Text-Independent Voice Conversion

David Sündermann

**Der Fakultät für Elektrotechnik und Informationstechnik
der Universität der Bundeswehr München
zur Erlangung des akademischen Grades eines**

**Doktors der Ingenieurwissenschaften
(Dr.-Ing.)**

vorgelegte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr.-Ing. Berthold Lankl
1. Berichterstatter:	Prof. Dr. phil. nat. Harald Höge
2. Berichterstatter:	Prof. Dr.-Ing. habil. Rüdiger Hoffmann

Tag der Prüfung: 11. Juli 2008

Neubiberg, der 11. Juli 2008

Abstract

This thesis deals with text-independent solutions for voice conversion. It first introduces the use of vocal tract length normalization (VTLN) for voice conversion. The presented variants of VTLN allow for easily changing speaker characteristics by means of a few trainable parameters. Furthermore, it is shown how VTLN can be expressed in time domain strongly reducing the computational costs while keeping a high speech quality.

The second text-independent voice conversion paradigm is residual prediction. In particular, two proposed techniques, residual smoothing and the application of unit selection, result in essential improvement of both speech quality and voice similarity.

In order to apply the well-studied linear transformation paradigm to text-independent voice conversion, two text-independent speech alignment techniques are introduced. One is based on automatic segmentation and mapping of artificial phonetic classes and the other is a completely data-driven approach with unit selection. The latter achieves a performance very similar to the conventional text-dependent approach in terms of speech quality and similarity. It is also successfully applied to cross-language voice conversion.

The investigations of this thesis are based on several corpora of three different languages, i.e., English, Spanish, and German. Results are also presented from the multilingual voice conversion evaluation in the framework of the international speech-to-speech translation project TC-Star.

Acknowledgments

My entire and perfect appreciation is due Prof. Dr. Harald Höge. Over the four years of this project and before, he played the outstanding roles of my teacher, my supervisor, my colleague, my financial backer, my patron, and my friend. Without him, most of the last years' exciting work would not have been possible.

This thesis brought me and my blessed family around the world: We lived in the Netherlands, in Barcelona, Munich, Los Angeles, and New York City, calling for a lifestyle that was only possible due to the extraordinary flexibility and absolute endorsement of my wife Christiane Sündermann-Oeft and the enthusiasm for aviation and foreign languages of my daughters Nele Theophanu, Hannah Eulàlia, and Kara-Lailach Igone.

Many researchers supported this dissertation scientifically, financially, and ideally – I am wholeheartedly indebted to Profs. Drs. Antonio Bonafonte, Julia Hirschberg, Shri Narayanan, and Hermann Ney for giving me the opportunity to work in their distinguished research laboratories. Further services for the successful completion of this work were rendered by my former teacher Prof. Dr. Rüdiger Hoffmann kindly serving as reviewer of this work, by Dr. Jaka Smrekar who provided valuable mathematical assistance, by Dr. Renko Geffarth proofreading the manuscript, and by my parents Sabine and Dr. Uwe Sündermann.

New York City, October the 28th, 2007

Contents

1 Introduction

Voice conversion is a technology that transforms characteristics of a source voice to those of a target voice. This means, we are given a source voice (in terms of recorded speech) and want it to be converted to another voice, the target voice whose characteristics are given either by recorded speech as well or by more general descriptors as the gender or age of the speaker, parameters as mean fundamental frequency or speaking rate, or just by the requirement of being sufficiently different from the source.

In addition to this vague definition of the target, we face vagueness in what we refer to as voice characteristics. Here we have two different levels:

- *voice quality* (or timbre, the set of segmental cues) describes the voice's sound influenced by formant locations and bandwidths, spectral tilt, and the contribution of the voice excitation by the vocal folds [Childers & Lee 91],

- *prosody* (or the set of supra-segmental cues) is related to the style of speaking and includes phoneme duration, evolution of fundamental frequency (intonation) and energy (stress) over the utterance [Horne 00].

This dissertation deals with two specific tasks out of the set of possibilities introduced above:

1. A source voice given by recorded speech is to be transformed so that the converted voice sounds different from the source in a specific way (e.g. change of gender).

2. A source voice given by recorded speech is to be transformed so that the converted voice sounds similar to a target voice given by recorded speech as well.

As will be discussed below, the voice conversion technology described in this dissertation is applied to both the intra-lingual and the cross-language task where source and target voice use different languages. Since prosody is highly language-dependent [Dutoit 93, Boudelaa & Meftah 96, Barbosa 97, Zhu & Zhang$^+$ 02], it is a very hard task to convert a speaker's prosodical characteristics across languages[1]. This is the reason for omitting prosodical conversion in this work. The only prosodical property to be transformed is the mean fundamental frequency having strong influence on the perception of a voice as shown in Section 3.4. Since voice quality seems to be much more independent of the language than prosody, this dissertation focuses on the transformation of voice quality[2].

Approaches presented in this thesis are evaluated using state-of-the-art error measures as discussed in Section 2.5. They deal with the success of voice conversion, i.e. with the

[1] Recent investigations in the framework of speech-to-speech translation use prosodical features of the source to produce a more natural target prosody across languages [Agüero & Adell$^+$ 06]. However, here, the target is synthetic speech produced by a text-to-speech system, and the proposed approach cannot be applied to natural speech, since it does not aim at detecting parallelisms between source and target.

[2] There is evidence that also certain voice quality aspects vary between languages [Braun & Masthoff 02], but this effect is much weaker than the language dependence of prosody.

similarity of converted and target voice, as well as with the overall quality of the converted voice. The present work reports on both types of measures, those based on human listening tests and also automatic ones.

1.1 Text-Independent and Cross-Language Voice Conversion

Looking at task **2** introduced above, there is no statement about the properties of the recorded speech being required. Obviously, characteristics as how much speech is available, whether it is based on phonetically balanced text, whether the recording is of high quality (in terms of signal-to-noise ratio, sampling frequency, and quantization), and whether the speakers are professional, are of vital importance for a successful voice conversion. However, as we will learn in Section 2.1, the most crucial point is that both source and target speech are time-aligned. Such an alignment can be produced by means of well-investigated techniques when both speech samples involved are based on the same text. This case is referred to as *text-dependent* as opposed to the case of the texts being different; this is called *text-independent*.

The task becomes even more complicated if the involved speech samples' languages differ. Here, we mostly face different phoneme sets, thus we are unable to find an appropriate time alignment to phonemes that are missing in the other phoneme set. Therefore, one also distinguishes between *intra-lingual* and *cross-language* voice conversion.

1.2 Application Areas of Voice Conversion

In literature, a vast number of applications for voice conversion is given. Obviously, only a few of them were ever carried out. Therefore, the author restricts his considerations to the most popular ones:

A. Definitely, the most important application of voice conversion is the use as a module in text-to-speech synthesis where we want to change the standard speaker's voice characteristics towards a well-defined target speaker [Kain 01]. This is used to rapidly adapt, personalize, or corporatize synthesized voices, e.g. in dialog systems [Fischer & Kunzmann 06].

B. Moreover, voice conversion has been applied to normalizing high quality speech databases to enlarge the amount of available speech data to build a concatenative speech synthesis system [Eide & Picheny 06].

C. In embedded environments, voice conversion can serve to manipulate voices in such a way that the original speaker cannot be recognized or the speaker's gender or age is changed. Due to computational limitations, conversion algorithms of embedded systems focus on manipulating only a few parameters as fundamental frequency and vocal tract length [Sündermann & Ney$^+$ 03b].

D. Cross-language voice conversion has been applied to dubbing tasks in the film and music industry [Turk & Arslan 02].

E. In the framework of speech-to-speech translation, a translated sentence is to be uttered with the source speaker's voice [Höge 02]. Here, in addition to speech recognition, machine translation, and speech synthesis, a text-independent cross-language voice conversion module is necessary.

F. For people who are speech-impaired or -disabled voice conversion can be a means leading to more natural and intelligible speech as for persons suffering dysarthria [Hosom & Kain$^+$ 03] or laryngectomy [Nakamura & Toda$^+$ 06].

G. Recently, voice conversion has been applied to speaker normalization for speech recognition. In particular, it was used for reducing the Lombard effect describing a voice's change in noisy environments, in order to improve the recognition performance [Bořil & Fousek$^+$ 06].

1.3 Objectives and Contributions

The main objectives of this work are

- to develop a technique for producing a target voice that is sufficiently different from the source voice and can be easily controlled by means of few parameters. It is to be investigated how many well-distinguishable voices can be created from a single source voice without affecting the naturalness of the resulting voice. Furthermore, the technique is to be optimized in terms of time and memory complexity to be applied in an embedded environment (application **C**),
- to improve the state of the art of the converted voice's naturalness and similarity to the target by investigating residual prediction techniques (applications **A**, **B**, **D**, **E**),
- to produce solutions to voice conversion that are text-independent which is mandatory for its application to speech-to-speech translation (**E**) and convenient for most of the aforementioned applications (**A**, **B**, **D**, **F**, **G**),
- to investigate solutions to cross-language portability that is crucial for applying voice conversion to speech-to-speech translation (**E**) and other tasks (**D**).

The main contributions of this work are

- the application of vocal tract length normalization (VTLN) to voice conversion,
- the generalization of VTLN warping functions by means of a piece-wise linear function with several segments which allows for applying dynamic programming to estimate its parameters,
- to prove that VTLN which is normally applied in frequency domain can be applied directly in time domain omitting two Fourier transformations and, hence, may strongly accelerate the voice conversion – the author refers to this concept as *time domain VTLN*,
- to show that VTLN-based voice conversion is able to produce several well-distinguishable voices (five or more) based on one source voice by varying two parameters (warping factor and mean fundamental frequency),

- two residual prediction techniques (one based on residual smoothing, the other based on unit selection) that outperform the state of the art in terms of overall speech quality and similarity to the target,
- a text-independent speech alignment technique based on unit selection that can be used as a preprocessor of conventional (text-dependent) voice conversion techniques,
- to show that the aforementioned text-independent speech alignment technique is robust with respect to the cross-language task and produces higher speech quality than conventional text-dependent alignment (dynamic time warping).

2 State of the Art

Having a look at an excerpt of the call for papers of the world's largest speech processing conference, Interspeech, held August 2007 in Antwerp, Belgium, one finds, among others, the following research fields:

a. speech coding,

b. speech synthesis,

c. speech recognition,

d. cross-lingual processing,

e. speaker recognition,

f. evaluation and standardization.

This dissertation deals with text-independent voice conversion, a topic positioned between all of these fields, borrowing and joining knowledge from all of them and making voice conversion an interdisciplinary research area. In particular, it loans

- the source-filter model, linear prediction, and residual processing from **a**,
- acoustic synthesis, the unit selection paradigm, and the approach of pitch-synchronous processing from **b**,
- vocal tract length normalization and the Gaussian mixture model from **c**,
- the concepts of text-independence and cross-language portability from **d**,
- linear transformation, vector quantization, and clustering from **e**,
- the concepts of objective and subjective evaluation as well as several standardized error measures from **f**.

This chapter is to introduce the aforementioned concepts building a bridge to their application to voice conversion. Furthermore, the respective application's state of the art in voice conversion research is briefly presented.

2.1 Text-Independent Voice Conversion

Speech is the human being's original communication medium. It carries threefold information:

- segmental information (related to voice quality),

- supra-segmental information (related to prosody),
- linguistic information (expressed by the series of phonemes uttered).

The first two of these, segmental and supra-segmental information, are related to voice characteristics introduced in Chapter 1, whereas the third is not to be covered under the scope of voice characteristics in this dissertation[3]. Although linguistic information is not focus of this work, its presence plays an important role for the following investigations. Most of the state-of-the-art voice conversion techniques which aim at converting a source towards a given target speaker (task **2** of Chapter 1) work in two phases, the training and the conversion phase. In training, speech of both the source and target voices is processed, and useful information (parameters) is extracted. These parameters are used in the conversion phase to change the voice characteristics of a new source utterance to sound similar to the target voice.

When introducing the speech production model in Section 2.3 it will be argued that the vocal tract shape is responsible for both

- the individual sounds (phonemes) produced by the speaker and
- an essential part of the speaker's voice characteristics related to voice quality.

Voice-quality-related information is extracted by comparing speech data of source and target speaker. In doing so, one has to compensate for variations which are due to the phonemes uttered. This compensation can be done by using *time-aligned* speech, i.e., where both speakers produced the same phonemes at exactly the same time.

The easiest way to achieve such an alignment is the text-dependent approach. Here, one uses speech of both involved voices based on the same text (also referred to as *parallel speech*). To produce an exact time alignment, there are standard techniques as dynamic time warping [Rabiner & Rosenberg$^+$ 78], or hidden-Markov-model-based forced alignment [Young & Woodland$^+$ 93] (which tends to be more exact than dynamic time warping [Inanoglu 03]), or a combination of both [Duxans & Bonafonte 03]. Forced alignment requires the underlying text as well as acoustic and language models, i.e., it is language-dependent, as opposed to dynamic time warping which is language-independent.

As argued in Section 1.3, several applications require a text-independent alignment, i.e., the speakers' utterances are based on different texts (i.e. *non-parallel* data). Interestingly, when the author started his investigations on text independence for voice conversion, to the best of his knowledge, there were no publications dealing with this issue yet. To clearly distinguish between his work and that of other researchers, he decided to discuss his first contribution, the automatic segmentation and mapping of artificial phonetic classes [Sündermann & Ney 03a], in Section 5.1 and dedicated the current section to another recent work on text independence of voice conversion.

The automatic segmentation and mapping does not require any phonetic knowledge about the considered language, i.e., it is language-independent. This fact is advantageous on the one hand, since the technique can be applied to arbitrary languages without additional data. On the other hand, more information about the phonetic structure of the processed speech could lead to a more reliable mapping between source and target speech.

[3] There are indeed opinions partially relating linguistic information (as the speaker's accent) to voice characteristics [Kain 01].

Consequently, [Ye & Young 04b] proposed to use a speaker-independent hidden-Markov-model-based speech recognizer to label each frame[4] with a state index such that each source or target speaker utterance is represented by a state index sequence. If the text of these utterances is known, this can be done by means of forced alignment resulting in more reliable state sequences.

In a second step, subsequences are extracted from the set of target sequences to match the given source state index sequences using a simple selection algorithm. This algorithm favors longer matching sequences to ensure a continuous spectral evolution of the selected target speech frames. The latter are derived from the state indices considering the frame–state mapping delivered by the speech recognizer. The result is two sequences of parallel speech frames.

2.2 Cross-Language Voice Conversion

The very first investigations on cross-language voice conversion in the beginning of the 1990s focused on the speech-to-speech translation task [Abe & Shikano+ 90]. At that time, ATR (Advanced Telecommunications Research Institute International in Kyoto, Japan) where the authors of the latter paper were working was developing a so-called *interpreting telephone*. This was the name of a speech-to-speech translation system applied to telephone conversations. It integrated a cross-language voice conversion module to preserve speaker recognizability across languages.

This first attempt was based on a codebook mapping that used a discrete representation of the acoustic feature space. To the best of the author's knowledge, there were no investigations carried out dealing with the technique's speech quality. [Stylianou & Cappé+ 95] claimed that there were considerable artifacts due to the discreteness of the codebook mapping approach in order to promote their linear transformation technique, see Section 2.4.2.

Besides, it was not sufficiently shown whether the codebook mapping approach is able to successfully convert voice characteristics. The results reported were based on objective error metrics that are not standardized and sometimes hardly correlate with the perceptive similarity, cf. Section 2.5. Subjective experiments using the described codebook mapping technique reported successful gender transformation from male to female and 61% successfully transformed examples for male-to-male conversion using an ABX[5] test [Abe & Nakamura+ 88]. Although this approach was text-independent, it did not produce an alignment between source and target speech. This is the reason for not considering this technique in Section 2.1.

More than a decade later, Japanese researchers (some of them also at ATR) continued the investigations on cross-language voice conversion and applied the linear transformation paradigm introduced in Section 2.4.2 [Mashimo & Toda+ 01]. They avoided the text independence problem by using bilingual (Japanese/English) speakers as source speakers. The conversion function was trained on parallel Japanese utterances of source and target speakers and applied to English source speech in conversion phase. The only difference to text-dependent intra-lingual voice conversion are the different phoneme sets of source and target language. The corresponding intra-lingual baseline system described in [Toda & Lu+ 00] achieved a fair speech quality (mean opinion score 2.9) and a conversion performance of about 90% on an ABX scale (for these evaluation metrics, see Section 2.5).

[4] For a definition of the term *speech frame*, see Section 2.3.1.

[5] For the definition of an ABX test, see Section 2.5.2.

For projects like the interpreting telephone, TC-Star[6], or Minnesang[7], the text-dependent solution to cross-language voice conversion is hardly applicable for the following reasons:

- At least one of the involved speakers would have to speak both source and target language to apply the text-dependent training.
- However, this would not be the source speaker, since otherwise there were no need for speech-to-speech translation.
- The target speaker is a synthetic voice generally based on a unit selection concatenative text-to-speech system involving a large speech corpus ($\geq$ 10h) of a professional and carefully selected speaker [Black & Lenzo 01]. It would be a severe restriction to the speaker selection to demand bilingual speakers. If the speech-to-speech translation system is used in several languages, one would require either a professional speaking all languages to be converted, or one would have to build a new bilingual voice for all language combinations to be considered.
- The introduction of a new language would mean to build a new text-to-speech voice from scratch.
- Since the system is text-dependent, a set of utterances based on a predefined text would be required by each source speaker who wants to use the system.

All of these drawbacks are to be overcome by the text-independent and cross-language solutions discussed in Chapter 5.

2.3 The Speech Production Model

2.3.1 Speech as a Sequence of Frames

Once again, the three types of information carried by human speech are to be consulted:

- segmental information (related to voice quality),
- supra-segmental information (related to prosody),
- linguistic information (expressed by the series of phonemes uttered).

Looking at a single utterance, voice quality can be regarded as constant or, at least, slowly changing, whereas phonemes change rapidly over time[8]. To capture all necessary information of a time-varying speech signal, it is therefore necessary to split the signal into small portions (frames) which themselves can be regarded as stationary. The smaller such a frame is, the more stationary are its contents.

Still, there is a natural lower limit to the frame size coming from the third speech information type, the prosody. Prosody covers the evolution of fundamental frequency for a

[6] See Section A.1.

[7] In the Minnesang project [Spelmezan & Borchers 06], an exhibition's visitor listens to his voice speaking a medieval German poem after having uttered a short verse in his mother tongue.

[8] As an example, [Hain 01] reports around 10 phonemes per second for British English spontaneous speech.

typical adult female reaching from 165 to 255Hz and for a typical adult male from 85 to 155Hz [Dolson 94]. Using smaller frames than those given by the signal's periodicity would mean to lose not only the fundamental frequency but also to reduce the amount of information contained in a frame to such a degree that feature extraction methods as linear prediction would fail, cf. Section 2.3.3. Typically, in speech coding and recognition, frame lengths between 10 and 30ms are used, e.g. 20ms for the GSM and UMTS codecs [Hillebrand 02], and 25ms for many speech recognizers, see e.g. [Ney & Welling$^+$ 98] or [Furui & Nakamura$^+$ 06].

For both mentioned fields, speech coding and recognition, state-of-the-art techniques are based on constant frame lengths [Dharanipragada & Gopinath$^+$ 98], whereas in speech synthesis, the frame lengths are usually linked to the fundamental frequency (pitch). That is, they make use of the pitch-synchronous paradigm which allows for applying standard pitch and duration modification techniques as pitch-synchronous overlap and add (PSOLA) [Charpentier & Stella 86]; for details refer to Section 2.3.4.

Due to the excitation of the voice by the periodically vibrating vocal fold during voiced sounds (see Section 2.3.2 for the source-filter model), speech can be regarded a pseudo-periodic signal in voiced regions. The pitch-synchronous paradigm is to cut the signal into frames each of which contains such a period. The automatic detection of the cutting points is referred to as pitch marking [Cheng & O'Shaughnessy 89] which turns out to be a rather difficult task keeping a number of researchers busy even nowadays [Bernadin & Foo 06, Germann 06, Kotnik & Höge$^+$ 06, Mattheyses & Verhelst$^+$ 06]. In unvoiced regions, the pitch marks are interpolated between neighbored voiced regions, or constant frame durations (e.g. 10ms) are assumed [Black & Lenzo 03].

Frequently discussed is the question which cutting point within the pseudo-periodic signal is the most reliable one while, at the same time, it maximally correlates with the periodicity of the laryngeal excitation. [Höge & Kotnik$^+$ 06] found that "the time instant at which the lower margins of the vocal folds are touching each other can be uniquely and consistently determined on the basis of the speech signal itself. Therefore, [...] the most negative peak of the speech signal will be defined as the [...] starting point of each new pitch period". Figure 2.1 shows three periods of a speech signal and the respective cutting points according to this definition.

At the beginning of his work on voice conversion, the author made use of the algorithm for "accurately marking pitch pulses in speech signals" by [Goncharoff & Gries 98] which is based on dynamic programming. In a recent evaluation [Kotnik 06], it was shown that this algorithm was clearly outperformed by the Praat algorithm working in the lag (autocorrelation) domain. Praat was claimed "to be several orders of magnitude more accurate than the methods commonly used" according to the algorithm's creator [Boersma 01].

The Praat software comes along with a voicing detector and only produces pitch marks in voiced signal parts. In unvoiced regions, pitch marks have to be padded as explained above, whereas the Goncharoff/Gries algorithm automatically produces pitch marks for the whole signal including unvoiced portions. However, for several investigations in this work (as for instance on PSOLA, cf. Section 2.3.4), it is necessary to be explicitly provided with voicing information. Thus, in connection with the Goncharoff/Gries pitch marking algorithm, the voicing detector used for the Adaptive Multi-Rate speech codec [ETSI 99] was used. This algorithm does not only *detect* but it also provides a continuous level of voicing $0 \leq v \leq 1$ where $v = 0$ is completely unvoiced and $v = 1$ is completely voiced. Such a continuous level of voicing will be of particular interest for the residual smoothing technique discussed in Section 4.1.

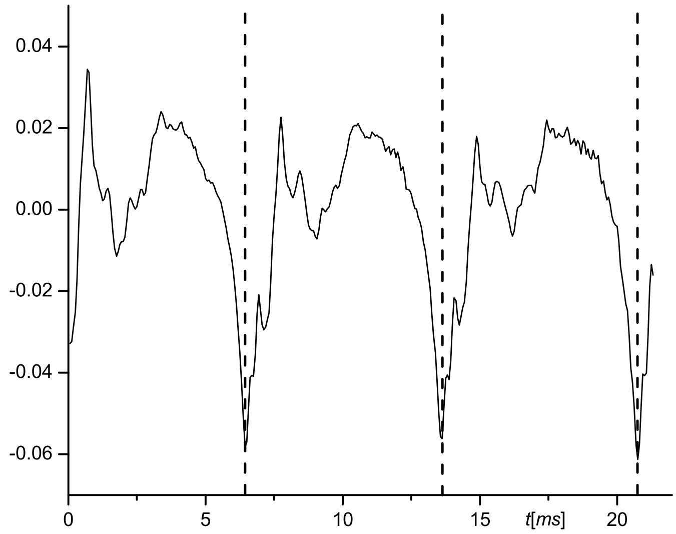

Figure 2.1: Example of a voiced speech portion of three pitch periods and the respective pitch marks (vertical lines).

2.3.2 The Source-Filter Model

The human production of speech is based on an air flow coming from the lungs and passing the whole vocal tract to mouth and nose where the speech sound is emitted. This sound is excited either by the periodically vibrating vocal fold resulting in voiced sounds or by flow turbulences at constrictions in the vocal tract producing frication noise or plosives. In addition to the excitation, the vocal tract shape plays an important role generating well-distinguishable sounds. This shape can be varied by the position of the tongue, opening or closing the mouth, lowering the soft palate (for producing nasal sounds), etc. Besides, the vocal tract shape is highly speaker-dependent [Nolan 83], a fact that will be exploited in most of the following investigations. For a sketch of the human vocal tract, see Figure 2.2.

Since using different vocal tract shapes but a constant excitation results in clearly different sounds (phonemes), human speech production was understood to be of a source-filter type where the signal source is the excitation and the filter is represented by the vocal tract [Fant 70]. This interpretation was the motivation to search for a mathematical model appropriately describing the production of human speech leading to the concept of linear prediction described in the following section.

2.3.3 Linear Prediction and Residual

Historically, linear prediction [Markel & Gray 76] is one of the most important speech analysis and generation techniques. It is based on the source-filter model where the filter is assumed to be all-pole, i.e., it produces only poles and no zeroes. This allows for predicting a sample by calculating a weighted sum of the past samples which also explains the term linear prediction.

The condition of being all-pole comes from the assumption that the vocal tract can

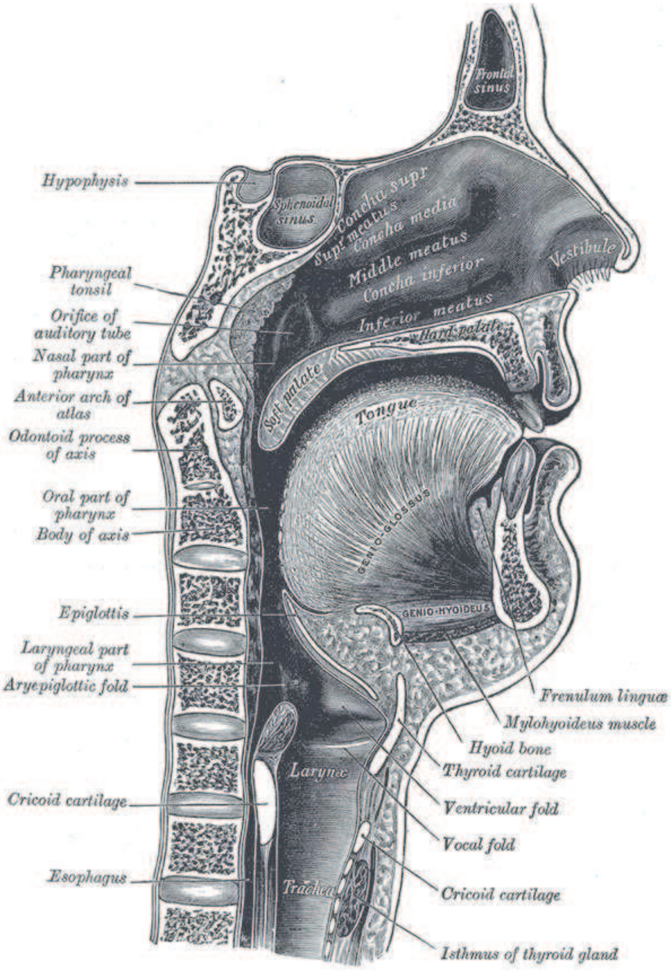

Figure 2.2: The human vocal tract, in [Gray 18].

be reasonably well modeled by a series of concatenated uniform lossless cylindrical acoustic tubes. The human vocal tract and the tube model differ in several aspects, though:

- the vocal tract is not built of cylinders,
- the vocal tract is not lossless,
- the vocal tract has a side passage (the nasal cavity).

However, the long experience in speech processing has shown that in spite of these drawbacks, linear prediction produces sufficiently reliable results applied to a variety of areas as speech coding, speech recognition, or voice conversion. This holds in particular when the linear prediction order (the number of filter coefficients or cylindrical tubes) is sufficiently large leading to a reasonable approximation of the vocal tract [Fallside & Woods 85].

As discussed in Section 2.3.1, in this work, the speech signal is assumed to be a sequence of frames. For each of these speech frames, a set (or vector) of linear predictive coefficients (also referred to as features), i.e., the weights mentioned above, is estimated. This estimation (or feature extraction) is done in a way that the coefficients produce an optimal prediction of the respective frame's samples. Here, optimality is defined as the minimization of the error, i.e. the differences between the predicted and the actual signal, also referred to as *residual*. Feeding the residual to the linear prediction model, i.e. filtering the residual by means of the above coefficients, always gives back the original signal [Smith 07]. Therefore, the residual stands for the excitation and the linear predictive filter for the vocal tract of the source-filter model of Section 2.3.2.

The error minimization for estimating the linear predictive coefficients is carried out using the ancient least squares approach [Gauss 09], according to which the minimization is applied to the sum of the squared sample differences. [Makhoul 75] shows that this minimization can be performed by applying the Levinson-Durbin algorithm [Durbin 60] to the autocorrelation sequence of the frame's power spectrum.

In addition to linear predictive coefficients, a number of other feature types has been widely used for speech processing:

- **Line spectral frequencies [Itakura 75].** When being applied to linear-transformation-based voice conversion linear predictive coefficients as defined above tend to produce artifacts that can be effectively suppressed by converting the coefficients to line spectral frequencies. This is due to their superior interpolation properties as compared to other linear prediction representations [Paliwal 95], a fact that is exploited in the linear transformation introduced in Section 2.4.2. An algorithm for converting linear predictive coefficients to line spectral frequencies and vice versa is given e.g. in [Deller & Proakis$^+$ 93].

- **Mel frequency cepstral coefficients [Picone 93].** This type of feature provides a compact representation of the speech amplitude spectrum in a form which is anatomically and perceptually motivated. Like in the cochlea, the amplitude spectrum is scaled according to the mel scale [Stevens & Volkman$^+$ 37] and filtered by means of a number of (triangular) overlapping bandpass filters. Now, the discrete cosine transformation is applied to the list of mel amplitudes finally producing the features which are the amplitudes of the resulting spectrum. According to [Ye & Young 04a], mel frequency coefficients, formerly used in the linear transformation framework [Stylianou & Cappé$^+$ 95], are outperformed by line spectral frequencies in terms of speech quality.

- **Cubic-spline-interpolated cepstrum [Ye & Young 03].** Here, the logarithmized spectral amplitudes are resampled to a certain number of frequency points which then are transformed to the cepstral domain. A similar feature type was used by the author in an investigation on text-independent voice conversion [Sündermann & Bonafonte$^+$ 04a], but [Ye & Young 04a] themselves worked out that these features produce a lower speech quality than line spectral frequencies.

For these reasons, line spectral frequencies will be used as features in this work unless otherwise noted.

2.3.4 Pitch-Synchronous Overlap and Add

Now, having introduced linear prediction as a principal feature extraction method, the far end of frame-based speech processing, the acoustic synthesis, is to be discussed. Purpose of the acoustic synthesis is to concatenate a sequence of speech frames and potentially change the underlying fundamental frequency or duration maintaining a good speech quality. At this point, it shall be mentioned that pitch modification is mostly limited to voiced speech portions, since its application to unvoiced portions may produce a certain level of voicing coming from the windowing and, in case of raising the pitch, repetition of frames [Ceyssens & Verhelst$^+$ 02]. Consequently, in unvoiced regions, the signal is simply copied.

Above all, the following acoustic synthesis techniques are discussed in literature:

- **Harmonic sinusoidal model [Macon 96].** This model represents a short segment of speech (a pitch-synchronous frame) by adding up a number of sinusoids (sine waves) with certain amplitudes and phases whose frequencies are integer multiples of the fundamental frequency. Extending the model by the assumption that every speech frame can be composed of a voiced and unvoiced spectral component separated by a voicing-dependent cutoff frequency, it is said to be capable of high-quality time and pitch modifications [Stylianou 01].

- **Multi-band re-synthesis overlap and add [Dutoit & Pagel$^+$ 96].** This algorithm is based on the multi-band excitation [Griffin 87] model allowing for spectral interpolation between voiced signal parts, though it is performed in time domain. However, unlike the harmonic sinusoidal model and PSOLA (described below), it does not require a preliminary marking of pitch periods. Due to a PSOLA-related patent of France Télécom, multi-band re-synthesis overlap and add is not a free algorithm; the respective software is only available as binary code of a complete speech synthesizer back-end, making it difficult to be applied to this work.

- **Pitch-synchronous overlap and add (PSOLA) [Charpentier & Stella 86].** The PSOLA algorithm is based on the pitch-synchronous paradigm introduced in Section 2.3.1. It is still probably the most popular acoustic synthesis technique in the speech processing community and will also be used in this work. Below, the fundamentals of PSOLA will be briefly introduced, and three different types will be presented.

For the following considerations, a frame is to be composed of two pitch periods, the frames' overlap is one period. I.e., when $p_1^{M+1} = p_1, p_2, \dots, p_{M+1}$ is the sequence of considered pitch periods then

$$x_1^M = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}, \begin{pmatrix} p_2 \\ p_3 \end{pmatrix}, \dots, \begin{pmatrix} p_M \\ p_{M+1} \end{pmatrix} \tag{2.1}$$

is the sequence of frames[9]. This avoids signal discontinuities in the overlap-and-add concatenation, see [Kain 01].

- **Time domain PSOLA [Hamon & Moulines$^+$ 89].** Given the time waveform of the frames to be processed, a Hanning window [Oppenheim & Schafer 89] is applied

$$x'(t) = \frac{x(t)}{2}\Big(1 - \cos\Big(2\pi\frac{t}{T}\Big)\Big), \; 0 \le t \le T, \tag{2.2}$$

[9] Here, the notation $x = \begin{pmatrix} p \\ q \end{pmatrix}$ means that vectors p and q, consisting of the speech samples of the respective frames, are concatenated yielding vector x.

and the frames are overlapped. Here, T is the frame's duration. Remember that each frame consists of two pitch periods, and the standard overlap is one period. Looking at the successive frames $x_m = \begin{pmatrix} p_m \\ p_{m+1} \end{pmatrix}$ and $x_{m+1} = \begin{pmatrix} p_{m+1} \\ p_{m+2} \end{pmatrix}$, the overlapping period is p_{m+1}. It is windowed by the falling half of the Hanning window for x_m and by the rising half for x_{m+1} [10]. Adding together both contributions produces

$$\begin{aligned} p'_{m+1}(t) &= \frac{p_{m+1}(t)}{2}\left(1-\cos\left(2\pi\frac{t+\frac{T}{2}}{T}\right)\right)+\frac{p_{m+1}(t)}{2}\left(1-\cos\left(2\pi\frac{t}{T}\right)\right) \\ &= \frac{p_{m+1}(t)}{2}\left(2-\cos\left(2\pi\frac{t}{T}+\pi\right)-\cos\left(2\pi\frac{t}{T}\right)\right) \\ &= p_{m+1}(t) \qquad (2.3) \end{aligned}$$

which leaves the signal unchanged in the standard case and, consequently, does not produce additional distortions. However, this does not apply if the fundamental frequency or the timing is changed. This is done by shifting the overlapping frames so that the resulting period length becomes

$$\tilde{T} = \frac{T}{\rho} \quad \text{with} \quad \rho = \frac{\bar{f}_{0,\mathrm{t}}}{\bar{f}_{0,\mathrm{s}}} \qquad (2.4)$$

where $\bar{f}_0$ is the mean fundamental frequency observed in the training data and s and t denote source and target, respectively. In order to keep or consciously change the temporal evolution of the speech signal, time frames have to be repeated or deleted, respectively. Figure 2.3 demonstrates the principle of TD-PSOLA for decreasing and increasing fundamental frequency.

- **Frequency domain PSOLA [Charpentier & Moulines 88].** If the considered frames are given in frequency domain, i.e. as magnitude and phase spectra, one can apply an interpolation to them as discussed in Section 2.4.3 to match the target number of spectral lines given by Equation 2.4. When one now returns to time domain using inverse discrete Fourier transformation the standard case introduced for time domain PSOLA applies, and the additional signal distortion is negligible. Here, the main signal deterioration is due to the interpolation.

- **Linear predictive PSOLA [Moulines & Charpentier 88].** If the residual time waveform is given as in the case described in Section 2.4.3, one can apply time domain PSOLA directly to the residual and filter the result by means of the corresponding frames' linear predictive coefficients derived from the features. The advantage compared to the application of time domain PSOLA to the signal after filtering is that the spectral distortions at formant frequencies are lower (time domain PSOLA has a bandwidth broadening effect) [de los Galanes & Savoji$^+$ 94].

  If the residual magnitude and phase spectra are given as e.g. in Section 2.4.3, another approach is to interpolate and join these spectra and multiply them with the spec-

[10] Since the length of the considered pitch periods is subject to change from m to $m+1$ to $m+2$ depending on the fundamental frequency, the term *half* is an approximation.

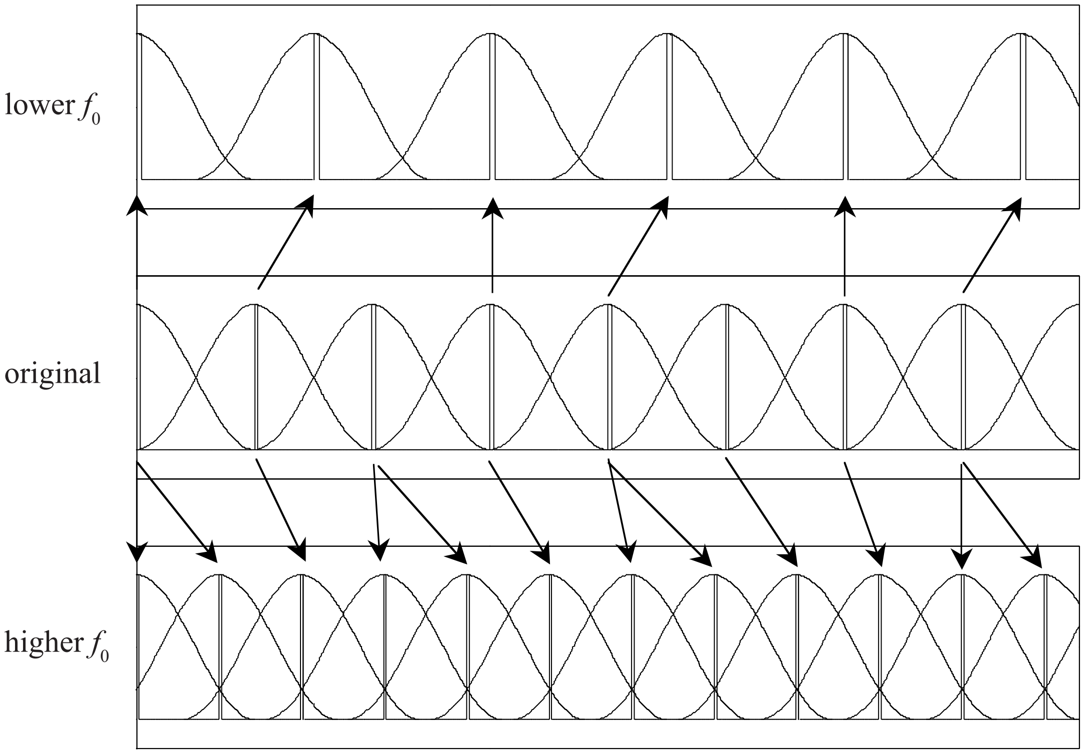

Figure 2.3: TD-PSOLA: Examples for decreasing and increasing the fundamental frequency. The figure shows how the Hanning windows overlap and how frames are repeated or deleted, respectively.

tra directly derived from the transformed feature vectors[11]. The remaining steps are equivalent to frequency domain PSOLA.

