailed in Section 2.2.3.

2.2.1 The High-Resolution Cepstrogram

In order to get the high-resolution cepstrogram, we first compute the logarithmic power density for every frame:

$$S(k) = log(\frac{1}{NU}|X(k)|^2)\,, \qquad 0 \leq k \leq N-1 \tag{2.1}$$

where $X(k)$ is the discrete Fourier transform of the frame, N the window length and U a constant to compensate for the window function. We used a Hamming window and hence $U = 0.3974$. The power density is defined for the discrete frequencies $f_k = f_s k/N$, with $k = 0, 1, \ldots, N-1$.

Subsequently, we eliminate the frequency components of the power density spectrum that are higher than $f_b = f_s b/N$. In order to get a cepstrum with sufficiently high resolution, we apply padding to the power density spectrum

$$\begin{aligned} S'(k) &= S(k)\,, & 0 \leq k \leq b \\ S'(k) &= S(b)\,, & b < k < M-b-1 \\ S'(k) &= S(M-k-1)\,, & M-b-1 \leq k < M \end{aligned} \tag{2.2}$$

where M is the number of points. Finally, we use the inverse Fourier transform to calculate the cepstrum with a quefrency resolution of $N/(Mf_s)$:

$$c(m) = \frac{1}{N}\sum_{k=0}^{N-1} S(k)e^{j(2\pi/N)km}\,, \qquad 0 \leq m \leq N-1 \tag{2.3}$$

For the cepstrogram shown in Fig. 2.3, we used a 50 ms Hamming window at a sampling frequency of 22.05 kHz, f_b and M were set to 5 kHz and 8192, respectively.

2.2.2 Local Declination and Curvature

We define the local declination $d(t)$ and the curvature $c(t)$ of the discrete time sequence $q(t)$ with sampling points at every time interval T_s as follows:

$$d(t) = \frac{(q(t) - q(t{-}2T_s))/2 + q(t) - q(t{-}T_s)}{2T_s} \tag{2.4}$$

$$c(t) = \frac{q(t) - 2q(t{-}T_s) + q(t{-}2T_s)}{T_s} \tag{2.5}$$

In our case $q(t)$ stands for a sequence of quefrencies that are computed from consecutive frames of a speech signal with a frame shift of T_s. In fact, we apply the logarithm on these quefrency values, therefore d and c are relative changes and will be expressed in the following as %/s (percent per second).

The 2-dimensional probability distribution of the declination and curvature was estimated from the voiced parts of some 17 hours of speech from various speakers and languages. In a first step, we detected the F_0 of all signals with an algorithm similar to YIN (see [dCK02a]) and converted the F_0 values to logarithmic T_0 values. Then for triples of consecutive log T_0 values the declination and curvature was computed. An overview of the resulting pairs of declination d and curvature c is shown as a normalised histogram in Fig. 2.1.

Since an empirical probability distribution as shown in Fig. 2.1 is not practically usable, we approximated it with a 2-dimensional Gaussian mixture model with two mixture components. The probability density function $p(d, c)$ of the resulting model is shown in Fig. 2.2.

2.2.3 Finding the Most Probable T_0 Contour

To determine the optimal T_0 contour in the log cepstrogram $C(t, l)$, where t and l are the discrete time and log quefrency, respectively, we devised a Viterbi-like procedure (inspired by [KdC05] and [JBB07]). This procedure evaluates the globally optimal sequence of log quefrency values over all discrete times t. The optimisation is based on a local score $\alpha(t, l)$ that considers $C(t, l)$ and the likelihood of the local declination and curvature. More formally, the local score is defined as

$$\alpha(t, l) = p(d, c) \cdot e^{wC(t,l)}, \tag{2.6}$$

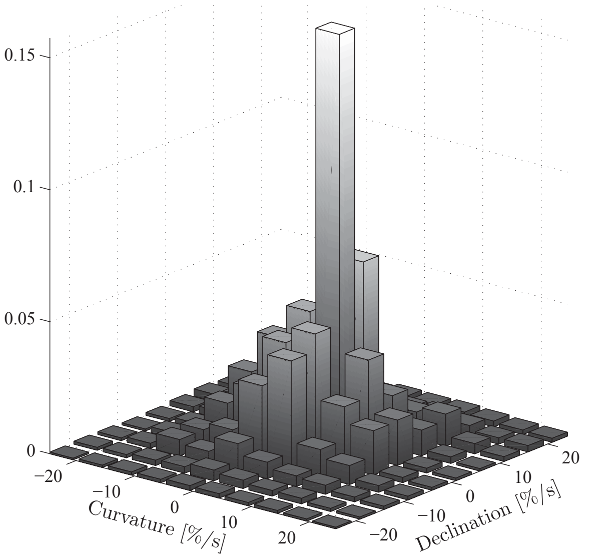

Figure 2.1: *Empirical probability distribution of the local declination and curvature of natural T_0 contours*

whereby w is a weighting factor that equals the signal power of the corresponding frame. Note that $p(d, c)$ depends on the two preceding points of the T_0 contour (cf. equations 2.4 and 2.5). The overall score of the optimal T_0 contour can then be found with the following iteration over all discrete quefrency and time values:

$$\delta(t, l) = \max_k \{\delta(t{-}T_s, k) \cdot \alpha(t, l)\}\,, \tag{2.7}$$

where $\delta(t, l) = 1$ for $t \leq 0$. To find the most probable sequence of (t, l) pairs at the end of the recursion, it is necessary to store the optimal predecessor l for every time t and quefrency l in $\Psi(t, l)$ during the recursion. Like in the Viterbi algorithm, the optimal sequence of (t, l) pairs can then be found by starting at the end point and going iteratively backwards. A resulting T_0 contour is shown as a continuous line in the cepstrogram in Fig. 2.3. The computation of this search algorithm is done in log domain for numerical reasons.

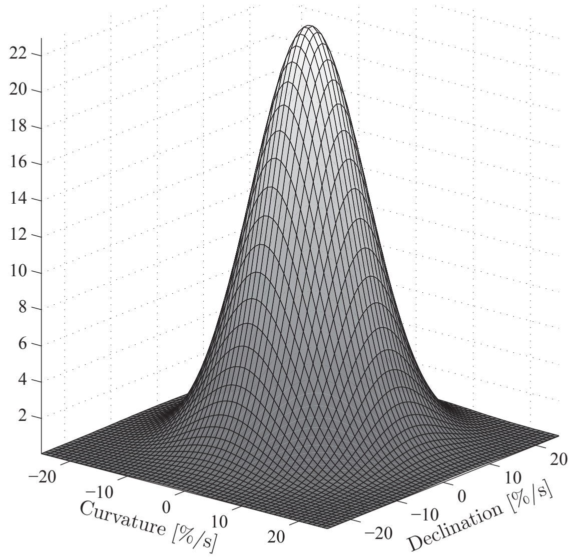

Figure 2.2: *Probability density function $p(d, c)$ of the GMM approximating the 2-dimensional distribution of local declination d and curvature c of natural T_0 contours*

2.3 Evaluation

We refrained from comparing our F_0 results with laryngograph-based estimates, as it is commonly done, because unclear frames are usually excluded from such tests for objectivity reasons. However, unclear frames are most interesting with respect to our application (prosodic modification of speech). Furthermore, it is questionable whether for such frames objectively more reliable values can be estimated from a laryngograph signal at all. Instead of a not very helpful comparison we demonstrated the quality of our results by means of examples enclosed with [EHP09][2]. As an example, the very robust behaviour of our method is illustrated in Fig. 2.3, a speech signal of a female Mandarin speaker, where even the extreme F_0 drop (shown as a T_0 increase) at the creaky speech segment at around 2.2 s is captured with our method.

[2] `http://www.isca-speech.org/archive/interspeech_2009/i09_0100.html`

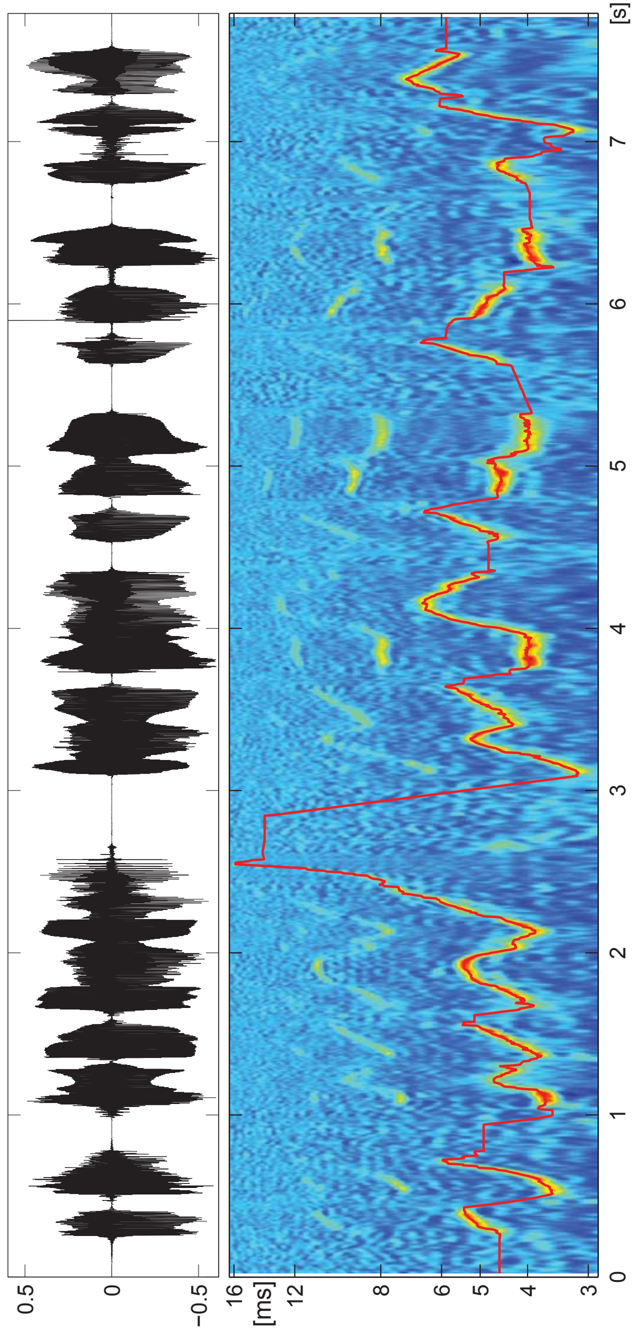

Figure 2.3: *Speech signal and the corresponding high-resolution cepstrogram with T_0 contour, drawn on a logarithmic quefrency axis. Because $F_0 = 1/T_0$, the F_0 contour can easily be achieved by vertically flipping the T_0 contour.*

Note that the large detour of the T_0 contour in Fig. 2.3 is in the pause after the creaky speech segment and is thus irrelevant as this part refers to silence and will not be used.

2.4 Discussion

The presented F_0 detection is based on a clear mathematical model and on statistical properties acquired from a large speech corpus. The global optimisation (in contrast to piecewise as in [KdC05] and [JBB07]) gives more robust results and also does not need any post-processing of the F_0 or T_0 contours. The estimated F_0 contours are virtually perfect: we have not seen any errors in all manually inspected F_0 contours. Furthermore, the F_0 detection also works well for expressive speech with very high F_0 dynamic which several authors reported to be a problem for their algorithms. The processing time for a signal is about real-time, but can be severely reduced by reducing the maximum F_0 change from 40 octaves per second, which is appropriate to catch the most extreme F_0 contours, which may appear in creaky voice segments.

Our F_0 contours are very well suited for the type of F_0 modelling described in [Rom09] and for the generation of a continuous fundamental wave as it is used in the frame classification method presented in Chapter 3 and in the pitch marking approach presented in Chapter 4.

Chapter 3

Frame Classification

This chapter is concerned with a frame classification method that classifies the frames of a speech signal into five classes: voiced, unvoiced, mixed, irregularly glottalized and silence. We present and motivate the features to describe the signal. These features are then fed into an Artificial Neural Network (ANN) that assigns one of the five classes to each signal frame. As an alternative classifier, Support Vector Machines (SVM) are investigated. Finally, the results of both classifiers are presented and discussed.

3.1 Introduction

Besides a precise estimate of the F_0 contour (see Chapter 2) and accurate pitch marking (see Chapter 4), prosodic modification depends on an accurate frame classification method.

First publications on frame classification date back to the 1970s [AR76, Sie79]. Until today, most frame classification approaches are restricted to binary classification, which means they only distinguish the two classes voiced and unvoiced [Sie79, AS99, LL03, SISY04, MJ08, MHMH07, MIH^{+}11]. The algorithms presented in these approaches comprise classification based on a combination of thresholds on various features [AS99], a matching pursuit algorithm used with Gabor decomposition [LL03], a Gaussian mixture classifier [SISY04], and a linear classification of features based on wavelet transforms. Furthermore, [MJ08] suggests an empirical mode decomposition model [MHMH07], which was augmented with an adaptive thresholding approach [MIH$^+$11].

One approach included silence as a third class (see [QH93]), and there are a few approaches that distinguish the four classes voiced, unvoiced, silence and mixed excitation. One of the first approaches that introduced mixed excitation as an additional class was suggested in [SB82]. The motivation to introduce a mixed excitation class in this work was twofold: on the synthesis side, that is for formant synthesis, the mixed excitation class would allow to better model voiced fricatives and devoiced vowels. On the recognition side, this class would allow to recognise voiced fricatives more accurately. This work explored two kinds of classifiers, a Gaussian classifier and a linear discrimination function. This linear discrimination function was introduced to avoid making simplified assumptions about the unknown statistical distribution of features but did not lead to better results than the Gaussian classifier in the speaker-independent case. Another approach that used four classes was presented in [CHL89]. This approach, however, used both the electroglottogram (EGG) and the speech signal and derived the classification from the differences in the characteristics of the two signals. However, this approach is limited to cases where the EGG signal is available, which is not the case for most recordings.

3.2 Frame Classification

It was outlined in Section 2.1 that in order to perform high quality prosodic modification we need information about the frame properties such as voicing, pitch regularity, etc. Motivated from this application, we defined five classes of frames that must be treated differently in prosodic modifications. Details on how the frame class affects the prosodic modification procedure of a signal period are presented in Chapter 5. The five classes are:

voiced: Speech with clearly perceptible voicing; harmonic signal; not noisy; low frequencies dominant (typical phonemes: vowels, nasals)

unvoiced: Speech without perceptible voicing; noisy; high frequencies above 2 kHz are dominant (typical phonemes: fricatives, unvoiced plosives)

mixed: Speech with voicing and noise; only lower harmonics visible in spectrum; higher spectral components noisy (typical phonemes: voiced fricatives; frequently in voiced-to-unvoiced transitions)

irregular: Speech with irregularly spaced glottal pulses; no significant fricative components; low frequencies dominant; also known as creaky or stiff voice or vocal fry; very frequent in some voices or languages (occurs often in voiced plosives and towards the end of utterances, when F_0 drops to very low frequencies or in Mandarin low tones)

silence: Signal segments with low energy; virtually not audible

These classification criteria are primarily motivated by the application (that is to perform prosodic modifications) rather than by linguistic arguments.

Such a classification can be realised with a classifier like a feedforward ANN or an SVM. Important input information for the classifier can be derived from the local properties of the fundamental wave, as shown in Section 3.2.2. First, we will explain how we generate a virtually continuous fundamental wave for a whole speech signal.

3.2.1 Generating the Fundamental Wave

The fundamental wave can be achieved from the convolution of the speech signal with a Hamming window of the size of the period length T_0 [Ohm94]. Formally, the fundamental wave for a sample i of a signal $x(\cdot)$ is computed as follows:

$$f(i) = \frac{\sum_j x(j) \cdot w(j-i)}{\sum_j w(j)}, \tag{3.1}$$

where $w(\cdot)$ is a Hamming window of length $T_0 = 1/F_0$. This convolution of the speech signal with a Hamming window of length T_0 corresponds to a low-pass filter with a cut-off frequency of F_0 and zeros at the harmonics at $k \cdot F_0, k > 1$. Thus, all harmonics are removed and only the fundamental wave remains (see Fig. 3.1). Since T_0 varies along the speech signal, the size of the Hamming window has to be adapted continuously, which means not only for each frame, but for each sample. Because a T_0 contour, as it results from the optimisation described in Section 2.2.3, is specified with one value per frame, this T_0 contour has to be interpolated to get a T_0 value for each sample of the speech signal.

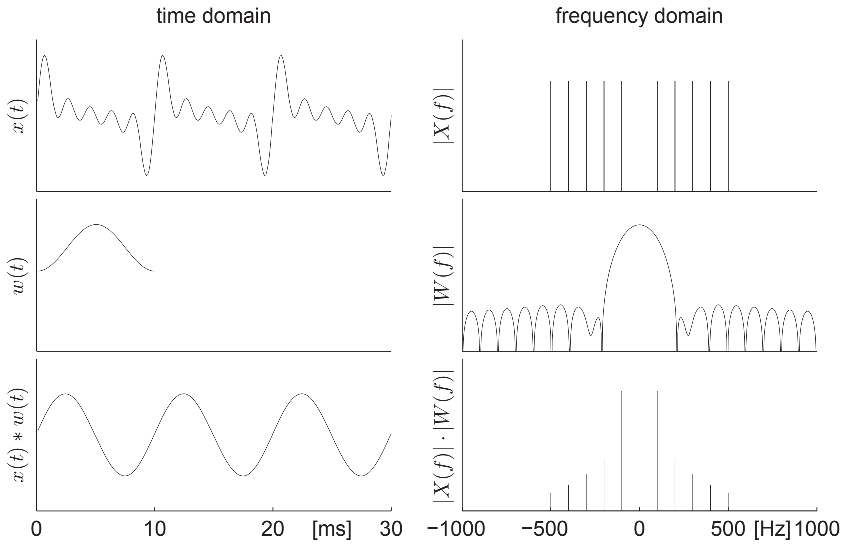

Figure 3.1: *Generating the fundamental wave with an artificial signal (F_0 = 100 Hz). The frequency domain signals are shown in decibel.*

The resulting fundamental wave is completely smooth and particularly shows no discontinuities at frame boundaries. A segment of such a fundamental wave computed from a speech signal is shown in Fig. 3.2.

3.2.2 Properties of the Fundamental Wave

It can easily be seen that the convolution that was sketched above produces a signal that is indeed equal to the fundamental wave of a quasi-stationary harmonic signal like voiced speech. This fundamental wave is close to sinusoidal and its period changes only slowly. Conversely, in clearly unvoiced sections of the speech signal the fundamental wave gets very irregular, in terms of amplitude as well as with respect to the period. Also sections of vocal fry are clearly visible, because their fundamental wave typically is neither close to sinusoidal nor regular periodic.

The local properties of the fundamental waveform can be described by a number of simple features that will be used by the frame classifiers. These and further features will be sketched in Section 3.2.3.

3.2.3 Classification Features

Related work

Most approaches to build a frame classifier consist of two steps. First, a set of speech features is selected that are adequate to the decision task, then some rule is established to classify the patterns based on the values of the features.

An extensive overview on the features that have been used in literature is given in [BS90]. Among the most frequent are the root mean square energy (RMSE), the zero crossing rate (ZCR), LPC predictor coefficients, the normalised autocorrelation coefficient and the ratio of high frequency signal energy (above 4kHz) to low frequency signal energy (below 2kHz). Some of the features are computed from the unfiltered signal and additionally from the preemphasized signal. More recent work also additionally applied cepstral peaks ([AS99]), MFCC ([SISY04]) and wavelet transforms ([MJ08]).

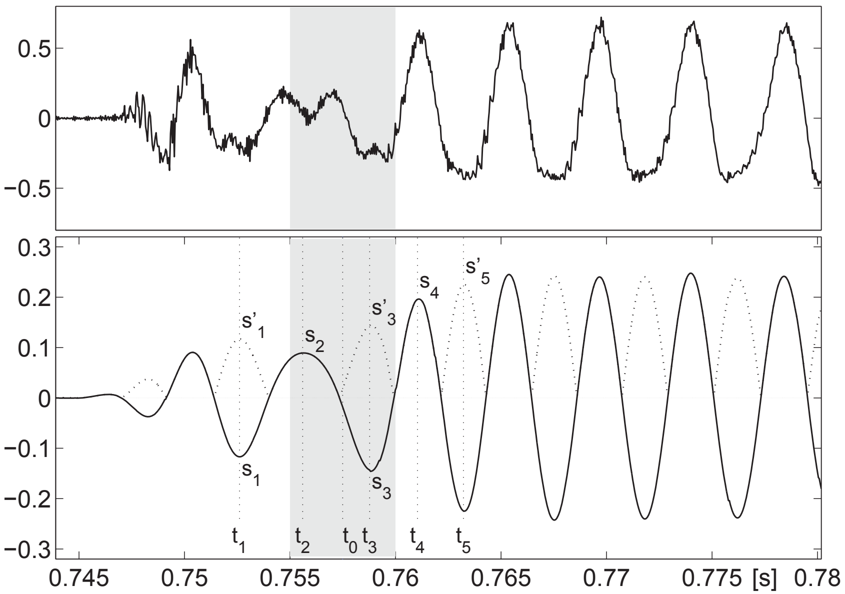

Figure 3.2: *Speech segment of the phones [iː] from a female Norwegian speaker; the derived fundamental wave is shown as a continuous line, the rectified fundamental wave as a dotted line; the current frame is marked in grey; its centre is at time t_0*

Feature list

In addition to the features listed below, we investigated various feature combinations that included wavelet transforms, energy in different frequency bands, and various MFCC. We computed cross-validation error rates on the training data that is presented in Section 3.2.5 and finally selected the set of features that is outlined below. Features number 1 to 5 include classic features for the voiced/unvoiced decision and to describe periodicity, number 6 to 10 describe the degree of regularity of the fundamental wave. The times and points indicated refer to the segment shown in Fig. 3.2. The features are illustrated for the frame interval 0.755 to 0.76 s where t_0 designates the middle of this frame.

1. Zero crossing rate of the speech signal
2. Speech signal power (in logarithmic scale)
3. Spectral tilt (first Mel frequency cepstral coefficient)
4. Dominance of central frequencies (second Mel frequency cepstral coefficient)
5. Periodicity: value of the cepstrogram at quefrency T_0
6. Amplitude of the fundamental wave in terms of distance between the two line segments defined by the points s_2 and s_4 and by the points s_1 and s_3 at time t_0
7. Dynamics of the fundamental waveform: $|1{-}f_1|$, where $f_1 = (t_2{-}t_1)/(t_3{-}t_2) \cdot T_0(t_3)/T_0(t_1)$ and $T_0(t)$ is the period of the speech signal at time t, estimated as described in Section 2.2.3
8. Similar to feature 7, but uses time points t_2, t_3 and t_4: $|1{-}f_2|$, where $f_2 = (t_3{-}t_2)/(t_4{-}t_3) \cdot T_0(t_4)/T_0(t_2)$
9. Regularity of the fundamental waveform: $|1{-}f_3|$, where $f_3 = 4(t_3{-}t_2)/(T_0(t_3) + T_0(t_2))$
10. Irregularity of increase or decrease of fundamental wave amplitude: mean square error of the quadratic regression of the points s'_1, s_2, s'_3, s_4 and s'_5

3.2.4 Training Data Generation

For the development of the classifiers we used 6 minutes of studio quality speech signal from 20 different voices covering a range of 12 European and Asian languages. These speech signals contain a great diversity of voice qualities including plenty of creaky and stiff voice segments. We manually classified these speech signals into segments of the five classes given at the beginning of Section 3.2 by looking at the waveform, spectrogram, the fundamental wave, the pitch contour and by listening to the speech segments. The manually classified speech signals were split into a training and a test set, whereby no speaker was in both sets. This allows to estimate the speaker-independent classification rate.

3.2.5 Training of an ANN-based Classifier

First, we trained a fully connected 2-layer ANN with 10 inputs, 6 nodes in the hidden layer and 5 nodes in the output layer. For the training the back-propagation algorithm was used with randomly selected subepochs. We balanced the training data by removing patterns belonging to over-represented classes. We used no evaluation set, as it is generally used for stopping the training at the optimal point, because previous experiments had shown that the training of our rather small ANN is not critical in terms of over-fitting.

3.2.6 Training of an SVM-based Classifier

As an alternative classification approach to ANN we investigated support vector machines (SVM) [Vap99]. As a significant advantage of SVM we considered the reduced training time compared to ANN. In the field of speech classification, SVM are used for voiced/unvoiced classification (see [QBL04]), for speech/non-speech discrimination (see [RYG$^+$06]) and voice activity detection (see [YRG$^+$06]).

Multi-class Classification

In general, SVM are binary classifiers. Various approaches exist to apply SVM on multi-class classification problems. The most commonly known are one-against-all, directed acyclic graph SVM (DAGSVM) and one-against-one, where it is still an open research question, which of these approaches is suited best [HL02]. In the one-against-all approach, k different SVM models are trained, where k is the number of classes. For the training of SVM i, training patterns from class i are assigned positive labels, all other training patterns are assigned negative labels.

class	unvoiced	silence	voiced	mixed	irregular	total
#data	7921	12611	29080	2760	2274	54646

Table 3.1: *Number of frames for the classes*

In the testing phase, a pattern $\boldsymbol{x}$ is classified with all of the k SVM and predicted to be in the class of the SVM with the largest decision value.

For the DAGSVM, $k(k-1)/2$ binary SVM models are trained, each one on data from two classes. In the testing phase, the DAGSVM uses a binary directed acyclic graph with $k(k-1)/2$ internal nodes and k leaves. Each node is a binary SVM of the classes i and j. For a test pattern $\boldsymbol{x}$, we start at the root node and evaluate the binary decision function. Then we move either left or right depending on the output value until we reach a leaf node, which indicates the predicted class.

In this work, however, we concentrated on the one-against-one method, as exploratory experiments had shown no improvement using one of the alternative methods and the one-against-one method requires considerably less training time compared to one-against-all ([HL02]). Just as the DAGSVM method, the one-against-one method constructs $k(k-1)/2$ binary SVM models, each one trained on data from two classes. For the classification of a pattern $\boldsymbol{x}$, however, a different voting strategy is applied: If the model trained with patterns from classes i and j predicts $\boldsymbol{x}$ to be in class i, the number of votes for class i is increased by one. Otherwise, the number of votes for class j is increased by one. Finally we predict $\boldsymbol{x}$ to be in the class with the highest number of votes. This voting approach is also called the “MaxWins” strategy. In case that two classes have an equal number of votes, the one with the smaller index (i or j) is selected, as it is proposed in [HL02].

We have also investigated alternative voting strategies that not only consider the class decision of a particular model for a pattern $\boldsymbol{x}$ but also the distance of the pattern $\boldsymbol{x}$ from the hyper plane that separates the classes in the feature space. The best results were obtained for a strategy which considers the distance of the pattern $\boldsymbol{x}$ from the hyper plane but caps the maximum distance to the value 1. The results of this voting strategy were in the same range as for the “MaxWins” strategy described above, but no fundamental improvement was achieved.

Coping with Unbalanced Data

The voicing data described in Section 3.2.5 is highly unbalanced with a strong bias towards voiced data (see Table 3.1). In addition to the approach that we used for the ANN training, namely to balance the training data by removing patterns that belong to over-represented classes, SVM offer an additional possibility. Weight parameters that are reciprocally proportional to the class sizes can be used to compensate over-represented classes (see [CL11]).

Kernel and Parameter Selection

The basic SVM is only able to separate patterns linearly in the original feature space. To allow also for non-linear decision functions, the feature data are mapped to a higher dimensional space with a so-called kernel function. There are four commonly used kernel functions from which we chose the radial basis function (RBF) kernel, which commonly considered the first choice for several reasons (see [HCL$^+$]): the RBF kernel can handle the case when the relation between class labels and features is non-linear, it has fewer numerical difficulties compared to other kernels and adds only one additional parameter to the model. An additional reason is that other kernels behave like the RBF kernel for certain parameters.

There are two parameters to be optimised for a SVM that uses an RBF kernel: C and γ. $C > 0$ is the penalty parameter for the error term in the objective function, $\gamma > 0$ is a parameter in the kernel function. The optimal parameter values for C and γ are not known beforehand, thus the best parameter combination must be determined. This can be achieved with a grid-search approach using cross-validation as described in [HCL$^+$]. Various parameter pairs of (C, γ) are tried and finally the pair (C, γ) with the highest cross-validation accuracy is selected. This parameter selection step is the final step in the configuration of a SVM and has to be taken after all other decisions on data scaling, data balancing and the kernel have been made.

3.2.7 Results

We evaluated the accuracy of the classifiers in a strictly frame-wise comparison of the output with the manual frame classification. The mean relative classification rate for ANN was 90.49%. The results per class can be seen in Table 3.2. By allowing a shift tolerance of 5 ms (one frame) for phone transitions, the total classification rate increased by 1.95 %.

	unvoiced	silence	voiced	mixed	irregular
unvoiced	85.28	2.60	0.00	9.68	2.44
silence	1.81	93.15	0.04	1.07	3.93
voiced	0.01	0.03	90.69	3.38	5.89
mixed	8.56	1.14	4.12	76.60	9.58
irregular	1.88	2.60	3.86	7.62	84.04

Table 3.2: *Confusion matrix for the ANN-based frame classifier*

The best mean relative classification rate for SVM was slightly worse with 90.01 %. For this result we used a weighted SVM with an RBF kernel and the parameter values $C = 256$ and $\gamma = 0.0078$. As voting strategy for multi-class decisions we used the “MaxWins” strategy presented in Section 3.2.6. The results per class can be seen in Table 3.3. Again, by allowing a shift tolerance of 5 ms for phone transitions, the total accuracy increased 2.35 %.

	unvoiced	silence	voiced	mixed	irregular
unvoiced	83.89	2.16	0.00	9.86	4.08
silence	1.39	95.25	0.01	0.32	3.02
voiced	0.01	0.08	91.30	3.71	4.89
mixed	8.59	0.93	3.20	79.35	7.92
irregular	1.66	2.28	2.58	5.32	88.16

Table 3.3: *Confusion matrix for the SVM-based frame classifier*

3.3 Discussion

A direct comparison of classification results on the same data is difficult as for many authors the test data is either not available or not sufficiently described to reproduce the results. Many authors recorded their own training and test data [SB82, CHL89, BS90, QH93]. One author used data from the HINT database [LL03] with manually classified labels. Other authors used data from the TIMIT database and either used corrected labels [AS99], original labels [SISY04] or do not specify [MIH$^+$11]. Some authors do not elaborate at all on the origin of their training data [MHMH07]. Most importantly, including additional classes into the classification which are scarcely represented in the training data leads to a higher error rate. Therefore, error rates from classifiers that only take a binary decision are hard to compare to those from classifiers that distinguish more classes.

Nevertheless, we give a short overview on the classification accuracy that is achieved in the literature for different numbers of classes. All the error rates below relate to the speaker-independent case.

For binary classification three approaches achieved very low error rates. [AS99] reports a V-UV error rate of 1.06 %, that is 1.06, % of the voiced frames were classified as unvoiced, and a UV-V error rate of 0.62 %, that is 0.62 % of the unvoiced frames were erroneously classified as voiced. Even lower error rates were reported in [MIH$^+$11] with a V-UV error rate of 0.03 % and a UV-V error rate of 0.4 %. In the first approach, training and test data was labelled by applying some kind of correction factors on the complete data, with the effect that there was some kind of system in the labelling. In the second approach, the frames were also labelled by the authors, however, they do not give any details on the labelling. We believe that error rates in this order of magnitude can only be achieved if the data is labelled with some system that can be reproduced algorithmically. For many frames arguments for both classes are present and thus it is difficult to draw a precise line based on visual and acoustic inspection. Therefore, inherent contradictions in manually labelled data should prohibit such low error rates. [BS90] also reports a very low error rate of 0.4 %. In this work, although manually classified frames were used for testing, it is not clear how the test set, which consists of a very small subset of the available data, was assembled. Other approaches report much higher error rates, [SISY04] mentions an error rate of 6 % on the original TIMIT data,[MHMH07] reports an error rate of 7.6 % on manually labelled speech, and [LL03] achieved an error rate of 16 % on manually classified data from the HINT database.

For the three classes voiced, unvoiced and silence, [QH93] obtained an error rate of 6 % on data that the authors recorded and classified manually.

For the four classes voiced, unvoiced, silence and mixed excitation, [SB82] reported an error rate of 6 %. The scenario described in that work is most comparable to our scenario as several classes are used and the data was manually labelled. However, there are still large differences between the scenario in [SB82] and our scenario. First, in contrast to the training and test data that we used, both training and test data in [SB82] was heavily unbalanced. In training and testing, only some 5 % of the frames were mixed. As a consequence, the linear discrimination function was heavily biased but yielded a relatively low overall error rate due to the few mixed frames in the test data. But also the Gaussian classifier benefits from the unbalanced test data for it is easier to discriminate voiced from unvoiced frames than mixed from unvoiced and mixed from voiced, which clearly shows in the error rates. Furthermore, only four classes were discriminated. Irregularly voiced frames, which are hard to distinguish from regularly voiced ones, were not considered separately. Finally, only English speech data was used, which is also the case for all other publications. Our approach, in contrast, is not only speaker-independent but also a language-independent as it was trained and tested on multiple languages. Considering these additional difficulties in our scenario, we believe that the results of our approach are comparable to the best results achieved in related work.

As can be seen from Tables 3.2 and 3.3, the voiced/unvoiced decision is nearly done perfectly and if we consider only this aspect, the classifier ranks among the best of those presented in literature. If confusions occur, they do between those classes that share signal qualities, as unvoiced and mixed and voiced and irregular. Inspection showed that these confusions mostly concern border cases where arguments for both classes are present.

Although the SVM performs comparably to the ANN, we found that the SVM encountered problems with signal segments that had been manually set to zero, for example to remove breathing noise. We therefore decided to use the ANN classifier for our further work as it requires no special treatment of zero-segments.

Chapter 4

Pitch Marking

This chapter presents a new approach to pitch marking. We will first give an outline of the new method. Then we will introduce the short-term energy and the fundamental wave as the two main features that we use as a basis for our procedure and enlarge upon their respective restrictions and advantages. Next, we present the complete procedure based on the combination of these two features and conclude with an experiment to evaluate the approach.

4.1 Introduction

TD-PSOLA is known to allow for high-quality pitch and time scale modification of speech segments for concatenative speech synthesis. The quality of TD-PSOLA-modified speech largely depends on how the speech signal is split into windowed double period segments. It is generally accepted and can easily be verified experimentally that the best results are achieved if double period segments start at a glottal closure instant (GCI) and end at the next but one GCI.

The GCI is commonly referred to as the maximum of the derivative of the airflow through the glottis. In terms of timing, the GCI is located towards the end of the closure of the glottis. It would be very easy to

detect the GCI from the glottal flow signal. However, this signal is generally not available. Therefore, the GCIs have to be estimated from the speech signal. These estimates are denoted here as pitch marks.

There are several known approaches to set pitch marks. One standard approach consists of two steps: First, pitch mark candidates are generated from local maxima either of the speech signal [CL02, LJ04, KHK09], some wavelet components [SS00] or the inverse filtered speech signal using frame-wise extracted LP coefficients [TR91][1]. In the second step, a subset of these candidates is selected according to optimisation criteria that mainly account for the smoothness of the fundamental frequency contour [CL02, LJ04, SS00, Vel00] and the waveform consistency of the signal around the pitch mark candidates [CL02, Vel00].

Another approach to pitch marking uses a pseudo-state space representation of speech frames and sets the pitch marks at the crossings of the trajectories with the Poincaré plane [HK06]. At the start of a voiced speech segment, this plane has to be placed according to some criterion, for example a local maximum of the signal or the point where the trajectories are most parallel. Both possibilities may result in pitch marks that are very bad estimates of the GCIs. A further problem of this approach is phase drift, which means in some signals the distances between the pitch marks may systematically be slightly too large or too small.

These methods are not robust enough for voiced segments with significant noise components as in voiced-to-unvoiced transitions or vice versa and in voiced fricatives. Furthermore, they are not designed to cope with irregular-pitched segments of vocal fry. Our new approach to pitch marking uses the maxima of the short-term energy contour as the main pitch marking criterion because the maximum of the derivative of the airflow through the glottis corresponds to the energy maximum in the speech signal. Our approach copes very well even with the difficult cases mentioned above and will be outlined in the next section.

[1] The method presented in that paper is available as the function *epochs* of the ESPS software package for speech analysis (`http://www.speech.kth.se/software/`)

4.2 Outline of New Pitch Marking

Estimating pitch marks from amplitude peaks often results in pitch periods with significant jitter, which needs to be reduced by smoothing. We argue that since the concept of TD-PSOLA is based on the idea that the energy should concentrate in the centre of the windowed double period segments, pitch marks should be placed at peaks of the energy contour rather than at peaks of the signal. However, the short-term energy is not always a reliable criterion to set pitch marks (see Section 4.4). But it is possible to detect the locations where the short-term energy criterion should not be used. In these cases, the fundamental wave of the speech signal is used as a robust fallback feature to determine the pitch mark positions. It has to be emphasised that neither of the two features can be used alone for pitch marking. Some problematic cases for the short-term energy feature are illustrated in Section 4.4. Also the fundamental wave alone is not usable for setting pitch marks, because there is no fixed relation between the GCIs as estimated from the short-term energy and the phase of the fundamental wave (see Fig. 4.1). Hence a solution that is based solely on the fundamental wave would fail in some cases. Only with a combination of both features, as described below, good results may be attained for arbitrary voices.

4.3 The Two Main Features for Pitch Marking

Our pitch marking algorithm is based on two features: the short-term energy contour and the fundamental wave. These features are extracted for each sampling point of the speech signal with an analysis window, which has the size of the local signal period. Thus, the resulting features are continuous and smooth. The signal period is derived from a continuous F_0 contour as resulting from the method described in Chapter 2.

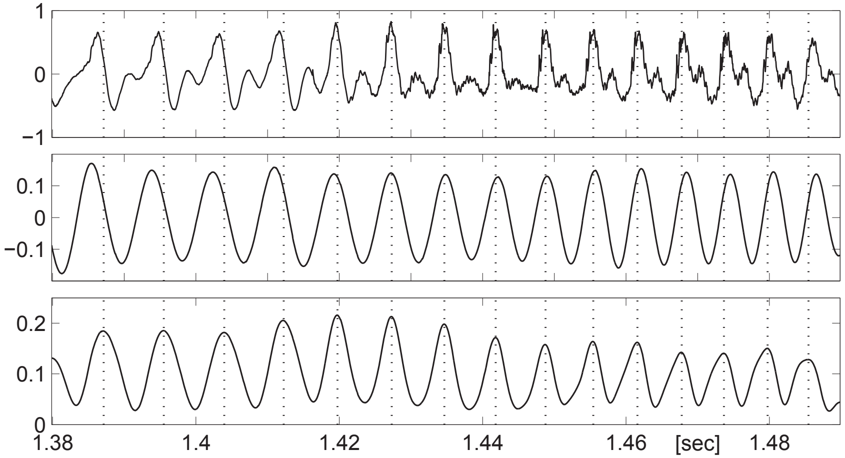

Figure 4.1: *Speech signal (top), fundamental wave and short-term energy (bottom) of the phones [mɪ]. GCIs as estimated from the short-term energy are denoted as dotted lines. The phase of the fundamental wave at the estimated GCIs varies from 248° to 107°.*

4.3.1 Continuous Short-Term Energy Contour

The short-term energy for a sample i of a signal $x(\cdot)$ is computed as follows:

$$E(i) = \frac{\sum_j \left[x(j) \cdot w(j-i)\right]^2}{\sum_j w(j)^2}, \tag{4.1}$$

where $w(\cdot)$ is a Hamming window of length T_0, centred at 0. The window of size T_0 is motivated by the fact that it is long enough to provide a good estimate for signals with noise components and short enough to provide the energy distribution within the length of one period. Since T_0 varies with time, the size of the Hamming window has to be adapted continuously, which means not only for each frame, but for each sample. Because a T_0 contour as it results from the optimisation described in Chapter 2 is specified with one value per frame, it has to be interpolated to get a T_0 value for each sample of the speech signal, in the same way as in the computation of the fundamental wave (see Section 3.2.1).

The resulting short-term energy contour is completely smooth and particularly shows no discontinuities at frame boundaries. A segment of such a short-term energy contour is shown at the bottom of Fig. 4.1.

4.3.2 Continuous Fundamental Wave

The fundamental wave can be computed by convolving the speech signal with a Hamming window of length T_0 as described in Section 3.2.1. A segment of such a fundamental wave is shown in the middle plot of Fig. 4.1.

4.4 Problems with Short-Term Energy

In most cases the short-term energy works fine for the pitch marks (see voiced frames of Fig. 4.4). However, there are cases where short-term energy peaks turn out to be not suitable for pitch marking. We illustrate such problematic cases below.

4.4.1 Pitch Doubling

For nearly sinusoidal speech signals, basing pitch marks on short-term energy peaks would be problematic because the energy contour shows two almost equally high peaks per period (see Fig. 4.3 from 0.45 s). If the energy peaks are chosen without any further consideration it will lead here to the well known effect of pitch doubling. This effect tends to appear more often with female voices where the fundamental frequency can be in the area of the first formant, leading to a dominating fundamental wave in the signal.

4.4.2 Jitter

The reason for jitter is that the energy contour does either not have clear peaks or again two relative maxima per period. Fig. 4.2 shows such a speech signal. After 0.7 s the short-term energy contour first shows dual peaks and later broad peaks with the maximum changing from right to left and to right again. If these short-term energy peaks are used, the resulting pitch marks alternate between the left and right peaks, which results in jitter.

