

Diplomarbeit

Recognition of Regional Variants of German using Prosodic Features

Martin Hagmüller

durchgeführt am Institut für Elektronische Musik und Akustik (IEM)
in Zusammenarbeit mit Forschungszentrum Telekommunikation Wien (ftw)

Betreuer: Mag. Dipl.-Ing. Dr.techn. Univ.-Prof. Robert Höldrich
 Dipl.-Ing. Dr.techn. Univ.-Prof. Gernot Kubin
Begutachter: Mag. Dipl.-Ing. Dr.techn. Univ.-Prof. Robert Höldrich

Graz, im Mai 2001

Abstract

This thesis attempts to recognize national variants of German in Austria and Germany using only prosodic features.

An introduction to speech processing with special consideration of speech recognition and language identification is given. Fundamental frequency (F0) and Intensity of the speech signal are analyzed. Parameterization of F0 (Fujisaki, Intofit) is carried out and statistical features (Standard deviation, skewness, kurtosis, percentiles) are calculated from signals such as derivative and correlation of F0 and Intensity.

The features are evaluated using the t-test and a simple classification algorithm using combinations of up to three features. Combinations with Fujisaki parameters yield the best results with recognition rates of 72%

The small size of the data-corpus is a drawback of the study.

Zusammenfassung

Diese Diplomarbeit versucht allein aufgrund prosodischer Merkmale die nationalen Varietäten des Deutschen in Österreich und Deutschland zu unterscheiden.

Eine Einführung in die Sprachverarbeitung mit besonderer Beachtung von Spracherkennung und Sprachenidentifikation wird gegeben.

Es werden die Sprachgrundfrequenz (F0) und Sprachintensität analysiert. Methoden der Parametrisierung der F0 aus der Sprachsynthese werden verwendet (Fujisaki, Intofit). Statistische Merkmale (Standardabweichung, Skewness, Kurtosis, Perzentile) von Signalen wie Ableitung, Korrelation von F0 und Intensität werden berechnet.

Die Merkmale werden mit dem t-Test und mit einem einfachen Klassifikationsalgorithmus evaluiert. Es werden bis zu 3 Features kombiniert, wobei Kombinationen mit Fujisaki Parameter die besten Ergebnisse mit Trefferquoten von 72% erzielen.

Ein Schwachpunkt der Untersuchungen ist der kleine Datensatz von 90 Sätzen pro nationaler Varietät.

Acknowledgements:

Robert Höldrich

Gernot Kubin

Micha Baum

Alois Sontacchi

Hannes Pirker

Hans Grassegger

Rudolf Muhr

Wolfgang Ring

Elsa & Heinz Hagmüller

Daniela Hagmüller

CONTENTS

1. INTRODUCTION.....	7
1.1 Why would one want to identify regional variants of German?	7
1.2 Structure of this thesis.....	8
2. SOME BACKGROUND	9
2.1 Human speech production	9
2.2 Prosody	11
2.2.1 Acoustic correlates of prosody	12
2.2.2 Role of prosody in speech.....	13
2.2.3 Language.....	14
2.2.4 ToBI-Tones	14
2.3 Regional variants of German	15
2.3.1 Dialect and Prosody	16
2.4 Differences between Austrian and German.....	17
2.4.1 Lexical and grammatical differences	17
2.4.2 Pronunciation differences.....	17
2.4.3 Prosodic differences.....	18
2.5 Summary	18

3. SPEECH PROCESSING.....	19
3.1 Speech Analysis	19
3.2 Speech synthesis.....	24
3.3 Automatic Speech Recognition	26
3.3.1 A General Model for Speech Recognition	28
3.3.2 Parametric representation and Feature extraction.....	29
3.3.3 Pattern matching and time alignment	30
3.3.4 Networks for speech recognition	32
3.3.5 Language modeling	35
3.3.6 Summary	35
3.4 Speaker Recognition.....	36
3.4.1 Prosodic cues for speaker identification	37
4. LANGUAGE IDENTIFICATION	38
4.1 Useful Cues for LID.....	38
4.2 Multi-language speech corpora.....	39
4.3 Human performance	39
4.4 Recent Approaches	40
4.4.1 Single phone recognizer followed by language modeling (PRLM)	40
4.4.2 Parallel phone recognizers followed by a language model	42
4.4.3 Prosodic and Duration Approaches	44
4.5 Accent/Dialect Identification.....	47
4.6 Summary	49
5. FEATURE EXTRACTION	50
5.1 Fundamental Frequency Tracking	50
5.1.1 Theoretical Background.....	51
5.1.2 Pitch-post-processing	53
5.2 Parametric Description of the F0-contour	53
5.2.1 TILT-Analysis	53

Contents

5.2.2 Intofit	54
5.2.3 The Fujisaki - Model.....	57
5.2.4 LPC-Coefficients	59
5.2.5 Peaks and Intervals	62
5.3 Intensity	63
5.4 Statistical Features	63
5.5 Summary	64
6. EVALUATION.....	66
6.1 Classification algorithm	66
6.2 Feature Evaluation	67
6.2.1 T-Test on all Statistical Features	67
6.2.2 Feature combination	69
6.2.3 Alternative Evaluation with MLP.....	72
6.3 Discussion	72
7. SUMMARY AND DISCUSSION.....	74
7.1 Discussion and Outlook.....	74
7.2 Summary	75
REFERENCES	77

1. INTRODUCTION

1.1 Why would one want to identify regional variants of German?

German is a so-called pluricentric language with different national and regional variants [Muhr2000]. For speech dialog systems and speaker-independent speech recognition in general, special consideration of these differences is necessary. Speech recognition engines trained on a special regional variant have difficulties to recognize speech from a speaker with a different variant than the one the model is trained with. Another application is to choose a regional similar synthetic speaker in human-computer speech dialog systems, because a familiar variant of the language is generally perceived with more sympathy than a very unfamiliar one. This is a fact that might influence ones willingness to spend money.

As part of the SpeechDat project a corpus for telephone speech was recently acquired for Austrian [Baum 2000]. This opens the way to Austrian models for speech recognition. A preprocessor prior to the phone recognizer should decide which model of a variant of German is needed. This thesis is exploring how prosodic differences can be used to distinguish Austrian German¹ from the variant of the language spoken in Germany.

¹ In the text the variant of German spoken in the Republic of Austria will be referred to as Austrian and the variant of German spoken in the Federal Republic of Germany will be referred to as German

1.2 Structure of this thesis

After some background in human speech production and linguistic basics, specially covering prosody, a short glance at the different variants of the German language and specially at the differences between Austrian and German is taken (chapter 2). In chapter 3 a short summary of the field of speech processing with focus on human-computer interfaces is given. General principles of speech recognition including major techniques will be covered. A survey on Language Identification (LID), which reviews common approaches for LID with special consideration of dialect and accent recognition will be brought in chapter 4.

Then the experiments trying to distinguish Austrian and German as spoken in the Federal Republic of Germany will be discussed. Extraction of F0 will be explained and different methods for parameterization of F0 will be introduced in chapter 5. A section is dedicated to the calculation of statistical features from different signals derived from F0 and intensity (derivative, correlation, multiplication)

Chapter 6 covers the evaluation of the features using the t-test for the statistical features and a simple classification algorithm for feature combination.

A summary and outlook concludes the thesis in chapter 7.

2. SOME BACKGROUND

2.1 Human speech production

There are three main elements for human speech production (see Figure 2.3).

- a) Power source (lungs),
- b) Phonation (larynx),
- c) Articulation (oral and nasal cavities).

The source for most speech sounds is air expelled from the lungs through muscular action. During normal breathing the vocal folds are held apart forming a gap (glottis) to let the air flow freely and unless in case of some pathology, no or little audible sound is created.

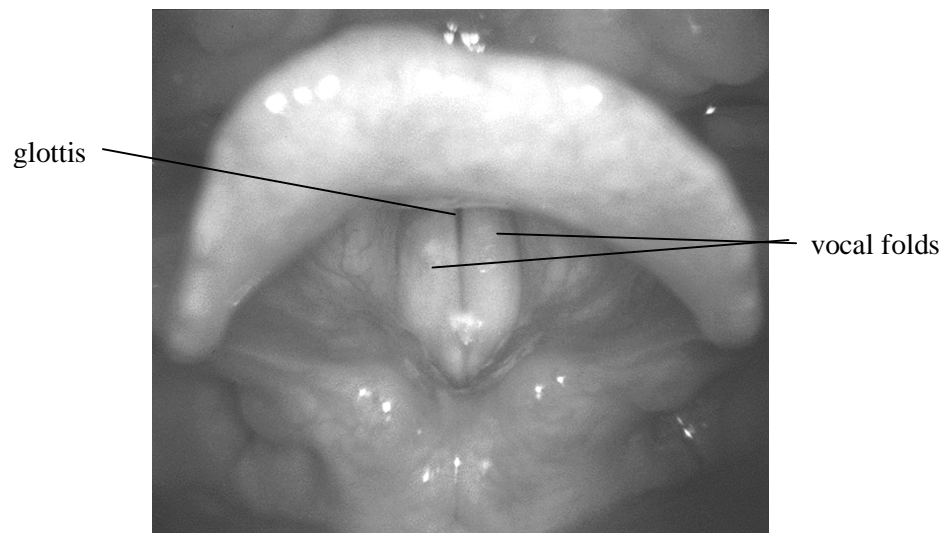


Figure 2.1: Vocal folds in phonation position (from [Putz+1998])

When speaking voiced sounds or singing the vocal folds (see Figure 2.1 and Figure 2.2) close the gap and higher pressure is built up in the lungs (subglottal pressure). This pressure

forces the vocal folds apart and thus lets the airflow through the glottis into the pharynx. The glottis forms a bottleneck for the air so the airflow speed is much higher than in the trachea. According to the Bernoulli law the air pressure between the vocal folds will be reduced. Consequently, the vocal folds get sucked together again and the cycle starts from the beginning. This vocal fold vibration is caused by both aerodynamics and the elasticity of muscle tissue, and is explained by the myoelastic aerodynamic theory of voicing. The fundamental frequency of this oscillation, which corresponds to the interruptions of the airflow, is determined by the length and mass of the vocal cords, and is controlled by its tension. For males, fundamental frequency is at about 80-200 Hz, for women at about 150-300 Hz. This frequency is not constant, but changes over the time of an utterance. This frequency pattern is called intonation.

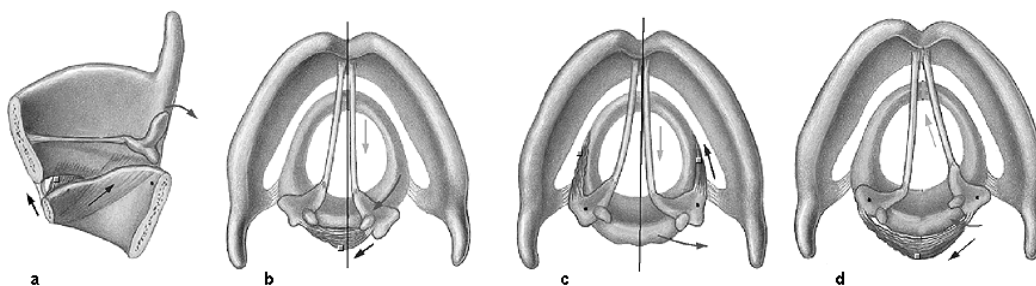


Figure 2.2: Movement of the vocal cords (right) (from [Putz+1998])

When whispering the vocal chords do not oscillate. They are close together, but build a triangular gap. The air flowing through the glottis causes noise that gives whispering its typical voiceless sound.

The acoustical signal, generated by the larynx, which is rich in harmonics in case of voicing and broadband noise for voiceless sounds, can be modified in the vocal tract by manipulation of the position of the velum, teeth, tongue, lips and jaw. Depending on the position of these parts, different resonance frequencies occur. The vocal tract can be seen as a filter for the source signal from the larynx. Those resonance frequencies are called formants and are essential for the intelligibility of speech because they do not change with the fundamental frequency of the glottis sound. For this reason we can recognize the vowel ‘a’ at every different pitch.

Non vowel like sounds are produced by narrowing the airflow passage (fricatives: e.g. ‘f, s’) or by blocking the flow altogether and then suddenly releasing it again (plosives: e.g. ‘p, t’).

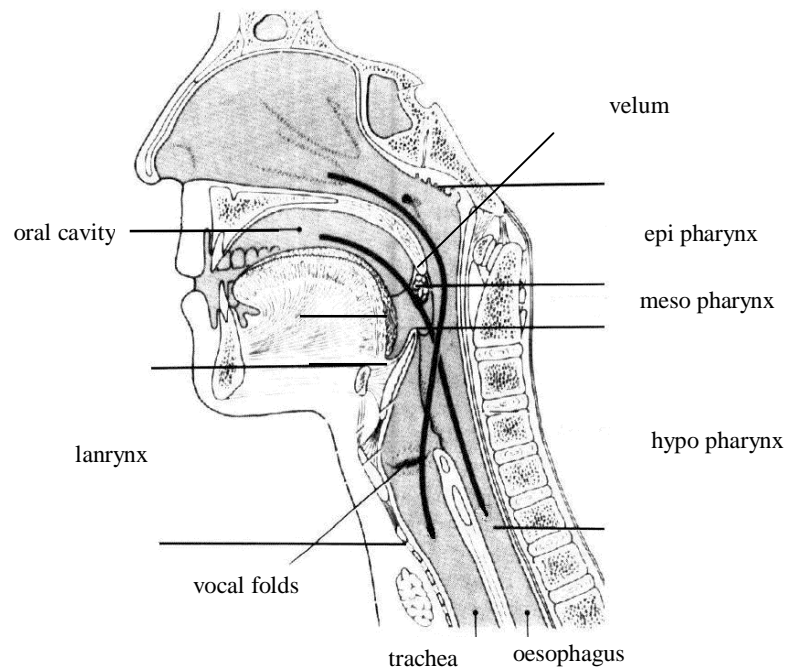


Figure 2.3: vocal tract

Combining those possibilities we are able to produce a theoretically infinite number of distinct sounds, though in every language only a certain amount of phonemes are used. Growing older we lose the spontaneous ability to articulate other sounds than those used in our mother tongue.

2.2 Prosody

From Merriam-Webster's dictionary online:

Main Entry: **prosody** [ˈprɒsədi], *noun*, plural: **-dies**

Etymology: Middle English, from Latin *prosodia* accent of a syllable, from Greek *prosOidia* song sung to instrumental music, accent, from *pros* in addition to + *Oide* song -- more at PROS-, ODE

Date: 15th century

1: the study of versification; *especially:* the systematic study of metrical structure

2: a particular system, theory, or style of versification

3: the rhythmic and intonational aspect of language

This section offers a short introduction to prosody (mostly according to [Neppert+1992] if not noted differently) and its applications to speech processing and specially dialect identification.