2.4 Changing Voice Quality: VTLN, Linear Transformation, and Residual Prediction

2.4.1 Vocal Tract Length Normalization

According to the source-filter model introduced in Section 2.3.2, one of the two components responsible for the speaker-dependent voice quality is the vocal tract shape. An approach to converting this shape is to change its *length*. This approach was first used in speech recognition to compensate for vocal tract length variations and is referred to as vocal tract length normalization (VTLN) [Cohen & Kamm$^+$ 95]. Here, VTLN is applied to speaker normalization and adaptation which can yield a considerable recognition performance gain [Pye & Woodland 97].

[11] For an algorithm to compute the spectrum from linear predictive coefficients, see [Markel & Gray 76] or [Rabiner & Juang 93].

[Eide & Gish 96] investigated the change of the vocal tract length based on the tube model described in Section 2.3.2. They showed that a division of the vocal tract length (that of the tube sequence) l by a factor α such that $\tilde{l} = \frac{l}{\alpha}$ means that the frequency axis of a sound's spectrum generated by such a vocal tract is warped reciprocally according to $\tilde{\omega} = \alpha\omega$. α is usually referred to as warping factor. Consequently, VTLN is performed by applying a linear warping function to the spectrum of the speech frames. Further investigations into VTLN have shown that there are more powerful warping functions than the purely linear one. Some of them are discussed in the following.

In speech recognition, only the *magnitude* spectrum is processed, since the later featurization (into mel frequency cepstral coefficients or perceptual linear predictive coefficients [Hermansky 90, Hermansky & Morgan 94]) neglects the phase spectrum which does not seem to be significant for recognition. For its use in voice conversion, however, the phase spectrum also plays an important role to produce natural speech and, hence, should also be subject to warping[12].

Let $0 \leq \omega \leq \pi$ be the normalized, continuous frequency, then an arbitrary warping function g that depends on a set of parameters $\{\xi_1, \xi_2, \ldots\}$ is applied to ω yielding the scaled frequency

$$\tilde{\omega} = g(\omega|\xi_1, \xi_2, \ldots) \text{ with } 0 \leq \tilde{\omega} \leq \pi. \tag{2.5}$$

Accordingly, when a given frame's (magnitude) spectrum as a function of ω is referred to as $X(\omega)$ [13] and the warped counterpart is $\tilde{X}$ then one obtains the following equality taking into account that the magnitude of the warped spectrum at the warped frequency $\tilde{\omega}$ is to equal the source magnitude at the original frequency ω:

$$\tilde{X}(\tilde{\omega}) = X(\omega). \tag{2.6}$$

In literature on speech recognition, there is a variety of warping functions, most of them based on just one parameter, the warping factor α. Table 2.1 gives an overview on commonly used functions. Figure 2.4 shows an example frame's magnitude spectrum warped by a bilinear warping function.

In speech recognition, the determination of the parameters (mostly the warping factor α), requires the execution of a forced Viterbi alignment on the whole training data involved for each parameter candidate [Lee & Rose 96]. The candidate minimizing the score is selected.

Referring to *candidates* suggests that the considered warping factors are taken from a discrete set, in relevant literature ([Welling & Ney$^+$ 02, Molau 02]) given by[14]

$$\alpha \in \{0.88, 0.9, \ldots, 1.12\}. \tag{2.7}$$

Restricting the set to these 13 candidates is supposed to reduce the computational effort for the parameter estimation and turns out to be a reasonable setting for the use in speech recognition. In Section 3.1, we will see that when applying VTLN to voice conversion upper and lower limit of the warping factor have to be altered.

[12] For details on phase interpolation, see Section 2.4.2.

[13] In digital speech processing, we do not have continuous spectra but, due to discrete time sampling and application of the discrete Fourier transformation, discrete spectra. However, for the current considerations and to simplify matters, we assume ω and $X(\omega)$ to be continuous. Discrete counterparts of the following derivations can be computed using interpolation methods, e.g. cubic splines [de Boor 78].

[14] This holds for the piece-wise linear warping function with two segments.

function	definition	reference
piece-wise linear (two segments)	$g_1(\omega\|\alpha) = \begin{cases} \alpha\omega : & \omega \leq \omega_0 \\ \alpha\omega_0 + \dfrac{\pi - \alpha\omega_0}{\pi - \omega_0}(\omega - \omega_0) : & \omega > \omega_0 \end{cases}$	
– asymmetric	$\omega_0 = \dfrac{7}{8}\pi$	[Wegmann & McAllaster+ 96]
– symmetric	$\omega_0 = \begin{cases} \frac{7}{8}\pi : & \alpha \leq 1 \\ \frac{7}{8\alpha}\pi : & \alpha > 1 \end{cases}$	[Uebel & Woodland 99]
power	$g_2(\omega\|\alpha) = \pi \left(\dfrac{\omega}{\pi}\right)^{\alpha}$	[Eide & Gish 96]
quadratic	$g_3(\omega\|\alpha) = \omega + \alpha \left(\dfrac{\omega}{\pi} - \left(\dfrac{\omega}{\pi}\right)^2\right)$	[Pitz & Ney 05]
all-pass transform	$g_4(\omega\|\alpha, \beta_1^I, \gamma_1^I) = g(z^{-1}(\tilde{z}(z(\omega))))$ with $z(\omega) = e^{\imath\omega} \quad \longrightarrow \quad z^{-1}(\zeta) = \omega = -\imath \log \zeta$ and $\tilde{z}(z) = \dfrac{z-\alpha}{1-\alpha z} \prod_{i=1}^{I} \left(\dfrac{z-\beta_i}{1-\beta_i^* z} \cdot \dfrac{z-\beta_i^*}{1-\beta_i z}\right) \left(\dfrac{1-\gamma_i^* z}{z-\gamma_i} \cdot \dfrac{1-\gamma_i z}{z-\gamma_i^*}\right)$ where β_i and γ_i are complex parameters with $\|\beta_i\|, \|\gamma_i\| < 1$.	[McDonough & Byrne+ 98]
bilinear	$\tilde{z}(z) = \dfrac{z-\alpha}{1-\alpha z}$ (special case of the all-pass transform, $I = 0$)	[Acero & Stern 91]
combinations	e.g. $g_5(\omega\|\alpha_1, \alpha_2) = g_2(g_1(\omega\|\alpha_1)\|\alpha_2)$	[Molau & Kanthak+ 00]

Table 2.1: Overview on warping functions.

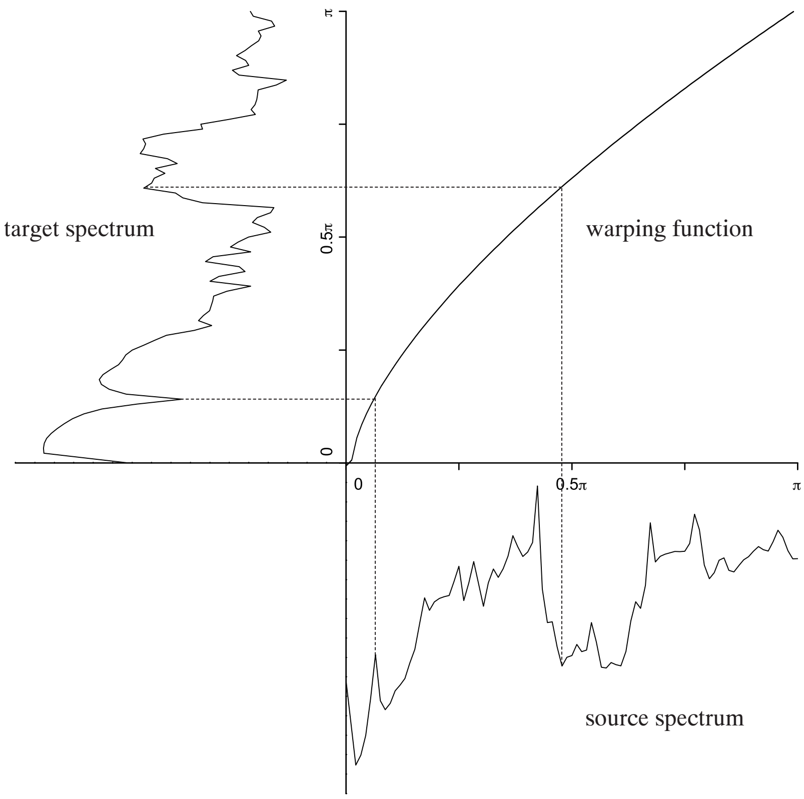

Figure 2.4: Example of a magnitude spectrum warped by a bilinear warping function.

2.4.2 Voice Conversion Based on Linear Transformation

The very first approaches to voice conversion [Abe & Nakamura$^+$ 88] used a discrete representation of the spectral envelopes of source and target speech. Each envelope was represented by a spectral vector clustered (quantized) into a certain number of artificial phonetic classes[15]. This approach generated considerable artifacts which, even after further refinements, could not be sufficiently suppressed [Kuwabara & Sagisaka 95].

Linear-transformation-based voice conversion [Stylianou & Cappé$^+$ 95] significantly reduces the shortcomings of the above described method by using a Gaussian mixture model [Duda & Hart 73], a formalism that models a probability density function as a sum of multivariate normal distributions. It has been very successfully applied to modeling speech given by spectral vectors. In literature, one finds applications to speech recognition [Duda & Hart 73], language identification [Zissman 93], speaker recognition [Reynolds 95], speaking rate estimation [Faltlhauser & Pfau$^+$ 00], gender classification [Tranter & Reynolds 04], and more. Applying the Gaussian mixture model to voice conversion

- produces continuous (smooth) classification indices as opposed to the vector quantization approach where discontinuities may occur when a vector jumps from one class to the other and

[15] In this dissertation, speech sounds (phones) are also referred to as *classes*. Natural phonetic classes are those which are recognized as distinctive by the International Phonetic Association and transcribed in their International Phonetic Alphabet [Ladefoged 90]. Artificial phonetic classes are sounds which are automatically clustered and do not necessarily represent a well-defined sound. They are not required to be perceptively distinctive.

- represents classes not only by their mean vectors but as a whole probability distribution including the covariance matrices containing more precise information about the voice characteristics.

After more than a decade, this approach is still state-of-the-art, as the vast number of recent publications suggests: E.g. at Interspeech, held in September 2006 in Pittsburgh, 10 articles on Gaussian-mixture-model-based voice conversion were published [Lee & Wu 06, Nakagiri & Toda$^+$ 06, Nakamura & Toda$^+$ 06, Nicolao & Drioli$^+$ 06, Nurminen & Tian$^+$ 06, Ohtani & Toda$^+$ 06, Sündermann & Höge$^+$ 06, Tian & Nurminen$^+$ 06, Toda & Ohtani$^+$ 06, Uto & Nankaku$^+$ 06].

Linear-transformation-based voice conversion uses the source-filter model introduced in Section 2.3.2 separating vocal tract (represented by feature vectors) from excitation (represented by the residual) as discussed in Section 2.3.3. The features are line spectral frequencies derived from the respective frame's spectrum X and arranged as the D-dimensional feature vector[16] x, see Section 2.3.3. Usually, x's dimensionality D is much smaller than that of X which is of importance for the robustness and mathematical stability of the statistical analysis explained below. Due to the lower dimensionality, x does not contain all the information contained in X but is a spectral envelope representation of X comprising its formant structure. This behaviour is particularly interesting for the application to voice conversion, since the spectral envelope is mainly influenced by the speaker-dependent vocal tract to be transformed.

Let us now consider two time-aligned[17] feature vector sequences x_1^M (source) and y_1^M (target) as training data. At first, a Gaussian mixture model is fitted to the source vectors, each mixture density $i \in \{1, \ldots, I\}$ representing an artificial phonetic class assumed to be Gaussian distributed. This is done by applying the expectation-maximization algorithm [Hogg & McKean$^+$ 95] which yields the parameters α_i (mixture weights, $\sum_i \alpha_i = 1; \alpha_i \geq 0$), μ_i (mean vectors), and Σ_i (covariance matrices) describing the probability density

$$p(x) = \sum_{i=1}^{I} \alpha_i \mathcal{N}(x|\mu_i, \Sigma_i) \tag{2.8}$$

where $\mathcal{N}(x|\mu, \Sigma)$ denotes the D-dimensional normal distribution

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\Big(-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\Big). \tag{2.9}$$

Now, the conditional probability of a mixture density i given an arbitrary source vector x is derived by

$$p(i|x) = \frac{p(x,i)}{p(x)} = \frac{\alpha_i \mathcal{N}(x|\mu_i, \Sigma_i)}{\sum\limits_{j=1}^{I} \alpha_j \mathcal{N}(x|\mu_j, \Sigma_j)}. \tag{2.10}$$

Here, $p(x, i)$ is the probability that x is generated by the i^{th} mixture.

[16] In this thesis, the variable x is used in two contexts: It either stands for a *time signal*, as for instance in the discussions of VTLN-based voice conversion, or a *feature vector*, as in the present case. In no case, both contexts will be used in parallel.

[17] As introduced in Section 2.1.

The voice transformation function F that converts a given source vector x to a vector $\tilde{x}$ supposed to represent spectral contents similar to the target voice is expressed by a linear transformation

$$\tilde{x} = F(x) = Ax + b \tag{2.11}$$

or, considering mixture-dependent transformations[18],

$$F(x) = F(x|A_1^I, b_1^I) = \sum_{i=1}^{I} p(i|x)(A_i x + b_i). \tag{2.12}$$

Finally, the unknown parameters A_i, matrices of size $D \times D$, and b_i, vectors of dimensionality D are to be estimated based on the least squares technique where the sum of the squared distances between the transformed vectors $\tilde{x}_m = F(x_m|A_1^I, b_1^I)$ and the aligned target vectors y_m for $m \in \{1, \dots, M\}$ is minimized[19]:

$$\sum_{m=1}^{M} |y_m - F(x_m|A_1^I, b_1^I)|^2 = \min! \tag{2.13}$$

In [Stylianou & Cappé$^+$ 98], a solution to this least square problem is carried out showing that A_1^I and b_1^I can be expressed as a linear combination of α_1^I, μ_1^I, and Σ_1^I.

In conversion phase, a new source speaker utterance is featurized, the resulting feature vector sequence is converted using the voice transformation function F. Section 2.4.3 is dedicated to how to transform the resulting feature vectors to time domain.

A schematic overview of the conversion phase components of a voice conversion system based on linear transformation is given in Figure 2.5. The input (source) speech is processed by a pitch marking (and voicing detection) algorithm and accordingly segmented into frames as described in Section 2.3.1. Features (linear predictive coefficients) are computed and converted to a better representation (line spectral frequencies), cf. Section 2.3.3. Now, the linear-transformation-based conversion function is applied based on the Gaussian mixture model parameters estimated in training phase as described above. The converted line spectral frequencies are converted back to linear predictive coefficients. After producing suitable residuals by means of residual prediction (see Section 2.4.3), coefficients and residuals are combined using linear predictive PSOLA generating the final waveform, cf. Section 2.3.4. In doing so, the target fundamental frequency is derived from the source fundamental frequency based on a prosody transformation which in this thesis is the transformation of the mean fundamental frequency as discussed in Chapter 1.

2.4.3 Residual Prediction

In Section 2.4.2, the paradigm of linear-transformation-based voice conversion was introduced dealing with the filter part of the source-filter model. In order to produce the final waveform, this filter has to be excited with an appropriate residual signal as introduced in Section 2.3.3. The determination of a proper residual for each processed speech frame is referred to as residual prediction [Kain & Macon 01].

[18] This is, since the voice transformation function is expected to be class-dependent and, consequently, mixture-dependent, as the mixtures provide an artificial phonetic distinction; see also Footnote 15.

[19] $|x|$ is the l^2-norm, i.e. the square root of the squares of x's elements.

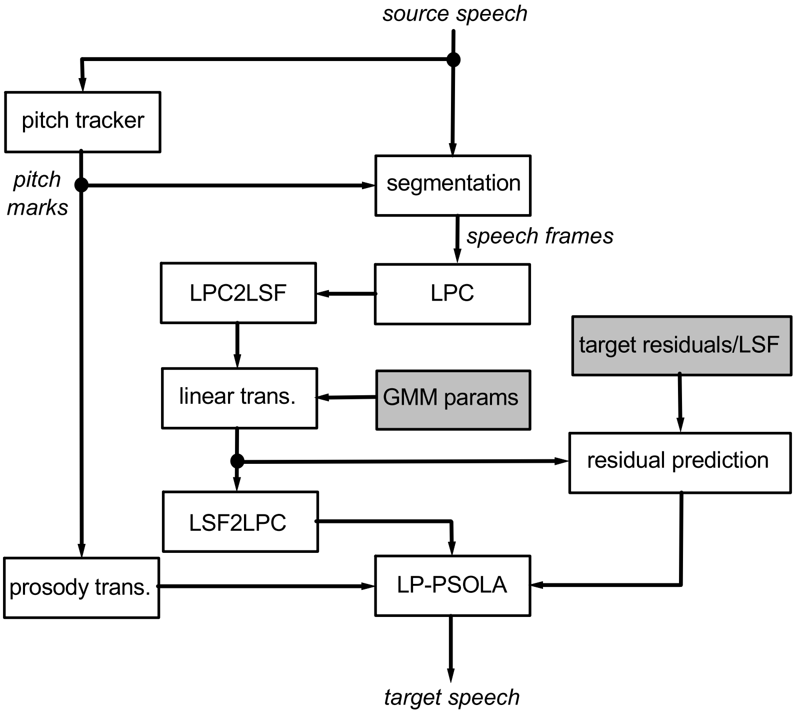

Figure 2.5: Components of a voice conversion system based on linear transformation in conversion phase. LPC stands for linear predictive coding/coefficients; LSF for line spectral frequencies; GMM for Gaussian mixture model.

In addition to its application to voice conversion, residual prediction is also useful for hidden-Markov-model-based speech synthesis [Masuko 02] where a sequence of feature vectors is generated which have to be excited by means of appropriate residuals. However, so far, in synthesis, mostly the vocoder approach is used.

In the following, several solutions to residual prediction are briefly described:

- **Vocoder.** The vocoder approach [Weinstein 75] is based on the standard formulation of linear predictive coding [Atal & Hanauer 71] where source signals are either a white noise or a pulse train thus resembling unvoiced or voiced excitations of the vocal tract, respectively. Both, white noise and pulse train have a flat (white) magnitude spectrum, but the former's phase spectrum is randomized, and the latter's is zero. A more advanced approach uses a mixed excitation model with continuous voicing degrees for several spectral bands [Cho & Kim$^+$ 98]. Drawback of the vocoder approach is the synthetic sound of the voice which is to be overcome by techniques involving natural residual waveforms.

  Recent advances to the vocoder in the STRAIGHT project, particularly focused on an enhanced excitation model, resulted in speech sounds which "are sometimes indistinguishable from the original speech sounds in terms of their naturalness" [Banno & Hata$^+$ 07]. These enhancements included a careful modeling of the signal's aperiodicities in conjunction with mixed-mode excitation and group delay manipulation [Kawahara & Estill$^+$ 01]. STRAIGHT, however, makes use of not only a white excitation but also processes the source residuals.

- **Copying source residuals.** When considering residual prediction for its application to voice conversion we are given natural speech as input whose time-alignment to the output speech is well known. Hence, for every output frame, at least the corresponding input residual is known. According to the ideal source-filter model, one might expect most of the speaker-dependent information to be represented by the vocal tract and, consequently, by the features. The excitation and, consequently, the residuals, might be expected to be less crucial for the voice characteristics, thus, finally, a simple solution would be to directly use the source as target residuals and apply the transformed features. This technique was used by [Kain & Macon 98], but they stated that "merely changing the [spectral envelope] is not sufficient for changing the speaker identity". Most of the listeners "had the impression that a 'third' speaker was created".

- **Residual codebook method.** In addition to the observation that the residual signal also contains speaker-dependent information, it turns out that spectral features and corresponding residuals are correlated [Kain 01]. This insight led to the idea that the residuals of the converted speech could be predicted based on the converted feature vectors and resulted in the following residual prediction technique.

  In training phase, similar to Section 2.4.2, the probability distribution of the target feature vectors y_1^M is modeled by a Gaussian mixture model. Here, the features are a cepstral representation of linear predictive coefficients [Kain 01]. The corresponding residual waveforms are converted to frequency domain, split into magnitude and phase spectra and normalized to a constant number of spectral lines expressed by the norm fundamental frequency f_n set to 100Hz as proposed by [Ye & Young 04a]. This is done by means of cubic spline interpolation [de Boor 78]. In doing so, one has to unwrap the phase spectra which are only given in the principal representation that is modulo 2π [McAulay & Quatieri 86, Yang & Koh$^+$ 93]. Another slightly different approach is to perform the interpolation in the complex domain, i.e. interpreting the spectral lines as two-dimensional vectors to be connected by cubic splines [Sündermann & Bonafonte$^+$ 04a].

  Now, typical residual magnitude spectra $\hat{r}_i$, referred to as *residual codebook*, are determined for each mixture density $i \in \{1, \ldots, I\}$ by computing a weighted average of all residual magnitude spectra r_m seen in training where the weights are the posterior probabilities $p(i|y_m)$ that a given feature vector y_m belongs to the mixture density i:

  $$\hat{r}_i = \frac{\sum\limits_{m=1}^{M} r_m p(i|y_m)}{\sum\limits_{k=1}^{M} p(i|y_k)}. \tag{2.14}$$

  During conversion phase, for each frame, we obtain a converted cepstral vector $\tilde{x}$ that serves as basis for the prediction of the residual magnitude spectrum $\tilde{r}$ by calculating a weighted sum over all mixture densities:

  $$\tilde{r} = \sum_{i=1}^{I} \hat{r}_i p(i|\tilde{x}). \tag{2.15}$$

  To predict the phase spectrum, a similar approach is used that does not express the spectrum as a sum of spectra, but selects the most likely entry of the codebook according to the posterior probability $p(i|\tilde{x})$ and finally applies a smoothing to avoid phase discontinuities.

- **Residual selection** The residual codebook method tries to represent an arbitrary residual by a linear combination of a limited number of prototype residuals. To better model the manifold characteristics of the residuals, [Ye & Young 04a] introduced the residual selection technique.

  All residuals r_m seen in training are stored into a table together with the corresponding feature vectors y_m that this time are composed of line spectral frequencies and their deltas. Given the transformed feature vector $\tilde{x}$, in conversion phase, a residual is selected from the table by minimizing the square error between $\tilde{x}$ and all feature vectors seen in training:

  $$\tilde{r} = r_{\tilde{m}} \quad \text{with} \quad \tilde{m} = \underset{m\ =\ 1,\ldots,M}{\arg\min} |\tilde{x} - y_m|. \tag{2.16}$$

  For the phase spectra, a similar technique as for the residual codebook method is used. Both approaches suffer from the inconsistency of the treatment of magnitude and phase spectra, resulting in audible phase mismatches.

2.5 Evaluation Metrics for Voice Conversion

When assessing the performance of a voice conversion technique one mostly focuses on two criteria:

- speech quality of the converted speech and
- speech similarity, i.e. the similarity of the converted source and the target.

To measure quality or similarity of the converted speech, there are two kinds of evaluation paradigms used in this work: *objective* and *subjective* evaluation.

Objective criteria are based on calculations on features derived from the assessed speech and are computed automatically. They feature the strong advantage of repeatability, i.e., they always produce the same result given the same data, and they are cheap for being executed on a computer. However, and details are given below, often they are not able to sufficiently represent the perception of a human listening to the same speech, and for certain questions, e.g. for rating the speech quality, there are no reliable objective metrics available.

Subjective criteria, on the other hand, are based on the opinion of human listeners (subjects) making these criteria more realistic, since the consumers of voice conversion technology are humans. Unfortunately, subjective criteria are very instable, they depend on[20]:

- The subject's mood. In a large number of subjective tests, the author observed the following effect: The test was always put on a web interface accessible via Internet. The test subjects were mostly colleagues from different institutions the author has worked at in Germany, Spain, and the USA who were asked for participation by an e-mail from the author.

  Some of the e-mail's addressees immediately performed the test, meaning they were willing and having sufficient time to participate. Since a subjective test is the more

[20] This list is not supposed to be exhaustive.

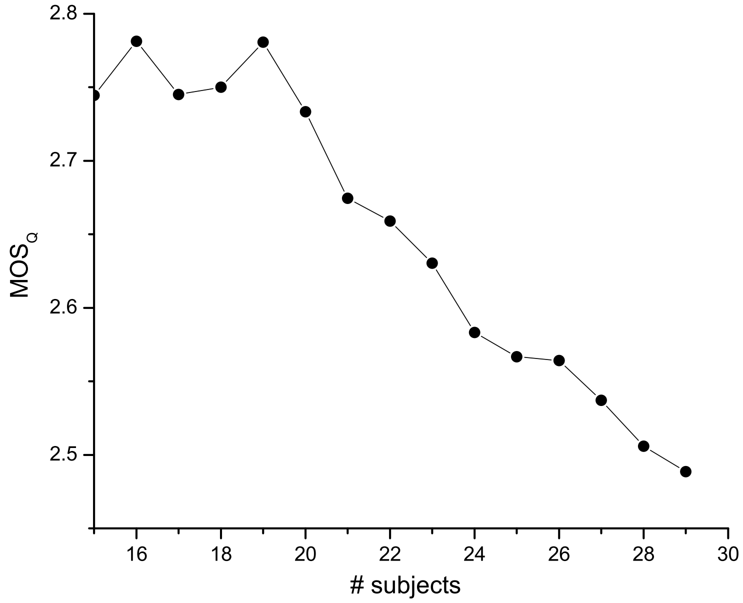

Figure 2.6: Mean opinion score depending on the number of subjects in the experiment.

significant the more people take part, the author used to ask the remaining colleagues by means of a second and later a third e-mail, and finally, he received data from most of them, even from those who were feeling disturbed by the test or being pressed for time due to their work.

It turns out that, consistently over most of the tests carried out in the framework of this dissertation, taking into account only the immediate responses produces significantly better outcomes than using the whole number of participants[21]. This is due to the deteriorating mood and attitude of the late participants.

Figure 2.6 shows an example of the mean opinion score[22] on speech quality averaged over all submissions to the experiment on residual prediction techniques discussed in Section 4.1.3 taking into account more and more subjects. The mean opinion score is almost 2.8 for 15 subjects and decreases to less than 2.5 for the final number of subjects (29).

- The subject's familiarity with speech technology. Professionals tend to be less critical than naïve subjects, see [Bennett 05].

- The characteristics of other speech samples assessed at the same time. If two speech samples are assessed in one test, the rating tends to be relative, even though the subject was asked to rate in absolute terms. I.e., if someone is asked to rate technique A and B where A is the technique in question and B is significantly worse than A the latter usually is rated higher than if B were clearly superior. An example of this contrasting effect is given in Section 5.2.3.

[21] To avoid misunderstandings: The results presented in this dissertation are, without exception, based on the average over **all** participants of the respective experiments.

[22] For the definition of mean opinion score, see Section 2.5.3.

	quality	similarity
objective	–	log-spectral distortion
subjective	mean opinion score	extended ABX, mean opinion score

Table 2.2: Evaluation metrics used in this dissertation.

As mentioned before, the evaluation metrics used in this work are distinguished in two dimensions: If they refer to quality or similarity and if they are objective or subjective. This twofold distinction allows for nicely grouping the evaluation metrics as shown in Table 2.2. The following sections discuss the details of these metrics.

2.5.1 Log-Spectral Distortion

This measure evaluates the similarity of two parallel spectral vector sequences. Since the objective of voice conversion is to convert the source speech in a way that it sounds similar to the target voice, the log-spectral distortion, a distance measure between the converted source and the target speech, is the smaller, the more successful the conversion is. In order to produce parallel vector sequences of converted source and target, parallel speech samples of the reference (target) speaker are required and are aligned to the converted source by means of dynamic time warping as for the text-dependent voice conversion training introduced in Section 2.1. According to [Hagen 94], the log-spectral distortion is defined as the distance between the cepstral coefficients of converted source $\tilde{x}_1^K$ and target y_1^K multiplied with a constant used for achieving compatibility to distance measures formerly introduced by [Gray & Markel 76]:

$$D_{\text{LSD}} = \frac{10\sqrt{2}}{K \ln 10} \sum_{k=1}^{K} |\tilde{x}_k - y_k|. \tag{2.17}$$

The log-spectral distortion is supposed to appropriately represent the subjective impression of spectral dissimilarity, although considerations about spectral details as e.g. required for residual prediction (see Section 4.3.1) show that, for certain investigations, this measure is not useful.

2.5.2 Extended ABX

An ABX test is a method of subjectively comparing two audio files to determine if any audible differences between the files can be found [Meilgaard & Civille$^+$ 99]. Its name is derived from the fact that there are two known samples, A and B, and one unknown sample, X, the ABX tester must identify as either sample A or sample B.

The ABX test was formerly used to show the high fidelity of the compact disc and other digital sound storage media as compared to analog ones and is very often applied to the evaluation of speech codecs. Already in one of the first voice-conversion-related publications [Abe & Nakamura$^+$ 88], this test was applied to assess the success of voice conversion. There, A and B were either source and target or vice versa (by random), and X was the converted source. The subjects were asked whether X sounds similar to A or to B. A large majority selecting the target was indicated as being successful, and, on the other hand, if most subjects selected the source, the voice conversion was said to be unsuccessful.

MOS	quality	similarity
5	excellent	identical
4	good	similar
3	fair	uncertain
2	poor	dissimilar
1	bad	different

Table 2.3: Mean opinion score rating scheme.

This paradigm was used over a decade until [Kain & Macon 98] observed that several of their experiments showed an almost uniformal distribution among source and target voice which could have been interpreted as being fairly successful. However, when interviewing the subjects they found "that many had the impression that a *third* [...] speaker was created" and, therefore, tended to decide by chance. In order to identify this category, this dissertation's author introduced a third choice to the test and called it extended ABX [Sündermann & Bonafonte$^+$ 04b]. Here, the subject is asked whether voice X sounded like

(1) voice A,

(2) voice B, or

(3) neither of them.

2.5.3 Mean Opinion Score

The mean opinion score (MOS) provides a numerical indication of perceived quality of multimedia such as audio, voice telephony, or video [ITU 96]. The MOS is expressed as a number between 1 and 5 with a rating scheme according to Table 2.3 and is the arithmetic mean of all individual scores. In speech science, the MOS is widely used for the evaluation of text-to-speech technology [Cole 98] and voice conversion [Moulines & Sagisaka 95].

A decade later, in the framework of the project TC-Star (cf. Section A.1), the author proposed to apply the MOS also to rate the similarity of converted and target voice replacing the extended ABX test discussed in Section 2.5.2 [Sündermann & Bonafonte$^+$ 05]. This was to produce a symmetric two-dimensional representation of speech quality and similarity in one diagram as shown in Figure 5.5 and to allow for deriving a single score by averaging both MOS scores in order to rank several systems. The respective rating scheme is given in Table 2.3.

3 VTLN-Based Voice Conversion

The considerable performance gain achieved when applying VTLN to speech recognition mentioned in Section 2.4.1 suggested that the proposed frequency warping changed an important part of the voice characteristics towards a given target (the norm voice). Since after warping, the spectra can be easily transformed back to an audible time signal by means of frequency domain PSOLA (see Section 2.3.4), its use for voice conversion seems promising.

3.1 Parameter Training

In Section 2.4.1, it was mentioned that the parameters of the considered warping functions are determined by performing a forced alignment on the whole training data for each parameter candidate and searching for the parameter resulting in the maximum score. Since in the case of voice conversion, we are given parallel speech of source and target (cf. Section 2.1), the optimal parameter candidate can be estimated by comparing converted source and target speech frame by frame, accumulating the distances between corresponding frames and choosing the candidate with the lowest result. In doing so, voiced frames are to be preferred, as, generally, they contain much more vocal-tract-related information than unvoiced frames[23]. Consequently the accumulation of the aforementioned distances is performed by weighting the respective frames' contributions by their voicing degree v_m as introduced in Section 2.3.4:

$$\xi_1^I = \arg\min_{{\xi'}_1^I} \sum_{m=1}^{M} v_m |Y_m - \tilde{X}_m({\xi'}_1^I)|^2 \tag{3.1}$$

where $|X|$ is the continuous equivalent to the definition of the l^2-norm according to Footnote 19:

$$|X| = \sqrt{\frac{1}{\pi}\int_0^\pi |X(\omega)|^2 \,\mathrm{d}\omega}. \tag{3.2}$$

Compared to the very time-consuming estimation of the warping factor in speech recognition, this procedure is rather fast, particularly when taking into account that the amount of speech considered in the voice conversion task rarely exceeds 15 minutes (cf. corpora descriptions in Section A.2), whereas in speech recognition, sometimes databases of several thousand hours of speech are employed [Evermann & Chan$^+$ 05]. At least this holds when the search space is kept as limited as mentioned earlier for applications to speech recognition in Equation 2.7:

$${\xi'}_1^I = \alpha' \in \{0.88, 0.9, \ldots, 1.12\}. \tag{3.3}$$

For two reasons, however, the search space has to be enlarged when being applied to voice conversion:

[23] [Ye & Young 04a] observed that "unvoiced sounds contain very little vocal tract information [...]. Hence, in common with other [voice conversion] systems, unvoiced frames [...] are simply copied to the target."

- According to the definition of VTLN (see Section 2.4.1), this technique aims at compensating for variations in the vocal tract length and normalizing it to a norm value which is located symmetrically between the extrema $\alpha = 0.88$ and $\alpha = 1.12$. When applied to voice conversion the objective is slightly different. Here, one aims at transforming the vocal tract length from a given voice to an arbitrary other voice. If the source voice is already an extreme voice (e.g. a child featuring a very short vocal tract), and the target voice is the other extremum (a bass voice with large vocal tract), the range of warping is about double of that expected in the normalization case of speech recognition[24]. Consequently, for the piece-wise linear warping function with two segments, the final range would be
$$\alpha' \in \{0.76, 0.78, \dots, 1.24\}. \tag{3.4}$$

- In Table 2.1, two warping functions depending on several parameters were introduced. They give more flexibility for changing the voice characteristics in the voice conversion framework on the one hand, but raise the problem of how to estimate these parameters on the other hand.

  For the case of combining k warping functions with one parameter each (see last row of Table 2.1) whose cardinality[25] is c, one obtains c^k different combinations of warping factors to be searched. This search is only tractable if k is small.

  In the case of the all-pass transform and given a considerable number of parameters, the estimation using a full search is not tractable anymore. [Sündermann & Ney$^+$ 03b] presented a parameter estimation algorithm based on the gradient descend method [Avriel 76]. Since this method delivers a local optimum in the neighborhood of a given initial parameter setting, a large number of Gaussian distributed initial parameter settings were applied in order to find the global optimum.

3.2 A Piece-Wise Linear Warping Function with Several Segments

The variety of mostly non-linear warping functions featuring different parameter ranges and characteristics as well as the computational effort to estimate the warping parameters calls for a generalized warping function with several parameters that can be computed fast and easily. For that purpose, the author defines a function composed of $I + 1$ linear segments fulfilling the following constraints common with the conventional warping functions of Table 2.1:

- The function is continuous.

- The function is monotonous.

- The function is bounded in domain as well as in co-domain to values between 0 and π also serving as starting and ending points.

[24] This suggests that the term vocal tract length *normalization* is not completely adequate to its application to voice conversion. Words like *adaptation*, *transformation*, or *conversion* seem to better represent the described technique. However, for the considerable popularity of the acronym VTLN in the speech community, the author decided to ignore this slight imprecision.

[25] The number of members of the set of parameters. E.g., the cardinality of the set of all possible α' in Equation 3.4 is 25.

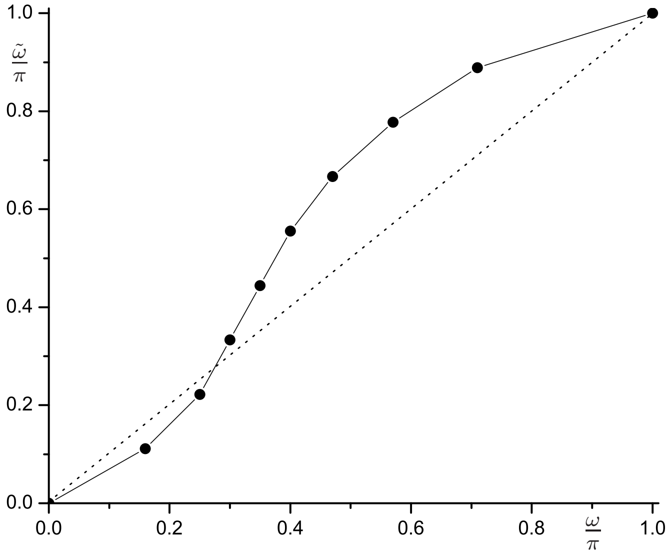

Figure 3.1: Example of a the piece-wise linear warping function with several segments.