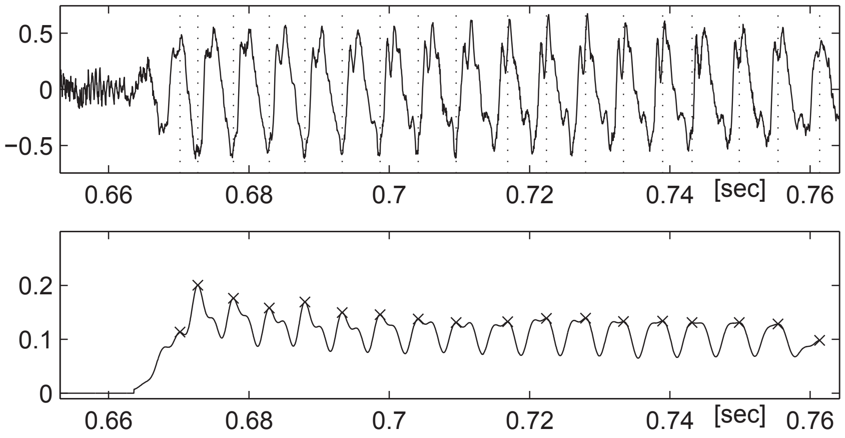

Figure 4.2: *Speech signal (top) and short-term energy (bottom) of the phones [tã]. The dotted lines show the pitch marks if based solely on short-term energy peaks (marked with the plus signs), jitter occurs after 0.71 s.*

4.4.3 Spurious Energy Peaks

High frequency noise can cause spurious energy peaks or can shift the energy peaks in the otherwise periodic signal. As can be seen in Fig. 4.3 from 0.30 to 0.36 s, the short-term energy contour exhibits irregular peaks, whereas the fundamental wave remains periodic and is not perturbed by the high frequency components. This is often observed with voiced fricatives next to vowels as it is the case in Fig. 4.3.

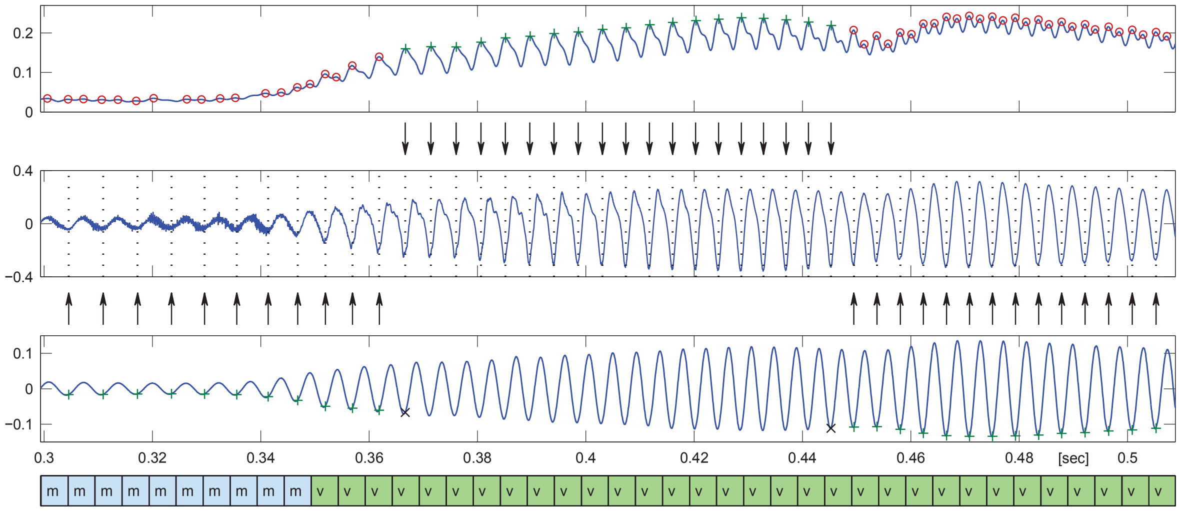

Figure 4.3: *Speech signal of the phones [zum] from the word zoomde uttered by a female Dutch speaker (middle plot). On top the short-term energy is shown (for illustrative purposes we applied the square root), at the bottom the fundamental wave. The plus signs and circles designate extracted peaks. The peaks specified by plus signs are used as pitch marks, either directly as in the case of the short-term energy or in combination with the phase as in the case of the fundamental wave. The pitch marks are shown as dotted lines in the waveform signal. In the bottom plot, the voicing information is shown frame-wise, where the letters v and m denote voiced and mixed frames, respectively.*

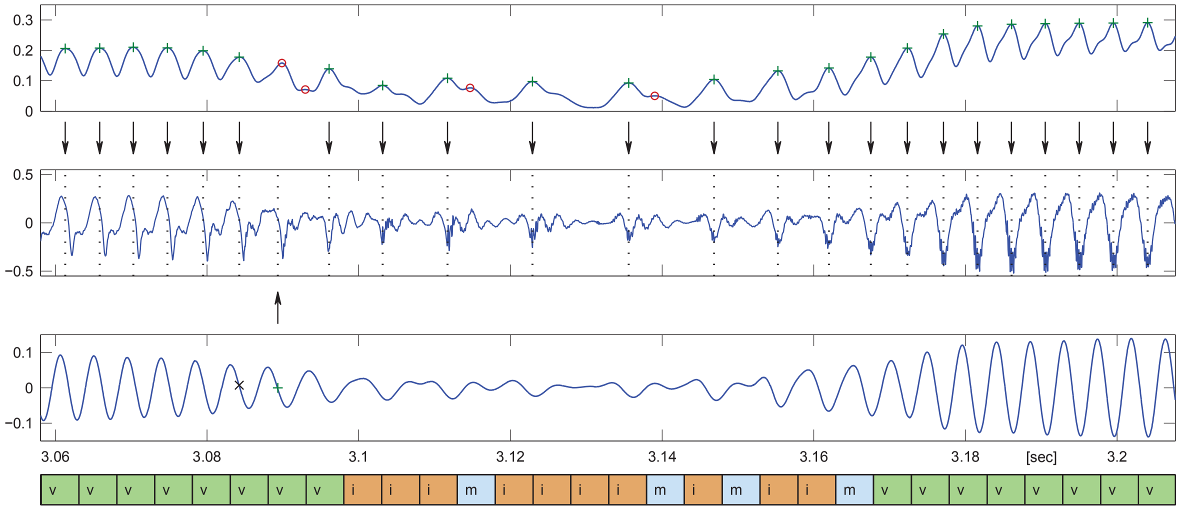

Figure 4.4: *Speech signal of the phones [ŋiː] from the words jiāng yī uttered by a female Mandarin speaker with the same information as described in Fig. 4.3. The letters v, i and m denominate voiced, irregular and mixed frames, respectively.*

4.5 Combining Short-Term Energy and Fundamental Wave

4.5.1 Reliability of the Short-Term Energy

As is apparent from the previous section, a careful estimation of the short-term energy peaks' reliability as pitch mark indicators is most important. We used a test set of nine speakers that were known to be problematic for pitch marking to establish criteria for determining unreliable short-term energy peaks. The speakers were female and male speakers of German, Mandarin, Norwegian, French, Dutch, Turkish, and American and British English. For the test set we selected two sentences from each speaker that we considered to be particularly interesting for pitch marking. Finally, 18 sentences were used, as for German we included both a female and a male voice.

For the development of the criteria we manually inspected the pitch period contour that was computed from the pitch mark positions to identify outliers that may hint towards erroneous pitch mark positions. We used another tool to visualise all the information that the pitch mark decisions were based on in order to assess the effect of individual criteria and their associated thresholds. Finally, the following criteria for determining unreliable short-term energy peaks emerged (for details on the thresholds see Table 4.1):

- Prominence: the peak is not prominent enough compared to adjacent valleys (Fig. 4.3, from 0.45 s).
- Amplitude: the peak's amplitude is below a noise threshold (Fig. 4.3, until 0.345 s).
- Position of adjacent valleys: adjacent valleys are too close or too far away (Fig. 4.4, from 3.09 to 3.16 s).
- Position of adjacent peaks: adjacent peaks are too close or too far away (Fig. 4.4, from 3.09 to 3.16 s).
- Form: the peak is too broad. We measure the extension of the peak at 85% of its maximum height. If this extension covers more than a certain fraction of the current fundamental period (44%), the peak is considered as too broad (Fig. 4.2, from 0.69 s).
- Quality of neighbours: one of the adjacent peaks is considered an insufficient indicator by meeting one of the above mentioned criteria.

Minimum distance to adjacent valley	20%
Maximum distance to adjacent valley	80%
Minimum distance to adjacent peak	70%
Maximum distance to adjacent peak	130%

Table 4.1: *Thresholds for the reliability criteria for short-term energy peaks. The distances are given in percent of the fundamental period and denote the distance from the current short term energy peak.*

If at least one of these negative criteria is fulfilled, the energy peak is considered an insufficient indicator and the algorithm resorts to the fallback methods.

4.5.2 Reliability of the Fundamental Wave

In many cases where the energy contour shows no usable peaks because high-frequency noise components are present in the signal (see for example the voiced fricative at the beginning of the signal in Fig. 4.3), the fundamental wave is almost perfectly periodic and does not exhibit any perturbation. In these cases the fundamental wave can be used as a fallback. However, there are cases where also the fundamental wave fails as a reliable pitch mark indicator, which is frequently the case with creaky voice segments (see Fig. 4.4 at 3.10–3.17 s). Therefore, again a set of criteria for the detection of bad fundamental wave segments was established, with thresholds given in Table 4.2:

- Amplitude: the amplitude is below a noise threshold (Fig. 4.4, from 3.1 to 3.15 s).

- Valley positions: valleys are too close or too far away from their adjacent peak (Fig. 4.4, from 3.1 to 3.165 s).
- Peak positions: peaks are too close or too far away from their neighbours (Fig. 4.4, from 3.09 to 3.155 s).
- Regularity of adjacent valleys: the ratio r of the distances between the peak and its two adjacent valleys is too different from the local increase of F_0.

The fundamental wave is locally considered as not reliable if at least one of the above criteria is fulfilled. The noise thresholds mentioned above are dynamically estimated for each signal using the energy-based loudness measure presented in Section 6.10. The rest of the parameters was manually determined using the set described in Section 4.5.1. The parameters should generalise well, as the test set that was used for their development included a wide range of languages, including Mandarin which is known to be problematic for voice quality, and a fair amount of nine different speakers. However, we did not test the parameters for noise robustness as the application scenario of our proposed pitch marking method included high quality studio speech only.

Minimum distance to adjacent valley	37.5%
Maximum distance to adjacent valley	65%
Minimum distance to adjacent peak	75%
Maximum distance to adjacent peak	130%
Distance ratio r	1.3

Table 4.2: *Thresholds for the reliability criteria for the fundamental wave. The distances are given in percent of the fundamental period and denote the distance from the current fundamental wave peak.*

4.5.3 Procedure to Set the Pitch Marks

As input to the pitch marking procedure the short-term energy and the fundamental wave (as described in Section 4.3) are extracted from the

signal, as well as frame-wise defined voicing information which distinguishes voiced, unvoiced, mixed-excitation, irregularly-glottalized and silence segments (see Chapter 3.2).

We first select glottalized (voiced, mixed or irregular) segments that are delimited by unvoiced or silent regions. Every glottalized segment is processed as follows:

1. We use the time points of the reliable short-term energy peaks in the glottalized segment as initial pitch marks.
2. If there is a region in the segment where short-term energy peaks are unreliable, we check if the fundamental wave can be used to set additional pitch marks. If this is the case, we detect at which phase of the fundamental wave the last valid pitch mark has been set (marked with x-signs in Fig. 4.3) and interpolate or extrapolate along the reliable part of the fundamental wave[2].
3. For regions where still no pitch marks are set, we regard the voicing information: For regions classified as irregular we use prominent short-term energy peaks (which may be irregularly spaced) to set the pitch marks. For regions classified otherwise we set the pitch marks by interpolation (if there are pitch marks to the left and right of the region) or by extrapolation (at the start or the end of the glottalized segment) with a period length derived from the continuous F_0 contour.

4.5.4 Pitch Marks in Unvoiced and Silent Segments

We implemented two strategies for placing the pitch marks in unvoiced and silent segments: one with focus on optimal pitch mark placing and the other with focus on efficient storage.

The strategy for the *optimal placing* takes into account the periods of the voiced segments neighbouring the unvoiced segments. The first approach is inspired by the method presented in [MVV06]. For

[2] In the rare case when no reliable short-term energy peak was found in step 1 for the whole glottalized segment, we select one of the short-term energy peaks that fulfil most of the reliability criteria.

segments that are neighboured by voiced segments we attempt a linear sweep using the period lengths of the neighbouring left and right voiced segment. If the remaining gap to be filled is too large (in the implementation we used 10 % of a period as threshold) we use equally spaced pitch marks with a period chosen to be closest to the mean of the left and right period instead of a linear sweep. At the beginning and at the end of the signal the adjacent right or left period is repeated.

The strategy for the *storage optimised placing* of pitch marks in unvoiced regions aims at using as many constant periods as possible to allow a more efficient encoding and thus save storage space. This strategy works as follows. The preceding voiced period is repeated a few periods (in our implementation we use 3 repetitions) towards the right, and the succeeding voiced period a few periods towards the left. The remaining unvoiced periods are set to a constant length. The constant length is chosen to be be a bit longer than the average period length of the voice.

4.6 Evaluation

Statistical evaluation can only be considered if evaluated pitch marks and reference pitch marks follow the same criteria. So we decided to evaluate our method through the performance of a diphone-based TTS system that applies TD-PSOLA. We built two systems each for four different voices, one male and one female Dutch voice, one female German and one male American English voice. The voices were recorded in studio quality with professional equipment. For the baseline system we used pitch marks generated with the cross-correlation algorithm from the Praat toolkit [Boe02], for the other system we used our proposed method. With the exception of the difference in pitch marking the systems were otherwise equal.

To illustrate some differences in the pitch marks produced with the two methods, two short speech segments are shown in Figs. 4.5 and 4.6. Generally, pitch mark positions coincide roughly in many cases with the Praat pitch marks often showing a systematic offset from the pitch marks of our proposed method and thus from the energy maxima. This offset can be as high as shown in Fig. 4.5. Larger differences can be observed in irregularly voiced segments, where Praat tends to produce equidistant pitch marks due to unvoiced voicing decision or pitch marks that are not set accurately as shown in Fig. 4.6.

The speech signals synthesised with our proposed method showed clear improvement for the two male voices, they contain less reverberation and sound less raspy. Most improvement was achieved for the bassy male American English voice, which sounds more sonorous and less frizzled. The female voices showed isolated improvements in mixed or irregular speech. We did not encounter any cases where our proposed method performed worse than the baseline system. We demonstrated the quality of our results by means of examples enclosed in [EP10][3]. We refrained from a formal evaluation because the improvements can be clearly followed by visual and acoustic inspection of the given examples. For the two male voices the baseline system contains, for example, clearly visible waveform irregularities, pitch irregularities and phase jumps that can be acoustically perceived even by non-experts. The examples include two sentences for each male voice which four to nine segments per sentence, which are highlighted, where differences can be clearly perceived. The sentence durations range between 6 and 14 seconds.

4.7 Discussion

In this chapter we introduced a new pitch marking method. The algorithm places pitch marks basically at the peaks of the short-term energy contour, which is a good estimate of GCIs. For speech segments where the energy peaks are not suitable, we use fallback methods based on the fundamental wave or on the F_0 contour. Whereas the method obtains robust results in voiced, unvoiced and mixed segments, we sometimes observed additional erroneous pitch marks in irregularly voiced segments. This problem can be attributed to the energy threshold, which is not estimated accurately enough for that particular segment. The loudness that the energy threshold is based on is computed over the entire sentence but sometimes drops strongly in irregularly voiced segments. Therefore, a further improvement of the algorithm could be to

[3] `http://www.tik.ee.ethz.ch/spr/pitch_marking_examples/`

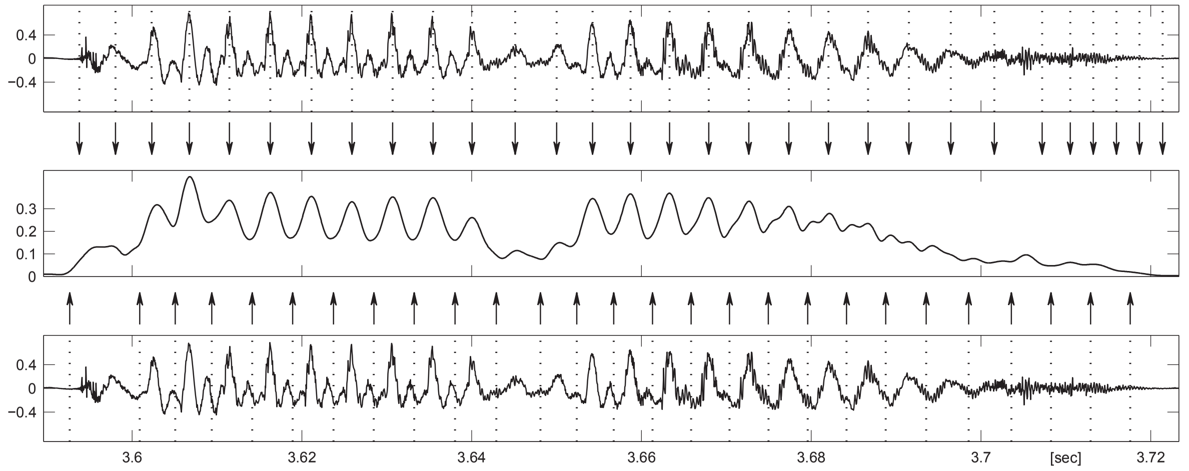

Figure 4.5: *Speech signals of the phones [brɛ] uttered by a female Norwegian speaker. On top, the signal is shown with pitch marks produced with our proposed method. In the bottom plot, the same signal is shown with pitch marks produced with the method from the Praat toolkit. In the middle, the short-term energy is shown (for illustrative purposes we applied the square root). A slight systematic offset can be observed in the pitch mark positions produced by the two methods.*

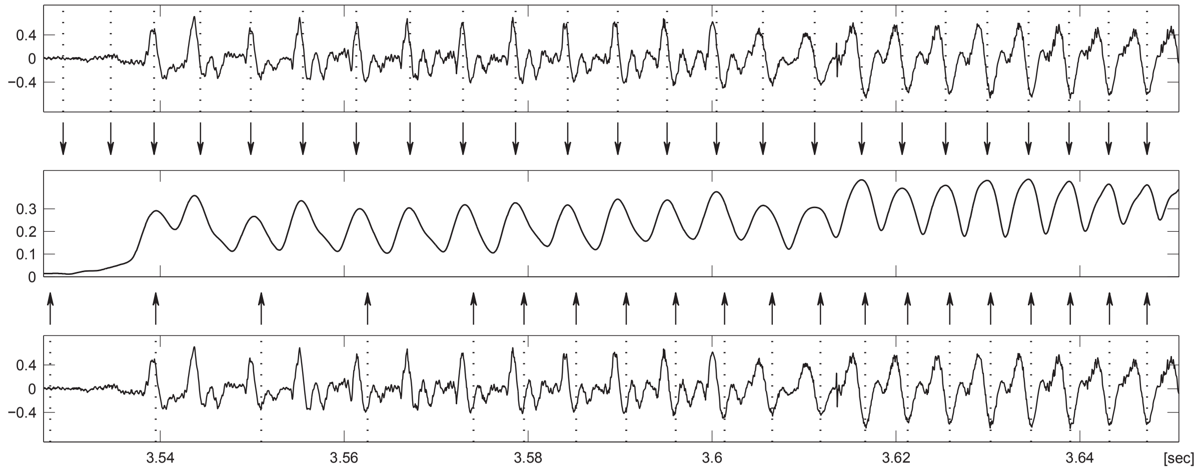

Figure 4.6: *Speech signals of the phones [ɛl] uttered by a female Norwegian speaker. On top, the signal is shown with pitch marks produced with our proposed method. In the bottom plot, the same signal is shown with pitch marks produced with the method from the Praat toolkit. In the middle, the short-term energy is shown (for illustrative purposes we applied the square root). In the result of the Praat method, some pitch marks are omitted at the vowel onset (from 3.54 to 3.57 s).*

investigate a more locally limited loudness measure to be used for the energy threshold. However, for our application, which is prosodic modification of speech segments, these occasional misplacements of pitch marks in irregularly voiced segments do not have any effect, as prosodic modifications of these segments are generally avoided.

Chapter 5

Speech Synthesis with an Extension of TD-PSOLA

This chapter explains how the signal information gathered with the methods presented in Chapters 2 to 4 is applied for prosodic modification. First, we give an overview on the standard TD-PSOLA approach. Then we explain how this comprehensive signal information is applied in the prosodic modification of mixed segments and irregularly glottalized segments and particular phones like plosives and affricates. Finally, we discuss the quality and the scope of the proposed method.

5.1 Introduction

TD-PSOLA is known to allow for high-quality pitch and time scale modification of speech segments for concatenative speech synthesis. However, the distinction of the signal into only two classes, voiced and unvoiced, does not give enough consideration to distinctive properties of a speech signal, like mixed excitation or irregular glottalisation or properties of plosives and affricates. Therefore, such signal parts are treated in a way that is not optimal, which means audible artifacts may result or even the intelligibility of phonemes such as plosives may be reduced. We therefore suggest an extended TD-PSOLA approach that not only distinguishes voiced and unvoiced segments, but is based on the detailed frame classification introduced in Chapter 3, which distinguishes voiced, unvoiced, mixed, irregular and silence frames. This detailed classification allows for a more appropriate modification of F_0 and duration to avoid artifacts that may otherwise arise. This treatment of mixed and irregularly glottalized frames is described in Section 5.4 and 5.5. Furthermore, minor enhancements in the treatment of diphone transitions and particular phones are added to the TD-PSOLA algorithm, which are described in Sections 5.6 to 5.9.

5.2 Re-Synthesis Experiment

In order to identify shortcomings of the signal handling in TD-PSOLA, we carried out a speech synthesis experiment that allowed us to identify artifacts introduced in the signal handling step and to exclude artifacts that may be introduced by either higher levels of synthesis or by the selection of possibly mismatching corpus elements. We used 10 sentences of natural speech from four voices (one female German voice: *fg*, one male Dutch voice: *md*, one male American English voice: *me* and one male Turkish voice: *mt*) and created a sentence diphone corpus from each sentence. Each sentence corpus contained exactly the diphones that made up the source sentence. We then re-synthesised these sentences using synthetic prosody and the sentence corpora.

These re-synthesised sentences were subsequently used to analyse artifacts that could have been introduced by either:

- noise, clicks or glitches in the source sentence
- errors in the F_0 extraction
- errors in the pitch marking
- errors in the frame classification
- errors in the diphone segmentation
- inappropriate synthetic prosody or strong prosodic modifications

- inappropriate signal handling in the synthesis procedure

Due to a mismatch between the voices that we used for the experiment and the prosody control of the SVOX system (see [Tra95]), that we used for prosody prediction, the application of the synthetic prosody caused rather strong prosodic modifications. As a consequence, the duration was increased sometimes by a factor of 3, often in combination with a strong increase of the F_0. The F_0 was also often increased or decreased by a factor of 2. In addition, the F_0 is very dynamic in the original recordings, but was predicted to be comparatively flat in the artificial prosody. Modifications to such an extent cannot be made unheard even by very good signal processing, so the synthesised signals sounded partly unnatural. However, these strong modifications were in fact an advantage in terms of analysis, as many problems get audible only with considerable prosodic modifications.

We analysed the signals and targeted errors stemming from inappropriate signal handling by iteratively extending the TD-PSOLA algorithm. For that purpose, we introduced the information retrieved from the signal analysis steps, especially the pitch marking and voicing classification, in the signal processing.

5.3 Outline of the Method

The proposed extension of the TD-PSOLA algorithm uses the standard TD-PSOLA approach on periods classified as voiced, unvoiced and silence whereas mixed and irregular periods are treated differently as explained in Sections 5.4 and Section 5.5.

In the standard TD-PSOLA approach, for voiced, unvoiced and silence segments, every two neighbouring period segments are multiplied with a Hanning window function. The F_0 of the resulting signal can be increased or decreased by moving the double period segments together or apart. The duration of the signal can be increased or decreased by either repeating or omitting double period segments. The algorithm distinguishes between voiced and unvoiced double period segments, because for unvoiced segments the time axis of every repeated double period segment has to be reversed. This prevents the algorithm from introducing artificial short-term correlations in the signal, which are perceived as audible buzziness.

Unlike the standard approach, we generally limit the number of period repetitions to 4. Experiments have shown that if a period is repeated more than 4 times, the negative segmental effect is more severe than the fact that the predicted duration is not entirely achieved. Although on the other hand, a limitation to 3 or less repetitions in many cases still leaves an almost perfect segmental impression, this limitation has a strong influence on prosody. Therefore, 4 repetitions seemed a reasonable compromise between prosody and segmental quality. This limitation on the number of period repetitions may not seem very elegant. However, in practise this limitation did not pose a problem for the kind of corpora we created as the diphones were created from phone instances with a duration well above the average. Therefore, excessive lengthening of these elements is not necessary at all, and even lengthening of more than a factor of two is required very rarely. Moreover, the duration of pauses and preplosive pauses, where the effect of this limitation would be most perceivable, is not affected by this limitation, as these pauses can be lengthened to an arbitrary duration by introducing zero samples.

5.4 Modification of Mixed Frames

5.4.1 Modification of Duration

Signal segments classified as mixed excitation often appear in voiced fricatives or transitions from unvoiced fricatives to voiced phones. These signal segments can only be shortened. A lengthening of the signal by repeating double period segments without time reversal of the repeated segment would lead to artificially introduced periodicity in the unvoiced frequency regions. Time reversal, however is not feasible as it would modify the periodicity of the voiced components in an unpredictable way (see [MC90]). Therefore, we refrain from lengthening signal frames that are classified as mixed. This is generally not a problem in synthesis, as most voiced fricatives consist of both mixed and unvoiced frames, where the latter can be repeated. Transitions from unvoiced fricatives to voiced stationary phones consist only of a few mixed frames, so this constraint does not cause any audible effect on prosody, above all, since the unvoiced fricatives and voiced stationary phones can normally be lengthened.

5.4.2 Modification of F_0

Need for F_0 Modification of Mixed Periods

One could argue that fundamental frequency modifications of frames classified as mixed may not be necessary, as the voiced components in the signal may be rather weak and the beneficial effect of modifying the fundamental frequency in this case may be marginal. However, mixed and voiced frames can alternate, for example if the voice quality is breathy. An example is shown in Fig. 5.1. In this case, the F_0 of the mixed frame has to be adapted to the same target F_0 as the voiced frames. Otherwise sudden jumps in F_0 with a strong disturbing effect occur as shown in the top plot of Fig. 5.1.

Compensation of Energy Variations

One problem with the F_0 modification of mixed frames is that energy variations are introduced, which are not extremely disturbing, but clearly audible (see Fig. 5.4). Therefore, the impact of the F_0 modification on the energy contour of the signal must be compensated for. More precisely, if the F_0 is decreased, the double periods are moved apart, which leads to a decrease of energy between the centres of these double periods. On the other hand, if the F_0 is increased, the double periods are moved closer, which leads to an increase in energy between the two double periods. A compensation of this effect can be achieved with a slight modification of the standard PSOLA algorithm.

With voiced frames and even more so with irregular frames, energy variation in time is high. The energy is concentrated at the glottal closure instants (GCIs, see Section 4.1). This is not very much the case with mixed frames, especially the high frequency energy components do not vary in time. If the signal is assumed to have constant energy

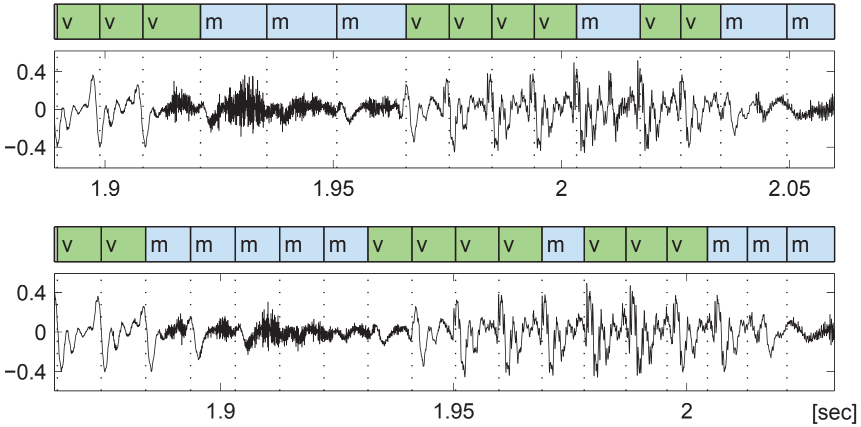

Figure 5.1: *Fricative components in the vowel [ɪ] (from about 1.97s to 2.05s) cause mixed and voiced frames to alternate. If the length of mixed periods is not modified, as shown in the top plot, sudden jumps in the the period lengths occur, causing the signal to sound disturbingly rough. The bottom signal shows the correct manipulation of the signal, where the lengths of the mixed periods are also modified.*

and the F_0 of the signal is modified, the resulting energy contour is not constant any more (see Fig. 5.2). With voiced signals a change of the energy contour is not a problem, because there is little energy around the middle between the pitch marks and therefore these manipulations do not have an audible effect. Although in case of increasing the F_0, the energy between the pitch marks is increased, the energy is still low in this area, thus normally no effect can be perceived for manipulations within a reasonable range.

With voiced fricatives, however, this variation of the energy contour becomes audible and causes a choppy impression. Therefore, mixed periods have to be treated in a special way, as shown in Fig. 5.3. First, the double period halves are taken from the original signal. These double period halves consist of a Hanning window half, which has the length of the original period (from time t_1 to t_2 in the top plot of Fig. 5.3) plus a part of the original signal of length $s/2$, where s is the difference

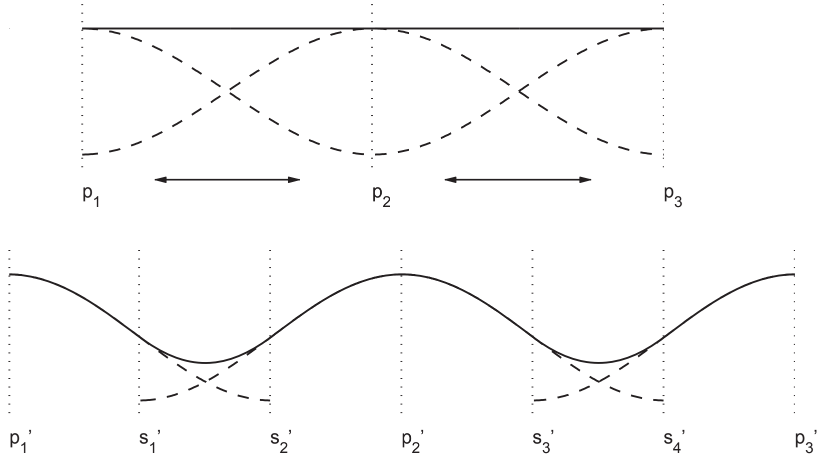

Figure 5.2: *Effect on the energy contour if the F_0 of the constant energy signal is decreased with the standard PSOLA method. On top as a solid line, the constant energy contour of the original signal is plotted. In the bottom plot, the F_0 of the signal is decreased by a factor of 0.5. The resulting energy contour shows "dents“ between the new pitch mark locations, denoted as p'_1, p'_2 and p'_3 in the modified signal.*

between the original period length and the modified period length. In this part no windowing is applied (from time t_2 to p_2 in the top plot of Fig. 5.3). Because the length of these double periods is more than two periods of the original signal, overlapping parts of the signal are used. To create the F_0-modified signal, these double periods are concatenated with an overlap of the original period length. As shown in the bottom plot of Fig. 5.3, the energy contour of the sum of the double periods corresponds to the original constant energy contour.

As an example, a speech signal where the correct F_0 modification of mixed frames is crucial is shown in Fig. 5.4. The signal contains the transition from the unvoiced fricative [f] to a vowel. In the top plot, the F_0 of the mixed frames was modified without compensating for the energy variability caused by this modification. In the bottom plot, the energy variability was compensated using the technique described above. The segmental quality of the pitch modified fricative in the

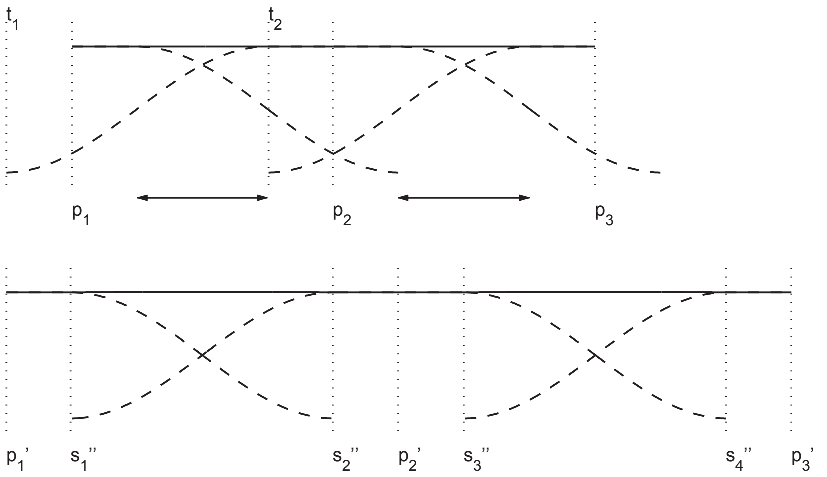

Figure 5.3: *Effect on the energy contour if the F_0 of the constant energy signal is decreased with the energy compensation taken into account. The resulting energy contour in the bottom plot is constant.*

bottom plot is as good as if fixed period lengths were used.

5.5 Modification of Irregular Frames

5.5.1 Modification of Duration

In a different context than the original one, irregularly glottalized segments often sound disturbing. Therefore, phone instances with irregularly glottalized frames should be avoided as much as possible in the creation of speech corpora. However, for some phones this is almost impossible, for example for glottal closures and often also for phones in the neighbourhood of glottal closures. Apart from avoiding irregularly glottalized segments in the first place, the best strategy to cope with these segments is to clearly identify them and modify them as little as possible. Glottal closures for example often contain strong energy spikes. If these energy spikes are repeated or if timing or context of these energy spikes is changed, for example by omission of periods, these energy spikes may suddenly be perceived as disturbing. However, if they are kept unchanged, they are mostly perceived as appropriate. Therefore, irregularly glottalized periods and their adjacent periods are neither repeated nor omitted but put exactly once.

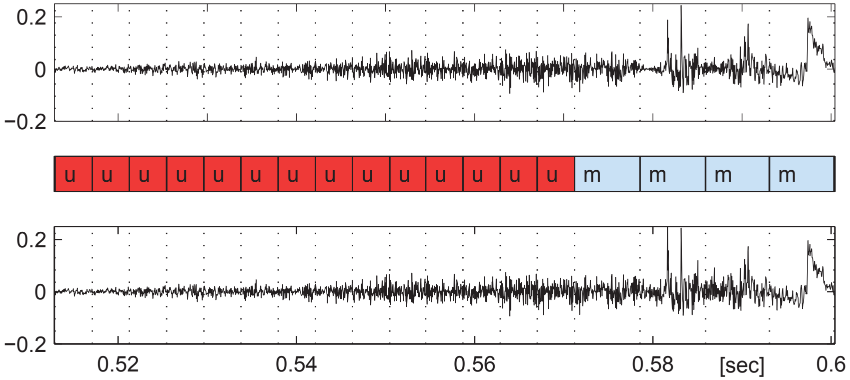

Figure 5.4: *Mixed frames in the transition from an unvoiced fricative [f] to a vowel. The dotted lines show the pitch marks, the letters u, and m in the middle plot denominate unvoiced and mixed frames, respectively. In top signal the F_0 of the mixed frames was modified without compensating for the energy variability caused by this modification, in the bottom signal the energy variability was compensated for.*

5.5.2 Modification of F_0

For irregularly glottalized segments, the fundamental frequency is not defined, in the sense that the signals are not shift-invariant. Furthermore, there is no clear impression of a particular fundamental frequency when listening to irregularly glottalized segments[1]. Therefore, a pitch

[1] Irregularly voiced segments frequently appear before pauses towards sentence or phrase boundaries. In these positions, the F_0 normally decreases, and if it decreases strongly, the signal often becomes irregularly glottalized. So in this case, these irregularly glottalized segments leave an impression of a very low, unspecific F_0.

scale modification of these segments is not necessary. On the other hand, a pitch scale modification like in the standard TD-PSOLA algorithm is also not advisable, as periods vary strongly in these segments and thus double period segments can be very asymmetric. In this case, the glottal pulse may be attenuated irregularly by the multiplication of the Hanning window with the asymmetric double period segment.

5.6 Modification of Plosives and Affricates

5.6.1 Modification of Duration

The repetition or omission of periods in the burst phase of plosives or affricates can have very strong effects on the quality of this phone and can even entirely change its nature, such that for example a [p] may be converted to a [t]. Consequently, the duration of the burst phase of plosives and affricates should not be modified at. This restriction applies to both voiced and unvoiced plosives. As it is hard to exactly determine the burst phase of a plosive or affricate, we estimated the burst phase as the first 30 ms after the release point. We also found that plosives should not be excessively shortened (by more than 40 % of their total duration), as this may impair their quality as well.

5.6.2 Modification of F_0

In terms of F_0, we found that it is beneficial to allow the modification of the period lengths to avoid sudden jumps in F_0 at the transition to neighbouring voiced phones. This applies mainly to voiced plosives, as for most unvoiced plosives and affricates, the frames are classified as unvoiced, and therefore period lengths are not modified. Nevertheless, we encountered quite a few examples, where unvoiced plosives were actually realised as voiced. This was especially the case for American English, where the intervocalic [t] is transcribed as an unvoiced [t], but is actually pronounced as a voiced flap (see [LM96], an example is shown in Fig. 5.5).

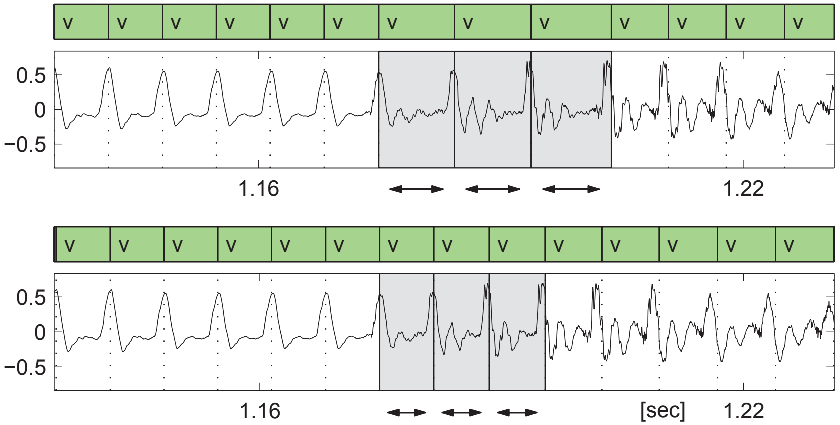

Figure 5.5: *Synthesised speech signals of the phone sequence [nt^{h}ɪ] where the [t^h] (shaded in grey) is realised as a voiced flap with all periods voiced. In the top plot, the period lengths of the plosive are not modified, which causes a sudden, clearly audible drop in F_0 around the [t^h]. In the bottom plot, the period lengths of the plosive are modified, thus avoiding F_0 jumps.*

5.7 Distinction of [ɦ]/[h] Variants

The characteristics of the glottal fricative [h] are highly influenced by its context. It may be articulated either as a voiced [ɦ] if preceded by a voiced phone, or as an unvoiced [h] if preceded by an unvoiced phone or by a pause. Furthermore, the differences in energy of its voiced and unvoiced articulations are quite high. This property was observed in particular for the speaker *mg*, which was not included in the resynthesis experiment (see Section 5.2), but in the later chapters on phone quality (see Section 6.4.3). As a consequence, combining the the first half of an unvoiced [h] with the second half of a voiced [ɦ] often creates a plosive impression due to a strong energy increase on the one hand and a sudden voicing onset on the other hand. Therefore, we distinguish voiced and unvoiced variants of the phone [ɦ]/[h] and use the appropriate variant in synthesis.

5.8 Prosodic Modification After Pauses

We noticed that for certain phones, apart from plosives and affricates, the repetition and omission of periods is very delicate in sentence-initial or phrase-initial position, in other words if those phones follow a pause. If, on the one hand, the beginning of a sentence-initial [l] is lengthened, often the impression of a schwa phone [ə] is introduced. On the other hand, if periods of a sentence-initial [l] are omitted, a plosive effect may be caused. Similar effects have been observed for the phones [h], [f], [s] and [ʃ], if they directly follow a pause. Therefore, we refrain from modifying the duration at the start of those phones if they follow a pause, which means we do neither repeat nor omit periods in the first 25 ms. Periods lengths may be manipulated, however.

One other technical detail that has to be considered with phones following a pause or a preplosive pause is that an unintentional click may be produced from period repetition. If the double period segment that combines the last period of the pause and the first period of the following phone is repeated and reversed because it is classified as unvoiced, then a click may be generated from the windowing and the signal in the beginning of the phone. This click becomes audible if the energy at the beginning of the following phone is high enough. This would, for example, be the case for most plosives, if the period repetition in the burst phase was allowed. But also for other phones this effect can occur (see Fig 5.6). To solve this problem, the first pitch mark of a phone following a pause must not be repeated for any phone.