Prosody has to do with speech features that are not segmental like phones, its domain of interpretation is well beyond phone boundaries, concerning words, phrases and sentences. Therefore, prosodic characteristics are often called supra-segmentals.

Prosody describes the relationships of amplitude, duration and fundamental frequency of speech. It provides very different cues for syntactic information (segmentation, resolving ambiguity, conversational structure), emotions, stress and dialect.

[Neppert+1992] stated the following elements as parts of prosody:

- | | |
|--------------------------|------------------|
| 1. Fundamental frequency | 6. Tempo |
| 2. Duration | 7. Voice quality |
| 3. Intensity | 8. Musicality |
| 4. Timbre | 9. Emphasis |
| 5. Pauses | |

2.2.1 Acoustic correlates of prosody

Fundamental Frequency:

Intonation is the variation of the fundamental frequency (F0) in a sentence or more general utterance. It is the most important part of prosody, because most of the prosodic information lies in the pitch contour. Therefore, it is the feature, which is most referred to.

In almost every language most utterances have a downward trend, called declination after the first stressed word. It is normally reset at major syntactic boundaries. This effect is assumed to be correlated to the declining air pressure in the lungs.

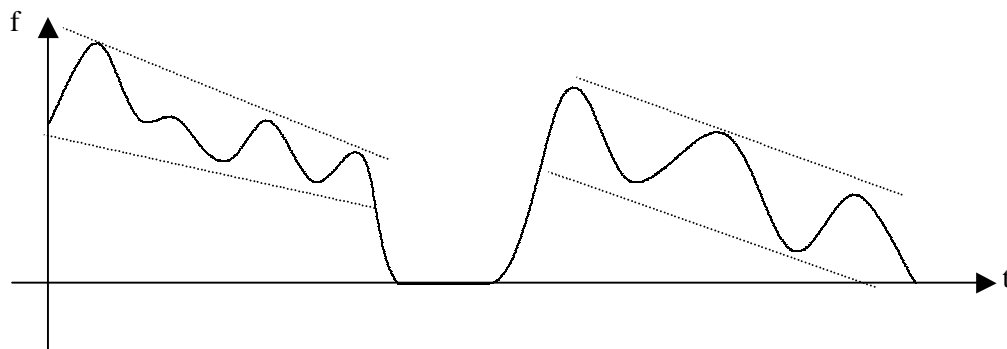


Figure 2.4: Declination of speech fundamental frequency

Lower F0 peaks at the end of a phrase are perceived as strong as higher peaks at the beginning of a phrase. This is because information lies in relative value (intervals) and slopes of the contour.

Fundamental Frequency is by far the most examined quality of prosody, possibly because it is easy to measure compared to other features mentioned above.

Intensity:

Intensity correlates to the loudness of the speech signal. Accented syllables normally have higher intensity than others. Whereas in Germanic languages intensity is important for placing accents, in Romance languages there are almost no changes in intensity at accents.

Duration:

Duration summarizes effects as speaking rate, phone and syllable duration and rhythm. It is an important feature, but difficult to obtain with automatic approaches.

2.2.2 Role of prosody in speech

The different roles of prosody in speech can be categorized in three groups (from [Mixdoff1997]).

linguistic (lexical, syntactic, semantic)	paralinguistic	non-linguistic
<ul style="list-style-type: none"> • sentence mode • discourse organization (focus) • segmentation (integration, delimitation) • disambiguation 	<ul style="list-style-type: none"> • speakers intention, attitude 	<ul style="list-style-type: none"> • age • gender • speakers's background (native language, dialect, sociolect) • emotional condition

Table 2-1: Roles of prosody in speech

The linguistic features refer to the way a message is formally coded and organized into units of a certain language. They correspond to the surface structure of the message on a still rather abstract level. The actual meaning of the message can often not be decoded without interpreting the underlying paralinguistic information. The question 'Are you tired?' is simply a request for being supplied with information on someone's psychological and physiological condition. If asked with a concerned undertone the message may be: 'Come on, you've been working so hard, you have to get yourself some sleep!' With an ironical undertone, it may mean: 'You lazy guy, you've been sleeping all day and still you're tired!'

Syntax: One basic function of prosody is to segment speech utterances into phrases and sentences, which help the listener to process speech in smaller units than the whole speech flow.

It also signals the function of a phrase or sentence such as a question or imperative and also the syntactic structure as main or subordinate clause. An important role of prosody is solving ambiguity.

- (1) 'Vielleicht. Am Montag bei mir. Paßt das?'
'Maybe. On Monday, at my place. Is that OK?'
- (2) 'Vielleicht am Montag. Bei mir paßt das.'

'Maybe on Monday. That's possible for me.'

Here two equal sets of words get a different meaning through prosodic variation, i.e. it is prosody that dissolves the ambiguity. This is used for speech recognition and understanding in the German VERBMOBIL [Nöth1997].

Accent - Stress: Another fundamental role of prosody is the placement of accents on words and phrases. Thus a sentence with the same words and different prosody placing the focus on different words can get completely different meanings.

Word accent can even have lexical importance. In German there are minimal pairs of words which are segmentally equivalent and only distinguished by the position of their word accent (e.g. 'umgehen' (to handle) vs. 'um'gehen' (to avoid) [Mixdorff1997]).

Highlighting stressed syllables against a background of unstressed syllables is a primary function of prosody. Stressed syllables are longer, more intense, and/or have F0 patterns that cause them to stand out against unstressed syllables.

Speaking style: Intonation is heavily influenced by the speaking style. Main categories are read, narrative and spontaneous speech. [Batliner1995] stated that spontaneous and non-spontaneous speech can be distinguished reasonably well by looking just at prosody.

Emotions: Different emotions and attitudes have big influence in prosody. Research is going on to detect emotions in speech with prosodic features [Waibel1996].

Personality: Prosody is a very personal characteristic [Mersdorf1997], so it is used for speaker identification [Carey1996]. Gender, age-group, health status and sometimes even vocational cues can be communicated via prosody [Neppert1992].

2.2.3 Language

Although many similar prosodic features can be found in different languages there are still differences between languages. [Thymé-Gobbel1996] used 220 features to distinguish English, Spanish, Japanese and Mandarin. These four languages were chosen since they represent the traditional categories of stress-timed, syllable-timed, mora-timed and tone languages.

2.2.4 ToBI-Tones

ToBI (Tones and Break Indices) -Tones is a system for transcribing and labeling the tones of a language. [Grice+1995] have adopted the original system, which was intended for American English, to German, calling it ToBIG, the Saarbrücken system. There are two main types of events, pitch accents and boundary tones. The system makes use of two tones, H and L. They can be grouped together into pitch accents and boundaries. What follows now is only a short introduction with the most important tone combinations.

There are two monotonal pitch accents:

H* 'peak accent' an apparent tone target on the accented syllable roughly in the upper 2/3 of the speakers range, often corresponding to a peak in F0.

L* 'low accent' an apparent tone target on the accented syllable low in the speaker's range, often corresponding to a dip in F0.

There are four bitonal ones ('*' indicates the tone of the accented syllable):

L*+H 'scooped accent' an apparent target low in the range followed by a peak high in the range.

L+H* 'rising peak accent' an apparent low target, followed by a high target on the accented syllable.

H+L* 'step-down to low' a preaccentual high or mid target followed by a target on the accented syllable which is clearly or very near at the bottom of the speakers range.

H+H* 'step-down to mid' a preaccentual high target followed by a target on the accented syllable which is in the middle of the range.

Boundaries can also occur in four bitonal combinations:

L-H% a low target roughly at the end of the accented word followed by a final rise to a level around the middle of the speaker's range.

L-L% an apparent target low in the range. It is not usual to discern two separate dips in F0 contour.

H-L% gives a high or mid plateau continuing at the same level as the most recent H tone in the phrase.

H-H% gives a plateau at the same level as the most recent H tone in the phrase, followed by a sharp rise at the end of the phrase.

A major disadvantage of the system is that the results are highly dependent on the person who labels the corpus.

2.3 Regional variants of German

German is as mentioned above (Section 1.1) a pluricentric language. [Muhr2000] considers Standard German as the common part of the three national variants, which are Austrian, German and Swiss (see Figure 2.5). Depending on the geographical area within the countries further regional variants can be observed.

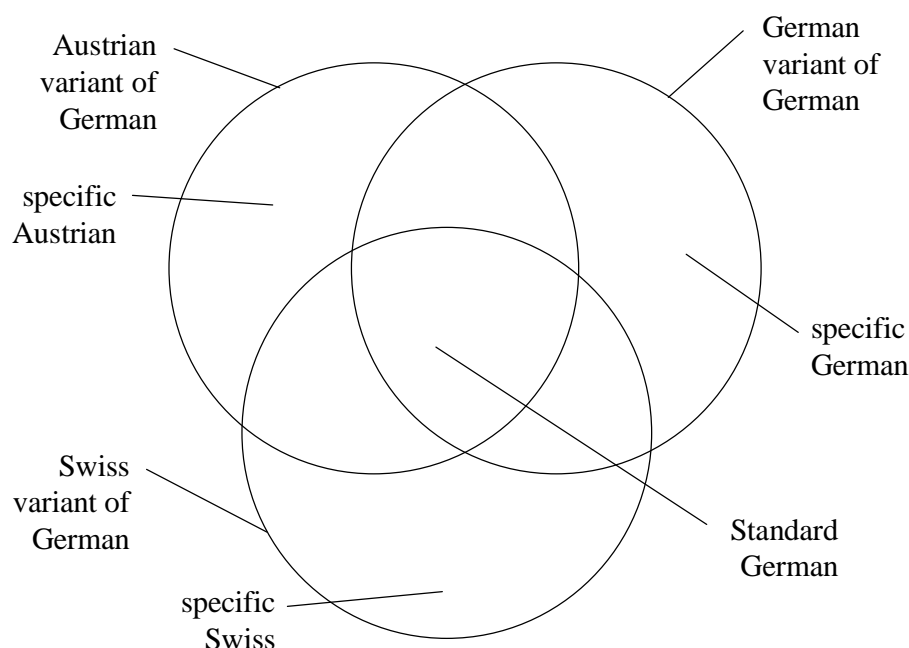


Figure 2.5: National variants of German

Concerning traditional dialect regions, in the Eastern part of Austria mainly the Middle- and Southern Bavarian is spoken, but the very western province Vorarlberg belongs to the Alemanic dialect. Between those regional variants there are major differences.

Additionally, one has to decide between an inner-standard that is spoken in a rather close communication form and an out-standard, which would be used in rather formal communication and with people speaking a different variant of German. The latter can be called the Austrian variant of standard German.

There is certain homogeneity because the same media, as television, radio programs and newspapers influences the whole Republic of Austria.

2.3.1 Regional variants and Prosody

[Gibbon1997] mentioned 10 regional standards associated with the cities Berlin, Hamburg, Hanover, Cologne, Frankfurt, Stuttgart, Munich, Leipzig, and Vienna for Austrian German, and Zurich for Swiss German sharing fundamentally the same prosodic properties with characteristic differences in the details. He states: 'In general, Southern dialects are associated with a right-displaced prominence peak; that is, the syllable perceived as being accented has low pitch, and a pitch rise, often followed by a peak, occurs on one of the following syllables (ToBI L*+H). In the standard pronunciation, the peak tends to occur on the accented syllable itself, though in some speech styles, such as telling stories to children, the right-displaced peak rhythm occurs.'

[Auer+1999] analyzed differences between the local dialects of Hamburg and Berlin by using natural discourses. They stated some prosodic patterns that made utterances typical for

a speaker of the city. This research is very context related and focuses on a few selected utterances.

[Schaeffler1999] performed perceptual experiments to find out whether there are prosodic cues in the speech signal of German dialects that are strong enough to identify the origin of the speaker in German. The speakers were classified in 7 dialect regions. Austrian and Bavarian speakers belonged to one dialect class. The utterances were delexicalized with a bandpass-filter between 70 Hz and 270 Hz. 10 speakers from each region were selected and the stimuli were 40 seconds of spontaneous monologue. Recognition rates for Bavarian/Austrian were well above chance level, but not clear. As a general statement it can be said, that there is a salient difference between northern and southern regions that could be used for dialect recognition by prosodic features.

2.4 Differences between Austrian and German

The listed differences do not claim completeness, but should give an overview of the dissimilarity between Austrian and German.

2.4.1 Lexical and grammatical differences

Most lexical differences between Austrian and German apply to cooking (Erdäpfel - Kartoffel, Paradeiser - Tomate, Karfiol - Blumenkohl, Faschiertes - Hackfleisch, Kren - Mehrrettich, ...) and public administration (Stellung - Musterung, Anrainer – Anlieger, ...). Other differences are different gender for the same word such as 'der Akt (m.) – die Akte (f.)' or different forming of plural 'die Erlässe – die Erlasse'. Another difference would be the different form of past perfect: 'ich bin gelegen – ich habe gelegen'. As mentioned above it is not so easy to specify a typical Austrian variant, so some of the examples would also be valid in southern Germany [Weiss1999]. Specific Austrian variants are coded in the 'Österreichisches Wörterbuch' [ÖWB1979]

2.4.2 Pronunciation differences

[Takahashi1996] investigated regional variants of German in Germany, Austria and Switzerland. He used two sources for his study. He started by using standard pronunciation dictionaries and their covering of regional variants. He then analyzed his own test speakers. To obtain Standard German, i.e. a correct pronunciation for the regional variant, he chose newsreaders and teachers of German as a foreign language.

In the following, some of the specific pronunciation features of the Standard German of Austria in contrast to the Standard German of Germany are documented (see also [Muhr2000]).

Vocals:

- The long open vowel [ɛ] becomes a closed [ə], such as ‘*erklärte*’, ‘*Auszählung*’, ‘*ordnungsgemäß*’.
- The short open vowels [ɪ,ɛ,ʏ] are pronounced closed [i,ə,ʏ], as in ‘*Mexiko*’, ‘*Ende*’, ‘*wird*’, ‘*dürfte*’.
- Articles are extremely shortened (e.g. ‘*die Amerikaner*’ [daməɐ̯i ka nɐ]).
- The suffix –er is pronounced extremely short [ɐ] (‘*linker, seiner*’)

Consonants:

- k and g become palatalized in front of a vowel or an l (‘*klar, keine, Gletscher*’)
- The final syllable –ig ist pronounced [ɪk] (e.g. ‘*gleichzeitig, vorläufig*’).
- The voiced [z] becomes a voiceless lenis [ʒ] or a voiceless [s] (e.g. ‘*seit*’ [ʒ], ‘*Somalia*’ [s]).
- A voiceless [t] at the end of a word or in a weak syllable becomes a voiceless lenis [d] (as in ‘*Zeit im Bild, sollte*’).
- There is no glottal stop in front of word and syllable boundaries (e.g. ‘*als einer*’ [al zɛɪnɐ], ‘*gab es*’ [ʁa bɛs]).