Such a function is given by

$$g(\omega|\omega_1^I, \tilde{\omega}_1^I) = \alpha_i\omega + \beta_i \quad \text{for} \quad \omega_i \le \omega < \omega_{i+1};\ i = 0, \dots, I \tag{3.5}$$

$$\text{with} \quad \alpha_i = \frac{\tilde{\omega}_{i+1} - \tilde{\omega}_i}{\omega_{i+1} - \omega_i}, \quad \beta_i = \tilde{\omega}_{i+1} - \alpha_i\omega_{i+1} \quad \text{and}$$

$$0 = \omega_0 < \omega_1 < \dots < \omega_I < \omega_{I+1} = \pi, \quad \text{for } \tilde{\omega}_i \text{ equivalent.}$$

An example of this function is shown in Figure 3.1 where the target frequency parameters are equidistantly distributed as

$$\tilde{\omega}_i = \frac{i\pi}{I+1}. \tag{3.6}$$

For the following considerations, the parameters $\tilde{\omega}_1^I$ are kept constant, so the estimation procedure given by Equation 3.1 is limited to the unknown parameters ω_1^I:

$$\begin{aligned} \omega_1^I &= \arg\min_{\omega'^I_1} \sum_{m=1}^{M} v_m |Y_m - \tilde{X}_m(\omega'^I_1)|^2 \\ &= \arg\min_{\omega'^I_1} \sum_{m=1}^{M} v_m \int_0^\pi |Y_m(\omega) - \tilde{X}_m(\omega|\omega'^I_1)|^2 \mathrm{d}\omega. \end{aligned} \tag{3.7}$$

Inverting Equation 2.5 produces

$$\omega = g^{-1}(\tilde{\omega}) \tag{3.8}$$

that can be applied to Equation 2.6 yielding

$$\tilde{X}(\omega) = X(g^{-1}(\omega)). \tag{3.9}$$

Inverting $g(\omega)$ as given by Equation 3.5 results in

$$g^{-1}(\omega|\omega_1^I) = \frac{\omega - \beta_i}{\alpha_i} \quad \text{for} \quad \tilde{\omega}_i \le \omega < \tilde{\omega}_{i+1};\ i = 0, \dots, I \tag{3.10}$$

which is inserted into Equation 3.9:

$$\tilde{X}(\omega|\omega_1^I) = X\left(\frac{\omega-\beta_i}{\alpha_i}\right) \quad \text{for} \quad \tilde{\omega}_i \le \omega < \tilde{\omega}_{i+1};\ i = 0,\dots,I. \tag{3.11}$$

This allows for breaking the integral of Equation 3.7 down into a sum of partial integrals:

$$\omega_1^I = \arg\min_{\omega'^I_1} \sum_{m=1}^{M} v_m \sum_{i=0}^{I} \int\limits_{\tilde{\omega}_i}^{\tilde{\omega}_{i+1}} \left| Y_m(\omega) - X_m\Big(\frac{\omega-\beta_i}{\alpha_i}|\omega'^I_1\Big)\right|^2 \mathrm{d}\omega. \tag{3.12}$$

Exploiting the fact that α_i and β_i only depend on the unknown variables ω_i and ω_{i+1} yields

$$\begin{aligned} \omega_1^I &= \arg\min_{\omega'^I_1} \sum_{i=0}^{I}\sum_{m=1}^{M} v_m \int\limits_{\tilde{\omega}_i}^{\tilde{\omega}_{i+1}} \left| Y_m(\omega) - X_m\Big(\frac{\omega-\beta_i}{\alpha_i}|\omega'_i,\omega'_{i+1}\Big)\right|^2 \mathrm{d}\omega \\ &= \arg\min_{\omega'^I_1} \sum_{i=0}^{I} G(\omega'_i,\omega'_{i+1}). \end{aligned} \tag{3.13}$$

Now, the auxiliary quantity Q is introduced as

$$\begin{aligned} Q(i,\nu) &= \min_{\substack{\nu_1^i;\\ \nu_i=\nu}} \sum_{j=1}^{i} G(\nu_{j-1},\nu_j) \\ &= \min_{\nu_{i-1}^i} \left[G(\nu_{i-1},\nu_i) + \min_{\substack{\nu'^{i-1}_1;\\ \nu'_{i-1}=\nu_{i-1}}} \sum_{j=1}^{i-1} G(\nu_{j-1},\nu_j)\right] \\ &= \min_{\nu_{i-1}^i} \Big[G(\nu_{i-1},\nu_i) + Q(i-1,\nu_{i-1})\Big]. \end{aligned} \tag{3.14}$$

Inserting this recursion formula into Equation 3.13 and setting the initial quantity $Q(0,0) = 0$ produces

$$\omega_1^I = \arg Q(I+1,\pi). \tag{3.15}$$

This shows that the problem can be broken down into simpler subproblems that require a minimization of two rather than I parameters, i.e., it is a typical dynamic programming problem [Cormen & Leiserson$^+$ 90]. The complexity of a straightforward search over all parameter combinations of ω'^I_1 is $\mathbf{O}(c^I)$, whereas solving the subproblems corresponds to only $\mathbf{O}(Ic^2)$. For c cf. Footnote 25.

3.3 Frequency Domain vs. Time Domain VTLN

So far, this chapter's considerations dealt with parameter estimation of VTLN-based voice conversion that takes place in the training phase. Often, the training can be performed offline so that there are no strong restrictions with respect to real-time ability and memory

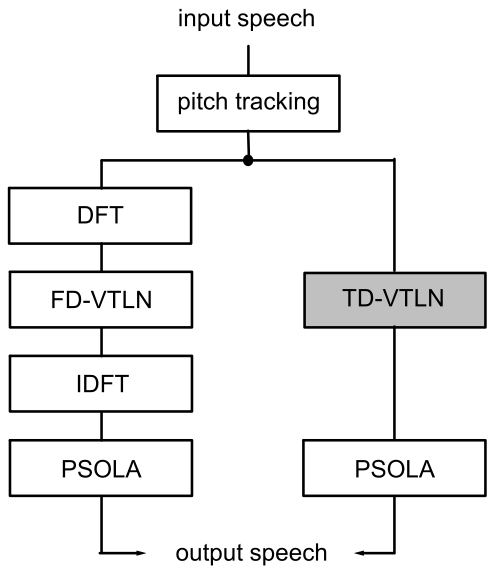

Figure 3.2: Frequency-domain- vs. time-domain-VTLN-based voice conversion.

requirements[26]. The conversion, however, should be real-time-able as crucial for several applications introduced in Chapter 1 (**E**, **F**, **G**). Even more demanding with respect to the real-time factor is the application to embedded devices (**C**) where the computational effort has to be reduced as much as possible.

Having a look at Figure 3.2, on the left hand side, we see the signal flow according to conventional VTLN-based voice conversion where, after breaking the input speech signal down into frames according to Section 2.3.4, phase and magnitude spectra are extracted applying discrete Fourier transformation. Then, spectra are warped as explained in Section 2.4.1, transformed back to time domain via inverse discrete Fourier transformation, and, finally, concatenated by means of PSOLA. Table 3.1 shows the computational time behaviour of this algorithm. Taking an example mean fundamental frequency of 160Hz and a sampling rate of 16kHz, we get $\bar{K} = 200$ and, finally, find that more than 97% of the computation is due to the Fourier transformation[27].

This outcome suggests to search for a VTLN technique that avoids the use of discrete Fourier transformation and directly manipulates the signal in time domain. The principle of this idea is to exploit the time correspondences of the discrete Fourier transformation that explain what the warping in frequency domain causes in time domain. Again, due to the advantageous characteristics of the piece-wise linear warping function with several segments introduced in Section 3.2, it will serve as prime example of the following considerations.

[26] There are exceptions like the Minnesang project mentioned in Section 2.2 involving nearly real-time cross-language training based on a short verse uttered by the visitor in his mother tongue. The voice conversion technology used in this project is based on developments of the author described in [Sündermann & Ney$^+$ 03b, Sündermann & Bonafonte$^+$ 04a] as well as in this dissertation.

[27] This result does not consider the computational contribution of the pitch tracker. In several applications, the pitch marks are computed offline and, hence, do not contribute to real-time ability.

	running time / operations		memory / 16bit	
	FD-VTLN	TD-VTLN	FD-VTLN	TD-VTLN
DFT	$4\bar{K}^2 - 2\bar{K}$		$\bar{K}$	
spline interp.	$40\bar{K}$	$40\bar{K}$	$6\bar{K}$	$6\bar{K}$
IDFT	$4\bar{K}^2 - 2\bar{K}$		$\bar{K}$	
total	$8\bar{K}^2 + 36\bar{K}$	$40\bar{K}$	$8\bar{K}$	$6\bar{K}$

Table 3.1: Frequency domain vs. time domain VTLN: running time and memory requirements; $\bar{K} = 2\frac{f_\mathrm{s}}{\bar{f}_0}$ is the average number of samples per frame (remember that a frame consists of two signal periods); for f_s and $\bar{f}_0$, see Sections 2.4.2 and 2.3.4. Operations include floating point additions and multiplications performed per frame. To explicitly estimate the number of operations per second, multiply the number of operations given in the table by the average number of frames per second, e.g.: $\bar{f}_0 \cdot (8\bar{K}^2 + 36\bar{K}) = 32\frac{f_\mathrm{s}^2}{\bar{f}_0} + 72f_s$.

3.3.1 Frequency Domain VTLN

Expressing the time frame $x(t);\ 0 \le t \le T$ introduced in Section 2.3.4 as time series $x_1^K;\ K = Tf_\mathrm{s}$, the discrete Fourier transformation is defined as [Walker 96]

$$X_n = \sum_{k=1}^{K} x_k e^{-2\pi\imath(n-1)\frac{k-1}{K}};\ n = 1,\ldots,K \tag{3.16}$$

or, shorter,

$$X = \mathcal{F}(x). \tag{3.17}$$

In the following, the abbreviation

$$\kappa = \pi\frac{k-1}{K} \tag{3.18}$$

is used. The coefficients X_2^K feature the symmetry

$$\begin{aligned} X_{K-n+2} &= \sum_{k=1}^{K} x_k e^{-2\kappa\imath(K-n+1)} \\ &= \sum_{k=1}^{K} x_k e^{2\kappa\imath(n-1)} \\ &= \sum_{k=1}^{K} \left(x_k e^{-2\kappa\imath(n-1)}\right)^* \\ &= X_n^* \quad \text{for} \quad n = 2,\ldots,K \end{aligned} \tag{3.19}$$

and, for the sake of completeness, it is defined

$$X_{K+1} = X_1^*. \tag{3.20}$$

This means, about half of the coefficients are redundant, whereas the other half corresponds to the spectrum $X(\omega)$, treated as being continuous so far, cf. discussion in Footnote 13. Let N be the number of non-redundant coefficients $X_1, \dots, X_N$ with

$$N = \begin{cases} \frac{K+1}{2}: & \text{odd } K \\ \frac{K+2}{2}: & \text{even } K, \end{cases} \tag{3.21}$$

then, in the following, a discretized variant of $X(\omega)$ is considered, given by

$$X_n := X(\omega_n) \text{ with } \omega_n = \pi \frac{n-1}{N-1};\ n = 1, \dots, N. \tag{3.22}$$

Now, Equation 3.11 yields $\tilde{X}$ which can be computed from discrete values for ω by means of interpolation techniques as cubic splines, cf. Section 2.4.1. From such a modified spectrum $\tilde{X}$, one can always return to the symmetric form $\tilde{X}_n,\ n = 1, \dots, K$ as introduced in Equation 3.16 using the equivalence derived in Equation 3.19. The only missing parameter at this point could be K, for instance, when pitch adaptation for (frequency domain) PSOLA is performed together with the frequency warping affecting the number of redundance-free spectral lines N and, hence, K. Equation 3.21 contains the relation between N and K, but it is unclear whether K is odd or even if only N is given. However, looking at the symmetry formula in Equation 3.19 for the case $n = \frac{K}{2} + 1$ that is only integer and, hence, valid, if K is even, one obtains

$$X_{K-\frac{K}{2}-1+2} = X_{\frac{K}{2}+1} \quad = \quad X^*_{\frac{K}{2}+1} \quad \longrightarrow \quad \Im(X_{\frac{K}{2}+1}) = \Im(X_N) = 0 \ \text{ for } \ K \in \{2, 4, \dots\}. \tag{3.23}$$

This outcome allows for inverting Equation 3.21[28]:

$$K = \begin{cases} 2N-1: & \Im(X_N) \neq 0 \\ 2N-2: & \Im(X_N) = 0. \end{cases} \tag{3.24}$$

Finally, inverse discrete Fourier transformation leads back to time domain and, hence, to the converted time frame

$$\tilde{x}_k = \frac{1}{K} \sum_{n=1}^{K} \tilde{X}_n e^{2\kappa\imath(n-1)};\ k = 1, \dots, K \tag{3.25}$$

or, simply,

$$\tilde{x} = \mathcal{F}^{-1}(\tilde{X}). \tag{3.26}$$

3.3.2 Time Domain VTLN

As motivated earlier in this section, the objective of time domain VTLN is to avoid the use of discrete Fourier transformation (Equations 3.16 and 3.25) by expressing the frequency warping of Equation 3.11 by means of direct modifications of the time signal x_1^K. Equation 3.11

[28] Theoretically, the condition $\Im(X_N) = 0$ is not sufficient for determining that K is even, since one can construct cases of odd spectra with $\Im(X_N) = \Im(X_{N+1}) = 0$; but in natural speech data, such cases are extremely rare.

can be rewritten as a sum of linearly warped spectra windowed by rectangular windows:

$$\begin{aligned}\tilde{X}(\omega|\omega_1^I,\tilde{\omega}_1^I) &= \sum_{i=0}^{I} X\left(\frac{\omega-\beta_i}{\alpha_i}\right) R(\omega|\omega_i,\omega_{i+1}). \\ &=: \sum_{i=0}^{I} X^{(i)}(\omega)R^{(i)}(\omega) \qquad (3.27)\end{aligned}$$

$$\text{with} \quad R(\omega|\omega_i,\omega_{i+1}) = \begin{cases} 1: & \omega_i < \omega < \omega_{i+1} \\ \frac{1}{2}: & \omega = \omega_i \vee \omega = \omega_{i+1} \\ 0: & \text{otherwise.} \end{cases} \qquad (3.28)$$

Interpreting X and R as discrete spectra, the time waveform $\tilde{x}$ is derived as[29]

$$\begin{aligned}\tilde{x} &= \mathcal{F}^{-1}\Big(\sum_{i=0}^{I} X^{(i)}R^{(i)}\Big) \\ &= \sum_{i=0}^{I} \mathcal{F}^{-1}(X^{(i)}R^{(i)}) \qquad (3.29)\end{aligned}$$

exploiting the linearity of the discrete Fourier transformation. Furthermore, by means of the convolution theorem [Smith 03], the multiplication of X and R in frequency domain can be expressed as a convolution in time domain where x and r refer to the respective inverse transforms:

$$\tilde{x} = \sum_{i=0}^{I} x^{(i)} * r^{(i)} \qquad (3.30)$$

with the following definition of the convolution operator:

$$(a*b)_k = \sum_{n=1}^{k} a_n b_{k-n+1} + \sum_{n=k+1}^{K} a_n b_{k-n+1+K} \quad \text{for} \quad k = 1,\ldots,K. \qquad (3.31)$$

The subsequent considerations apply some foundations of the (continuous) Fourier transformation, that is why, in the following paragraph, $x^{(i)}$ is treated as a function of the continuous time t:

$$x^{(i)}(t) = \mathcal{F}^{-1}\left\{X\left(\frac{\omega-\beta_i}{\alpha_i}\right)\right\}(t). \qquad (3.32)$$

Section A.3.1 provides a derivation of the equality

$$x^{(i)}(t) = \alpha_i e^{\imath\beta_i t} x(\alpha_i t). \qquad (3.33)$$

There are some interesting aspects about this result:

- A scaling in time domain is uncomplicated as long as the continuous Fourier transformation is considered that is defined for $-\infty < t < \infty$. However, when returning to the discrete case where the source frame consisted of K samples a time scaling by α_i means that

[29] Here, XR refers to the element-wise multiplication of the spectra X and R.

- the time waveform consists of the sample values $x_{\alpha_i k}$ with $\alpha_i k$ being integer and, hence, has to be derived by interpolation.
- For the same reason, the number of time samples K has to be changed to the integer next to $\frac{K}{\alpha_i}$, in the following referred to as $K^{(i)}$, to make sure that no part of the waveform is missing[30]:

$$K^{(i)} = \left\lfloor \frac{K}{\alpha_i} \right\rfloor . \tag{3.34}$$

- Since for each addend in Equation 3.29, the value $K^{(i)}$ can be different due to different values of α_i, it is not clear how the summation is to be performed properly. This problem is overcome by waiting with the summation until the final waveform concatenation in the time domain PSOLA framework where each of the I contributions is treated separately yielding a continuous waveform featuring the target pitch. Ultimately, these contributions are summed up resulting in the final speech signal.

- The term $e^{\imath\beta_i t}$ is complex, and therefore $x^{(i)}$ becomes complex. This is due to the symmetry characteristics of the discrete Fourier transformation discussed in Section 3.3.1 which, so far, have not been taken into account. Later in this section when joining the contributions of X and R this issue will be addressed in detail.

- These considerations lead to the discrete representation

$$x_k^{(i*)} = \alpha_i e^{2\kappa^{(i)}\imath\beta_i} x_{\alpha_i k} \quad \text{for} \quad k = 1, \ldots, K^{(i)}. \tag{3.35}$$

The superscript $(*)$ is to denote that this time series is complex-valued. To become real-valued, it first will be convolved with the time correspondence $r^{(i*)}$ of that part of the rectangular window R representing the spectral segment where $\omega_i \le \omega \le \omega_{i+1}$ (in discrete terms expressed by $n_i \le n \le n_{i+1}$, cf. below). Then, it is summed up with its symmetric counterpart $x^{(i*-)}$ convolved with R's analogon $r^{(i*-)}$. Such a counterpart x^- to a given (complex-valued) waveform x is calculated by applying the inverse discrete Fourier transformation (Equation 3.25) to the spectrum $X^*_{K-n+2},\ n = 1, \ldots, K$ that is motivated by the symmetry assumption derived in Equation 3.19:

$$x_k^- = \frac{1}{K} \sum_{n=1}^{K} X^*_{K-n+2}\ e^{2\kappa\imath(n-1)}. \tag{3.36}$$

Substituting $n = K - n' + 2$ and using Equation 3.20 yields

$$\begin{aligned} x_k^- &= \frac{1}{K} \sum_{n'=1}^{K} X^*_{n'}\ e^{2\kappa\imath(K-n'+1)} \\ &= \frac{1}{K} \sum_{n'=1}^{K} X^*_{n'} \left(e^{2\kappa\imath(n'-1)}\right)^* . \end{aligned} \tag{3.37}$$

By means of the relation

$$a^* b^* = (ab)^*, \tag{3.38}$$

[30] The symbol $\lfloor a \rfloor$ refers to the arithmetic rounding of a.

one ultimately obtains

$$\begin{aligned} x_k^- &= \frac{1}{K}\sum_{n'=1}^{K}\left(X_{n'}\ e^{2\kappa\imath(n'-1)}\right)^* \\ &= x_k^*. \end{aligned} \tag{3.39}$$

Section A.3.2 provides the contribution of $R^{(i)}$ to Equation 3.29:

$$\begin{aligned} r_k^{(i*)} &= \frac{1}{K^{(i)}}\left[\frac{\sin\left(\kappa^{(i)}(n_{i+1}-1)\right)}{\sin(\kappa^{(i)})}e^{\kappa^{(i)}\imath(n_{i+1}-2)} - \frac{\sin(\kappa^{(i)}n_i)}{\sin(\kappa^{(i)})}e^{\kappa^{(i)}\imath(n_i-1)}\right. \\ &\left. +\frac{1}{2}e^{2\kappa^{(i)}\imath(n_i-1)} + \frac{1}{2}e^{2\kappa^{(i)}\imath(n_{i+1}-1)}\right] \end{aligned} \tag{3.40}$$

Having derived the complex-valued time correspondences $x^{(i*)}$ and $r^{(i*)}$, the convolution according to Equation 3.30 is to be carried out, likewise, the symmetric counterparts $x^{(i*-)}$ and $r^{(i*-)}$ derived using Equation 3.39 are convolved, and both contributions are summed up:

$$\begin{aligned} \tilde{x} &= \sum_{i=0}^{I}\left[x^{(i*)} * r^{(i*)} + x^{(i*-)} * r^{(i*-)}\right] \\ &= \sum_{i=0}^{I}\left[x^{(i*)} * r^{(i*)} + \left(x^{(i*)}\right)^* * \left(r^{(i*)}\right)^*\right]. \end{aligned} \tag{3.41}$$

Referring to Equation 3.31, the convolution of two conjugated numbers a^* and b^* can be computed (also using Equation 3.38) by

$$\begin{aligned} (a^* * b^*)_k &= \sum_{n=1}^{k} a_n^* b_{k-n+1}^* + \sum_{n=k+1}^{K} a_n^* b_{k-n+1+K}^* \\ &= \sum_{n=1}^{k}(a_n b_{k-n+1})^* + \sum_{n=k+1}^{K}(a_n b_{k-n+1+K})^* \\ &= (a*b)_k^* \end{aligned} \tag{3.42}$$

finally yielding

$$\begin{aligned} \tilde{x} &= \sum_{i=0}^{I}\left[x^{(i*)} * r^{(i*)} + \left(x^{(i*)} * r^{(i*)}\right)^*\right] \\ &= 2\sum_{i=0}^{I}\Re\left(x^{(i*)} * r^{(i*)}\right). \end{aligned} \tag{3.43}$$

3.3.3 Time Behaviour

In the beginning of Section 3.3, the derivation of time domain VTLN was motivated by its supposedly superior computation time behaviour compared with the conventional frequency

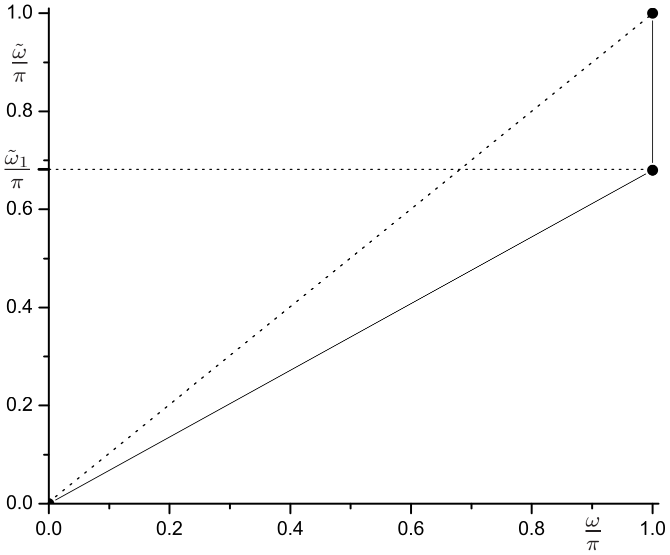

Figure 3.3: Special case of a the piece-wise linear warping function with two segments and $\omega_1 \to \pi$.

domain VTLN. The latter consumed most of the time for the domain transformation from time to frequency and back to time – discrete Fourier transformation has a computational complexity of $\mathbf{O}(K^2)$.

However, the final formula of the time domain VTLN (Equation 3.43) involves a convolution that, likewise, is of complexity $\mathbf{O}(K^2)$. Hence, it seems that the avoidance of the domain transformation did not have a positive impact on the time behaviour of VTLN-based voice conversion.

To refute this argument, a special case of the piece-wise linear warping function is to be taken into account where two segments are considered, i.e., $I = 1$. Equation 3.5 yields

$$\alpha_0 = \frac{\tilde{\omega}_1}{\omega_1}; \qquad \alpha_1 = \frac{\pi - \tilde{\omega}_1}{\pi - \omega_1}; \qquad \beta_0 = 0; \qquad \beta_1 = \pi \frac{\tilde{\omega}_1 - \omega_1}{\pi - \omega_1}. \tag{3.44}$$

Now, it is considered that ω_1 approaches π, as shown in Figure 3.3. Consequently, α_1 approaches infinity, and Equation 3.34 produces $K^{(1)} = 0$. This means, segment $i = 1$ is omitted, and the only remaining free parameter is $\tilde{\omega}_1$. Since $\beta_0 = 0$, Equation 3.35 yields

$$x_k^{(0*)} = \frac{\tilde{\omega}_1}{\pi} x_{\frac{\tilde{\omega}_1}{\pi} k}. \tag{3.45}$$

This function is real-valued and, considering the omission of the segment $i = 1$, allows for rewriting Equation 3.43:

$$\tilde{x} = 2x^{(0*)} * \Re\left(r^{(0*)}\right). \tag{3.46}$$

As shown in Section A.3.3, the above real part can be expressed by

$$\Re\bigl(r_k^{(0*)}\bigr) = \begin{cases} \frac{1}{2}: & k = 1 \\ 0: & k > 1. \end{cases} \tag{3.47}$$

Now, the convolution operator's definition given by Equation 3.31 is inserted into Equation 3.46 yielding the final result

$$\begin{aligned}\tilde{x}_k &= 2\left[\sum_{n=1}^{k} x_n^{(0*)}\Re(r_{k-n+1}^{(0*)}) + \sum_{n=k+1}^{K} x_n^{(0*)}\Re(r_{k-n+1+K}^{(0*)})\right]\\ &= 2x_k^{(0*)}\cdot\frac{1}{2} \;=\; \frac{\tilde{\omega}_1}{\pi}x_{\frac{\tilde{\omega}_1}{\pi}k}.\end{aligned} \qquad (3.48)$$

This means, for the special case considered in the above derivations, the convolution is omitted, and the computational complexity is reduced to $\mathbf{O}(K)$. An exact breakdown of the number of operations for this special case compared with frequency domain VTLN is given in Table 3.1. Taking into account the example at the very beginning of Section 3.3, i.e., $\bar{K} = 200$, the proposed time domain VTLN accelerates computation by a factor of about 40. In absolute terms, this corresponds to 1.28Mops/s.

3.3.4 Experiments

In the last section, it was shown that, for a special case, time domain VTLN significantly accelerates the computation. Now, it was important to investigate if this paradigm change would perceptually affect the speech output. Therefore, both frequency- and time-domain-VTLN-based voice conversion were compared in terms of speech quality and performance to change the voice identity. This was done using the MOS and the extended ABX test described in Sections 2.5.2 and 2.5.3. For these subjective tests, 14 subjects, 12 of whom specialists in speech processing, were invited.

From the Spanish synthesis corpus described in Section A.2.1, ten parallel utterances (about 40s) were randomly selected and aligned by means of dynamic time warping. For both gender combinations, the parameter training according to Section 3.1 was performed. Then, frequency and time domain VTLN were applied to a random selection of 8 parallel utterances of both genders (about 30s) which were different from those in the training data obtaining a total of 32 converted utterances. From these, 8 sentences were randomly selected in a way that each gender/VTLN-type combination was represented by exactly two sentences. This randomization was carried out again for each of the 14 participants. An overview about the training and testing data as well as the compared systems' properties is given in Table 3.2. Table 3.3 reports the results of the MOS test and Table 3.4 those of the extended ABX test depending on the VTLN technique and the gender combination.

3.3.5 Conclusion

The outcomes of the MOS test on speech quality in Table 3.3 certify a fair speech quality for both frequency- as well as time-domain-VTLN-based voice conversion. The conversion from the female to the male voice performs significantly better than the other way around. An analysis of the distortions and artifacts produced by the conversion algorithm showed that most of them could be attributed to errors of the pitch tracker that achieved a higher performance for female than for male speech confirming results of [Höge & Kotnik$^+$ 06].

This outcome, however, does not allow for the general conclusion that male-to-female conversion is the harder task, since comparing with experiments on databases involving other voices suggests that the conversion has a stronger dependence on the particular processed

	frequency domain VTLN	**time domain VTLN**
conversion type	text-dependent	
	intra-lingual	
source/target language	Spanish/Spanish	
alignment technique	dynamic time warping	
corpus	Spanish synthesis corpus	
voices	1 female, 1 male	
amount of training data	38.7s female, 42.7s male	
amount of test data	30.6s female, 34.0s male	
number of test subjects	14	
pitch mark extraction	Goncharoff/Gries (automatic)	
voicing determination	Adaptive Multi-Rate (automatic)	
warping function	piece-wise linear, two segments	piece-wise linear, two segments
	(symmetric, cf. Table 2.1)	(cf. Figure 3.3)
free parameters	α, ρ^*	$\tilde{\omega}_1{}^\dagger$, ρ
acoustic synthesis	frequency domain PSOLA	time domain PSOLA

Table 3.2: Frequency domain vs. time domain VTLN: system properties.

* For ρ, see Equation 2.4.

† According to Equation 3.44, the parameter $\tilde{\omega}_1$ can be normalized by $\omega_1 = \pi$ resulting in the warping factor α_0 making frequency and time domain VTLN equivalent also in terms of the nature of their free parameters.

	FD VTLN	TD VTLN
female to male	3.3	3.4
male to female	2.6	2.6
total	3.0	3.0

Table 3.3: Frequency domain (FD) vs. time domain (TD) VTLN: results of the MOS test.

[%]	FD VTLN	TD VTLN
source speaker	20	16
target speaker	29	36
neither	**50**	**48**

Table 3.4: Frequency domain vs. time domain VTLN: results of the extended ABX test.

voices than on the gender. This is confirmed in literature where one finds examples claiming that “the conversion from female to male is much better than from male to female” [Qin & Chen$^+$ 05] but also vice versa – “male to female conversion achieves better quality than female to male conversion” [Tang & Wang$^+$ 01].

Considering the outcomes of the extended ABX test shows that VTLN-based voice conversion features the capability to manipulate a given voice in such a way that the result is sufficiently different from the original to be perceived as another voice (task **1**): $\leq 20\%$ of the example sentences were recognized as being spoken by the source speaker, see Table 3.4.

On the other hand, VTLN-based voice conversion is not appropriate to imitate a certain speaker's voice (task **2**): The results report that only around a third of the examples were perceived to be uttered by the target speaker.

Finally, the experimental results show that the two compared conversion techniques based on frequency and time domain VTLN perform very similarly in terms of both speech quality and voice identity. The time-domain-VTLN-based technique, however, is considerably faster (in the example discussed above by a factor of about 40).

3.4 On Generating Several Well-Distinguishable Voices

The results of the last section showed that VTLN-based voice conversion is able to change a source speaker's voice such that it is perceived to be another voice. However, would it be possible to produce *two* voices that are perceived to be different from the source voice and also from each other, or even more? This capability could play a role for scenarios like in speech-to-speech translation for parliamentary speeches as in the project TC-Star (cf. Section A.1) where several politicians have to be clearly distinguished.

Consequently, this section is to investigate the characteristics of VTLN-based voice conversion with respect to the production of well-distinguishable voices. In doing so, the special case discussed in the last section is to be used, i.e., the number of parameters is limited to two, the warping factor α and the ratio ρ between the mean fundamental frequencies after and before the conversion. To simplify matters, they are combined in the vector $v = (\alpha, \rho)$.

3.4.1 On the Naturalness of Converted Voices

It is obvious that only parameter values inside a certain range result in reasonable, i.e. naturally sounding voices. E.g., setting $v = (1, 1)$ does not change the voice at all and should result in the maximum naturalness when distortions by the analysis-synthesis system can be neglected. On the other hand, extreme values as $r \to \infty$ produce artificial or even irrecognizable voices.

Hence, at the beginning, the relation between the parameter settings and the voices' naturalness was to be investigated. This was done by performing a subjective test on the naturalness of the converted speech. The corpus used in the following experiments is the German synthesis corpus described in Section A.2.2. From both the female and the male speaker, 25 utterances were randomly selected and converted based on the parameters displayed in Figure 3.4 which were determined by means of informal listening tests.

Now, 11 subjects were asked to rate the naturalness of converted speech samples on an MOS scale between 1 (artificial) and 5 (natural) based on the following instructions:

> "You listen to 50 speech samples. In this experiment, you are to rank the naturalness of voices. Please, do not pay attention to the recording conditions, the sound or synthesis quality of the samples, or to the speaking style; we are just interested in whether you think the voices sound as if uttered by a real person."

Table 3.5 shows the system properties of this experiment. From the resulting naturalness scores belonging to the 25 parameter vectors of the female and the male speaker, respectively, an estimate for the score in the whole v space was produced by applying a two-dimensional interpolation based on Delaunay triangulation [Preparata & Shamos 85] to the scattered data. Figure 3.5 shows the results.

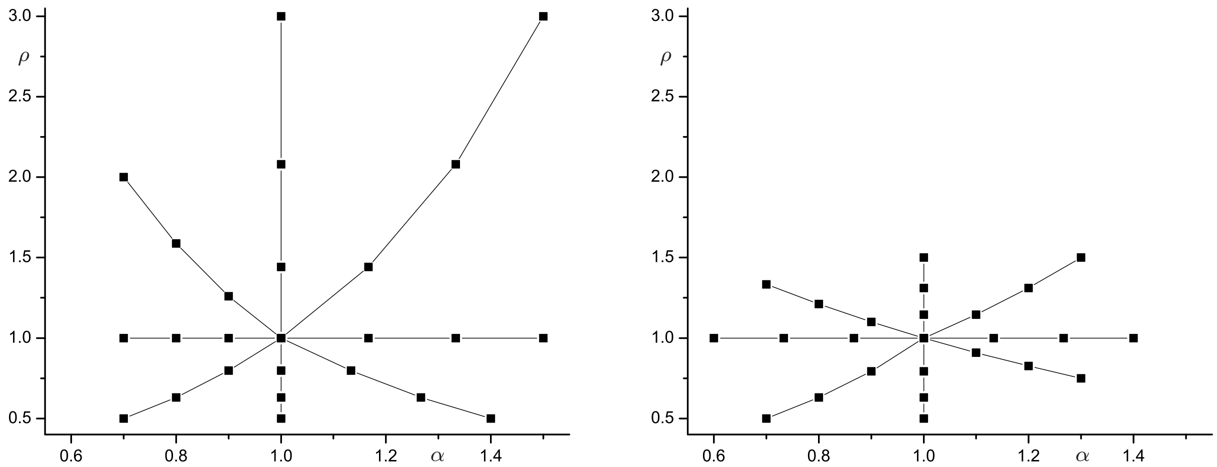

Figure 3.4: Naturalness of converted voices: parameter settings for the female (left) and the male (right) voice.

conversion technique	time domain VTLN
corpus	German synthesis corpus
voices	1 female, 1 male
amount of test data	98.6s female, 55.8s male*
number of test subjects	11 [10, 13]
pitch mark extraction	Goncharoff/Gries (automatic)
warping function	piece-wise linear, two segments
free parameters	α, ρ
acoustic synthesis	time domain PSOLA

Table 3.5: Naturalness [dissimilarity, well-distinguishability] of converted voices: system properties.

* The time difference is due to considerable differences in the speaking rate of both speakers.

In both examples, the female as well as the male voice, the naturalness as a function of α and ρ resembles a somewhat elliptical and tilted shape. The tilt is due to the fact that vocal tract length (represented by the warping factor α) and fundamental frequency (represented by ρ) are correlated. Speakers with low voices mostly feature long vocal tracts and high voices short vocal tracts. Low voices with short vocal tracts, or the opposite, are much rarer and, hence, are rated less natural.

3.4.2 On the Dissimilarity of Converted Voices

The demand for well-distinguishable voices leads to the question, how the subjective dissimilarity of two voices produced by warping a source voice depends on the objective difference between the two parameter vectors used, $v_1 = (\alpha_1, \rho_1)$ and $v_2 = (\alpha_2, \rho_2)$. To describe the

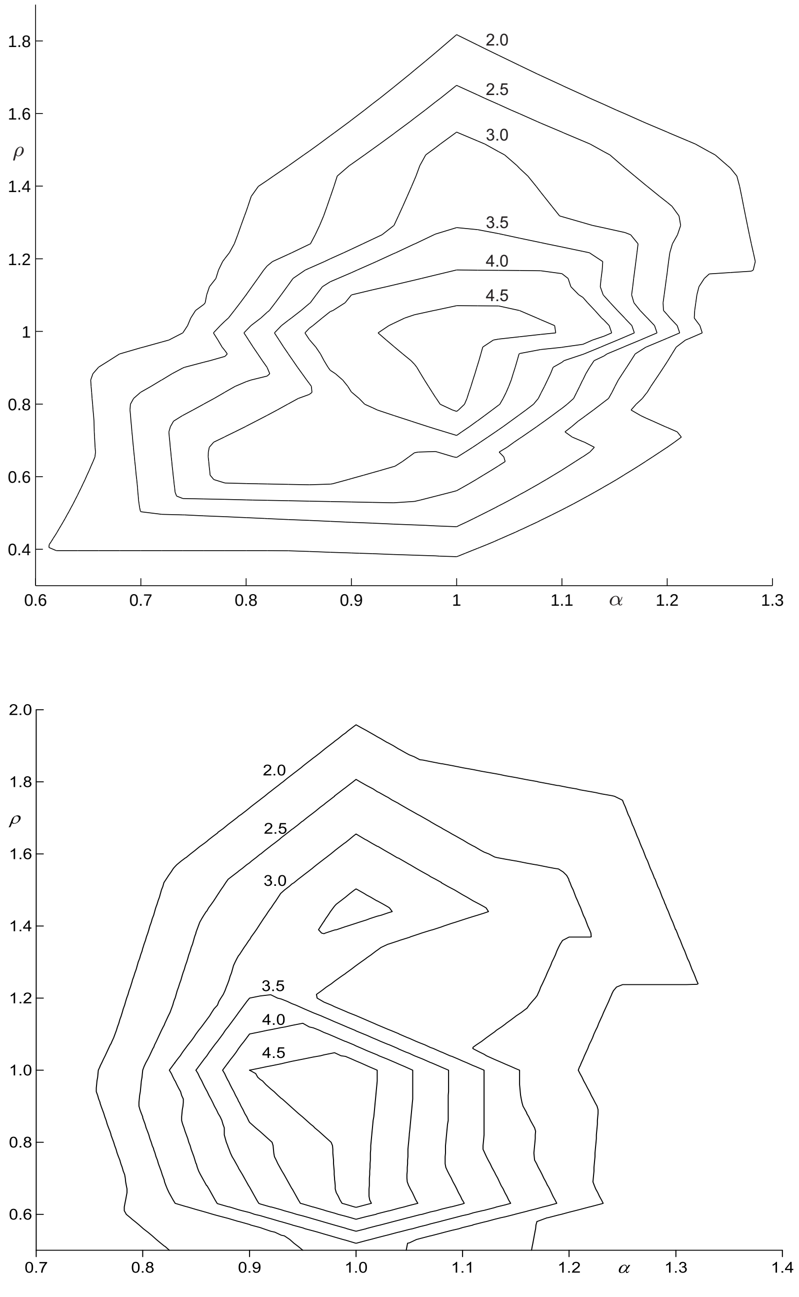

Figure 3.5: Dependence of the naturalness of a converted voice on the parameters α and ρ: Examples of a female (top) and a male (bottom) voice. Lines describe equal naturalness for a given MOS value in the range of $2, 2.5, \ldots, 4.5$.

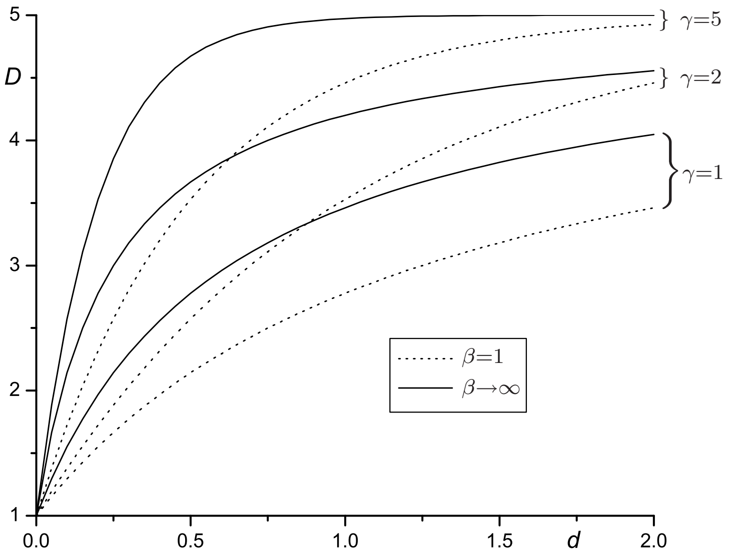

Figure 3.6: The function $D(d)$ as given by Equation 3.50 for some example parameter settings.

dependence between both measures, the following model is introduced:

- When the Euclidean distance is used as measure in parameter space one has to make sure that the contributions of the involved parameters are equivalent. E.g., it is not clear that a listener feels the same dissimilarity when facing voices generated using $v_1 = (0.9, 1)$ and $v_2 = (1.1, 1)$ as when the parameters were $v_1 = (1, 0.9)$ and $v_2 = (1, 1.1)$, although, in both cases, the Euclidean distance is $|v_2 - v_1| = 0.2$. Therefore, the involved parameters are scaled using the weights w and $(1 - w)$, $0 \leq w \leq 1$, respectively. In doing so, it is to be taken into account that the logarithmic frequency scale better represents the human perception than the linear one. I.e., by logarithmizing the fundamental frequency ratio, the parameter vectors $v_2 = (1, 0.5)$ and $v_2 = (1, 2)$ result in the same distance from the vector $v_1 = (1, 1)$ (one octave).