5.9 Power Smoothing at Concatenation Points

A sudden increase in energy at the concatenation point of stationary phones often causes disturbances and even plosives to be perceived. Most problematic are transitions within fricatives, e.g. [h], [f] or [s], where plosives may be perceived due to sudden energy increases (see Fig. 5.7). To avoid these effects, a linear energy smoothing is applied for the concatenation of stationary phones. The smoothing is applied within a very short period from 15 ms before to 15 ms after the concatenation point with some limitation on the gain factor that is applied on the signal. Furthermore, it has to be ensured that a smoothing period cannot overlap with a previous or subsequent smoothing period.

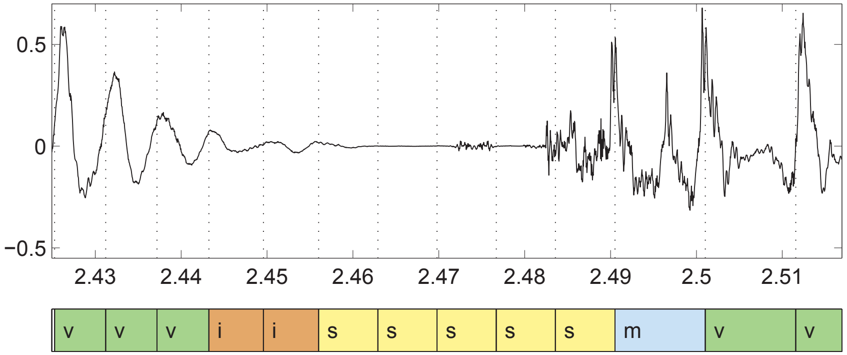

Figure 5.6: *Synthesised segment of the phones [əɛ] from a male Dutch speaker (md). Between 2.47 and 2.48 s an unintentionally produced click is visible that was produced through period repetition and time reversal.*

5.10 Discussion

In this chapter, the extended TD-PSOLA approach was presented. It considers not only voiced and unvoiced but also mixed, irregular and silence periods and, furthermore, uses phonological information. The most important findings were the careful treatment of irregularly glottalized segments and the abdication of duration modifications of the burst phase. These enhancements of the TD-PSOLA algorithm had clearly the largest effect on speech quality, which may be due to the prominence of plosives and irregularly glottalized segments, which in general have high intensity.

In related work we did not encounter any specific treatment of mixed or irregularly voiced segments in TD-PSOLA. Some authors mention specific pitch marking approaches for mixed excitation segments (see [MVV06]) but do not propose any special treatment in the

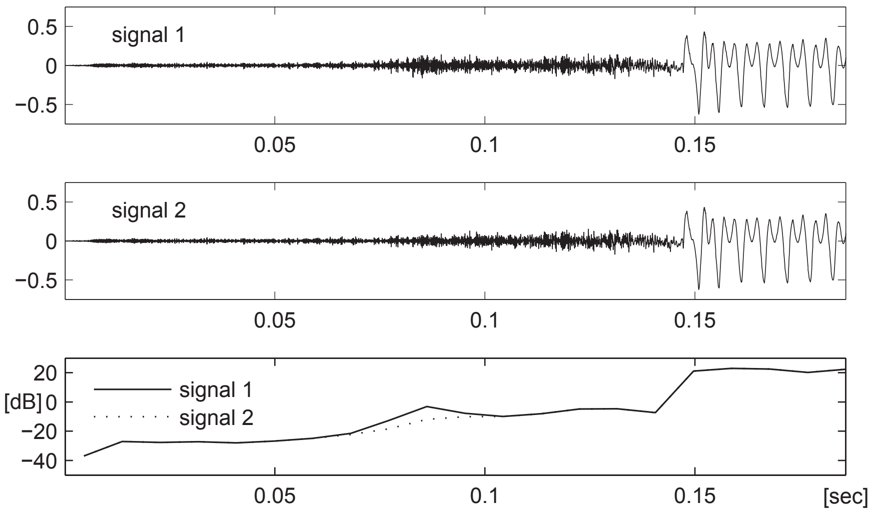

Figure 5.7: *Synthesised speech signals of an [s] followed by an [ɔː], spoken by a female English speaker (fe). In the top plot, no power smoothing is applied at the concatenation point. An energy bump can be seen at 0.08s. Although this energy bump may not seem very pronounced, listening to this signal gives the clear impression of a [t͡s]. The middle plot contains the signal with power smoothing applied, which gives the unambiguous impression of a [s]. The bottom plot shows the energy contours of both signals, which differ only where smoothing is applied.*

synthesis step. There are variants of TD-PSOLA which use modified approaches of the overlap-add principle. MBR-PSOLA (see [DL93]), for example, circumvents the pitch marking problem by coding the speech database using a Multiband Excitation (MBE) model and then re-synthesising a speech diphone database. In the synthesis step, a simple overlap-add algorithm is then used on this database which does not have to distinguish any voicing classes. Speech quality, however, is considerably degraded through this pre-processing step, resulting in a metallic sound or buzziness in voiced segments [EHM$^+$99].

In principle, the enhancements presented in this chapter enabled us to perform prosodic modifications within a certain range without any artifacts to be perceived. In an exploratory experiment we abstracted the prosodic information (one duration value and five F_0 values per phone) from natural sentences and re-synthesised these sentences using this information on the diphone elements taken from these sentences. The resulting synthetic sentences were indistinguishable from the original ones. In addition to this informal exploratory experiment, a more extended evaluation to compare results achieved with the standard TD-PSOLA method with those achieved with our extended TD-PSOLA approach in a formal listening test would be interesting.

The findings of this study are partly restricted to diphone synthesis as many problematic issues do not come about if larger units are concatenated. In unit selection, for example, problematic segments in principle remain within their original context where they are perceived as natural. However, if prosodic modifications are applied, which is the case for many high-quality unit selection systems, then accurate signal analysis and proper signal treatment is indispensable.

Chapter 6

Phone Quality Aspects

This chapter outlines different characteristics that contribute to the quality of a phone. First, we report on an exploratory listening experiment, which we conducted to obtain data on phone quality. After that, we detail on the measurement of various phone characteristics, including characteristics related to spectrum, phase, fundamental frequency, duration, voicing. Furthermore, signal intensity as a quality aspect at selection time will be investigated. A digression will follow about a method to monitor loudness already at recording time. Finally, we investigate characteristics of particular phones like plosives and fricatives and will lay out a few methods on the treatment of some peculiarities, which can arise with these phones.

6.1 Introduction

In concatenative speech synthesis, corpus generation still involves tedious manual or semi-automatic selection of units. In *diphone synthesis*, the segments for a diphone set are selected manually from a speech database with typically up to 100,000 phones or even more to choose from, if they are not extracted from designated diphone carrier words with one or two diphones embedded in one carrier word (see [LB00]).

In the same way, creating a *unit selection* voice involves manual work in the post-processing of speech recordings to identify low quality segments, labelling errors or pronunciation variants.

This high demand of manual effort is required because there exists no quality measure that could help to decide which phone segments are appropriate to be selected. Automatic phone quality judgement for corpus creation was only considered to a small extent so far. In [TH99], the best diphone variant is selected using the cepstral distance between the two semi-diphones and the corresponding phone centroids as the only automatic measure. Unit selection does not directly consider the quality of the selected units in their target costs during synthesis, because no acoustic properties for the target units are known. Phone quality is considered only indirectly through the concatenation costs, which only take into account spectral discontinuities (see [CB96]). Various measures to detect these spectral discontinuities were proposed in [SS01] and [Don01]. A detailed review of these spectral distance measures is presented in Section 6.4.1.

However, phone quality, what concatenative speech synthesis is concerned, not only depends on spectral quality, but has several, partly orthogonal aspects. In the following, we present these aspects and present features that we determined to describe and quantify these aspects. These features can finally be combined to constitute a phone quality measure that can be used to automatically select diphones from a speech database. For each diphone the following criteria are important:

1. The two involved phones must be heard as clearly articulated and unambiguously identifiable instances of these phones.
2. The signal of the phones has to be suitable for prosodic modification (for example with TD-PSOLA) without impairing the perceived speech quality.
3. No audible artifacts may occur at the concatenation points if the chosen diphones are concatenated.

In Chapters 7 and 8 we will apply this phone quality measure to automatically select diphones from a given speech database. Our phone quality measure can not only be used for diphone selection but for concatenation synthesis in general. We applied it here in the context of diphone synthesis because the high number of concatenation points immediately points to possible weaknesses of the method.

6.2 Exploratory Listening Experiment

6.2.1 Method

In the beginning of our phone quality research, a first, rather naive experiment was conducted to find a measure for phone quality. The experiment was arranged as follows: subjects were prompted with signals of isolated vowels, which were cut out of German diphone carrier words. These diphone carrier words were originally recorded from a female German speaker to create a multi-lingual diphone corpus (see [TH99]). For each prompt, the subjects had to choose one of three levels of quality, which were defined as:

1. Low: not correct phone or not identifiable
2. Medium: phone is intelligible but not well articulated or contains artifacts
3. High: phone clearly articulated; no artifacts

No further instructions about our notion of phone quality was given in the experiment.

The subjects were able to listen to the prompts repeatedly and to navigate through all phones. Four subjects (German speakers, experienced listeners) evaluated between 73 and 89 phone instances, depending on the phone. We used prompts from only one voice, a female German speaker, and each subject evaluated up to 4 different phones.

6.2.2 Results

Originally, the objective of this experiment was to create a standard to evaluate different measures for phone quality. However, the correlations between the different subjects were very low. For example, for the phone [ε], the inter-subject correlations were: 0.48, 0.32, 0.04 and 0.37. For the rest of the phones, the inter-subject correlations were even lower.

We traced the apparently large differences in the subjects' notions of phone quality to the very open definition of phone quality that was given for the test. As a consequence of this very open definition, there was no common understanding of the subjects on how to weight the different aspects of phone quality. Some of the listeners considered irregularly glottalized voice segments as unsuitable, some considered clicks as unsuitable, some considered pressed voice as unsuitable and some considered phones with abnormal F_0 contours as unsuitable. Furthermore, the voice that was used for the test was perceived as rather unpleasant by the listeners, as phones were often articulated in a rather pressed and strained way. It seemed to us that this aspect also influenced each of the subjects' classifications to a different extent.

As a result, the material that resulted from this listening test combined many aspects of phone quality in the subjects' classifications, so it was very hard to derive consistent information about single one aspect from that material. Consequently, we implemented an analysis tool to better investigate and isolate the influence of various aspects. This tool is described in the following section.

6.3 Identification of Phone Quality Aspects

As mentioned above, several, partly orthogonal aspects contribute to the overall subjective impression of the quality of a phone. To identify these aspects, we implemented an interactive graphical tool, which allows diphone instances to be selected from a ranked list and to be used in synthesis. Each diphone can be played in different contexts to subjectively judge its quality, not absolutely but only with regard to the rank order. In this way, we were able to identify aspects that strongly influence synthesis quality and therefore had to be integrated into our phone quality measure. These aspects are now described in detail in the following sections.

6.4 Spectral Characteristics

6.4.1 Introduction

In the following, we denote a concrete realisation of a phoneme as a *phone instance*. A phone instance is a concrete realisation of a particular phoneme. This phone instance normally differs from another phone instance of the same phone, as it may be influenced not only by the phonetic context, but also lexical stress, speech rate, mood of the speaker, etc.

Spectral properties play a crucial role in assessing the quality of phones, in particular of stationary ones. In the context of diphone synthesis we are not only interested in excluding phone instances that are not unambiguously identifiable as a particular phone but also instances that are not typical for this phone as pronounced by a certain speaker. We score the spectral appropriateness of a particular phone instance using the distance between the spectral description of this phone instance and its corresponding centroid.

Centroids were first used in [Kae85] to characterise typical spectral properties of phones. The spectral description of a phone is generally given as a sequence of spectral feature vectors which move over time in its corresponding feature space. This movement describes a trajectory, whose course is determined by the given phone sequence but also by the speaker and by random variability, because humans cannot reproduce speech perfectly. For different phone instances, these trajectories concentrate in an area of the feature space which is characteristic for that phone. These trajectories, with their origins and destinations depending on the neighbouring phones, typically do not intersect. But still, this area of trajectory concentration describes the typical spectral properties of the phone, the centroid, which is defined as follows:

> The centroid vector is the vector with the smallest mean distance to the minimum distance vectors of every trajectory.

As minimum distance frames, a weighted sequence of frames with the smallest distance to the centroid was used. For details on the iterative computation of the centroid, see Section 6.4.4.

In [Kae85], the log area ratio distance measure was used to compute the centroid distances. This measure was used because first, results from speech coding suggested the log area ratio as a good measure to predict distortion (see [QB82]) and second, it was efficiently computed. However, since then, a lot of research has been conducted on distance measures. Early works by [GM76] investigated spectral and cepstral distances and showed that the cepstral distance is a computationally efficient estimate of the log spectral distance of cepstrally smoothed spectra. One of the first studies to investigate the application of distance measures or, more precisely, distortion measures on speech coding was [GBGM80]. Whereas the first studies focused predominantly on the area of speech coding, later studies investigated the application of distance measures on speech recognition. In 1985, [NSRK85] investigated the effects of various distortion measures on the performance of a standard dynamic time warping based isolated word recogniser.

Finally, and most interesting for us, distance measures were studied in the area of speech synthesis to objectively measure spectral discontinuity in concatenative speech synthesis, for example to determine join costs in unit selection systems. Although we do not directly measure the spectral discontinuity at concatenation points, these studies are most closely related to our problem, as these measures describe the human perception of allophonic differences. Coding distortions, in contrast, may have a different effect on speech than allophonic variations and measures used for speech recognitions may not capture some of these allophonic variations as they are not relevant for recognition.

According to [KV98], a symmetrical Kullback-Leibler distance measure on LPC power-normalised spectra is suited best to detect concatenation discontinuities in vowels. A study on join costs for unit selection investigates Euclidean, SKL, Mahalanobis, and absolute distance on three different features (MFCC, line spectral frequencies and MCA coefficients) to identify discontinuities in diphthongs ([VKT02]). None of the combinations really outperformed the other. A comprehensive study was conducted in [SS01], where 13 combinations of spectral descriptions and distance measures were investigated, also to identify discontinuities in vowels. They found that the Kullback-Leibler distance on power spectra has the highest detection rate followed by the Euclidean distance on MFCC. Studies in [WM98] confirmed that the Euclidean distance on MFCC is a good predictor for perceptual discontinuities in vowels.

The limitation of all these studies for our purposes is that they restrict their measures to vowels and diphthongs in one case, and, the studies are based largely on English speech data. Although the results of these studies are not entirely consistent, some trend can be deduced. This trend allowed us a preselection of the measures that we investigated further in a listening experiment and a classification task to identify the measure that is most suitable to characterise spectral phone quality in terms of a centroid distance.

6.4.2 The Classification Task

We formulated a classification task to determine the most appropriate method to discriminate suitable from unsuitable phone instances in terms of spectral properties. Isolated phones were manually classified into two classes, *suitable* and *unsuitable*. These data were then used to identify the measure that discriminates the two classes best.

6.4.3 The Data

The data for that classification task were generated in a similar way as for the exploratory experiment described in Section 6.2. However this time, all the shortcomings of that previous experiment were taken into consideration. Various voices were used to mitigate effects caused by peculiarities of a particular voice, the concept of phone quality was strictly defined, the author, who knew the characteristics of the voices very well, manually classified the data.

The four voices we used were a German male and German female, and an English male and English female voice (*fg*, *fe*, *mg* and *me*). For each voice, 100 instances of all stationary phones (including some context) were cut out of natural speech. These phone instances were not prosodically manipulated. Unlike in the first experiment, problematic phone instances, where different aspects than spectral properties strongly influenced phone quality, were excluded from the classification. Examples for that were irregularly glottalized phone instances, instances with very low F_0 (which tend to be irregularly glottalized and often sound strange if perceived out of context), instances with extremely high F_0 (which also often sound annoying if perceived out of context). Phone instances that were too short for a reasonable classification were also excluded. Furthermore, the author had a very good knowledge of the voices to decide which phone instances were typical for the particular speaker and which not. For the manual classification a graphic selection tool was implemented as a Wavesurfer Plug-in (see [SB00]). In total, a number of 11547 phone instances were manually classified and double checked with the classification results.

The phone instances were classified into one of the categories *suitable* or *unsuitable* phone instances. The category *unsuitable* contained phone instances that may still be the same indicated phone, but which differ from the optimal phone in terms of articulation, which means they may be articulated too open or too closed, a strong influence of phonetic context may be present, inappropriate nasalisation, etc.

A number of phones was manually classified but not used in the subsequent classification task:

- The fricatives ([f], [ʃ], [s], [z], [ʒ], [ç]) were excluded, because the way they were classified could not be reproduced by considering only spectral properties. In manual classification, fricatives that contained disturbing noise (like sibling noise, clicks etc) were classified as unsuitable. However, those phones can still have a small distance from the centroid as long as a part of the phone that is not affected by the noise is long enough.
- Syllabic phones [n̩], [l̩] were excluded as we found them not to be so stationary that an unambiguous centroid could be determined.

Difficulties with British and American English Transcriptions

For the American English *me* voice only a British English transcription was available to us. In this transcription, words like “job” were noted as [d͜ʒɑb], although they were pronounced as [d͜ʒaːb]. However, in American English the phone [aː] replaces the phone [ɑ] in many words (see [Jon03]). Furthermore, spoken variants were present in American English recordings: for example the word *long* can be either pronounced as [lɔːŋ] or [laːŋ]. Both variants are possible according to [Jon03], but it was not feasible to manually correct the transcriptions as to which variant was actually used[1]. Therefore, the phone instances transcribed as [ɔː] contained instances which were pronounced very open and were actually instances of the phone [aː]. Thus, careful listening was required in judging those phones.

On the other hand, the British English voice was partly transcribed with American English variants. E.g. the word *forget* was transcribed with the American English form [fr̩ˈget] instead of [fəˈget], as it would be correct in British English. Again, it was not possible to manually correct all these transcription inconsistencies, however we tried to determine and exclude phone instances that were not labelled correctly.

6.4.4 Features and Distance Measures

Computation of the Centroid

The centroid as it is presented in Section 6.4.1 is computed iteratively: First, an initial centroid is computed by averaging the vectors of all frames of all phones that represent a particular phoneme. Then, for each phone those five consecutive frames are determined whose added distances to the centroid are minimal. These five frames, which cover a total length of 60 ms, are denoted *minimum distance frames*. The new centroid is then computed by taking the average of the minimum-distance frames of all phones. This step is repeated until convergence occurs, i.e. the minimum-distance frames remain constant, or a maximum number of iterations is reached. These steps are applied for every stationary phoneme.

In earlier experiments we iteratively removed outliers, recomputed the centroid and investigate the effect on the centroid. However, no significant change in the centroid position was observed, which may be due to either the large percentage of suitable phones compared to

[1] The databases cover 1000 and 2800 sentences, each sentence with a duration of some 5 s.

outliers or possibly also to some symmetric distribution of the outliers around the centroid such that their deviations cancel out.

Features and Distance Metrics

We compared combinations of distance measures and spectral features that proved to be most suitable according to literature on concatenation discontinuities (see Section 6.4.1). As features we used MFCC, Straight cepstral coefficients, LPC cepstral coefficients, line spectral frequencies (lsf) and the log area ratios (lar). The distances we used on these features were the Euclidean distance, SKL distance, Mahalanobis distance and absolute distance. For three of the four voices, the signal were resampled from 22 kHz to 16 kHz sampling frequency, such that for all voices the same spectral information was available.

Mel Frequency Cepstral Coefficients

We used 12-dimensional cepstral vectors, whereby the zeroth cepstral coefficient was neglected. For the number of filters at 16 kHz sampling frequency we chose 32, which in the lower frequencies corresponds to the 24 filters which are normally used at 8 kHz.

Cepstral Coefficients Based on Straight Spectrum

We used 12-dimensional cepstral vectors derived from Mel-sampled interference-free Straight spectra presented in [KMT$^+$08], whereby the zeroth cepstral coefficient was neglected.

LPC Cepstral Coefficients

We used 12-dimensional LPC cepstral coefficients, whereby the zeroth cepstral coefficient was neglected. To compute the LPC coefficients, order 20 was used.

Line Spectral Frequencies

According to [VKT02], line spectral frequencies give results comparable to MFCC in terms of correlation between perceptual scores and objective scores so they were also included as features. Line spectral frequencies are computed from the LPC filter coefficients and represent LPC spectral information the same way the reflection coefficients or log area ratios do. For further details see [SJ84].

Log Area Ratio

The log area ratio distance was used in the first study on diphone selection based on centroid distances ([Kae85]). Apart from its good perceptual properties, in those days, computational efficiency had been an argument to use this measure. We included it in our evaluations for comparison reasons. The log area ratios can be computed from the reflection coefficients:

$$g_i = log \frac{1 + k_i}{1 - k_i}, \tag{6.1}$$

where g_i is the log area ratio and k_i the reflection coefficient i.

Euclidean Distance

The Euclidean distance between two feature vectors is defined as:

$$d_{eu}(X, Y) = \sqrt{\sum_{i=1}^{N} (X_i - Y_i)^2} \tag{6.2}$$

Symmetrical Kullback-Leibler Distance

According to [VK03], a symmetrical version of the Kullback-Leibler distance gives good results on the identification of concatenation discontinuities. However, this distance measure is not applied directly on the cepstral features but on the power-normalised spectral envelopes that are computed from the cepstral features. Let X and Y denote power-normalised spectral envelopes, i.e.

$$\sum_{n=1}^{N} X_n = 1; \sum_{n=1}^{N} Y_n = 1. \tag{6.3}$$

The Kullback-Leibler distance between these two spectra is defined as

$$d_{KL}(X, Y) = \sum_{n=1}^{N} X_n \log \frac{X_n}{Y_n}. \tag{6.4}$$

The symmetrical Kullback-Leibler distance between these two spectra is defined as

$$d_{SKL}(X, Y) = \sum_{n=1}^{N} (X_n - Y_n) \log \frac{X_n}{Y_n}. \tag{6.5}$$

The spectral envelopes that X and Y are based on, are computed from the according cepstral features. In other words, for this distance measure we reconstruct spectra that have some kind of smoothing applied. This smoothing depends on the cepstral feature that we use as an input for the reconstruction.

Mahalanobis Distance

The Mahalanobis distance weights features with the inverse of their standard deviation. Thus, features with low variance are boosted and have more influence on the total distance. We computed the standard deviations for the Mahalanobis distance for each phone context. The Mahalanobis distance between two feature vectors is defined as:

$$d_{ma}(X, Y) = \sum_{i=1}^{N} \frac{(X_i - Y_i)^2}{\sigma_i^2}, \tag{6.6}$$

where σ_i is the standard deviation of the i^{th} feature vector element.

Absolute Distance

The absolute distance between two points, also more colourfully known as Manhattan distance, is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes. This distance measure results in higher distances for points differing more equally in all dimensions compared to the Euclidean distance. We included the absolute distance as [VKT02] reported that the absolute distance on MFCC is comparably and partly even stronger correlated to perceptual scores than the Euclidean distance on MFCC. Formally, the absolute distance is defined as:

$$d_{abs}(X,Y) = \sum_{i=1}^{N} |X_i - Y_i| \tag{6.7}$$

6.4.5 Feature and Distance Measure Evaluation

We used the measures described above to classify the phones into the classes suitable and unsuitable according to their distance from the centroid. A typical distribution of suitable and unsuitable phone instances is illustrated in Fig. 6.1. The best classification results are achieved by minimising the total number of misclassification based on the a posteriori probabilities $p(C_k|\mathbf{x})$. In this case, error rates of about 9 - 12 % for the best measure can be achieved. Detailed results for all voices can be found in Appendix A in Tables A.1 to A.4. However, these low error rates are partly due to the fact that the two classes of suitable and unsuitable phones are very unequally distributed with a lot more suitable phone instances than unsuitable ones. This leads to a high a priori probability for the suitable phones. Consequently, to minimise the number of misclassification, the best results are obtained by setting the decision boundary to a very high value and, thus, classifying all or nearly all phone instances as suitable.

However, for a phone quality measure applied in the selection of diphone elements from speech databases, it is not beneficial to find the globally best classification but the best classification for a single phone, independently from the a priori probabilities of the classes. Two types of mistakes can be made, with very different consequences. False positives, that means unsuitable phones that were classified as suitable have more serious consequences than false negatives, which are suitable phones that were classified as unsuitable. False negatives can have the effect that possibly suitable phones are ignored, whereas false positives can have the effect that unsuitable phones are taken into account as corpus candidates. In general, it is thus better to ignore a potentially suitable phone than to favour an unsuitable phone. Therefore, we introduced a cost ratio r to set the decision boundary at a value such that the number of false negatives is r times the number of false positives. In other words, the decision boundary is moved such that more suitable phones are classified as unsuitable and less unsuitable phones are classified as suitable. This is achieved by considering the two probability distributions: $P(d_u < d)$, which is the probability that the centroid distance of a unsuitable phone instance is smaller than the value d, and $P(d_s > d)$, which is the probability that the centroid distance of a suitable phone instance is larger than the value d. An example of the two probability distributions for the phone [o] of a female German speaker is shown in Fig. 6.2.

Equal error probability, that means an equal number of false positives and false negatives, is achieved if the decision boundary is set to the distance value of the point where the two curves intersect, depicted as a vertical line in Fig. 6.2. To achieve a number of false negatives that is r times the number of false positives, the decision boundary has to be moved to the left. Obviously, with increased r, on the one hand, the overall error rate increases, as can be seen in Fig. 6.3, but, on the other hand, the number of false positives decreases. In detail, this effect can be observed in the classification results in Tables 6.1 to 6.7, where a cost ratio of $r = 10$ is applied.

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	20.38 %	31.01 %	21.05 %	19.30 %
straight	26.22 %	39.10 %	25.12 %	26.40 %
lsf	26.18 %	26.55 %	25.71 %	26.63 %
lpc	24.07 %	28.15 %	27.23 %	25.17 %
lar	28.68 %	28.64 %	30.42 %	29.23 %

Table 6.1: *Mean error rates for all phone instances for the female German voice (fg),* $r = 10$

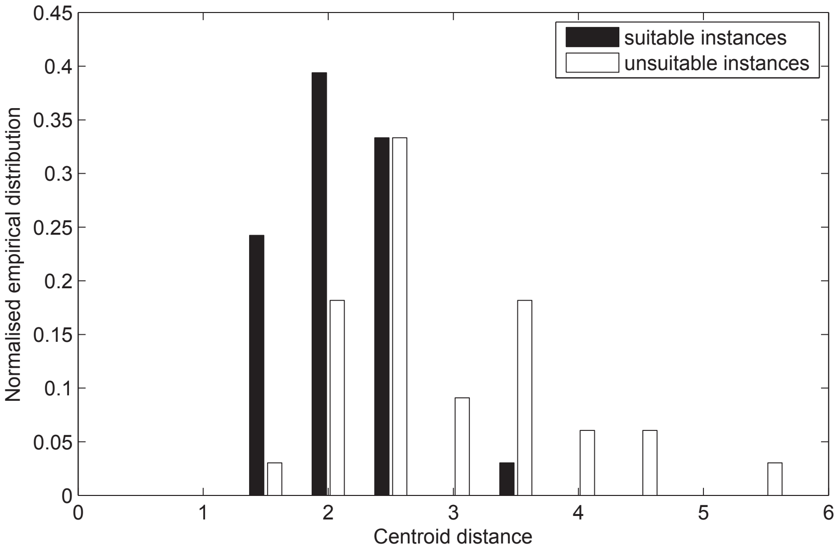

Figure 6.1: *Empirical distribution of the centroid distances of suitable and unsuitable phone instances of phone [o] for the female German fg voice. Euclidean distance was used on MFCC to compute the centroid distances.*

The overall results for the four voices can be seen in Tables 6.9 and 6.10 with MFCC in combination with the Euclidean distance yielding the lowest error rate and the lowest false acceptance rate. A different number of phones was evaluated for the 4 voices, so the overall result is not just the average of the 4 individual results. In fact, the overall result is computed by taking the average of the error rate or false acceptance rate of all phones from all voices.

6.4.6 False Positive Errors

Because the error rates for the best measure is still high, an error analysis was performed for the combination of MFCC and the Euclidean distance. For every voice, we considered the false positives that still remain and analysed the reason for their manual classification as unsuitable to conclude on the effects if these phones were used in a corpus. We found that these false positives were often phones that were articulated too open or too closed or had some inappropriate colouring; a few times they were really completely different phones. The detailed results of this analysis are listed in Appendix A in Tables A.5 to A.8.

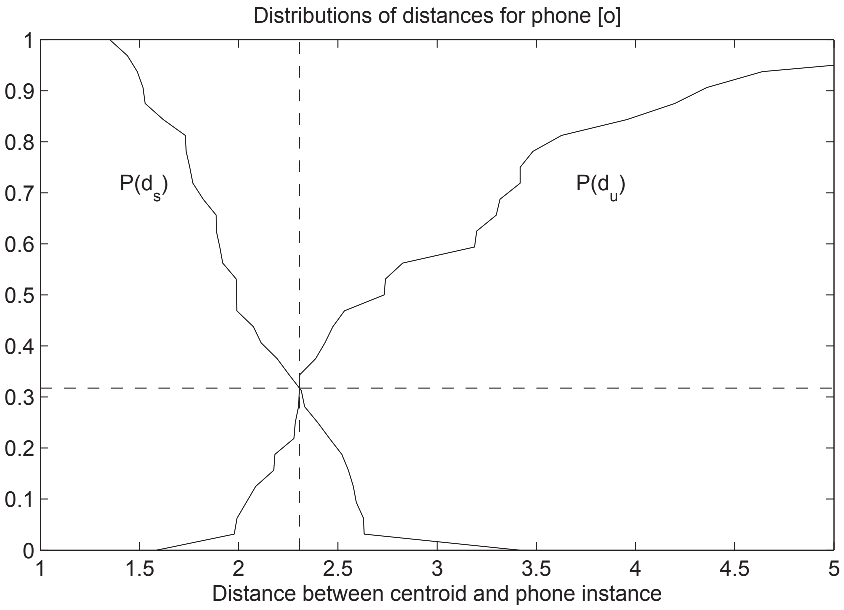

Figure 6.2: *Probability distributions $P(d_s > d)$ and $P(d_u < d)$ for phone [o] of the female German fg voice. The Euclidean distance measure was used on MFCC, where distances for suitable and unsuitable phones are denominated as d_s and d_u, resp. The horizontal dashed line shows the equal error probability, the vertical dashed line the equal error distance.*

6.4.7 False Positives in Corpus Generation

In the previous section, false positives that are actually problematic in terms of spectral quality (those contained in the last four rows of Ta-

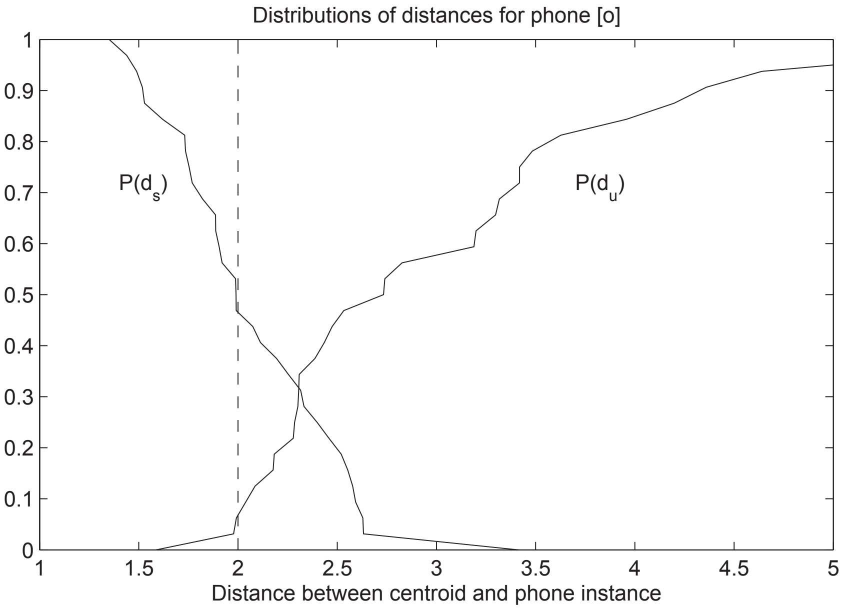

Figure 6.3: *Two probability distributions $P(d_s > d)$ and $P(d_u < d)$ for phone [o] of the female German fg voice. The cost ratio r was set to 10. MFCC and Euclidean distance measure was used.*

bles A.5 to A.8 in Appendix A) have been identified. Next, we checked how many of these false positives are finally selected in a diphone corpus. To that aim, we have created three corpora from the voices *fe*, *mg* and *me* using the method described in Chapter 7. For every corpus we determined for every problematic phone if it was part of a selected diphone and analysed the available alternative candidates to establish why such a phone was chosen. Below, the detailed analyses for the thee corpora are described. From these analyses it can be concluded that if one of the false positives was finally selected in a diphone corpus, it was selected due to a lack of more suitable candidates.

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	2.30 %	3.17 %	2.30 %	2.19 %
straight	2.88 %	3.95 %	2.62 %	2.78 %
lsf	2.79 %	2.84 %	2.84 %	2.79 %
lpc	2.46 %	2.90 %	2.84 %	2.68 %
lar	3.01 %	3.01 %	2.90 %	3.06 %

Table 6.2: *False acceptance rates for all phone instances for the female German voice (fg), $r = 10$*

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	19.58 %	38.36 %	21.27 %	19.55 %
straight	29.57 %	41.88 %	27.44 %	26.15 %
lsf	32.78 %	32.09 %	30.62 %	30.40 %
lpc	30.45 %	24.12 %	29.86 %	31.48 %
lar	33.23 %	32.78 %	33.55 %	32.40 %

Table 6.3: *Mean error rates for all phone instances for the female English voice (fe), $r = 10$*

fe Voice

Out of the 2021 diphones that were selected for the *fe* voice corpus, two diphones contained phone instances that had been manually classified as unsuitable:

The first one, an instance of the phone [ɔː], is part of the diphone [gɔː]. We analysed the alternative candidates for this phone pair to investigate the reason why this unsuitable phone instance was chosen: In total there were 7 phone pairs [gɔː] to choose from. In 5 of these pairs, the burst of the [g] was determined to be not distinctive enough as it was below a certain energy threshold (see Section 7.1.6). Non-existence of the burst leads to a high penalty for the diphone candidate, therefore none of these phone pairs was selected. From the remaining two candidates, one was very irregularly glottalized (9 of 30 frames classified as irregular for the phone [ɔː] and had a higher distance from the centroid than the one that was finally selected. It can be concluded that the candidate that was manually classified as unsuitable was still the best choice for that particular diphone.

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	2.09 %	4.09 %	2.25 %	2.25 %
straight	3.13 %	4.20 %	2.90 %	2.82 %
lsf	3.37 %	3.29 %	3.29 %	3.13 %
lpc	3.05 %	2.49 %	3.21 %	3.29 %
lar	3.37 %	3.37 %	3.53 %	3.29 %

Table 6.4: *False acceptance rates for all phone instances for the female English voice (fe), $r = 10$*

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	15.34 %	27.06 %	16.91 %	17.04 %
straight	18.08 %	42.77 %	19.14 %	18.09 %
lsf	20.44 %	21.86 %	18.77 %	20.74 %
lpc	18.06 %	14.15 %	19.71 %	17.89 %
lar	24.31 %	27.45 %	26.89 %	24.87 %

Table 6.5: *Mean error rates for all phone instances for the male German voice (mg), $r = 10$*

The second phone instance that was manually classified as unsuitable and nevertheless selected was an instance of the phone [v] and was part of the diphone [zv]. In total there were 10 phone pairs [zv] to choose from. Looking at the [z] part we found that 3 of the candidates that were not selected had a maximum phase offset of 0.5, which results in a high penalty, and had only half or less of the frames classified as mixed, which leads to a high penalty for voiced fricatives. Considering the [v] part, 4 candidates had the maximum phase offset of 0.5, and for 8 out of 10 phones less than half of the frames were classified as mixed leading to a high penalty just like for the phone [z]. This leaves only one other candidate, which had a smaller centroid distance but a higher phase offset, and contained an irregularly glottalized frame, which is also penalised in voiced fricatives, as irregular frames often indicate noise. In addition, the duration of the [z] part of this candidate is only half that of the one that was selected. Consequently, also for this case it can be concluded that this candidate that was manually classified as unsuitable, was still a good choice for that particular diphone.

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	1.89 %	2.86 %	2.04 %	2.09 %
straight	2.12 %	4.48 %	2.17 %	2.12 %
lsf	2.38 %	2.43 %	2.04 %	2.47 %
lpc	2.09 %	1.70 %	2.23 %	2.04 %
lar	2.62 %	2.86 %	2.86 %	2.67 %

Table 6.6: *False acceptance rates for all phone instances for the male German voice (mg), $r = 10$*

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	22.22 %	30.82 %	22.16 %	23.09 %
straight	27.22 %	49.59 %	28.49 %	27.74 %
lsf	31.36 %	32.61 %	29.85 %	30.04 %
lpc	29.47 %	26.10 %	30.71 %	27.63 %
lar	32.36 %	33.51 %	35.74 %	32.22 %

Table 6.7: *Mean error rates for all phone instances for the male American English voice (me), $r = 10$*

mg Voice

Out of the 1187 diphones that were selected for the *mg* voice corpus, 5 contained phone instances that were manually classified as unsuitable, where for 1 phone pair ([a͜iɐ]) there was no alternative.

The selected phone pair [ir] contained the phone [r], which was manually classified as unsuitable. This classification can be attributed to inaccurate transcription as for all 4 candidates, the actual pronunciation is [iɐ̯] with the [r] actually sounding like an [ɐ]. All of the 4 candidates are pronounced within several recordings of the same word (”Schamir“), and the phone pair that was chosen was actually the best version in terms of both [i] and [r]/[ɐ̯]. The other candidates tended to sound more pressed, one of them showed slight nasalisation. The same transcription inaccuracy that led to a manual classification of the phone [r] as unsuitable applies for the pair [rs]. The phone pair was actually pronounced as [ɐ̯s] for all 7 candidates, (one candidate was taken from the phone sequence [tɛɐ̯s], the rest was taken from the phone sequence [bɛɐ̯s]).

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	2.41 %	2.90 %	2.41 %	2.55 %
straight	2.89 %	4.82 %	3.03 %	2.89 %
lsf	3.33 %	3.26 %	3.19 %	3.19 %
lpc	2.97 %	2.76 %	3.12 %	2.97 %
lar	3.19 %	3.40 %	3.47 %	3.19 %

Table 6.8: *False acceptance rates for all phone instances for the male American English voice (me), $r = 10$*

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	**19.09 %**	29.56 %	20.14 %	19.54 %
straight	24.62 %	42.99 %	24.47 %	24.10 %
lsf	26.84 %	27.50 %	25.44 %	26.30 %
lpc	24.71 %	22.77 %	26.52 %	24.91 %
lar	29.03 %	30.13 %	31.09 %	29.16 %

Table 6.9: *Mean error rates for all phone instances for all voices, $r = 10$. We can see that MFCC as features in combination with the Euclidean distance yields the lowest error rate.*

Another phone, which was manually classified as unsuitable and nevertheless selected, was an instance of the phone [r] as part of the diphone [ru]. This selection can be attributed to the fact that 6 out of a total of 7 phone pairs stem from the same word ("Februar“), where the [u] has a very strong tendency towards [ɔ]. This can be also seen in the centroid distances, which are considerably high with values around 10. The remaining candidate next to the selected one has very low intensity and a considerable amount of irregularly glottalized frames, which makes it also unsuitable for selection. In this case, the selected candidate was clearly the best choice.

The last phone that was manually classified as unsuitable and nevertheless selected was an instance of the phone [u] as part of the diphone [mu]. There is only one word in the recordings with that phone pair, which is the word "kommunistische“. In all 4 recordings of this word, the [u] has a tendency towards [ə]. Two of the candidates which were not selected are spoken very low with irregularly glottalized frames. From the two remaining candidates which are spoken with normal intensity,

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	**2.17 %**	3.07 %	**2.17 %**	**2.17 %**
straight	2.70 %	4.35 %	2.62 %	2.61 %
lsf	2.89 %	2.89 %	2.75 %	2.84 %
lpc	2.57 %	2.46 %	2.82 %	2.61 %
lar	2.99 %	3.11 %	3.13 %	3.01 %

Table 6.10: *False acceptance rates for all phone instances for all voices, $r = 10$.*

no difference can be perceived in quality.

me Voice

It resulted that none of the unsuitable phone instances that was classified as suitable was part of any of the selected diphones. This is due to the other criteria that influenced the final scores of these unsuitable phone instances so that they were finally excluded.

6.4.8 Centroids from Manually Selected Phones

One natural assumption is that precisely estimated centroids should improve the classification of the spectral quality of phones and therefore the quality of a corpus. To investigate that open question we used centroids based on manually selected phones to compute the centroid distances. As manually selected phones we used the phones that were manually classified as suitable as described in Section 6.4.3. We then used the centroids in the selection of diphone elements from the speech databases of these two voices and created test sentences from these corpora. We compared these test sentences with test sentences that were created from corpora that were based on automatically computed centroids based on all phones. We used the *me* voice because in previous experiments it had shown to be most problematic with respect to spectral mismatches. Thus, we assumed that a speech corpus from this voice would profit most from improved centroid quality. We also used the *fe* voice to include a female voice in this investigation.