2.4.3 Prosodic differences

Concerning prosodic differences there is still much to explore, but maybe one of the most obvious dissimilarities is related to word accent. There is quite a list of words where the Austrian version has the accent on a different syllable as the German version, as in Kaffee A: [kɑ fə] vs. G: [kɑfə] (coffee) or Platin A: [ʔlɑ ti n] vs. G: [ʔlɑ ti n] (platinum)

2.5 Summary

We have discussed human speech production consisting of power source (lungs), phonation (larynx) and articulation (oral and nasal cavities).

We then explored the term prosody giving special consideration to its acoustic correlates and its role in speech. A short paragraph was dedicated to prosody and language.

The problem of national and regional variants of German was presented, focusing on the differences between Standard Austrian and Standard German. Those differences, though only briefly covered, show that appropriate modeling of Austrian pronunciation could help to improve speech recognition performance [Baum+2000].

3. SPEECH PROCESSING

Speech processing has become a broad field of research including many specific areas, which are sometimes closely related. This chapter gives an introduction to the main areas of speech processing. Most of the material is taken from [O'Shaughnessy2000] if not noted differently. The book is an excellent introduction to speech processing, provides extensive references, and therefore can be used for more in-depth studies.

The following topics are covered:

- Speech Analysis
- Speech Synthesis
- Automatic Speech Recognition
- Speaker Recognition
- Language Recognition
- Accent/Dialect Recognition

3.1 *Speech Analysis*

Speech Analysis is maybe the technical foundation for all other disciplines. It involves a transformation of a speech signal $s(n)$ into another signal, a set of signals or a set of parameters, with the objective of simplification and data reduction.

Speech analysis tries to extract relevant features while suppressing redundancy or irrelevance. Another goal of speech analysis is the finding of efficient representation of speech. Since speech analysis cannot be covered on a few pages, only methods that were actually used during this research are mentioned.

There are a few main assumptions about speech signals. First, it is usually assumed that the signal properties change relatively slow with time. This allows examination of speech with short-time windows presuming the parameters remain constant for the duration of the window. Usually speech sound is assumed to stay constant for at least 10ms. This opens the

subject of windowing signal. But this is not to be explored here. Background can be found in [Oppenheim+1995].

In the time-domain, analysis transforms a speech signal into a set of parameter signals, which usually vary much more slowly in time than the original signal.

The *Zero-crossing rate* (ZCR) provides very simple analysis in the time domain for spectral measures. In a signal $s(n)$ such as speech, a zero-crossing occurs when $s(n)=0$, i.e., the waveform crosses the time axis. For narrowband signals (e.g. sinusoids), ZCR is an accurate spectral measure.

The ZCR can be defined as

$$T[s(n)] = 0.5 |\text{sgn}(s(n)) - \text{sgn}(s(n-1))|,$$

where the algebraic sign of $s(n)$ is

$$\text{sgn}(s(n)) = \begin{cases} 1 & \text{for } s(n) \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

The ZCR can help in voicing decisions. Most energy in voiced speech is at low frequency, since the spectrum of voiced glottal excitation decays at about -12 dB/Oct. In unvoiced sounds, broadband noise excitation excites mostly higher frequencies, due to effectively shorter vocal tracts. While speech is not a narrow-band signal (and thus the sinusoid example does not hold), the ZCR correlates well with the average frequency of major energy concentration. Thus, high and low ZCR correspond to unvoiced and voiced speech, respectively.

Short-time energy or *amplitude* can help segment speech into smaller phonetic units, which can e.g. approximately correspond to syllables. The short-time energy is defined as

$$E[k] = \sum_{m=0}^{N-1} x[m]^2 w[n-m]$$

Short-time *autocorrelation* gives information about energy and periodicity of the signal. It is used for F0 determination and Linear Prediction.

$$R_n[k] = \sum_{m=-\infty}^{\infty} s[m]w[n-m]s[m-k]w[n-m+k]$$

Frequency domain parameters provide the most useful parameters for speech processing. The basic model of speech production is a noisy or periodic waveform that excites a vocal tract filter. This corresponds well to separate spectral models for the excitation and for the vocal tract (see Figure 3.1).

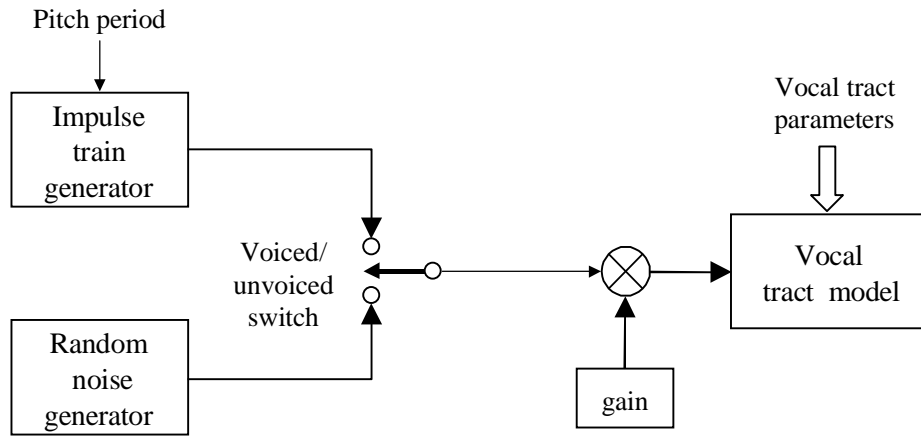


Figure 3.1: source-filter speech model

Human hearing appears to pay much more attention to spectral aspects of speech than to phase or timing aspects. Thus, spectral speech analysis generally receives much more attention.

Short-Time (discrete) Fourier transform (STFT) applies the discrete Fourier transform (DFT) to successive windows:

$$S[k] = \sum_{m=0}^{N-1} s[m] e^{-j2\pi km/N} w[n-m]$$

The choice of N (window length) is crucial for STFT. Low values for N give poor frequency resolution, but good time resolution. Large N, on the other hand, gives poor time resolution and good frequency resolution. As an optical speech analysis tool, the spectrogram provides a three dimensional representation of speech utterances using the STFT. For speech analysis there are two main representations. Wideband analysis displays individual pitch periods as vertical striations corresponding to the large amplitude at vocal cords closure. It smoothes the harmonic amplitudes under each formant across a range of 300Hz, displaying a band of darkness for each formant. The center of each band is a good estimate for the formant frequency.

Narrowband spectrograms display separate harmonics instead of pitch periods.

They can help to analyze F0 and vocal tract excitation (see Figure 3.2).

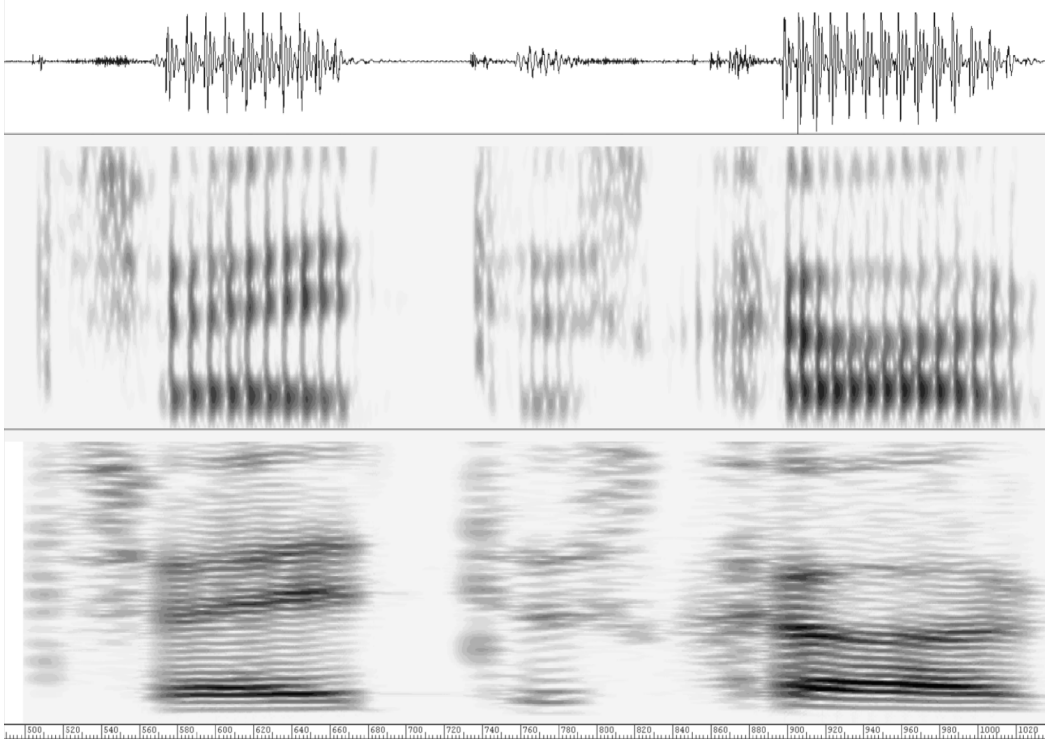


Figure 3.2: Wideband and narrowband spectrograms of a sentence

Linear Predictive Coding (LPC) is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate. It provides accurate estimates of speech parameters, and is relatively efficient for computation. The underlying assumption of LPC is a speech model as in Figure 3.1, which is a source-filter model with an excitation signal of either an impulse train or random noise. This excitation signal is then filtered by the vocal tract transfer function.

$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{1 - \sum_{k=1}^P a_k z^{-k}}$$

The predicted signal is calculated using an FIR filter:

$$\tilde{x}[n] = \sum_{k=1}^P \alpha_k x[n-k]$$

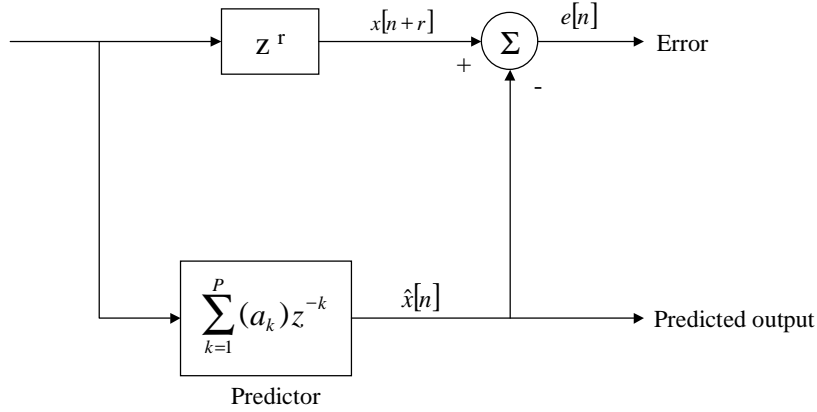


Figure 3.3: Block diagram for linear prediction

Linear predictive analysis is a technique aimed at finding the set of prediction coefficients $\{\alpha_k\}$ that minimize the mean-squared prediction error between a signal $x[n]$ and a predicted signal based on a linear combination of past samples; that is

$$\langle (f[n])^2 \rangle = \left\langle \left(x[n] - \sum_{k=1}^P \alpha_k x[n-k] \right)^2 \right\rangle,$$

where $\langle \cdot \rangle$ represents averaging over a finite range of values of n . Usually, the autocorrelation method is used to find the optimum predictor coefficients $\{\alpha_k\}$. Small segments of speech (usually approx. 10ms) are used to ensure that the signal doesn't change significantly during analysis.

If the speech signal $x[n]$ is filtered by an inverse or predictor filter (the inverse of an all-pole $V(z)$)

$$A(z) = 1 - \sum_{k=1}^P \alpha_k z^{-k}$$

the output $e[n]$ is an error signal

$$f[n] = x[n] - \tilde{x}[n] = x[n] - \sum_{k=1}^P \alpha_k x[n-k] \equiv Ge[n].$$

This signal (called residual) as input will synthesize the original signal perfectly.

$$x[n] = \sum_{k=1}^P \alpha_k x[n-k] + Ge[n]$$

From $f[n]$ it can be determined whether the signal is voice or unvoiced and, if voiced, the fundamental frequency. This is how basic LPC encodes the residual. This encoding, however, causes quality loss, because the predicted filter is never ideal.

Problems arise when the excitation is not purely voice or unvoiced, but something in between. Various efforts have been made to code the error signal. For example, CELP (Code

Excited Linear Prediction) is a version where the residual signal is matched to entries in a codebook and the entries are used for coding.

Determining the *fundamental frequency* (F_0) is important in many speech applications. It is the primary acoustic cue to intonation and stress in speech. Most low-rate voice coders require accurate F_0 estimation for good reconstructed speech. F_0 patterns are useful in speaker recognition and synthesis. Time domain F_0 detectors have three components: a preprocessor (to filter and simplify the signal via data reduction), a basic F_0 extractor (to locate pitch epochs in the waveform), and a postprocessor (to locate errors).

Frequency-domain methods for F_0 estimation involve correlation, maximum likelihood, and other spectral techniques where speech is examined over a short-term window. In Section 5.1 an auto-correlation F_0 estimation method will be discussed.

3.2 Speech synthesis

Text-to-speech synthesis (TTS) is the automatic generation of a speech signal, starting from a textual (or conceptual) input and using previously analyzed digital speech data.

Two main steps are required, first, the linguistic analysis which transfers a text to a rather phone orientated description including prosodic information and second, the actual speech wave generation.

The text analysis consists of different stages (see Figure 3.4). The **preprocessor** does the normalization and segmentation of the text. Numbers, special symbols, abbreviations, acronyms, control characters, etc. have to be handled. Punctuation is examined as a cue for sentence end detection.

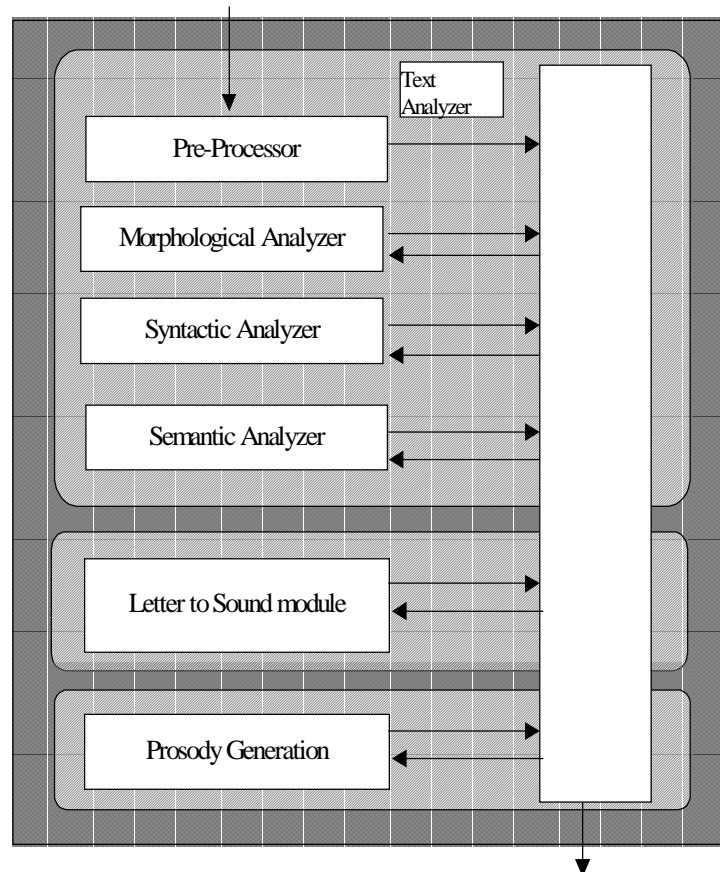


Figure 3.4: Text analysis for speech synthesis (from [Dutoit1997])

The **morphological analyzer** decomposes all words into their elementary units (Morphemes) to be able to have a dictionary at a reasonable size.