- In the following, the subjective distance D that is an MOS, rating the dissimilarity of two voices on a scale between 1 (identical) and 5 (different)[31], and the objective distance

$$d = d(w, v_1, v_2) = \sqrt{\left[w(\alpha_2 - \alpha_1)\right]^2 + \left[(1 - w)\log\left(\frac{\rho_2}{\rho_1}\right)\right]^2} \qquad (3.49)$$

  resulting from the above considerations are to be used. If the parameter vectors v_1 and v_2 are identical, i.e., $d = 0$, the expected dissimilarity is minimum ($D = 1$). If the distance between the vectors approaches infinity, the voices are expected to be totally different ($D = 5$). A relation between d and D that fulfills these boundary conditions is the following:

[31] The objective here is to generate *dissimilar* voices as opposed to task **2** of Chapter 1 where target and converted voices are to resemble each other. That is why the rating scheme used here is inverted compared to that of Table 2.3 assuring that the optimum is always 5.0.

	female	male
γ	4.7	3.2
w	0.5	0.7
$\varepsilon(w,\infty,\gamma)$	0.3	0.4

Table 3.6: Parameters of the dissimilarity model.

$$D(d) = 5 \cdot \left[1 - \left(\frac{\gamma}{\beta} d + \sqrt[\beta]{\frac{5}{4}} \right)^{-\beta} \right]; \quad \gamma, \beta > 0. \tag{3.50}$$

Using the maximum and minimum parameter values for α and ρ given in Figure 3.4 and inserting these values into Equation 3.49 yields $d < 2$ for the current experimental setup. Figure 3.6 displays the function $D(d)$ for $0 \leq d \leq 2$ for some example parameter settings.

In order to determine the unknown model parameters w, β, and γ, another subjective test was performed; for the system settings, see Table 3.5. This time, 10 subjects were asked to rate the dissimilarity of $J = 24$ pairs of voices derived from the aforementioned corpus. The compared voices were the same as generated for the naturalness experiment in Section 3.4.1. A pair always consisted of voices derived from opposite parameter vectors with respect to the neutral element $v = (1, 1)$, see Figure 3.4. Averaging over the participants, this resulted in a score D_j for each parameter vector pair v_{1j}, v_{2j}, 12 for each gender. Applying Equations 3.49 and 3.50, the model parameters are estimated by minimizing the root mean square between the model $D(d(v_{1j}, v_{2j}))$ and the subjective scores D_j:

$$(w, \beta, \gamma) = \arg\min_{w', \beta', \gamma'} \varepsilon(w', \beta', \gamma') \quad \text{with} \tag{3.51}$$

$$\varepsilon(w, \beta, \gamma) = \sqrt{\frac{1}{J} \sum_{j=1}^{J} \Big[D\big(d(w, v_{1j}, v_{2j}), \beta, \gamma\big) - D_j \Big]^2} \, . \tag{3.52}$$

It turns out that for the considered scores, β becomes sufficiently large to approximate Equation 3.50 by its limit for $\beta \to \infty$:

$$\begin{aligned} \lim_{\beta\to\infty} D_\beta(d) &= 5 \cdot \left[1 - \lim_{\beta\to\infty} \left(\frac{\gamma}{\beta} d + \sqrt[\beta]{\frac{5}{4}} \right)^{-\beta} \right] \\ &= 5 \cdot \left[1 - \exp\left(\lim_{\beta\to\infty} \left[\ln \left(\frac{\gamma}{\beta} d + \sqrt[\beta]{\frac{5}{4}} \right)^{-\beta} \right] \right) \right] . \end{aligned} \tag{3.53}$$

Substituting $\beta = \frac{1}{b}$ and applying l'Hôpital's rule yields

$$\begin{aligned}
\lim_{\beta\to\infty} D_\beta(d) &= 5 \cdot \left[1 - \exp\left(\lim_{b\to 0} \frac{\ln\left[\gamma d b + \left(\frac{5}{4}\right)^b\right]}{-b}\right)\right] \\
&= 5 \cdot \left[1 - \exp\left(\lim_{b\to 0} \frac{\gamma d + \left(\frac{5}{4}\right)^b \ln\frac{5}{4}}{-\gamma d b - \left(\frac{5}{4}\right)^b}\right)\right] \\
&= 5 \cdot \left[1 - \exp\left(-\gamma d - \ln\tfrac{5}{4}\right)\right] \\
&= 5 - 4e^{-\gamma d}; \quad \gamma > 0. \qquad (3.54)
\end{aligned}$$

In Table 3.6, for both genders, the determined parameters are displayed. In order to assess the performance of the model, we also include the (absolute) model error $\varepsilon(w, \infty, \gamma)$, i.e. the root mean square obtained by using the optimized parameters γ and w, see Equations 3.51 and 3.52.

3.4.3 Generating Well-Distinguishable Voices

As argued in Section 3.4.1, the objective of the current investigations is to generate a certain number of well-distinguishable voices, while taking into account that the produced voices feature a reasonable naturalness as discussed in Section 3.4.2. As an example, for each of the given voices, it is to be shown that 5 voices can be created by applying VTLN-based voice conversion to the source voice such that their naturalness and dissimilarity scores are at least 3.0:

- At first, the area of the parameter space that provides a naturalness score greater than 3.0 is determined (cf. contour lines in Figure 3.5).

- Then, 5 vectors v_1^5 are distributed inside the region so that the minimum distance between two of these vectors $d(v_i, v_j)$ becomes maximal:

  $$v_1^5 = \arg\max_{v'^5_1} \min_{\substack{i,j=1,\ldots,5 \\ i\neq j}} d(v'_i, v'_j)\ . \qquad (3.55)$$

  This is done by

  - rasterizing the region obtaining a discrete set of allowed vectors,
  - determining the centroid of these vectors, and,
  - as initial setting, assigning the centroid position to all the vectors v'^5_1.
  - Now, the vector v'_1 is moved to the position of that allowed vector which has the maximum distance to the closest neighbor.
  - This is done equivalently with the remaining vectors v'^5_2.
  - The process is iterated by moving v'_1 to the position of that allowed vector which has the maximum distance to the closest neighbor, and so on, until v'^5_1 remain constant.

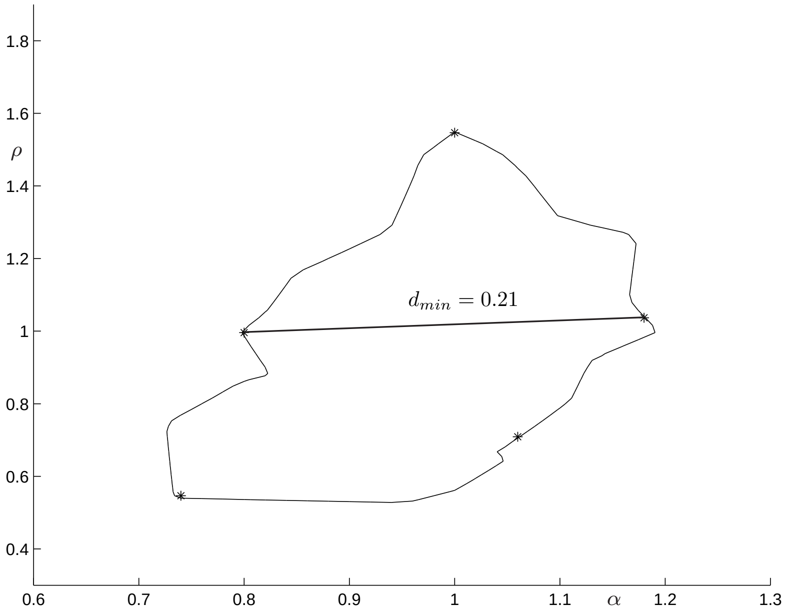

Figure 3.7: Distributing five maximally distant vectors in the area of a minimal naturalness score of 3.0 (example for the female voice). The applied distance measure is given by Equation 3.49.

This algorithm does not necessarily find the optimal positions of v'^{5}_{1}, but, as for the author's experience, produces very reasonable outcomes. A full search through all possible combinations of positions becomes intractable even for medium raster resolutions. Figure 3.7 displays the vectors for the female voice.

- Applying the distance d_{min} between the closest vectors to Equation 3.54 produces an estimate of the minimal subjective dissimilarity D_{min}. For the experimental data given in Tables 3.5 and 3.6, we have $d_{min} = 0.21$ or $D_{min} = 3.5$ for the female voice and $d_{min} = 0.22$ or $D_{min} = 3.0$ for the male, i.e., the requirement of having a minimum dissimilarity score of $D = 3.0$ is fulfilled, if we trust the dissimilarity model introduced in Section 3.4.2.

- To control the fulfillment of the requirement of a minimum dissimilarity score 3.0 independently of the dissimilarity model, a third subjective test was performed where 13 subjects were asked to rank the dissimilarity of all voice combinations. Since five parameter vectors result in ten unique parameter vector pairs, there were 20 speech samples to be prepared, if no samples are to be repeated. Taking into account that there are two source voices, 40 converted speech samples had to be produced and pairwisely presented to the subjects.

  The conversion system's properties are given in Table 3.5. The experiment results in a minimum dissimilarity score of $D_{min} = 4.4$ for the female and $D_{min} = 3.4$ for the male voice, respectively.

3.4.4 Conclusion

This section's considerations aimed at investigating the ability of VTLN-based voice conversion to produce several well-distinguishable voices from one source voice. For that purpose, the dependence between the parameter settings (the warping factor α and the ratio between the means of the fundamental frequency of two compared voices ρ) and naturalness as well as dissimilarity of the produced voices was experimentally investigated. For the latter, i.e. the relation between the respective parameters of two compared voices (α_1, ρ_1 and α_2, ρ_2) and the perceived dissimilarity in terms of an MOS, a model equation was introduced that resulted in a model error of a root mean square of less than 0.4 on the German synthesis corpus which is relatively small compared with the range of the MOS (1.0 to 5.0).

It was shown that in the region of a naturalness of at least 3.0 on an MOS scale, five parameter sets could be distributed such that the dissimilarity model certified a minimum dissimilarity of 3.0. This result could be confirmed by conducting an experiment where for each voice pair the dissimilarity was subjectively assessed.

4 Residual Prediction

In Section 2.4.3, four approaches to residual prediction were described, and their drawbacks were identified as summarized in Table 4.1. Both the residual codebook technique and residual selection were based on the assumption that the correlation between feature vectors and underlying residual is strong enough to reliably predict the residuals based on the features only. However, the following gedankenexperiment weakens this assumption, in particular with respect to residual selection:

Let us consider a number of natural residuals extracted from the frames of a speech sample and just one feature vector derived from a randomly chosen frame. Now, we filter the residuals by means of the single available feature vector's coefficients producing a number of time frames that, dependent on the nature of the underlying residual, feature individual characteristics[32]. When, however, such candidates, i.e. frames that have identical feature vectors but varying residuals, appeared in the residual selection table, this would have a fatal effect: If the residual of a feature vector identical to the one used in the experiment was to be predicted, one of the many varying residuals would be selected by chance, maybe an inadequate one, resulting in unintentional perceptive effects as artifacts.

This gedankenexperiment challenges the validity of the correlation between feature vectors and residuals and is an exaggeration of the real-world situation, since many of the artificially created feature-vector/residual combinations do not appear in real speech data. Still, the phenomenon of selecting inappropriate residuals given a sequence of natural feature vectors has to be taken into account, as experiments on residual selection resulting in a considerable number of artifacts in voiced speech regions show, see report in Section 4.1.

In addition to the correlation insufficiency, both the residual codebook method and residual selection deal with magnitude and phase spectra in different processes frequently leading to mismatches between both spectral components and, finally, to a "rough" [Kain & Macon 01] or "harsh" [Ye & Young 04a] quality of the converted signal. This dissertation investigates three solutions to these dilemmas.

4.1 Residual Smoothing

4.1.1 Residual Smoothing as an Integral Approach

Residual smoothing tries to simultaneously handle inaccuracies of the residual selection due to a lack of correlation and mismatches between magnitude and phase spectra of the predicted residuals. Similar to the integral approach to frequency warping based on complex-valued spline interpolation (cf. Section 2.4.3), the author assumed that the most reliable approach

[32] Informal listening tests even suggest that, if these time frames are concatenated using PSOLA, the text content of the resulting speech can still be recognized, and much of the speaker individuality survives. This confirms that the residuals contain important speaker-dependent information.

technique	drawbacks
vocoder	synthetic sound
source residuals	a third speaker is created
residual codebook	only a few prototypes, phase mismatches
residual selection	artifacts, phase mismatches

Table 4.1: Residual prediction techniques and their drawbacks.

for dealing with the phase spectra is to not separate them from the magnitude spectra, but, instead, treat them as one complex-valued spectrum.

The proposed time-variant residual smoothing is based on the residual selection technique, in particular on Equation 2.16 which delivers the most probable residual spectrum $\tilde{r}$ for a given feature vector $\tilde{x}$ by selecting from a table with all residuals r_1^M seen in training. Here, r_m are complex-valued spectra normalized to a constant number of spectral lines as discussed in Section 2.4.3. In the following, $\tilde{r}_1^K$ is a sequence of residual spectra generated by means of residual selection given the sequence of converted feature vectors $\tilde{x}_1^K$.

As already mentioned in Section 2.4.3, in unvoiced frames, the residual is expected to be white with random phases, thus, “chaotic”, whereas in voiced parts, the signal is expected to be pseudo-periodic. Hence, significant changes in the residual waveform within voiced parts, e.g. due to an insufficient correlation between the residual and the converted feature vector, would be fatal and result in clearly audible artifacts.

The presented solution to this problem is a voicing-dependent residual smoothing which makes sure that neighbored residuals resemble each other in voiced regions but does not change anything in unvoiced regions hardly affected by the aforementioned signal deteriorations. Again, as in Sections 2.3.4 and 3.1, these considerations are not based on a hard voiced/unvoiced decision but on the voicing degree $0 \leq v_k \leq 1$ estimated for each frame $k = 1, \ldots, K$ to be converted.

The proposed smoothing assumes that the residual at a position k is influenced by all its neighbors. This influence is the stronger the closer the neighbor is to the position k. This is achieved by summing up over all residuals produced by means of the residual selection paradigm weighted with a normal distribution whose mean coincides with the position k, to make sure that the contribution of the current position is strongest, and whose standard deviation linearly depends on the voicing degree of the k^{th} frame

$$\sigma_k = v_k \sigma_0 \tag{4.1}$$

where the norm standard deviation σ_0 determines the smoothing strength. It is an empirically determined constant adjusting the tradeoff between smoothing away the typical residual selection artifacts and producing a too smooth sound of the voice by raising the voicing of hardly voiced speech portions, similar to the effect of applying PSOLA to unvoiced sounds discussed in Section 2.3.4. Thus, the smoothed residual at the position k is expressed by:

$$\tilde{r}'_k = \frac{\sum\limits_{\kappa=1}^{K} \mathcal{N}(\kappa|k, \sigma_k) \cdot \tilde{r}_\kappa}{\sum\limits_{\kappa=1}^{K} \mathcal{N}(\kappa|k, \sigma_k)}, \tag{4.2}$$

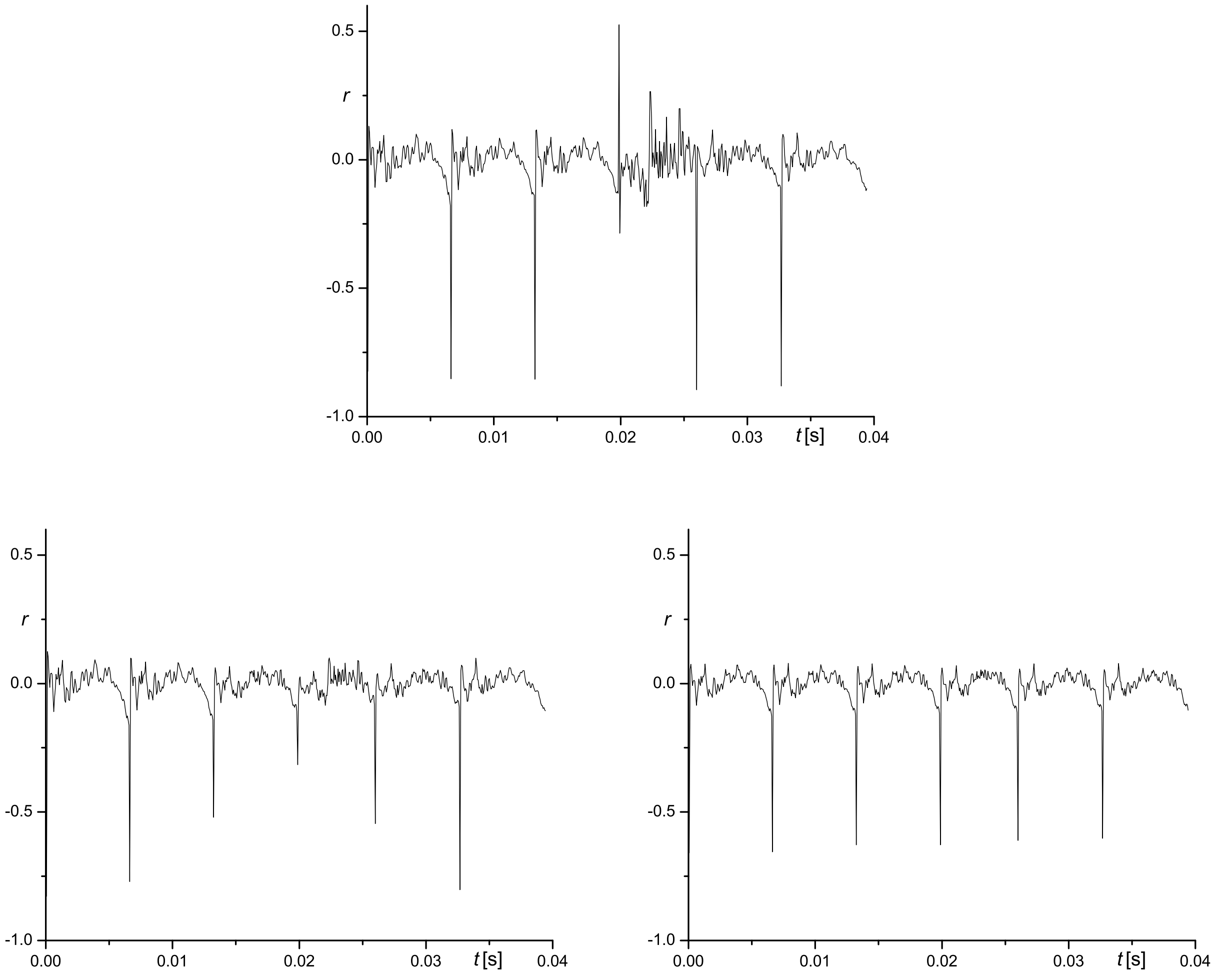

Figure 4.1: Residual smoothing: original and smoothed residuals for $\sigma_0 = 1$ and $\sigma_0 = 3$.

with the one-dimensional special case of Equation 2.9

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{4.3}$$

The normalization denominator in Equation 4.2 is to make sure that the residual weights sum up to one.

For voiced sounds, i.e. a voicing degree of around one, the standard deviation would become maximum and the contribution of the neighbored residuals would cover many of the undesired artifacts. An example is displayed in Figure 4.1. Here, six residuals of consecutive voiced ($v_k = 1$) frames are shown, the fourth obviously misplaced. Now, the residual smoothing is performed with settings $\sigma_0 = 1$ and $\sigma_0 = 3$, respectively, clearly removing the misplaced positive peak[33]. In this example, the aforementioned tradeoff becomes evident: For $\sigma_0 = 1$, there is still a change of the amplitude of the negative peak in the region of the artifact which disappears for $\sigma_0 = 3$. However, in the latter case, many of the frames'

[33] Incorrectly polarized residuals may be due to pitch tracking errors in the training corpus. They also may occur at sound transitions where the correct pitch marks do not necessarily coincide with the most negative peak of the speech signal (see Section 2.3.1).

individual characteristics have disappeared, and all displayed residuals look fairly similar to each other tending to produce less natural speech.

In the opposite case, i.e. for unvoiced frames, the voicing degree and, consequently, the standard deviation σ_k approach 0, and Equation 4.2 becomes

$$\lim_{\sigma_k \to 0} \tilde{r}'_k \quad = \quad \frac{\sum\limits_{\kappa=1}^{K} \Delta(\kappa - k) \cdot \tilde{r}_\kappa}{\sum\limits_{\kappa=1}^{K} \Delta(\kappa - k)} \quad = \quad \tilde{r}_k, \tag{4.4}$$

with the Dirac delta function as the limit of the normal distribution according to [Arfken 85]:

$$\lim_{\sigma \to 0} \mathcal{N}(x|\mu, \sigma) = \Delta(x - \mu). \tag{4.5}$$

That is, in unvoiced frames, the residuals remain unchanged as delivered by the residual selection.

4.1.2 On the Correlation between Source and Target Residuals

The above example (Figure 4.1) suggests that residual smoothing is a strong technique to suppress artifacts. As argued in Section 4.1.1, the drawback of this technique is the tendency to oversmoothing, yielding a loss of naturalness. If there were means to avoid the artifacts already before applying the smoothing, the smoothing strength σ_0 could be reduced leading to a higher naturalness of the generated speech.

As mentioned before, there are two error sources of the residual selection, the potential weakness of the conversion function, and consequently erroneous converted feature vectors, and the insufficient correlation between feature vectors and respective residuals. The detour via imperfect converted feature vectors could be avoided by predicting the target residuals based on the source residuals only, expecting a considerable correlation between them. The idea was to derive a vector representation from the residuals themselves and use them instead of the line-spectral-frequency-based feature vectors $\tilde{x}$ and y required in residual selection Equation 2.16.

After testing a set of residual vector representations using informal listening tests, it turned out that particularly reliable is a vector based on the component-wise absolute value of the time domain representation r of a given residual vector reduced by the constant component (its mean $\bar{r}$) and normalized to a constant number of spectral lines and to an energy of 1.0:

$$R(r) = \frac{\mathrm{abs}(r - \bar{r})}{|r - \bar{r}|}. \tag{4.6}$$

Returning to Equation 2.16, the converted vector $\tilde{x}$ would be replaced by the current source residual vector $R(r)$ and the target vector y_m by the vector representation of the raw target residual $R(r_m)$:

$$\tilde{r} = \mathop{\arg\min}_{r'=r_1,\ldots,r_M} \big|R(r) - R(r')\big|. \tag{4.7}$$

4.1.3 Experiments

This section describes a comparative study of most of the techniques introduced in Section 2.4.3 and in this chapter investigating both the techniques' speech quality and their capability of changing the voice identity from source to target. The following techniques are compared:

- copying *source residuals*,
- *residual codebook* method,
- *residual selection*,
- *residual smoothing* based on the feature criterion (Equation 2.16),
- *residual smoothing** based on the residual criterion (Equation 4.7).

Evaluation corpus was the Spanish synthesis corpus described in Section A.2.1. From this corpus, ten parallel utterances were randomly selected for training and three utterances for testing. The training was performed according to the paradigm described in Section 2.4.2 estimating the linear transformation parameters. The free parameters of featurization and Gaussian mixture model (line spectral frequency order, number of Gaussian mixtures, type of covariance matrix) were chosen in accordance with relevant literature [Kain 01] and taking into account numerical instabilities that arise when the number of parameters to be estimated is too large [Sündermann & Bonafonte$^+$ 04a]. Details of this experiment's system properties can be found in Table 4.2.

As subjective evaluation criteria, the mean opinion score was used for assessing speech quality (see Section 2.5.3), and the extended ABX test served for determining the similarity of converted and target voice (cf. Section 2.5.2). The results of the evaluation are shown in Tables 4.3 and 4.4.

4.1.4 Conclusion

As for the speech quality reported in Table 4.3, a well-known speech synthesis rule claiming that less (signal processing) is more (speech quality) [Beutnagel & Conkie$^+$ 99] lets unprocessed source residuals achieve a 0.9 MOS points better result than the best of its competitors. Among the remaining techniques, residual smoothing outperforms the others in terms of speech quality, although the absolute MOS (2.6) shows that there is still a need for improvement.

Addressing the similarity of converted and target voice, one finds that all assessed techniques succeed in converting the source voice to the target voice to more than 70%, cf. Table 4.4. The only exception are the source residuals where the majority of the listeners had the impression of hearing a third speaker as also found in former studies, see Section 2.4.3. Residual smoothing shows the highest conversion performance.

Finally, the question on which of the correlations, that between converted feature vectors and underlying residuals (conventional) or that between source and target residuals (proposed), achieves better results, is to be addressed. The outcomes of MOS as well as

conversion type	text-dependent
	intra-lingual
source/target language	Spanish/Spanish
alignment technique	dynamic time warping
corpus	Spanish synthesis corpus
voices	1 female, 1 male
amount of training data	42.1s ($M = 6325$) female, 39.2s ($M = 4629$) male
amount of test data	14.0s ($K = 1901$) female, 13.4s ($K = 1703$) male
number of test subjects	10 [29]
sampling rate	$f_\mathrm{s} = 16$kHz
norm frequency	$f_\mathrm{n} = 100$Hz
pitch mark extraction	Goncharoff/Gries (automatic)
voicing determination	Adaptive Multi-Rate (automatic)
conversion technique	linear transformation
features	$D = 16$ order line spectral frequencies
number of mixtures	4
covariance matrix	diagonal
residual codebook	8 mixtures (magnitude prediction)
	16 mixtures (phase prediction)
residual selection	16 mixtures (phase prediction)
residual smoothing	$\sigma_0 = 3.0$ [1.5]
acoustic synthesis	linear predictive PSOLA

Table 4.2: Comparison of residual prediction techniques [of residual smoothing and unit selection]: system properties.

	m2f	f2m	total
source residuals	3.2	3.7	3.5
residual codebook	1.6	1.9	1.8
residual selection	1.7	2.3	2.0
residual smoothing	2.2	2.9	2.6
residual smoothing*	2.2	2.8	2.5

Table 4.3: Comparison of residual prediction techniques: results of the MOS test. *m2f* refers to male-to-female conversion; *f2m* vice versa.

[%]	source	target	neither
source residuals	20	10	70
residual codebook	0	70	30
residual selection	0	70	30
residual smoothing	0	85	15
residual smoothing*	0	80	20

Table 4.4: Comparison of residual prediction techniques: results of the extended ABX test.

extended ABX test do not show significant differences between either criteria, both correlations seem to be equivalently appropriate to the task. Since the conventional paradigm has a wider range of application (in addition to its use for voice conversion, it can be applied to HMM-based speech synthesis where only target feature vectors are available to predict from, cf. comments in Section 2.4.3), the following investigations of this thesis will be based on the feature criterion, i.e. the conventional approach.

4.2 Unit Selection

The conclusion of Section 4.1.4 was that both the transformed feature vector's and the source residual's correlation to the target residual is not sufficient to produce artifact-free speech based on the residual selection paradigm. To suppress these artifacts, residual smoothing has to be applied tending to reduce the speech naturalness by oversmoothing.

If there were a technique for more reliably selecting residuals from the training database so that the selected residual sequence $\tilde{r}_1^K$ already contains less artifacts, one could reduce the smoothing strength σ_0 and, hence, improve the quality of the converted speech. Certainly, the most appropriate residual sequence is one that was seen in training and that fulfills the optimization criterion of Equation 2.16 at the same time. Of course, this will only apply if the converted feature sequence $\tilde{x}_1^K$ is identical to a feature sequence seen in training. Since this will hardly be the case, these conditions should be weakened, and the residual sequence should be allowed to be composed of several subsequences seen in training whose endings fit together. Furthermore, also suboptima of Equation 2.16 should be taken into account in order to obtain subsequences of reasonable lengths. This approach is called *unit selection*, a technique that is widely used in concatenative speech synthesis [Hunt & Black 96].

Generally, in the unit selection framework, two cost functions are defined. The target cost $C^{\mathrm{t}}(u_k, t_k)$ is an estimate of the difference between the database unit u_k and the target t_k it is supposed to represent. The concatenation cost $C^{\mathrm{c}}(u_{k-1}, u_k)$ is an estimate of the quality of a join between the consecutive units u_{k-1} and u_k.

In speech synthesis, the considered units are subphones [Donovan & Woodland 95], phones [Taylor & Black+ 98], or syllables [Lewis & Tatham 99], whereas in the residual prediction case, the base unit length is set to be a single speech frame, since this allows for being independent of additional linguistic information about the processed speech as the phonetic segmentation. Furthermore, the cost functions can easily be defined by interpreting the residuals as database units, i.e., $u = r$.

In the following sections, the properties of the cost functions used for unit-selection-based residual prediction are described. Moreover, it is shown how the final residual sequence is produced.

4.2.1 Target and Concatenation Cost Function

Similar to the residual selection described in Equation 2.16, the appropriateness of a residual r_m seen in training for being selected for the k^{th} frame is determined based on the distance between the corresponding feature vector y_m and the one representing the properties of the k^{th} converted frame, $\tilde{x}_k$. Furthermore, fundamental frequency and energy of the considered residual are to be taken into account. This is to minimize the extent of signal processing to produce the prosodic characteristics of the converted speech and, thus, avoid distortions of the natural waveform. According to [Hunt & Black 96], the target cost C^{t} is calculated as the weighted sum of the distances between the considered features of the target and the candidate:

$$\begin{aligned} C^{\mathrm{t}}(u,t) &= C^{\mathrm{t}}\big(r_m, (\tilde{x}_k, \tilde{f}_{0k}, \tilde{S}_k)\big) \\ &= w_1 d\big(y_m, \tilde{x}_k\big) + w_2 d\big(f_0(r_m), \tilde{f}_{0k}\big) + w_3 d\big(|r_m|^2, \tilde{S}_k\big). \end{aligned} \tag{4.8}$$

$\tilde{f}_{0k}$ and $\tilde{S}_k$ are target fundamental frequency and energy that, in the case of residual prediction for voice conversion, can be derived from the respective parameters of the source speech frame:

- the fundamental frequency is linearly normalized to match the target average by means of the factor ρ (cf. Section 2.3.4),
- the energy is expected to be equivalent to the source residual energy $|r_k|^2$.

This finally yields[34]

$$\begin{aligned} C^{\mathrm{t}}(u,t) &= w_1 d\big(y_m, \tilde{x}_k\big) + w_2 d\big(f_0(r_m), \rho f_0(r_k)\big) + w_3 d\big(|r_m|^2, |r_k|^2\big) \\ &= C^{\mathrm{t}}(k,m). \end{aligned} \tag{4.9}$$

For the weights holds (see also Equation 4.12)

$$w_1 + w_2 + w_3 \leq 1; \quad w_1, w_2, w_3 \geq 0. \tag{4.10}$$

This makes sure that the sum of the cost functions' weights including that of the below discussed concatenation cost is always 1.

d is the Mahalanobis distance that compensates for differences of range and amount of variation between the features used in Equation 4.8:

$$d(x,y) = \sqrt{(x-y)'\Sigma^{-1}(x-y)} \tag{4.11}$$

where Σ is the covariance matrix computed using the respective features of all residuals seen in training.

The cost for concatenating the residuals r_m and r_n is defined using the normalized residual representation $R(r)$ given in Equation 4.6:

$$C^{\mathrm{c}}(m,n) = (1 - w_1 - w_2 - w_3) \cdot \big|R(r_m) - R(r_n)\big|^2. \tag{4.12}$$

When m and n refer to successive frames in the training data, i.e., $n = m + 1$ the concatenation should be optimal, hence, it is defined

$$C^{\mathrm{c}}(m, m+1) = 0. \tag{4.13}$$

For the sake of completeness, furthermore, it is defined

$$C^{\mathrm{c}}(0, m) = 0. \tag{4.14}$$

[34] Note that r_m is the m^{th} target residual of the training data and r_k is the k^{th} source residual of the data to be converted. Consequently, the case $m = k$ does not mean that both residuals are identical.

4.2.2 Finding the Optimal Residual Sequence

The sought-after residual sequence $\tilde{r}_1^K$ is determined by minimizing the sum of the target and concatenation costs applied to an arbitrarily selected sequence from the set of residuals seen in training:

$$\tilde{r}_k = r_{m_k} \quad \text{for} \quad k = 1, \ldots, K \quad \text{with}$$

$$m_1^K = \arg\min_{m'^K_1} \sum_{k=1}^{K} \Big[C^{\mathrm{t}}(k, m'_k) + C^{\mathrm{c}}(m'_{k-1}, m'_k) \Big]; \quad m'_k \in \{1, \ldots, M\}. \tag{4.15}$$

The arbitrarily selected residual sequence is given by the index sequence m'^K_1. As for its definition, there are M^K different index sequences making the problem intractable at first glance. However, it is observed that this formula which is a sum over addends that only depend on the two successive variables m'_{k-1} and m'_k has the same structure as Equation 3.13 proven to be resolvable by means of dynamic programming. Nevertheless, it turns out that it still requires a high computational effort: The full solution of Equation 4.15 results in

$$O \approx K \cdot M^2 \cdot \left(8D^2 + 4D + 6\frac{f_{\mathrm{s}}}{f_{\mathrm{n}}} \right) \text{ops.} \tag{4.16}$$

For this formula's respective parameter values from the experimental corpus used in this work, see Table 4.2. As example, male-to-female conversion is to be considered which is computationally more expensive, since Equation 4.16 is dominated by the variable M (quadratic) as opposed to K (linear).

When these example parameter settings are used and the computation is run on a computer that executes $1^{\mathrm{Gops}}/_{\mathrm{s}}$ it would take around 58 hours, i.e., it operates with 15,600 times real time. In order to build a real-time system, ways to simplify the algorithm must be searched for:

- By introducing a pruning that only considers the 5 best hypotheses, the real-time factor can be reduced to RTF = 12.3.
- Instead of utilizing the Mahalanobis distance (Equation 4.11), the Euclidean distance can be applied without noticeably affecting the residual prediction quality (RTF = 4.4).
- The calculation of the concatenation cost in Equation 4.12 is restricted to the first of the two signal periods contained in the processed residuals, cf. Section 2.3.4 (RTF = 2.1).
- The norm frequency f_{n} used to transform the residuals to vectors of identical lengths, cf. Section 2.4.3, can be duplicated without noticeably affect the behavior of the concatenation cost function (RTF = 1.3).
- Taking into account that almost half of the durations of the speech signal to be converted and of that used for training is actually non-speech (silence or noise), we obtain a real-time factor of 0.3.

Consequently, at least on fast computers, the algorithm is real-time-able.

	m2f	f2m	total
residual selection & smoothing	2.4	2.8	2.6
unit selection & smoothing	3.0	3.1	3.0

Table 4.5: Comparison of residual smoothing and unit selection: results of the MOS test.

[%]	source	target	neither
residual selection & smoothing	2	78	20
unit selection & smoothing	2	83	15

Table 4.6: Comparison of residual smoothing and unit selection: results of the extended ABX test.

4.2.3 Residual Smoothing

As predicted at the beginning of Section 4.2 where the application of the unit selection paradigm to residual prediction was motivated, informal listening tests showed that the output of the unit selection features essentially less artifacts than that of the residual selection discussed in Section 2.4.3. However, since there are still audible signal discontinuities, the application of the residual smoothing described in Section 4.1 is still recommendable. It turns out that the smoothing strength σ_0 can be effectively decreased due to the already smoother input residual sequence. It was determined $\sigma_0 = 1.5$ for the unit selection approach (as opposed to $\sigma_0 = 3.0$ for the residual selection), thus, the output features a higher naturalness and better articulatory properties. In the following section, this statement is to be confirmed by means of a subjective evaluation.

4.2.4 Experiments

The present experiments are to compare the proposed unit-selection-based residual prediction approach with the winner of the evaluation in Section 4.1.3, i.e. with the residual smoothing based on the conventional feature criterion, cf. discussion in Section 4.1.4.

This time, 29 subjects, 25 of whom specialists in speech processing, participated in the tests on speech quality (MOS) and speech similarity (ABX). For details on speech corpus and system properties, refer to Table 4.2. Table 4.5 reports the results of the MOS test and Table 4.6 those of the extended ABX rating depending on the residual prediction technique and the gender combination.

4.2.5 Conclusion

Having a look at the speech quality results, one notes that the unit selection approach clearly outperforms the residual selection: In particular, for the obviously harder task of this experiment, the male-to-female conversion, unit selection achieved a considerable gain of 0.6 MOS points. The overall result (MOS = 3.0, i.e. a *fair* speech quality) is still essentially worse than natural speech (MOS = 4.8), but already suitable for many applications that do not require high fidelity speech, for instance in telecommunications.

Also the outcomes of the extended ABX test show improvements of unit selection over residual selection: In 83 % of the cases, unit selection successfully converted the source voice to the target voice. This is 5 % absolute more than the result achieved by residual selection.

4.3 Applying VTLN to Residuals

Although the development of the unit-selection-based residual prediction led to a strong performance gain, in particular with respect to speech quality (cf. experiments in Sections 4.1.3 and 4.2.4), there is still a considerable gap between unprocessed residuals (source residuals achieved MOS = 3.5 in Table 4.3) and unit selection (MOS = 3.0 in Table 4.5). Unfortunately, the former showed a much worse performance as for the similarity score, reported in Tables 4.4 and 4.6.

Some applications, however, require a higher speech quality while accepting a lower voice similarity. It was therefore to be investigated if the source residuals could be *transformed* keeping the natural time order, energy, and pitch trajectory instead of *selecting* from the training pool of target residuals as done by residual selection as well as by the unit selection approach. Such a transformation paradigm is VTLN-based voice conversion (see Chapter 3) which the author intended to apply only to the residuals, whereas the features were to be transformed by means of the linear transformation paradigm as usual. This idea faces two contradictions, though:

1. As the name implies, VTLN (vocal tract length normalization) is to change the length, or more generally, the shape of the vocal tract. According to the source-filter model discussed in Section 2.3.2, the vocal tract is represented by the features, whereas the residuals are supposed to be an approximation of the excitation. Consequently, an application of VTLN to the residuals should hardly change the speech characteristics.

2. Furthermore, according to experiences in speech recognition, applying VTLN in conjunction with a linear transformation in feature space should not help since the linear transformation already compensates for the effect of speaker-dependent vocal tract lengths and shapes [Pitz & Ney 05].

These statements suggest the thesis

Applying VTLN to residuals does not change the input speech.