Spectral Properties of Manually Selected Phones (*me* Voice)

In order to gain some insight into the spectral properties of the manually selected phone instances compared to all instances of a phone, we compared the mean distance μ'_d of the manually selected phone instances from the centroid to the mean distance μ_d of all instances of a phone from the centroid. The centroid is computed from all instances of a phone. The mean distances μ'_d and μ_d for each phoneme of the *me* voice are listed in Table 6.11. As expected, the mean centroid distance for each phone is lower for the manually selected phones than for all phones with a few exceptions of the phones [f], [ʃ], and [s]. The reason for that could be that only fricatives were classified as suitable phones that contained no disturbing noise (like sibling noise, clicks etc). However, phones that contain such noise can still have a low distance from the centroid as long as the correctly pronounced part of the phone is long enough. By excluding phones with disturbing noise, is seems that also phones with a low distance to the centroid were excluded. So it may be concluded that for the unvoiced fricatives the automatic selection method is more beneficial than the manual selection. Other exceptions where no or no large improvement was reached by using the manually selected phones were [n̩] and [l̩]. We conjecture that these phones were actually not so stationary that an unambiguous centroid could be determined. Sometimes these syllabic phones consisted in fact of a combination of two phones, a short [ə] and an [n] or [l].

Phone	μ'_d	μ_d
[ə]	1.77	2.33
[ɜ͜r]	1.67	1.75
[ɑ]	1.51	1.92
[ɛ]	1.68	1.83
[f]	1.86	1.91
[i]	1.25	1.76
[ɪ]	1.66	2.10
[l]	1.78	2.06
[l̩]	1.71	1.64
[m]	1.64	1.73
[n]	1.45	1.73
[n̩]	1.45	1.43
[ŋ]	1.55	1.60
[o]	1.43	1.66
[ɔ]	1.39	1.73
[r]	1.78	2.07
[ə͜r]	1.82	2.03
[s]	1.63	1.67
[ʃ]	1.74	1.64
[uː]	1.75	2.09
[ʊ]	1.73	2.34
[v]	1.73	2.09
[ʌ]	1.60	1.93
[w]	1.48	1.78
[z]	1.73	2.25
[ʒ]	1.48	1.83

Phone	μ'_d	μ_d
[ə]	1.68	2.22
[ɜ]	1.42	1.61
[ɑ]	1.93	2.55
[ɑː]	1.44	1.84
[e]	1.41	1.89
[f]	1.51	1.58
[iː]	1.63	1.93
[ɪ]	1.67	2.45
[l]	2.00	2.75
[m]	1.37	1.52
[n]	1.27	1.70
[ŋ]	1.31	1.63
[ɔː]	1.58	1.86
[æ]	1.58	2.27
[r]	1.93	2.26
[s]	1.67	1.81
[ʃ]	1.48	1.46
[uː]	1.59	2.18
[ʊ]	1.93	2.45
[v]	1.58	2.17
[ʌ]	1.62	2.11
[w]	1.66	2.41
[z]	1.73	2.29
[ʒ]	1.68	2.03

Table 6.11: *Comparison of the mean centroid distances μ'_d of the manually selected phones and the mean centroid distances μ_d of all instances of a phone, on the left for the me voice and on the right for the fe voice. Euclidean distance on MFCC was used.*

Evaluation of Corpora Based on Manually Selected Phone Instances (*me* voice)

We created a diphone corpus where the phone centroids that are normally computed from all phone instances were replaced by centroids computed from manually selected phone instances. For the corpus creation we used the approach described in Chapter 7. We synthesised two sets of test sentences using the two corpora based on the respective types of centroids.

We found for the two corpora that about 20-25 % (depending on the threshold values of the centroid penalty functions) of the diphones were selected differently when using centroids based on manually selected phone instances. However, the differences in the selected phone instances were almost not noticeable by listening to the two sets of test sentences.

Spectral Properties of Manually Selected Phones (*fe* Voice)

As for the *me* voice, for the *fe* voice we also compared the mean distance μ'_d of the manually selected phone instances from the centroid to the mean distance μ_d of all instances of a phone from the centroid. As can be seen in Table 6.11, the manually selected phone instances have a lower mean distance from the centroid than all instances of the phone. One exception occurs for the phone [ʃ], which may also be due to disturbing noise as in the case of the *me* voice.

Evaluation of Corpora Based on Manually Selected Phone Instances (*fe* Voice)

We also created a diphone corpus based on centroids computed from manually selected phone instances and synthesised two sets of test sentences. We found for the two corpora that about 15-20 % of the diphones are selected differently when using centroids based on manually selected phone instances. However, as for the *me* voice, the differences in the selected phone instances were almost not noticeable when listening to the two sets of test sentences.

Comparison of the Two Centroid Types

To estimate the effect the manual selection of phone instances on the centroid position, we investigated the difference between the respective centroids. More precisely, we measured the Euclidean distances between the centroids that were computed from the manually selected phones and those computed from all phones.

Centroid Distances for the *me* Voice The results for the *me* voice can be seen in Table 6.12. Higher distances between the two kinds of centroids can be seen for example for the phones [ʒ] and [z]. This is because for these phones, only really voiced phones were chosen in the manual selection and only 38 % for [ʒ] and 44 % for [z] of all phone instances were really voiced. That means that a lot of these voiced fricatives are clearly unvoiced, so that in this case the centroid of the manually selected phone instances represents a phone which is quite different to the centroid which represents all phones. A higher distance can also be observed for the phone [uː]. This may be due to the fact that only 46 % of the phone instances were labelled as suitable during manual classification. This means, a large share of the phones was considered as unsuitable, mainly because they sounded like the German [y] and not like a [uː]. Maybe this rigorous classification is due to the German native speaker background of the author and does not correctly represent the impression of a well-pronounced English [uː]. In any case, this large distance for the phone [uː] is most likely due to a mismatch between how the phones are actually pronounced by the speaker and the conception of that phone of the author who did the classification. For the *me* voice it can be concluded that the centroids computed from all phone instances correspond to a large extent to those computed only from manually selected phone instances.

phone	$d(z, z')$
[ɜ͜r]	0.56
[l̩]	0.37
[n̩]	0.62
[ə]	0.56
[ə͜r]	0.41
[ɑ]	0.49
[ɛ]	0.29
[ɪ]	0.64
[ŋ]	0.42
[ɔ]	0.36
[ʃ]	0.77
[ʊ]	0.42
[ʌ]	0.60
[ʒ]	1.03
[f]	0.59
[i]	0.71
[iː]	0.53
[l]	0.68
[m]	0.44
[n]	0.58
[o]	0.78
[r]	0.51
[s]	0.32
[uː]	1.19
[v]	0.48
[w]	0.29
[z]	1.18

phone	$d(z, z')$
[ɜ]	0.25
[ə]	0.57
[ɑ]	2.29
[ɑː]	0.47
[ɪ]	1.51
[ŋ]	0.39
[ɔː]	0.64
[ʃ]	0.63
[ʊ]	1.03
[ʌ]	0.94
[ʒ]	0.99
[e]	0.73
[f]	0.22
[iː]	0.69
[l]	1.43
[m]	0.26
[n]	0.49
[æ]	1.35
[r]	0.71
[s]	0.63
[uː]	0.59
[v]	0.63
[w]	1.21
[z]	1.75

Table 6.12: *Euclidean distances on MFCC between centroids based on manually selected phone instances z' and centroids of all phone instances z. On the left side, results for the me voice are shown, on the right for the fe voice.*

Centroid Distances for the *fe* Voice The results for the *fe* voice can be seen in Table 6.12. Differences in the centroids are larger than for the *me* voice. The large difference of the [ɑ] centroids is due to the fact that the phone [ɑ] deviates considerably from the phone [ɑː]. In the manual classification, however, phone instances that sounded like [ɑː] were chosen as good representatives of the phone [ɑ] because the [ɑ] phones varied to such an extent that no common other representative sound could be identified. In total, only about 10 % of the phone instances of the phone [ɑ] were classified as suitable, which led to the mismatch between the centroid based on all phone instances and the centroid based on these 10 % of the phone instances. Although distances for the *fe* voice are larger than for the *me* voice, they are still in the range of the variation that is typical for studio recordings of professional speakers. Especially the first two MFCC tend to vary more than the rest of the coefficients (see [PK08], Fig. 10.5). To that effect, we observed that if the first two MFCC are excluded from the distance measure, the distances drop considerably. Therefore, also for the *fe* voice the centroids computed from all phone instances are still close to those computed only from manually selected phone instances.

6.4.9 Discussion

MFCC in combination with the Euclidean distance measure proves to be most appropriate to represent spectral quality as it best classifies the manually created training data (see Table 6.9). This result is in line with the work in [WM98], where the Euclidean distance on MFCC performed very good and also with [SS01], a study which is related to our work as distance measures to detect concatenation discontinuity were investigated. According to that work, the Kullback-Leibler distance on power spectra had the highest detection rate of discontinuities followed by the Euclidean distance on MFCC. It is remarkable that features originally designed for speech recognition are also suited best for measuring spectral phone quality, where one would assume that a more detailed spectral description would be more appropriate.

The cepstra computed from the Mel-warped Straight spectrum seem not to be well suited for our task. Looking at the Straight spectra, we conjecture that these spectra give a too detailed representation of the spectral properties.

What alternatives to the Euclidean distance are concerned, we could confirm the result from [WM98] that Mahalanobis distance improves results for cepstral distances on a linear spectrum (like the LPC cepstral coefficients) but not for cepstral distances on a mel-warped spectrum like MFCC or the cepstra based on the Mel-warped Straight spectrum. However, in contrast to that work, we could not find any improvement using the Mahalanobis distance on line spectral frequencies.

Although it is not possible to completely separate suitable and unsuitable instances with any measure (see Fig. 6.1 for an illustration), experiments in Section 6.4.7 show that in practise false positives appear actually only very rarely in the final corpora. Generally, for corpus creation, we aim to select phone instances with centroid distances as low as possible. If a phone instance at the classification boundary is selected, it is very often because the alternatives show grave deficiencies in some other phone quality aspect. Section 6.4.8 showed that there is no need to manually select phone instances to improve the centroids, because centroids based on all phone instances are situated closely to those based on the manually selected ones.

However, to evaluate phone quality, we have to be aware that a spectral measure like the Euclidean distance on MFCC can only be used for stationary phones and neither for voiced nor unvoiced plosives[2]. Furthermore, such a measure can only cover spectral aspects. As we saw in the exploratory experiment described in Section 6.2, other aspects are as important and will be described in the following sections.

6.5 Phase Characteristics

For some speakers it can be observed that instances of the same phone have very diverse waveforms although they sound similar and the cepstral distance between them is quite small. This diversity can be attributed to considerable differences in the phase characteristics of these phones. If such phones happen to be concatenated, an artifact may be

[2] For affricates, as partly stationary phones, the centroid measure can be applied to the fricative phase, which is actually stationary.

clearly audible from the resulting signal. An example of such a case is shown in Fig. 6.4. A similar effect may be caused by erroneous pitch marks. E.g. for a nearly sinusoidal speech signal it is not always clear whether the pitch marks have to be set at the energy maxima that coincide with the positive or with the negative maxima of the fundamental wave (see Chapter 4).

Both these problems can be detected from the position of the pitch marks relative to the phase of the fundamental wave. If this phase value φ for a phone instance differs considerably from the average phase value μ_φ over all instances of that phone, one of the above mentioned cases applies and this phone instance should be avoided.

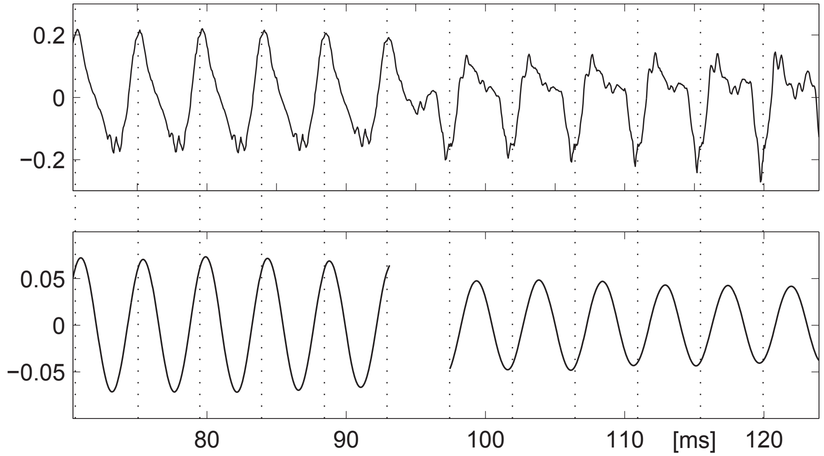

Figure 6.4: *Artifact at around 95 ms in a speech signal resulting from the concatenation of the diphones [tʏ] and [ʏɐ]. The speech signal is shown on top, the corresponding fundamental waves are shown at the bottom. The pitch marks in [ʏ] of the left diphone are set near the positive maximum of the fundamental wave, whereas they are near the minimum in [ʏ] of the second diphone.*

6.6 Fundamental Frequency Characteristics

If TD-PSOLA-based F_0 and duration modification is applied to speech segments that are to be concatenated, the F_0 characteristics of these segments may cause several issues. If the F_0 at the end of one speech segment deviates considerably from the F_0 at the beginning of the next segment that is going to be concatenated, the degree or even the direction of F_0 modification required to realise a smooth contour changes abruptly at the concatenation point. Furthermore, speech segments with rapidly rising or falling F_0 are not suited to be transformed into segments with constant F_0 or even with an opposite direction of F_0 movement. Thus, extreme F_0 values as well as rapidly rising or falling F_0 contours have a negative effect on phone quality.

6.7 Duration Characteristics

In concatenation synthesis, longer phones are preferred over shorter ones since, generally, shortening impairs the quality much less than lengthening. However, in automatically segmented speech signals, phone instance durations that are much higher than the mean phone duration may originate from segmentation errors. Mean phone durations and variances are computed for each phoneme, except for plosives where preplosive pauses and the burst part are treated separately. These two parts of a plosive can be easily split and are later assigned to separate diphones, so separate treatment is appropriate.

6.8 Voicing Characteristics

In the context of speech synthesis, voicing has two aspects. First, breathy or pressed vowels are not desirable because they do not sound clear, and second, irregularly glottalized speech is problematic for prosodic modification with TD-PSOLA (see Fig. 6.5, see also Section 5.5). Therefore, we want to penalise voiced stationary phones with

these properties. We used the output of the frame classifier presented in Chapter 3, that decides if speech frames are voiced, unvoiced or mixed and distinguishes between regularly and irregularly glottalized frames and included the number of mixed frames and the number of irregularly glottalized frames as an aspect of phone quality.

Whereas mixed frames in a voiced stationary phones give a strong hint at breathy voice quality, mixed frames naturally occur in unvoiced phones and plosives. Therefore, the number of mixed frames is only considered for voiced stationary phones. Irregular frames, however, have to be considered with voiced and unvoiced stationary phones. Irregular frames in voiced and unvoiced fricatives hint at noise often caused by strong low frequency signal components.

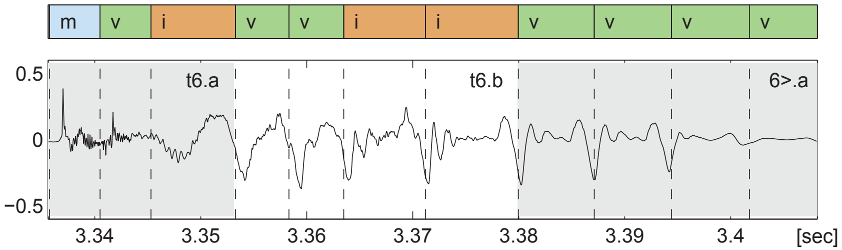

Figure 6.5: *Synthesised speech signal of the phone sequence [tɐ]. The periods are marked with v for voiced, m for mixed, and i for irregular according their frame classification. The creaky segment around 3.37 s is clearly perceived as disturbing.*

6.9 Signal Intensity

In natural speech, the signal intensity varies considerably even between instances of the same phone. In concatenative speech synthesis, strong variation in signal intensity at concatenation points often causes a disturbing effect and has to be avoided. We learnt from listening experiments that an intensity difference of 6 dB is the limit of what is perceived to be still tolerable. In fact, the degree of irritation caused by the intensity change depends on its position within the word. In the example speech signal shown in Fig. 6.6, a 6 dB increase of the signal belonging to diphone [iːn] sounds strongly disturbing, a 6 dB increase of the signal belonging to diphone [ən] is perceived only moderately disturbing. Generally, a sudden intensity jump is perceived as more irritating than a sudden intensity drop. One method to mitigate the effect of an intensity mismatch at the concatenation point was presented in Section 5.9, where a power smoothing approach during a very short period around the concatenation point was presented. However, the possibilities to mitigate such a mismatch are limited, especially as a change in intensity is often an indicator for a change in speaking style. As a consequence, phone instances that strongly deviate from the mean intensity of the phone should be avoided.

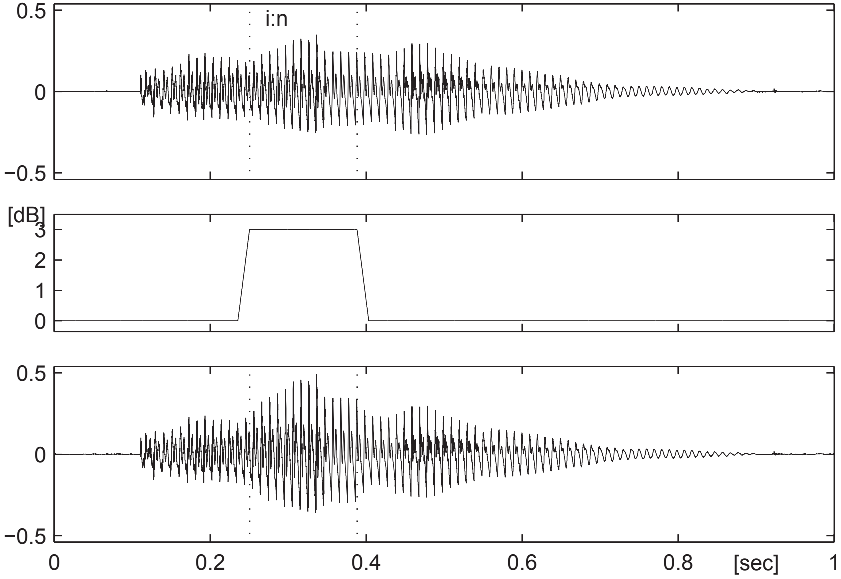

Figure 6.6: *Speech signal of the word "ihnen", pronounced as [|iːnən] where the intensity of the diphone [nə] has been increased by 6 dB, which corresponds to some doubling of the signal amplitude. On top the original signal is shown, in the middle the applied gain, at the bottom the modified signal.*

6.10 Loudness

6.10.1 Online Loudness Monitoring

In contrast to the aspects of phone quality mentioned up to now, loudness is an aspect that has to be measured and monitored at recording time, because loudness inconsistencies in the recordings can hardly be corrected later. This is because speakers tend to increase their vocal effort when speaking louder, which not only changes the loudness itself, but also the spectral characteristics of the voice. Since a speech database is recorded typically for a couple of days, if not weeks, the speaking style and with it loudness may change over that period and must hence be monitored to ensure consistency of the recorded corpus.

6.10.2 Related Work

Loudness perception has been extensively researched in psychoacoustics with synthetic signals, which mostly consisted of pulsed or continuous sine waves or noise of various bandwidths [ZF99]. Nevertheless, the application of this knowledge on audio material such as speech proves to be difficult.

There have also been comprehensive studies on loudness perception of typical radio broadcast signals containing a heterogeneous mixture of music and speech including even extreme speaking styles like shouting [SQN04, SN04]. In contrast to that, we aimed at devising a measure for a narrowly drawn scenario, which is monitoring of speech recordings. The signals used in this study were confined to high quality professional speech recordings without any noticeable background noise. Furthermore, the loudness measure is intended to be used only in a speaker-dependent way. This means that only relative loudness values along the recordings of a single speaker are relevant. In particular it is not required to compare loudness values from different speakers. One constraint provided by the application was a very efficient computation of the measure, as the loudness has to be monitored on-line together with other properties of the speech during the recordings.

6.10.3 Measuring Loudness

Our aim was to design a loudness measure that corresponds as well as possible to the subjectively perceived loudness and can be applied to speech signals with a length of at least a few seconds. Preliminary investigation had shown that the power of the speech signal is not sufficient as a measure of perceived loudness. The long-term power or similar measures like different variants of the *equivalent continuous sound level* L_{eq} [ANS94] depend on the amount of pauses, or if those are excluded, on the pause/speech discrimination threshold. The signal's short-term power again is strongly dependent on the phoneme. Some investigation had been conducted on the effects of vocal loudness on signal properties like spectral tilt [SN06], yet measuring only vocal loudness would neglect the effect of the speaker's distance to the microphone, which the measure should also account for.

The Corpus

The corpus that we used for our investigations was recorded with a French professional female speaker. The recording session took place in an anechoic room and professional equipment was used. The level of background noise is therefore reduced to a minimum. The sentences were recorded with a sampling frequency of 44,100 Hz. The material contains approximately 9 minutes of speech, split into 72 sentences. These sentences were recorded on five different days with two different sound engineers. On the first day, the distance between the microphone and the speaker was too big, so that the intensity of the recorded speech was insufficient. This was partly corrected on the second day, but it is only from the third day on that the distance between the microphone and the speaker was chosen in an appropriate manner. This material allowed us to make first plausibility checks of our measures.

Distribution of Power

In a first step, we decided to study the distribution of the short-term power within the different types of sentences. We took the 72 sentences of the data set and computed the short-term power of the segments containing speech (with silence removed). We used a window length of 5 ms and a window shift of 5 ms. The results are shown in Fig. 6.7.

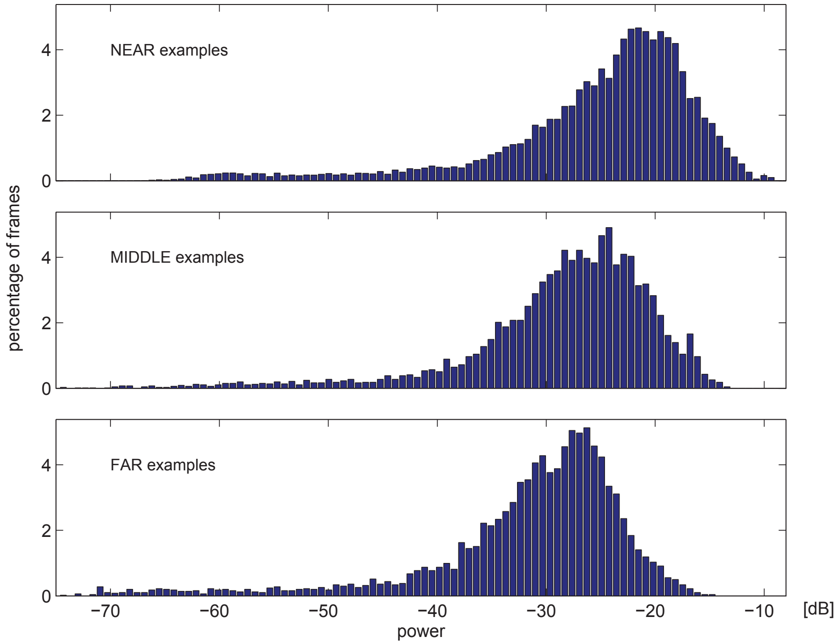

Figure 6.7: *Histogram of the power for utterances recorded on the first day ('far'), on the second day ('middle'), and on the following days ('near').*

The x-axis is in logarithmic scale, since this better matches human perception. The x-axis does not actually correspond to any physical unit, since the microphone was not calibrated in any way. However, the absolute value of power (in Watts) is of no practical relevance to us, since we are only interested in how the measurements relate to each other. In that respect, we can clearly observe a shift of the histograms towards the left when going from the 'near' sentences to the 'far' sentences.

Different Measure Proposals

In a second step, we proposed six different loudness measures, that were to be correlated with subjective evaluations. These measures were based on the following considerations:

1. Human perception of loudness is roughly proportional to the logarithm of the power of the acoustic signal.
2. The subjective impression of loudness seems to depend more on the segments with high intensity than on the ones with low intensity.
3. Spectral tilt is influenced by vocal loudness.

The six measures are presented below. The short-term power p_i for frame i is defined as follows:

$$p_i = \frac{1}{N} \sum_{j=0}^{N-1} |x(j)|^2. \tag{6.8}$$

Measure 3 uses the segmental alpha ratio defined as

$$\alpha_i = \frac{I_{HF}^{(i)}}{I_{LF}^{(i)}}, \tag{6.9}$$

where $I_{HF}^{(i)}$ and $I_{LF}^{(i)}$ are the signal power of frame i above and below 1 kHz, respectively. For each of these measures, a window length of 50 ms and a shift of 5 ms were chosen. P denotes the set of p_i of all frames. The coefficient α_i together with the set A is defined accordingly.

Measure 1: A first attempt was based on the histograms presented in Fig. 6.7. The mode of the histograms seemed a good indication for the category the utterance belongs to ('far', 'middle' or 'near') and furthermore was considered a stable quantity regarding outliers:

$$m_1(P) = \operatorname*{mode}_P \log(p_i). \tag{6.10}$$

Measure 2: Measure 2 is based on the assumption that only the high power segments are relevant. We therefore take the logarithm of the 90 % percentile of P:

$$m_2(P) = \log(P_{0.9}). \tag{6.11}$$

Measure 3: Instead of completely ignoring low-power segments, an intensity-dependent weighting was applied with the factor p_i^n, where n was set to 0.9:

$$m_3(P) = \frac{\sum p_i^n \log(p_i)}{\sum p_i^n}. \tag{6.12}$$

Measure 4: Measure 4 investigates the correlation between spectral tilt and loudness of speech. It uses the alpha ratio as defined in Equation (6.9) as follows:

$$m_4(A) = \log(A_{0.5}), \tag{6.13}$$

where $A_{0.5}$ denotes the median of the alpha ratios α_i in set A.

Measure 5: Measure 5 combines the ideas presented so far. First it discards the lower intensities using a threshold which is defined as a fraction of the maximal power. We define a set:

$$P_c = \{p_i \mid p_i > c \max_P p_i\}. \tag{6.14}$$

We set c to 0.08. The power of the remaining segments in the set P_c is then averaged and logarithmised:

$$m_5(P_c) = \log\left(\frac{\sum_{s_i \in P_c} s_i}{|P_c|}\right). \tag{6.15}$$

Measure 6: Measure 6 is very similar to measure 5 with the difference that logarithm is taken from the geometric mean instead of the arithmetic mean. This confers lesser importance to the p_i having a higher value:

$$m_6(P_c) = \log\left(\sqrt[|P_c|]{\prod_{s_i \in P_c} s_i}\right). \tag{6.16}$$

6.10.4 Evaluation Method

In order to assess the measures, we correlated them with the results from subjective evaluations. For those evaluations, a loudness matching experiment was conducted using the *method of adjustment* [ZF99]. This method requires the subjects to adjust the loudness of a comparison stimulus until it matches a reference stimulus, since it is impossible for humans to give an absolute value for the loudness of an utterance. The power of the comparison stimulus could be changed with a slider, applying a gain from -5 dB to +5 dB (with discrete intervals of 1 dB) until the loudness of the two stimuli was deemed equivalent. The two stimuli were randomly drawn from the data set presented in Section 6.10.3, which consists of 72 sentences. Each stimulus was used exactly once, therefore the complete test consisted of 36 comparisons where the stimuli for each comparison were randomly drawn from the data set. By employing this pair-matching method, no fixed reference stimulus is used, making the method less prone to bias.

The listening test was completed by 10 different subjects, leading to a total of 360 comparisons. Fig. 6.8 shows a summarisation of the evaluations. As can be seen from the plot, the agreement between the subjects was quite high with an inter-subject cross-correlation $\hat{r}$ of 0.94. Inter-subject cross-correlation stands for the average correlation between the evaluation of one subject and the average evaluation made by the other subjects.

Linear regression was used with an examination of the the residual error and the correlation coefficient to estimate the measures' goodness-of-fit and thus to assess their aptitude to model relative human loudness perception. Linear regression was also used to re-scale the measures in such a way that the difference of loudness measured between two utterances reflects the actual difference of intensity evaluated in dB. Linear regression was preferred to a more complex model due to our limited amount of data.

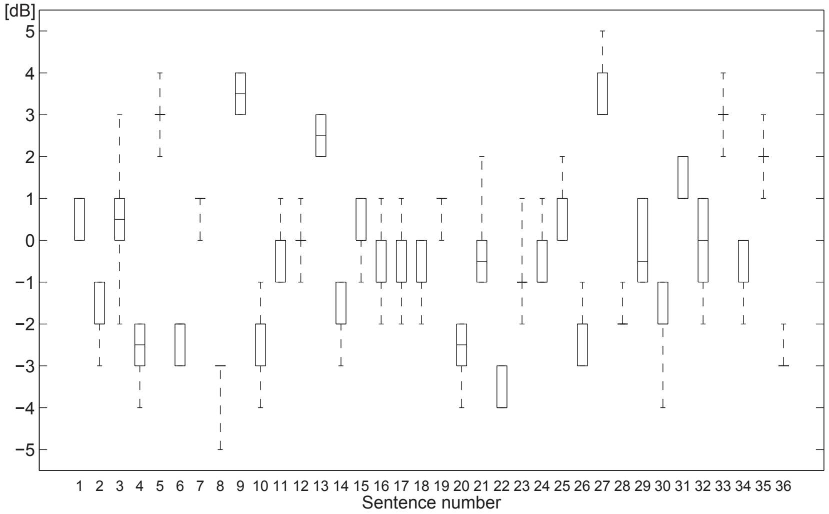

Figure 6.8: *Boxplot summarising the perceived difference of loudness in dB by 10 subjects for the 36 comparisons. The boxes show the interquartile interval while the whiskers show the minimal and maximal values.*

6.10.5 Results

The results of the regression analysis for the six measures proposed in Section 6.10.3 are displayed in Fig. 6.9. To facilitate comparison of the measures, the linearly re-scaled measures are plotted versus the average evaluations of the test subjects. This implies that a perfect measure (one with all error terms equal to 0) would only produce points on the line $y = x$.

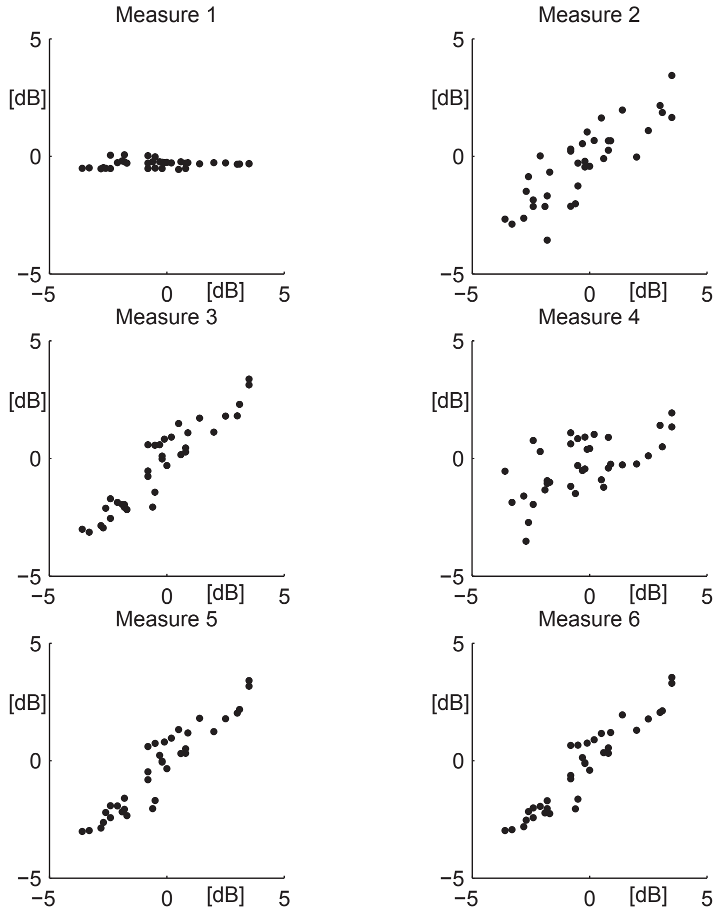

Figure 6.9: *Results of the six intensity measures plotted against the average of the evaluations made by the test subjects. The measures x_i were linearly re-scaled as $\hat{\beta}_0 + \hat{\beta}_1 x_i$ to facilitate their comparison.*

The numerical results yielded by the regression analysis are presented in Table 6.13. Since this study is concerned with measuring relative loudness, the results for the intercept β_0 are not displayed, as β_0 only provides information about the absolute value of the measure. The second column of Table 6.13 presents the slope estimate $\hat{\beta}_1$, which was computed with the least-squares method. The 95 % confidence intervals are also shown. These confidence intervals are given as a percentage of $\hat{\beta}_1$, so that they can be compared between the different measures. The estimate s of the standard deviation of the error terms in decibels is given in the third column. The value of s can be used to derive a 95 % confidence interval for the value returned by the measure. In the last column the correlation r is given between the measure and the evaluations done by the human beings.

The six measures are reviewed below:

Measure 1: The results produced by measure 1, which takes the mode of the discrete power distribution, are particularly surprising, as the correlation coefficient is almost 0. The resolution of the power distribution is extremely low when computed for one sentence only, such that the mode has a high random component. This measure may yield different results when considering several minutes of speech, but it is ineffective for short sentences.

Measure 2: Measure 2 delivers reasonably good results. However, choosing the right percentile (90 % in this case) is critical: lowering the threshold to 75 % impairs the results significantly.

Measure 3: This measure, which weights the segments according to their power, is among the measures that perform best, along with measure 5 and 6.

Measures 4: Measure 4 shows a positive correlation between the α-ratio and the loudness of speech. This corresponds to the results presented in [SN06]. However, this correlation is not strong enough for the α-ratio to be used as a loudness measure.

Measures 5 and 6: These are the two measures that performed best together with measure 3. The two measures are identical except for the order in which the mean and the logarithm are taken.

The results obtained by these two measures are also similar, with correlation coefficients of respectively 0.94 and 0.95. The value of the threshold c, which was set to 0.08 is not critical, as long as it remains between 0.01 and 0.1. Below values of 0.01, measure 6 is susceptible to be impaired by large negative values that result from the log of low energy segments.

Measure	$\hat{\beta}_1$	s **(in dB)**	r
Measure 1	-0.07 (±401.42 %)	1.96	-0.09
Measure 2	3.66 (±22.08 %)	1.06	0.84
Measure 3	3.82 (±12.56 %)	0.67	0.94
Measure 4	1.20 (±43.36 %)	1.54	0.63
Measure 5	4.01 (±12.34 %)	0.66	0.94
Measure 6	4.08 (±11.98 %)	0.64	0.95

Table 6.13: *Regression analysis for the different loudness measures showing the slope $\hat{\beta}_1$ with the 95 % confidence intervals, the standard deviation estimate of the error s and the correlation coefficient r.*

6.10.6 Loudness Correction of Recorded Corpora

For two voices, that were recorded in different recording sessions, which lead to rather different speaking style and loudness, we made an attempt to apply the loudness measure 6 to correct at least the loudness aspect. We decided to use measure 6 since the test showed that it is the best performing measure.

Loudness Correction of Female German Corpus

The first voice, a female German voice (*fg2*), was recorded in at least six sessions with different equipment and different recording settings (see [TH99]). Especially the sentences recorded for prosody training are much lower than the recordings of the diphone carrier words (see Table 6.14). We adapted the loudness of all sentence/word recordings towards a target loudness, which was the loudness of the diphone carrier words. After compensation of the loudness differences, it was possible to combine the recordings to generate a diphone corpus.

Recording session	l_μ [dB]	l_σ [dB]
German prosody sentences	1.24	0.61
French prosody sentences	0.99	0.52
German diphone carrier words	2.69	0.90
English diphone carrier words	2.47	0.75
French diphone carrier words	1.41	0.65
Italian diphone carrier words	2.21	0.83

Table 6.14: *Mean loudness and standard deviations measured for the different recording sessions of the speaker fg2.*

Loudness Correction of Male German Corpus

The male German voice (*mg*) had been recorded in at least three sessions with analogue equipment (see [Kae85]). Later these recordings were digitised. The speech signals from these recordings show some differences in intensity, as shown in Table 6.15. We applied the same loudness compensation described above and created diphone corpora from all possible subsets of these three recording sessions. However, we found that neither subset from the three recording sessions could be combined because the remaining differences in speaking style were still too big. An additional reason for that incompatibility most probably was the use of different recording equipment for the three sessions.

Recording session	l_μ [dB]	l_σ [dB]
Diphone carrier words	1.65	0.83
Prosody sentences 1	2.21	0.74
Prosody sentences 2	1.52	0.56

Table 6.15: *Mean loudness and standard deviations measured for the different recording sessions of the speaker mg.*

6.10.7 Discussion

The results of our study on loudness showed that measure 6 proved useful to efficiently measure relative loudness during speech recordings. In combination with a surveillance of the speaker's distance from the microphone, this method can be used to ensure constant loudness along with constant vocal effort. We have implemented the proposed measure in a monitoring tool for studio speech recordings. This tool is now actually used for speech recordings and was found to work much better than the previous one.

Other studies in this area differ in the fact that they measured absolute loudness on the one hand and on the other hand used either synthetic signals [ZF99] or broadcast material [SQN04, SN04]. The latter contained a heterogeneous mixture of music and speech whereas our study only used high-quality speech recordings. As a result of that specific application scenario, our measure has not been tested for noise robustness and sensitivity.

6.11 Characteristics of Plosives

6.11.1 Burst intensity

The most important characteristic of plosives is the burst, which is the sudden air flow after the release of the closure. The burst has to be strong enough to be clearly identified by the listener. In order to determine the burst intensity, we apply a heuristic, which requires two conditions to be met by the burst phase. First, intensity boundaries are determined by looking at large increases of the intensity, according to the method described in [GO07, HP10]). If no such boundary is found, the burst is considered as too weak. Second, we look at the intensity of the plosive in the band between 2 and 8 kHz and compare it to a noise threshold, which is determined from the whole signal with the loudness measure described in Section 6.10.3. If the plosive frame with the highest intensity is below this threshold, the burst is also considered as too weak.

6.11.2 Plosive Aspiration

In some languages, for example in German, correct pronunciation requires to articulate unvoiced plosives either with or without aspiration, depending on the context. Therefore, we created a method to decide from the speech signal which plosives are aspirated and which ones are not.

From phonetics it is known (see e.g. [Lis63]) that unvoiced plosives with a voice onset time (VOT) greater than some 50 ms are clearly heard as aspirated, whereas no aspiration is heard if the VOT is less than 20 ms. To assess the aspiration of unvoiced plosives, we have to detect the release point and the start of voicing. This is illustrated in Fig. 6.10. The release point, which is the boundary between the closure and the burst, is determined by looking for the point of maximum increase of the power in the band from 2 to 8 kHz (see [GO07, HP10]).

The start of voicing is detected from the intensity of the fundamental wave. The fundamental wave can be achieved from the convolution of the speech signal with a Hamming window of the size of the period length T_0 as described in Section 3.2.1. The intensity of fundamental wave is then computed as follows:

$$e(i) = \sqrt{\frac{\sum_j [f(j) \cdot u(j-i)]^2}{\sum_j u(j)}}, \tag{6.17}$$

where $f(j)$ is the fundamental wave at sample j and $u(\cdot)$ is a Hamming window of length $2T_0$ centred at 0. From this intensity curve the start of voicing is detected by means of a threshold that depends on an estimate of the speech loudness. Finally, the difference between the start of voicing and the release point yields the required VOT.

The information about the plosive aspiration, which is retrieved with this method, is used to improve the segmentation of the speech databases by correcting the labels of aspirated and non aspirated plosives if necessary.

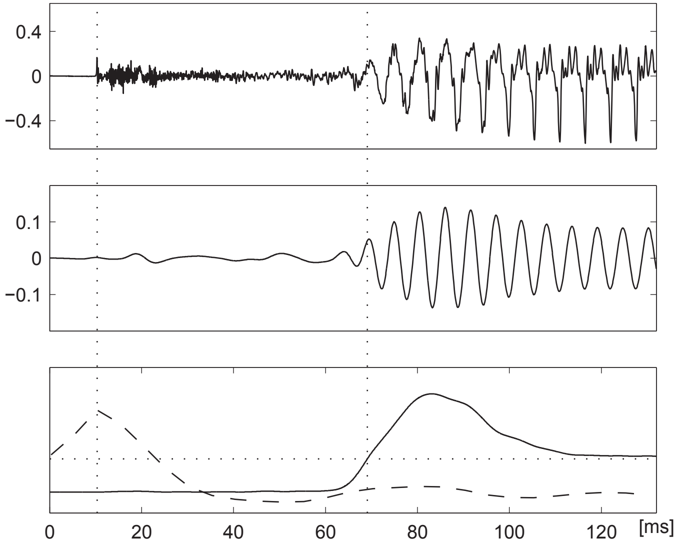

Figure 6.10: *Estimation of the VOT in a speech signal (top plot) with the phones [ta͜i]. The fundamental wave is shown in the middle plot. The maximum power change curve (dashed, bottom plot) defines the release point at 10 ms. The intensity of the fundamental wave (bottom plot) crosses the threshold at 68 ms. Therefore, the VOT is greater than 50 ms and the unvoiced plosive is considered to be aspirated.*

6.11.3 Determining Spectral Quality of Plosives

An experiment to use the centroid measure to characterise the quality of unvoiced plosives has been conducted with *fg* voice. Although for the plosives [t] and [$\mathrm{t^h}$] the centroid method seemed to give some reasonable results, for all other plosives the results depended strongly on the subsequent phone. We refrained from further experiments, because of strong hints that the centroid method for plosives would be very susceptible to coarticulation effects.