The **syntactic analysis** helps to identify the part-of-speech of every word and can then structure the text or sentence to be able to extract prosodic information.

Semantic information would help to solve most of the current problems of text interpretation, but so far and probably for the near future there is no efficient tool to handle this task.

The **Letter-to-sound module** then uses the already obtained information for the phonetization of the text.

The most difficult part is to generate proper prosody for unrestricted text. This is one of the main reasons that synthetic speech still doesn't sound natural.

Speech-wave production methods can be broadly divided into two categories. – those which predominantly model the speech signal and those which predominantly concatenate prerecorded speech signal.

There are two prominent members of modeling speech – ‘articulatory synthesizers’ and ‘formant synthesizers’. Both generate a synthetic speech signal purely from parametric information, which drive an abstract model of production.

Articulatory synthesizers attempt to produce a synthetic speech by modeling the characteristics of the vocal tract and the speech articulators. There is still not enough information about the exact mechanism of the speech process, partly because there are no appropriate models of the speech production available. Research is going on, but no commercial use is made so far, because excessive computational power is necessary for real time applications.

Formant synthesis looks at the acoustical properties of speech and tries to model them. The underlying model is most of the times a source-filter-approach (Figure 3.1).

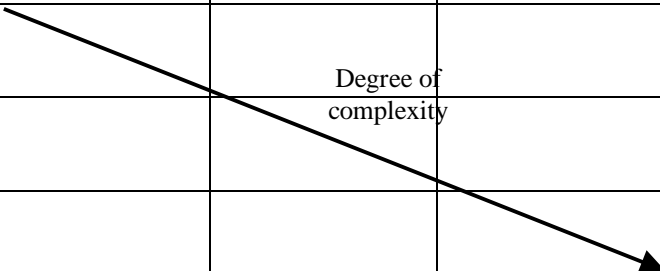
The second approach is to encode pieces of natural speech and put them together as needed. The advantage is that no complicated model of the speech signal is necessary and with the possibility of advanced techniques for modifying the signal and cheap computer memory **concatenative synthesis** is now the common approach for commercial TTS Systems.

3.3 Automatic Speech Recognition

Automatic speech recognition (ASR) has been much more difficult to achieve than TTS. When trying to let a computer recognize human speech several problems arise. Among them are variability (time, speaker, etc.), vocabulary, continuous speech, etc.

In theory, ASR could be as simple as a large dictionary where each entry is a digitized stored waveform labeled with a text pronunciation. Given an input utterance, the system would only search the dictionary for the closest match and find the corresponding text from a lookup table. Whereas this approach works for speaker dependent discrete utterance, small vocabulary application, for more complex systems this procedure is not feasible due to the immense complexity. Different applications yield very different complexity of the ASR system.

	Discrete utterance	Connected speech	Continuous speech
Speaker Dependent			
Multi Speaker			
Speaker Independent			



Degree of complexity

Table 3-1: Complexity of different recognition tasks (From [Morgan+1991])

The many sources of **variability** in human speech are a major reason for this complexity. All recognizer are influenced by environmental variability due to background and channel noise

(specially for phone applications) Speaker-dependent recognizers have the least variability, but still every word is not pronounced the same all the time (Intra-speaker variability). It still takes training time to adapt the system to a specific speaker. On the other hand, speaker-independent systems have to cope with inter-speaker variability. Here many factors have to be taken into account such as gender, age, accent or dialect, style of speaking and different anatomy of the vocal tract. Those sources of variability are the key problems of ASR.

Excursion on Problems of Dialect and Accent in ASR

With the example of German, the different influences of dialectal coloring on ASR will be considered. Dialectal differences are a major problem in ASR. Depending on the region realization of phonemes and hence, pronunciation of words differ a lot. Not only acoustic deviations, but also lexical differences make it necessary to include many dialectal regions into an ASR system. [König1981] described several dialectal subdivisions and boundaries. Foreign accents may rise the problem that phonemes of the target language do not exist in the original language, so people might not be able to pronounce those phonemes correctly

To keep the complexity, which is rising with utterance length, at a low level, segmenting speech into smaller units such as words syllables or phones is an important task. However, it is hard to find reliable cues for this task. Sudden large changes in speech spectrum or amplitude help to estimate unit boundaries. For example, silence can be between words, but it can also be before plosives or at glottal stops. Correct endpoint detection helps to improve error rates and keep down computational costs.

For performance evaluation, error rates (e.g. the percentage of words not correctly recognized of those spoken) or accuracy (the percentage of correctly identified words) are used. Cost, speed and the likelihood of an input being rejected are other important factors.

Speaker dependent isolated word recognition with a small vocabulary often reaches accuracy of >99%, but may fall to 90-95% for speaker independent connected speech applications.

One crucial tool for speech recognition is the use of databases of speech labeled with textual transcriptions as training data and as evaluation tools. For German, the Bavarian Archive for Speech Signals (BAS) [Schiel+1999] provides different corpora including RGV1 (A Database for Regional Variants of German) [Burger+1998] and the SpeechDat project which includes an Austrian corpus [Baum+2000].

The crucial point of speech recognition is the problem of pattern recognition. An utterance has to be compared with reference data obtained through training. This data can be divided into two main approaches. One can view ASR from either a *cognitive view*, which is a knowledge based or expert system, or an *information theoretic view*. The first one tries to model through finding relationships between speech signals and their corresponding text

messages and postulating phonetic rules to explain the phenomena. It is, however, very difficult to model the complexity of speech in a knowledge-based system. The information theoretic view tries to describe speech using information derived through statistic analysis. It uses statistical models that maximize the likelihood of choosing the correct symbols corresponding to the input.

3.3.1 A General Model for Speech Recognition

A general speech recognition model is illustrated in Figure 3.5. The major components of this model include:

- Preprocessing to normalize the speech signal
- Parameterization and feature extraction to identify the key components of a parametric representation and eliminating redundant information.
- Time alignment and pattern matching algorithm for performing word detection
- Language processing to select a linguistically valid word string

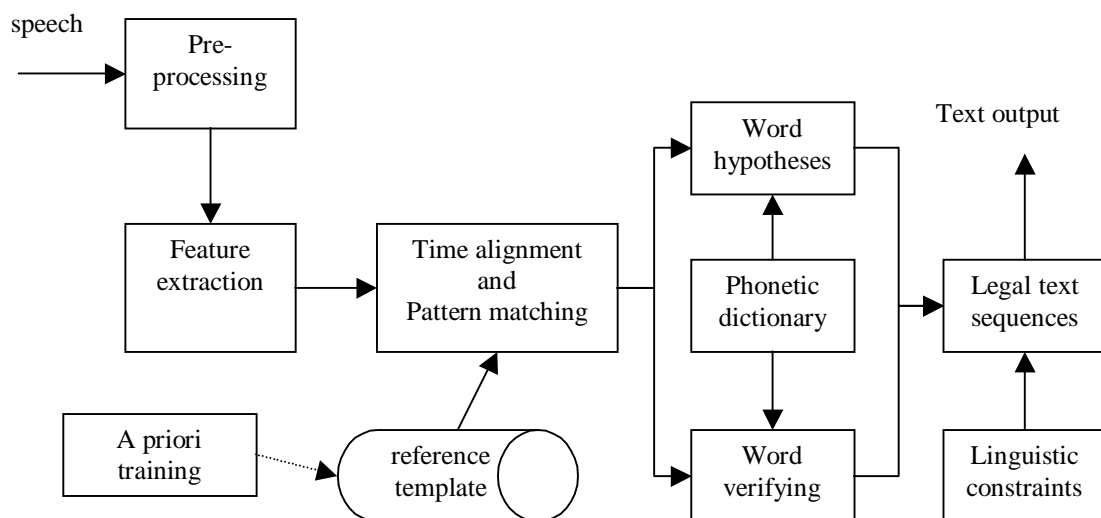


Figure 3.5: General speech recognition model

a) Preprocessing

First, the speech utterances have to be normalized for example with automatic gain control (AGC) to reduce the influence of different recording conditions (for example through different distance to the microphone) or transmission channels. It has to be applied with long time constants to preserve prosodic information in amplitude changes. Normalization of temporal variations is not done at this level.

b) Parametric representation and Feature extraction

The parameterization of the speech signal has the goal of efficient data reduction without losing information relevant for ASR.

c) Pattern matching and time alignment

At the heart of ASR lies the measurement of similarity between two windowed speech patterns, i.e. the representation of a frame of the input speech and a frame from a set of reference patterns or models (obtained during training). The comparison or evaluation involves finding the best match in terms of a distance between templates or deciding which reference model is the most likely.

d) Linguistic Evaluation

Most speech corresponds to texts, which follow linguistic rules (e.g. lexical, syntactic, semantic constraints). Exploiting these rules is crucial for ASR performance.

3.3.2 Parametric representation and Feature extraction

Usually successive frames of 10ms distance (see Chapter 3.1) are parameterized. and a source-filter model as in Figure 3.1 is used for the parameterization. But very often the excitation parameters such as the voicing decision, amplitude and pitch are ignored in ASR, though recent research tries to employ those prosodic parameters to aid recognition at a linguistic level (Figure 3.6). [Nöth+1997] use a combination of prosodic information such as fundamental frequency and energy, and a word hypothesis generator following a phoneme recognizer to gain positions of accents and sentence boundaries. [Strom1995] uses only Energy and F0 of a speech signal to place the accents and sentence boundaries. Both are part of VERBMOBIL [Walster1996] a multidisciplinary research project by several research institutions in Germany. Its goal is to develop a tool for machine translation of spoken language from German into English and Japanese.

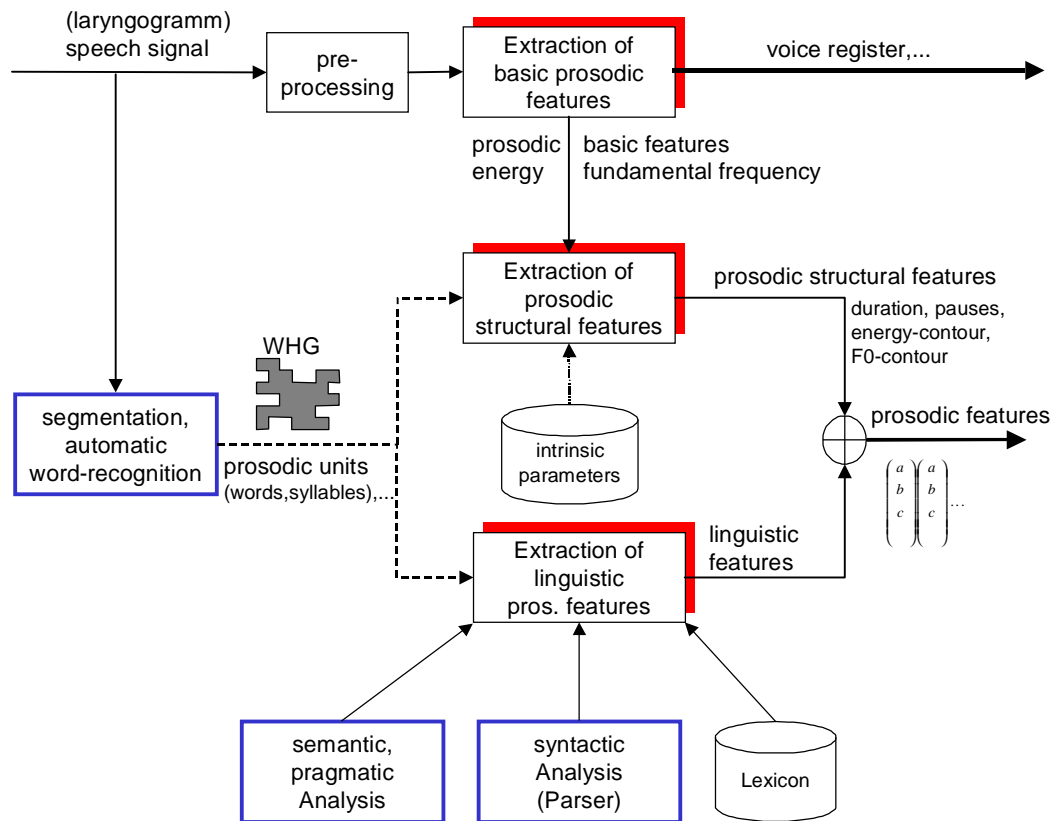


Figure 3.6: Incorporation of prosody for speech recognition (from [Nöth1997])

The spectral envelope provides the primary ASR parameters. The most common ASR parameters are mel-based cepstral coefficients, but LPC analysis, energies from a channel vocoder, reduced forms of DFT and zero-crossing rates in bandpass channels are other examples. They all attempt to capture in about 10 parameters enough spectral information to identify spoken phones.

Now, these parameters can be reduced to features in order to decrease redundancies of the parameters. Such features may be subdivided into acoustic and phonetic features, depending of the degree of data reduction. Phonetic features have a discrete range and assign sound to linguistic categories, e.g. voiced or fricative; they represent major data reduction thus leading toward a phonemic decision. Acoustic features (e.g. formants, F0) represent an intermediate step between parameters and phonetic features.

Features are fewer in number than parameters and therefore potentially more efficient for ASR; they are speech specific and require classification that can be erroneous.

3.3.3 Pattern matching and time alignment

Each parametric (or feature) pattern for a frame of speech can be viewed as an N-dimensional vector, having N parameters/frame. If the parameters are well chosen, then separate regions can be established in the N-space for each segment.

The similarity between two patterns is often expressed via a distance, measuring how close the patterns are in N-space. Another popular method is statistical, where the reference models store probability density functions (PDF) and similarity is judged in terms of the likelihood of the test pattern for each PDF. To handle multiframe utterances (i.e. all practical cases), local (frame) distance measures typically sum up to yield a global (utterance) distance. The reference pattern yielding the smallest distance or highest probability is usually chosen for the ASR output.