In spite of the above contradictions, the author conducted several informal listening tests and found that, perceptively, the antithesis seems to hold:

Applying VTLN to residuals is equivalent to applying VTLN to the input speech.

corpus voices amount of test data	first TC-Star evaluation corpus 2 female, 2 male 208.0s female1, 160.7s female2, 189.9s male1, 157.8s male2
sampling rate norm frequency	$f_\mathrm{s} = 16\mathrm{kHz}$ $f_\mathrm{n} = 100\mathrm{Hz}$
pitch mark extraction	Goncharoff/Gries (automatic)
conversion technique features	(a) none; (b) time domain VTLN 16th order line spectral frequencies
residual prediction	(a) time domain VTLN; (b) none

Table 4.7: Applying VTLN to residuals: system properties.

4.3.1 Experiments

In this section, the aforementioned paradox is to be investigated by objective means. According to the antithesis, two speech samples are compared:

(a) the output speech (with the spectral vector sequence $\tilde{X}^{(a)K}_{\;\;1}$) that results from splitting the input speech X_1^K into features and residuals, applying VTLN to the residuals, and filtering the latter by means of the (unchanged) features (residual-VTLN-converted speech),

(b) the output speech $\tilde{X}^{(b)K}_{\;\;1}$ that results from applying VTLN directly to the input speech (VTLN-converted speech).

Often when comparing parallel speech samples a distance measure is used that is related to the human perception as for instance the log-spectral distortion, cf. Section 2.5.1. This distance measure, however, is applied to the spectra of parallel speech frames and measures the quadratic difference of the respective cepstral coefficients. According to [Papamichalis 87], the latter can be converted to linear predictive coefficients and, finally, to line spectral frequencies, the feature type used in this study. Consequently, the log-spectral distortion does not show any difference between $\tilde{X}^{(a)}$ and X, since the features were not changed. This would produce the trivial result $D^{(a)} = 0$ for the standard of comparison defined in Equation 4.18 if it were based on log-spectral distortion.

In order to take into account spectral details not covered by the features, the Euclidean distances between the full magnitude spectra are computed, summed up over all speech frames, and normalized by the number of speech frames and spectral lines[35]:

$$D^{(ab)} = \frac{f_\mathrm{n}}{K f_\mathrm{s}} \sum_{k=1}^{K} \left| \mathrm{abs}\big(\tilde{X}_k^{(a)}\big) - \mathrm{abs}\big(\tilde{X}_k^{(b)}\big) \right|. \tag{4.17}$$

To have a standard of comparison, also the sum of the distances between the spectra of the input speech X_1^K and those of the residual-VTLN-converted speech $\tilde{X}^{(a)}$ is calculated:

$$D^{(a)} = \frac{f_\mathrm{n}}{K f_\mathrm{s}} \sum_{k=1}^{K} \left| \mathrm{abs}\big(\tilde{X}_k^{(a)}\big) - \mathrm{abs}\big(X_k\big) \right|. \tag{4.18}$$

$D^{(b)}$ for the VTLN-converted speech is computed equivalently.

[35] The component-wise absolute value abs is used, since the distance between complex-valued spectra sometimes is misleading. For instance, the complex-valued distance between identical spectra rotated by 180° can be a considerable number, although, perceptively, they are identical.

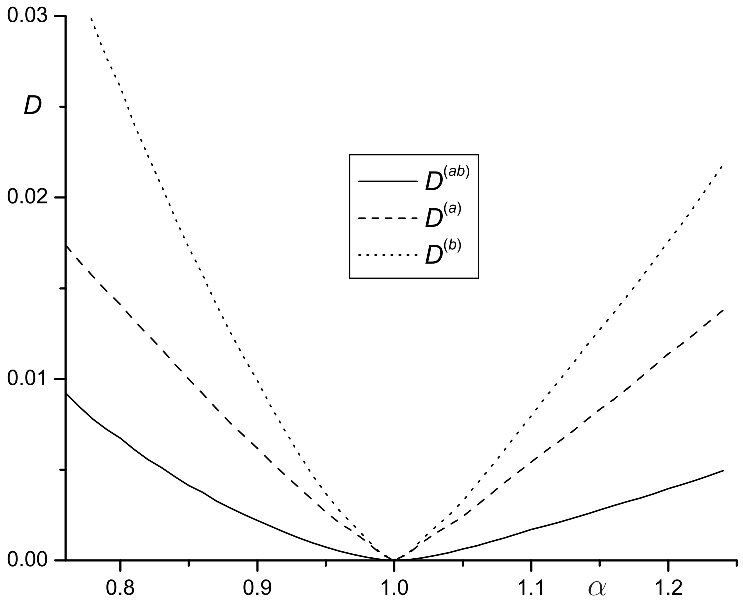

Figure 4.2: Applying VTLN to residuals vs. applying VTLN to input speech: inter-comparison ($D^{(ab)}$) and comparison with the input speech ($D^{(a)}$ and $D^{(b)}$).

The experiments are based on the first TC-Star evaluation corpus described in Section A.2.3. The system properties are given in Table 4.7. Figure 4.2 displays the measures $D^{(ab)}$, $D^{(a)}$, and $D^{(b)}$ as a function of the warping factor α in the range suggested in Equation 3.4. Since the outcomes were very similar for all considered voices, the figure shows only the results for one of them (male1).

4.3.2 Conclusion

Figure 4.2 shows that, independent of the warping factor α, the following relation holds

$$D^{(b)} > D^{(a)} > D^{(ab)}. \tag{4.19}$$

The outcome that $D^{(b)}$, i.e. the distance between input speech and VTLN-converted input speech, is greater than $D^{(a)}$, the distance between input speech and residual-VTLN-converted input speech, is not surprising, since in the latter case, an essential component of the speech representation, the spectral features, were kept constant. However, as opposed to the thesis formulated in Section 4.3, input speech and residual-VTLN-converted input speech are not similar at all. The corresponding measure $D^{(a)}$ does approach zero except for values of α around 1.0 (slight warping) where it shows the same behavior like VTLN-converted input speech. Consequently, the thesis is disproved.

As for the antithesis, when looking at the measure $D^{(ab)}$ which is considerably smaller than the standards of comparison $D^{(a)}$ and $D^{(b)}$ it turns out that, indeed, VTLN-converted and residual-VTLN-converted speech are relatively similar to each other. In particular, for warping factors α around 1.0, $D^{(ab)}$ clearly approaches zero. This means, the antithesis could be objectively confirmed to a certain extent.

So, what is wrong with the two contradictions given in Section 4.3?

1. The source-filter model assumes that the excitation (represented by the residual) is white, i.e., it has a flat magnitude spectrum (cf. Section 2.4.3). In this case, an application of VTLN would not change the residual magnitude spectrum, and the magnitude spectra of input speech and residual-VTLN-converted input speech were identical. In the real world, however, the residuals are not white. They contain a considerable amount of spectral fine structure which is affected by both the application of VTLN to the residuals and to the input speech itself.

2. The realization that a linear transformation in feature space already includes the effects of VTLN is of interest in speech recognition research that mostly does not look at the residuals at all. But again, since the residuals carry a lot of voice-dependent information not covered by the features, the aforementioned equivalence does not apply to a voice conversion algorithm that processes the residuals by means of VTLN and the features by means of linear transformation.

5 Text-Independent and Cross-Language Voice Conversion

In Sections 2.1 and 2.2, the author argued that an approach to text-independent voice conversion is to search for similar phonetic contents in the speech of source and target speaker. After aligning these contents, the respective speech data can be considered as being parallel which is the condition for applying conventional voice conversion training techniques as discussed in Chapters 3 and 4.

Since in addition to text independence, also cross-language portability and fast adaptability of the respective algorithms to new languages are major objectives of this work, algorithms had to be developed that

1. avoid the use of language-specific databases or algorithms,

2. avoid the use of linguistic information such as the underlying text, pronunciation lexica, or language models of source or target language,

3. avoid the use of large speech databases, and

4. are robust with respect to changes of source or target language, i.e., differences in the phoneme sets of the involved languages.

None of these criteria is fulfilled by the speech-recognition-based text-independent approach discussed in Section 2.1 which is, to the best of the author's knowledge, the only considerable work ever addressed to the subject of text-independence of voice conversion.

The above formulated paradigm for producing an alignment between non-parallel speech mentions the search "for similar phonetic contents in the speech of source and target". But how can similar phonetic contents be found, if language-specific knowledge, i.e. linguistic information, databases, and algorithms, must not be used?

This work's approach for answering this question is based on the fact that phonetic contents are manifested in the speech data itself; phonemes can be distinguished based on their differing spectral structure. These spectral differences between phonemes are the fundament of automatic speech recognition which, however, suffers a significant lack of correctly recognizing the phonemes from speech (on standard databases as the TIMIT corpus [Fisher & Doddington$^+$ 86], state-of-the-art phoneme recognizers produce error rates between 25% and 40% [Reynolds 94, Tan & Fu$^+$ 96, Ali & Spiegel$^+$ 99]). This normally is compensated by using contextual information derived from pronunciation lexica and language models. Since in the case of text-independent voice conversion neither the underlying text nor the phonemes themselves are of interest but the proper alignment of speech segments, this lack may potentially be ignored.

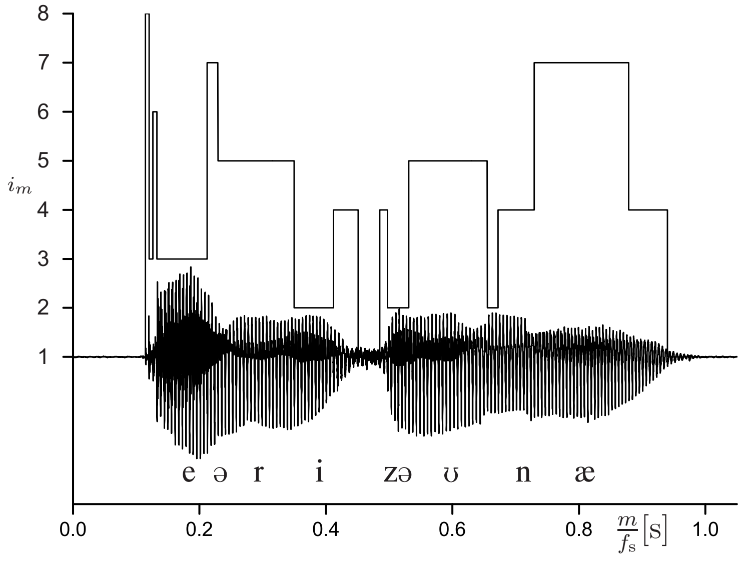

Figure 5.1: Automatic segmentation of artificial phonetic classes: an example; $I = 8$.

5.1 Automatic Segmentation and Mapping of Artificial Phonetic Classes

5.1.1 Segmentation

Assuming a phonetic structure in a given speech sample that is manifested in the differences of their spectral characteristics, a first step towards a phonetic alignment as discussed in the introduction of this chapter is the following: Given the vector sequence X_1^M of magnitude spectra derived from pitch-synchronous frames whose number of spectral lines is constant (see Section 2.4.3), a clustering algorithm such as k-means or expectation-maximization [Seber 84] is performed splitting these vectors into I artificial phonetic classes. As a result, each spectrum X_m is assigned one of these I classes, i_m. An example is shown for of a female voice uttering the word "Arizona" and distinguishing $I = 8$ classes in Figure 5.1.

Now, the class centroid, i.e. the most representative spectrum of each class, can be determined:

$$\hat{X}_i = X_{m_i} \quad \text{with} \quad m_i = \arg\min_{m \in M_i} \sum_{m' \in M_i} |X_m - X_{m'}|^2 \quad \text{and} \quad M_i = \{m : i_m = i\}. \tag{5.1}$$

Although, generally, the class center, i.e. the mean of all vectors of a class,

$$\bar{X}_i = \frac{1}{|M_i|} \sum_{m' \in M_i} X_{m'}, \tag{5.2}$$

is located close to the centroid, in the following, the centroid is used, since it is a *natural* spectrum rather than the center that, for the fact of being a sum of spectra, features considerably less spectral details which, as shown in Chapter 4, can be of high importance for the application to voice conversion. An example of such a centroid vs. center is shown in Figure 5.2.

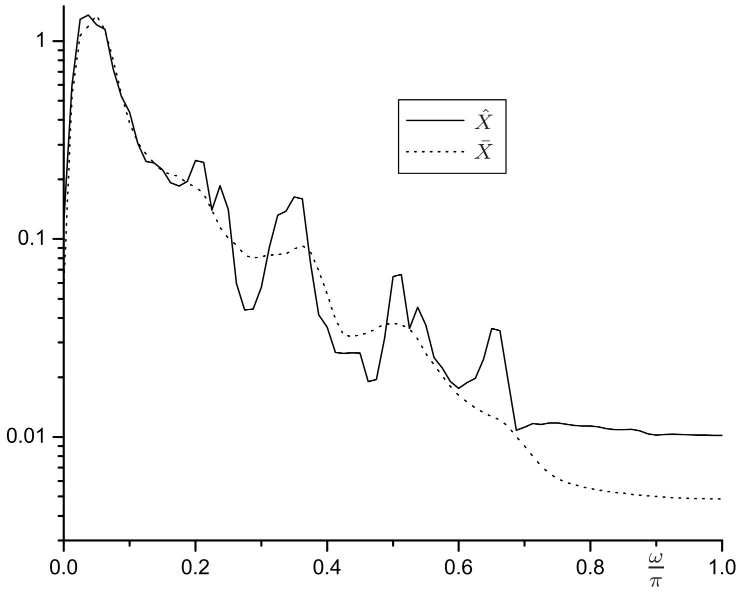

Figure 5.2: Automatic segmentation of artificial phonetic classes: class centroid $\hat{X}$ vs. class center $\bar{X}$.

This segmentation of speech data into artificial phonetic classes is applied to both source speech, given by the vector sequence X_1^M, and target speech, given by the vector sequence Y_1^N which is non-parallel to X_1^M. This produces the class centroids $\hat{X}_i$ and $\hat{Y}_j$ and the class sets M_i and N_j which contain the indices of all vectors of input sequences belonging to the class i or j, respectively, as introduced in Equation 5.1. Here, M is the number of vectors of the source speech data and N that of the target data, and $i \in \{1, \dots, I\}$ are the source classes and $j \in \{1, \dots, J\}$ the target classes, respectively.

5.1.2 Class Mapping

As argued above, the objective is to find an artificial phonetic mapping, i.e. a mapping between the classes i and j. The unique mapping $j(i)$ which assigns exactly one target class j to a given source class i can be found by comparing the respective class centroids

$$j(i) = \underset{j'=1,\dots,J}{\arg\min}\, d(\hat{X}_i, \hat{Y}_{j'}), \quad i = 1, \dots, I. \tag{5.3}$$

Similar to the consideration in Section 2.4.1, the distance measure d should take into account vocal tract differences of source and target speaker to produce a reliable mapping. Therefore, VTLN is applied to compensate for effects of different vocal tract lengths. The warping parameters are optimized such that the Euclidean distance becomes minimum:

$$d(X, Y) = \min_{\xi_1, \xi_2, \dots} \left| \tilde{X}(\xi_1, \xi_2, \dots) - Y \right|. \tag{5.4}$$

As a limit case of this approach, the piece-wise linear warping function introduced in Section 3.2 is considered based on as many segments as there are frequency lines in the spectrum. This leads to the highest possible resolution of the warping function and the closest match between target and warped source spectrum. In relevant literature, this approach is referred to as dynamic frequency warping which has already been applied to vowel classification [Ainsworth & Paliwal$^+$ 84] and normalization [Matsumoto & Wakita 86] i.e. problems related to that of mapping phonetic contents.

5.1.3 From Class to Frame Mapping

So far, a mapping between source and target centroids has been produced and, consequently, a mapping between the respective classes. However, for the training procedures explained in Sections 2.4.2 and 3.1, parallel sequences of source and target vectors are required. Such an alignment can be obtained by normalizing the target class to optimally fit the corresponding source class by reducing each of its member vectors by the distance between source and target centroid[36].

This is done as follows: For a given source vector X, one takes the respective class i and its corresponding target class determined by means of Equation 5.3, $j(i)$, calculates the distance between the classes which is the difference of their centroids $\hat{Y}_{j(i)} - \hat{X}_i$ and normalizes the given vector accordingly:

$$\tilde{X} = X + \hat{Y}_{j(i)} - \hat{X}_i. \tag{5.5}$$

Now, the vector of the target class $j(i)$ being closest to the normalized source vector $\tilde{X}$ is searched for:

$$\tilde{Y} = \underset{Y \in \{Y_n:\ n \in N_{j(i)}\}}{\arg\min} |\tilde{X} - Y|. \tag{5.6}$$

This is done for every vector X_m, $m = 1, \ldots, M$ of the source speech data yielding the corresponding target vectors $\tilde{Y}_m$. Now, this frame alignment is finally used to conduct conventional voice conversion parameter training according to Sections 2.4.2 and 3.1.

5.1.4 Experiments

The initial experiments were to compare the performances of text-dependent alignment (dynamic time warping as used in this thesis so far) and text-independent alignment (automatic segmentation and mapping) in the general linear transformation framework introduced in Section 2.4.2. As experimental corpus, the Spanish synthesis corpus described in Section A.2.1 was used. The amount of training data was varied by selecting between 1 and 64 parallel utterances from the female and the male speaker, and also the number of mixture densities of the Gaussian mixture model was varied between 1 and 8, since the author expected that more training data would support using more mixture densities which would better represent the voice characteristics. The utterances were aligned by means of both text-dependent and text-independent techniques. Then, in both cases, the standard training according to Section 2.4.2 was carried out. The system properties of these experiments are shown in Table 5.1.

[36] The technique described in the following assumes similar behavior of the vectors of a target class and the normalized source class (micro level) which may not be the case. This assumption would be remedied, if the number of classes was high enough such that the vector variations within a class became infinitesimal. However, as investigated in [Sündermann & Ney 03a], the more classes are used for the mapping, the more similar the classes become (macro level) weakening the strength of the voice conversion. Consequently, as a trade-off between having too similar behavior on the micro and on the macro level, $I = J = 8$ turned out to be a good choice for the present data.

conversion type	text-dependent	text-independent
source/target language	intra-lingual Spanish/Spanish	
alignment technique	dynamic time warping	autom. segmentation, mapping number of classes: $I = J = 8$
corpus	Spanish synthesis corpus	
voices	1 female, 1 male	
amount of training data	1 to 64 utterances	
amount of test data	10 utterances	
sampling rate	$f_\mathrm{s} = 16\mathrm{kHz}$	
norm frequency	$f_\mathrm{n} = 100\mathrm{Hz}$	
pitch mark extraction	Goncharoff/Gries (automatic)	
conversion technique	linear transformation	
features	16th order line spectral frequencies	
number of mixtures	1, 2, 4, or 8	
covariance matrix	diagonal	
residual prediction	vocoder	
acoustic synthesis	frequency domain PSOLA	

Table 5.1: Automatic segmentation and mapping: system properties.

	text-dependent	text-independent
female-to-male	0.38	0.49
male-to-female	0.39	0.47

Table 5.2: Automatic segmentation and mapping: log-spectral distortion.

Then ten parallel utterances of both speakers which were different from the training data were selected for testing and also aligned by means of dynamic time warping. This was to produce parallel test data to apply the log-spectral distortion as objective error measure as described in Section 2.5.1. Remember that the log-spectral distortion is the smaller, the more successful the conversion is.

Figure 5.3 shows the log-spectral distortion of the male-to-female conversion as a function of the number of training utterances and the number of mixture densities. The behavior of the female-to-male conversion was very similar. In accordance with the observation in Section 4.1.3, for too great numbers of mixture densities applied to a too small amount of training data, the expectation-maximization algorithm estimating the conversion parameters does not converge or produces unreasonable outcomes. In this case, the respective data points in Figure 5.3 are missing. Obviously, such estimation problems arise more frequently in text-independent alignment. In Table 5.2, the best results of the respective techniques are shown, this is 64 training utterances and 8 mixture densities for the text-dependent case and two training utterances and one mixture density for the text-independent case.

5.1.5 Conclusion

The automatic segmentation and mapping presented in this section produces an alignment between source and target spectral vectors based on the idea of artificial phonetic classes and correspondences between such classes of the source and the target speech data. Within

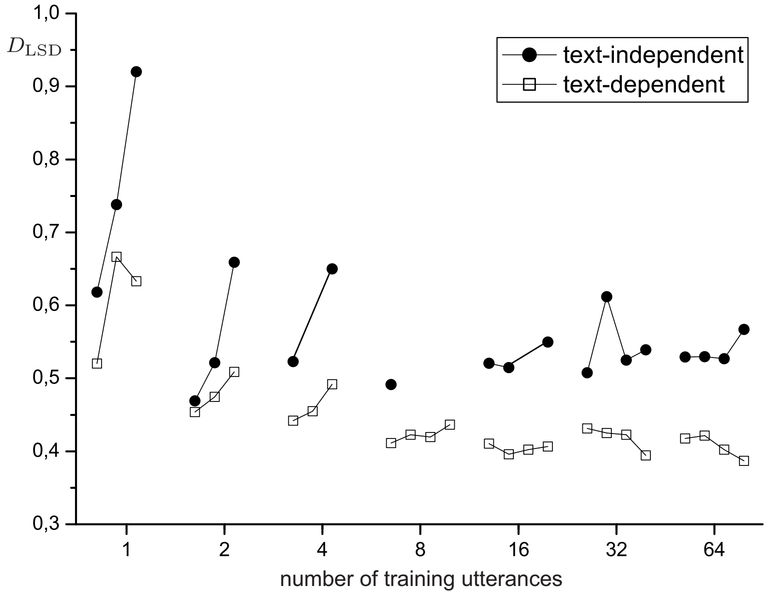

Figure 5.3: Automatic segmentation and mapping: log-spectral distortion as a function of amount of training data and number of mixture densities. For each set of training utterances, from the left to the right, the number of mixture densities is 1, 2, 4, 8.

corresponding classes, variations of the source spectra are mapped to variations of the target spectra.

The objective comparison of the conventional text-dependent alignment paradigm (dynamic time warping) and the proposed text-independent approach shows that in case of male-to-female conversion, the text-dependent approach achieves a result that is around 17% better than the text-independent one as shown in Table 5.2. For the female-to-male conversion, the difference is about 22%. In Figure 5.3, it can be seen that already for two training utterances, the text-independent training technique produces a saturation effect, and using more than one Gaussian mixture density does not improve the performance at all.

Together with the observation that the text-independent training suffers a higher tendency to instabilities (as mentioned in Section 5.1.4), it can be concluded that the alignment has to be particularly improved in terms of the naturalness, i.e. continuity of speech. The automatic segmentation and mapping technique allowed arbitrary discontinuities in the aligned target speech leading to the mentioned negative effects.

5.2 Text-Independent Voice Conversion Based on Unit Selection

So far, the algorithms for text-independent speech alignment, i.e. the one applying a speech recognizer as described in Section 2.1 and the automatic segmentation and mapping introduced in Section 5.1, followed a common approach: They tried to

- extract a phonetic structure underlying the processed speech,
- find a mapping between the phonetic classes of source and target speech,

- and transform the class mapping back to frame mapping.

All these steps may produce errors that accumulate and may lead to a poor parameter estimation or even to a convergence failure of the parameter estimation algorithm as reported in Section 5.1.4. Hence, it would be helpful to avoid the detour through the class layer and, instead, find the mapping only using frame-based features.

Given the source speech feature sequence x_1^M, one would straightforwardly be able to determine the best-fitting sequence of corresponding target feature vectors $\tilde{y}_1^M$ by selecting from an arbitrary target feature sequence y_1^N. This could be done by minimizing the Euclidean distance between the compared feature vectors[37]:

$$\tilde{y}_m = \underset{y \in \{y_1,\ldots,y_N\}}{\arg\min} |x_m - y|, \quad m = 1, \ldots, M. \tag{5.7}$$

However, as concluded in Section 5.1.5, there is a lack as for the continuity of the aligned target speech which has to be improved by taking into account also the frame context. Equation 5.7 ignores the context, since it determines the target feature vector $\tilde{y}_m$ just looking at the corresponding source feature vector x_m. Now, also distances to the neighbor vectors $\tilde{y}_{m-1}$ and $\tilde{y}_{m+1}$ are to be considered to improve continuity. This is done by once again using the unit selection framework introduced in Section 4.2.

To avoid the detour via the underlying phonetic structure, the base unit length is set to be a single frame. Interestingly, this time, both target and concatenation cost are computed in feature space as opposed to the application of unit selection to residual prediction, discussed in Section 4.2. There, the target costs were computed in feature space, whereas the concatenation costs were computed in residual space. Consequently, in the case of unit-selection-based speech alignment, the feature vectors can be interpreted directly as database units, i.e. $t := x$ and $u := y$.

For determining the target vector sequence $\tilde{y}_1^M$ best-fitting a given source sequence x_1^M, the sum of the target and concatenation costs applied to an arbitrarily selected sequence of vectors taken from the non-parallel target sequence y_1^N has to be minimized. Since all compared units are of the same structure (they are feature vectors) and dimensionality, the cost functions may be represented by Euclidean distances finally yielding

$$\tilde{y}_1^M = \arg\min_{y_1^M} \sum_{m=1}^{M} \left\{ w|y_m - x_m|^2 + (1-w)|y_{m-1} - y_m|^2 \right\}. \tag{5.8}$$

Here, the parameter w is for adjusting the tradeoff between fitting accuracy of source and target sequence and the spectral continuity criterion. As opposed to residual prediction based on unit selection, this time, fundamental frequency and energy are not taken into account, since they do not directly influence the feature vectors. Finally, only the feature vectors are used for the training of the linear-transformation-based conversion function ignoring fundamental frequency and energy. Consequently, their contribution weights (w_1 and w_2 according to Equation 4.8) are set to zero and $w := w_1$. The weight w is determined by conducting informal listening tests; the author found that a rather reliable setting is $w = 0.3$ (for the dependence of w on the amount of training data, see the discussion on the speech alignment paradox in Section 5.4.3).

[37] As opposed to Equation 5.3, this time, no VTLN is applied, since this would result in severe computational problems. Even without VTLN, the algorithm described in the following faces real-time factors greater than 100 as pointed out in Section 5.2.1. Nonetheless, the presented technique is rather robust against vocal tract differences as the experimental results of Section 5.2.2 show.

5.2.1 Time Behavior

As already discussed in Section 4.2.2, unit selection using single speech frames as base unit turns out to be very time-consuming. The structure of Equation 5.8 allows for applying dynamic programming making the problem tractable. However, unlike well-known applications of dynamic programming (e.g. to dynamic time warping), in the case of unit selection-based speech alignment the search space is considerably larger: Conventionally, in the former case, the possible successors of a feature vector y_n are limited to the set $\{y_n, y_{n+1}, y_{n+2}\}$, whereas in the latter case all vectors are allowed. This leads to a time complexity of $\mathbf{O}(M \cdot N^2)$.

Let us consider an example: For the male-to-female conversion described in Section 5.2.2, about 80s speech data of both source and target speaker was used. Taking into account the different fundamental frequencies of the speakers (or rather: their different frame lengths), this resulted in $M = 11400$ and $N = 16800$ vectors. According to the above complexity, the expression in the curly braces of Equation 5.8 had to be computed about $3.2 \cdot 10^{12}$ times. After performing some of the steps discussed in Section 4.2.2 to reduce the complexity, the computation still took more than 3 hours on a 3GHz Intel Xeon processor corresponding to a real-time factor of about 150. It should, however, be mentioned that in several applications of voice conversion, the training can be performed offline accepting real-time factors above 1.0. The unit selection can also be parallelized which allows for further accelerating the computation.

5.2.2 Experiments

The following experiments were to compare text-dependent voice conversion using dynamic time warping with the unit selection-based text-independent conversion. This was done by conducting subjective tests assessing both the voice similarity between converted and target speech and the overall speech quality of the conversion output. The experiments were based on the first TC-Star evaluation corpus. From each of the 50 utterances, 10 utterances were randomly selected to be used as test data.

For the text-independent case, the remaining 40 parallel utterances were randomly split yielding 20 different utterances for the source and for the target speaker, respectively, to have a real-world scenario where source and target texts are completely different. For the text-dependent case, 20 randomly selected parallel utterances were used for training. From the possible twelve source/target speaker combinations, each gender combination was selected once.

In the test, 13 subjects (exclusively speech processing specialists) participated, the system properties are shown in Table 5.3. As a standard of comparison, the results were compared to those produced in the first evaluation campaign of the project TC-Star (cf. Section A.1). Here, only the text-dependent data was evaluated. Both the assessed speech data and the evaluation metrics were the same as in the present experiment. The number of test subjects was 17. In addition to the text-dependent system according to Table 5.3, in the TC-Star evaluation, another voice conversion system, in the following referred to as X, was evaluated.

As subjective error measures served mean opinion scores on both voice similarity and overall speech quality as defined in Section 2.5.3. The outcomes of the experiments are shown in Table 5.4.

conversion type	text-dependent	text-independent
source/target language	intra-lingual Spanish/Spanish	
alignment technique	dynamic time warping	unit selection $w = 0.3$
corpus	first TC-Star evaluation corpus	
voices	2 female, 2 male (non-professionals)	
amount of training data	≈ 80s per speaker	
amount of test data	≈ 40s per speaker	
number of test subjects	13 professionals [17 non-professionals]	
sampling rate	$f_\mathrm{s} = 16\mathrm{kHz}$	
norm frequency	$f_\mathrm{n} = 100\mathrm{Hz}$	
pitch mark extraction	Goncharoff/Gries (automatic)	
conversion technique	linear transformation	
features	16$^{\mathrm{th}}$ order line spectral frequencies	
number of mixtures	4	
covariance matrix	diagonal	
residual prediction	VTLN	
acoustic synthesis	time domain PSOLA	

Table 5.3: Text-independent voice conversion based on unit selection: system properties of the comparison text-dependent with text-independent voice conversion [the first TC-Star evaluation campaign].

			TC-Star			
	source	target	X	text-dependent		text-independent
quality	4.6	4.7	1.6	3.2	2.7	2.5
similarity	1.8		2.9	2.0	2.6	2.9

Table 5.4: Text-independent voice conversion based on unit selection: MOS tests on speech quality and similarity. The table contains the results for clean source and target speech (and the comparison of them in the case of similarity), text-independent and text-dependent system, and those of the first TC-Star campaign on the very same text-dependent system and another system X.

5.2.3 Conclusion

When comparing the results for speech quality and similarity of text-independent and text-dependent techniques it turns out that, in terms of both measures, they achieve similar performances. The similarity scores are comparable to those reported on other systems as shown in the evaluation discussed in Section 5.3.1, whereas the speech quality was rated worse than expected. Altogether, there is a large gap between the results of the identical tests of the text-dependent case and the TC-Star evaluation. The speech quality scores differ by -0.5 MOS points which could be interpreted as a considerable deterioration of speech quality. As for the speech similarity, the difference is 0.6 MOS points corresponding to a considerable improvement. However, as the speech samples in both tests were identical, the reason for this virtual deterioration and improvement must be due to the evaluation framework. The major differences between the two subjective tests are

- the subjects' scientific background (naïve subjects in the case of TC-Star vs. speech processing experts in the other case) and

- the fact that the second system assessed in the TC-Star evaluation, X, featured considerably different characteristics – it achieved a higher voice similarity but a worse speech quality. As motivated in Section 2.5, when being presented two techniques of clearly distinguishable quality, obviously, the test subjects apply a *relative* rating: They tend to emphasize the contrast between the compared techniques by exaggerating the scores, i.e. increasing the score of the better and decreasing that of the worse candidate. This contradicts the idea of the MOS which is intended to produce *absolute* scores certifying a certain level of performance. To overcome this tendency, it could be helpful to define standards of comparison by presenting speech samples which are supposed to be rated MOS = 1, MOS = 2, etc. to the subjects, before the test starts. However, the definition of these standards of comparison itself is highly subjective and would require extensive subjective tests or even a standardization by a recognized consortium like the International Telecommunication Union which originally defined the MOS.

  The similar outcomes of text-independent and text-dependent voice conversion suggest that the subjects seemed to cluster them into one single class and compare them to the only other class evaluated, namely the clean speech. Taking into account this contrasting effect, the scores of the converted voice were decreased leading to lower speech quality MOSs. On the other hand, in the case of the TC-Star evaluation, the presence of a third class with clearly worse speech quality led to increased scores for the better sounding technique explaining the gap of the outcomes of both text-dependent techniques in Table 5.4. As for speech similarity, a similar conclusion can be drawn.

5.3 From Text-Independent to Cross-Language Voice Conversion

In the beginning of this chapter, a list with four criteria was given which are necessary to build a text-independent cross-language voice conversion module as motivated in Section 2.2, e.g. for the use in the speech-to-speech translation framework. The techniques presented in this chapter so far fulfill the criteria **1**, the avoidance of language-specific databases and algorithms, **2**, the avoidance of linguistic information, and **3**, the avoidance of large speech databases, such that the only remaining question is item **4**, the robustness of the presented algorithms with respect to differences of the phoneme sets of the processed languages.

In former studies [Sündermann & Ney 03a, Sündermann & Ney$^+$ 03b], the author investigated the application of the automatic segmentation and mapping approach to the cross-language task on the two languages German and English. However, due to the drawbacks of the automatic segmentation and mapping discussed in Section 5.1.5, the following investigations focus on the unit selection paradigm introduced in Section 5.2.

In training, one is given speech data of a source speaker in the target language and data of a target speaker in any other language. The speech is aligned according to the unit

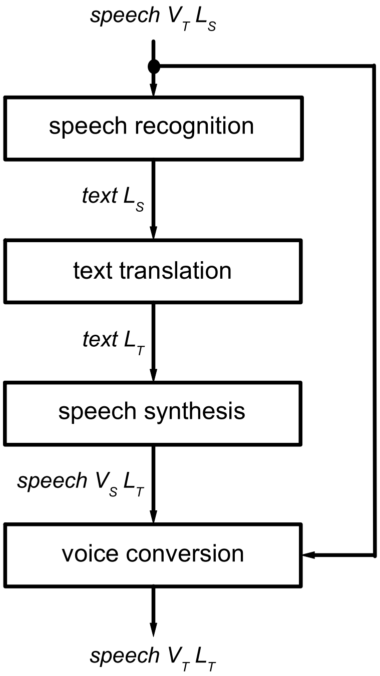

Figure 5.4: The components of a speech-to-speech translation system with voice conversion. V and L stand for voice and language, S and T for source and target.

selection paradigm, and the training is performed as described in Section 2.4.2 resulting in the conversion parameters.

In conversion phase, an utterance of the source speaker in the target language is transformed by means of the trained conversion parameters producing speech in the target language carrying the voice characteristics of the target speaker. To illustrate this procedure, Figure 5.4 shows the components of a speech-to-speech translation system such as the one used in the project TC-Star (see Section A.1) with a text-independent cross-language voice conversion module. In the following, experiments are described which aim at investigating the applicability of the unit-selection-based voice conversion to the cross-language task.

5.3.1 Experiments

The following experiments were conducted in the framework of the second evaluation campaign of the project TC-Star. The aim was to compare several approaches to voice conversion – text-dependent and text-independent, intra-lingual and cross-language – of different research groups. The second TC-Star evaluation corpus (cf. Section A.2.4) consists of four voices, two female and two male which are of bilingual (English and Spanish) speakers. They uttered about 160 British English and 200 Spanish phrases (about 800s of speech). From this corpus, 10 utterances were selected for testing, the remaining data served for training. As in Section 5.2.2, from the possible twelve source/target speaker combinations, each gender combination was selected once.

Taking into account that English was to be target language, the bilinguality of all involved voices allows for the following training paradigms (cf. Section 1.1):

A. text-dependent intra-lingual – only English speech data is used for training

B. text-independent intra-lingual – only English speech data is used for training

C. text-dependent cross-language – the training is exclusively based on the Spanish speech data of the involved speakers, but the conversion is applied to the English speech of the respective speakers (cf. the example discussed in Section 2.2)

D. text-independent cross-language – the training is based on the English speech of the source speaker and the Spanish speech of the target speaker

The author submitted two systems, one based on paradigm **A**, the other based on **D**. In total, 9 systems participated in the evaluation.

To achieve the highest possible speech quality, the following changes were applied to the baseline system of the first evaluation campaign discussed in Section 5.2.2:

- **Pitch tracking.** Correct and consistent pitch marks are crucial for a good acoustic synthesis based on time domain PSOLA. Also in training, pitch mark errors can lead to a poor estimation of the conversion parameters. Since recent studies on the reliability of pitch mark determination algorithms [Höge & Kotnik$^+$ 06, Kotnik & Höge$^+$ 06] showed that the pitch tracker used so far (Goncharoff/Gries) features a rather poor performance, it was replaced by the Praat software [Boersma 01]. Additional Laryngograph data was used to achieve a high pitch marking accuracy. After automatically determining the pitch marks, only those training speech files were selected for the parameter training whose pitch marks could be reliably determined. This decision was made based on informal listening tests on the PSOLA-modified speech samples. For the test speech files where correct pitch marks are much more important some parts were manually corrected.

- **Voicing information.** According to the discussion in Section 3.1, only voiced speech portions were transformed, whereas unvoiced portions were generally copied to the target, or, in a few cases, slightly adapted by means of VTLN-based conversion.

- **Feature dimensionality.** According to the considerations in Section 2.4.2, due to the low dimensionality of the processed feature vectors, spectral details are neglected by the features which have to be reintroduced by residual prediction. Since the latter decreases speech quality, it is preferable to use a higher dimensionality of the feature vector. On the other hand, one faces data sparseness leading to parameter estimation problems when using too large dimensionalities. For the relatively large speech corpus available in the second TC-Star evaluation campaign, the dimensionality could be increased to 32 without perceptively affecting speech quality.

- **Voices.** Finally, it is to be mentioned that the voices of the second TC-Star evaluation corpus are of professional speakers as opposed to those of the first TC-Star evaluation corpus. This fact can be of importance as experiences in speech synthesis research show [Black & Lenzo 03]. Professional speakers' voices are more consistent and clear and tend to behave better when being automatically processed e.g. by means of a pitch tracker.

conversion type	text-dependent	text-independent
	intra-lingual	cross-language
source/target language	English/English	Spanish/English
alignment technique	dynamic time warping	unit selection
		$w = 0.3$
corpus	second TC-Star evaluation corpus	
voices	2 female, 2 male (bilingual professionals)	
amount of training data	$\approx$ 400s per speaker	
amount of test data	$\approx$ 50s per speaker	
number of test subjects	14	
sampling rate	$f_\mathrm{s} = 16\mathrm{kHz}$	
norm frequency	$f_\mathrm{n} = 100\mathrm{Hz}$	
pitch mark extraction		
- of training data	Praat supported by Laryngograph signal	
- of test data	ditto; partially manually corrected	
conversion technique	linear transformation	
features	32^{nd} order line spectral frequencies	
number of mixtures	4	
covariance matrix	diagonal	
residual prediction	VTLN	
acoustic synthesis	time domain PSOLA	

Table 5.5: Text-independent cross-language voice conversion: system properties.

	source	target	text-dependent intra-lingual	text-independent cross-language
quality	4.8	4.8	3.1	3.4
similarity	1.6		2.4	2.0

Table 5.6: Text-independent cross-language voice conversion: MOS tests on speech quality and similarity. The table also contains the results for clean source and target speech (and the comparison of them in the case of similarity).