6.11.4 Preplosive Pauses

The first diphone corpora did not distinguish between transitions of phones to different plosives (see [Kae85]), primarily to keep the number of different diphones low to reduce corpus size. However, it can be observed that in most cases only from the articulation of the phone that precedes the preplosive pause the character of the subsequent plosive can be determined[3]. Therefore, in the context of diphone synthesis, it is important to distinguish the transitions from a phone to different preplosive pauses, otherwise misleading hints of a different plosive than the one following may disturb the listener. To distinguish these transitions, we introduced individual transcriptions for different preplosive pauses in our system. We extended the IPA symbols by denoting a preplosive pause that pertains to a plosive by placing a dot under the corresponding plosive, as dots denote pauses in IPA. So the preplosive pause of the plosive [b] is written as [ḅ].

6.11.5 Plosive Elisions

Elisions are most common in the simplification of consonant clusters (see [Jon03]). If we consider the case of two subsequent plosives where the first plosive is elided, we observe that the character of the elided phone is determined from the articulation of the phone that precedes the preplosive pause. Only the second plosive is articulated with a

[3] This is because already during the articulation of the preceding phone, the articulators move according to the plosive's place of articulation.

proper burst. This means, in plosive elisions the phone that precedes the preplosive pause gives the impression of a different burst to follow than actually follows because of the elision. E.g. the plosives [p] and [t] of the at the word “uptake” are realised first with a transition from [ɑ] to a [p], where the [p] is not realised and second, with the plosive [t^{h}].

Preplosive Pauses of Elisions

As a first idea, the usage of explicit preplosive pauses as presented in Section 6.11.4 seems sufficient. However, the mean durations of preplosive pauses differ considerably depending on whether an elision follows or not. Depending on the speaker and the language, preplosive pauses that are followed by an elision are lengthened by up to 124 % compared to preplosive pauses followed directly by their plosive (see Tables B.3 to B.2 in Appendix A). Thus, for the selection of diphone elements it is preferable to distinguish preplosive pauses that are followed by elisions from preplosive pauses that are followed directly by their plosives.

If no such distinction was made, preplosive pauses would have to be strongly lengthened. This would lead to audible artifacts, as preplosive pauses in many cases are no real pauses that contain exclusively silence. Especially with voiced plosives, the preplosive pauses often contain clearly audible speech. However, if preplosive pauses from the context of elisions are used, no extensive lengthening is required.

Introduction of Additional Preplosive Pause Transcriptions

As a solution to the problem described in the previous section, particular transcriptions for preplosive pauses that appear in the context of elisions were added to the notation that is used in our system. Bursts that are elided are denoted with two dots under the corresponding plosive. For example, an elided [d] is noted as [d̤]. Using again the example mentioned above, in which the phone [p] is elided in the context of [ɑ] and [e͜ɪ], the plosive [p], that is not audible, is written as [p̤]. So the phone sequence for the word “uptake” is written as [ʌp̤t^{h}e͜ɪḳk^{h}]. For the selection of diphone elements from speech databases, the bursts that are inaudible due to an elision had also to be introduced in the

segmentations of the different recordings.

6.11.6 Voiced Plosives after Pauses

Humming Sound Effect

For some speakers, voiced plosives after pauses, mostly at the beginning of sentences, show a short phase of “humming” (see Fig. 6.11). This humming sounds like an [m] before [b] and like an [n] before [d] and like an [ŋ] before [g]. The segmentation erroneously considers this humming to be part of the burst and sets the phone boundaries accordingly. If these plosives are later used as diphones, this humming sound has a disturbing effect on the listener. Even worse, if the diphone that contains the plosive is shortened, the shortening takes place at the end of the plosive. This is because the beginning of the plosive is assumed to contain the beginning of the burst, which should not be modified. In this special case, however, this assumption causes the the burst to be cut off as the burst is actually located at the end of the plosive. As a consequence, only the humming sound and no burst is audible in the synthesised signal.

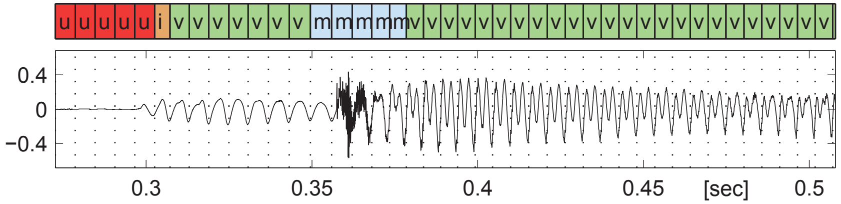

Figure 6.11: *Voiced plosive [d] following silence at the beginning of a sentence, recording from the fg voice. The humming sound before the burst occurs from 0.3 to 0.35 s. Voicing information per period is shown on top: unvoiced (u), voiced (v), mixed (m) and irregular (i) periods.*

Detection of Humming Sound Cases

To detect these cases of humming sound in the context of voiced plosives after a pause we rely on the frame classification introduced in Chapter 3. We found that most of the affected plosives show a particular voicing pattern, that is irregular frames, followed by voiced frames, again followed by unvoiced or mixed frames. More precisely, formulated as a regular expression, the pattern is $(\texttt{irregular}^*\texttt{voiced}^+)^+(.)^*$ where the dot stands for any classification.

We found that for the *fg* voice, which was most problematic in this respect, there was a clear improvement through this heuristic humming sound detection. Bursts, that were occasionally cut off before, were clearly audible when penalties were applied to plosives that had the typical humming frame pattern. One typical example from the *fg* voice is shown below. Before the correction, the problematic diphone shown in Fig. 6.11 was used with the burst part of the [deː].a diphone cut off. After the correction, a more suitable plosive diphone was selected with a proper burst (see Fig. 6.12).

6.12 Characteristics of Fricatives

6.12.1 Glottalisation in Fricatives

Glottal stops often introduce noise towards the end of preceding fricatives. This noise is caused by creaky voice phonation, which can often be observed in phones that precede a glottal stop (see [LM96]). A typical example can be seen in Fig. 6.13. The frame classification method presented in Chapter 3 enables us to characterise fricatives in terms of glottalisation quite reliably.

6.12.2 High Frequency Noise in Fricatives

For a particular voice, a French female speaker, we observed that an exceptional high number of fricatives contained high frequency noise like whistling or hizzing. We developed a noise detection method for that particular voice (see [Sim08]), where we distinguished sibling, fizzling, buzzing, and gliding noises. This detection method was based on spectral features and autocorrelation in high frequency components and reached a classification rate of 73.48 % on a training and test set of 265 and 132 fricatives, resp. However, this classification method was speaker-dependent, as it relied on the estimation of some parameters from the noisy fricatives of this voice. In an attempt to extend this method to speaker-independence, we faced the fact that all the other voices that were available to us almost did not contain any high frequency noise in fricatives, thus making it impossible to find a sufficient amount of training material for a speaker-independent classifier. On the other hand, this lack of noise in fricatives was also good news, as for the four voices that we used for corpus generation high frequency noise in fricatives did not pose any problem.

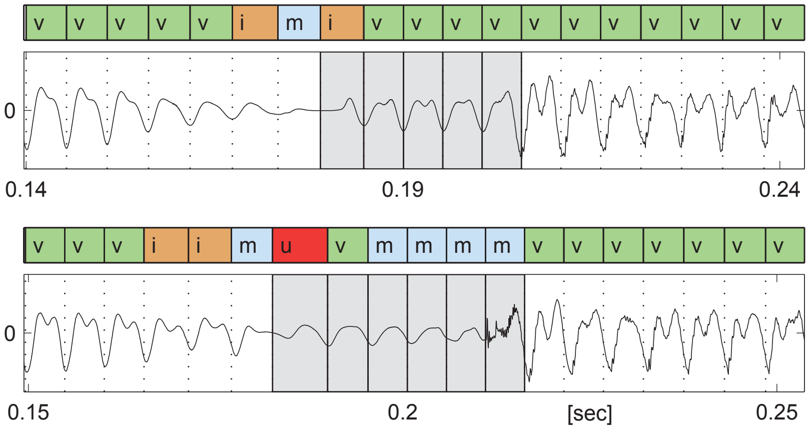

Figure 6.12: *Synthesised words “in der“ ([ɪndeːɐ̯]) with the burst phase of the semi-diphone [deː].a cut off in the top plot (shaded grey). In the bottom plot, the same words synthesised with a more suitable diphone (shaded grey). Voicing information per period is shown on top: unvoiced (u), voiced (v), mixed (m) and irregular (i) periods.*

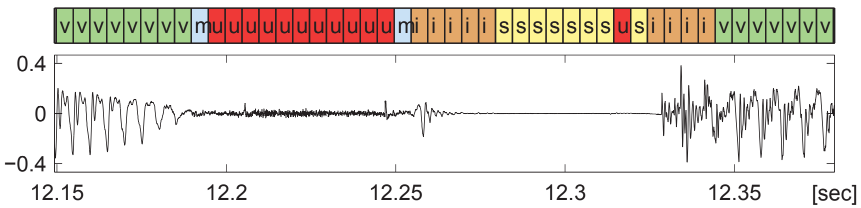

Figure 6.13: *Speech segment [uf|a] of female German fg voice. Voicing information per frame is shown on top: unvoiced (u), silence (s), voiced (v), mixed (m) and irregular (i) periods. The noise after 12.25 s in the fricative [f] followed by glottal stop around 12,33 s is classified as irregular speech.*

6.12.3 Mixed Excitation in Voiced Fricatives

Voiced fricatives contain both voiced and unvoiced signal components. Some speakers, however, tend to substitute voiced fricatives by their unvoiced counterparts[4]. These fricatives do not contain voiced signal components and can be identified by a low rate of mixed frames. Voiced fricatives that are properly realised, on the other hand, contain a high rate of mixed frames, although typically not all frames will be classified as mixed. Therefore, the rate of mixed frames can be used as a feature to measure quality of voiced fricatives (see also Section 6.8).

6.13 Discussion

We extended the concept of phone quality from a purely spectral aspect to a set of several, partly orthogonal aspects. Furthermore, features were determined to quantify these aspects so they can be combined to a single phone quality score. Actually, two possibilities for this combination will be proposed in Chapters 7 and 8.

Apart from a precise characterisation of phone properties in the post-processing step of speech recordings, the flip side of this approach was also touched, namely to improve recording quality in the first place. Loudness emerged as an important point, as it may not be possible to correct loudness variations belatedly as we reported in Section 6.10.6. Loudness is one feature to be monitored during speech recordings, another property could be speaking style, including speech rate and pitch. Vocal fatigue could also be monitored to determine the point of time for the recordings to be interrupted, as vocal fatigue influences the speaking style as well. Actually, a tool to determine vocal fatigue was intended to be implemented in the framework of this work, however, the lack of sufficient data refrained us from completing this task[5]. Finally, recording quality could be improved with an automated detection of click and smack sounds that speaker often tend to produce if they did not drink enough during the recordings. However, this task is a very challenging one, as these clicks and smacks are very subtle and hard to distinguish from intentional plosives and glottal closures.

[4] This is a frequent phenomenon for speakers of German from Switzerland, Austria and the southern regions of Germany.

[5] Speech recordings normally are interrupted as soon as vocal fatigue is noticed by any of the recording staff and continued once the voice talent has refreshed.

Chapter 7

Combining Phone Quality Aspects With a Linear Approach

In this chapter, we present a simple method to combine the phone quality features described in Chapter 6. We used penalty functions on these features and designed these functions in a way they can be added to create a single measure for each phone. We then used this measure to select diphone sets from four different speech databases. The quality of these diphone sets was demonstrated by means of synthesis examples enclosed in [EP11] and showed that the proposed measure can be applied to automatically select from a speech database all necessary diphones for high-quality speech synthesis. The use of penalty functions described in this approach is a simple heuristic method, that does not require training data and gives reproducible and transparent results. A more complex approach to combine phone quality features based on machine learning will be presented in Chapter 8.

7.1 Penalty Functions

The penalty functions presented in this chapter transform the values of the features described in Chapter 6 into a penalty value. For phone quality assessment in the context of diphone selection, we want to assign a very small penalty value (less than 1) to a phone if the aspect under consideration is within limits of acceptability from a perceptual point of view, and a high penalty otherwise. This concept is reflected in the various exponentiations that are used in the computation of these penalty values. To obtain one value for the overall quality of a phone, the sum of these penalty values is taken.

7.1.1 Spectral Penalty

According to the results on spectral characteristics presented in Section 6.4, we used the Euclidean distance on 12-dimensional MFCC, whereby the zeroth cepstral coefficient was neglected. From experiments we found that phone instances with a distance $h < 2$ from their corresponding centroid still are perceived as very clear. Therefore, we designed a function that strongly penalises phones with higher distance values:

$$P(h) = \exp(0.4(h - 2)) \tag{7.1}$$

The exponentiation factor of 0.4 accounts for the steepness of the penalty function increase, with a higher factor leading to a more rapid increase. As we subtract 2 from the centroid distance value h the penalty function takes the value 1 for a centroid distance of $h = 2$.

7.1.2 Phase Penalty

Considerable differences between the phases of phones should be avoided (see Section 6.5). An example of such a case is illustrated in Fig. 6.4. Therefore, considerable deviations of the phase value φ for a phone instance from the average phase value μ_φ over all instances of that phone should be penalised. With the phase values expressed in radians, we apply the following penalty function, which is illustrated in

Fig. 7.1:

$$P(\varphi) = (3 \cdot (\varphi - \mu_\varphi))^4 \tag{7.2}$$

The exponentiation factor of 4 accounts for the steepness of the penalty function increase. The function should be relatively flat for phase values that are still tolerable (from -0.3 to 0.3) and then rise sharply to penalise higher phase values. This function behaviour should model the acoustic effect of the phase differences: small differences in the phases of phones are inaudible, whereas a phase difference of 0.5 has a strong disturbing effect. Of course, the factor of 4, which is used here, is only a rough estimate to obtain the desired function behaviour and not a precise value. Finally, the multiplicative factor of 3 is chosen to scale the function in a way that the value 1 is reached when the values for the phase difference start to become unsuitable.

7.1.3 Fundamental Frequency Penalty

To account for the effects of unsuitable F_0 contours as described Section 6.6, we have defined two penalty functions. The first one penalises phones with a fundamental frequency that considerably deviates from μ_f, which is the mean value over all instances of that phone:

$$P_1(F_0) = (10 \cdot |f - \mu_f|)^3 \tag{7.3}$$

Note that the logarithm of F_0 is used, which makes the formula equally valid for male and female voices. Therefore, the mean F_0 value over all N frames of a phone instance is $f = \frac{1}{N}\sum_n(\log F_0(n))$. The exponential and multiplicative factors are used in the same way as in Formula 7.2. Again, a penalty value of 1 should be reached for deviations from μ_f that may cause a disturbing acoustic effect.

The variation of F_0 within a phone instance is expressed with the temporal derivative and results in the second penalty:

$$P_2(F_0) = \exp(200(f' - \mu_{f'} + \sigma_{f'})) + \exp(100(\hat{f}' - \mu_{\hat{f}'} + \sigma_{\hat{f}'}))\,, \tag{7.4}$$

where $f' = \frac{1}{N}\sum_n |\log F_0(n) - \log F_0(n-1)|$ is the mean absolute derivative of the F_0 of this phone instance and $\mu_{f'}$ and $\sigma_{f'}$ are the mean and standard deviation of f' over all instances of that phone. Similarly, $\hat{f}'$ is the mean over the highest 25 % of the components that contribute to f', and $\mu_{\hat{f}'}$ and $\sigma_{\hat{f}'}$ are the mean and standard deviation of $\hat{f}'$ over all instances of that phone.

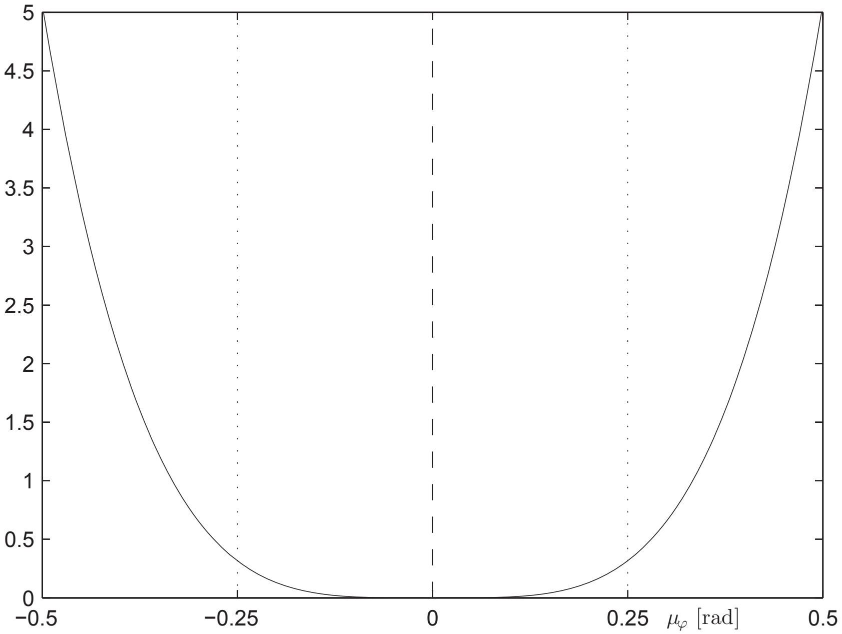

Figure 7.1: *Penalty function for the phase characteristics. The x-axis represents the difference of pitch mark phase φ from the average phase value μ_φ in radians.*

7.1.4 Duration Penalty

To prefer longer phones over shorter ones and at the same time to reject phone instance durations that are much higher than the average phone duration the following function was designed:

$$P(d) = 10 \cdot |\log d - (\mu_d + \sigma_d)|^3 \tag{7.5}$$

This function prefers phones that are one standard deviation σ_d longer than the mean duration value μ_d (see Fig. 7.2). Note that durations are in seconds and are used in the log domain, thus the mean duration of J phone instances is $\mu_d = \frac{1}{J}\sum_j \log d_j$ with $j = 1 \ldots J$.

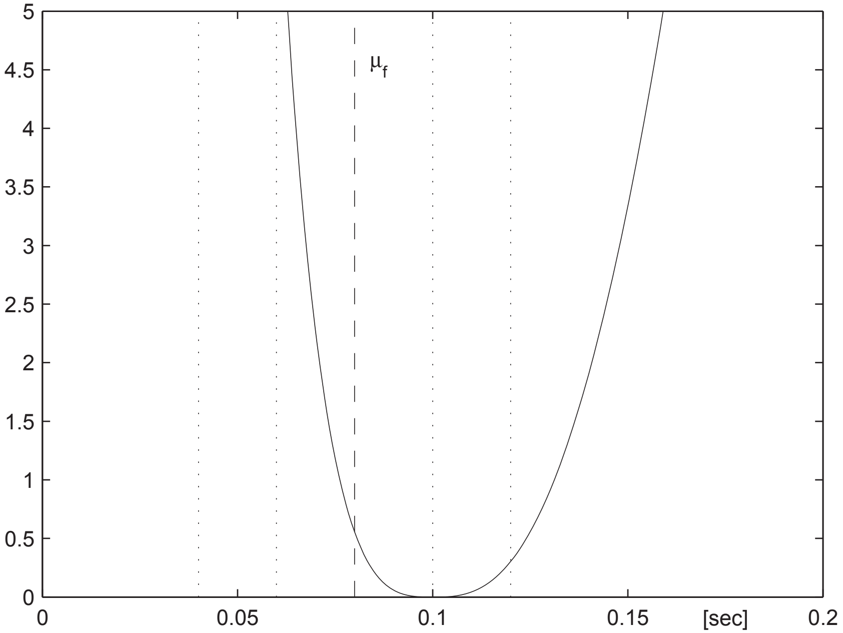

Figure 7.2: *Penalty function for the duration characteristics, plotted on a linear time scale. The vertical dotted lines represent the standard deviations from the mean duration μ_f at 0.08 s, shown by the dashed line. If a log scale instead of a linear scale is used for the duration, the penalty function is symmetric around $\mu_f + \sigma$.*

7.1.5 Voicing Penalty

We used the number of mixed frames and the number of irregularly glottalized frames of a phone as described in Section 6.8 to design two penalty functions. The number of mixed frames N_m and the number of irregularly glottalized frames N_i of a phone with N frames is considered as follows:

$$P(v) = 20\frac{N_m + N_i}{N} \tag{7.6}$$

In this penalty function the number of mixed and irregularly glottalized frames is weighted linearly. A penalty value of 1 is reached if one out of 20 frames is mixed or irregularly glottalized. This penalty function is applied on voiced stationary phones except voiced fricatives. As irregularly pitched frames of neighbouring phones are often a sign of non-standard voice quality of a phone, like creaky or stiff voice, the same penalty function is applied to neighbouring phone instances. As the influence of the irregularly glottalized frames of neighbouring phones however is not as strong as from the irregularly glottalized frames of the current phone, this penalty is weighted with a factor of 0.2. For unvoiced stationary phones, only the number of irregularly glottalized frames N_i, is considered in the penalty function as mixed frames naturally occur in those phones.

7.1.6 Penalty for Plosives

In Section 6.11.1 we presented a heuristic to determine whether the burst of a plosive is distinctive enough. If this is not the case, a high penalty constant is added.

7.1.7 Penalty for Fricatives

As described in Section 7.1.5, for voiced fricatives the penalty function only included the number of irregular and not the number of mixed frames. Otherwise fricatives are treated as regular stationary phones, thus no additional features are applied.

7.1.8 Signal Intensity Penalty

From the experiment in Section 6.9 we have seen that a difference in short-term signal intensity of more than 6 dB is perceived as disturbing. Therefore, we penalise stationary phones with a short-term signal intensity g that differs more than 3 dB from the average intensity μ_g over all instances of this phone:

$$P(g) = (0.2 \cdot (g - \mu_g))^4 \qquad (7.7)$$

7.2 Combination of Penalty Functions

From the above described penalty functions that penalise various phone aspects individually, an overall score has to be derived. As already mentioned, the set of aspects to be applied depends on the type of phone. This is not a problem, because in the context of diphone selection we only have to compare instances of the same phone and not arbitrary phones. Furthermore, the overall score does not have to represent an absolute or even interpretable value. The overall score is only needed to rank instances of the same phone.

As overall score we used the sum of the penalty values resulting from the functions given in Chapter 7.1. No normalisation of these penalty functions was applied as the penalty functions are designed in such a way that the limits of acceptability for each aspect that the penalty functions describe range around the same value.

For the development of the penalty functions we used an interactive tool, which allows diphone instances to be selected from a ranked list and to be used in synthesis. Each diphone can be played in different contexts to subjectively judge its quality, not absolutely but only with regard to the rank order. In this way, we were able to identify aspects that strongly influence synthesis quality and therefore had to be integrated into the phone quality measure.

7.3 Automatic Diphone Set Extraction

Given a speech database, which means a sufficiently large collection of recorded sentences and the corresponding text, the automatic process for diphone extraction comprises the following steps:

1. Phonetic transcription: The phonetic transcription of the text is generated with the SVOX speech synthesiser. This synthesiser allows for various types of outputs, amongst others a phonological transcript that includes the phonetic transcription of the words augmented with abstract prosodic information such as syllable stress level and phrase boundaries.

2. Segmentation into phones: Based on the phonetic transcription, a fully automatic HMM-based segmentation of the speech signals is performed as described in [HP10].

3. Definition of diphone set: The list of all phone transitions in the recordings is extracted from the segmentation. Note that this list cannot be compiled from the phonetic transcription, because only after the segmentation we know which words are separated by pauses. In these cases there is no direct transition from the last phone of a word to the first phone of the next word.

4. Computation of phone scores: The scores of all phone segments are computed as described above[1].

5. Computation of diphone scores: The score of a diphone is based on the scores of the respective phone instances that are combined with a sum of squares. The sum of squares results in higher values for extreme scores. As we take the phone instances with the smallest combined score, extreme values for phone instances are disfavoured.

6. Setting diphone boundaries: A diphone boundary, which should be somewhere in the middle of a phone, is defined as the point with minimal cepstral distance from the corresponding phone centroid. We have found that for robustness reasons the distance measure has to consider several weighted frames. This method is applicable only for stationary phones. For plosives the diphone boundary is set right before the release point.

[1] The computation of these scores requires random access to the data of all phone instances of one complete recording. Randomly accessing this data, which consist of more than 250 MByte per recording, in memory proved to be prohibitive. Therefore, that data was stored in a MySQL database enabling us to use nested SQL queries to speed up the the computation of the scores.

Finally the best-scored instance of each diphone is extracted from the speech signals.

7.4 The Speech Databases

We applied our diphone extraction method to four speech databases: one from a female German speaker (*fg*), one from a male German speaker (*mg*), one from a female British English speaker (*fe*) and one from a male American English speaker (*me*). These databases contained sentences of various length. The overall length of the speech signals were 150 minutes for the female German and the male English database, 85 minutes for the female English and 45 minutes for the male German one. The two English databases and the female German database were recorded recently with professional studio equipment. The male German database was recorded in the early 1980s in the framework of the work described in [Kae85] with analogue equipment and later digitised.

7.5 Evaluation

From the four databases listed in the previous section, we created diphone sets in a fully automatic way as described in Section 7.3. In order to assess the quality of the resulting diphone sets, we used them to synthesise example sentences. In order to exclude possible artifacts in the synthesised speech signal that may originate from weaknesses of other components of the synthesis system, for example from prosody control, we synthesised the example sentences as follows: First, we selected a small set of sentences from each of the four databases. Note that these sentences were excluded from the above described diphone extraction. From these sentences we then extracted the prosody, which comprises the durations and the F_0 values of the phones. This allowed us to use diphone concatenation to generate synthetic speech with natural prosody.

The results from this experiment were as follows: The example sentences produced with the female German and with the female English diphone sets showed very little distortion and sounded quite natural. More distortions were audible in some of the examples from the male English diphone set, others were virtually free of defects. The example sentences from the male German diphone set showed more defects than those from the other three diphone sets, possibly because the size of the male German database, which contains only 45 minutes of speech, is rather limited. We demonstrated our results by means of examples enclosed with [EP11][2].

This evaluation clearly lacks formality. But the linear approach to combine phone quality aspects, which was presented in this chapter, is only an intermediate solution, which is included for its simplicity and effectiveness. In a next step, it is replaced by a machine learning approach, which is presented in the following chapter. From a comparison of example sentences it is clear that this machine learning approach results in much higher quality. Therefore, we performed a formal evaluation only for the machine learning approach.

7.6 Discussion

The quality of the synthesis examples shows that our proposed phone quality measure can be applied to select a high-quality diphone set from a speech database. This simple heuristic method is effective and practicable as it does not require training data and gives reproducible and transparent results. Moreover, weights and parameters can be easily included for example to shift penalty to one aspect that should be especially avoided with a particular voice. Nevertheless, this simple way of feature combination may not entirely represent the perceptual impression as some features may exhibit interdependencies that are not accounted for in this approach.

Thus, a machine learning approach would be desirable to weight and combine the features. However, the approach to create an appropriate data set for the training of such an approach is not obvious: many problems like few training data, unbalanced data and undefined feature values will be tackled in the following chapter.

[2] `http://www.tik.ee.ethz.ch/spr/test_sentences/`

Chapter 8

Combining Phone Quality Aspects With a Machine Learning Approach

The approach presented in Chapter 7 has two serious limitations. First, the penalty functions for the phone characteristics are motivated by acoustic inspection, and second, merely taking the sum of the penalties may not entirely represent the perceptual impression. Therefore, we aimed at replacing these penalty functions by a machine learning approach to weight the features and combine them in a non-linear way.

The first section of this chapter describes the generation of the training data and the difficulties associated with this data. Next, we elucidate the problem of how to classify variable-sized feature vectors and introduce a modification to ANN that can handle this kind of feature vectors. The features that were finally used as inputs to the ANN are motivated and characterised afterwards. Eventually, it will be outlined how the optimal network configuration and parametrisation was determined. A detailed account on this elaborate process can be found in

Appendix C.

8.1 Training Data

The approach to create an appropriate data set for the training of such a machine learning approach is not obvious. As training data semi-diphone or phone instances that are manually classified as suitable or unsuitable for diphone corpora are required. First, the phone instances should be evaluated in a prosodically modified context, because one important aspect of phone quality is suitability for prosodic modification. Then, the scenario should be realistic, in other words, the phone instances should be classified in a similar context as the one they are later used in. Furthermore, all aspects of phone quality (not just spectral quality, for example) should simultaneously play a role in the manual classification. This is a crucial point, because if we judge the particular aspects in an isolated way, we do not gain any information about how to combine these aspects for an overall score, leaving us in the same situation as with the linear approach. Eventually, the manual classification process should be relatively simple and fast to allow the generation of a sufficiently large number of training data. The classification should be reproducible and should allow phone instances to be excluded for one or the other reason.

8.1.1 Generation of Training Data

We extracted the phone sequence and natural prosody from recorded sentences and used randomly selected diphones to resynthesise these sentences. We used the complete recordings from each speaker to create corpora that contained all diphones that can be possibly extracted from these recordings. Natural prosody was deemed advantageous as negative effects that could arise from errors in the predication of artificial prosody were avoided.

We used the four voices listed in Section 7.4 with a set of 10 to 14 sentences per voice and synthesised various versions of these sentences with different randomly selected elements. The author then manually classified individual phone elements (semi-diphones) into the categories suitable and unsuitable. To allow fast and convenient classification, a Wavesurfer plug-in was written that allows listening to the semi-diphones in varying context and shows some information about the semi-diphones. For a detailed description on the number of sentences and number of phones that has been classified for each voice, see Table 8.1. In total, 15,761 phones were manually evaluated, 4,269 were classified as unsuitable instances, 10,879 were classified as suitable instances and the rest (613 instances) was excluded, mostly because the phones were too short for their quality to be properly judged. The author, who evaluated the phones, was very familiar with the four voices and their particular characteristics and weaknesses.

voice	total number of sentences	total number of phones
fg	52	4533
me	30	2868
fe	40	3707
mg	42	4653

Table 8.1: *The number of sentences and number of phones that have been manually classified for each voice.*

The voices that were used to generate the training data cover in fact different aspects of phone quality. The *mg* voice contains many mixed segments, which nevertheless sound acceptable. Also the intensity of the *mg* recordings varies considerably more than for the *fg* and *fe* voices. The *fg* and *fe* recordings contain many irregularly glottalized segments that sound unacceptable if they are used out of their original context. The *me* voice has a high variability what the pronunciation of stationary phones is concerned.

8.1.2 Data Difficulties

Although using prosodically modified diphones as basis for the training data is clearly better than using unmodified diphones, the problem remains that some undesired properties of phones can only be detected through a specific kind of prosodic modification. A semi-diphone may seem suitable within one specific context, however, in a different context it may turn out problematic. E.g. a very short phone may seem suitable in a context where it is used as a short phone, however when it is used as a long phone it turns out to be absolutely unsuitable. The same aspect applies for fundamental frequency, where a diphone with very low F_0 may sound acceptable in a context with low F_0 yet sound disturbing in a context with high F_0. Also diphones with irregularly glottalized periods may sound acceptable within a certain context and turn out to be disturbing in a different context.

The manual classification task required a high level of attention and proved to be rather tedious, as it was necessary to listen very carefully to correctly identify the semi-diphones that were responsible for the artifacts. This was impeded by psychoacoustic effects that made the localisation of unpleasant effects difficult, often it was necessary to listen several times very carefully. We manually double-checked the whole data after training a classifier as described in the remainder of the chapter and investigating contradictions between manual and automatic classification. But still, as a consequence of the difficulties listed above, there were inherent contradictions in the data.

Another downside of this approach, which arose during the classification, was that due to the random selection of the diphone elements, the training data are unbalanced in terms of suitable and unsuitable phones, as for an average voice suitable phone instances outnumber unsuitable phone instances.

8.2 Classification of Variable-Sized Feature Vectors

We distinguish different types of phones (voiced stationary, unvoiced stationary phones, voiced plosives and unvoiced plosives), which have common features like their duration, mean energy or maximum energy but also features that only apply to one particular type, like the phase offset for voiced stationary phones or the centroid distance for stationary phones. For the linear combination approach this did not pose any problem, because for every phone type the scores were computed with different combinations of penalty functions. However, if one single neural network should be used for the suitable/unsuitable classification, we had to cope with variable-sized feature vectors due to actually undefined feature values depending on the type of phone.

A related issue to variable-sized feature vectors is the problem that some of the input values may be missing for some of the input patterns. The simplest solution is to discard those patterns [RM99], if first, the data quantity is large enough, and second, the mechanism which is responsible for the omission of data values is independent of the data itself. However, neither is the case for our data. Actually, *every* input pattern has at least one missing input value, thus there are no patterns which have valid input for all values and we would have to drop all patterns.

The second requirement, that the mechanism which is responsible for the omission of data values is independent of the data, also applies for the common heuristics, to 'fill in' the missing values with, for example, the mean of the corresponding variable over those patterns for which its value is available. In addition, creating artificial data from some more or less elaborate model would on the one hand blur information and on the other hand would content-wise not be intuitive, as we would for example generate artificial burst duration values for all kinds of phones or F_0-derived values for unvoiced phones.

8.3 Gate Neural Networks

To solve the problem of undefined feature values we devised a neural network variant based on back-propagation. The requirement was that non-existent input values should not have any influence in the computation of the output and that any weights that originate from the input node with the undefined input value should not be updated. Conceptually, gates that are connected with every input are either open with no effect if the input feature exists or inhibit the input if the feature does not exist. The inhibition of a particular input i in the forward-propagation phase has the effect that for the weighted sum of inputs

$$a_k = \sum_{j=1}^{n} w_{jk} x_j, \tag{8.1}$$

that is fed into the activation function $g(a_k)$ of every node k in the second layer, the term $w_{ik}x_i$ is excluded. In the back-propagation phase, the weight w_{ik} of the inhibited unit is not updated.

This gate function can be easily implemented by setting the input $x_i = 0$ if the feature is not defined. Note that this step has to happen after normalisation, if any normalisation is applied to the input values. As a consequence, the influence of the inhibited input i in the weighted sum of inputs in Formula 8.1 becomes zero. Also in the back-propagation phase, the weight update $\Delta w_{ij} = -\eta \delta_j x_i$ is zero because $x_i = 0$. Therefore, the weight that originates from the prohibited unit x_i is not modified. This implementation leads to a different interpretation of the gate aspect in combination with normalisation. If we normalise the input values to mean value 0 and set the value of a undefined input to 0, we do not indicate a bias in any direction for that particular undefined input.

Actually, we chose neural networks because they proved to be the only classifier that could be modified to cope with undefined inputs in a suitable way without using one of the approaches listed in Section 8.2. As a proof of concept for this kind of implementation using gate functions, we created several simple classification problems, where for a certain percentage of the input patterns one of the input dimensions was not defined. The Bayes error rate can be computed for the different input dimensions, thus the overall Bayes error rate can be computed from these individual error rates. One of these experiments is given in Section C.1 in Appendix C.

So-called input gates had been used for neural networks before, for example in [MNH98]. However, the objective of input gates in that work is completely dissimilar to our work. In that work, functions, that were called input gates, were used to produce weights for the neural network input values. The input values were multiplied with these weights that take values between 0 and 1, depending on the relevance of the input. The objective of that work was to identify relevant input (in terms of interpretation) to the network and to avoid superfluous inputs, in order to make the network more robust.

Another type of ANN that involves gates is also completely unrelated to the way we used input gates ([JJ94]). In that work, gates were used to treat problems in a divide-and-conquer manner. The input is processed independently by several expert subnetworks and the gates provide weights to blend the results of these expert networks.

8.4 The Neural Network Inputs

8.4.1 Phone Context

The probability distributions of phone properties are strongly influenced not only by the phone type itself, but also by the context of the particular phone. For example, the phone duration is on average higher if a phone is followed by a pause. Therefore, we not only considered different phone types, but also their context for the classification. This context information, which is highly relevant for the classification, is described with predicates, which are used as additional features to the neural network.

We analysed the data described in Section 8.1 for systematic differences between suitable and unsuitable phone instances depending on their context and designed predicates to describe that context. The context information is based on the phonetic information that we obtained from the phone segmentation of the speech data (see Chapter 7.3). In the end, we used the following 14 predicates to characterise phones and their context:

isGlottalClosure Glottal closures tend to be irregularly glottalized, much more than any other phones.

isVoicedFricative Voiced fricatives contain a high share of mixed frames.

isPause Pauses differ from all other phones in that intensity should be as low as possible. High intensity in pauses may be due to noise caused by breathing. In contrast, for all other phones, the intensity of a suitable phone instance should be not too far from the mean intensity to avoid intensity discontinuities.

For pauses, it can be seen from histograms of the data described in Section 8.1 that intensity is an important criterion for quality. Suitable pause instances have significantly lower mean and maximum intensity. On the other hand, for non-pause phones, this is not the case.

isPreplosivePause As for pauses, also for preplosive pauses, the intensity should be as low as possible. High intensity in preplosive pauses may be due to noise like clicks and smacks. Furthermore, often the fundamental wave of the preceding phone (if voiced) is still present at low intensity in the preplosive pause. If duration has to be manipulated for PSOLA, unpleasant effects can occur is these fundamental wave periods are repeated too often.

isPhoneTrill We observed from the data described in Section 8.1 that for trills like the phone [r] more deviation from the means is tolerated in terms of irregularly glottalized frames, mixed frames, centroid distance, and energy.

isPlosive For plosives some stationary features are not defined, like the centroid distance.

isNasal For the phones [n] and [m] and [ŋ] more deviation from the means seems to be tolerated in terms of centroid distance and fundamental frequency.

isFollowedByPause Diphones that have a pause as a second part are used after strong phrase boundaries and therefore tend to be synthesised with longer durations. Therefore, this context has to be known to select phones with a longer duration, which are more suitable in that particular context.

isVoiced This predicate groups the phone instances into voiced and unvoiced phones.

isAffricate Affricates are plosives with a prolonged and stronger period of frication after the release than ordinary plosives. For this stationary fricative part, a centroid distance can be computed, which can be used to characterise the affricate.

isPhoneV As for trills, we noticed from the data described in Section 8.1 that for the phone [v] a higher deviation from the means is tolerated for certain features.

isFollowedByPreplosivePause Phones before a plosive are often influenced by that plosive. We observed that these phones show different spectral properties and a tendency to contain more irregular and mixed frames.

isPhoneSchwa We had the impression that for schwas larger deviations from the centroid are tolerated.

diphonePart This predicate contains the information whether the first or the second part of the phone instance is used in the diphone. This predicate allows a more accurate description of the context. For instance, the information that a phone is followed by a pause has a stronger influence on the duration if the second part of the phone instance in combination with the following pause is used.

8.4.2 Phone Properties

Basically we wanted to use all the information that was used in Chapter 7 while keeping the total number of features and with it the number of weights in the network minimal. We finally chose the following 14 features:

centroidDistance The same cepstral distance is used as described in Section 7.1.1.

nCreaky, nCreakyLeft, nCreakyRight We use the number of irregular frames of the current, preceding and subsequent phone over the corresponding total number of frames (in analogy to Section 7.1.5). Note that for glottal closures irregular frames are not a sign for unsuitable phones as we can see from the data described in Section 8.1. Therefore, the information whether the phone is a glottal closure or not is included as a predicate (see Section8.4.1).

nMixed This feature is defined as the number of mixed frames (see Section 6.8 over the total number of frames. In contrast to other phones, for voiced fricatives, a high number of mixed periods is desirable. Therefore, the information whether the phone instance is a voiced fricative is included as a predicate (see Section 8.4.1).

duration, durationStd To describe duration, we used two features: First, we used the deviation of the log duration from the mean log duration μ_d of all instances of the phone: $\log d - \mu_d$. Second, we used the log of the standard deviation of the mean duration μ_d.

meanIntensity To characterise the mean signal intensity, we used the energy deviation of the phone instance from the mean intensity μ_g over all phones:

$$e_{mean} = g - \mu_g, \tag{8.2}$$

where g is the mean intensity of the phone instance in dB.

maxIntensity This feature is defined in the same way as the previous one, except that we used the maximum intensity instead of the mean:

$$e_{max} = g_{max} - \mu_{g_{max}}, \tag{8.3}$$

where g_{max} is the maximum intensity frame of the phone instance in dB.

plosiveExists We used the same feature as described in Section 6.11.1 to determine plosives with their intensity below some threshold.

f0penalty1 We used the same preprocessed F_0 properties that the penalty functions described in Section 7.1.3 are based on. However, for the neural network we did not apply the exponentiations:

$$P_1(F_0) = |f - \mu_f| \tag{8.4}$$

Note that the logarithm of F_0 is used, which makes the formula equally valid for male and female voices. Therefore, the mean F_0 value over all N frames of a phone instance is $f = \frac{1}{N}\sum_n(\log F_0(n))$.

f0penalty2 The variation of F_0 within a phone instance is expressed with the temporal derivative and results in the feature

$$P_2(F_0) = f' - \mu_{f'} + \sigma_{f'} \tag{8.5}$$

where $f' = \frac{1}{N}\sum_n |\log F_0(n) - \log F_0(n{-}1)|$ is the mean absolute derivative of the F_0 of this phone instance and $\mu_{f'}$ and $\sigma_{f'}$ are the mean and standard deviation of f' over all instances of that phone.

f0penalty3 A second feature to describe the F_0 variation is defined in a similar way:

$$P_3(F0) = \hat{f}' - \mu_{\hat{f}'} + \sigma_{\hat{f}'}\,, \tag{8.6}$$

where $\hat{f}'$ is the mean over the highest 25 % of the components that contribute to f', and $\mu_{\hat{f}'}$ and $\sigma_{\hat{f}'}$ are the mean and standard deviation of $\hat{f}'$ over all instances of that phone.

phaseOffset The phase offset feature is defined in Section 7.1.2 as the position of the pitch marks relative to the phase of the fundamental wave.