Euclidean and Mahalanobis distances:

In ASR involving templates, each unknown test utterance is converted to an N-parameter test template, to be compared with reference templates to find the closest match. The similarity of two templates is inversely proportional to the distance in N-space between points corresponding to the templates. The most common distance measure is the Euclidean distance (or L2-norm):

$$d_2(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T (\vec{x} - \vec{y})} = \sqrt{\sum_{i=1}^N (x_i - y_i)^2}$$

Another common speech distance is the Mahalanobis or covariance-weighted distance,

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T W^{-1} (\vec{x} - \vec{y})},$$

where W is a positive-definite matrix that allows different weighting for individual parameters depending on their utility in identifying the speech segments in N-space.

Despite the advantages of the Mahalanobis distance in weighing properly, ASR often uses the Euclidean distance or a LPC distance, because it is difficult to reliably estimate W from limited training data. Moreover the latter two distances require only N multiplications for an N -dimensional parameter vector vs. N^2 multiplications with the Mahalanobis distance.

Stochastic similarity measures:

The Mahalanobis distance has origins in statistical decision theory. If each utterance of a word represents a point in N-space, the many possible pronunciations of that word describe a multivariate PDF in N-space. Assuming ASR among equally likely words and maximum likelihood as the decision criterion, Bayes' rule specifies choosing the word whose PDF is most likely to match the test utterance. Because of the difficulty of estimating accurate PDFs from a small amount of training data, many systems assume a parametric form of PDF, e.g. Gaussian, which can be simply and fully described by a mean vector μ and a covariance matrix W . Since ASR parameters often have unimodal distributions resembling Gaussians, the assumption can be reasonable.

The distance measures in the previous paragraph are general and can be used with many sets of parameters such as LPC or Cepstral parameters.

Dynamic time warping

The dynamic time warping (DTW) procedure combines alignment and distance computation in one dynamic programming procedure. DTW finds an optimal path through a network of possibilities in comparing two multiframe templates, using the Bellman optimality principle. DWT aligns a test template as a whole with each reference template by finding a time warping that minimizes the total distance measure, which sums the individual frame distances in the template.

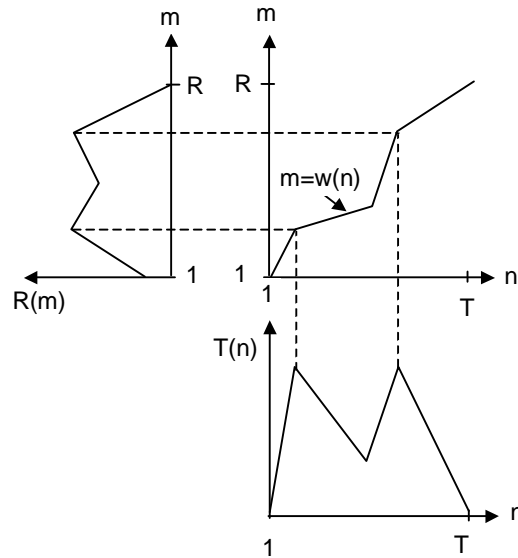


Figure 3.7: Dynamic time warping from [O'Shaughnessy2000]

3.3.4 Networks for speech recognition

Networks in ASR employ a rather statistical representation of acoustic information. Let us have isolated word recognition (IWR) as an example: model each word with a succession of phonetic states i (corresponding roughly to phones), linked by transitions specified by likelihoods a_{ij} . This probability of a phonetic segment j following segment i governs the transition between the states representing those two sounds. Consider *pass* as an example word, where states for $[\nu]$ closure (silence), $[\nu]$ burst, $[\nu \cdot \text{æ}]$ aspiration $[\text{æ}]$, and $[\text{s}]$ might be chosen via coarse segmentation or other technique. To allow for the chance that the $[\nu]$ burst and/or aspiration may be missing the a_{ij} may vary considerably, they typically correspond to the frequency of actual transitions in the training data.

For IWR each input utterance is evaluated by each word network, to find the network most likely to have generated the word. Instead of searching a DTW space for a path of minimal distance, each network is searched for the path maximizing the product of all transition probabilities between states corresponding to the test utterance.

Popular approaches are Hidden Markov Models [Rabiner1989] or Neural Networks [Morgan1991].

Hidden Markov Models

Rabiner presents an excellent introduction to Hidden Markov Models (HMMs) in his tutorial [Rabiner1989].

The key assumption of the statistical approach to speech recognition is that speech can be modeled statistically during an automatic process. By examining an ensemble of training speech data, a probabilistic that characterizes the entire ensemble is created. The resulting model, which represents each speech unit (word or sub unit), is more powerful and general than a template.

In the HMM formalism, speech is assumed to be a two-stage probabilistic process. In the first part of the two-stage process, speech is modeled as a sequence of transitions through states. The states are not themselves directly observable (hidden), but are manifest by observations, or features. Second, the observations in any state are not deterministic, but are specified by a probabilistic density function over the space of features. The power and flexibility of the statistical approach comes from this two-stage modeling procedure.

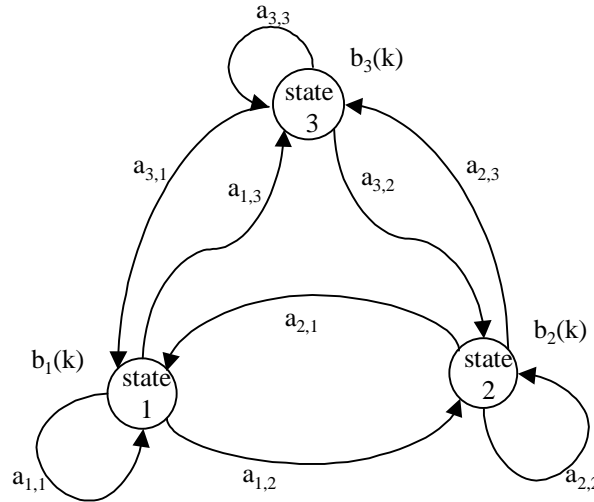


Figure 3.8: 3-state Hidden Markov Model

As shown in Figure 3.8, a HMM is characterized by the following:

1. N , the number of states in the model. We denote the individual states as $S = \{S_1, S_2, \dots, S_N\}$, and the state time t as q_t .
2. M , the number of distinct observations per state. We denote the individual symbols as $V = \{v_1, v_2, \dots, v_M\}$.
3. The state transition probability distribution $A = \{a_{ij}\}$ where

$$a_{ij} = P[q_{t+1} = S_i | q_t = S_j] \quad 1 \leq i, j \leq N$$

For the special case where any state can reach any other state in a single step, we have $a_{i,j} > 0$ for all i,j . This is called an ergodic HMM. In ASR left-right HMMs are very common because time depended properties can be modeled very well. Only state transitions from left to right are allowed (Figure 3.9).

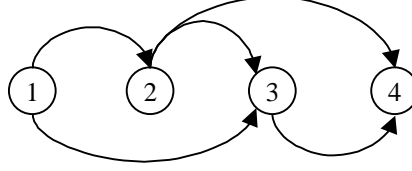


Figure 3.9: Left-right HMM

4. The observation symbol probability distribution in state j , $B=\{b_j(k)\}$, where

$$B_j(k) = P[v_k \text{ at } t \mid q_t = S_j] \quad \begin{matrix} 1 \leq j \leq N \\ 1 \leq k \leq M \end{matrix}$$

5. The initial state distribution $\pi = \{\pi_i\}$, where

$$\pi_i = P[q_1 = S_i] \quad 1 \leq i \leq N$$

Given appropriate values of N , M , A , B and π , the HMM can be used as a generator to give an observation sequence:

$$O = O_1 O_2 \dots O_T$$

For convenience usually a compact notation for a HMM is used:

$$\lambda = (A, B, \pi)$$

Given a HMM there are three basic problems which have to be solved should the model be useful in real-world applications:

- Problem 1: Given the observation sequence $O = O_1 O_2 \dots O_T$, and a model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O|\lambda)$, the probability of the observation sequence, given the model. \rightarrow *The Baum-Welch forward-backward algorithm may be used to find the probability $P(O|M_k)$ of generating the observation sequence O from the model M_k*
- Problem 2: Given the observation sequence $O = O_1 O_2 \dots O_T$, and a model λ , how do we choose a corresponding state sequence $Q = q_1, q_2, \dots, q_T$ which is optimal in some meaningful sense (i.e., best ‘explains’ the observations)? \rightarrow *There is a well know Viterbi dynamic programming solution.*
- Problem 3: How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$ \rightarrow *The model parameters are found by an iterative procedure known as Baum–Welch re-estimation. An initial HMM is assumed, and the Baum-Welch forward-backward algorithm is carried out to find the state occupation probabilities as a function of time.*

HMMs can be used to model phones, words or larger speech units. A phone HMM might use three states to represent to, in order, an initial transition spectrum from a prior phone, a spectrum from the phone's presumed steady state, and a final transition spectrum to a following phone.

3.3.5 Language modeling

Language models (LM) appear on a lexical (or phonotactic), a syntactic and possibly on a semantic level.

Since the phonemic composition of words in most languages is highly restricted (e.g., the sequences /tz/ and /sd/ are illegal in English syllables), a word level language model is applied. Normally given a history of prior (recognized) words in an utterance, the number of words P that an ASR must consider as possibly coming next is much smaller than the vocabulary size V . P is called the perplexity of a language model. LMs are stochastic descriptions of text, usually involving likelihoods of local sequences of N consecutive words in training texts.

n-gram Models

Typically, N -gram models estimate the likelihood of each word, given the context of the preceding $N-1$ words, e.g. bigram models use statistics of word pairs and trigrams model word triplets. Unigrams are simply prior likelihoods for each word, independent of context. These probabilities are determined by analysis of large amounts of text, and are incorporated into a Markov language model.

3.3.6 Summary

We have covered some basic principles of speech recognition. Even though only some topics were covered, the complexity of the task is obvious.

Most current ASR uses statistical pattern recognition, applying general models as structures to incorporate knowledge about speech in terms of reference models. The parameters of the models are estimated during a training procedure, in which speakers utter words or sentences, which may be repeated during actual ASR.

Alternative approaches such as cognitive methods, which incorporate knowledge on human speech production and perception, decreasing hardware (e.g. memory) cost will improve future ASR performance.

3.4 Speaker Recognition

There are two main Speaker recognition applications:

- Verifying a persons identity prior to admission to a secure facility or to a transaction over the telephone
- Association a person with a voice, e.g. in audio-conferences

Other applications could include identification of the persons gender, emotions in speech [Waibel1996], accent or dialect of a speaker and the language being spoken. The latter will be covered extensively in Chapter 4.

There are two main areas in speaker recognition, first *automatic speaker verification* (ASV) and second *automatic speaker identification* (ASI).² ASV only has to evaluate the test pattern with one reference model and a binary decision whether the test speech matches the model of the claimant has to be made. ASI, on the other hand, requires choosing which of N known voices best matches a test voice.

When focusing on the identity of a speaker there are three sources of variation among speakers: differences in vocal cords and vocal tract shape, differences in speaking style (including variations in both target positions for phonemes and dynamic aspects of coarticulation such as speaking rate), and differences in what speakers choose to say.

Two classes of error occur: *false acceptance* when the system incorrectly accepts an impostor during ASV or identifies a wrong person during ASI, and *false rejections*, when the system rejects a true claimant in ASV or incorrectly finds no match in ASI.

Analysis techniques are similar for speech and speaker recognition since both involve pattern recognition of speech signals, but in speaker recognition only one decision is necessary compared to ASR where decisions are made for every phone or word. Templates or models are not focused on text, but on speakers. The considered features include prosodic properties, LPC and cepstral coefficients.

Main approaches are either utilizing features using long time averages (e.g. means and variances of F0, amplitude or LPC coefficients) or comparing specific sound with a test template of e.g. phones. Two other categories are text dependent or text independent solutions.

² ASV/I will be used for discussions applying both to ASV and ASI

3.4.1 Prosodic cues for speaker identification

[Carey+1996] have utilized prosodic features based on pitch and energy contour for speaker identification. Gender was identified with 98% accuracy using the mean pitch parameter alone. Consequently, mean pitch was also used for identification of unknown speakers.

Additionally they used the first four statistics, mean, variance, skewness, kurtosis, of the pitch and energy and their first two derivatives. The mean and variance of the length of the voiced speech segments were added to these. Those features were tested to draw the seven best ones for classification using Linear Discriminant Analysis.

They then combined the prosodic system with a system using spectral envelope parameters employing a filterbank with 19 filters. The log power outputs were transformed into twelve cepstral coefficients and twelve delta cepstral coefficients at a frame rate of 10 ms. Hidden Markov Models were used for classification. The spectral envelope yielded better results than the prosodic features. However, the latter were much more robust to signal degradation than the spectral envelope.

[Waibel1996] used two sets of features to recognize the emotional state of a speaker. The first set consists of 7 global statistics of the pitch signal such as mean, standard deviation, minimum, maximum, range, slope and speaking rate.

The pitch contour for the second feature set was smoothed using piecewise cubic splines. The derived features were statistics related to rhythm, the smoothed pitch signal and its derivative, individual voiced parts and slopes.

4. LANGUAGE IDENTIFICATION

Language Identification (LID) can be seen as a subset of the field of speaker recognition, because it gains personal information about the speaker. A more specialized application is the problem of dialect or accent identification.

Applications for LID are dialog systems, especially in multilingual countries or multilingual translation systems. Instead of trying to recognize a speech utterance in all possible language and choosing the most likely output it is computationally more efficient to first identify the language spoken and then to apply the right ASR system.

Another application would be for emergency telephone services to aid operators if an LID front-end can route a call to the appropriate person.

4.1 Useful Cues for LID

In comparison to ASR where most of the information lies in small portions of the speech, for LID these units are not enough and larger segments have to be considered. [Muthusamy+1994] list several sources of information for language identification:

- **Acoustic Phonetics:** Phonetic inventories differ from language to language. Even when languages have identical phones, the frequencies of occurrence of phones differ across languages.
- **Prosodics:** Languages vary in terms of the duration of phones, speech rate and the intonation (pitch contour). Tonal languages (i.e. languages in which the intonation of a word determines its meaning) such as Mandarin and Vietnamese have very different intonation characteristics than stress languages such as English or German.
- **Phonotactics:** Phonotactics refers to the rules that govern the combinations of the different phones in a language. There is a wide variance in phonotactic rules across

languages. For example the phone cluster /sr/ is very common in the Dravidian language Tamil, whereas it is not a legal cluster in English.

- **Vocabulary:** Conceptually the most important difference between languages is that they use different sets of words – that is, their vocabularies differ. Thus, a non-native speaker of English is likely to use the phonemic inventory, prosodic patterns and even (approximately) the phonotactics of her/his native language, but will be judged to speak English if the vocabulary used is that of English

4.2 Multi-language speech corpora

A major reason for research progress in the last 10 years was the availability of multi-language speech corpora [Muthusamy+1992] to capture the many sources of variability within and across languages. These include variability due to speaker differences (e.g. age, gender, dialect), microphones, telephone handsets, communication lines, background noise and the language being spoken. It is also important that the corpus contains a wide variety of speech from each speaker, ranging from fixed-vocabulary utterances to natural, continuous speech. The availability of such a corpus in the public domain enables researchers to study languages and to develop, evaluate and compare multi-language recognition algorithms.