In the comparison of text-dependent with text-independent voice conversion in Section 5.2, MOSs for both speech quality and similarity were used. The same metrics were applied to the second TC-Star evaluation campaign whose results were derived based on the opinion of 14 non-speech-professional subjects. The system properties for both submitted systems are given in Table 5.5 and the evaluation results in Table 5.6.

To simplify interpreting these outcomes, in Figure 5.5, the performance of all competing systems is displayed as points in a coordinate system on the MOSs of quality and similarity. In addition to the two submissions of the author (**intra-lingual** and **cross-language**), the following points are denoted to serve as standards of comparison:

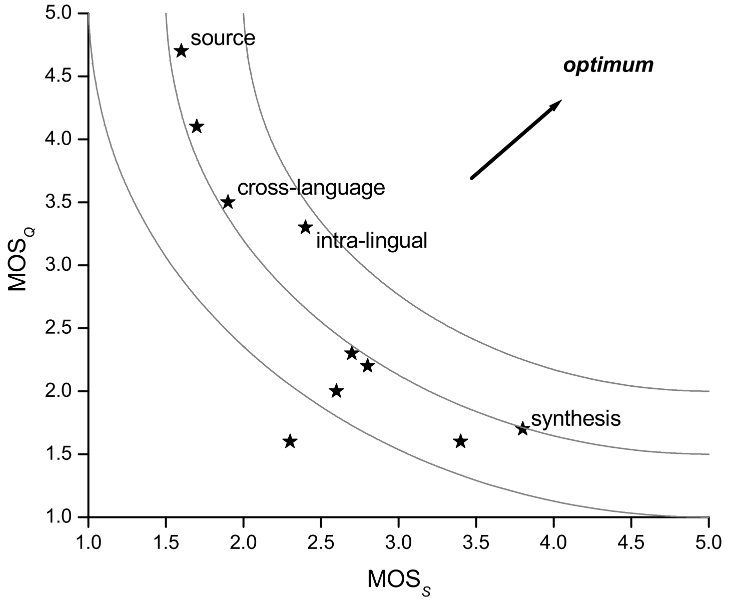

Figure 5.5: Results of the second TC-Star evaluation campaign. MOS$_Q$ and MOS$_S$ are the MOSs for speech quality and similarity, respectively. The gray lines are the set of points with distance $d = \{3, 3.5, 4\}$ from the theoretical optimum MOS$_Q$ = MOS$_S$ = 5. Points without designation belong to other competing systems whose details are disclosed in [Bonafonte & Banchs$^+$ 06].

- **Source.** This is the source speech that naturally achieved the highest speech quality but, at the same time, the lowest similarity to the target.
- **Synthesis.** The speech processing group at the Technical University of Catalonia built a text-to-speech synthesis system exclusively based on the second TC-Star evaluation corpus. As the amount of available speech data was very limited compared to conventional speech synthesis corpora that usually incorporate several hours of data, the achieved sound quality was relatively poor. On the other hand, it did not convert source speech to sound like the target but directly took speech segments (units) from the given target speech and concatenated them. Therefore, the similarity score was very high.
- ***Optimum.*** This is the region where an optimal voice conversion system is located.

5.3.2 Conclusion

The speech quality of both submissions was between good and fair fulfilling the goal of the TC-Star project formulated in [Bonafonte & Höge$^+$ 05a] ($MOS_Q \geq 3.0$). Interestingly, as opposed to the studies in Section 5.2, the text-independent cross-language voice conversion achieved a higher speech quality than the text-dependent intra-lingual conversion. In terms of voice similarity, it was the other way around.

Both effects might be attributed to the nature of the text-independent training method: The cost minimization described in Equation 5.8 encourages low target costs, i.e. low distances between source and corresponding target vector. The more training data is available, the smaller these distances become (the experiments of Section 5.2 were based on about 80s of speech, whereas in the current case, it was around 400s). For an infinite amount of training data, the author expected the distances to tend to zero or a relatively small limit. However, the more similar corresponding source and target vectors are, the less speaker-dependent information can be trained from them. For the limit case where source and target vectors are equivalent one gets zero vectors and identity matrices as parameters of the linear transformation. In this case, the converted feature vectors are equivalent to the source vectors, i.e., one produces the source speech as output featuring the trivial baseline similarity but a high speech quality.

The described effect contradicts common experience in machine learning where the availability of more training material usually allows for the extraction of more knowledge, i.e., in the aforementioned case, one would expect that more speaker-specific information would be trained. In the following section, this effect which the author refers to as *speech alignment paradox* is investigated.

Returning to the issue of robustness (item **4** of the list of criteria given at the very beginning of this chapter), the outcomes of the experiments described in this section show that both the text-dependent intra-lingual and the text-independent cross-language voice conversion achieved similar results on the task investigated in the second TC-Star evaluation campaign. The frequently asked question, if the different phoneme sets of source and target language lead to parameter estimation problems, speech with audible accent, or signal distortions, found a satisfying answer also in experiments on other languages as for instance Medieval German or Chinese which the author conducted in the course of other projects.

Obviously, the search for similar feature vectors as performed by Equation 5.8 can also be successful even when the source vector comes from a phoneme that does not exist in the target language. The algorithm straightforwardly selects the frame best resembling the source frame, independent of the phoneme it comes from. The more data is available, the more successful is this search, since the numerous phoneme transitions in the speech produce a large variety of sounds not available in the standard phoneme set.

It has also to be taken into account that the linear transformation paradigm described in Section 2.4.2 is applied to source vectors that come from canonical source speech. This means, the baseline is not accented or distorted, and this also applies to the residuals, if they are predicted by means of the VTLN technique. Considering the small number of Gaussian mixtures (4 in this chapter's experiments) suggests that the linear transformation is not completely phoneme-dependent but rather represents the speech characteristics by four global classes. Hence, the transformation does not affect single phonemes but converts the voice characteristics in a more global way avoiding the aforementioned effects.

5.4 The Speech Alignment Paradox

In the previous section, it was found that when applying unit-selection-based speech alignment to the training of a voice conversion system one faces the following effect:

> The more training data is available, the less speaker-specific information is trained.

The author expected that in the limit case, i.e., when infinitely much data was available source and aligned target speech would be identical such that the training would be trivial. In this section, this effect, referred to as *speech alignment paradox*, is to be investigated by experimental and mathematical means.

conversion type	text-independent
	intra-lingual
source/target language	English/English
alignment technique	unit selection
	$w \in \{0, 0.1, \ldots, 1.0\}$
corpus	second TC-Star evaluation corpus
voices	2 female, 2 male
amount of training data	1, 2, 4, 8, 16, 32, 64, or 128 utterances per speaker
sampling rate	$f_\mathrm{s} = 16\mathrm{kHz}$
pitch mark extraction	Praat

Table 5.7: Speech alignment paradox: system properties.

5.4.1 Experiments

The expectation that for an increasing amount of training data the aligned target speech would more and more resemble the source can be verified by an experiment determining the similarity of both speeches as a function of the available amount of data. The similarity of the two parallel feature vector sequences x_1^M (source speech) and $\tilde{y}_1^M$ (the aligned target speech selected from the target training data y_1^N) can be objectively expressed by their distance, or more precisely, by the mean squared Euclidean distance of their members

$$D = \frac{1}{M} \sum_{m=1}^{M} |\tilde{y}_m - x_m|^2. \tag{5.9}$$

It is to be investigated if this function approaches zero (maximum similarity) when the amount of available target training data, represented by the number of vectors in the target data, N, or by the corresponding duration t, tends to infinity.

Due to the time behavior of the unit selection algorithm addressed in Section 5.2.1, it is not possible to perform this experiment for very large amounts of data. Therefore, the considered data was limited to around 600s of speech taken from the second TC-Star evaluation corpus. As in Section 5.2.2, each gender combination was considered. The experiments were carried out taking 1, 2, 4, 8, 16, 32, 64, or 128 English training utterances of the respective speakers. In addition to the amount of training data, also the trade-off parameter of Equation 5.8 was varied $w \in \{0, 0.1, \ldots, 1.0\}$. The experiment's properties are shown in Table 5.7.

Independent of the voice combinations used, the outcomes were very similar. As an example, the results of the female-male voice combination are displayed in Figure 5.6 in double logarithmic representation. Independent of the trade-off parameter w, the values of D almost constantly decrease[38]. To simplify matters, in the following, the special case $w = 1$ is to be investigated, the respective diagram is shown in Figure 5.7.

For the considered amounts of data, the test samples are almost located on a straight line in double logarithmic representation. Consequently, the relation between D and t can be approximated by

$$\log D = c - b \log t \quad \text{with} \quad b > 0. \tag{5.10}$$

[38] Except for $w = 0$ which does not lead to a useful alignment, since no target costs are considered.

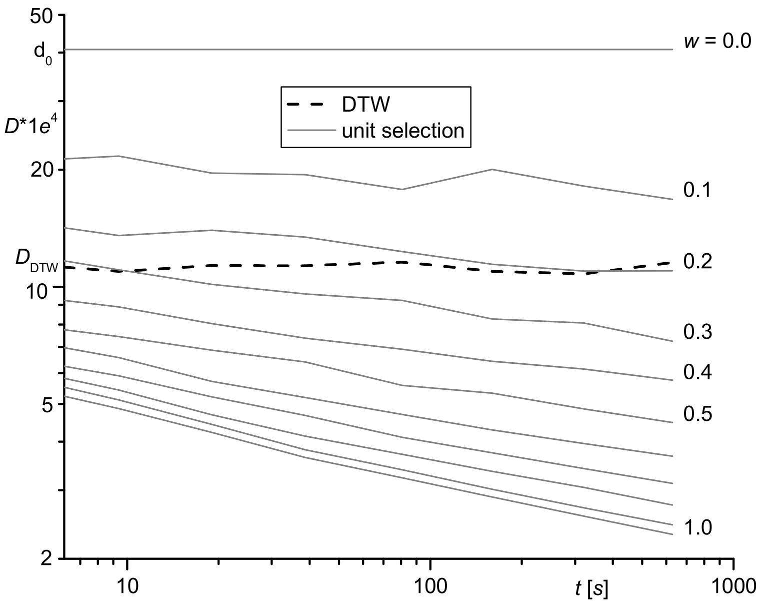

Figure 5.6: Speech alignment paradox: mean distance between corresponding source and target feature vectors, D, depending on the amount of data, t, and the trade-off parameter w. The function D_{DTW} indicates the mean distance produced by text-dependent speech alignment based on dynamic time warping.

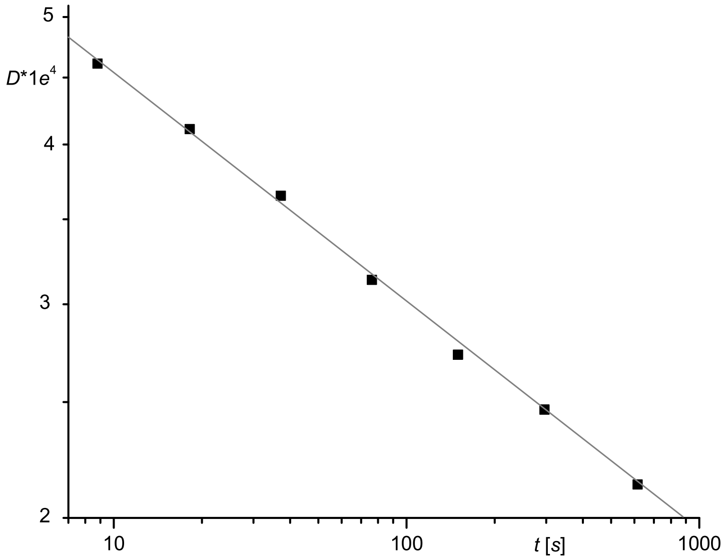

Figure 5.7: Speech alignment paradox: mean distance between corresponding source and target feature vectors, D, depending on the amount of data, t for the special case $w = 1$.

D	t	disk space
5	5.6 s	174 kB
2	900 s = 15 min	27 MB
1	$4.2 \cdot 10^4$ s = 11.7 h	1.3 GB
0.5	$2.0 \cdot 10^6$ s = 22.8 d	59 GB
0.2	$3.2 \cdot 10^8$ s = 10.3 a	9.2 TB

Table 5.8: Speech alignment paradox: required amount of data (t) for certain degrees of similarity (D) and the corresponding hard disk space necessary for storing a 16kHz/16bit PCM version of the data.

Exponentiation yields

$$D = e^{c-b\log t} = e^c e^{\log t^{-b}} = at^{-b} \quad \text{with} \quad a, b > 0. \tag{5.11}$$

If validity of Equation 5.11 is assumed also for amounts of data beyond the experiment's scope, one gets the limit

$$\lim_{t\to\infty} D = \lim_{t\to\infty} at^{-b} = 0. \tag{5.12}$$

This means, for very large amounts of data, the aligned speech samples become very similar to each other (for the limit case even identical) providing evidence for the speech alignment paradox.

Due to the computational complexity of the involved unit selection, it is not possible to massively increase the amount of data. This is the main reason for describing the speech alignment paradox by mathematical means as done in the next section.

5.4.2 On a Proof of the Speech Alignment Paradox

Although the empirical investigations of Section 5.4.1 were confirmed by several experimental cycles, doubts arose on the validity of the limit value shown in Equation 5.12, since it could be interpreted as follows:

> If enough speech data is available, an arbitrary utterance of an arbitrary voice can be produced only by selecting and concatenating units from this data.

However, the crucial point in the statement is the word *enough*. Applying the parameters $a = 6.8$ and $b = 0.18$ determined on the data of Figure 5.7 to Equation 5.11, the required amounts of data for several degrees of similarity were estimated, cf. Table 5.8. The amount of necessary data extremely grows when the mean distance between source and aligned target feature vectors becomes smaller and soon exceeds the limits of the technical possible. Nonetheless, since the validity of the statement phrased above could be of interest to the speech processing community, e.g. for the design of speech synthesis systems based on large databases of pooled speakers as proposed in [Eide & Picheny 06], in the following, the alignment technique's behavior for very large amounts of data is investigated by mathematical means.

Applying the Gaussian Mixture Model

As discussed in Section 2.4.2 speech which in the case of unit-selection-based alignment is given by sequences of feature vectors is very often described by means of the Gaussian mixture model – examples include speech recognition, language identification, speaker recognition, speaking rate estimation, and gender classification. The success of the Gaussian mixture model in these speech processing fields also suggests its application to the investigation of the speech alignment paradox.

In order to keep things manageable, the degrees of freedom are reduced as follows:

- The dimensionality of the feature vectors is reduced to $D = 1$ (w.l.o.g.).
- Pooled covariance matrices identical for source and target are assumed for the feature vector sequences to be aligned, i.e., for $D = 1$, the standard deviation σ is used.

The A-Priori Alignment

When having a look at the speech samples without performing any alignment one can determine an a-priori value for the mean vector distance D: the expected value $E_1(D)$[39]. The latter is the expected distance between the two random vectors x and y that are distributed according to a Gaussian mixture model, i.e. a sum of I normal distributions weighted by $\alpha_x^{(1)}, \ldots, \alpha_x^{(I)}$ for the source voice and J distributions weighted by $\alpha_y^{(1)}, \ldots, \alpha_y^{(J)}$ for the target voice, respectively (cf. Equation 2.8),

$$E_1(D) = \int\limits_{-\infty}^{\infty} E_1(D|x) \sum_{i=1}^{I} \left[\alpha_x^{(i)} \mathcal{N}(x|\mu_x^{(i)}, \sigma)\right] \mathrm{d}x \tag{5.13}$$

where $E_1(D|x)$ is the expected value of D, if x is fixed, and $\mathcal{N}(x|\mu_x^{(i)}, \sigma)$ is the probability density function of a normal distribution given by Equation 4.3. In the following, the *standard* normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{5.14}$$

is used, and Equation 5.13 is modified accordingly:

$$E_1(D) = \frac{1}{\sigma} \sum_{i=1}^{I} \alpha_x^{(i)} \int\limits_{-\infty}^{\infty} E_1(D|x)\, f\left(\frac{x - \mu_x^{(i)}}{\sigma}\right) \mathrm{d}x. \tag{5.15}$$

[39] The subindex 1 is due to the fact that this expected value is a special case of $E_N(D)$ described later in this section.

Now, the expected value of D, if x is fixed, is calculated:

$$\begin{aligned}
E_1(D|x) &= \frac{1}{\sigma}\int\limits_{-\infty}^{\infty}|x-y|\sum_{j=1}^{J}\left[\alpha_y^{(j)} f\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)\right]\mathrm{d}y \\
&= \frac{1}{\sigma}\sum_{j=1}^{J}\alpha_y^{(j)}\left[\int\limits_{-\infty}^{x}(x-y)\,f\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)\mathrm{d}y-\frac{1}{\sigma}\int\limits_{x}^{\infty}(x-y)\,f\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)\mathrm{d}y\right] \\
&= \sum_{j=1}^{J}\alpha_y^{(j)}\left[\sigma\,f\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)+(x-\mu_y^{(j)})\,\Phi\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)\right]_{y=-\infty}^{x} \\
&\quad-\sum_{j=1}^{J}\alpha_y^{(j)}\left[\sigma\,f\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)+(x-\mu_y^{(j)})\,\Phi\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)\right]_{y=x}^{\infty} \\
&= \sum_{j=1}^{J}\alpha_y^{(j)}\left[2\sigma\,f\left(\frac{x-\mu_y^{(j)}}{\sigma}\right)+(x-\mu_y^{(j)})\left(2\Phi\left(\frac{x-\mu_y^{(j)}}{\sigma}\right)-1\right)\right] \qquad (5.16)
\end{aligned}$$

where $\Phi(x)$ is the standard normal cumulative density function thus it holds $\frac{\mathrm{d}\Phi(x)}{\mathrm{d}x}=f(x)$. By inserting the result into Equation 5.15, one gets

$$\begin{aligned}
E_1(D) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\alpha_x^{(i)}\alpha_y^{(j)}\left[2\int\limits_{-\infty}^{\infty}f\left(\frac{x-\mu_y^{(j)}}{\sigma}\right)f\left(\frac{x-\mu_x^{(i)}}{\sigma}\right)\mathrm{d}x\right. \\
&\quad+2\int\limits_{-\infty}^{\infty}\frac{x-\mu_y^{(j)}}{\sigma}\,\Phi\left(\frac{x-\mu_y^{(j)}}{\sigma}\right)f\left(\frac{x-\mu_x^{(i)}}{\sigma}\right)\mathrm{d}x \\
&\quad\left.-\int\limits_{-\infty}^{\infty}\frac{x-\mu_y^{(j)}}{\sigma}\,f\left(\frac{x-\mu_x^{(i)}}{\sigma}\right)\mathrm{d}x\right] \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\alpha_x^{(i)}\alpha_y^{(j)}\left[T_1^{(i,j)}+T_2^{(i,j)}+T_3^{(i,j)}\right]. \qquad (5.17)
\end{aligned}$$

For $T_1^{(i,j)}$ and $T_3^{(i,j)}$, there are straightforward solutions:

$$\begin{aligned}
T_1^{(i,j)} &= \frac{1}{\pi}\int\limits_{-\infty}^{\infty}e^{-\frac{\left(x-\mu_y^{(j)}\right)^2+\left(x-\mu_x^{(i)}\right)^2}{2\sigma^2}}\mathrm{d}x \\
&= \frac{1}{\pi}\int\limits_{-\infty}^{\infty}e^{-\frac{\left(2x-\mu_x^{(i)}-\mu_y^{(j)}\right)^2+\left(\mu_y^{(j)}-\mu_x^{(i)}\right)^2}{4\sigma^2}}\mathrm{d}x \\
&= 2\,f\left(\frac{\mu_y^{(j)}-\mu_x^{(i)}}{\sqrt{2}\sigma}\right)\int\limits_{-\infty}^{\infty}f\left(\frac{2x-\mu_x^{(i)}-\mu_y^{(j)}}{\sqrt{2}\sigma}\right)\mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&= 2\,f\biggl(\frac{\mu_y^{(j)}-\mu_x^{(i)}}{\sqrt{2}\sigma}\biggr)\left[\frac{\sigma}{\sqrt{2}}\,\Phi\biggl(\frac{2x-\mu_x^{(i)}-\mu_y^{(j)}}{\sqrt{2}\sigma}\biggr)\right]_{x=-\infty}^{\infty}\\
&= \sqrt{2}\sigma\,f\biggl(\frac{\mu_y^{(j)}-\mu_x^{(i)}}{\sqrt{2}\sigma}\biggr); \qquad (5.18)
\end{aligned}
$$

$$
\begin{aligned}
T_3^{(i,j)} &= \left[(\mu_x^{(i)}-\mu_y^{(j)})\,\Phi\biggl(\frac{x-\mu_x^{(i)}}{\sigma}\biggr)-\sigma\,f\biggl(\frac{x-\mu_x^{(i)}}{\sigma}\biggr)\right]_{x=-\infty}^{\infty}\\
&= \mu_y^{(j)}-\mu_x^{(i)}, \qquad (5.19)
\end{aligned}
$$

whereas $T_2^{(i,j)}$ requires a more complex derivation given in Section A.3.4:

$$
T_2^{(i,j)} = \sqrt{2}\sigma f\biggl(\frac{\mu_y^{(j)}-\mu_x^{(i)}}{\sqrt{2}\sigma}\biggr)+2(\mu_y^{(j)}-\mu_x^{(i)})\left[\Phi\biggl(\frac{\mu_y^{(j)}-\mu_x^{(i)}}{\sqrt{2}\sigma}\biggr)-1\right] \qquad (5.20)
$$

yielding the requested expected value of D (in the following, the variable $\delta^{(i,j)} = \mu_y^{(j)} - \mu_x^{(i)}$, the difference between the means of the i^{th} source density and the j^{th} target density, is used):

$$
E_1(D) = \sum_{i=1}^{I}\sum_{j=1}^{J}\alpha_x^{(i)}\alpha_y^{(j)}\left[2\sqrt{2}\sigma\,f\biggl(\frac{\delta^{(i,j)}}{\sqrt{2}\sigma}\biggr)+2\delta^{(i,j)}\,\Phi\biggl(\frac{\delta^{(i,j)}}{\sqrt{2}\sigma}\biggr)-\delta^{(i,j)}\right]. \qquad (5.21)
$$

For some of the following considerations, the special case $I = J = 1$ is to be considered where the quantity $\delta = \delta^{(1,1)}$ is used and Equation 5.21 becomes

$$
E_1(D) = 2\sqrt{2}\sigma\,f\biggl(\frac{\delta}{\sqrt{2}\sigma}\biggr)+2\delta\,\Phi\biggl(\frac{\delta}{\sqrt{2}\sigma}\biggr)-\delta. \qquad (5.22)
$$

Figure 5.8 shows $E_1(D)$ as a function of δ for $\sigma \in \{0.5, 1, 2\}$ and indicates the lower bound of $E_1(D)$ given by the limit

$$
\lim_{\delta/\sigma\to\pm\infty} E_1(D) = |\delta|. \qquad (5.23)
$$

This limit can also be calculated using Equation 5.15: When $\frac{\delta}{\sigma}$ approaches infinity the deviation of x's distribution function becomes infinitely small as compared with its mean's distance to y's mean. Consequently, the normal distribution can be replaced by the Dirac delta function Δ yielding

$$
\begin{aligned}
\lim_{\delta/\sigma\to\pm\infty} E_1(D) &= \lim_{\delta/\sigma\to\pm\infty}\frac{1}{\sigma}\int_{-\infty}^{\infty} E_1(D|x)\,f\biggl(\frac{x-\mu_x}{\sigma}\biggr)\,\mathrm{d}x\\
&= \lim_{\delta/\sigma\to\pm\infty}\frac{1}{\sigma}\int_{-\infty}^{\infty} E_1(D|x)\,\Delta\biggl(\frac{x-\mu_x}{\sigma}\biggr)\,\mathrm{d}x\\
&= \lim_{\delta/\sigma\to\pm\infty}\frac{1}{\sigma}\int_{-\infty}^{\infty} E_1(D|\sigma\xi+\mu_x)\,\Delta(\xi)\sigma\mathrm{d}\xi\\
&= \lim_{\delta/\sigma\to\pm\infty} E_1(D|\mu_x) \qquad (5.24)
\end{aligned}
$$

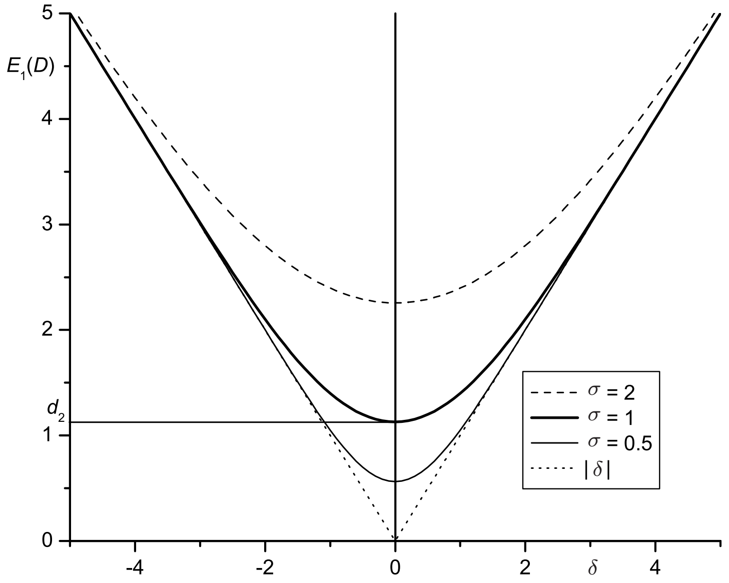

Figure 5.8: Speech alignment paradox: expected value of the mean distance between two feature vectors $E_1(D)$ as a function of the difference of the distribution means δ and of the standard deviation σ.

Applying Equation 5.16 yields

$$\lim_{\delta/\sigma\to\pm\infty} E_1(D) = \lim_{\delta/\sigma\to\pm\infty} 2\sigma f\left(\frac{\delta}{\sigma}\right) + \delta\left[\Phi\left(\frac{\delta}{\sigma}\right) - 1\right] = |\delta|. \tag{5.25}$$

Incidentally, the author found that for the special case $\delta = 0$ and $\sigma = 1$, Equation 5.21 becomes the closed form solution of what in statistical process control is referred to as the constant d_2 (the expected distance between two instances of a standard normally distributed random process) whose value 1.1284 determined by numerical means is given in textbooks on process control, e.g. [Montgomery 96]:

$$d_2 = \frac{2}{\sqrt{\pi}}. \tag{5.26}$$

Towards a Closed Form Solution

After studying the a-priori alignment, the above approach is to be extended to the unit selection-based alignment considering several mixture densities. The principal difference to the former is the minimization in Equation 5.8 to search for the optimal sequence of vectors. This directly affects the expected value of D given x (cf. Equation 5.16), since now the probability density function of y is not a sum over normal distributions but over the more complicated terms $p_N(y|x,j)$:

$$E_N(D|x) = \int_{-\infty}^{\infty} |x-y| \sum_{j=1}^{J} \left[\alpha_y^{(j)} p_N(y|x,j)\right] \mathrm{d}y. \tag{5.27}$$

As mentioned earlier, N denotes the number of feature vectors in the target feature vector sequence y_1^N serving as a pool appropriate units are selected from, see Section 5.2. Again, it is assumed that these vectors are distributed according to a Gaussian mixture model with J mixtures and the parameters $\alpha_y^{(j)}$, $\mu_y^{(j)}$ and σ are independent of each other.

Given a mixture j, for each possible y, the probability density of the n^{th} target feature vector being equal to y and closest to x is calculated. The sum over all these vectors from 1 to N yields the sought-after density $p_N(y|x,j)$.

In detail: The probability density of the n^{th} vector being equal to y is

$$P_n^{(j)} = \frac{1}{\sigma} f\left(\frac{y-\mu_y^{(j)}}{\sigma}\right). \tag{5.28}$$

The probability of the n^{th} vector being closest to x means that the distance to all other vectors y_ν for $\nu \in \{1,\ldots,N\}$, $\nu \neq n$ is greater than that to y_n or, given $y_n = y$, that $|y_\nu - x| > |y - x|$:

$$\begin{aligned}
Q_n^{(j)} &= p\Big(\bigwedge_{\substack{\nu=1\\ \nu\neq n}}^{N} |y_\nu - x| > |y-x|\Big)\\
&= \prod_{\substack{\nu=1\\ \nu\neq n}}^{N} p(|y_\nu - x| > |y-x|)\\
&= p(|\psi - x| > |y-x|)^{N-1}\\
&= \begin{cases} p(\psi < y \;\vee\; \psi > 2x-y)^{N-1} & \text{for } y < x\\ p(\psi > y \;\vee\; \psi < 2x-y)^{N-1} & \text{otherwise}\end{cases}\\
&= \begin{cases} \left(\Phi\left(\frac{y-\mu_y^{(j)}}{\sigma}\right) + 1 - \Phi\left(\frac{2x-y-\mu_y^{(j)}}{\sigma}\right)\right)^{N-1} & \text{for } y < x\\ \left(1 - \Phi\left(\frac{y-\mu_y^{(j)}}{\sigma}\right) + \Phi\left(\frac{2x-y-\mu_y^{(j)}}{\sigma}\right)\right)^{N-1} & \text{otherwise}\end{cases}
\end{aligned}$$

Here, ψ is a y-like distributed random variable replacing y_ν for $\nu \in \{1,\ldots,N\}, \nu \neq n$. Accordingly, Equation 5.27 becomes (also cf. Equation 5.16)

$$\begin{aligned}
E_N(D|x) &= \sum_{j=1}^{J} \alpha_y^{(j)} \int_{-\infty}^{\infty} |x-y| \sum_{n=1}^{N} (P_n^{(j)} Q_n^{(j)}) \mathrm{d}y \qquad (5.29)\\
&= \frac{N}{\sigma} \sum_{j=1}^{J} \alpha_y^{(j)} \Bigg[\int_{-\infty}^{x} (x-y) f\left(\frac{y-\mu_y^{(j)}}{\sigma}\right) \left(1 + \Phi\left(\frac{y-\mu_y^{(j)}}{\sigma}\right) - \Phi\left(\frac{2x-y-\mu_y^{(j)}}{\sigma}\right)\right)^{N-1} \mathrm{d}y\\
&\quad - \int_{x}^{\infty} (x-y) f\left(\frac{y-\mu_y^{(j)}}{\sigma}\right) \left(1 - \Phi\left(\frac{y-\mu_y^{(j)}}{\sigma}\right) + \Phi\left(\frac{2x-y-\mu_y^{(j)}}{\sigma}\right)\right)^{N-1} \mathrm{d}y \Bigg].
\end{aligned}$$

As already mentioned in Footnote 39, for $N = 1$, this becomes identical to Equation 5.16. One can show that Equation 5.29 can be simplified to the problem of solving

$$\int f^m(x)\,\Phi^n(x-\delta)\,\mathrm{d}x \quad \text{for} \quad m \in \{1,2\} \text{ and } n \in \{0,1,2,\ldots\} \tag{5.30}$$

whose closed-form solution does not seem to exist for $n > 1$.
However, if large values of $\frac{\delta}{\sigma}$ are considered (cf. Equation 5.23), for the special case $I = J = 1$, one gets a very exact approximation of the sought-after expected value of D as already discussed in Equation 5.24 by

$$\lim_{\delta/\sigma\to\pm\infty} E_N(D) = \lim_{\delta/\sigma\to\pm\infty} E_N(D|\mu_x). \tag{5.31}$$

Considering also Equation 5.29 where y is substituted by $z + \mu_x$ yields

$$\begin{aligned}\lim_{\delta/\sigma\to\pm\infty} E_N(D) &= \lim_{\delta/\sigma\to\pm\infty} \frac{N}{\sigma}\left[-\int_{-\infty}^{0} z\, f\!\left(\frac{z-\delta}{\sigma}\right)\left(1+\Phi\!\left(\frac{z-\delta}{\sigma}\right)-\Phi\!\left(\frac{-z-\delta}{\sigma}\right)\right)^{N-1}\mathrm{d}z\right.\\ &\quad\left.+\int_{0}^{\infty} z\, f\!\left(\frac{z-\delta}{\sigma}\right)\left(1-\Phi\!\left(\frac{z-\delta}{\sigma}\right)+\Phi\!\left(\frac{-z-\delta}{\sigma}\right)\right)^{N-1}\mathrm{d}z\right].\end{aligned} \tag{5.32}$$

For $\frac{\delta}{\sigma} \to \infty$, the first integral of this equation becomes zero, since

$$\lim_{\delta/\sigma\to\infty} f\left(\frac{z-\delta}{\sigma}\right) = 0 \quad \text{for} \quad z < 0. \tag{5.33}$$

Moreover, for the remaining integral holds

$$\lim_{\delta/\sigma\to\infty} \Phi\left(\frac{-z-\delta}{\sigma}\right) = 0 \quad \text{for} \quad z > 0, \tag{5.34}$$

and, taking into account Equation 5.33, the remaining integral's lower limit can be extended to $-\infty$, since this adds zero. Consequently, for $\frac{\delta}{\sigma} \to \infty$, Equation 5.32 can be expressed by

$$\lim_{\delta/\sigma\to\infty} E_N(D) = \lim_{\delta/\sigma\to\infty} \frac{N}{\sigma}\int_{-\infty}^{\infty} z\, f\left(\frac{z-\delta}{\sigma}\right)\left(1-\Phi\left(\frac{z-\delta}{\sigma}\right)\right)^{N-1}\mathrm{d}z. \tag{5.35}$$

Applying the above steps to the case $\frac{\delta}{\sigma} \to -\infty$ and using the relations $\Phi(x) = 1 - \Phi(-x)$ and $f(x) = f(-x)$ yields

$$\lim_{\delta/\sigma\to\pm\infty} E_N(D) = \lim_{\delta/\sigma\to\pm\infty} \frac{N}{\sigma}\int_{-\infty}^{\infty} z\, f\left(\frac{z-|\delta|}{\sigma}\right)\left(1-\Phi\left(\frac{z-|\delta|}{\sigma}\right)\right)^{N-1}\mathrm{d}z. \tag{5.36}$$

N	$\mu(N)$	closed form of $\mu(N)$
1	0	0
2	0.56	$\frac{1}{\sqrt{\pi}}$
3	0.85	$\frac{3}{2\sqrt{\pi}}$
4	1.03	$\frac{6}{\sqrt{\pi^3}} \arctan\sqrt{2}$
5	1.16	$\frac{15}{\sqrt{\pi^3}} \arctan\sqrt{2} - \frac{5}{2\sqrt{\pi}}$
10	1.54	
100	2.51	
1 000	3.24	
1 000 000	4.86	

Table 5.9: Speech alignment paradox: the offset constant $\mu(N)$ for different values of N.

Substituting z by $|\delta| - \sigma\xi$ and taking into account that the quotient $\frac{\delta}{\sigma}$ disappears which allows for removing the limit leads to

$$\begin{aligned}
\lim_{\delta/\sigma\to\pm\infty} E_N(D) &= \frac{N}{\sigma}\int_{\infty}^{-\infty}(|\delta| - \sigma\xi) f(-\xi)\left(1-\Phi(-\xi)\right)^{N-1}(-\sigma \mathrm{d}\xi) \\
&= N\int_{-\infty}^{\infty}(|\delta| - \sigma\xi) f(\xi)\Phi(\xi)^{N-1}\mathrm{d}\xi \\
&= N|\delta|\left[\frac{\Phi(\xi)^N}{N}\right]_{\xi=-\infty}^{\infty} - \sigma N\int_{-\infty}^{\infty}\xi f(\xi)\Phi(\xi)^{N-1}\mathrm{d}\xi \\
&= |\delta| - \sigma\mu(N). \qquad (5.37)
\end{aligned}$$

The structure of this formula gives a qualitative overview about some of the expected value's characteristics. It consists of the term $|\delta|$ being independent of the standard deviation σ and a term which is a constant with respect to $|\delta|$ but linearly depends on σ.

[David & Nagaraja 03] give solutions to $\mu(N)$ for $N \in \{1, ..., 5\}$, but so far, for $k > 5$, the author did not succeed in finding a closed form. Table 5.9 gives some example values of $\mu(N)$, and Figure 5.9 shows a plot of $E_N(D)$ as a function of δ for several values of N. For large δ, the graphs approach $|\delta| - \sigma\mu(N)$ as derived in Equation 5.37, and for $N = 1$, one obtains the special case discussed in Equations 5.23 and 5.25.

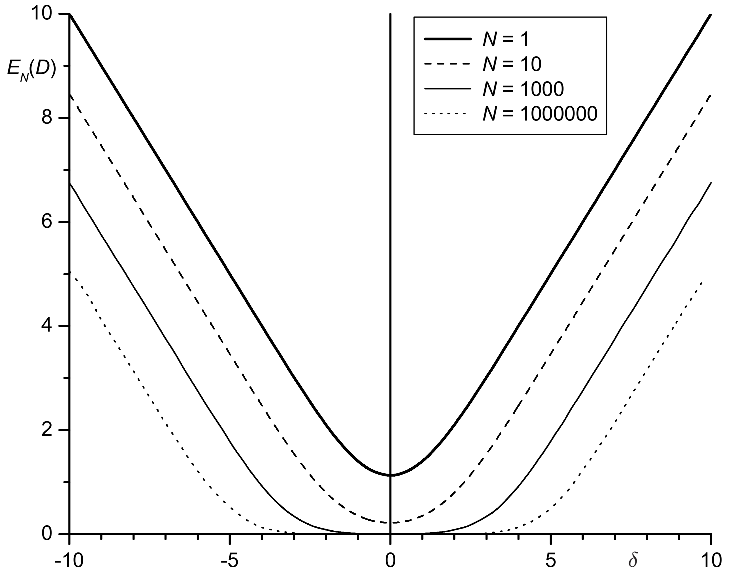

Figure 5.9: Speech alignment paradox: expected value of the minimum distance between a source feature and N target feature vectors $E_N(D)$ as a function of the difference of the distribution means δ and the number of available target feature vectors N; $\sigma = 1$.