In total, the 14 features for the phone context, which are described with predicates, and the 14 features for the phone properties are used as inputs for the neural network.

8.5 Parameter optimisation and network choice

Before training any neural network to classify phone instances into suitable and unsuitable as in the proposed scenario, a number of open questions had to be examined. First, it had to be determined whether our fairly limited data set would suffice to train such a network. After that, we had to determine the best network configuration and optimisation method, and subsequently select the parameters associated with that method.

To resolve the questions mentioned above, we estimated some realistic distribution of the training data to generate synthetic data of arbitrary size for training and a sufficiently large test set that is not biased (for details see Appendix C.2). Based on this data, we were able to estimate an adequate training data size and came to the conclusion that some 15,000 training patterns would suffice to train the network in a way that its error rate would approximate the Bayes error rate (for details see Appendix C.3). What the training configuration was concerned, we needed to decide on how to cope with of heavily unbalanced data (see Appendix C.4.1 to C.4.3), the network layer structure (see Appendix C.4.4 to C.4.5), the target coding scheme (see Appendix C.4.6), and the optimisation approach itself (see Appendix C.4.7). Each optimisation approach by itself has a certain number of parameters to be tuned, where the number of iterations, the learning rate, and the momentum term are the most common.

After this meta-optimisation process, as it is called in [RM99], we decided to use a network with two hidden layers and a layer structure of 28/12/5/2. We used an educational learning method with learning rate adaptation, as described in Appendix C.4.3. For the educational learning method a learning rate factor of $\lambda = 2$ was used, and the training was stopped after 60,000 iterations. The learning rate η_c for the correctly classified patterns was set to the value of 0.005, and no momentum term was used. For further details on the network selection see Appendix C.4.8.

Chapter 9

Evaluation

This chapter presents the evaluation of diphone corpora that we created based on the machine learning approach described in Chapter 8. To measure the quality of that approach, we evaluated two different aspects of quality with two types of listening test. The first aspect was intelligibility, the second aspect was segmental quality.

Intelligibility was determined with rhyme tests. We describe the design and implementation of the rhyme tests and analyse the results on various levels: after showing the total intelligibility rates for each corpus, we drill down to phone positions and their influence on the error rate and finally to the most frequent phone confusions.

Segmental quality was evaluated with a listening test, where the subjects were required to determine the number of artifacts that they perceived in the signal. By this means, we compared the number of artifacts for two corpora, where the diphones of the corpora were selected from the same speech database, but once manually and once automatically.

9.1 Evaluation Criteria

Until now, no objective methods exist to derive the quality of a speech signal from the signal itself. So far, the only possibility to evaluate speech quality and thus to evaluate our selection method are subjective listening tests. In those tests, the quality of a speech signal can be measured using different criteria. The most commonly used are intelligibility, naturalness, pleasantness, pronunciation, articulation, listening effort, audio flow and acceptance. However, many of these criteria depend on factors that are beyond the selection of synthesis elements.

Naturalness, listening effort and audio flow strongly depend on prosody. Prosody control was not part of this study and furthermore, the large number of non-native subjects in this evaluation would have made it difficult to receive reliable results (see [JC10]). Pronunciation depends on the text analyser whereas pleasantness is strongly related to voice itself. In order to measure acceptance, a clear application scenario and target group would have to be defined, which was not the case for our study. The remaining criteria intelligibility and articulation are influenced most by the selection of synthesis elements, which amounts to segmental quality. Thus, we restrained to test for the criteria intelligibility and segmental quality. In Section 9.3, rhyme tests as intelligibility tests are discussed and presented and in Section 9.4 the segmental quality test is presented.

9.2 Creation of the Corpora

For the evaluation we created four diphone corpora from the four speech databases listed in Section 7.4. We used the network that was deemed to perform best (see Section C.4.8) to score the phones and created the diphone corpora as described in Section 7.3.

9.3 Intelligibility Tests

There are various ways to test intelligibility[1]: closed response set tests encompass articulation tests with nonsense monosyllabic words (logatoms) and rhyme tests with monosyllabic words. Tests with an open response set, where listeners have to write down what they think they heard may employ isolated words like in the Bellcore corpus (see [SAMW89]), meaningful sentences (semantically predictable sentences, SPS) or anomalous sentences (semantically unpredictable sentences, SUS).

Also opinion tests like the MOS test [VV05] are used to evaluate intelligibility, normally on a sentence level with SPS [Gol95]). However, opinion tests do not test intelligibility directly but rather ask the subjects to rate intelligibility on a scale. Therefore, they do not provide any diagnostic value and suffer the usual problems with interval metrics like the influence of the number of categories (5 or 11) and the choice of circuit conditions, that means, the quality range of the systems that are evaluated influences the evaluation results. Moreover, extensive training for the subjects would be necessary to convey the categories, and finally the demands placed on the listeners are much higher compared to direct testing.

9.3.1 Rhyme Tests

Originally, rhyme tests had been designed to measure the quality of signal transmission (see [Sot82, Hou63]). In the meantime, rhyme tests are widely applied to evaluate intelligibility of synthetic speech (see [LGP89, BT05]). Rhyme tests encompass word lists of well-known, normally monosyllabic words, which are organised in sets of 6 related words, so-called ensembles. As a rule, the words are of the form consonant-vowel/diphthong-consonant (CVC); some few words take the form CV or VC. The words in each set differ from each other only in one phoneme, either in the initial or the final consonant or in the vowel that is the nucleus of each word. When the test is given, one word from each ensemble is played to the subject as acoustic stimulus. The subject then marks which word he thinks he heard on a multiple-choice answer sheet offering all 6 words in each set.

[1] The term intelligibility test refers to the correct recognition of words, whereas the term articulation test refers to the recognition of phonemes or logatoms (see [Gol95]).

One example of an ensemble in which each word contains the vowel [i] and ends with a consonant [t] is

meat, feat, heat, seat, beat, neat

We decided to evaluate intelligibility with a rhyme test with a closed response set because:

- Results by [Sot82, Hou63, Kae85] showed that groups of 10 to 20 persons already deliver reliable results.
- Spelling of logatoms may vary from subject to subject, there is no authoritative spelling, especially for English, therefore an error-prone interpretation step would be required.
- A closed set test takes less effort and time as the subjects do not need to write down what they perceived.
- Rhyme tests have a diagnostic value, since they can reveal phonetic features that cannot be discriminated enough in a confusion matrix.
- Unlike for SPS and SUS, prosody plays a very minor role in intelligibility.
- Rhyme tests place low metric demands on the listener (they do not have to sort words or assign scores).
- A low mental effort is required, there is only a small load on short-term memory, compared to SUS or SPS.

9.3.2 Rhyme Test for German

The German rhyme test created by [Sot82] helps to identify confusion of consonants as well as confusion of vowels. The test words are phonetically balanced, which means that the phone frequencies correspond to the mean frequencies in German. As a consequence, the intelligibility results are representative for those that are achieved from synthesising common German text. One exception, however applies. The phone schwa is not represented in the test words as it does not exist in monosyllabic words. The test encompasses 100 ensembles with 6 words each, where 34 of them differ in the initial consonant, 33 in the final consonant and the remaining 33 in the mid-word vowel. We generated the test words for the two German voices, the female German *fg* voice and the male German *mg* voice. The diphone corpora for these two voices were created as described in Section 9.2. The prosody of the test words was created with the German prosody control of the SVOX speech synthesiser, version 5.

9.3.3 Rhyme Test for English

For the English voices we used the Modified Rhyme Test (MRT) (see [Hou63]). The MRT uses 50 ensembles of 6 words, where 25 ensembles differ in the initial consonant and 25 ensembles in the final consonant. Mid-word vowels and consonant clusters are not tested, so the MRT allows only consonant combinations to be identified that are hard to discriminate. We generated the test words for the two English voices, the female British English *fe* voice and the male American English *me* voice. We observed that the English test partly checks for very subtle differences, for example as in the pairs "piece"/"peas", "save/safe" or "vest"/"west", that may be hard to identify for non-natives. The diphone corpora for these two voices were created as described in Section 9.2, the prosody of the test words was created with the English prosody control of the SVOX speech synthesiser, version 5.

9.3.4 Implementation of the Tests

A tool with a graphical user interface was implemented to present the test. The tool offered written instructions for the rhyme test, so all the subjects were presented the same information. Furthermore, the subjects could take a trial test, repeatedly if desired, to familiarise with the procedure.

After the subject has chosen one of the voices for the rhyme test, each of the ensembles is presented in the same way:

- First, an acoustic stimulus is presented, which is randomly chosen from the ensemble of 6 words.
- Not till after the stimulus is finished playing, the response set of the six words is presented, also in random order.
- The subject marks his choice by selecting a radio button (with a mouse click or keyboard shortcut).
- By pressing a *Next* button, the test proceeds to the next ensemble.

The audio signals were presented over headphones in a quiet but not specially sound insulated room. The test setting ensured that the subject could hear each stimulus only once and could not go back to modify his choice. During the rhyme test, the subject could pause any time by marking his choice but postponing to press the *Next* button.

For the test, 5 female and 11 male subjects volunteered kindly enough without receiving any remuneration. The age of the test subjects was 32 years on average, ranging from 20 to 61. Of the test subjects 13 were native speakers of German (3 Swiss and 10 standard German), one subject was native Persian, one native Italian and one native Spanish speaker. For the Spanish speaker, the test results for all rhyme tests were excluded, and for the Persian speaker the test results for the German rhyme tests were excluded due to insufficient language proficiency.

9.3.5 Results

If a closed response set of m choices is used for a test, a minimum accuracy of $1/m$ can already be achieved by guessing. The conventional scoring formula to correct for chance success is:

$$v_c = \frac{mv_r - 1}{m - 1}, \tag{9.1}$$

where v_c is the corrected accuracy, and the raw accuracy v_r is:

$$v_r = \frac{\text{number of correct answers}}{\text{number of all answers}}. \tag{9.2}$$

In the following, the corrected version of the intelligibility rate will be used. The overall intelligibility rates for the four voices are shown in Fig. 9.1. In Fig. 9.2, the intelligibility rates are displayed with respect to the phone positions. For the German voices, intelligibility rates for initial consonants, middle vowels and final consonants are shown, for the English voices, initial consonants and final consonants are distinguished. This more detailed illustration reveals that for the German voices the middle vowels and for the English voices final consonants are most problematic. A detailed analysis of the phone confusions for all voices is listed in Appendix D.

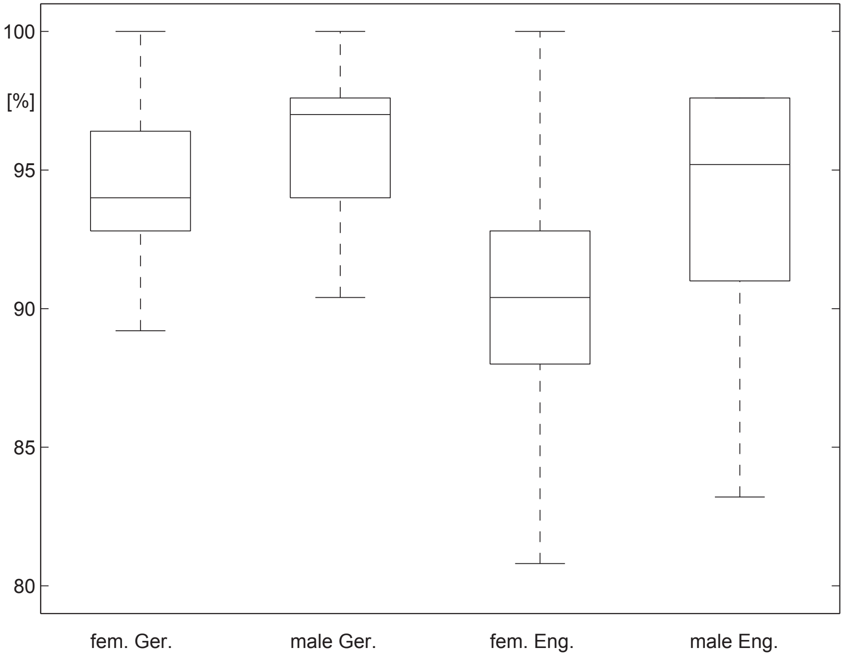

Figure 9.1: *Intelligibility rates of the four voices, corrected for chance.*

As can be seen from Fig. 9.1 (and in more detail in Table 9.1), results for English voices are considerably worse than for German. This may be partly due to the test subjects' insufficient knowledge of the subtleties of English. There was no native English speaker among the test subjects and several users reported problems about either not knowing many of the English words used in the test at all and about not knowing the pronunciation differences of word pairs that were presented as alternatives, like safe/save or vest/west.

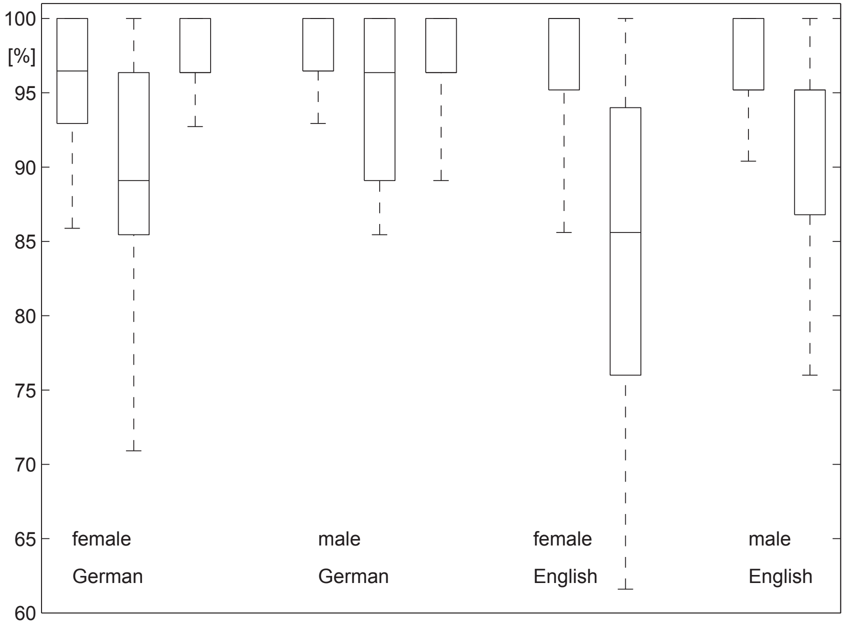

Figure 9.2: *Position-wise intelligibility rates of the four voices, corrected for chance. For each of the German voices, intelligibility rates for initial consonants, middle vowels and final consonants are shown. For the English voices only initial consonants and final consonants are shown.*

For the female German *fg* speaker, many confusions occur between the vowels [ɛ], [e] and [ɛː], and also [iː] and [ɪ] to a minor extent (see confusion matrix in Table D.2 in Appendix D). The relatively high number of confusions is probably due to the high inconsistency of the speaker's pronunciation of the open and closed [ɛ]/[e]. Many, but not all words with an open [ɛ] or [ɛː] are actually pronounced as a closed [eː], as it is common in the Northern variants of German. As a con-

voice	accuracy
fg	94.26 % (2.95)
mg	96.14 % (2.71)
fe	90.56 % (5.40)
me	94.08 % (4.04)

Table 9.1: *Mean intelligibility rates for all voices in percent (standard deviations in parenthesis).*

sequence, the centroids for the phones [ɛ]/[ɛː] become more vague as they are influenced by the many [eː] that are labelled as [ɛ]/[ɛː]. Therefore, in the selection process, both open and closed versions of [ɛ]/[e] are selected. This causes two effects. First, words that are transcribed with an [ɛː] are actually pronounced with an [eː], like for example in the word ”rät”, which is transcribed with an [ɛː], but was confused several times with the word ”Reet”, which is pronounced with an [eː]. Interestingly enough, confusion also occurs the other way round. For example the word ”Scheel”, pronounced with an [eː], was taken for the word ”schäl”, pronounced with an [ɛː], several times. One explanation could be that the listeners became more tolerant to the notion of that speaker's [ɛː] and therefore chose the word “schäl”, which is, in addition, more common than the very rare word “Scheel”. The second effect occurs if a closed and an open version of the phone [ɛ]/[e] are concatenated. In this case, often the impression of the diphthong [ai̯] is created. We observed this effect in the test words and it also shows in the confusion matrix in Table D.2 where some [ɛ] and [ɛː] are taken for an [ai̯].

The *mg* voice performed best concerning the total intelligibility rate, which may be due to the very consistent speaking style of the voice talent. Nevertheless, some minor confusions occurred. As in the *fg* voice, the phone [eː] is sometimes taken for an [ɛ], [ɛː], [iː] or [ɪ], but to a much lesser extent (see Table D.5). Furthermore, the final [ŋ] is sometimes confused with an [n] (see Table D.6). This problem can be traced to one badly selected diphone [ŋ/], which exhibits only a very slight nasalisation. Furthermore, the affricate [t͡s] is confused a few times with [t$^{\mathrm{h}}$]. There is no obvious explanation for this confusion as the fricative part of the affricate is clearly audible in the affected test words.

The *fe* voice was the voice that performed worst. Most problems occurred with the final consonants (see Table D.7), but there were also confusions of initial consonants (see Table D.8). The initial [h] was sometimes taken for a [p], which may be explained with spectral mismatches at the concatenation point that can introduce a plosive effect. Furthermore, the voiced fricative [v] is sometimes taken for a [w]. However, this confusion occurs with only one word pair ”vest”/”west”, where the the voiced fricative in the word “vest“ is clearly audible. This confusion can be explained with the test subjects' native German background, as in German the word “West“ is pronounced with a [v] exactly as the English word “vest“ and many of the subjects stated not to have been aware of this fact. More problematic than the intelligibility of the initial consonants is the intelligibility of the final consonants, where often voiced and unvoiced phone variants were mixed up, like in the words pairs ”peas”/”peace” or ”bad”/”bat”. We found that the *fe* speaker tended to pronounce final voiced plosives very strongly and tended to devoice voiced plosives. The speaker compensated this devoicing with a lengthening of the preceding vowels. However, this strong pronunciation of the voiced plosives made them almost impossible to distinguish from their unvoiced counterparts if perceived in isolation. In contrast, the American English speaker articulates voiced plosives much more softly. However, for synthesising the words for the rhyme tests, for both voices a prosody control was used that does not differentiate the durations of the preceding vowels very much depending on the voicing of the following consonant as it was trained on an English prosody corpus recorded by a German speaker. Thus, duration differences were only as little as 15 % for the vowels of word pairs like ”bad”/”bat”. Consequently, the words had to be distinguished mainly by the voicing of the final consonant. This made it particularly hard for the test subjects with non-native background which are not trained to the subtleties of the English pronunciation.

The intelligibility results of the *me* voice are comparable to those of the *fg* voice. Concerning initial consonants, similarly to the *fe* voice, the [h] was confused with the plosives [p] and [b], presumably for the same reasons (see Table D.9). In analogy to the *fe* voice, the voiced/unvoiced discrimination of final consonants was problematic also for the *me* voice.

The confusions were not as pronounced as for the *fe* voice. One reason for that could be, in contrast to the *fe* speaker, that the *me* voice talent pronounced voiced plosives very softly, which made the voiced variants easier to identify.

9.4 Segmental Quality Test

A diphone corpus with manually selected elements and one with automatically selected elements that were extracted from the same recordings were available to us. This enabled us to compare the two corpora and to therefore compare the quality of a manually created corpus to the quality of an automatically created one. The automatically created corpus was the one created from the *mg* voice, that was created as described in Section 9.2. The reference corpus had been created in 1985 by manual selection of diphones (see [Kae85]) from the same recordings (see Section 7.4). For this corpus, the diphones were extracted interactively considering visual and auditive criteria and, in addition, frequency domain information that had been extracted from the signals to identify the optimal cut points. This manual extraction process comprised several months of tedious work.

For the test, a list of 100 isolated words and short groups of words was synthesised with each corpus. During the test, the subject were presented with 100 acoustic stimuli, where every stimulus was randomly chosen from one of the two corpora. The subjects then marked the number of artifacts they perceived, distinguishing between weak, medium and strong artifacts. These categories were defined as follows:

- **Weak:** The artifact(s) can be perceived in the signal as unnatural. However, they do not sound disturbing and do not affect intelligibility.
- **Medium:** The artifact(s) are perceived in the signal as disturbing. They do sound disturbing but do not affect intelligibility.
- **Strong:** The artifact(s) are perceived very strongly in the signal. They sound disturbing and do affect intelligibility.

9.4.1 Implementation of the Test

The same GUI tool described in Section 9.3.4 was used to present the segmental quality test. Again written instructions and a trial test were offered to familiarise the subjects with the test. During the test, each of the test words was presented in the same way.

- First, the test word from the list was randomly chosen from one of the two corpora and played to the subject
- The subject could mark how many artifacts of the categories *weak*, *medium* and *strong* he perceives. During that time, the subject could listen to the stimulus repeatedly.
- By pressing a *Next* button, the test proceeded to the next ensemble

The test conditions were the same as described in Section 9.3.4, except that the subjects were able to listen to the stimuli as often as they liked.

9.4.2 Results

The distribution of weak, medium and strong artifacts for the corpora with manually and automatically selected elements is shown in Fig. 9.3. The corpus with automatically selected elements performs better in all three categories. The difference is most pronounced for strong artifacts, where the number of artifacts for the corpus with automatically selected elements is only about half of those for the one with manually selected elements. The detailed results are shown in Table 9.2. The relatively high number of weak artifacts for the corpus with automatically selected elements may be caused by the clearly audible background noise that is present in the recordings that we used. The diphones from the corpus with manually selected elements do not exhibit this noise, probably some kind of noise reduction had been applied. At least one subject marked this noise, which is clearly audible especially in the silence segments, consistently as a minor artifact.

	weak	medium	strong
manually sel. elements	0.44 (0.31)	0.24 (0.31)	0.07 (0.12)
automatically sel. elements	0.35 (0.24)	0.16 (0.29)	0.04 (0.05)

Table 9.2: *Mean number of weak, medium and strong artifacts per sample for the corpora with manually and the automatically selected elements (standard deviations in parentheses).*

9.5 Discussion

For the *mg* corpus, the intelligibility rates achieved with the corpus with automatically selected elements are comparable to those achieved with a corpus from the same voice where the elements had been selected manually, as is is presented in [Kae85]. In that work, each of 15 subjects took the rhyme test 20 times. Over the course of these 20 listening sessions a significant learning effect was observed, the intelligibility increased from 90.1 % at the first session to almost 97.5 % at the last session. The existence of such a learning effect was also observed by [Sot82]. Our subjects took the test only once and achieved a mean intelligibility rate of 96.14 % for the *mg* corpus, which is already some 6 % higher than the intelligibility rate reached in the first session of the evaluation in [Kae85].

Also the intelligibility rates of the other voices were constantly higher than the intelligibility rate achieved for the corpus with manually selected elements in the first session, despite the fact that the English corpora were more difficult to evaluate for the non-native test subjects than the German corpora. We refrained from multiple test sessions out of consideration for the test subjects, but based on the results of [Kae85] and [Sot82], we could also assume a learning effect if multiple sessions were undertaken.

The intelligibility tests revealed some weaknesses of the corpora. The voiced/unvoiced distinction of final consonants especially for the female English *fe* speaker is not always easily perceivable. Besides the fact that subtle differences in word-final voicing are difficult to identify for untrained non-native speakers, the higher intelligibility of the male American English voice indicates that this probably was not the only reason. We found that the *fe* speaker tends to devoice final voiced plosives and compensates for this de-voicing with a lengthening of the preceding vowels. However, the durations of the preceding vowels were not differentiated accordingly by the duration control we used, thus making the voiced/unvoiced distinction quite difficult. An important point to conclude is that the prosody control must match the voice to avoid such inconsistencies.

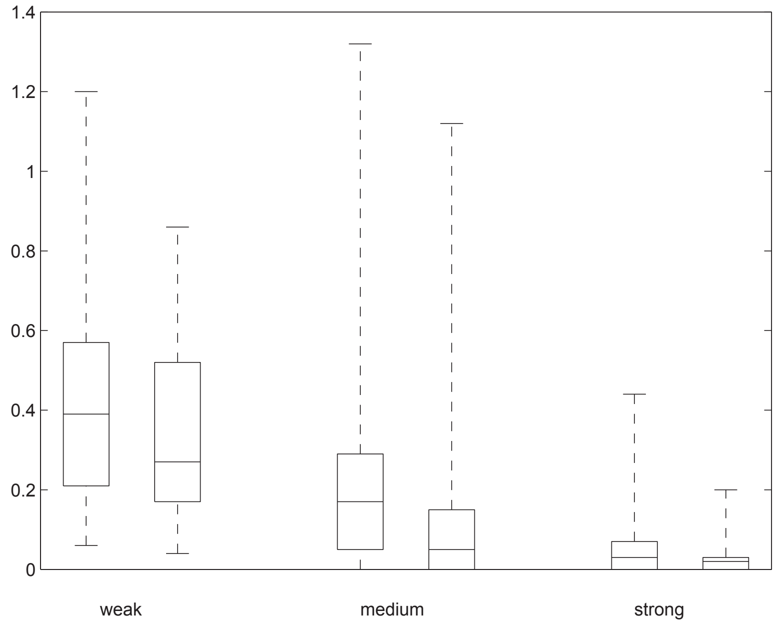

Figure 9.3: *For every subject the distribution of weak, medium and strong artifacts per sample is shown. The left boxplots describe the corpus with manually selected elements and the right boxplot the one with automatically selected elements.*

Unfortunately, we could not compare our results directly to those from the Blizzard Challenge evaluations. Although for these evaluations a MRT is used to compute a word error rate (WER), this MRT uses an open response set instead of a closed response set. In this case, the subjects have to type in what they have heard (see [BB06]), what makes the test considerably more difficult. Furthermore, the WER listed in the Blizzard Challenge are based on two tests, the MRT and a semantically unpredictable sentences (SUS) test. As a consequence, the WER in these tests range from around 10 % for the best systems to more than 50 %.

Chapter 10

Conclusion

10.1 Discussion

The primary aim of this thesis was to devise a phone quality measure that should not only consider spectral phone quality but should also include further characteristics to ensure that a high quality phone is clearly articulated, unambiguously identifiable, and that the signal of the phones is suitable for prosodic modification.

We proposed a phone quality measure that is based on various constituents. Besides spectral characteristics, the quality measure includes numerous properties that are based on fundamental frequency, duration, pitch marks and voicing characteristics. This phone quality measure was applied to create four diphone corpora from different voices in a fully automatic way. These four corpora were evaluated for intelligibility with rhyme tests, one of the corpora was additionally evaluated for segmental quality. For the speech database where a corpus with manually selected elements was available for comparison, the corpus with automatically selected elements reached higher intelligibility rates. Indeed, also the segmental quality of the automatically created corpus was perceived as higher compared to the the manually created one. For the other voices, intelligibility rates were also comparable. These results show that our proposed phone quality measure can be applied to select a high-quality diphone set from a speech database. As a consequence, tedious manual work in the creation of such diphone sets can largely be eliminated.

This study mainly addresses the question of how to qualify phones from existing recordings. One exception is the proposed loudness measure, which can be used for online monitoring of recordings. However still, the quality of the recordings heavily influences the corpus quality, which is also reflected in the evaluation results. The recordings for the male German voice, which performed best in the intelligibility test, are very consistent, the speaking style is very homogeneous, and almost no variants are used. It is still an open research question how to effectively monitor speech recordings to avoid pronunciation variants.

The effects of pronunciation variants can be observed in the evaluation of the female German *fg* voice, where the pronunciation of the closed/open [ɛ]/[e] varies considerably. As a consequence, the vowel intelligibility for these vowels is considerably reduced. Similarly, the American English *me* voice varies strongly in the pronunciation of vowels, as for many words there is no authoritative pronunciation and thus several transcriptions are possible. In this case, a complete avoidance of pronunciation variants during the recordings may be difficult and may constrain the recordings too much. An alternative solution could be to identify pronunciation variants in the segmentation step. How this can be implemented effectively and which restriction have to be applied to successfully detect these variants is still an open question.

A problem with pronunciation variants, which is even harder to solve is the detection of short diphthong hints in vowels caused also by pronunciation variants. In some variants of English, for example in the word “book” the vowel [o] is followed by a short hint of an [A̯], so the word is pronounced like [boA̯k], with a very short [A̯]. If the segment that contains the pure [o] is long enough, the phone instance is considered suitable in terms of spectral quality. In diphone concatenation, however, this short hint of a different phone can cause considerable disturbance if taken out of context. It is still an open question of how to reliably determine these very short variations and how to tell them apart from ordinary coarticulation effects.

In the post-processing and signal analysis step, minor improvements can still be added. One could be to measure softness and voicing of voiced plosives. This is especially important for speakers that tend to devoice voiced plosives and compensate for this devoicing with a lengthening of the preceding vowels. The effect of ambiguous voicing of plosives, however, can be alleviated if the prosody control is fully compatible with the voice and also marks the voicing differences prosodically in the way the speaker does.

Certain difficulties caused by pronunciation variants or by restricted context, however, are limited to diphone synthesis. E.g. if a diphone element consisting of a phone to silence transition is selected from continuous speech, this element must almost certainly be selected from the end of a sentence. Typically this diphone element has low energy, including the first part. If this diphone element is concatenated with another element that is spoken with average intensity, intensity jumps are unavoidable. This problem, however, cannot be solved due to the limited context of diphones.

10.2 Outlook

Although this work concentrated on the selection of diphone elements, our phone quality measure is not limited to diphone synthesis. We believe that other synthesis methods can also benefit from our phone quality measure, as well as from the proposed methods for signal analysis and modification. In unit selection, the phone quality measure may be used as a criterion in database pruning to reduce the size of the system, and in the selection step as an additional feature for the target costs of candidate units. Also in the post-processing of speech recordings, the quality measure can be used to identify low quality segments, labelling errors or pronunciation variants. This may be even useful for synthesis methods other than concatenative speech synthesis. This preselection of training material could perhaps be used to improve statistical parametric speech synthesis.

The precise pitch marking method could be used in combination with the proposed extension to TD-PSOLA to synthesise expressive speech, even to synthesise emphatic accents, without any perceivable artifacts. Many application would benefit greatly from improved expressive speech, most notable reading applications for long texts, like books.

Another application scenario for this quality measure would be a semi-automatic tool, that could be used in corpus development. The users could re-rank the list of segments that are proposed by the algorithm if some negative phone property is not sufficiently weighted. This re-ranking information could be used to extend the training data and to re-train the gate neural network in order to self-improve the method while it is used.

Appendix A

Evaluation of Spectral Distance Measures

This appendix gives the detailed results of the spectral distance evaluations presented in Section 6.4 and a detailed investigation of the false positives that are still produced if the best measure is applied.

A.0.1 Evaluation without Cost Ratio

If we consider the number of misclassifications based on the a posteriori probabilities $p(C_k|\mathbf{x})$, where $p(C_k, \mathbf{x}) = p(C_k|\mathbf{x})p(\mathbf{x})$ and $p(\mathbf{x})$ is common to all terms, we receive the classification results given in Tables A.1 to A.4. These tables clearly show that if we consider only the number of misclassifications without applying a cost ratio to reduce the number of false positives, the error rates are quite low.

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	10.29 %	18.14 %	10.27 %	10.51 %
straight	13.31 %	26.14 %	13.38 %	12.60 %
lsf	17.85 %	25.87 %	16.97 %	16.59 %
lpc	16.09 %	14.04 %	15.30 %	15.24 %

Table A.1: *Mean error rates for all phone instances for the female German voice (fg)*

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	10.60 %	23.28 %	12.12 %	11.41 %
straight	15.31 %	24.57 %	15.68 %	15.46 %
lsf	18.12 %	32.64 %	17.67 %	17.53 %
lpc	16.95 %	14.50 %	16.77 %	15.66 %

Table A.2: *Mean error rates for all phone instances for the female English voice (fe)*

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	8.92 %	14.48 %	10.63 %	9.10 %
straight	10.58 %	19.77 %	11.89 %	11.79 %
lsf	12.99 %	18.61 %	13.10 %	12.96 %
lpc order 28	13.38 %	8.11 %	11.69 %	11.81 %
lpc order 20	13.61 %	15.48 %	11.87 %	12.14 %

Table A.3: *Mean error rates for all phone instances for the male German voice (mg)*

	d_{eu}	d_{SKL}	d_{ma}	d_{abs}
mfcc	11.58 %	16.35 %	11.97 %	11.16 %
straight	14.31 %	27.11 %	14.56 %	14.30 %
lsf	17.56 %	22.54 %	16.66 %	17.21 %
lpc	19.60 %	19.22 %	16.89 %	17.22 %

Table A.4: *Mean error rates for all phone instances for the male American English voice (me)*

A.0.2 Error Analysis

We analysed the false positives produced by the Euclidean distance on MFCC, which proved to be the most suitable measure, and give the results in Tables A.5 to A.8. The false positives in the first 6 columns do not affect synthesis quality (they are either articulated correctly (except the mislabelled phone instances in column 6) or exhibit unsuitable characteristics besides spectral characteristics which would lead to high penalty scores) whereas the false positives in the last three columns may affect synthesis quality.

phone	number of false positives	too short for proper auditory impression	short, otherwise OK	high/low F_0, otherwise OK	creaky/breathy/noisy, otherwise OK	coarticulatory influence on parts of phone, otherwise OK	phone completely OK (mislabelled as bad)	same phone but articulated too open/closed	colouring of different phone	completely different phone (misclassified)	other
[œ]	1							1			
[ŋ]	1		1								
[m]	0										
[ɐ]	1									1	
[ɪ]	0										
[aː]	0										
[l]	0										
[ɛː]	0										
[ʏ]	1									1	
[iː]	0										
[øː]	0										
[ʊ]	1							1			
[ɛ]	1								1		
[eː]	0										
[ɔ]	1								1		
[oː]	0										
[ç]	0										
[uː]	0										
[ə]	2								2		
[x]	0										
[r]	2										
[o]	2								2		
[v]	1									1	
[yː]	1								1		
[n]	0										
[ãː]	0										

phone	number of false positives	too short for proper auditory impression	short, otherwise OK	high/low F_0, otherwise OK	creaky/breathy/noisy, otherwise OK	coarticulatory influence on parts of phone, otherwise OK	phone completely OK (mislabelled as bad)	same phone but articulated too open/closed	colouring of different phone	completely different phone (misclassified)	other
[e]	1								1		
[l]	2	1								1	
[ɔː]	1							1			
[iː]	0										
[ɑː]	0										
[ʌ]	1			1							
[ʊ]	1								1		
[ɑ]	0										
[ŋ]	0										
[r]	1									1	
[æ]	0										
[ɪ]	0										
[ə]	0										
[v]	2									1	
[n]	1										
[uː]	0										
[w]	0										
[m]	0										
[ɜ]	1									1	

Table A.6: *Error analysis for the spectral classification with the Euclidean distance on MFCC and a cost ratio of $r = 10$ applied on all stationary phone instances of the fe voice.*

phone	number of false positives	too short for proper auditory impression	short, otherwise OK	high/low F_0, otherwise OK	creaky/breathy/noisy, otherwise OK	coarticulatory influence on parts of phone, otherwise OK	phone completely OK (mislabelled as bad)	same phone but articulated too open/closed	colouring of different phone	completely different phone (misclassified)	other
[ɛ]	0										
[ʊ]	0										
[eː]	0										
[m]	0										
[l]	0										
[ç]	1										
[ə]	1									1	
[ɔ]	0										
[e]	1									1	
[œ]	0										
[ŋ]	0										
[iː]	0										
[i]	1							1			
[ɐ]	1							1	1		
[ɪ]	0										
[aː]	1								1		
[øː]	0										
[a]	0										
[oː]	0										
[ɛː]	0										
[r]	1									1	
[ʏ]	1									1	
[o]	0										
[v]	1									1	
[n]	0										
[uː]	0										

phone	number of false positives	too short for proper auditory impression	short, otherwise OK	high/low F_0, otherwise OK	creaky/breathy/noisy, otherwise OK	coarticulatory influence on parts of phone, otherwise OK	phone completely OK (mislabelled as bad)	same phone but articulated too open/closed	colouring of different phone	completely different phone (misclassified)	other
[ə͜r]	1									1	
[ʌ]	0										
[l]	1									1	
[ʊ]	0										
[ə]	1									1	
[ɔ]	1							1			
[iː]	0										
[ŋ]	1			1							
[o]	0										
[i]	1								1		
[ɜ͜r]	1									1	
[ɪ]	1									1	
[n]	1										
[r]	2									2	
[m]	0										
[uː]	1								1		
[ɛ]	0										
[v]	1									1	
[w]	1									1	
[ɑ]	0										

Table A.8: *Error analysis for the spectral classification with the Euclidean distance on MFCC and a cost ratio of $r = 10$ applied on all stationary phone instances of the me voice.*

Appendix B

Preplosive Pauses and Elisions

In order to decide if separate diphone elements are required for preplosive pauses of elisions, we examined the durations of preplosive pauses depending on whether an elision follows or not. Detailed statistics on these durations are presented in this Appendix.

Tables B.3 to B.2 present the results that clearly show a strong lengthening of the preplosive pause if an elision follows. We introduced additional symbols to denote preplosive pauses and preplosive pauses before an elision (see Appendix E).

preplosive pause	elision (yes/no)	mean duration [ms]	lengthening	# of phone instances
p̣	no	48.63	-	1918
p̤	yes	69.84	+43.61 %	4
ṭ	no	41.69	-	27223
t̤	yes	42.18	+1.16 %	22
ḳ	no	50.25	-	3290
k̤	yes	63.66	+26.67 %	20
ḅ	no	44.31	-	3373
b̤	yes	-	-	0
ḍ	no	78.73	-	7839
d̤	yes	-	-	0
g̣	no	41.28	-	3341
g̤	yes	-	-	0

Table B.1: *Mean duration of preplosive pauses for the female German voice (fg). The lengthening in case of a subsequent elision compared to no elision is given in percent. The mean duration of the [t̤] is actually reduced if the elision is followed by a [d]: the mean duration for [t̤] if followed by [d] is 35.71 ms. This is the case for 19 out of 22 phones, thus there is almost no difference in duration between [ṭ] and [t̤]. No elisions exist for voiced plosives as elisions only take place at word junctions and plosives are devoiced in terminal position according to German pronunciation rules.*

preplosive pause	elision (yes/no)	mean duration [ms]	lengthening	# of phone instances
p̣	no	85.47	-	1494
p̤	yes	194.67	+127.75 %	3
ṭ	no	64.53	-	32055
t̤	yes	81.13	+25.73 %	112
ḳ	no	69.18	-	3510
k̤	yes	108.43	+56.73 %	7
ḅ	no	67.40	-	3266
b̤	yes	-	-	0
ḍ	no	71.53	-	9286
d̤	yes	-	-	0
g̣	no	51.67	-	3497
g̤	yes	-	-	0

Table B.2: *Mean durations of preplosive pauses for the male German voice (mg). The lengthening in case of a subsequent elision compared to no elision is given in percent. No elisions exist for voiced plosives as elisions only take place at word junctions and plosives are devoiced in terminal position according to German pronunciation rules.*

preplosive pause	elision (yes/no)	mean duration [ms]	lengthening	# of phone instances
p̣	no	59.84	-	1644
p̤	yes	108.21	+80.84 %	18
ṭ	no	37.13	-	6119
t̤	yes	83.07	+123.70 %	43
ḳ	no	50.76	-	2243
k̤	yes	75.00	+47.75 %	41
ḅ	no	62.79	-	1726
b̤	yes	111.90	+78.22 %	1
ḍ	no	36.60	-	3631
d̤	yes	74.08	+102.38 %	49
g̣	no	47.32	-	635
g̤	yes	81.56	+72.37 %	4

Table B.3: *Mean durations of preplosive pauses for the female British English voice (fe).The lengthening in case of a subsequent elision compared to no elision is given in percent.*

preplosive pause	elision (yes/no)	mean duration [ms]	lengthening	# of phone instances
p̣	no	51.87	-	4063
p̤	yes	-	-	0
ṭ	no	40.16	-	14649
t̤	yes	-	-	0
ḳ	no	46.33	-	5911
k̤	yes	60.55	+30.71 %	1
ḅ	no	61.10	-	3799
b̤	yes	-	-	0
ḍ	no	38.38	-	8572
d̤	yes	-	-	0
g̣	no	50.39	-	1086
g̤	yes	-	-	0

Table B.4: *Mean durations of preplosive pauses in milliseconds for the male American English voice (me). The lengthening in case of a subsequent elision compared to no elision is given in percent. Elisions were only transcribed in one case.*

Appendix C

Gate Neural Network Experiments

This appendix presents experiments with synthetic data that were used first, to determine the network configuration and parameters for classification with variable-sized feature vectors and second, to investigate if our fairly limited data set is sufficient to train a corresponding network.