The OGI Multi-language Telephone Speech Corpus was designed specifically for language ID research. It consists of spontaneous and fixed-vocabulary utterances in 11 languages: English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese.

4.3 Human performance

Human performance was studied by [Muthusamy+1994a]. One experiment with monolingual English speakers was done with 10 languages. After some training, the people were able to identify languages with accuracy ratings from 39% (Korean) to 86% (German and French) using just 6-second excerpts. English scored 100%. An additional experiment was performed where some subjects were able to speak more than one language and there were native speakers of every language tested. Overall performance increased and listeners who knew more languages tended to perform better. The subjects noted to use the following cues for the recognition task:

- Special phones or phone-combinations were linked to certain languages.
- East Asian languages were associated with special intonation (tones).

However, it is still not clear which cues humans use to identify or distinguish unknown languages.

4.4 Common Approaches

[Muthusamy+1994] described several approaches to LID. Much progress has been made in ASR using stochastic models such as Hidden Markov Models (HMM) or Artificial Neural Networks (NN). These approaches are being used in recent LID as well.

The easiest approach is to model an entire language by a single stochastic model, such as an ergodic HMM. Because a single HMM cannot model the complexity of a language, this approach has not been very successful.

The basic idea of the system by [Lamel+1994] is similar to the above approach. It is to train not just one but a set of large phone-based ergodic HMMs for each language and to identify the language as that associated with the model set having the highest acoustic likelihood. Using the 10-language OGI telephone speech corpus, the overall identification rate is 59.2% with 10s of signal. They note that this technique has also been successfully applied to gender and speaker identification and has other possible applications such as dialect identification.

The most popular approach to LID is to look at the phoneme inventory of the languages. Some phonemes do only exist in a certain language while others have subtle differences in realizations in languages. There are also differing frequencies of occurrence of the same phonemes.

4.4.1 Single phone recognizer followed by language modeling (PRLM)

One way would be to use a phone recognizer which can be either language independent [Hazen+1997], [Corredor-Ardoy+1997] or for one specific language [Zissmann1996], [Caseiro+1998]. The language independent implementation uses a phone inventory that covers all the phones of the language to be identified, whereas for the language specific approach a phone recognizer for a specific language, for example English is used for all speech utterances.

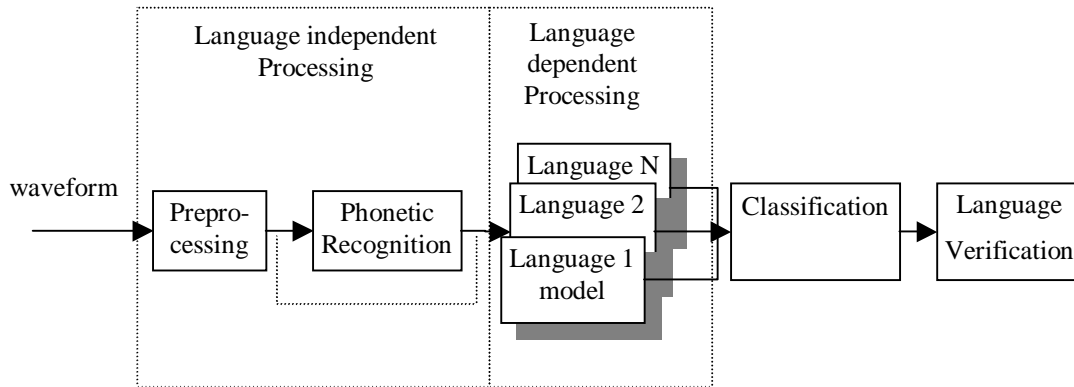


Figure 4.1: Phone Recognizer followed by language modeling

The phone recognizer offers a phone string that includes either all phonemes of the languages or just the phones the recognizer is trained with. This phone string is then fed into language specific models. Those can be trained for language l by running training speech for language l into the phone recognizer and computing a model for the statistics of the phones and phone sequences that are produced by the recognizer. N-grams can be used to model the language. [Zissmann1996] counted the occurrence of n-grams, which are subsequences of n symbols. Training was performed by accumulating a set of n-gram histograms, one per language, under the assumption that different languages will have different n-gram histograms. They then used interpolated n-gram language models to approximate the n-gram distribution as the weighted sum of the probabilities of the n-gram, the (n-1)-gram, etc. Then the log-likelihood for every language was calculated and the decision was made with a maximum likelihood classifier.

[Hazen+1997] employed the concept as shown in Figure 4.1 but additionally incorporated the fundamental frequency and segment duration. They then used three different models for the language likelihoods. The *language model* is very similar to the implementation by [Zissmann1996], as described above. The *acoustic model* accounts for the different acoustic realizations of the phonetic elements that may occur across languages.

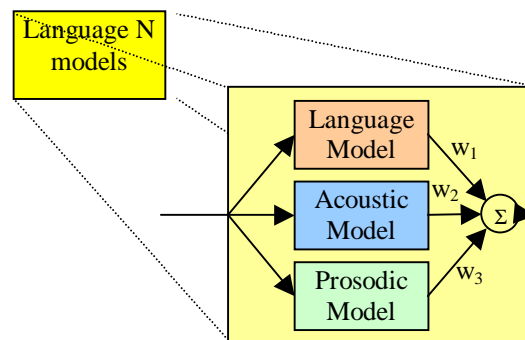


Figure 4.2: Language Model

The *prosodic model* captures the differences that can occur in prosodic structures of different languages due to the stress or tone patterns created by variations in the phone durations and F0 contour. For each frame, a fundamental frequency and a voicing probability are estimated. Then the logarithm is taken for all voice frames and the mean is subtracted. Additionally a delta F0 is calculated. Since no well-developed techniques for automatic capturing and understanding of word- and sentence- level prosodic information were available, their prosodic model only captured simple statistical information about the fundamental frequency and segment duration information of an utterance. They used two separate models for fundamental frequency and segment duration, assuming statistical independence.

They incorporated all models into the system by optimizing weighting factors for different utterance lengths. As the length of the test utterance increases, the weights of the acoustic, duration and prosodic models generally decrease. This effectively gives the language model more weight for longer utterances.

Evaluation by the NIST1994³ test yielded the following results with 11 languages of the OGI Multi-Language telephone speech corpus:

Set of models	<i>10 s utterances</i>	<i>45 s utterances</i>
	<i>Accuracy</i>	<i>Accuracy</i>
Complete system	65,3%	78,1%
Language model	62,7%	77,5%
Acoustic model	49,0%	53,5%
Duration model	31,7	44,4%
F0 model	12,4%	20,9%

Table 4-1: Performance of complete system and individual components

Performance for very short utterances (~ 1 s) was dominated by the acoustic model. The F0-model generally yielded a rather poor performance. For increased performance [Hazen+1997] suggested models that are more sophisticated.

4.4.2 Parallel phone recognizers followed by a language model

Parallel phone recognizers followed by a language model (PRLM) have either a phone recognizer for each language to be identified or any number of phone recognizers of arbitrary languages. [Zissman1996] used phone recognizers for English, Japanese and

³ A standardized test by the US National Institute of Standards and Technology

Spanish to identify Farsi, French and Tamil. A phoneme string for N languages is calculated and then modeled with each of the language models. Computationally this approach is of course much more intensive than the previous solution. [Navrátil1999] employed a similar model but like [Hazen+1997] he additionally used an acoustic model to take into account different pronunciations of the same phone in different languages and a prosodic model which uses segment duration.

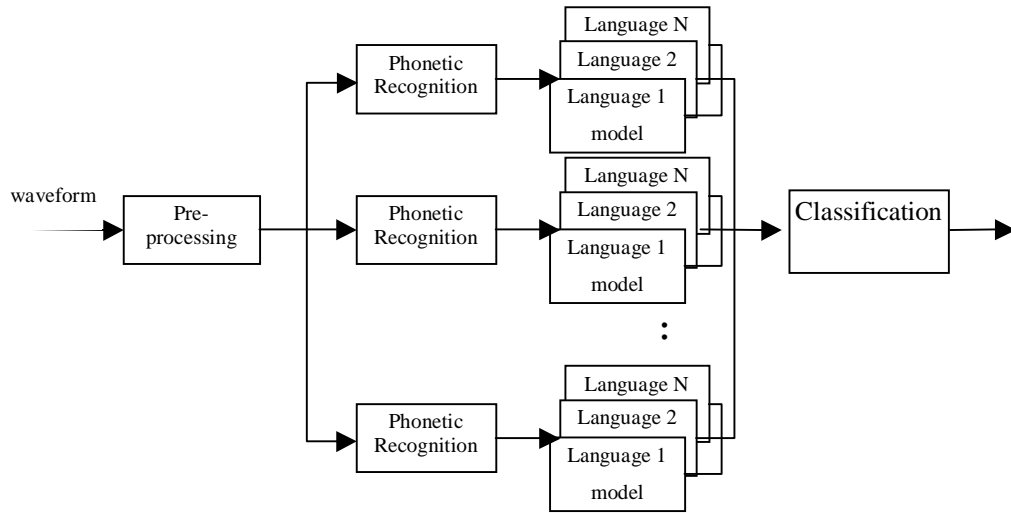


Figure 4.3: Parallel Phoneme Recognizer followed by Language modeling

[Zissmann1996] introduced another model, the parallel phone recognition (PPR) that allows the phone recognizer to use the language-specific phonotactic constraints during the Viterbi decoding process rather than applying those constraints after phone recognition is complete, the most likely phone sequence identified during recognition is optimal with respect to some combination of both the acoustics and phonotactics. The disadvantage of this system is that it needs phonetically labeled speech for every language to be recognized.

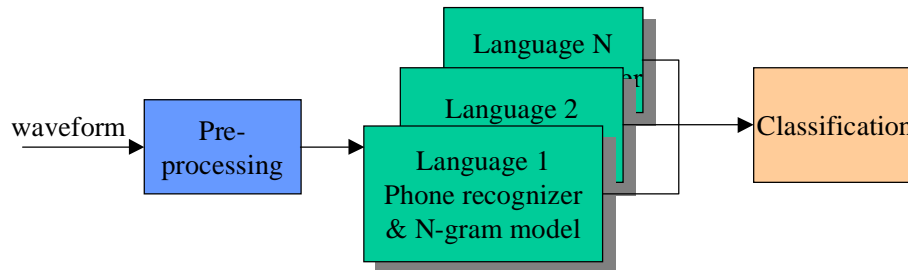


Figure 4.4: Parallel Phone Recognizer (PPR)

[Zissmann1996] also performed experiments using gender dependent acoustic models for phone recognition, including duration tagging, where average phone durations were compared with tested phones. Both improved the performance of the above systems.

4.4.3 Prosodic and Duration Approaches

As mentioned above [Hazen+1997] stated, that the prosodic model didn't contribute much to the overall performance of their system.

[Muthusamy1994] considered more complex prosodic models, which take into account the pitch variation within and across the different segments marked by a broad-category classifier. He also extracted features indicative of speech rate and syllabic timing. Again, these prosodic features were found to be marginally useful.

[Foil1986] examined both formant and prosodic feature vectors, finding that formant features were generally superior.

[Itahashi+1999] used two different methods of parameterizing F0-contours and combined it with statistical analysis. Their approach was to approximate the fundamental frequency by a set of polygonal lines

$$y_k = a_k(t - t_{k-1}) + b_k \quad k = 1, 2, \dots, K$$

where a_k is the slope of the line segment k , b_k is the intercept, and t_{k-1} is the boundary between the adjacent line segment. The parameters a_k and b_k were determined so as to minimize the mean square error between $y(t)$ and $F_0(t)$.

F_0 patterns show hat-like shapes suggesting that an exponential function is expected to be more suitable for approximation:

$$y(t) = a \frac{e}{\tau} e^{-\frac{t}{\tau}} + bt + c$$

Their statistical features were related to F0 and speech power (standard deviation, skewness and kurtosis) additionally correlation coefficients of F0 and speech power were used. For the parameterization, statistical features were calculated as well. Additionally to the F0-contour 12 mel cepstral, 12 delta mel cepstral coefficients and a delta power were calculated (referred to as MCC). This method is based on an ergodic HMM using MCC as segmental information. One HMM was used for each language. For training and evaluation the OGI-TS corpus (10 languages) [Muthusamy1992] was used. Recognition rates were 25.5% / 28.0% for the F0-line / exponential model and 55.5% / 56% for a 32 / 64 state HMM MCC model. The best combined result gave 68.5% accuracy. A suitable weighting factor for the influence of F0 and MCC is important for optimal recognition rates. Performance of MCC HMMs is far better than the F0-contour, but as an additional feature, the latter is still improving performance.

[Thymé-Gobbel+1996] presented the most promising approach. They performed syllable segmentation and extracted pitch and amplitude contour information on a syllable-by-syllable basis and included a statistical module, which computes inter-syllable relationships

in the pitch and amplitude information. They used 224 individual features such as moving averages, deltas, standard deviation, and correlation of measures in the following classes:

- Pitch Contour (shape of pitch contour on a syllable)
- Differential Pitch (pitch differences between syllables)
- Size (distance between syllables and syllable duration)
- Differential size (differenced distance between syllables and syllable duration)
- Amplitude (shape of amplitude contour on a syllable)
- Differential Amplitude (amplitude differences between syllables)
- Rhythm (low frequency FFT of amplitude envelope, syllables per second within breath group)
- Phrase Location (initial/mid/final in breath group; relative phrase position based on syllable distance rations)

Pair-wise language discrimination was performed between English, Spanish, Japanese and Mandarin. These languages represent the traditional categories of stress-timed, syllable-timed, mora-timed and tone language.

The most prominent feature is the pitch (and to a lesser extend deltaPitch). Combinations of location and pitch and delta pitch seem to be most important for LID. The weakest distinctions involve amplitude and differential amplitude, suggesting that using amplitude features is a very poor LID strategy.

The best result was Mandarin versus Spanish using pitch features and phrase location scoring 86% recognition rate.

[Cummins+1999] used ΔF_0 and Δ Amplitude-envelope modulation for discriminating among languages. They do not compute a featural representation of the speech signal in advance; instead, the variables were presented as a time series to a novel recurrent neural network. It employed a Long Short-Term Memory model, to overcome the shortcomings of recurrent neural networks when including temporal information.

[Thymé-Gobbel+1996] chose their four languages, because they include stress-, syllable-, mora-timed and tone languages. This research included German, because it is considered to have a prosodic system very similar to English and it thus allows testing the expectation of maximal confusability for prosodically similar languages.