Although there is no closed form solution to $\mu(N)$ for an arbitrary N, one can derive partial results exploiting the symmetries of f and Φ:

$$\begin{aligned}
\mu(N) &= N \int_{-\infty}^{\infty} x f(x) \Phi(x)^{N-1} \mathrm{d}x \\
&= -N \int_{-\infty}^{\infty} \xi f(\xi)(1-\Phi(\xi))^{N-1} \mathrm{d}\xi \\
&= -N \sum_{k=0}^{N-1} (-1)^k \binom{N-1}{k} \int_{-\infty}^{\infty} \xi f(\xi) \Phi(\xi)^k \mathrm{d}\xi \\
&= -\sum_{k=0}^{N-1} (-1)^k \binom{N}{k+1} \mu(k+1) \\
&= \sum_{\kappa=1}^{N-1} (-1)^\kappa \binom{N}{\kappa} \mu(\kappa) + (-1)^N \mu(N). \qquad (5.38)
\end{aligned}$$

This finally yields

$$\mu(N) = \frac{1}{2} \sum_{k=1}^{N-1} (-1)^k \binom{N}{k} \mu(k) \quad \text{for} \quad N \in \{1, 3, \ldots\}.$$

This formula enables to recursively compute $\mu(N)$ from $\mu(1), \ldots, \mu(N-1)$. Unfortunately, it holds only for odd N, so it does not allow a general statement unless a description for even

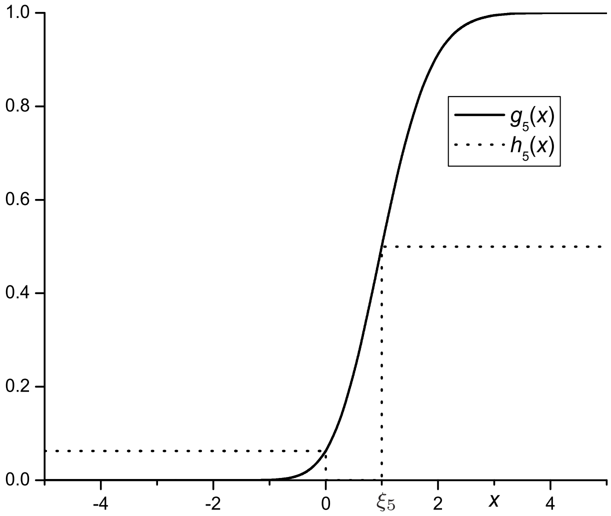

Figure 5.10: Speech alignment paradox: example of the functions $g_N(x)$ and $h_N(x)$ for $N = 5$; $\xi_5 = 0.998$, cf. Equation 5.41.

N is found. However, there is a way to study the behaviour of $\mu(N)$ when N approaches infinity:

$$\begin{aligned}\lim_{N\to\infty}\mu(N) &= \lim_{N\to\infty} N\int_{-\infty}^{\infty} xf(x)\Phi(x)^{N-1}\mathrm{d}x\\ &= \lim_{N\to\infty} N\left[\int_{-\infty}^{0} xf(x)\Phi(x)^{N-1}\mathrm{d}x + \int_{0}^{\infty} xf(x)\Phi(x)^{N-1}\mathrm{d}x\right]. \end{aligned}\qquad(5.39)$$

$f(x)$ is an even function, and x is odd, so $xf(x)$ is also odd. $\Phi(x)$ is a strictly positive, monotonous, and bounded function, hence $\int_{-\infty}^{0} xf(x)\Phi(x)^{N-1}\mathrm{d}x < 0$ and $\int_{0}^{\infty} xf(x)\Phi(x)^{N-1}\mathrm{d}x > 0$. Consequently, both terms become smaller, if $g_N(x) = \Phi(x)^{N-1}$ is replaced by a function $h_N(x)$ which is greater than the former for $x < 0$ and smaller for $x > 0$:

$$h_N(x) = \begin{cases} g_N(0) = \frac{1}{2^{N-1}} & \text{for } x \le 0\\ 0 & \text{for } 0 < x \le \xi_N\\ \frac{1}{2} & \text{for } \xi_N < x.\end{cases}\qquad(5.40)$$

Here, $\xi_N > 0$ [40] is the position where $g_N(x)$ becomes $\frac{1}{2}$:

$$\xi_N = \Phi^{-1}(2^{\frac{1}{1-N}}).\qquad(5.41)$$

Figure 5.10 shows an example of the functions $g_N(x)$ and $h_N(x)$. Applying the definition of

[40] This relation is only true for $N > 2$ which is, however, no additional constraint, since the current investigation concerns the limit for $N \to \infty$.

$h_N(x)$ to Equation 5.39 yields

$$\begin{aligned}\lim_{N\to\infty}\mu(N) &\geq \lim_{N\to\infty} N\int\limits_{-\infty}^{\infty} xf(x)h_N(x)\mathrm{d}x\\ &= \lim_{N\to\infty}\left[\frac{N}{2^{N-1}}\int\limits_{-\infty}^{0} xf(x)\mathrm{d}x+\frac{N}{2}\int\limits_{\xi_N}^{\infty} xf(x)\mathrm{d}x\right]\\ &= \lim_{N\to\infty}\left[\frac{-N}{2^{N-1}\sqrt{2\pi}}+\frac{Nf(\xi_N)}{2}\right]\\ &= \frac{1}{2}\lim_{N\to\infty} Nf(\xi_N). \end{aligned}\tag{5.42}$$

Employing Equation 5.41, N can be expressed as a function of ξ_N:

$$N=1-\frac{\log 2}{\log\Phi(\xi_N)}.\tag{5.43}$$

Equation 5.41 shows that when N approaches infinity ξ_N also tends to infinity. Consequently, Equation 5.42 can be rewritten as

$$\begin{aligned}\lim_{N\to\infty}\mu(N) &\geq \frac{1}{2}\lim_{\xi_N\to\infty} f(\xi_N)\left(1-\frac{\log 2}{\log\Phi(\xi_N)}\right)\\ &= -\frac{\log 2}{2}\lim_{\xi_N\to\infty}\frac{f(\xi_N)}{\log\Phi(\xi_N)}.\end{aligned}\tag{5.44}$$

Application of l'Hôpital's rule produces

$$\begin{aligned}\lim_{N\to\infty}\mu(N) &\geq -\frac{\log 2}{2}\lim_{\xi_N\to\infty}\frac{-\xi_N f(\xi_N)}{\frac{1}{\Phi(\xi_N)}f(\xi_N)}\\ &= \frac{\log 2}{2}\lim_{\xi_N\to\infty}\xi_N\Phi(\xi_N)\\ &= \infty.\end{aligned}\tag{5.45}$$

Hence, the limit value for $\mu(N)$ is infinity, if N approaches infinity. However, this means that for very large N, the approximation Equation 5.37 is not useful, since the expected value of D is non-negative. Consequently, when N approaches infinity one must not apply the simplifications derived in Equation 5.24, but has to consider the original definition of D's expected value (cf. Equation 5.15):

$$\lim_{N\to\infty}E_N(D)=\lim_{N\to\infty}\frac{1}{\sigma}\sum_{i=1}^{I}\alpha_x^{(i)}\int\limits_{-\infty}^{\infty}E_N(D|x)\,f\!\left(\frac{x-\mu_x^{(i)}}{\sigma}\right)\mathrm{d}x\tag{5.46}$$

where $E_N(D|x)$ is declared in Equation 5.29.

In Section A.3.5, it is shown that several substitutions and the application of Lebesgue's dominated convergence theorem considering the fact that $E_1(D)$ is finite lead to the result

$$\lim_{N\to\infty}E_N(d)=0.\tag{5.47}$$

This means, under the above given conditions and modeling the aligned speech by means of Gaussian mixture densities, for infinitely much training data, the expected distance between two aligned feature vectors becomes zero, i.e., the aligned feature vector sequences become identical such that no speaker-specific information can be trained based on this alignment.

5.4.3 Conclusion

Experiments on several speaker pairs with different amounts of training data and different settings of the weight w influencing the trade-off between target and concatenation costs for the unit-selection-based speech alignment were conducted and assessed by means of the objective measure D, the mean distance between aligned feature vectors. Experimental evidence was found that, independent of the trade-off parameter w, the measure D approaches zero, if the amount of training data tends to infinity. Due to the computational time behavior of the speech alignment based on unit selection, the experiments could not be applied to very large amounts of data. Therefore, a mathematical investigation of a special case of the speech alignment paradox was carried out showing that the expected value of D approaches zero if the number of target vectors in the training data becomes infinite.

To avoid this effect and to make sure that also for large amounts of training data speaker-specific information can be trained, the trade-off weight w should be set dependent on the amount of available data. In Figure 5.6, the mean distance between corresponding feature vectors of source and target aligned by means of the unit selection paradigm is displayed as a function of the amount of training data and w. Furthermore, the mean distance curve produced by means of text-dependent speech alignment based on dynamic time warping is given. As expected, this curve is almost constant with respect to the amount of training data, since dynamic time warping performs a search limited by strong local constraints (see Section 5.2.1) not allowing for long distance jumps and time discontinuity. The dynamic time warping curve cuts the $w = 0.3$ unit selection curve at $t \approx 10$s and the $w = 0.2$ curve at $t \approx 400$s. For an increasing amount of data, the trade-off weight w should be decreased accordingly.

6 Achievements and Conclusions

The voice conversion techniques discussed in this dissertation were assessed according to

- speech quality and
- similarity of converted voice and target.

Both of these aspects can be evaluated by means of mean opinion scores which allow for displaying a technique's performance in a two-dimensional diagram as shown in Figure 5.5. This figure shows very clearly that there is a trade-off between these two aspects. The stronger the similarity between converted voice and target is, the weaker the speech quality becomes, and vice versa. The techniques discussed in this work followed both paths and led to several remarkable outcomes:

The author applied VTLN to voice conversion and showed that

- it produces fair to good speech quality (MOS = 3.4 for female-to-male voice conversion based on time domain VTLN in Table 3.3),
- it is able to strongly change the source voice's characteristics (only 16% of the subjects recognized the source voice in the experiment based on time domain VTLN in Table 3.4; in Section 3.4.3, five well-distinguishable[41] voices were produced based on one source voice),
- signal processing can be exclusively applied in time domain (proof in Section 3.3.2),
- it is very resource-efficient making it directly applicable to embedded devices (It requires $40\bar{K}$ operations per frame of the length $\bar{K}$ and $12\bar{K}$ bytes, cf. Table 3.1; considering an average frame length of $\bar{K} = 200$ and a sampling rate of $f_s = 16\text{kHz}$, this relates to $1.28\,{}^{\text{Mops}}\!/_{\text{s}}$ and 2.4kB of memory),
- it can be used as residual prediction technique in conjunction with linear transformation producing a superior speech quality (see Section 4.3).

Furthermore, the author's investigations into residual prediction, particularly the novel techniques of residual smoothing and the application of unit selection led to a considerable gain in both

- speech quality (unit selection and smoothing increased the MOS by 1.0 points as compared to residual selection, the best competing technique, see Tables 4.3 and 4.5) and
- voice similarity (85% of the subjects recognized the target voice listening to speech produced by residual smoothing compared to the best competing techniques, residual codebook and residual selection only achieving 70%).

[41] Minimum naturalness and dissimilarity scores of 3.0.

The development of text-independent solutions to voice conversion was focus of this dissertation. All techniques introduced in this work (VTLN-based and linear-transformation-based voice conversion and all residual prediction techniques) are compatible with the text-independent paradigm introduced in Chapter 5.

Starting from a technique based on automatic segmentation and mapping of artificial phonetic classes, the author developed a completely data-driven approach based on unit selection which proved to produce similar speech quality and voice similarity compared with text-dependent voice conversion (see Table 5.4). The same result was obtained when applying text-independent voice conversion to the cross-language task as shown in Table 5.6.

Experimental observations (see Figure 5.5) showed that text-independent voice conversion produces speech located closer to the unconverted source voice than that produced by text-dependent voice conversion. Further experimental and mathematical investigations confirmed this effect which lets the converted voice sound the more similar to the source, the more speech data was used for training. For its paradoxical nature, the author called the effect *speech alignment paradox* and drew the conclusion that, if enough speech data is available, an arbitrary utterance of an arbitrary voice can be produced only by selecting and concatenating units from this data.

The research discussed in this dissertation has significantly contributed to the scientific progress in the field of voice conversion as indicated by

- the number of the author's peer-reviewed publications related to this thesis (about 20),
- the number of citations to the author's publications (more than 100),
- the number of invited speeches at international conferences, universities, and other scientific venues (about 40),
- the fact that the author was granted a patent on voice conversion based on time domain VTLN.

Potential future extensions of this work aim at further improving the two key properties assessing the presented voice conversion technology:

- *speech quality* may be further improved by using more sophisticated acoustic synthesis models as for instance the STRAIGHT method mentioned in Section 2.4.3. Furthermore, as discussed in Section 5.3.1, precision and robustness of the pitch tracker play a crucial role for the quality of the output speech. Consequently, a systematic study on pitch tracking techniques for voice conversion could be helpful, in particular for applications dealing with non-professional recordings.
- *voice similarity* may gain by extending the variety of properties describing the individuals' voice quality. For example, parameters such as jitter (random variation of the source periodicity) or shimmer (random variation of the excitation amplitude) can have a significant impact on the perceived voice quality (breathiness, creakiness, softness) [van Santen & Sproat^{+} 96]. Last but no least, it is worthwhile to investigate speech production models which explicitly consider the interaction between voice excitation and vocal tract [Rothenberg 84, Childers & Wong 94, Titze & Story 97] being the main drawback of linear predictive coding, cf. the discussion in the gedankenexperiment in Chapter 4.

A Appendix

A.1 TC-Star

An important part of this dissertation's work was performed within the framework of the project TC-Star (Technology and Corpora for Speech-to-Speech Translation) financed by the European Commission. Project members are 12 prestigious research institutions from industry and universities (for details, see `http://tc-star.org`).

The project's main goal is to significantly reduce the gap between human and machine translation performance. The 36 month project starting on April 1, 2004 was to support basic research in the areas speech recognition, machine translation, speech synthesis, and voice conversion (cf. also Figure 5.4) in the domain of parliamentarian and other political speeches. The long-term goal of this research is (among others) its application to real-time translation tasks in European and other parliaments. At first, the world's most frequently spoken languages, Mandarin, English, and Spanish, were considered.

In the framework of TC-Star, large text and speech corpora for all of the four research areas were compiled. E.g., a part of the recognition and translation data was derived from about 40 hours of speech recorded at the European parliament plenary sessions [Gollan & Bisani$^+$ 05]. Detailed information about the voice-conversion-related corpora is given in Sections A.2.3 and A.2.4.

In the first two TC-Star evaluation campaigns, in 2005 and 2006, impressive results were reported in all of the aforementioned research fields as given in [Mostefa & Garcia$^+$ 06] and [Bonafonte & Banchs$^+$ 06]. Some of the voice-conversion-related outcomes were discussed in Chapters 4 and 5.

A.2 Corpora

A.2.1 Spanish Synthesis Corpus

This corpus was recorded for the use in the unit-selection-based speech synthesis system of the Technical University of Catalonia [Bonafonte & Esquerra$^+$ 98]. It consists of speech uttered by two professional speakers (both genders) and the corresponding Laryngograph data, recorded in an acoustically isolated environment. For the current work, the original sampling rate of 32kHz was reduced to 16kHz. The total corpus size is more than one hour for each speaker. For details, see Table A.1.

A.2.2 German Synthesis Corpus

The German synthesis corpus was generated for the use of diphone-based speech synthesis in embedded systems [Hoffmann & Jokisch$^+$ 03]. It was designed to cover all possible German diphones (1212). A professional female and a semi-professional male speaker were recorded using 16kHz, 16bit resolution in a professional sound studio. Details can be found in Table A.2.

speakers	
speaker profile	professional native Spanish
number of speakers	1 female, 1 male
text	
language	Spanish
phonetic coverage	the number of different triphones is maximized
number of utterances	≈ 1000, length about nine words
recording setup	
recording environment	acoustically isolated
input devices	large-membrane microphone and Laryngograph
sampling rate	32kHz downsampled to 16kHz
resolution	16bit

Table A.1: Properties of the Spanish synthesis corpus.

speakers	
speaker profile	(semi-)professional native German
number of speakers	1 female, 1 male
text	
language	German
phonetic coverage	each diphone is contained at least once
number of utterances	≈ 300
recording setup	
recording environment	acoustically isolated
input devices	large-membrane microphone
sampling rate	16kHz
resolution	16bit

Table A.2: Properties of the German synthesis corpus.

A.2.3 First TC-Star Evaluation Corpus

This corpus was produced following the specifications of the TC-STAR project [Bonafonte & Höge$^+$ 05b]. The text had to be phonetically balanced (each diphone had to appear at least once), and a special recording paradigm called the *mimicking approach* was applied. The latter is to produce an optimal time alignment and similar prosody evolutions of identical sentences uttered by different speakers [Kain & Macon 01]. This is to achieve a higher quality of the text-dependent time alignment based on dynamic time warping. Furthermore, the mimicking approach reduces the contribution of prosody to voice dissimilarity when used as test data, as the present work focuses on voice quality, see Section 1.

According to the mimicking approach, first, a template speaker is recorded uttering the whole text corpus to be considered. In a second step, the final speakers (in the case of the first TC-Star evaluation corpus two female and two male) listen to an utterance of the template speaker and then are asked to repeat this utterance following his pitch and timing pattern.

The recording was carried out in an acoustically isolated room whose reverberation time was smaller than 0.3s. The signal-to-noise ratio of the recorded speech is greater than

speakers	
speaker profile	non-professional native Spanish
number of speakers	2 female, 2 male
text	
language	Spanish
phonetic coverage	each diphone is contained at least once
number of utterances	50, length about nine words
recording setup	
recording environment	acoustically isolated
input devices	large-membrane and close-talk microphone, Laryngograph
sampling rate	96kHz downsampled to 16kHz
resolution	24bit downsampled to 16bit

Table A.3: Properties of the first TC-Star evaluation corpus.

40dB. A large-membrane microphone, a close-talk microphone, and a Laryngograph served as input devices; only the first was used for the experiments in this work's scope. The very high recording resolution of 96kHz at 24bit was downsampled to 16kHz at 16bit to fulfill the TC-Star evaluation specifications [Sündermann & Bonafonte$^+$ 05]. Details are given in Table A.3.

A.2.4 Second TC-Star Evaluation Corpus

In the second TC-Star evaluation campaign, as a new task, cross-language voice conversion was considered. To be able to compare text-dependent with text-independent alignment techniques for the cross-language task, the corpus was required to be bilingual, see discussion in Section 2.2. For this purpose, four professional speakers (two female, two male), all bilingual English/Spanish, were carefully selected according to the specifications in [Bonafonte & Höge$^+$ 05b]. They read the same text and the corresponding English text which was produced to maximize the occurrence of rare phonemes as described in [Sündermann 05]. In doing so, they followed the mimicking approach introduced in Section A.2.3. The recording conditions and device characteristics were the same as in the case of the first TC-Star evaluation corpus. For details, see Table A.4.

A.3 Proofs

A.3.1 Time Correspondence of a Scaled and Shifted Discrete Fourier Spectrum

The time correspondence $\tilde{x}(t)$ of a Fourier spectrum scaled by α and shifted by β is to be derived:

$$\tilde{x}(t) = \mathcal{F}^{-1}\left\{X\left(\frac{\omega-\beta}{\alpha}\right)\right\}(t) \qquad \text{(A.1)}$$

Taking into account the auxiliary spectrum

$$X'(\omega) = X\left(\frac{\omega}{\alpha}\right) \qquad \text{(A.2)}$$

speakers	
speaker profile	professional native bilingual English/Spanish
number of speakers	2 female, 2 male
text	
language	English, Spanish
phonetic coverage	each phoneme occurs at least 10 times
number of utterances	159 for English, 196 for Spanish
recording setup	
recording environment	acoustically isolated
input devices	large-membrane and close-talk microphone, Laryngograph
sampling rate	96kHz downsampled to 16kHz
resolution	24bit downsampled to 16bit

Table A.4: Properties of the second TC-Star evaluation corpus.

implies

$$\tilde{x}(t) = \mathcal{F}^{-1}\left\{X'\left(\omega - \beta\right)\right\}(t). \tag{A.3}$$

Equation A.2 can be expressed in time domain using the scaling theorem of the Fourier transformation [Smith 03]:

$$\begin{aligned} x'(t) &= \alpha\mathcal{F}^{-1}\{X(\omega)\}(\alpha t) \\ &= \alpha x(\alpha t) \end{aligned} \tag{A.4}$$

Applying the shift theorem to Equation A.3 yields

$$\begin{aligned} \tilde{x}(t) &= e^{\imath\beta t}\mathcal{F}^{-1}\left\{X'(\omega)\right\}(t) \\ &= e^{\imath\beta t}x'(t) \\ &= \alpha e^{\imath\beta t}x(\alpha t). \end{aligned} \tag{A.5}$$

A.3.2 Time Correspondence of a Rectangular Window

The time correspondence of the rectangular window $R^{(i)}$ (see Equation 3.28) is to be investigated. At first, Equation 3.22 is reshaped considering the segment-dependent frame length $K^{(i)}$ introduced in Equation 3.34 and the relation between $K^{(i)}$ and $N^{(i)}$ according to Equation 3.21. From the continuous parameters ω_i, discrete counterparts n_i can be approximated[42]:

$$n_j = \left\lfloor \omega_j \frac{K^{(i)}}{2\pi} + 1 \right\rfloor. \tag{A.6}$$

[42] The subtle distinction between indices i and j is due to the fact that the parameters n_i and n_{i+1} are to be extracted while keeping the frame length $K^{(i)}$ constant.

According to the considerations on $x_k^{(i*)}$ in Section 3.3.2, when calculating the inverse discrete Fourier transform of $R^{(i)}$, at first, only one part of the two symmetric contributions of Equation 3.19 is to be considered denoted by the $(*)$ superscript[43]:

$$
\begin{aligned}
r_k^{(i*)} &= \frac{1}{K}\sum_{n=1}^{K} R_n(n_i, n_{i+1}) e^{2\kappa\imath(n-1)} \\
&= \frac{1}{K}\left[\frac{1}{2}e^{2\kappa\imath(n_i-1)} + \sum_{n=n_i+1}^{n_{i+1}-1} e^{2\kappa\imath(n-1)} + \frac{1}{2}e^{2\kappa\imath(n_{i+1}-1)}\right] \\
&= \frac{1}{K}\left[\sum_{n=0}^{n_{i+1}-2} e^{2\kappa\imath n} - \sum_{n=0}^{n_i-1} e^{2\kappa\imath n} + \frac{1}{2}e^{2\kappa\imath(n_i-1)} + \frac{1}{2}e^{2\kappa\imath(n_{i+1}-1)}\right] \\
&= \frac{1}{K}\left[\sum_{n=0}^{n_{i+1}-2} \cos(2\kappa n) - \sum_{n=0}^{n_i-1} \cos(2\kappa n) + \sum_{n=0}^{n_{i+1}-2} \imath\sin(2\kappa n) - \sum_{n=0}^{n_i-1} \imath\sin(2\kappa n)\right. \\
&\quad \left. + \frac{1}{2}e^{2\kappa\imath(n_i-1)} + \frac{1}{2}e^{2\kappa\imath(n_{i+1}-1)}\right]. \qquad \text{(A.7)}
\end{aligned}
$$

Finally, the following summation formulae [Gradshteyn & Ryzhik 80]

$$
\sum_{\nu=0}^{n} \cos(2\kappa\nu) = \frac{\sin(\kappa(n+1))\cos(\kappa n)}{\sin(\kappa)}, \tag{A.8}
$$

$$
\sum_{\nu=0}^{n} \sin(2\kappa\nu) = \frac{\sin(\kappa(n+1))\sin(\kappa n)}{\sin(\kappa)} \tag{A.9}
$$

are applied and yield

$$
\begin{aligned}
r_k^{(i*)} &= \frac{1}{K}\left[\frac{\dfrac{\cos(\kappa(n_{i+1}-2)) + \imath\sin(\kappa(n_{i+1}-2))}{\sin(\kappa)}}{\sin(\kappa(n_{i+1}-1))} - \frac{\dfrac{\cos(\kappa(n_i-1)) + \imath\sin(\kappa(n_i-1))}{\sin(\kappa)}}{\sin(\kappa n_i)}\right. \\
&\quad \left. + \frac{1}{2}e^{2\kappa\imath(n_i-1)} + \frac{1}{2}e^{2\kappa\imath(n_{i+1}-1)}\right] \\
&= \frac{1}{K}\left[\frac{\sin(\kappa(n_{i+1}-1))}{\sin(\kappa)}e^{\kappa\imath(n_{i+1}-2)} - \frac{\sin(\kappa n_i)}{\sin(\kappa)}e^{\kappa\imath(n_i-1)} + \frac{1}{2}e^{2\kappa\imath(n_i-1)} + \frac{1}{2}e^{2\kappa\imath(n_{i+1}-1)}\right]
\end{aligned}
\tag{A.10}
$$

with the special case (the window's mean)

$$
r_1^{(i*)} = \frac{n_{i+1}-n_i}{K}. \tag{A.11}
$$

[43] To simplify matters, in the following, the segment-dependent frame length $K^{(i)}$ will be referred to as K. The same applies to κ. $R_n(n_i, n_{i+1})$ is the discrete counterpart to $R(\omega|\omega_i, \omega_{i+1})$, cf. Equation 3.28.

A.3.3 Real Part of a Rectangular Window's Time Correspondence

Considering $\omega_0 = 0$ and $\omega_1 = \pi$, Equation A.6 produces

$$n_0 = 1, \qquad n_1 = \frac{K^{(0)}}{2} + 1. \tag{A.12}$$

Inserting these values in Equation A.10 gives (considering Footnote 43)

$$\begin{aligned}
\Re\big(r_k^{(0*)}\big) &= \frac{1}{K}\left[\frac{\sin\left(\pi\frac{k-1}{K}\left(\frac{K}{2}+1-1\right)\right)}{\sin\left(\pi\frac{k-1}{K}\right)}\cos\left(\pi\frac{k-1}{K}\left(\frac{K}{2}+1-2\right)\right) - \frac{\sin(\kappa)}{\sin(\kappa)}\cos(\kappa(1-1))\right.\\
&\quad \left. + \frac{1}{2}\cos(2\kappa(1-1)) + \frac{1}{2}\cos\left(2\pi\frac{k-1}{K}\left(\frac{K}{2}+1-1\right)\right)\right]\\
&= \frac{1}{K}\left[\frac{\sin\left(\pi\frac{k-1}{2}\right)}{\sin\left(\pi\frac{k-1}{K}\right)}\cos\left(\pi\frac{k-1}{2} - \pi\frac{k-1}{K}\right) - \frac{1}{2} + \frac{1}{2}\cos\left(\pi(k-1)\right)\right]\\
&= \frac{1}{K}\left[A_k - \frac{1}{2} + B_k\right].
\end{aligned} \tag{A.13}$$

Now, four cases have to be differentiated:

- If k is odd and greater than 1, the expression $\frac{k-1}{2}$ becomes integer and yields $\sin\left(\pi\frac{k-1}{2}\right) = 0$ and, hence, $A_k = 0$. Furthermore, from the fact that $k-1$ is even, one gets $\cos(\pi(k-1)) = 1$ and $B_k = \frac{1}{2}$. Finally, summing up the contributions produces $\Re\big(r_k^{(0*)}\big) = 0$.
- If $k \in \{2, 6, 10, \ldots\}$, one has $\frac{k-1}{2} \in \{\frac{1}{2}, \frac{1}{2}+2, \frac{1}{2}+4, \ldots\}$ and, hence,

  $$\sin\left(\pi\frac{k-1}{2}\right) = 1. \tag{A.14}$$

  Furthermore, for an arbitrary ξ holds

  $$\cos\left(\xi + \pi\frac{k-1}{2}\right) = \cos\xi + \frac{\pi}{2} = -\sin\xi. \tag{A.15}$$

  Setting $\xi = -\pi\frac{k-1}{K}$ yields $A_k = \frac{1}{\sin(-\xi)}(-\sin(\xi)) = 1$. Taking into account that $\cos(\pi(k-1)) = -1$ for even k, one gets $B_k = -\frac{1}{2}$ and $\Re\big(r_k^{(0*)}\big) = 0$.
- If $k \in \{4, 8, 12, \ldots\}$, one gets $\frac{k-1}{2} \in \{\frac{3}{2}, \frac{3}{2}+2, \frac{3}{2}+4, \ldots\}$. This gives

  $$\sin\left(\pi\frac{k-1}{2}\right) = -1, \qquad \cos\left(\xi + \pi\frac{k-1}{2}\right) = \cos\xi + \frac{3}{2}\pi = \sin\xi. \tag{A.16}$$

  Using the above definition of ξ yields $A_k = \frac{-1}{\sin(-\xi)}\sin(\xi) = 1$. Summing up with the above derived B_k for even k produces $\Re\big(r_k^{(0*)}\big) = 0$.
- For $k = 1$, Equation A.11 applies:

  $$\Re\big(r_k^{(0*)}\big) = \frac{n_{i+1} - n_i}{K} = \frac{\frac{K}{2} + 1 - 1}{K} = \frac{1}{2}. \tag{A.17}$$

The following equality resumes these cases:

$$\Re\big(r_k^{(0*)}\big) = \begin{cases} \frac{1}{2}: & k = 1\\ 0: & k > 1. \end{cases} \tag{A.18}$$

A.3.4 Definite Integral of Gaussian Distribution and Error Function

It is to be proven that

$$2\int_{-\infty}^{\infty} \frac{x-\mu_y^{(j)}}{\sigma}\,\Phi\left(\frac{x-\mu_y^{(j)}}{\sigma}\right) f\left(\frac{x-\mu_x^{(i)}}{\sigma}\right)\mathrm{d}x \tag{A.19}$$

$$= \sqrt{2}\sigma f\left(\frac{\mu_y^{(j)}-\mu_x^{(i)}}{\sqrt{2}\sigma}\right) + 2(\mu_y^{(j)}-\mu_x^{(i)})\left[\Phi\left(\frac{\mu_y^{(j)}-\mu_x^{(i)}}{\sqrt{2}\sigma}\right)-1\right].$$

By substituting x by $\sigma\xi+\mu_y^{(j)}$, and defining

$$\delta = \frac{\mu_x^{(i)}-\mu_y^{(j)}}{\sigma}, \tag{A.20}$$

the proof can be rewritten as

$$2\int_{-\infty}^{\infty} \xi\Phi(\xi)\,f(\xi-\delta)(\sigma\mathrm{d}\xi) = \sqrt{2}\sigma f\left(\frac{-\delta}{\sqrt{2}}\right) - 2\sigma\delta\left[\Phi\left(\frac{-\delta}{\sqrt{2}}\right)-1\right]. \tag{A.21}$$

For further simplification, this equation is divided by 2σ, and the same symmetry relations as for Equation 5.36 are applied leading to the following expression to be proven[44]:

$$\int_{-\infty}^{\infty} \xi\Phi(\xi)\,f(\xi-\delta)\mathrm{d}\xi = \frac{1}{\sqrt{2}}f\left(\frac{\delta}{\sqrt{2}}\right)+\delta\Phi\left(\frac{\delta}{\sqrt{2}}\right). \tag{A.22}$$

Look at the first term and substitute $\xi = y+\delta$:

$$\int_{-\infty}^{\infty} y\Phi(y+\delta)\,f(y)\mathrm{d}y + \delta\int_{-\infty}^{\infty}\Phi(y+\delta)\,f(y)\mathrm{d}y \quad = \quad I_1(\delta)+\delta I_2(\delta). \tag{A.23}$$

To determine $I_1(\delta)$, integration per partes is applied with $u(y)=\Phi(y+\delta)$ and $v(y)=-f(y)$, and, finally, considering the derivation in Equation 5.18 produces

$$I_1(\delta) \quad = \quad \Big[-\Phi(y)f(y)\Big]_{y=-\infty}^{\infty} + \int_{-\infty}^{\infty} f(y)\,f(y+\delta)\mathrm{d}y \quad = \quad \frac{1}{\sqrt{2}}f\left(\frac{\delta}{\sqrt{2}}\right). \tag{A.24}$$

As for $I_2(\delta)$, again, integration per partes is used with $u(y)=\Phi(y+\delta)$ and $v(y)=\Phi(y)$ deducing

$$I_2(\delta) = \Big[\Phi(y+\delta)\Phi(y)\Big]_{y=-\infty}^{\infty} + \int_{-\infty}^{\infty}\Phi(y)\,f(y+\delta)\mathrm{d}y \tag{A.25}$$

whose first term is 1; and for the second term, substituting $y=z-\delta$ yields

$$\int_{-\infty}^{\infty}\Phi(y)\,f(y+\delta)\mathrm{d}y \quad = \quad \int_{-\infty}^{\infty}\Phi(z-\delta)\,f(z)\mathrm{d}z \quad = \quad I_2(-\delta). \tag{A.26}$$

[44] The following proof is based on an unpublished work of Dr. Jaka Smrekar.

Inserting this into Equation A.25 gives

$$I_2(\delta) + I_2(-\delta) = 1. \tag{A.27}$$

Now, the term $I_2(\delta) - I_2(-\delta)$ is to be computed which is by definition

$$\begin{aligned} I_2(\delta) - I_2(-\delta) &= \int\limits_{-\infty}^{\infty} f(x)\big[\Phi(x+\delta) - \Phi(x-\delta)\big]\mathrm{d}x \\ &= \int\limits_{-\infty}^{\infty} f(x) \int\limits_{-x-\delta}^{x+\delta} f(u)\mathrm{d}u \; \mathrm{d}x. \end{aligned} \tag{A.28}$$

Substituting $u = y + x$, changing the integration order, applying Equation 5.18, and substituting $z = \frac{y}{\sqrt{2}}$ produces

$$\begin{aligned} I_2(\delta) - I_2(-\delta) &= \int\limits_{-\infty}^{\infty} f(x) \int\limits_{-\delta}^{\delta} f(y+x)\mathrm{d}y \; \mathrm{d}x \\ &= \int\limits_{-\delta}^{\delta} \int\limits_{-\infty}^{\infty} f(x)f(y+x)\mathrm{d}x \; \mathrm{d}y \\ &= \int\limits_{-\delta}^{\delta} \frac{1}{\sqrt{2}} f\left(\frac{y}{\sqrt{2}}\right)\mathrm{d}y \\ &= \int\limits_{-\frac{\delta}{\sqrt{2}}}^{\frac{\delta}{\sqrt{2}}} f(z)\mathrm{d}z \quad = \quad 2\int\limits_{0}^{\frac{\delta}{\sqrt{2}}} f(z)\mathrm{d}z. \end{aligned} \tag{A.29}$$

Adding Equations A.27 and A.29 yields

$$\begin{aligned} I_2(\delta) &= \frac{1}{2} + \int\limits_{0}^{\frac{\delta}{\sqrt{2}}} f(z)\mathrm{d}z \\ &= \int\limits_{-\infty}^{0} f(z)\mathrm{d}z + \int\limits_{0}^{\frac{\delta}{\sqrt{2}}} f(z)\mathrm{d}z \\ &= \Phi\left(\frac{\delta}{\sqrt{2}}\right). \end{aligned} \tag{A.30}$$

The latter and Equation A.24 add up to the expression in Equation A.23 finishing the proof.

A.3.5 Limit of Expected Minimal Distances of Random Vectors Distributed According to the Gaussian Mixture Model

It is to be proven that

$$\lim_{N\to\infty} E_N(D) = \lim_{N\to\infty} \frac{1}{\sigma}\sum_{i=1}^{I} \alpha_x^{(i)} \int\limits_{-\infty}^{\infty} E_N(D|x)\, f\left(\frac{x-\mu_x^{(i)}}{\sigma}\right)\mathrm{d}x = 0 \tag{A.31}$$

with

$$\begin{aligned} E_N(D|x) \;&=\; \frac{N}{\sigma}\sum_{j=1}^{J}\alpha_y^{(j)}\left[\int\limits_{-\infty}^{x}(x-y)f\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)\left(1+\Phi\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)-\Phi\left(\frac{2x-y-\mu_y^{(j)}}{\sigma}\right)\right)^{N-1}\mathrm{d}y\right.\\ &\left.-\int\limits_{x}^{\infty}(x-y)f\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)\left(1-\Phi\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)+\Phi\left(\frac{2x-y-\mu_y^{(j)}}{\sigma}\right)\right)^{N-1}\mathrm{d}y\right]. \end{aligned} \tag{A.32}$$

Changing the orders of limit, summation, and integration, Equation A.31 becomes

$$\lim_{N\to\infty} E_N(D) = \sum_{i=1}^{I}\sum_{j=1}^{J}\alpha_x^{(i)}\alpha_y^{(j)}\lim_{N\to\infty} E_N^{(i,j)}(D) \tag{A.33}$$

with

$$\begin{aligned} E_N^{(i,j)}(D) = \frac{N}{\sigma^2}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{x} f\left(\frac{x-\mu_x^{(i)}}{\sigma}\right)(x-y)f\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)\left(1+\Phi\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)-\Phi\left(\frac{2x-y-\mu_y^{(j)}}{\sigma}\right)\right)^{N-1}\mathrm{d}y\;\mathrm{d}x\\ -\frac{N}{\sigma^2}\int\limits_{-\infty}^{\infty}\int\limits_{x}^{\infty} f\left(\frac{x-\mu_x^{(i)}}{\sigma}\right)(x-y)f\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)\left(1-\Phi\left(\frac{y-\mu_y^{(j)}}{\sigma}\right)+\Phi\left(\frac{2x-y-\mu_y^{(j)}}{\sigma}\right)\right)^{N-1}\mathrm{d}y\;\mathrm{d}x. \end{aligned} \tag{A.34}$$

Substituting $u = \frac{y-\mu_y^{(j)}}{\sigma}$ and $v = 2\frac{x-\mu_y^{(j)}}{\sigma} - u$ and using Equation A.20, this can be rewritten as[45]

$$\begin{aligned} E_N^{(i,j)}(D) \;&=\; \frac{N\sigma}{4}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{v} f\left(\frac{u+v}{2}-\delta\right)(v-u)f(u)(1+\Phi(u)-\Phi(v))^{N-1}\mathrm{d}u\;\mathrm{d}v \quad (\text{A.35})\\ &-\frac{N\sigma}{4}\int\limits_{-\infty}^{\infty}\int\limits_{v}^{\infty} f\left(\frac{u+v}{2}-\delta\right)(v-u)f(u)(1-\Phi(u)+\Phi(v))^{N-1}\mathrm{d}u\;\mathrm{d}v. \end{aligned}$$

Applying integration per partes in the inner integrals [the terms $Nf(u)(1+\Phi(u)-\Phi(v))^{N-1}$ and $-Nf(u)(1-\Phi(u)+\Phi(v))^{N-1}$ are differentials of $(1+\Phi(u)-\Phi(v))^{N}$ and $(1-\Phi(u)+\Phi(v))^{N}$ with respect to u] yields the equality

$$E_N^{(i,j)}(D) \;=\; \frac{\sigma}{4}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{v}(1+\Phi(u)-\Phi(v))^{N} f\left(\frac{u+v}{2}-\delta\right)\left(\left(\frac{u+v}{2}-\delta\right)\frac{u-v}{2}+1\right)\mathrm{d}u\;\mathrm{d}v$$

[45] The following proof is based on an unpublished work of Dr. Jaka Smrekar.

$$+\frac{\sigma}{4}\int\limits_{-\infty}^{\infty}\int\limits_{v}^{\infty}(1-\Phi(u)+\Phi(v))^N f\left(\frac{u+v}{2}-\delta\right)\left(\left(\frac{u+v}{2}-\delta\right)\frac{v-u}{2}+1\right)\mathrm{d}u\ \mathrm{d}v \tag{A.36}$$

Switching the roles of u and v in the second summand applying Fubini's theorem [Rudin 87], one obtains

$$E_N^{(i,j)}(D) \quad = \quad \frac{\sigma}{2}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{v}(1+\Phi(u)-\Phi(v))^N f\left(\frac{u+v}{2}-\delta\right)\mathrm{d}u\ \mathrm{d}v. \tag{A.37}$$

Note that for $u \leq v$, the expression $1+\Phi(u)-\Phi(v)$ is at most 1, and therefore

$$\lim_{N\to\infty}(1+\Phi(u)-\Phi(v))^N = \begin{cases} 1: & u=v \\ 0: & \text{otherwise.} \end{cases} \tag{A.38}$$

A double application of Lebesgue's dominated convergence theorem [Browder 96] yields

$$\lim_{N\to\infty} E_N^{(i,j)}(D) = 0 \tag{A.39}$$

which, inserted into Equation A.33, finishes the proof.