C.1 Experiments with Simple Two-Class Problem

This first section presents a small introductory classification problem to investigate the classification of variable-sized feature vectors. In this simple classification problem, for a certain percentage of the input patterns one of the input dimensions was not defined. Different Bayes error rates can be computed for the different number of defined input dimensions and the overall Bayes error rate can be computed from these individual Bayes error rates.

C.1.1 The Problem

The classification problem consists of two classes in a three-dimensional space. The distribution of each of the two classes is described by two components of Gaussian distributions with diagonal covariance matrices. The third dimension, however, is defined only for 50 % of the data. The class distributions for all three dimensions is shown in Fig. C.1.

The Bayes error rate considering all three dimensions is 5.27 %. If the third dimension is ignored, the Bayes error rate rises to 9.92 %. Thus, for the case that the third dimension only exists for 50 % of the data, the overall Bayes error rate is:

$$\frac{5.27\% + 9.92\%}{2} = 7.59\%.$$

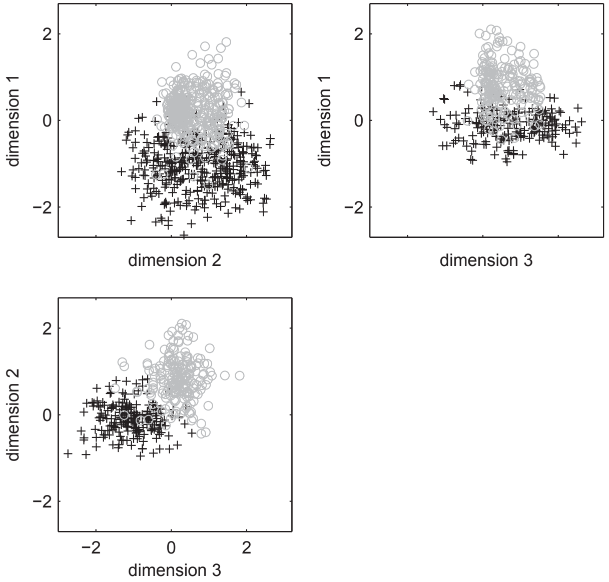

Figure C.1: *Distributions of the two classes, plotted in grey circles and black crosses, in all three dimensions*

C.1.2 The Neural Network Configuration

For the gate neural network we used the layer structure 3/12/7/2 with two hidden layers. To describe the layer structure of a network with N layers we use the notation $N_1/N_2/.../N_L$. This is merely a list of the number of nodes in each layer. A 3/12/7/2 network, for example, has 3 inputs, 12 nodes in the first hidden layer, 7 nodes in the second hidden layer, and 2 outputs. Each of these layers if presumed to be fully connected to its preceding and following layer with no short-cut or feedback connections. For the training of the network, a learning rate of 0.005 was applied while the momentum term was omitted. We stopped the training after 90,000 iterations. Some experiments on early stopping were conducted, however they lead to worse results than completing the full number of iterations[1]

C.1.3 Results

Fig. C.2 shows the results for various training data size. For every size, 10 sets of n patterns were used for the training. A large test set of 350,000 patterns was used to avoid bias that may be caused by a small test set that possibly represents the underlying distributions badly. The results show that the Bayes error rate of 7.59 % can be approximated given a sufficiently large number of training patterns. For 20,480 patterns a mean error rate of 7.99 % was reached, which is close to the overall Bayes error rate of 7.59 %. This shows clearly that the information contained in the third dimension, which is only defined for part of the data, can be used to decrease the classification error.

[1] Validation sets were used to implement the GL_2 early stopping criterion described in [Pre98], that estimates the relative increase of the validation error in percent over the minimum-so-far. The GL_2 criterion was chosen as a good trade-off between solution quality and training time. On the one hand, this criterion offers a high probability of finding a good solution and, on the other hand, does not need excessive training time.

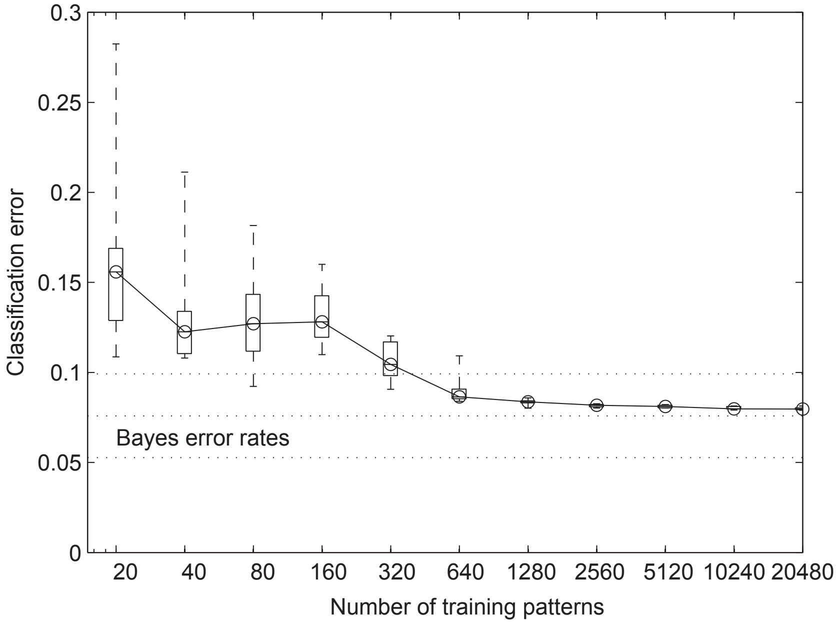

Figure C.2: *Classification error of a two-class problem in a three-dimensional space where the third dimension is only defined for 50 % of the input patterns. The number of training patterns is shown on a logarithmic scale on the x-axis. For each data size 10 different data sets were used for training, with the results represented as boxplot. Three Bayes error rates are shown as dotted horizontal lines: the Bayes error rate for 2 dimensions (top line), the Bayes error rate for all three dimensions (bottom line) and the Bayes error rate for data where the third dimension is only defined for 50 % of the input patterns (middle line).*

C.2 Training Data Simulation

The training of a neural network to classify phone instances into suitable and unsuitable, as in the proposed scenario, is a complex task due to the unlimited possibilities of network and training configurations. To find the optimal configuration for the neural network training and to find out if the available data are sufficient to realise an ANN-based classifier, we estimated some realistic distribution of the data to generate synthetic training data. The Bayes error rate of this synthetic training data is known as its distribution is known and thus we were able to an arbitrary number of data for training and a sufficiently large test set that is not biased.

C.2.1 Feature Probability Distributions

For almost all features, we estimated their distributions based on subsets of the data described in Section 8.1. As these subsets we determined all the predicate combinations that showed relevant differences in the feature distributions. A total of 67 subsets was found and from these subsets 67 distributions were estimated. In the following we describe how these distributions were estimated in detail.

Distribution of Centroid Distance

For the distribution of the centroid distances for suitable and unsuitable phones we were able to use more accurate data than the data from the test sentences described in Section 8.1. Instead, we used the data described in Section 6.4.3 to infer distributions and thresholds.

We computed the centroid from all phone instances (suitable and unsuitable) and then estimated the distributions of centroid distances for suitable and unsuitable phone instances. The centroid distance distributions for suitable and unsuitable phone instances of the *fg* voice are illustrated in Fig. C.3 for particular phones. To model the centroid distance distributions, however we did not differentiate between particular phones but for robustness reasons took the average of the means and variances over all phones, which were the following:

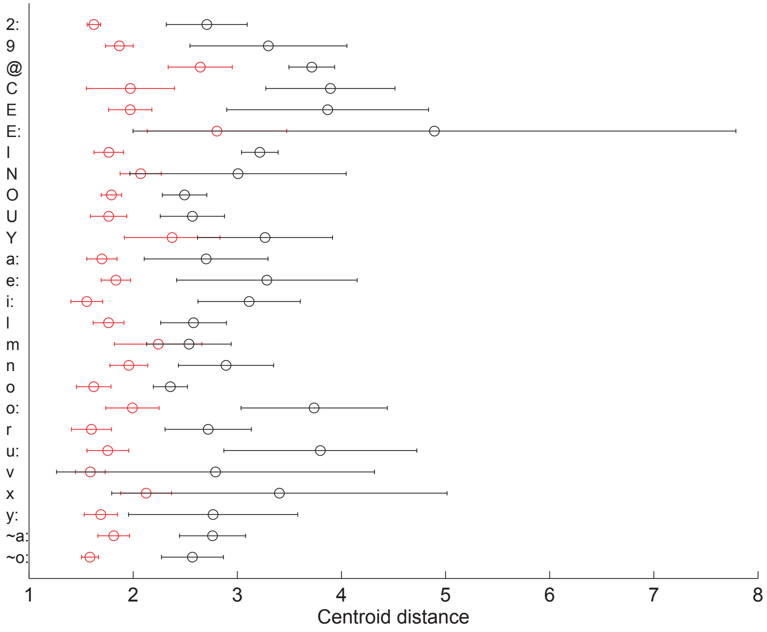

Figure C.3: *Estimated mean centroid distances and variances for suitable (on the left) and unsuitable (on the right) phone instances of phones of the fg voice. It can be seen that not only the mean distances for unsuitable phones are higher but also the variances.*

	mean centroid distance	variance
Suitable phone instances	1.56	0.15
Unsuitable phone instances	2.72	0.63

The centroid distances are only modelled for stationary phones, not for plosives, neither voiced nor unvoiced. One exception is made for affricates (voiced and unvoiced), where a separate centroid for the frication phase is modelled.

Distributions of Remaining Features

All the other feature distributions apart from the centroid distance were estimated from the data described in Section 8.1.1. We fitted different distributions on the data because Gaussian mixture models (GMM) do not apply well to all features. The distributions used to model the features are shown in Table C.1.

Probability distribution	Feature
Normal distribution	duration centroid distance f0penalty2
	pmOffset maxEnergy meanEnergy
Exponential distribution	nCreaky nMixed nCreakyLeft nCreakyRight
Lognormal distribution	f0penalty3
Beta distribution	f0penalty1
Bernoulli distribution	plosiveExists

Table C.1: *List of distributions used for the different features*

Covariance Matrices of Features Modelled by Gaussian Distributions

For those features that are modelled with Gaussian distributions, theoretically a full covariance matrix could be computed. However, we refrained from computing a full covariance matrix and compute only a diagonal covariance matrix instead, for two reasons: first and most important, the data are not sufficient to properly estimate a full covariance matrix. Second, some features of the data are mutually exclusive, that means, every sample contains some undefined features, making it impossible to estimate a common covariance matrix based on all data.

There are methods to compute the element (i,j) of the covariance matrix using all pairs of data points for which values of both x_i and x_j are available. However, such an approach may lead to poor results (see [GJ94]) and moreove, can be applied only to models based on a single Gaussian distribution (see [RM99]), not GMM.

Estimating Distributions for Data Partitions

We estimated the feature distributions for suitable and unsuitable phone instances considering particular partitions of the data. These partitions consisted of phones or phone types which can be described by combinations of the predicates described in Section 8.4.1. The feature distributions were estimated from partitions because the distinctive character of particular features would have been lost if these distributions were estimated from the whole data. If we consider for example the feature *duration* for the phone instances where the predicates `isVoicedPhone` and `isFollowedByPause` is true and the predicate `isPlosive` is false, we can see that the feature *duration* has a very strong distinctive character to distinguish suitable from unsuitable phone instances. If we instead consider the duration of all phone instances, its distinctive character is far less. However, it does not make any sense to consider different partitions of phones for features which are not influenced by these subset properties. E.g. the feature *phaseOffset* is not affected by the predicate `isFollowedByPause`, its distribution is the same for all data, no matter which value the predicate `isFollowedByPause` takes.

To decide which features are influenced by which combination of predicates, we investigated the data with the help of histograms and estimated separate distributions for the corresponding features if the predicate combinations appeared relevant.

C.2.2 Generating Complex Synthetic Data

The data, consisting of n patterns, were created in two steps. In a first step, we generated data containing the predicate values. More precisely, a number of n patterns was generated with the predicate values set according to their ratio in the manually annotated training data. For example, let the ratio of voiced plosives in the training data be r_{vp} and the number of patterns to be generated n. Then $m = r_{vp} * n$ patterns were generated with values `true` for the the predicates `isVoiced` and `isPlosive` and `false` for the rest of the predicates. This way we ensured that the synthetic data had the same ratio of phones and phone classes, for example the same ratio of glottal closures or voiced plosives as the training data.

In a second step, we went through the partitions, defined by the combinations of predicate values $p_{1..k}$, where k is the number of predicates, and determined the appropriate distribution for each feature f_i. That could be either a distribution of f_i that was estimated exactly from the subset described with the predicate sequence $p_{1..k}$ or a distribution that was estimated from a superset that included that subset described with $p_{1..k}$. A superset described with $q_{1..k}$ includes a predicate $p_{1..k}$ if some or all of the $q_l, 1 \leq l \leq k$ are either equal to p_l or undefined. Subsequently, this distribution was used to generate sample values for the feature f_i of that subset.

We generated two sets of synthetic data. One set contained equal members of each class (suitable and unsuitable phone instances), the other set took into consideration the unbalance of the two classes for each data partition.

C.2.3 Determining the Bayes Error

Data Partitions

To compute the Bayes error, we again used the predicate partitions that we used to generate the synthetic data (see C.2.2). Because these partitions are described by different distributions, the Bayes error has to be computed for each of these partitions separately using their individual distributions of feature values. Then a weighted mean of these errors (according to the ratio of the partitions with respect to all data) is taken to compute the overall Bayes error.

For the unbalanced data set, the a priori probabilities of the two classes have to be considered for each partition, therefore different

Bayes error rates result for the balanced and for the unbalanced data set.

Classic Approach to Determine the Bayes Error

The computation of the Bayes error rate in a high dimensional space proved to be not trivial. As a first classic approach, we computed the probabilities for the two classes for every sampling point on a grid. This grid was confined to a certain range (depending on the variances) around the means in each dimension, making sure that a certain percentage of the probability space of the distribution was covered[2]. If known, the possible range of the features was considered. E.g. the feature *nCreaky* only takes values between 0 and 1, therefore only patterns were generated with values in that range. However, this first approach was only used for a reduced number of features during development. Covering a range of 5 standard deviations would require to compute the probabilities of some 2,000,000 sampling points in up to 14 dimensions. A rough estimation showed that computing the probabilities for all these points would take some $3.5 \cdot 10^{68}$ years.

Monte-Carlo Sampling to Determine the Bayes Error

As described above, the computation of the class probabilities for every point in a 14-dimensional grid within a certain range is prohibitive. As a consequence, we applied Monte-Carlo sampling to compute the class probabilities for randomly chosen points in the n-dimensional feature space and used the ratio between correctly classified points and incorrectly classified points to determine the Bayes error. However, also for this approach the curse of dimensionality applied and the number of patterns has to grow exponentially with the number of dimensions to yield a reliably stable Bayes error estimate. We found that with more than 4 dimensions the computation time also grew beyond feasibility.

[2] The inverse cumulative distribution function can be used to compute the necessary range to include at least $k\%$ of the sampling points.

Using Test Pattern Probabilities to Determine the Bayes Error

As final approach we computed the class probabilities for the patterns that we finally used as a test set to determine the Bayes error. The advantage is that on the one hand we receive a very accurate estimate of the error rate that can be reached on that particular test set with an optimal classifier and on the other hand, the Bayes error computation is still computationally feasible for higher dimensions. The test set size should be large enough to be a good representative of the underlying distribution. However, the test set size is also confined by the evaluation time of the classifier.

C.3 Estimating the Number of Training Data

The crucial question for training a classifier, before one starts thinking about the details of parameter optimisation, is whether the number of training data that can be generated manually is sufficient for any classification at all. To answer that question, we used the distributions we gained in Section C.2 to generate data sets of different sizes using the same number of patterns for each class. For each training set size we generated 10 training sets and trained corresponding networks. Each network was tested on the same test set of 1,000,000 patterns. We present the 10 results for each training data size as boxplots in Fig. C.4. Note that we use a logarithmic scale on the x-axis which represents the training data sizes.

We actually trained several network configurations with different layer structures and number of nodes. The results for these several network configurations were similar, therefore we only show the error rates from one representative experiment in Fig. C.4.

We observed that the Bayes error rate is approached asymptotically with increased training data size. Between 5,000 and 10,000 patterns should suffice to achieve a reasonable classification result. This number of training pattern also lay in the order of magnitude of what we were able to classify manually.

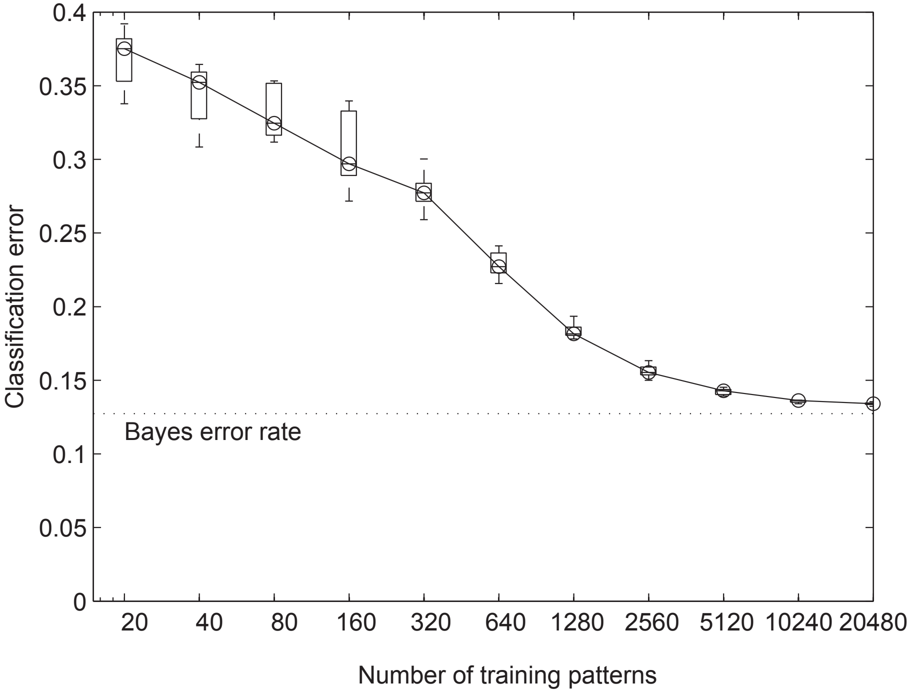

Figure C.4: *The error rates for each training data size are shown as boxplots, where the black circles denote the median error rate over 10 runs, the edges of the box denote the 25th and 75th percentiles, and the whiskers the most extreme data points. We used a network with 28 inputs, 9 nodes in the first, and 5 nodes in the second hidden layer, and 2 outputs (one for each class) and stopped the training after 50,000 iterations.*

C.4 Parameter Optimisation for Synthetic Data

Once the training data and the features for a neural network are known, many choices are to be made about the optimisation method and the parameters associated with each method. These choices include the way to cope with of heavily unbalanced data, the network layer structure, the target coding scheme, and the optimisation approach itself (backpropagation or some other approach). Each optimisation approach by itself has a certain number of parameters to be tuned, where the number of iterations, the learning rate, and the momentum term are the most common. Furthermore, if training sessions take too long, some form of early stopping or annealing technique may be considered. In fact, this meta-optimisation processis an iterative process: because of the large search space, it is prohibitive to optimise the parameters all at the same time. Furthermore, there exists no logical order in which to treat different aspects in this meta-optimisation process. In the following description, we treat these aspects in the order of relevance and start with the treatment of heavily unbalanced data in Section C.4.1.

C.4.1 Treatment of Unbalanced Data

The Situation

The percentage of suitable and unsuitable phone instances of the manually classified training data is very uneven, especially for some phones. One of the most unevenly distributed phones are affricates, unvoiced stationary phones that are followed by a preplosive pause (*unv. st. phones, followed by ppp* in Fig. C.5) and unvoiced preplosive pauses (*unv. ppp*), which are preplosive pauses preceding an unvoiced plosive. In the latter case, the uneven distribution can be attributed to the fact that, first, a preplosive pause is not a critical phone (after all, it is a pause) and second, a backward masking effect[3] may mask artifacts that occur in the preplosive pause. The backward masking effect occurs

[3] Obscuration of a sound immediately preceding the masker is denoted backward masking.

in this context because the preplosive pause is normally followed by a high intensity burst. From this example we can conclude that some phone instances are more problematic than others and that the uneven distribution is probably due to the phone class characteristics. Be that as it may, the classifier has to cope with this partly extreme uneven distribution of training data.

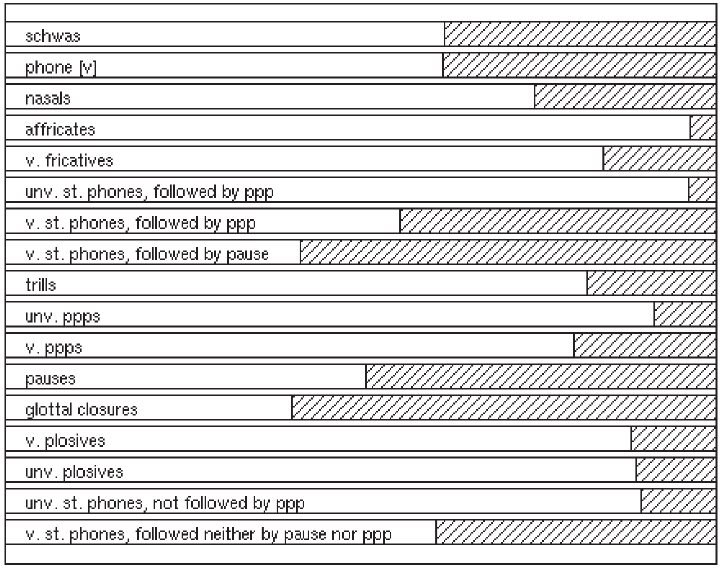

Figure C.5: *Distribution of the training data, on the left the suitable, on the right the unsuitable phone instances. Abbreviations are: v. for voiced, st. for stationary and ppp for preplosive pause.*

Approaches to Cope with Unbalanced Data

Conventional classification approaches express a posteriori probabilities through Bayes' theorem in the form:

$$P(C_k|\boldsymbol{x}\ ) = \frac{p(C_k)P(C_k)}{p(\boldsymbol{x}\ )}. \tag{C.1}$$

The neural network approach provides direct estimates of the a posteriori probabilities $P(C_k|\boldsymbol{x}\ )$. That means that it learns a combination of the class-conditional densities $p(C_k)$ and the a priori probabilities $P(C_k)$. To cope with unbalanced data, two approaches are presented. In the Section C.4.2, a data balancing approach and in Section C.4.3 an educational learning approach is demonstrated.

C.4.2 Balancing Training data

We can use different degrees of balancing, from not balancing data at all to full balancing of the training data.

No Training Data Balancing If we do not balance the training data, the network learns the a priori probabilities $p(C_k)$ of the training data. Thus, the network will be biased, and, for our training data, will produce much more false positives than false negatives. For our application, however, we rather want to have a suitable phone instance classified as unsuitable than vice versa. That means, we prefer a larger number of false negatives. Therefore, having the network learn the a priori probabilities is not an option for our application.

Full Training Data Balancing One approach to balance two classes is to replicate the patterns of the smaller class until both classes have the same size. As a consequence, the a priori probabilities $P(C_k)$ will be equal for both classes.

Partly Training Data Balancing As with full training data balancing, we replicate the patterns of the smaller class, yet impose some restriction on how often a pattern may be replicated. Thus, an over-representation of certain training patterns can be avoiced. As a consequence, the different data subsets will be balanced to a different extent, and therefore also a priori probabilities will be learnt to a different extent for different subsets of data.

Experiments with Different Degrees of Balancing

For experiments, we created unbalanced synthetic data from the same parameters as described in Section C.2 with the same distribution of suitable vs unsuitable phone instances as in the original training data. Because of the changed a priori probabilities, the Bayes error rate changed to 7.71 % as the a priori probabilities have to be considered when computing the Bayes error rate.

We used different layer structures and different numbers of iterations to achieve optimal results for different degrees of balancing. Only the parameter choice for the best results are shown in Table C.2.

data replication factor	layer structure	iterations	error rate on test data [%]
1	28/15/7/2	60,000	**12.31**
3	28/12/5/2	60,000	14.51
5	28/12/5/2	60,000	14.77
8	28/12/5/2	60,000	14.06
∞	28/9/5/2	40,000	13.54

Table C.2: *Best network parameters for different degree of balancing. Data replication factor describes how often a pattern may be used for balancing the data. A factor of 1 means each pattern is exactly used once, which means no balancing at all whereas ∞ means there is no restriction on how often a pattern may be used, so the data is completely balanced. The Bayes error rate for this data is 7.71 %.*

We clearly see that best results are achieved if the data is not balanced at all. However, the number of false positives is highest in this case. In total, the error rate is still far from the Bayes error rate of 7.71 %, no matter which degree of balancing is used.

C.4.3 Educational Learning

What our application is concerned, does focusing on a priori probabilities lead to the desired results? On the one hand, we want to order the phone instances to find the most suitable one. To achieve such an ordering, however, we do not need a (a priori-wise) balanced classifier, so we could ignore the number of false positives and leave the data unbalanced. But on the other hand, our classifier should learn rare events, so it would be advantageous to balance the data in some way, to have a large number of, in our case, unsuitable phone instances.

One approach to focus on rare events is educational learning. Educational learning is a type of pattern weighting heuristic (see [RM99]) that can be used to tackle the data balancing problem. We focus on the patterns which are misclassified, thus modifying the error function by giving more emphasis on those patterns.

In the following sections, we present two different heuristic approaches of educational learning: one is *learning rate adaptation* and the other is the *repeated application of wrongly classified patterns*.

Learning Rate Adaptation

In this approach two different learning rates are applied. A higher learning rate is applied if the pattern has been wrongly classified and a lower learning rate (the one that proved to be optimal in the preceding experiments) is applied if the pattern has been correctly classified:

$$\eta_w = \lambda \cdot \eta_c, \tag{C.2}$$

where η_w is the learning rate for wrongly classified pattern, η_c is the learning rate for correctly classified pattern and λ the learning rate factor. This learning rate factor describes the ratio between the learning rate for correctly classified patterns and wrongly classified patterns.

We used different values for λ to train networks and evaluated them on test data, see Table C.3.

λ	Layer structure	Iterations	Error rate on test data [%]
2	28/12/5/2	40,000	**8.55**
3	28/12/5/2	40,000	8.71
5	28/12/5/2	40,000	8.90
7	28/12/5/2	40,000	9.31
10	28/12/5/2	40,000	9.71
1.5	28/12/5/2	60,000	8.67
2	28/12/5/2	60,000	**8.55**
3	28/12/5/2	60,000	8.63
5	28/12/5/2	60,000	8.90
2	28/12/5/2	80,000	8.81
2	28/15/7/2	40,000	8.75
2	28/15/7/2	60,000	8.80
2	28/9/5/2	40,000	8.74
2	28/9/5/2	60,000	8.71
1.5	28/12/5/2	40,000	8.70

Table C.3: *Best network parameters for educational learning with online-learning and the adaption of the learning rate. The Bayes error rate for this data is 7.71 %. The error rates are the mean error rates over 5 runs.*

Repeated Application of Wrongly Classified Patterns

In this method, we applied subset training where a part of the subset consists of patterns that have not been learnt correctly in the previous iteration and the rest of the subset consists of randomly selected patterns. After a pattern has been applied for the second time, it is removed from the training to avoid that contradictory data remains in the subset for too long.

Experiments were conducted for different subset sizes and different numbers of iterations, see Table C.4.

Subset size	Layer structure	Iterations	Error rate on test data [%]
200	28/12/5/2	3,000,000	8.58
200	28/12/5/2	4,500,000	**8.54**
300	28/12/5/2	3,000,000	8.56
200	28/12/5/2	2,000,000	8.68

Table C.4: *Best network parameters for educational learning with subset learning and repeated application of wrongly classified patterns. The Bayes error rate for this data is 7.71 %. The error rates are the mean error rates over 5 runs.*

Choice of Educational Learning Method

For both educational learning methods, the lowest error rates achieved on the synthetic data were about equal (see Tables C.3 and C.4). However, the training time for the second method, the re-usage of wrongly classified patterns, is considerably higher. The final decision on the educational learning method was made after a set of informal listening tests (see Section C.4.8).

C.4.4 Network Size and Number of Iterations

Theoretical Considerations about Network Size

One substantial question in designing a network is how many nodes to place in each layer. On the one hand, the representational capacity of the network should cover the problem. In other words, the network should be sufficiently large so it can fit the function we want to approximate. On the other hand, generalisation criteria become the limiting factor on performance if networks grow over a certain size. A large network can often fit the training data exactly but will most probably not fit the data in a way that represents the underlying function that generated the data (see [RM99]).

In this section we focus on the number of nodes, the question whether to use one or two hidden layers is treated in Section C.4.5. There are various rules on the connection between network size and capacity but these rules either require certain preconditions (like only one hidden layer or linear threshold functions) or are only of theoretical significance, for example how many nodes are needed for the exact representation of a certain function. Generalisation, however, depends on many factors so statements based only on network size and number of patterns cannot predict generalisation performance (see [RM99]).

To approximate a reasonable number of nodes so the network would be capable to represent our data on the one hand, and on the other hand not be prone to over-fitting, we trained several network configurations and observed possible over-fitting effects after a large number of iterations. As a side product of these experiments we were also able to estimate a reasonable number of iterations, large enough to learn the training data but small enough to avoid over-training.

We observed from previous experiments that an increase in the number of nodes lead to a decrease in error rates on the test set, even with very high numbers of nodes. But we also saw that once a certain network size has been reached, the error rate could only be improved very slightly by increasing the number of nodes. The downside of increasing the number of nodes, however, is a drastic increase in training time. This saturation effect of the error rate concerning the number of nodes is even stronger if more iterations are used.

Network Size for Balanced Training Data

Before tackling the network design for unbalanced training data, we investigated the influence of network size using the classic learning method on balanced data. The data sizes were 4,000 and 8,000 patterns, which is in the order of magnitude of the around 15,000 training patterns of real, but heavily unbalanced data that were available. We used 4 different network structures starting from a small network with 28/7/5/2 to a large network with 28/15/7/2 and evaluated the networks on an evaluation set every 200 iterations until we observed clear over-fitting effects after some 200,000 iterations.

We found that an increase in the number of nodes lead to a slight decrease in the minimal error rate that is reached on the evaluation set during the training as can be seen comparing Table C.5 and C.6. However, especially for the larger networks and the smaller training data size of 4,000 patterns, we observed that the error rate on the evaluation set increased considerably having passed a certain number of iterations. E.g. considering the largest network trained with 4,000 patterns we can see in Table C.5 that the mean error rate goes up from 14.78% to 17.09% after 200,000 iterations. This over-fitting effect is smaller for more training patterns where the error rate only goes up from 14.12% to 14.64% (see Table C.6). These experiments with balanced data show that with a small sacrifice in error rate, considerable safety in terms of over-fitting can be achieved.

Layer structure	Mean min error rate	Mean error rate after 200,000 iterations
28/7/5/2	14.89 %	15.53 %
28/9/5/2	14.76 %	16.06 %
28/12/5/2	14.74 %	16.46 %
28/15/7/2	14.78 %	17.09 %

Table C.5: *Comparison of minimal error rates and error rates after 200,000 iterations for different network structures. Mean errors taken over 5 runs for each network structure. 4,000 training patterns were used. The Bayes error rate is 12.73 %.*

Layer structure	Mean min error rate	Mean error rate after 200,000 iterations
28/7/5/2	14.15 %	14.21 %
28/9/5/2	14.09 %	14.21 %
28/12/5/2	14.06 %	14.46 %
28/15/7/2	14.12 %	14.64 %

Table C.6: *Comparison of minimal error rates and error rates after 200,000 iterations for different network structures. Mean errors taken over 5 runs for each network structure. 8,000 training patterns were used. The Bayes error rate is 12.73 %.*

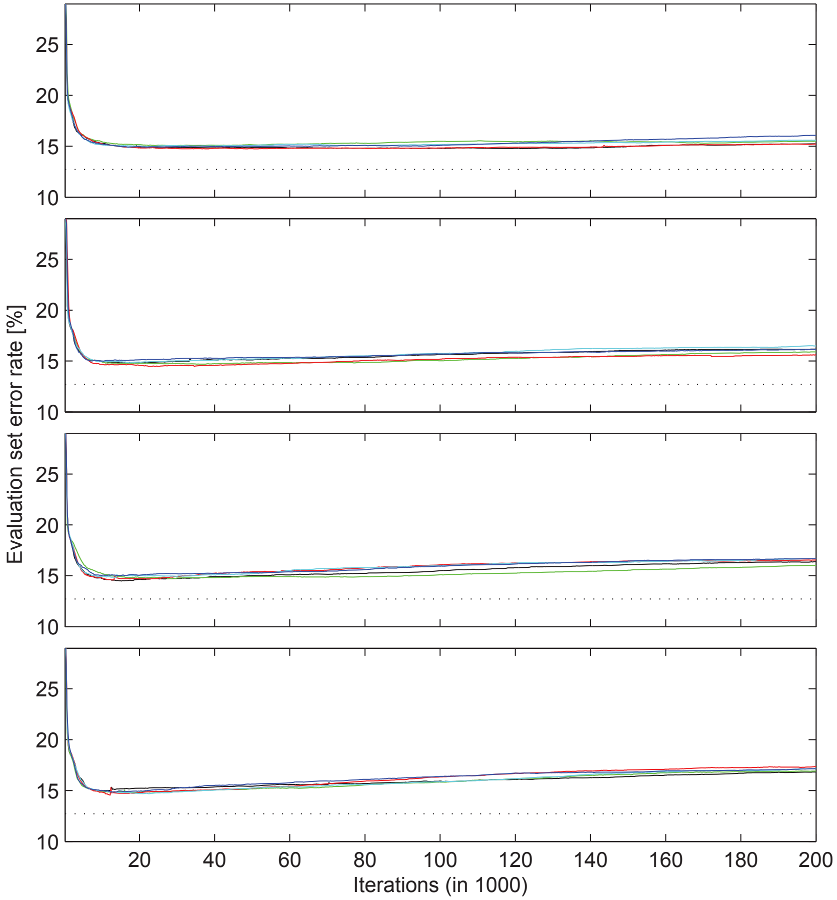

Figure C.6: *Evaluation set errors for different layer structures during the course of a training with 200,000 iterations and 4,000 balanced training patterns. The layer structures from top to bottom: 28/7/5/2, 28/9/5/2, 28/12/7/2 and 28/15/7/2. The Bayes error rate is shown as a dotted line at 12.73 %. First, the error rates drop sharply during the first few iterations, then start to increase slightly again.*

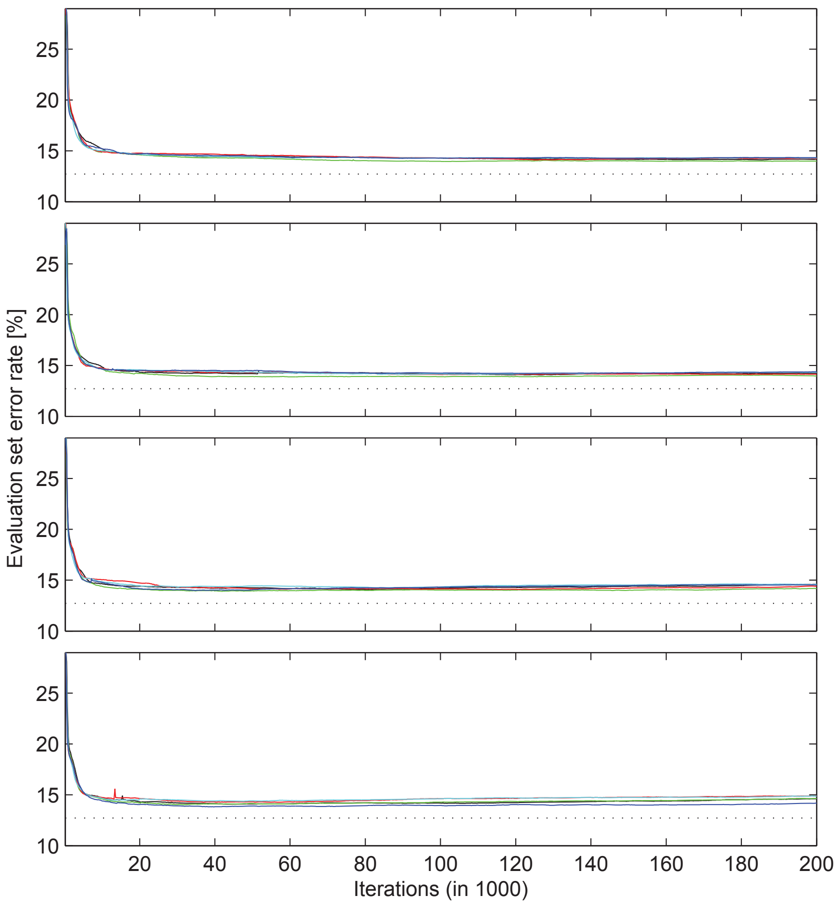

Figure C.7: *Evaluation set errors for different layer structures during the course of a training with 200,000 iterations and 8.000 balanced training patterns. The layer structures from top to bottom: 28/7/5/2, 28/9/5/2, 28/12/7/2 and 28/15/7/2. The Bayes error rate is shown as a dotted line at 12.73 %.*

Network Size for Unbalanced Data

We used 15,000 patterns of unbalanced data (see Section C.4.2) to investigate the possible effect of over-fitting on different network structures if educational learning was used.

We used the same network structures as in Section C.4.4 and again evaluated the networks on an evaluation set every 200 iterations. We used educational training with learning rate adaption to account for the unbalanced data. Just as for balanced data, we observed a slight error rate increase for the largest network after having reached a certain number of iterations (see Fig. C.8). As for the balanced data, we can see that the larger networks are slightly more prone to over-fitting. However, this effect is relatively small for the unbalanced data sets due to the larger number of data.

Layer structure	Mean min error rate	Mean error rate after 200,000 iterations
28/7/5/2	8.47 %	8.53 %
28/9/5/2	8.43 %	8.49 %
28/12/5/2	8.44 %	8.50 %
28/15/7/2	8.47 %	8.63 %

Table C.7: *Comparison of minimal error rates and error rates after 200,000 iterations for different network structures. Mean errors taken over 5 runs for each network structure. 15,000 unbalanced training patterns were used. The Bayes error rate is 7.71 %.*

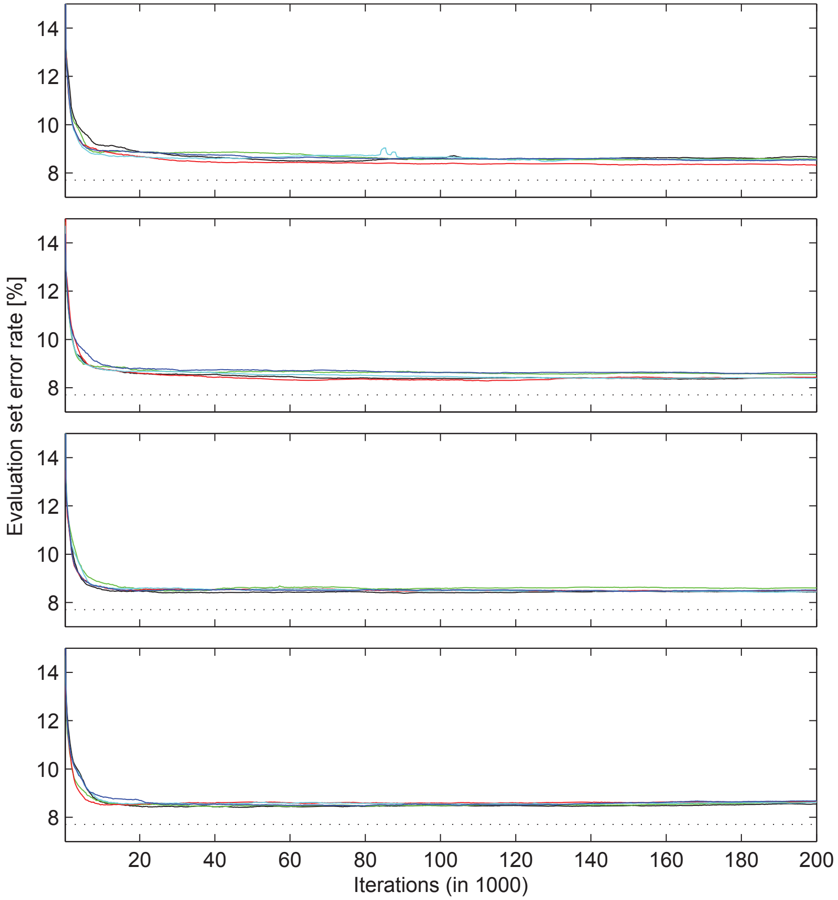

Figure C.8: *Evaluation set errors for different layer structures during the course of a training with 200,000 iterations and 15,000 training patterns. The layer structures from top to bottom: 28/7/5/2, 28/9/5/2, 28/12/7/2 and 28/15/7/2. The Bayes error rate is shown as the dotted line at 7.71 %. First, the error rates drop sharply, then they start to increase slightly again.*

C.4.5 Number of Hidden Layers

Theoretical results show that a network with three layers of weights[4] can generate arbitrary decision regions, which may be non-convex and disjoint (see [Lip87], [Bis95]). All the same, a two-layer network with sigmoidal activation functions can approximate any given decision boundary arbitrarily closely (see [Bis95]). However, one has to bear in mind that these theoretical results are to be considered statements about the power of a class of networks and do not guarantee that a particular network will be able to learn a particular set of data and generalise properly. Moreover, these theoretical results are based on asymptotic analyses that are valid only for large sets of data or large input dimensions (see [RM99]). Therefore, we conducted two experiments, one with balanced data and one with unbalanced data to study whether a two or a three-layer network proves more suitable for our problem.