Pair-wise discrimination based on ΔF_0 and/or Δ Env yielded the following mean results (50% is chance performance, standard deviation in brackets):

% correct $\Delta F0$ and ΔEnv				
	German	Spanish	Japanese	Mandarin
English	55.7 (2.3)	53.8 (5.0)	64.9 (1.2)	63.2 (1.8)
German	-	52.7 (1.3)	67.3 (2.1)	68.7 (2.6)
Spanish	-	-	67.3 (1.9)	72.7 (2.7)
Japanese	-	-	-	60.8 (2.0)

Table 4-2: Results of Neural Nets Prosodic Approach using $\Delta F0$ and ΔEnv

It is evident that the network is quite successful at discriminating among typologically distinct languages that is any pair from Mandarin, Japanese and either English, German or Spanish. However, performance is much worse within the group of the three Indo-European languages. Perhaps surprisingly, Spanish is not easily distinguished from English and German, despite the rhythmic difference.

% correct $\Delta F0$				
	German	Spanish	Japanese	Mandarin
English	53.8 (2.1)	54.3 (3.5)	65.9 (1.4)	63.3 (0.8)
German	-	54.2 (2.1)	73.3 (1.6)	71.7 (2.1)
Spanish	-	-	73.1 (2.2)	68.0 (1.3)
Japanese	-	-	-	52.0 (2.2)

Table 4-3: Results of Neural Nets Prosodic Approach using $\Delta F0$

Given only $F0$ as input, performance in most discrimination tasks is as good or better as in the two-input model. In particular, performance on comparisons involving any one of the Indo-European languages and either Mandarin or Japanese is still reliably above chance, often showing slight, though hardly significant improvement.

% correct ΔEnv				
	German	Spanish	Japanese	Mandarin
English	51.4 (2.5)	60.0 (2.5)	51.7 (2.6)	59.8 (2.5)
German	-	53.9 (2.1)	59.2 (2.0)	60.7 (4.3)
Spanish	-	-	60.7 (1.6)	52.1 (1.6)
Japanese	-	-	-	60.4 (1.4)

Table 4-4: Results of Neural Nets Prosodic Approach using ΔEnv

A point to note is the improvement shown in discriminating Spanish and English when ΔEnv alone is used as input. It is likely that the poorer performance in the 2-input task can be attributed to the harder task of both learning to discriminate based on ΔEnv and

simultaneously learning to ignore the apparently irrelevant ΔF_0 input. It is surprising that the German/Spanish task does not show a similar improvement in the single input case.

To sum up, F_0 is the most useful prosodic discriminant as noted by [Thymé-Gobbel+1996], but discrimination performance is highly dependent on language specific factors. It is interesting that English and German are not being discriminated by neither of all three approaches.

4.5 Accent Identification⁴

There have been very few attempts to identify dialects or accents. The task differs from language identification by the fact that all speakers are speaking the same target language. However, the speakers with foreign accents are expected to import some of the acoustic and phonological features from their first languages into the speech production process. So differences can be acoustic and phonotactic due to the phoneme substitutions and approximations.

[Kumpf+1996] described a system for automatic foreign accent identification for Australian English speech. The classifier is designed to process continuous speech and to discriminate between native Australian English speakers and two migrant speaker groups with foreign accents, whose first languages are Lebanese Arabic (LA) and South Vietnamese (SA).

The system is a Parallel Phoneme Recognition as described in Section 4.4.2 and [Zissman1996]. The speech signal is represented by the observation sequence of feature vectors $\mathbf{O} = \{ \mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_T \}$ with T being the number of frames in the utterance. The feature vector consisted of 12 MFCC coefficients, 12 delta MFCC coefficients, log energy and delta log energy. For each accent dependent recognizer a phoneme HMM set λ_A (3 state left to right topology) and a language model (phoneme bigram model) L_A are trained on the speech of accent A . During testing a Viterbi decoder finds for each recognizer the most likely state sequence representing the speech utterance incorporating the HMM and language model and assigns the log likelihood scores $S_A = \log P(\mathbf{O} | \lambda_A, L_A)$ to the proposed phoneme sequences. The maximum likelihood criterion is then applied to choose the recognizer with the highest likelihood score as the most probable to represent the accent of the test utterance

$$A = \arg \max_A \{ \log P(\mathbf{O} | \lambda_A, L_A) \} \quad A \in \{ \text{AuE}, \text{LA}, \text{SV} \}$$

⁴ Accent will be referred to people who have a different native language than the spoken one; dialect will be seen as a regional variant of one language.

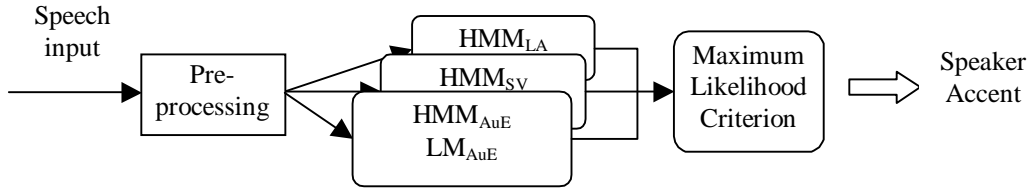


Figure 4.5: Accent classification system [Kumpf+1996]

This system reached an average classification rate for three accents of up to 84.2% correct. [Hansen+1995] presented a system for foreign accent identification of American English. They suggested that normal speech production consists of a sequence of movements in some articulatory feature space from one source generator to another. Actual speech production consists of a ‘neutral’ speech feature production path which must be traversed to produce a given word or utterance.

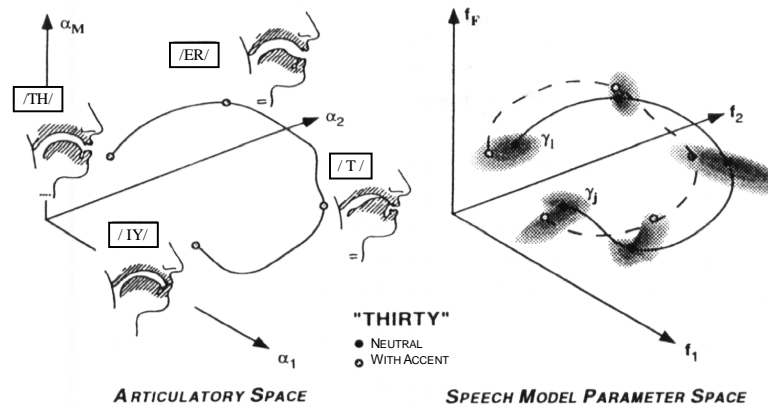


Figure 4.6: Sample Source Generator paths for American English under neutral and foreign accent conditions as projected an articulatory feature space (from [Hansen1995])

They claim that one develops a speaking style while acquiring language skill up to the age of 16, which consists of phoneme production, articulation, tongue movement and other physiological phenomena related to the vocal tract. In general a foreign speaker preserves this speaking style while learning a second-language, and therefore substitutes phonemes from his native language when he encounters a new phoneme in the second language. For accented speech this path through the feature space is somewhat deviated from the normal path.

In order to characterize the change in speech production due to accent in an articulatory space, a series of features was considered: Frame power, zero-crossing rate, LP reflection coefficients, autocorrelation lags, log-area-ratios, line-spectral pair frequencies, LP and FFT cepstrum coefficients, F_0 , formants location and bandwidths. Though there are significant

variations in pitch for the different accent, the most distinct features for classification were on phonemic level.

For an unknown open speech sequence for Neutral American English, German, Turkish and Chinese accent a recognition rate of 81,5% was achieved.

[Teixeira+1996] proposed a three-stage recognition system, in which the first stage decides about the speaker's gender, the second stage classifies the speaker's accent, and the final stage uses recognizer systems corresponding to the decisions made in the previous stages. Concerning the accent identification stage, they used an HMM technique similar to [Lamel+1994]. Global score for discrimination of 6 European language accents of English is 65.4%.

4.6 Summary

Language Identification including Accent and Dialect Identification was covered in this chapter. Human performance and potential cues for LID were discussed. A multi-language corpus, which enhances comparability of research, was presented.

Then common approaches to LID were covered, such as a single phone recognizer followed by language modeling, a parallel phone recognizer followed by a language model and a parallel phone recognizer. Then some specific prosodic approaches were presented. The chapter concluded by introducing work about Accent identification.

5. FEATURE EXTRACTION

The Data used for this research is provided by Forschungszentrum Telekommunikation Wien⁵ (ftw). The database GeveAT is taken from the SpeechDat-AT, a telephone speech database for Austrian German [Baum2000]. The German speakers are taken from the SpeechDat-AT database as well, since the SpeechDat criteria allow 5% non-native speakers. The speech file format is 8bit, 8kHz, A-law speech files, uncompressed. 17 speakers from each Austria and Germany were provided. For each of the speakers 10 sentences were recorded. Every sentence was spoken by at least both an Austrian and a German speaker. Because some of the speakers were not optimal for the task of discerning the origin of the speaker the data set was reduced to 10 speakers (7 male / 3 female).

The first task was to extract prosodically relevant parameters. As seen above, fundamental frequency and intensity contour are the most obvious features. In order to extract the intensity contour the a-law files were converted to wav-files. Duration features are much harder to obtain, because in this case segmentation information such as phoneme or syllable length are needed. Since these features had not been provided, duration features were not tested.

5.1 Fundamental Frequency Tracking

Extracting the fundamental frequency of a speech signal can be achieved in several ways. For this task, *Praat* phonetics tools⁶ were used. This is a toolbox for speech research, including features as spectrogram, LPC, Cepstral-analysis, PSOLA, formant and pitch

⁵Vienna Telecommunication Research Center

⁶ Praat can be obtained from: <http://www.fon.hum.uva.nl/praat/>

tracking. The pitch-tracking algorithm performs acoustic periodicity detection based on an accurate auto-correlation method, as described in [Boersma1993].

This method is more accurate, noise-resistant and robust than methods based on cepstrum or combs, or the original auto-correlation methods. Its key point is the fact that if one wants to estimate a signal's short-term auto-correlation function on the basis of a windowed signal, the auto-correlation function of the windowed signal has to be divided by the auto-correlation function of the window.

5.1.1 Theoretical Background

Ideally, the best candidate for the acoustic pitch period can be found using the position of the maximum of the auto-correlation function of the sound, and the degree of periodicity from the relative height of this maximum. However, the problem is, that sampling and windowing cause inaccuracies concerning the position and height of the maximum.

The auto-correlation of a time signal $x(t)$ as a function of the lag τ is defined as:

$$r_x(\tau) \equiv \int x(t)x(t+\tau)dt$$

If there is a maximum outside 0 and the height of the harmonic strength $r_x(\tau_{\max})$ is large enough the signal is periodic and in consequence there exists a lag T_0 , called the period. The fundamental frequency is then defined as $F_0=1/T_0$.

The short-term auto-correlation is estimated from a short windowed segment of the signal. This gives estimates $F_0(t)$ for the local fundamental frequency and $R_0(t)$ for the harmonic strength.

If there are strong harmonic components in the signal, the highest maximum of the auto-correlation of the windowed signal is rather at a lag that corresponds to the first formant than to the fundamental frequency. Therefore, the pitch estimate from the auto-correlation would be too high. A solution to this problem is to compute the normalized auto-correlation of the window function and to divide the auto-correlation of the signal by the auto-correlation of the window (See Figure 5.1).

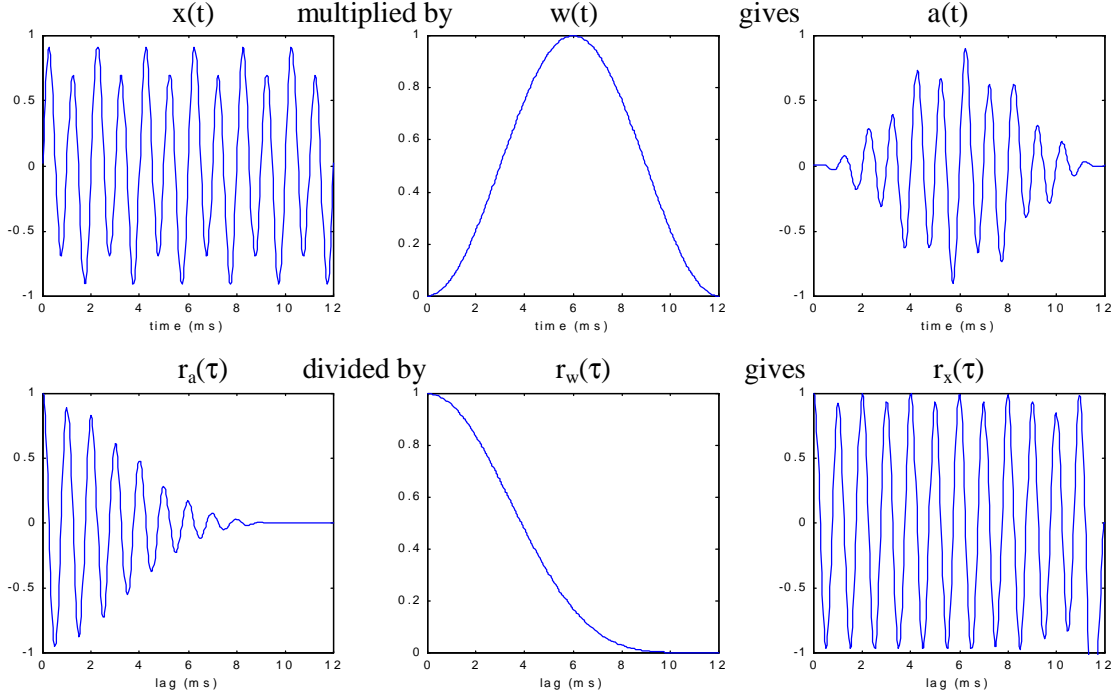


Figure 5.1: Pitch tracking using corrected autocorrelation (from [Boersma1993])

The corrected auto-correlation is:

$$r_x(\tau) \approx \frac{r_a(\tau)}{r_w(\tau)}$$

For every frame, a certain number of candidates (peaks in the auto-correlation function) are stored. A post-processing algorithm, considering cost for voicing threshold, octave jumps, voiced/unvoiced changes, etc., seeks the best path through the candidates.

The following arguments had to be applied:

Time step: the measurement interval, in seconds.

Minimum pitch: candidates below this frequency will not be recruited. This parameter determines the length of the analysis window.

The default arguments for the best path algorithm were used as suggested by *praat*. For male speakers the minimum frequency was 75 Hz and the maximum was 220 Hz. For female speakers the expected fundamental frequency was between 100 Hz and 350 Hz. For all samples a time step of 10 ms was used.

After the pitch tracking a smoothing algorithm was applied with either 8 Hz or 1 Hz bandwidth.