References

[Abe & Nakamura$^+$ 88] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara: Voice Conversion through Vector Quantization. In *Proc. of the ICASSP*, New York, USA, 1988.

[Abe & Shikano$^+$ 90] M. Abe, K. Shikano, H. Kuwabara: Cross-Language Voice Conversion. In *Proc. of the ICASSP*, Albuquerque, USA, 1990.

[Acero & Stern 91] A. Acero, R. Stern: Robust Speech Recognition by Normalization of the Acoustic Space. In *Proc. of the ICASSP*, Toronto, Canada, 1991.

[Agüero & Adell$^+$ 06] P. Agüero, J. Adell, A. Bonafonte: Prosody Generation for Speech-to-Speech Translation. In *Proc. of the ICASSP*, Toulouse, France, 2006.

[Ainsworth & Paliwal$^+$ 84] W. Ainsworth, K. Paliwal, H. Foster: Problems with Dynamic Frequency Warping as a Technique for Speaker-Independent Vowel Classification. *Proc. of the Institute of Acoustics*, Vol. 6, No. 4, 1984.

[Ali & Spiegel$^+$ 99] A. Ali, J. Spiegel, P. Mueller, G. Haentjens, J. Berman: An Acoustic-Phonetic Feature-Based System for Automatic Phoneme Recognition in Continuous Speech. In *Proc. of the ISCAS*, Orlando, USA, 1999.

[Arfken 85] G. Arfken: *Mathematical Methods for Physicists.* Academic Press, Orlando, USA, 1985.

[Atal & Hanauer 71] B. Atal, S. Hanauer: Speech Analysis and Synthesis by Linear Prediction. *Journal of the Acoustical Society of America*, Vol. 50, No. 2, 1971.

[Avriel 76] M. Avriel: *Nonlinear Programming: Analysis and Methods.* Prentice Hall, Englewood Cliffs, USA, 1976.

[Banno & Hata$^+$ 07] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, H. Kawahara: Implementation of Realtime STRAIGHT Speech Manipulation System: Report on Its First Implementation. *Acoustical Science and Technology*, Vol. 28, No. 3, 2007.

[Barbosa 97] A. Barbosa. A New Mexican Spanish Voice for the Festival Text to Speech System. Master's thesis, University of the Americas, Puebla, Mexico, 1997.

[Bennett 05] C. Bennett: Large Scale Evaluation of Corpus-Based Synthesizers: Results and Lessons from the Blizzard Challenge 2005. In *Proc. of the Interspeech*, Lisbon, Portugal, 2005.

[Bernadin & Foo 06] S. Bernadin, S. Foo: Wavelet Processing for Pitch Period Estimation. In *Proc. of the Southeastern Symposium on System Theory*, Cookeville, USA, 2006.

[Beutnagel & Conkie$^+$ 99] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, A. Syrdal: The AT&T Next-Gen TTS System. In *Proc. of the Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, 1999.

[Black & Lenzo 01] A. Black, K. Lenzo: Optimal Data Selection for Unit Selection Synthesis. In *Proc. of the ISCA Workshop on Speech Synthesis*, Perthshire, UK, 2001.

[Black & Lenzo 03] A. Black, K. Lenzo: *Building Synthetic Voices.* Carnegie Mellon University, Pittsburgh, USA, 2003.

[Boersma 01] P. Boersma: Praat, a System for Doing Phonetics by Computer. *Glot International*, Vol. 5, No. 9/10, 2001.

[Bonafonte & Banchs$^+$ 06] A. Bonafonte, R. Banchs, A. Moreno: *TC-STAR Workshop on Speech-to-Speech Translation.* Barcelona, Spain, 2006.

[Bonafonte & Esquerra$^+$ 98] A. Bonafonte, I. Esquerra, A. Febrer, J. Fonollosa, F. Vallverdú: The UPC Text-to-Speech System for Spanish and Catalan. In *Proc. of the ICSLP*, Sydney, Australia, 1998.

[Bonafonte & Höge$^+$ 05a] A. Bonafonte, H. Höge, I. Kiss, A. Moreno, D. Sündermann, U. Ziegenhain, J. Adell, P. Agüero, H. Duxans, D. Erro, J. Nurminen, J. Pérez, G. Strecha, M. Umbert, X. Wang. TC-STAR: TTS Progress Report. Technical report, 2005.

[Bonafonte & Höge$^+$ 05b] A. Bonafonte, H. Höge, H. Tropf, A. Moreno, H. v. d. Heuvel, D. Sündermann, U. Ziegenhain, J. Pérez, I. Kiss. TC-Star: Specifications of Language Resources for Speech Synthesis. Technical report, 2005.

[Bořil & Fousek$^+$ 06] H. Bořil, P. Fousek, D. Sündermann, P. Červa, J. Žd'ánský: Lombard Speech Recognition: A Comparative Study. In *Proc. of the Czech-German Workshop*, Prague, Czech Republic, 2006.

[Boudelaa & Meftah 96] S. Boudelaa, M. Meftah: Cross-Language Effects of Lexical Stress in Word Recognition: The Case of Arabic English Bilinguals. In *Proc. of the ICSLP*, Philadelphia, USA, 1996.

[Braun & Masthoff 02] A. Braun, H. Masthoff: *Phonetics and Its Applications.* Steiner, Stuttgart, Germany, 2002.

[Browder 96] A. Browder: *Mathematical Analysis: An Introduction.* Springer, New York, USA, 1996.

[Ceyssens & Verhelst$^+$ 02] T. Ceyssens, W. Verhelst, P. Wambacq: A Strategy for Pitch Conversion and Its Evaluation. In *Proc. of the SPS*, Leuven, Belgium, 2002.

[Charpentier & Moulines 88] F. Charpentier, E. Moulines: Text-to-Speech Algorithms Based on FFT Synthesis. In *Proc. of the ICASSP*, New York, USA, 1988.

[Charpentier & Stella 86] F. Charpentier, M. Stella: Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation. In *Proc. of the ICASSP*, Tokyo, Japan, 1986.

[Cheng & O'Shaughnessy 89] Y. Cheng, D. O'Shaughnessy: Automatic and Reliable Estimation of Glottal Closure Instant and Period. *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 37, No. 12, 1989.

[Childers & Lee 91] D. Childers, C. Lee: Vocal Quality Factors: Analysis, Synthesis, and Perception. *Journal of the Acoustical Society of America*, Vol. 90, No. 5, 1991.

[Childers & Wong 94] D. Childers, C.F. Wong: Measuring and Modeling Vocal Source-Tract Interaction. *IEEE Trans. on Biomedical Engineering*, Vol. 41, No. 7, 1994.

[Cho & Kim$^+$ 98] Y. Cho, M. Kim, S. Kim: A Spectrally Mixed Excitation (SMX) Vocoder with Robust Parameter Determination. In *Proc. of the ICASSP*, Seattle, USA, 1998.

[Cohen & Kamm$^+$ 95] J. Cohen, T. Kamm, A. Andreou: Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability. *Journal of the Acoustical Society of America*, Vol. 97, No. 5, 1995.

[Cole 98] R. Cole: *Survey of the State of the Art in Human Language Technology.* Giardini Editori e Stampatori, Pisa, Italy, 1998.

[Cormen & Leiserson$^+$ 90] T. Cormen, C. Leiserson, R. Rivest: *Introduction to Algorithms.* MIT Press, Cambridge, USA, 1990.

[David & Nagaraja 03] H. David, H. Nagaraja: *Order Statistics.* Wiley, New York, USA, 2003.

[de Boor 78] C. de Boor: *A Practical Guide to Splines.* Springer, New York, USA, 1978.

[de los Galanes & Savoji$^+$ 94] F. de los Galanes, M. Savoji, J. Pardo: New Algorithm for Spectral Smoothing and Envelope Modification for LP-PSOLA Synthesis. In *Proc. of the ICASSP*, Adelaide, Australia, 1994.

[Deller & Proakis$^+$ 93] J. Deller, J. Proakis, J. Hansen: *Discrete-Time Processing of Speech Signals.* Prentice Hall, Englewood Cliffs, USA, 1993.

[Dharanipragada & Gopinath$^+$ 98] S. Dharanipragada, R. Gopinath, B. Rao: Techniques for Capturing Temporal Variations in Speech Signals with Fixed-Rate Processing. In *Proc. of the ICSLP*, Sydney, Australia, 1998.

[Dolson 94] M. Dolson: The Pitch of Speech as a Function of Linguistic Community. *Music Perception*, Vol. 11, No. 1, 1994.

[Donovan & Woodland 95] R. Donovan, P. Woodland: Improvements in an HMM-Based Speech Synthesiser. In *Proc. of the Eurospeech*, Madrid, Spain, 1995.

[Duda & Hart 73] R. Duda, P. Hart: *Pattern Classification and Scene Analysis.* Wiley, New York, USA, 1973.

[Durbin 60] J. Durbin: The Fitting of Time-Series Models. *Revue de l'Institut International de Statistique*, Vol. 28, No. 3, 1960.

[Dutoit & Pagel$^+$ 96] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, O. v. d. Vrecken: The MBROLA Project: Towards a Set of High Quality Speech Synthesizers Free of Use for Non Commercial Purposes. In *Proc. of the ICSLP*, Philadelphia, USA, 1996.

[Dutoit 93] T. Dutoit. *High Quality Text-To-Speech Synthesis of the French Language.* Ph.D. thesis, Polytechnic Faculty of Mons, Mons, Belgium, 1993.

[Duxans & Bonafonte 03] H. Duxans, A. Bonafonte: Estimation of GMM in Voice Conversion Including Unaligned Data. In *Proc. of the Eurospeech*, Geneva, Switzerland, 2003.

[Eide & Gish 96] E. Eide, H. Gish: A Parametric Approach to Vocal Tract Length Normalization. In *Proc. of the ICASSP*, Atlanta, USA, 1996.

[Eide & Picheny 06] E. Eide, M. Picheny: Towards Pooled-Speaker Concatenative Text-to-Speech. In *Proc. of the ICASSP*, Toulouse, France, 2006.

[ETSI 99] ETSI. Adaptive Multi-Rate (AMR) Speech Transcoding. Technical report, European Telecommunications Standards Institute, Sophia Antipolis, France, 1999.

[Evermann & Chan$^+$ 05] G. Evermann, H. Chan, M. Gales, B. Jia, D. Mrva, P. Woodland, K. Yu: Training LVCSR Systems on Thousands of Hours of Data. In *Proc. of the ICASSP*, Philadelphia, USA, 2005.

[Fallside & Woods 85] F. Fallside, W. Woods: *Computer Speech Processing.* Prentice Hall, Englewood Cliffs, USA, 1985.

[Faltlhauser & Pfau$^+$ 00] R. Faltlhauser, T. Pfau, G. Ruske: On-Line Speaking Rate Estimation Using Gaussian Mixture Models. In *Proc. of the ICASSP*, Istanbul, Turkey, 2000.

[Fant 70] G. Fant: *Acoustic Theory of Speech Production.* Mouton, The Hague, Netherlands, 1970.

[Fischer & Kunzmann 06] V. Fischer, S. Kunzmann: From Pre-Recorded Prompts to Corporate Voices: On the Migration of Interactive Voice Response Applications. In *Proc. of the Interspeech*, Pittsburgh, USA, 2006.

[Fisher & Doddington$^+$ 86] W. Fisher, G. Doddington, K. Goudie-Marshall: The DARPA Speech Recognition Research Database: Specifications and Status. In *Proc. of the DARPA Workshop on Speech Recognition*, Palo Alto, USA, 1986.

[Furui & Nakamura$^+$ 06] S. Furui, M. Nakamura, K. Iwano: Why is automatic recognition of spontaneous speech so difficult? In *Proc. of the Symposium on Large-Scale Knowledge Resources*, Tokyo, Japan, 2006.

[Gauss 09] C. Gauss: *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium.* Perthes and Besser, Hamburg, Germany, 1809.

[Germann 06] U. Germann. An Iterative Approach to Pitch-Marking of Speech Signals without Electroglottographic Data. Technical report, University of Toronto, Toronto, Canada, 2006.

[Gollan & Bisani$^+$ 05] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, H. Ney: Cross Domain Automatic Transcription on the TC-Star EPPS Corpus. In *Proc. of the ICASSP*, Philadelphia, USA, 2005.

[Goncharoff & Gries 98] V. Goncharoff, P. Gries: An Algorithm for Accurately Marking Pitch Pulses in Speech Signals. In *Proc. of the SIP*, Las Vegas, USA, 1998.

[Gradshteyn & Ryzhik 80] I. Gradshteyn, I. Ryzhik: *Table of Integrals, Series and Products.* Academic Press, New York, USA, 1980.

[Gray & Markel 76] A. Gray, J. Markel: Distance Measures for Speech Processing. *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 24, No. 5, 1976.

[Gray 18] H. Gray: *Anatomy of the Human Body.* Lea and Febiger, Philadelphia, USA, 1918.

[Griffin 87] D. Griffin. *Multi-Band Excitation Vocoder.* Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, USA, 1987.

[Hagen 94] R. Hagen: Spectral Quantization of Cepstral Coefficients. In *Proc. of the ICASSP*, Adelaide, Australia, 1994.

[Hain 01] T. Hain. *Hidden Model Sequence Models for Automatic Speech Recognition.* Ph.D. thesis, Cambridge University, Cambridge, UK, 2001.

[Hamon & Moulines$^+$ 89] C. Hamon, E. Moulines, F. Charpentier: A Diphone Synthesis System Based on Time-Domain Prosodic Modification of Speech. In *Proc. of the ICASSP*, Glasgow, UK, 1989.

[Hermansky & Morgan 94] H. Hermansky, N. Morgan: RASTA processing of speech. *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 4, 1994.

[Hermansky 90] H. Hermansky: Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal of the Acoustical Society of America*, Vol. 87, No. 4, 1990.

[Hillebrand 02] F. Hillebrand: *GSM and UMTS: The Creation of Global Mobile Communications.* Wiley, New York, USA, 2002.

[Hoffmann & Jokisch$^+$ 03] R. Hoffmann, O. Jokisch, D. Hirschfeld, G. Strecha, H. Kruschke, U. Kordon: A Multilingual TTS System with Less than 1 Megabyte Footprint for Embedded Applications. In *Proc. of the ICASSP*, Hong Kong, China, 2003.

[Höge & Kotnik$^+$ 06] H. Höge, B. Kotnik, Z. Kacic, H. Pfitzinger: Evaluation of Pitch Marking Algorithms. In *Proc. of the ITG*, Kiel, Germany, 2006.

[Höge 02] H. Höge: Project Proposal TC-STAR - Make Speech to Speech Translation Real. In *Proc. of the LREC*, Las Palmas, Spain, 2002.

[Hogg & McKean$^+$ 95] R. Hogg, J. McKean, A. Craig: *Introduction to Mathematical Statistics.* Prentice Hall, Englewood Cliffs, USA, 1995.

[Horne 00] M. Horne: *Prosody: Theory and Experiment.* Kluwer Academic Publishers, Dordrecht, Netherlands, 2000.

[Hosom & Kain$^+$ 03] J. Hosom, A. Kain, T. Mishra, J. v. Santen, M. Fried-Oken, J. Staehely: Intelligibility of Modifications to Dysarthric Speech. In *Proc. of the ICASSP*, Hong Kong, China, 2003.

[Hunt & Black 96] A. Hunt, A. Black: Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In *Proc. of the ICASSP*, Atlanta, USA, 1996.

[Inanoglu 03] Z. Inanoglu. Transforming Pitch in a Voice Conversion Framework. Master's thesis, University of Cambridge, Cambridge, UK, 2003.

[Itakura 75] F. Itakura: Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals. *Journal of the Acoustical Society of America*, Vol. 57, 1975.

[ITU 96] ITU. Methods for Subjective Determination of Transmission Quality. Technical Report ITU-T Recommendation P.800, International Telecommunication Union, Geneva, Switzerland, 1996.

[Kain & Macon 98] A. Kain, M. Macon: Spectral Voice Conversion for Text-to-Speech Synthesis. In *Proc. of the ICASSP*, Seattle, USA, 1998.

[Kain & Macon 01] A. Kain, M. Macon: Design and Evaluation of a Voice Conversion Algorithm Based on Spectral Envelope Mapping and Residual Prediction. In *Proc. of the ICASSP*, Salt Lake City, USA, 2001.

[Kain 01] A. Kain. *High Resolution Voice Transformation.* Ph.D. thesis, Oregon Health and Science University, Portland, USA, 2001.

[Kawahara & Estill$^+$ 01] H. Kawahara, J. Estill, O. Fujimura: Aperiodicity Extraction and Control Using Mixed Mode Excitation and Group Delay Manipulation for a High Quality Speech Analysis, Modification and Synthesis System STRAIGHT. In *Proc. of the MAVEBA*, Firenze, Italy, 2001.

[Kotnik & Höge$^+$ 06] B. Kotnik, H. Höge, Z. Kacic: Evaluation of Pitch Detection Algorithms in Adverse Conditions. In *Proc. of the Speech Prosody*, Dresden, Germany, 2006.

[Kotnik 06] B. Kotnik: First PMA/PDA Evaluation Campaign. In *Proc. of the AST*, Maribor, Slovenia, 2006.

[Kuwabara & Sagisaka 95] H. Kuwabara, Y. Sagisaka: Acoustic Characteristics of Speaker Individuality: Control and Conversion. *Speech Communication*, Vol. 16, No. 2, 1995.

[Ladefoged 90] P. Ladefoged: The Revised International Phonetic Alphabet. *Language*, Vol. 66, No. 3, 1990.

[Lee & Rose 96] L. Lee, R. Rose: Speaker Normalization Using Efficient Frequency Warping Procedures. In *Proc. of the ICASSP*, Atlanta, USA, 1996.

[Lee & Wu 06] C. Lee, C. Wu: Map-Based Adaptation for Speech Conversion Using Adaptation Data Selection and Non-Parallel Training. In *Proc. of the Interspeech*, Pittsburgh, USA, 2006.

[Lewis & Tatham 99] E. Lewis, M. Tatham: Word and Syllable Concatenation in Text-to-Speech Synthesis. In *Proc. of the Eurospeech*, Budapest, Hungary, 1999.

[Macon 96] M. Macon. *Speech Synthesis Based on Sinusoidal Modeling.* Ph.D. thesis, Georgia Institute of Technology, Atlanta, USA, 1996.

[Makhoul 75] J. Makhoul: Linear Prediction: A Tutorial Review. *Proceedings of the IEEE*, Vol. 63, No. 4, 1975.

[Markel & Gray 76] J. Markel, A. Gray: *Linear Prediction of Speech.* Springer, New York, USA, 1976.

[Mashimo & Toda$^+$ 01] M. Mashimo, T. Toda, K. Shikano, N. Campbell: Evaluation of Cross-Language Voice Conversion Based on GMM and STRAIGHT. In *Proc. of the Eurospeech*, Aalborg, Denmark, 2001.

[Masuko 02] T. Masuko. *HMM-Based Speech Synthesis and Its Applications.* Ph.D. thesis, Tokyo Institute of Technology, Tokyo, Japan, 2002.

[Matsumoto & Wakita 86] H. Matsumoto, H. Wakita: Vowel Normalization by Frequency Warped Spectral Matching. *Speech Communication*, Vol. 5, No. 2, 1986.

[Mattheyses & Verhelst$^+$ 06] W. Mattheyses, W. Verhelst, P. Verhoeve: Robust Pitch Marking for Prosodic Modification of Speech Using TD-PSOLA. In *Proc. of the SPS-DARTS*, Antwerp, Belgium, 2006.

[McAulay & Quatieri 86] R. McAulay, T. Quatieri: Speech Analysis-Synthesis Based on a Sinusoidal Representation. *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 34, 1986.

[McDonough & Byrne$^+$ 98] J. McDonough, W. Byrne, X. Luo: Speaker Normalization with All-Pass Transforms. In *Proc. of the ICSLP*, Sydney, Australia, 1998.

[Meilgaard & Civille$^+$ 99] M. Meilgaard, G. Civille, B. Carr: *Sensory Evaluation Techniques.* CRC Press, Boca Raton, USA, 1999.

[Molau & Kanthak$^+$ 00] S. Molau, S. Kanthak, H. Ney: Efficient Vocal Tract Normalization in Automatic Speech Recognition. In *Proc. of the ESSV*, Cottbus, Germany, 2000.

[Molau 02] S. Molau. *Normalization in the Acoustic Feature Space for Improved Speech Recognition.* Ph.D. thesis, RWTH Aachen, Aachen, Germany, 2002.

[Montgomery 96] D. Montgomery: *Introduction to Statistical Quality Control.* Wiley, New York, USA, 1996.

[Mostefa & Garcia$^+$ 06] D. Mostefa, M. Garcia, O. Hamon, N. Moreau. TC-Star: Evaluation Report. Technical report, 2006.

[Moulines & Charpentier 88] E. Moulines, F. Charpentier: Diphone Synthesis Using a Multipulse LPC Technique. In *Proc. of the FASE Symposium*, Edinburgh, UK, 1988.

[Moulines & Sagisaka 95] E. Moulines, Y. Sagisaka: Voice Conversion: State of the Art and Perspectives. *Speech Communication*, Vol. 16, No. 2, 1995.

[Nakagiri & Toda$^+$ 06] M. Nakagiri, T. Toda, H. Kashioka, K. Shikano: Improving Body Transmitted Unvoiced Speech with Statistical Voice Conversion. In *Proc. of the Interspeech*, Pittsburgh, USA, 2006.

[Nakamura & Toda$^+$ 06] K. Nakamura, T. Toda, H. Saruwatari, K. Shikano: Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech. In *Proc. of the Interspeech*, Pittsburgh, USA, 2006.

[Ney & Welling$^+$ 98] H. Ney, L. Welling, S. Ortmanns, K. Beulen, F. Wessel: The RWTH Large Vocabulary Continuous Speech Recognition System. In *Proc. of the ICASSP*, Seattle, USA, 1998.

[Nicolao & Drioli$^+$ 06] M. Nicolao, C. Drioli, P. Cosi: Voice GMM Modelling for FESTIVAL/MBROLA Emotive TTS Synthesis. In *Proc. of the Interspeech*, Pittsburgh, USA, 2006.

[Nolan 83] F. Nolan: *The Phonetic Bases of Speech Recognition.* Cambridge University Press, Cambridge, UK, 1983.

[Nurminen & Tian$^+$ 06] J. Nurminen, J. Tian, V. Popa: Novel Method for Data Clustering and Mode Selection with Application in Voice Conversion. In *Proc. of the Interspeech*, Pittsburgh, USA, 2006.

[Ohtani & Toda$^+$ 06] Y. Ohtani, T. Toda, H. Saruwatari, K. Shikano: Maximum Likelihood Voice Conversion Based on GMM with STRAIGHT Mixed Excitation. In *Proc. of the Interspeech*, Pittsburgh, USA, 2006.

[Oppenheim & Schafer 89] A. Oppenheim, R. Schafer: *Discrete-Time Signal Processing.* Prentice Hall, Englewood Cliffs, USA, 1989.

[Paliwal 95] K. Paliwal: Interpolation Properties of Linear Prediction Parametric Representations. In *Proc. of the Eurospeech*, Madrid, Spain, 1995.

[Papamichalis 87] P. Papamichalis: *Practical Approaches to Speech Coding.* Prentice Hall, Englewood Cliffs, USA, 1987.

[Picone 93] J. Picone: Signal Modeling Techniques in Speech Recognition. *Proc. of the IEEE*, Vol. 81, No. 9, 1993.

[Pitz & Ney 05] M. Pitz, H. Ney: Vocal Tract Normalization Equals Linear Transformation in Cepstral Space. *IEEE Trans. on Speech and Audio Processing*, Vol. 13, No. 5, 2005.

[Preparata & Shamos 85] F. Preparata, M. Shamos: *Computational Geometry - an Introduction.* Springer, New York, USA, 1985.

[Pye & Woodland 97] D. Pye, P. Woodland: Experiments in Speaker Normalization and Adaptation for Large Vocabulary Speech Recognition. In *Proc. of the ICASSP*, Munich, Germany, 1997.

[Qin & Chen$^+$ 05] L. Qin, G. Chen, Z. Ling, L. Dai: An Improved Spectral and Prosodic Transformation Method in STRAIGHT-Based Voice Conversion. In *Proc. of the ICASSP*, Philadelphia, USA, 2005.

[Rabiner & Juang 93] L. Rabiner, B. Juang: *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, USA, 1993.

[Rabiner & Rosenberg$^+$ 78] L. Rabiner, A. Rosenberg, S. Levinson: Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 26, No. 6, 1978.

[Reynolds 94] D. Reynolds: An Application of Recurrent Nets to Phone Probability Estimation. *IEEE Trans. on Neural Networks*, Vol. 5, No. 2, 1994.

[Reynolds 95] D. Reynolds: Automatic Speaker Recognition Using Gaussian Mixture Models. *Lincoln Laboratory Journal*, Vol. 8, No. 2, 1995.

[Rothenberg 84] M. Rothenberg: Source-Tract Acoustic Interaction and Voice Quality. In *Proc. of the Symposium: Care of the Professional Voice*, New York, USA, 1984.

[Rudin 87] W. Rudin: *Real and Complex Analysis.* McGraw Hill, New York, USA, 1987.

[Seber 84] G. Seber: *Multivariate Observations.* Wiley, New York, USA, 1984.

[Smith 03] J. Smith: *Mathematics of the Discrete Fourier Transform (DFT), with Music and Audio Applications.* W3K Publishing, Menlo Park, USA, 2003.

[Smith 07] J. Smith: *Physical Audio Signal Processing.* Stanford University, Stanford, USA, 2007.

[Spelmezan & Borchers 06] D. Spelmezan, J. Borchers: Minnesang: Speak Medieval German. In *Proc. of the CHI*, Montreal, Canada, 2006.

[Stevens & Volkman$^+$ 37] S. Stevens, J. Volkman, E. Newman: A Scale for Measurement of the Psychological Magnitude of Pitch. *Journal of the Acoustical Society of America*, Vol. 8, No. 3, 1937.

[Stylianou & Cappé$^+$ 95] Y. Stylianou, O. Cappé, E. Moulines: Statistical Methods for Voice Quality Transformation. In *Proc. of the Eurospeech*, Madrid, Spain, 1995.

[Stylianou & Cappé$^+$ 98] Y. Stylianou, O. Cappé, E. Moulines: Continuous Probabilistic Transform for Voice Conversion. *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 2, 1998.

[Stylianou 01] Y. Stylianou: Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis. *IEEE Trans. on Speech and Audio Processing*, Vol. 9, No. 1, 2001.

[Sündermann & Bonafonte$^+$ 04a] D. Sündermann, A. Bonafonte, H. Ney, H. Höge: A First Step Towards Text-Independent Voice Conversion. In *Proc. of the ICSLP*, Jeju Island, South Korea, 2004.

[Sündermann & Bonafonte$^+$ 04b] D. Sündermann, A. Bonafonte, H. Ney, H. Höge: Frequency Domain vs. Time Domain VTLN. In *Proc. of the AST*, Maribor, Slovenia, 2004.

[Sündermann & Bonafonte$^+$ 05] D. Sündermann, A. Bonafonte, H. Duxans, H. Höge: TC-STAR: Evaluation Plan for Voice Conversion Technology. In *Proc. of the DAGA*, Munich, Germany, 2005.

[Sündermann & Höge$^+$ 06] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, J. Hirschberg: Text-Independent Cross-Language Voice Conversion. In *Proc. of the Interspeech*, Pittsburgh, USA, 2006.

[Sündermann & Ney 03a] D. Sündermann, H. Ney: An Automatic Segmentation and Mapping Approach for Voice Conversion Parameter Training. In *Proc. of the AST*, Maribor, Slovenia, 2003.

[Sündermann & Ney$^+$ 03b] D. Sündermann, H. Ney, H. Höge: VTLN-Based Cross-Language Voice Conversion. In *Proc. of the ASRU*, Virgin Islands, USA, 2003.

[Sündermann 05] D. Sündermann: A Language Resources Generation Toolbox for Speech Synthesis. In *Proc. of the AST*, Maribor, Slovenia, 2005.

[Tan & Fu$^+$ 96] B. Tan, M. Fu, A. Spray, P. Dermody: The Use of Wavelet Transforms in Phoneme Recognition. In *Proc. of the ICSLP*, Philadelphia, USA, 1996.

[Tang & Wang$^+$ 01] M. Tang, C. Wang, S. Seneff: Voice Transformations: From Speech Synthesis to Mammalian Vocalizations. In *Proc. of the Eurospeech*, Aalborg, Denmark, 2001.

[Taylor & Black$^+$ 98] P. Taylor, A. Black, R. Caley: The Architecture of the Festival Speech Synthesis System. In *Proc. of the ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

[Tian & Nurminen$^+$ 06] J. Tian, J. Nurminen, V. Popa: Efficient Gaussian Mixture Model Evaluation in Voice Conversion. In *Proc. of the Interspeech*, Pittsburgh, USA, 2006.

[Titze & Story 97] I. Titze, B. Story: Acoustic Interactions of the Voice Source with the Lower Vocal Tract. *Journal of the Acoustical Society of America*, Vol. 101, No. 4, 1997.

[Toda & Lu$^+$ 00] T. Toda, J. Lu, H. Saruwatari, K. Shikano: Straight-Based Voice Conversion Algorithm Based on Gaussian Mixture Model. In *Proc. of the ICSLP*, Beijing, China, 2000.

[Toda & Ohtani$^+$ 06] T. Toda, Y. Ohtani, K. Shikano: Eigenvoice Conversion Based on Gaussian Mixture Model. In *Proc. of the Interspeech*, Pittsburgh, USA, 2006.

[Tranter & Reynolds 04] S. Tranter, D. Reynolds: Speaker Diarisation for Broadcast News. In *Proc. of the Odyssey Speaker and Language Recognition Workshop*, Toledo, Spain, 2004.

[Turk & Arslan 02] O. Turk, L. Arslan: Subband Based Voice Conversion. In *Proc. of the ICSLP*, Denver, USA, 2002.

[Uebel & Woodland 99] L. Uebel, P. Woodland: An Investigation into Vocal Tract Length Normalization. In *Proc. of the Eurospeech*, Budapest, Hungary, 1999.

[Uto & Nankaku$^+$ 06] Y. Uto, Y. Nankaku, T. Toda, A. Lee, K. Tokuda: Voice Conversion Based on Mixtures of Factor Analyzers. In *Proc. of the Interspeech*, Pittsburgh, USA, 2006.

[van Santen & Sproat$^+$ 96] J. van Santen, R. Sproat, J. Olive, J. Hirschberg: *Progress in Speech Synthesis*. Springer, New York, USA, 1996.

[Walker 96] J. Walker: *Fast Fourier Transforms*. CRC Press, Boca Raton, USA, 1996.

[Wegmann & McAllaster$^+$ 96] S. Wegmann, D. McAllaster, J. Orloff, B. Peskin: Speaker Normalization on Conversational Telephone Speech. In *Proc. of the ICASSP*, Atlanta, USA, 1996.

[Weinstein 75] C. Weinstein: A Linear Prediction Vocoder with Voice Excitation. *EASCON Proceedings*, Vol., 1975.

[Welling & Ney$^+$ 02] L. Welling, H. Ney, S. Kanthak: Speaker Adaptive Modeling by Vocal Tract Normalization. *IEEE Trans. on Speech and Audio Processing*, Vol. 10, No. 6, 2002.

[Yang & Koh$^+$ 93] H. Yang, S. Koh, P. Sivaprakasapillai: Phase Unwrapping Methods for Quadratic Phase Interpolation in Voiced Speech Synthesis. In *Proc. of the SICON/ICIE*, Singapore, Singapore, 1993.

[Ye & Young 03] H. Ye, S. Young: Perceptually Weighted Linear Transformations for Voice Conversion. In *Proc. of the Eurospeech*, Geneva, Switzerland, 2003.

[Ye & Young 04a] H. Ye, S. Young: High Quality Voice Morphing. In *Proc. of the ICASSP*, Montreal, Canada, 2004.

[Ye & Young 04b] H. Ye, S. Young: Voice Conversion for Unknown Speakers. In *Proc. of the ICSLP*, Jeju Island, South Korea, 2004.

[Young & Woodland$^+$ 93] S. Young, P. Woodland, W. Byrne: *The HTK Book, Version 1.5*. Cambridge University, Cambridge, UK, 1993.

[Zhu & Zhang$^+$ 02] W. Zhu, W. Zhang, Q. Shi, F. Chen, H. Li, X. Ma, L. Shen: Corpus Building for Data-Driven TTS Systems. In *Proc. of the IEEE Speech Synthesis Workshop*, Santa Monica, USA, 2002.

[Zissman 93] M. Zissman: Automatic Language Identification Using Gaussian Mixture and Hidden Markov Models. In *Proc. of the ICASSP*, Minneapolis, USA, 1993.

Publication List Related to this Thesis

[1] H. Bořil, P. Fousek, D. Sündermann, P. Červa, J. Žd'ánský: Lombard Speech Recognition: A Comparative Study. In *Proc. of the Czech-German Workshop*, Prague, Czech Republic, 2006.

[2] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, J. Hirschberg: Text-Independent Cross-Language Voice Conversion. In *Proc. of the Interspeech*, Pittsburgh, USA, 2006.

[3] D. Sündermann, J. Smrekar, H. Höge: Towards a Mathematical Proof of the Speech Alignment Paradox. In *Proc. of the AST*, Maribor, Slovenia, 2006.

[4] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, J. Hirschberg: TC-Star: Cross-Language Voice Conversion Revisited. In *Proc. of the TC-Star Workshop*, Barcelona, Spain, 2006.

[5] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, A. Black, S. Narayanan: Text-Independent Voice Conversion Based on Unit Selection. In *Proc. of the ICASSP*, Toulouse, France, 2006.

[6] D. Sündermann, H. Höge, T. Fingscheidt: Breaking a Paradox: Applying VTLN to Residuals. In *Proc. of the ITG*, Kiel, Germany, 2006.

[7] D. Sündermann. German Patent DE 10 2004 048 707 B3: Voice Conversion Method for a Speech Synthesis System, 2005.

[8] D. Sündermann, H. Höge, A. Bonafonte, H. Duxans: Residual Prediction. In *Proc. of the ISSPIT*, Athens, Greece, 2005.

[9] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, A. Black: Residual Prediction Based on Unit Selection. In *Proc. of the ASRU*, San Juan, Puerto Rico, 2005.

[10] D. Sündermann, G. Strecha, A. Bonafonte, H. Höge, H. Ney: Evaluation of VTLN-Based Voice Conversion for Embedded Speech Synthesis. In *Proc. of the Interspeech*, Lisbon, Portugal, 2005.

[11] D. Sündermann: A Language Resources Generation Toolbox for Speech Synthesis. In *Proc. of the AST*, Maribor, Slovenia, 2005.

[12] A. Bonafonte, H. Höge, I. Kiss, A. Moreno, D. Sündermann, U. Ziegenhain, J. Adell, P. Agüero, H. Duxans, D. Erro, J. Nurminen, J. Pérez, G. Strecha, M. Umbert, X. Wang. TC-STAR: TTS Progress Report. Technical report, 2005.

[13] A. Bonafonte, H. Höge, H. Tropf, A. Moreno, H. v. d. Heuvel, D. Sündermann, U. Ziegenhain, J. Pérez, I. Kiss. TC-Star: Specifications of Language Resources for Speech Synthesis. Technical report, 2005.

[14] D. Sündermann, A. Bonafonte, H. Ney, H. Höge: A Study on Residual Prediction Techniques for Voice Conversion. In *Proc. of the ICASSP*, Philadelphia, USA, 2005.

[15] D. Sündermann, A. Bonafonte, H. Duxans, H. Höge: TC-STAR: Evaluation Plan for Voice Conversion Technology. In *Proc. of the DAGA*, Munich, Germany, 2005.

[16] D. Sündermann: Voice Conversion: State-of-the-Art and Future Work. In *Proc. of the DAGA*, Munich, Germany, 2005.

[17] D. Sündermann, A. Bonafonte, H. Ney, H. Höge: Time Domain Vocal Tract Length Normalization. In *Proc. of the ISSPIT*, Rome, Italy, 2004.

[18] I. Esquerra, J. Adell, P. Agüero, A. Bonafonte, H. Duxans, A. Moreno, J. Pérez, D. Sündermann: Els Talps També Parlen. In *Proc. of the CELC*, Andorra la Vella, Andorra, 2004.

[19] D. Sündermann, A. Bonafonte, H. Ney, H. Höge: A First Step Towards Text-Independent Voice Conversion. In *Proc. of the ICSLP*, Jeju Island, South Korea, 2004.

[20] D. Sündermann, A. Bonafonte, H. Ney, H. Höge: Voice Conversion Using Exclusively Unaligned Training Data. In *Proc. of the ACL/SEPLN*, Barcelona, Spain, 2004.

[21] D. Sündermann, A. Bonafonte, H. Ney, H. Höge: Frequency Domain vs. Time Domain VTLN. In *Proc. of the AST*, Maribor, Slovenia, 2004.

[22] D. Sündermann, H. Ney: VTLN-Based Voice Conversion. In *Proc. of the ISSPIT*, Darmstadt, Germany, 2003.

[23] D. Sündermann, H. Ney, H. Höge: VTLN-Based Cross-Language Voice Conversion. In *Proc. of the ASRU*, Virgin Islands, USA, 2003.

[24] D. Sündermann, H. Ney: An Automatic Segmentation and Mapping Approach for Voice Conversion Parameter Training. In *Proc. of the AST*, Maribor, Slovenia, 2003.

Biographical Note

David Sündermann was born on August the 8$^{\text{th}}$, 1977 in Magdeburg (Germany). He received his Master of Science degree in Electrical Engineering (with Distinction) from the Dresden University of Technology in 2002 and started his Ph.D. project at the RWTH Aachen in 2003. In the same year, he received a Ph.D. Fellowship by Siemens Corporate Technology in Munich and relocated to Barcelona early in 2004, working at the Technical University of Catalonia. In 2005, he was visiting scientist at the University of Southern California in Los Angeles and, later that year, at the Columbia University in New York City. Since March 2007, David Sündermann is with SpeechCycle, Inc. in New York City. He has written 30 papers and holds a patent on voice conversion (for further information see `http://suendermann.com`).