Experiments for Balanced Data

We trained different networks with an approximately equal number of weights, distributed over two and three layers, respectively.

layer structure	number of weights	error rate [%]
28/20/8/2	736	**13.26**
28/24/2	720	13.48
28/25/2	750	13.48

Table C.8: *Mean error rates for different layer configurations with approximately equal number of weights. For each layer configuration we trained 10 different networks with different training data and tested them on the same test set of 1,000,000 patterns. We used 20,480 patterns of balanced training data and stopped the training after 50,000 iterations.*

[4] As an L-layer network we consider a network with L active layers, which comprise $L-1$ hidden layers and one output layer. Inputs are excluded as they do no computation.

Experiments for Unbalanced Data

For the experiment with unbalanced training data less nodes were used in the networks, as the training data was unbalanced and consisted of less training patterns. We used educational training with learning rate adaption (see Section C.4.3) for all layer configurations.

layer structure	number of weights	error rate [%]
28/12/5/2	406	**8.53**
28/13/2	390	8.63
28/14/2	420	8.61

Table C.9: *Mean error rates for different layer configurations with approximately equal number of weights. For each layer configuration we trained 10 different networks with different training data and tested them on the same test set of 1,000,000 patterns. We used 15,000 patterns of unbalanced training data and stopped the training after 60,000 iterations.*

The results clearly show that layer configurations with two hidden layers achieve best results for balanced as well as for unbalanced data. Networks with one hidden layer perform still worse even if they contain a larger number of weights. A network with only one hidden layer could probably reach similar results with a substantially larger number of nodes. However this network would be very prone to over-training.

C.4.6 Choosing a Target Coding Scheme

Theory

The target values can be chosen according to various schemes. We investigated the two most common schemes for a two-class problem, the 1-of-c scheme and a single-output target coding scheme.

The coding with two outputs is called 1-of-c target coding scheme. In this scheme, we have $t_k^n = \delta_{kl}$ for an input vector $\mathbf{x}^n$ from class $\mathcal{C}_l$ where δ_{kl} is the Kronecker delta symbol. The Kronecker delta symbol is defined as $\delta_{kl} = 1$ if $k = l$ and $\delta_{kl} = 0$ otherwise. If a sum-of-squares error function is used (as we did), it can be shown that the outputs of the network correspond to Bayesian a posteriori probabilities of class membership (see [Bis95], p.225):

$$y_k(\mathbf{x}) = P(\mathcal{C}_k|\mathbf{x}). \tag{C.3}$$

As second approach we used a single output y with a target coding which sets $t^n = 0$ if $\mathbf{x}^n$ is from class $\mathcal{C}_1$ and $t^n = 1$ if $\mathbf{x}^n$ is from class $\mathcal{C}_2$. In this case it can be shown (see [Bis95], p.226) that the network output $y(\mathbf{x})$ represents the a posteriori probably of the input vector $\mathbf{x}$ belonging to class $\mathcal{C}_2$:

$$y_k(\mathbf{x}) = P(\mathcal{C}_2|\mathbf{x}). \tag{C.4}$$

The corresponding probability for class $\mathcal{C}_1$ is then given by $P(\mathcal{C}_1|\mathbf{x}) = 1 - y(\mathbf{x})$.

In the literature, both approaches were described as equally valid alternatives, so we conducted a number of experiments to investigate which approach gives better results when applied to our problem.

We conducted experiments to compare layer structures that only differed in the number of outputs and differed as little as possible in the total number of weights. The difference in number of weights was 5 (406 weights for 2 outputs, 401 weights for 1 output), which is a difference of 1.2 %.

Experiments with Balanced Data

From Table C.10 we notice that for balanced data of 10,240 training patterns, the error rates with 2 outputs are slightly lower but with an increased number of iterations, the error rates converge.

Experiments with Unbalanced Data

For networks trained on unbalanced data with error rate adaptation as learning method, we observed that with the number of iterations, the error rates approximate (see Table C.11).

layer structure	total number of weights	error rate after 40,000 iterations	error rate after 60,000 iterations
28/12/5/1	401	13.72 %	13.58 %
28/12/5/2	406	13.60 %	13.58 %

Table C.10: *Comparison two network structures that only differ in the number of nodes in the output layer. After 40,000 iterations, the network with two output nodes shows lower error rates, after 60,000 iterations, equal error rates are reached. We used 10,240 patterns for all runs and took the mean error taken over 5 runs for each network structure. The Bayes error rate is 12.73 %.*

layer structure	total number of weights	error rate after 40,000 iterations	error rate after 60,000 iterations
28/12/5/1	401	8.73 %	8.57 %
28/12/5/2	406	8.59 %	8.55 %

Table C.11: *Comparison two network structures that only differ in the number of nodes in the output layer. The network with two output nodes shows lower error rates, however the error rates approximate with a higher number of iterations. We used 15,000 unbalanced patterns for all runs, the mean error taken over 5 runs for each network structure. The Bayes error rate is 7.71 %.*

Results

Although we know from theoretical results that the number of outputs does not make any difference in the network performance, we discovered that networks with two outputs tend to learn a little bit faster. In addition to the experiments shown above, we conducted further experiments with a smaller number of iterations and larger training data sets, and we did not observe any result where one output performed better than two outputs. Therefore, we chose to use a target coding scheme with two outputs.

C.4.7 Optimisation Algorithms

Balanced Data

We use the *batch version* of gradient descent (see [Bis95]). The sequential approach (also pattern-based approach, where the error function gradient is evaluated for only one pattern at a time and the weights updated accordingly) is recommended when there is large redundancy in the data (i.e. lots of similar pattern with basically the same information). However, that is not the case with our data so we applied the batch version of gradient descent for reasons of computational efficiency. For more details on the batch and sequential approaches see [RM99].

We examined alternative optimisation methods to back-propagation, which is often criticised for its slow rate of convergence. Specifically, we examined implementations of RPPROP [RB93], scaled conjugate gradient [Møl93], Levenberg Marquardt training ([Mar63, Lev44] and quickprop [Fah88] from an open source Java framework [Hea10]. However, with none of these methods we could accomplish the error rates we were able to achieve with basic back-propagation. This may be due to the well tuned parameters that we used for back-propagation, which in this case is comparable to alternative methods even in terms of speed (see [RM99]).

Unbalanced data

For the educational learning method based on learning rate adaption we used a sequential approach of gradient descent. This is necessary, as the learning rate is modified according to whether the current training pattern was classified correctly or not. Therefore, the learning rate is set immediately after the evaluation of the pattern and before the weight update.

For the educational learning method that used repeated application of wrongly classified patterns, we used a batch version of gradient descent where the subsets are composed partly from previously misclassified patterns (see Section C.4.3).

C.4.8 Final Network Choice

To decide on the final network, 20 networks were trained with different network configurations using both educational learning methods. These networks were then used to select corpora and to synthesise between 10 and 14 test sentences per voice. Finally, these test sentences were evaluated with an informal listening test. We found that, although the error rates were approximately equal for both educational learning methods, the networks that had been trained with learning rate adaptation provided slightly better results when used for the selection of diphone elements.

In the end, we used a network with layer structure 28/12/5/2, trained it with a learning rate factor $\lambda = 2$ and stopped the training after 60,000 iterations. The learning rate η_c for the correctly classified patterns was set to a value of 0.005, no momentum term was used (7th configuration in Table C.3).

Appendix D

Intelligibility Test Results

This appendix presents in detail the evaluation results of the intelligibility tests described in Section 9. A general discussion of the results can be found in Section 9.3.5. In the following, the confusion matrices of the evaluations are listed. For German, the rhyme test checks the distinction of initial consonants, mid-word vowels and final consonants. Thus, three confusion matrices are listed for each German voice. The English rhyme test encompasses only initial and final consonants. Thus, two confusion matrices are shown for each English voice.

D.1 *fg* Voice

	[pʰ]	[tʰ]	[kʰ]	[b]	[d]	[g]	[t͡s]	[f]	[ʃ]	[z]	[h]	[r]	[v]	[m]	[n]	[l]	[j]
[pʰ]	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[tʰ]	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[kʰ]	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[b]	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0
[d]	0	0	0	0	33	0	0	0	0	0	0	1	0	0	0	0	0
[g]	0	0	0	2	0	17	0	0	0	0	0	2	1	0	0	0	0
[t͡s]	0	0	0	0	0	0	17	1	0	0	0	0	0	0	0	0	0
[f]	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0
[ʃ]	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0
[z]	0	0	0	0	0	0	1	0	0	38	0	0	0	0	0	0	0
[h]	1	0	1	2	0	0	0	0	0	0	12	1	0	0	0	0	0
[r]	0	0	0	0	0	0	0	0	0	0	1	49	0	0	0	0	0
[v]	0	0	0	0	0	0	0	0	0	0	0	0	49	0	0	0	0
[m]	0	0	0	0	0	0	0	0	0	0	0	0	0	42	0	0	0
[n]	0	0	0	0	0	0	0	0	0	0	0	0	0	2	29	0	0
[l]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0
[j]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13

Table D.1: *Confusion matrix for the initial consonants of the fg voice.*

	[a]	[aː]	[eː]	[ɛ]	[ɛː]	[iː]	[ɪ]	[oː]	[ɔ]	[uː]	[ʊ]	[øː]	[œ]	[ɔy]	[aj̯]	[au̯]	[yː]	[ʏ]
[a]	36	1	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0
[aː]	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[eː]	0	0	24	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0
[ɛ]	0	0	2	51	0	0	0	0	0	0	0	0	0	0	1	0	0	0
[ɛː]	0	0	4	0	3	2	4	0	0	0	0	0	0	0	1	0	0	0
[iː]	0	0	0	0	0	37	1	0	0	0	0	0	0	0	0	0	0	0
[ɪ]	0	0	1	1	0	2	50	0	0	0	0	0	0	0	0	0	0	0
[oː]	0	0	0	0	0	0	0	3	0	1	0	0	0	0	0	0	0	0
[ɔ]	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
[uː]	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0
[ʊ]	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0
[øː]	0	0	0	0	0	0	0	0	0	0	0	15	2	0	0	0	0	0
[œ]	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0
[ɔy]	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0
[aj̯]	2	0	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0
[au̯]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0
[yː]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0
[ʏ]	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	24

Table D.2: *Confusion matrix for the middle vowels of the fg voice.*

	[p]	[pʰ]	[t]	[tʰ]	[k]	[kʰ]	[t͜s]	[ç]	[s]	[ʃ]	[ɐ̯]	[f]	[x]	[r]	[l]	[m]	[n]	[ŋ]
[p]	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[pʰ]	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[t]	0	0	32	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
[tʰ]	0	0	0	83	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[k]	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0
[kʰ]	0	0	0	0	0	34	0	0	0	0	0	0	0	0	0	0	0	0
[t͜s]	0	0	0	2	0	0	11	0	0	0	0	0	0	0	0	0	0	0
[ç]	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
[s]	0	0	0	0	0	0	1	0	41	0	0	0	0	0	0	0	0	0
[ʃ]	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0
[ɐ̯]	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0
[f]	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0
[x]	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0
[r]	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0
[l]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0
[m]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0
[n]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	1
[ŋ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13

Table D.3: *Confusion matrix for the final consonants of the fg voice.*

D.2 *mg* Voice

	[pʰ]	[tʰ]	[kʰ]	[b]	[d]	[g]	[t͜s]	[f]	[ʃ]	[z]	[h]	[r]	[v]	[m]	[n]	[l]	[j]
[pʰ]	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[tʰ]	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[kʰ]	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[b]	0	0	0	25	0	0	0	0	0	0	1	0	1	0	0	0	0
[d]	0	0	0	0	46	0	0	0	0	0	0	0	0	0	0	0	0
[g]	0	0	3	0	0	22	0	0	0	0	0	0	0	0	0	0	0
[t͜s]	0	0	0	0	0	0	17	0	0	3	0	0	0	0	0	0	0
[f]	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0
[ʃ]	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0
[z]	0	0	0	0	0	0	0	0	0	46	0	0	0	0	0	0	0
[h]	1	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0
[r]	0	0	0	0	0	0	0	0	0	0	0	43	0	0	0	0	0
[v]	0	0	0	0	0	0	0	0	0	0	0	0	41	0	0	0	0
[m]	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0
[n]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0
[l]	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	36	0
[j]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7

Table D.4: *Confusion matrix for the initial consonants of the mg voice.*

	[a]	[aː]	[eː]	[ɛ]	[ɛː]	[iː]	[ɪ]	[oː]	[ɔ]	[uː]	[ʊ]	[øː]	[œ]	[ɔ͜y]	[ai̯]	[au̯]	[yː]	[ʏ]
[a]	50	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[aː]	3	29	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[eː]	0	0	20	1	1	2	3	0	0	0	0	0	0	0	0	0	0	0
[ɛ]	0	0	0	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[ɛː]	0	0	0	2	8	0	1	0	0	0	0	0	0	0	0	0	0	0
[iː]	0	0	0	0	0	28	1	0	0	0	0	0	0	0	0	0	0	0
[ɪ]	0	0	1	0	0	0	49	0	0	0	0	0	0	0	0	0	0	0
[oː]	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0
[ɔ]	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0
[uː]	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0
[ʊ]	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0
[øː]	0	0	0	0	0	0	0	0	0	0	0	16	0	1	0	0	0	0
[œ]	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0
[ɔ͜y]	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0
[ai̯]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0
[au̯]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0
[yː]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0
[ʏ]	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	19

Table D.5: *Confusion matrix for the middle vowel of the mg voice.*

	[p]	[pʰ]	[t]	[tʰ]	[k]	[kʰ]	[t͜s]	[ç]	[s]	[ʃ]	[ɐ̯]	[f]	[x]	[r]	[l]	[m]	[n]	[ŋ]
[p]	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[pʰ]	1	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[t]	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[tʰ]	1	0	0	90	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[k]	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
[kʰ]	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0
[t͜s]	0	0	0	3	0	0	6	0	0	0	0	0	0	0	0	0	0	0
[ç]	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0
[s]	0	0	0	2	0	0	0	0	38	0	0	0	0	0	0	0	0	0
[ʃ]	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0
[ɐ̯]	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
[f]	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0
[x]	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0
[r]	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0
[l]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0
[m]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28	0	0
[n]	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	48	0
[ŋ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	5	11

Table D.6: *Confusion matrix for the final consonants of the mg voice.*

D.3 *fe* Voice

	[p^h]	[t^h]	[k^h]	[b]	[d]	[g]	[θ]	[ð]	[ʃ]	[s]	[d͜ʒ]	[f]	[h]	[l]	[m]	[n]	[r]	[v]	[w]	[ɔ͜ɪ]	[iː]
[p^h]	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[t^h]	0	33	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
[k^h]	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[b]	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[d]	0	2	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[g]	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[θ]	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[ð]	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
[ʃ]	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0
[s]	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0
[d͜ʒ]	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0
[f]	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0
[h]	3	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0
[l]	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0
[m]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0
[n]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0
[r]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0
[v]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	0	0
[w]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0
[ɔ͜ɪ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
[iː]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9

Table D.7: *Confusion matrix for the initial consonants of the fe voice.*

	[p^h]	[t^h]	[k^h]	[b]	[d]	[g]	[θ]	[t͜ʃ]	[s]	[z]	[d͜ʒ]	[f]	[l]	[m]	[n]	[ŋ]	[r]	[v]	[ə]	[e͜ɪ]
[p^h]	30	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[t^h]	0	40	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[k^h]	1	0	47	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0
[b]	2	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[d]	0	12	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[g]	0	0	3	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[θ]	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	2	0	0
[t͜ʃ]	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0
[s]	0	0	0	0	0	0	0	0	35	1	0	0	0	0	0	0	0	0	0	0
[z]	0	0	0	0	0	0	0	0	5	2	0	0	0	0	0	0	0	0	0	0
[d͜ʒ]	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
[f]	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	2	0	0
[l]	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0
[m]	0	0	0	0	0	0	0	0	0	0	0	0	0	19	1	0	0	0	0	0
[n]	0	0	0	0	0	0	0	0	0	0	0	0	0	8	33	0	0	0	0	0
[ŋ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	5	0	0	0	0
[r]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
[v]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0
[ə]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
[e͜ɪ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8

Table D.8: *Confusion matrix for the final consonants of the fe voice.*

D.4 *me* Voice

	[pʰ]	[tʰ]	[kʰ]	[b]	[d]	[g]	[θ]	[ð]	[ʃ]	[s]	[dʒ]	[f]	[h]	[l]	[m]	[n]	[r]	[v]	[w]	[ɔː]	[iː]
[pʰ]	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[tʰ]	0	35	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[kʰ]	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[b]	1	0	0	32	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
[d]	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[g]	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[θ]	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[ð]	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
[ʃ]	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
[s]	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0
[dʒ]	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0
[f]	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0
[h]	1	0	0	2	0	0	0	0	0	0	0	1	31	0	0	0	0	0	0	0	0
[l]	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0
[m]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0
[n]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0
[r]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	3	0	0
[v]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[w]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0
[ɔː]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
[iː]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5

Table D.9: *Confusion matrix for the initial consonants of the me voice.*

	[pʰ]	[tʰ]	[kʰ]	[b]	[d]	[g]	[θ]	[tʃ]	[s]	[z]	[dʒ]	[f]	[l]	[m]	[n]	[ŋ]	[r]	[v]	[ɛː]
[pʰ]	20	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[tʰ]	0	32	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
[kʰ]	0	0	42	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
[b]	0	0	0	16	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
[d]	1	3	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[g]	0	0	2	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0
[θ]	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	1	0
[tʃ]	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0
[s]	0	0	0	0	0	0	0	0	41	2	0	0	0	0	0	0	0	0	0
[z]	0	0	0	0	0	0	0	0	4	4	0	0	0	0	0	0	0	0	0
[dʒ]	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
[f]	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	2	0
[l]	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0
[m]	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0
[n]	0	0	0	0	0	0	0	0	0	0	0	0	0	1	51	0	0	0	0
[ŋ]	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	12	0	0	0
[r]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0
[v]	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	15	0
[ɛː]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7

Table D.10: *Confusion matrix for the final consonants of the me voice.*

Appendix E

Phone Inventories

In this appendix, the phone inventories for German and English are defined as they are used in the SVOX speech synthesiser. The phones are represented with IPA symbols and illustrated with some examples in graphemic and phonetic form. For technical reasons we introduced preplosive pauses which are denoted according to the plosive they belong to. Therefore, we added symbols for preplosive pauses and preplosive before elisions, that are denoted with one or two dots below the corresponding plosive.

German phone inventory (incl. Swiss German diphthongs)

IPA	Example	
aː	Bahn	[ˈbaːn]
a	hat	[ˈhaṭt]
ɐ	Ober	[ˈ|oːḅbɐ]
ɐ̯	Uhr	[ˈ|uːɐ̯]
a͜i	weit	[ˈva͜iṭt]
a͜u	Haut	[ˈha͜uṭt]
ḅ	Liebe	[ˈliːḅbə]
b	Ball	[ˈbal]
ç	ich	[ˈ|ɪç]
d	dann	[ˈdan]
ḍ	Band	[ˈbanḍd]
d͜ʒ	Gin	[ˈd͜ʒɪn]
eː	Beet	[ˈbeːṭt]
e	Methan	[meˈṭtaːn]
ɛː	wähle	[ˈvɛːlə]
ɛ	hätte	[ˈhɛṭtə]
e͜ɪ	Frey	[ˈfre͜ɪ] [2]
ə	halte	[ˈhalṭtə]
f	Fass	[ˈfas]
g̣	Magen	[ˈag̣gn]
g	Gast	[ˈgasṭt]
g͜g	Rüegger	[ˈry͜əg̣g͜gər] [3]
h	hat	[ˈhaṭt]
iː	viel	[ˈfiːl]
i	vital	[viˈṭtaːl]
ɪ	bist	[ˈbɪsṭt]
i̯	Studie	[ˈʃṭtuːḍdi̯e]
i͜ə	Dietikon	[ˈdi͜əṭtiˌḳkoːn] [2]
j	ja	[ˈjaː]
k	Skandal	[sḳkanˈḍdaːl]
kʰ	kalt	[ˈkʰalṭt] [1]
ḳ	Macke	[ˈmaḳkə] [4]
k̤	Viktor	[ˈvɪk̤tʰoɐ̯] [5]
l	Last	[ˈlasṭt]
l̩	Nabel	[ˈnaːḅbl̩]
m	Mast	[ˈmasṭt]
m̩	grossem	[ˈgroːsm̩]

IPA	Example	
øː	Öl	[ˈ|øːl]
ø	Ökonom	[ˌ|øḳkoˈnoːm]
œ	göttlich	[ˈgœṭtlɪç]
p	Spatz	[ˈʃp̣paṭt͜s]
p͜f	Pfahl	[ˈp͜faːl]
pʰ	Pakt	[ˈpʰaḳkṭt] [1]
p̣	Mappe	[ˈmap̣pə] [4]
p̤	Abgabe	[ˈap̤gaːbə] [5]
r	Rast	[ˈrasṭt]
r͜r	Karren	[ˈkar͜rən]
s	Hast	[ˈhasṭt]
ʃ	Schal	[ˈʃaːl]
t	Stier	[ˈʃṭtiːɐ̯]
tʰ	Tal	[ˈtʰaːl] [1]
ṭ	Patt	[ˈpʰaṭtʰ] [4]
t̤	Mitbewohner	[ˈmɪt̤bəvoːnɐ] [5]
t͜s	Zahl	[ˈt͜saːl]
t͜ʃ	Matsch	[ˈmaṭt͜ʃ]
uː	Hut	[ˈhuːṭt]
u	kulant	[kuˈlanṭt]
ʊ	Pult	[ˈpʰʊlṭt]
u̯	aktuell	[aḳkˈṭtu̯ɛl]
ʊ͜ə	Ruedi	[ˈrʊ͜əḍdi] [2]
v	was	[ˈvas]
x	Bach	[ˈbax]
yː	Rübe	[ˈryːḅbə]
y	Mykene	[myˈḳkeːnə]
ʏ	füllt	[ˈfʏlṭt]
y̆	Etui	[|eˈṭty̆iː]
y͜ə	Blüemlisalp	[ˌbly͜əmlisˈalp̣p] [2]
z	Hase	[ˈhaːzə]
ʒ	Genie	[ʒeˈniː]
|	beamtet	[bəˈ|amṭtəṭt]

[1] aspirated plosive

English phone inventory

IPA	Example	
ə	another	[əˈnʌðə]
ə͜ʊ	nose	[ˈnə͜ʊz] [1]
æ	hat	[ˈhæṭt]
ɑ	got, frog	[ˈgɑṭt], [ˈfrɑg̣g] [2]
ɑː	stars	[ˈsṭtɑːz] [1], [ˈsṭtɑːrz] [2]
ʌ	cut, much	[ˈkʌṭt], [ˈmʌṭt͜ʃ]
a͜ɪ	rise	[ˈra͜ɪz]
a͜ʊ	about	[əˈḅba͜ʊṭt]
b	bin	[ˈbɪn]
ḅ	baby	[ˈbe͜ɪḅbɪ]
b̤	webpage	[ˈweb̤pe͜ɪd͜ʒ] [3]
ð	this, other	[ˈðɪs], [ˈʌðər]
d	din	[ˈdɪn]
ḍ	made	[ˈme͜ɪḍd]
d̤	Adkinson	[ˈˈɛd̤kinsən] [3]
d͜ʒ	Gin	[ˈd͜ʒɪn]
ɜː	bird, furs	[ˈbɜːḍd], [ˈfɜːz] [1]
ɜ	bird, furs	[ˈbɜrḍd], [ˈfɜrz] [2]
e	get	[ˈgeṭt]
e͜ɪ	raise	[ˈre͜ɪz]
ɛ͜ə	stairs	[ˈsṭtɛ͜əz] [1], [ˈsṭtɛ͜ərz] [2]
f	fit	[ˈfɪṭt]
g	give, bag	[ˈgɪv], [ˈbæg̣g]
g̣	beggar	[ˈbeg̣gər]
h	hit	[ˈhɪṭt]
ɪ	witch	[ˈwɪṭt͜ʃ]
iː	ease	[ˈiːz]
ɪ͜ə	fears	[ˈfɪ͜əz] [1], [ˈfɪ͜ərz] [2]
j	youth, yes	[ˈjuːθ], [ˈjes]
k	skat	[ˈsḳkɑːṭt]
kʰ	kin	[ˈkʰɪn]
ḳ	make	[ˈme͜ɪḳk]
k̤	acta	[ˈˈɛk̤tə] [3]
l	life, field	[ˈla͜if], [ˈfiːlḍd]
m	mean	[ˈmiːn]
ŋ	thing	[ˈθɪŋ]
n	fine, net	[ˈfa͜ɪn], [ˈneṭt]
ɔː	abroad	[əˈḅbrɔːḍd]

IPA	Example	
ɒ	got, frog	[ˈgɒṭt], [ˈfrɒg̣g]
o͜ʊ	nose	[ˈno͜ʊz] [2]
p	speed	[ˈspiːḍd
pʰ	pin	[ˈpʰɪn]
p̣	tip	[ˈtɪp̣p]
p̤	wept	[ˈwep̤t] [3]
r	ring, stress	[ˈrɪŋ], [ˈsṭtres]
ʃ	shine, brush	[ˈʃa͜ɪn], [ˈbrʌʃ]
s	sin, mouse	[ˈsɪn], [ˈma͜ʊs]
θ	thin, method	[ˈθɪn], [ˈmeθəḍd
t	street	[ˈsṭtriːṭt]
tʰ	time	[ˈtʰa͜ɪm]
ṭ	mat	[ˈmɛṭt]
t̤	Eastbourne	[ˈˈiːst̤boːn] [3]
t͜ʃ	chin	[ˈt͜ʃɪn]
ʊ	book	[ˈbʊḳk]
uː	lose	[ˈluːz]
ʊ͜ə	durable	[ˈdjʊ͜ərəḅbl]
v	very, heavy	[ˈverɪ], [ˈhevɪ]
w	well	[ˈwel]
x	loch	[ˈlɒx] [1]
ʒ	vision	[ˈvɪʒən]
z	zoo, fees	[ˈzuː], [ˈfiːz]

[1] British English
[2] American English
[3] preplosive pause before a
elision in rapid speech

Bibliography

[ANS94] ANSI. *American National Standard: Acoustical Terminology, ANSI S1.1-1994*. American National Standards Institute, 1994.

[AR76] B. Atal and L. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(3):201–212, 1976.

[AS99] S. Ahmadi and A.S. Spanias. Cepstrum-based pitch detection using a new statistical v/uv classification algorithm. *Speech and Audio Processing, IEEE Transactions on*, 7(3):333–338, 1999.

[BB06] C. L. Bennett and A. W. Black. The Blizzard Challenge 2006. In *Proc. Blizzard Challenge*, 2006.

[Bis95] C. M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[Boe02] P. P. G. Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2002.

[BS90] A. Bendiksen and K. Steiglitz. Neural networks for voiced/unvoiced speech classification. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 521–524. IEEE, 1990.

[BT05] A. W. Black and K. Tokuda. The Blizzard Challenge-2005: Evaluating corpus-based speech synthesis on common datasets. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[CB96] N. Campbell and A. W. Black. *Prosody and the selection of source units for concatenative synthesis.*, chapter 22, pages 279 – 292. Springer, 1996.

[CHL89] DG Childers, M. Hahn, and JN Larar. Silent and voiced/unvoiced/mixed excitation (four-way) classification of speech. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 37(11):1771–1774, 1989.

[CL02] V. Colotte and Y. Laprie. Higher precision pitch marking for TD-PSOLA. In *Proceedings of the European Signal Processing Conference*, pages 419–422, Toulouse, 2002.

[CL11] C. C. Chang and C. J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[CM89] F. Charpentier and E. Moulines. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In *Proceedings of Eurospeech*, pages 13–19, Paris, September 1989.

[dCK02a] A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[dCK02b] Alain de Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *JASA*, 111:1917–1930, 2002.

[DL93] T. Dutoit and H. Leich. MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, 13:435–440, 1993.

[Don01] R. Donovan. A new distance measure for costing spectral discontinuities in concatenative speech synthesisers. In *ISCA Tutorial and Research Workshop (ITRW)*, 2001.

[EHM+99] B. Etxebarria, I. Hernáez, I. Madariaga, E. Navas, JC Rodríguez, and R. Gándara. Improving quality in a speech synthesizer based on the mbrola algorithm. In *Proceedings of Eurospeech, Budapest*, pages 2299–2302, 1999.

[EHP09] T. Ewender, S. Hoffmann, and B. Pfister. Nearly perfect detection of continuous F0 contour and frame classification for TTS synthesis. In *Proc. of Interspeech*, pages 100–103, Brighton, September 2009.

[EP10] T. Ewender and B. Pfister. Accurate pitch marking for prosodic modification of speech segments. In *Proceedings of Interspeech*, pages 178–181, Makuhari (Japan), September 2010.

[EP11] T. Ewender and B. Pfister. Automatically creating a diphone set from a speech database. In *Proc. of Interspeech*, Florence, August 2011.

[Fah88] S. E. Fahlman. Faster-learning variations on back-propagation: An empirical study. In *Proceedings of the 1988 connectionist models summer school*, pages 38–51. Morgan Kaufmann, 1988.

[GBGM80] R. Gray, A. Buzo, A. Gray, Jr., and Y. Matsuyama. Distortion measures for speech processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):367–376, 1980.

[GJ94] Z. Ghahramani and M. I. Jordan. Supervised learning from incomplete data via an em approach. In *Advances in Neural Information Processing Systems 6*. Citeseer, 1994.

[GM76] A. Gray, Jr. and J. Markel. Distance measures for speech processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(5):380–391, 1976.

[GO07] L. Golipour and D. O'Shaughnessy. A new approach for phoneme segmentation of speech signals. In *Proceedings of Interspeech*, pages 1933–1936, 2007.

[Gol95] M. Goldstein. Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. *Speech communication*, 16(3):225–244, 1995.

[HCL$^+$] C. W. Hsu, C. C. Chang, C. J. Lin, et al. A practical guide to support vector classification.

[Hea10] J. Heaton. *Programming neural networks with encog 2 in Java*. Heaton Research, Inc., 2010.

[Hes83] W. Hess. *Pitch determination of speech signals: algorithms and devices*. Springer-Verlag Berlin and Heidelberg, 1983.

[Hes08] W.J. Hess. *Springer Handbook of Speech Processing*, chapter Pitch and voicing determination of speech with an extension toward music signals, pages 181–211. Springer, 2008.

[HK06] M. Hagmüller and G. Kubin. Poincaré pitch marks. *Speech Communication*, 48(12):1650–1665, 2006.

[HL02] C. W. Hsu and C. J. Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[Hou63] A. S. House. Psychoacoustic speech tests: A modified rhyme test. Technical report, DTIC Document, 1963.

[HP10] S. Hoffmann and B. Pfister. Fully automatic segmentation for prosodic speech corpora. In *Proceedings of Interspeech*, pages 1389–1392, Makuhari (Japan), September 2010.

[JBB07] D. Joho, M. Bennewitz, and S. Behnke. Pitch estimation using models of voiced speech on three levels. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 1077–1080, Honolulu, Hawaii, USA, April 2007.

[JC10] A. C. Janska and R. A. J. Clark. Native and non-native speaker judgements on the quality of synthesized speech. In *Proc. Interspeech*. Citeseer, 2010.

[JJ94] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214, 1994.

[Jon03] D. Jones. *Cambridge English Pronouncing Dictionary*. Cambridge University Press (ISBN 0-521-01712-2), 16th edition, 2003.

[Kae85] H. Kaeslin. Systematische Gewinnung und Verkettung von Diphonelementen für die Synthese deutscher Standardsprache, 1985.

[KdC05] H. Kawahara and A. de Cheveigné, et al. Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. In *Proceedings of Interspeech*, pages 537–540. ISCA, 2005.

[KEP09] T. Kaufmann, T. Ewender, and B. Pfister. Improving broadcast news transcription with a precision grammar and discriminative reranking. In *Proceedings of Interspeech*, pages 356–359, Brighton (United Kingdom), September 2009.

[KHK09] B. Kotnik, H. Höge, and Z. Kacic. Noise robust F0 determination and epoch-marking algorithms. *Signal Processing*, 89(12):2555 – 2569, 2009.

[KMT$^+$08] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *Proceedings of ICASSP*, pages 3933–3936, 2008.

[KV98] E. Klabbers and R. Veldhuis. On the reduction of concatenation artefacts in diphone synthesis. In *International Conference on Spoken Language Processing ICSLP*, volume 98, pages 1983–1986. Citeseer, 1998.

[LB00] K. A. Lenzo and A. W. Black. Diphone collection and synthesis. In *Proceedings of ICSLP*, 2000.

[Lev44] K. Levenberg. A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, 5:164–168, 1944.

[LGP89] J. S. Logan, B. G. Greene, and D. B. Pisoni. Segmental intelligibility of synthetic speech produced by rule. *The Journal of the Acoustical Society of America*, 86(2):566–581, 1989.

[Lip87] R. Lippmann. An introduction to computing with neural nets. *ASSP Magazine, IEEE*, 4(2):4–22, 1987.

[Lis63] L. Lisker. Cross-Language Study of Voicing in Initial Stops. *JASA*, 35:384–422, 1963.

[LJ04] C. Y. Lin and J. S. R. Jang. A two-phase pitch marking method for TD-PSOLA synthesis. In *Proceedings of Interspeech/ICSLP*, pages 1189–1192, Jeju Island (Korea), October 2004.

[LL03] A.P. Lobo and P.C. Loizou. Voiced/unvoiced speech discrimination in noise using gabor atomic decomposition. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–820. IEEE, 2003.

[LM96] P. Ladefoged and I. Maddieson. *The Sounds of the World's Languages*. Blackwell Publishers, 1996.

[Mar63] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[MC90] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453 – 467, December 1990.

[MHMH07] KI Molla, K. Hirose, N. Minematsu, and K. Hasan. Voiced/unvoiced detection of speech signals using empirical mode decomposition model. In *Information and Communication Technology, 2007. ICICT'07. International Conference on*, pages 311–314. IEEE, 2007.

[MIH$^+$11] M. Molla, K. Islam, K. Hirose, S.K. Roy, and S. Ahmad. Adaptive thresholding approach for robust voiced/unvoiced classification. In *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, pages 2409–2412. IEEE, 2011.

[MJ08] A. E. Mahdi and E. Jafer. Two-feature voiced/unvoiced classifier using wavelet transform. *The Open Electrical and Electronic Engineering Journal*, 2:8–13, 2008.

[MNH98] J. Murata, T. Noda, and K. Hirasawa. Artificial neural networks with input gates. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 1, pages 480–485. IEEE, 1998.

[Møl93] M. F. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, 6(4):525–533, 1993.

[MVV06] W. Mattheyses, W. Verhelst, and P. Verhoeve. Robust pitch marking for prosodic modification of speech using TD-PSOLA. In *Proc. of the IEEE Benelux/DSP Valley Signal Processing*, 2006.

[NSRK85] N. Nocerino, F. K. Soong, L. R. Rabiner, and D. H. Klatt. Comparative study of several distortion measures for speech recognition. *Speech Communication*, 4(4):317–331, 1985.

[Ohm94] H. Ohmura. Fine pitch contour extraction by voice fundamental wave filtering method. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume 2, pages II–189. IEEE, 1994.

[PK08] B. Pfister und T. Kaufmann. *Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. Springer Verlag (ISBN: 978-3-540-75909-6), 2008.

[Pre98] L. Prechelt. Early stopping-but when? *Neural Networks: Tricks of the trade*, pages 553–553, 1998.

[QB82] S. R. Quackenbush and T. P. Barnwell. An Analysis of Objectively computable Measures for Speech Quality Testing, 1982.

[QBL04] F. Qi, C. Bao, and Y. Liu. A novel two-step SVM classifier for voiced/unvoiced/silence classification of speech. In *2004 International Symposium on Chinese Spoken Language Processing*, 2004.

[QH93] Y. Qi and B.R. Hunt. Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier. *Speech and Audio Processing, IEEE Transactions on*, 1(2):250–255, 1993.

[RB93] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *Neural Networks, 1993., IEEE International Conference on*, pages 586–591. Ieee, 1993.

[RM99] R. D. Reed and R. J. Marks, II. *Neural Smithing*. The MIT Press, 1999.

[Rom09] H. Romsdorfer. *Polyglot Text-to-Speech Synthesis: Text Analysis & Prosody Control*. PhD thesis, No. 18210, ETH Zurich. Shaker Verlag Aachen (ISBN 978-3-8322-8090-1), February 2009.

[RYG$^+$06] J. Ramírez, P. Yélamos, J. M. Górriz, J. C. Segura, and L. García. Speech/non-speech discrimination combining advanced feature extraction and svm learning. In *Proceedings of Interspeech 2006*, Pittsburgh, (USA), September 2006. ISCA.

[SAMW89] M. Spiegel, M. J. Altom, M. Macchi, and K. Wallace. A monosyllabic test corpus to evaluate the intelligibility of synthesized and natural speech. In *Speech Input/Output Assessment and Speech Databases*, 1989.

[SB82] L. Siegel and A. Bessey. Voiced/unvoiced/mixed excitation classification of speech. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 30(3):451–460, 1982.

[SB00] K. Sjölander and J. Beskow. Wavesurfer-an open source speech tool. In *Sixth International Conference on Spoken Language Processing*, 2000.

[Sie79] L. Siegel. A procedure for using pattern classification techniques to obtain a voiced/unvoiced classifier. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(1):83–89, 1979.

[Sim08] S. Simonet. Detektion und Elimination von Störgeräuschen bei Frikativlauten. Semester thesis, 2008.

[SISY04] J.K. Shah, A.N. Iyer, B.Y. Smolenski, and R.E. Yantorno. Robust voiced/unvoiced classification using novel features and gaussian mixture model. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 17–21, 2004.

[SJ84] F. Soong and B. Juang. Line spectrum pair (LSP) and speech data compression. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, volume 9, pages 37–40. IEEE, 1984.

[SN04] E. Skovenborg and S. H. Nielsen. Evaluation of different loudness models with music and speech material. In *Proceedings of the 117th AES convention*, October 2004.

[SN06] J. Sundberg and M. Nordenberg. Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *J. Acoust. Soc. Am.*, 120(1):453–457, 2006.

[Sot82] J. Sotschek. Ein Reimtest für Verständlichkeitsmessungen mit deutscher Sprache als ein verbessertes Verfahren zu Bestimmung der Sprachübertragungsgüte. *Der Fernmeldeingenieur*, 55, Heft 4/5, April 1982.

[SQN04] E. Skovenborg, R. Quesnel, and S. H. Nielsen. Loudness assessment of music and speech. In *Proceedings of the 116th AES convention*, May 2004.

[SS00] M. Sakamoto and T. Saitoh. An automatic pitch-marking method using wavelet transform. In *Proceedings of Interspeech/ICSLP*, Beijing, September 2000.

[SS01] Y. Stylianou and A. K. Syrdal. Perceptual and objective detection of discontinuities in concatenative speech synthesis. In *Proceedings of ICASSP*, 2001.

[Tal95] D. Talkin. A robust algorithm for pitch tracking (RAPT). In *Klein, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis.*, pages 495–518, Elsevier Science B.V., Amsterdam, 1995.

[TH99] C. Traber, K. Huber, et al. From multilingual to polyglot speech synthesis. In *Proceedings of Eurospeech'99*, pages 835–838, Budapest, September 1999.

[TR91] D. Talkin and J. Rowley. Pitch-synchronous analysis and synthesis for its systems. In *The ESCA Workshop on Speech Synthesis*, 1991.

[Tra95] C. Traber. *SVOX: The Implementation of a Text-to-Speech System for German.* PhD thesis, No. 11064, Computer Engineering and Networks Laboratory, ETH Zurich, TIK-Schriftenreihe Nr. 7 (ISBN 3 7281 2239 4), March 1995.

[Vap99] V. N. Vapnik. An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10(5):988–999, 1999.

[Vel00] R. Veldhuis. Consistent pitch marking. In *Proceedings of Interspeech/ICSLP*, pages 207–210, Beijing, September 2000.

[VK03] R. Veldhuis and E. Klabbers. On the computation of the Kullback-Leibler measure for spectral distances. *Speech and Audio Processing, IEEE Transactions on*, 11(1):100–103, 2003.

[VKT02] J. Vepa, S. King, and P. Taylor. Objective distance measures for spectral discontinuities in concatenative speech synthesis. In *Proc. ICSLP*. Citeseer, 2002.

[VV05] M. Viswanathan and M. Viswanathan. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech & Language*, 19(1):55–83, 2005.

[WM98] J. Wouters and M. Macon. A perceptual evaluation of distance measures for concatenative speech synthesis, 1998.

[YRG$^+$06] P. Yélamos, J. Ramírez, J. Górriz, C. Puntonet, and J. Segura. Speech event detection using support vector machines. *Computational Science–ICCS 2006*, pages 356–363, 2006.

[ZF99] E. Zwicker and H. Fastl. *Psychoacoustics.* Springer-Verlag, Berlin, second edition edition, 1999.