5.1.2 Pitch-post-processing

Human pitch perception is rather logarithmical. Consequently, the logarithm of the fundamental frequency is taken and transformed to MIDI-Numbers ($A2=110\text{Hz}=\text{MIDI}\#36$). This explains why peaks at the end of the downward trend (declination) of an utterance (see Figure 2.4) are perceived as strong as the higher peaks at the beginning.

For statistical analysis, the usual downward trend of an utterance is removed by subtraction of the 1st order regression line. This is also to reduce speaker dependencies, specially concerning gender.

5.2 Parametric Description of the F₀-contour

The following pages describe attempts to model the F₀ contour using a parametric description. All of the models are originated in speech synthesis and they are used for modeling the pitch contour for synthesized speech.

5.2.1 TILT-Analysis

The Tilt-Analysis⁷ is a phonetic model of intonation parametric representation of a pitch contour using three parameters for intonational events: *duration*, *amplitude* and *tilt* (for the shape) [Taylor2000]. Intonational events can be either pitch accents (denoted *a*) or boundary tones (*b*).

The Tilt-Analysis is based on the rise/fall/connection (RFC) model. In this model, presented by [Taylor1995], each intonational event is characterized by four parameters: *rise amplitude*, *rise duration*, *fall amplitude* and *fall duration*. If an event has only a rise component, its fall amplitude and duration are set to 0. Likewise, when an accent only has a fall. The sections of contour between events are called *connections* (denoted *c*) and are also described by amplitude and duration.

In [Taylor2000] it is shown, that the RFC mechanism is not ideal in that the RFC parameters for each contour are not as easy to interpret and manipulate as one might like. Additionally, the parameters are highly correlated and therefore it is possible to reduce the set of parameters to three by transforming the four RFC parameters into three Tilt parameters, namely *duration*, *amplitude* and *tilt* itself.

A single parameter can be used to model the shape of the event. This *tilt* value is calculated as:

⁷ Tilt is part of the Edinburgh Speech Tools Library, which is provided by the Centre for Speech Technology, University of Edinburgh. It can be obtained from http://www.cstr.ed.ac.uk/projects/speech_tools.html.

$$tilt = \frac{|A_{rise}| - |A_{fall}|}{2(|A_{rise}| + |A_{fall}|)} + \frac{D_{rise} - D_{fall}}{2(D_{rise} + D_{fall})}$$

The *amplitude* parameter is the size of the F0 excursion of the event:

$$A_{event} = |A_{rise}| + |A_{fall}|$$

The *duration* is the sum of the rise and fall duration:

$$D_{event} = D_{rise} + D_{fall}$$

F0 position is the F0 distance from the baseline (usually 0 Hz) to the middle of the event.

Time position is where the event is located in time.

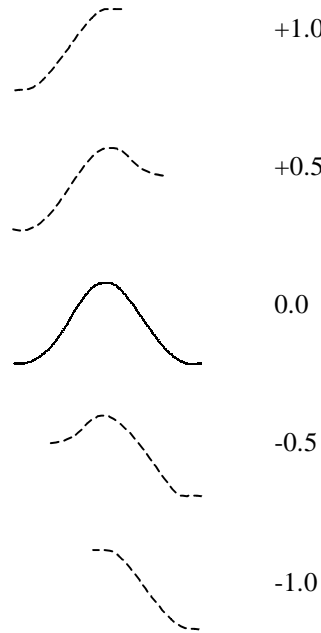


Figure 5.2: Examples of 5 events with varying values of tilt

Using Tilt, the problem arise that a label file for the intonational events is needed. This file is normally created by hand, which cannot be consistently done by different labelers. I have not succeeded in finding an automatic labeling tool.

5.2.2 Intofit

Intofit⁸ is, like Tilt, a parametric description of F0-contours, originally to be used for speech synthesis. It is a maximum-based model, assuming the F0-maxima to be the most important points of the intonation contour [Heuft+1995]. Each F0-contour is parameterized describing

⁸ Intofit was developed at the ‘Institut für Kommunikationsforschung und Phonetik’ at the University of Bonn, Germany and can be obtained from <http://www.ikp.uni-bonn.de/~tpo/intofit.html>.

only its maxima: for each maximum, approximated by \cos^2 functions, four parameters are given. First, the maximum is located precisely in time, relative to the onset of the accented vowel assigned to it. This distance is called *delay*.

The second parameter is, the height of the maximum (*amplitude*) is described as a percentage value between top and baseline. The third and fourth parameter describe the steepness of the contours preceding (*left slope*) and following (*right slope*) the maximum.

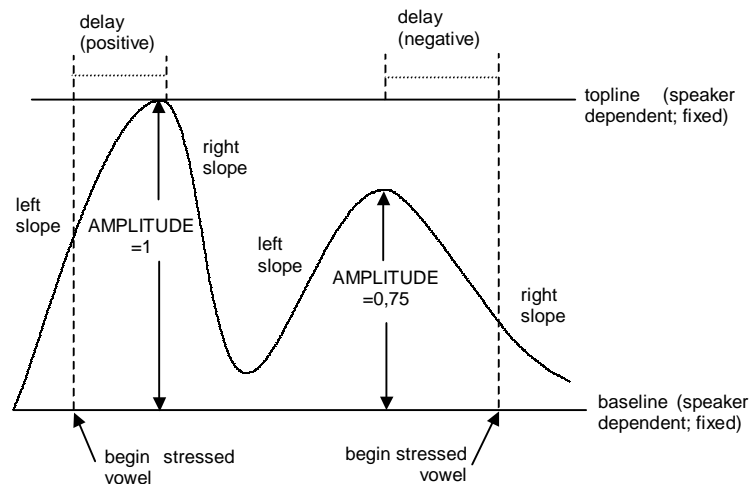


Figure 5.3: Intofit parameters (from [Heuft+1995])

The intonational events for Intofit were obtained assuming all accents were represented by pitch maxima down to a certain threshold. This is linguistically not correct. For this reason the delay parameter is always set to zero, but the F0-contour could be approximated quite correctly (Figure 5.4).

In Table 5-1 the Algorithm for finding the Intofit parameters is described. For the slope parameters an optimization by minimizing mean squared errors was performed. A weighing factor emphasizes the distance close to the maximum, because deviations close to the minimum are perceptual less relevant.

1. Finding the maximum close to the accented vocal
2. Calculation of distance between vocal onset and position of maximum
3. Determination of relative amplitude, related to top and baseline
4. Calculation of optimal slope parameter between previous minimum and current maximum.
5. Calculation of optimal slope parameter between current maximum and the following minimum.

Table 5-1: Calculation of Intofit parameter (from [Portele+1995])

It is to mention though, that quite often the approximation algorithm didn't find a correct fitting. Most of the time, the peaks of the fitted-contour were below the original peaks.

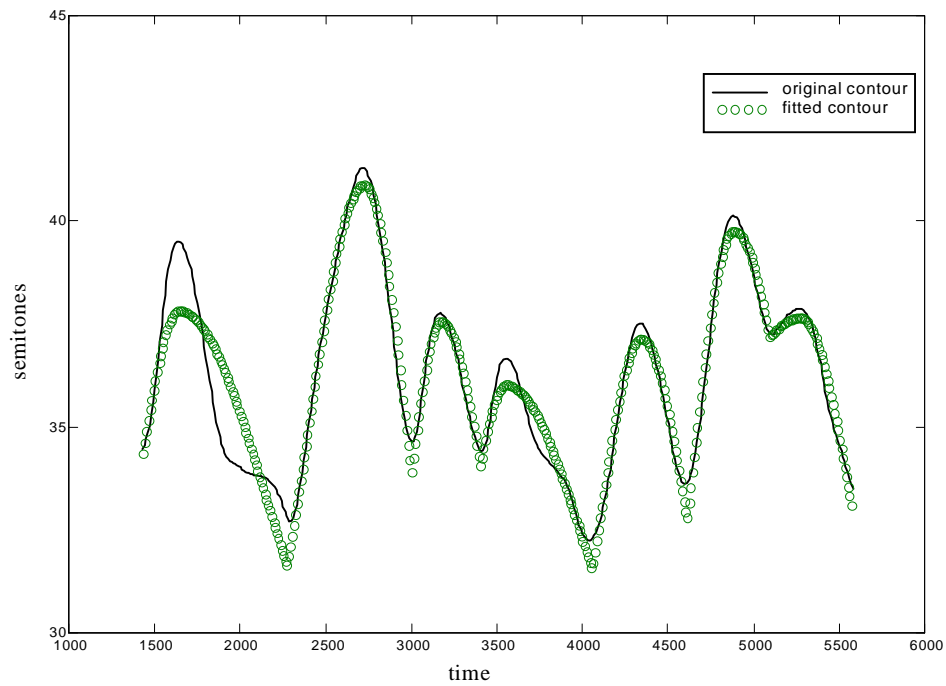


Figure 5.4: Original and (Into-)fitted F0-contour (log-domain)

Originally, Intofit uses linear frequency, but because speech is perceived rather logarithmically, a logarithmical input was used as well. Then for every speaker one mean intofit feature set was calculated (see Figure 5.5).

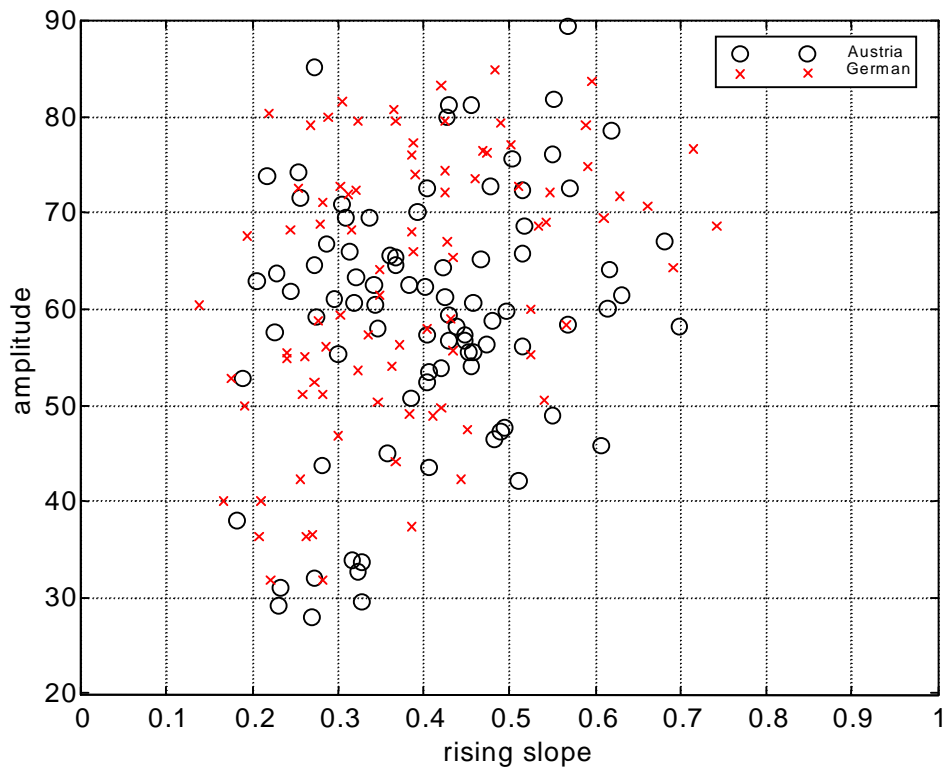


Figure 5.5: Log- Intofit-features: amplitude vs. falling slope

5.2.3 The Fujisaki - Model

The Fujisaki-model ([Fujisaki1983] and [Mixdorff1997]) aims at modeling the generation process of F0 and explaining the physical and physiological properties behind it. It views a F0 contour as the filtered sum of two components: word level accent commands and phrase-level utterance commands. Thus the F_0 contour, $F_0(t)$, of a sentence can be expressed by

$$\ln F_0(t) = \ln F_{\min} + \sum_{i=1}^I A_{pi} [G_{pi}(t - T_{0i}) - G_{pi}(t - T_{3i})] + \sum_{j=1}^J A_{aj} [G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})] \quad \text{Equation 5-1}$$

where

$$G_{aj}(t) = \min\{[1 - (1 + \beta_j t)e^{-\beta_j t}], \gamma\} u(t) \quad \text{Equation 5-2}$$

and

$$G_{pi}(t) = \alpha_i t e^{-\alpha_i t} u(t) \quad \text{Equation 5-3}$$

$u(t)$ = unit step function

respectively indicate the step response function of the corresponding control mechanism to the phrase and accent commands. The α_i 's and β_j 's are expected to be fairly constant within a sentence, or among utterances of an individual speaker. I and J are the number of phrase and accent commands, T_{0i} and T_{3i} denote the onset and end, respectively, of the i th phrase command, while T_{1j} and T_{2j} denote the onset and end, respectively, of the j th accent command. In the absence of pauses within a spoken sentence, the offset times T_{3i} for all phrase commands are assumed to be the same for all i 's within an utterance. On the other hand, the accent commands are constrained not to overlap each other.

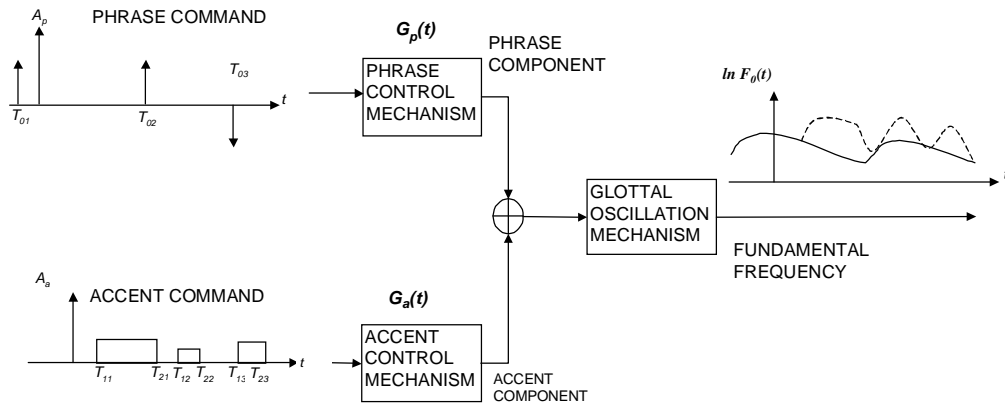


Figure 5.6: Block diagram of the Fujisaki-model

To extract the features the program ‘autofuji’ by Hansjörg Mixdorff was used [Mixdorff1999]. Fujisaki-model parameters are estimated from an ESPS-waves-based F0 contour⁹ in a multi-step procedure, consisting of a quadratic spline stylization, a component separation by filtering followed by command initialization. Then the initial parameter configuration is optimized in a three-pass hill-climb. In the latter part of the procedure, parameters for accent and phrase components are first optimized separately, then further optimized together using the spline contour as the target and ultimately fine-tuned with a weighted version of the extracted contour as the target (See Figure 5.7).

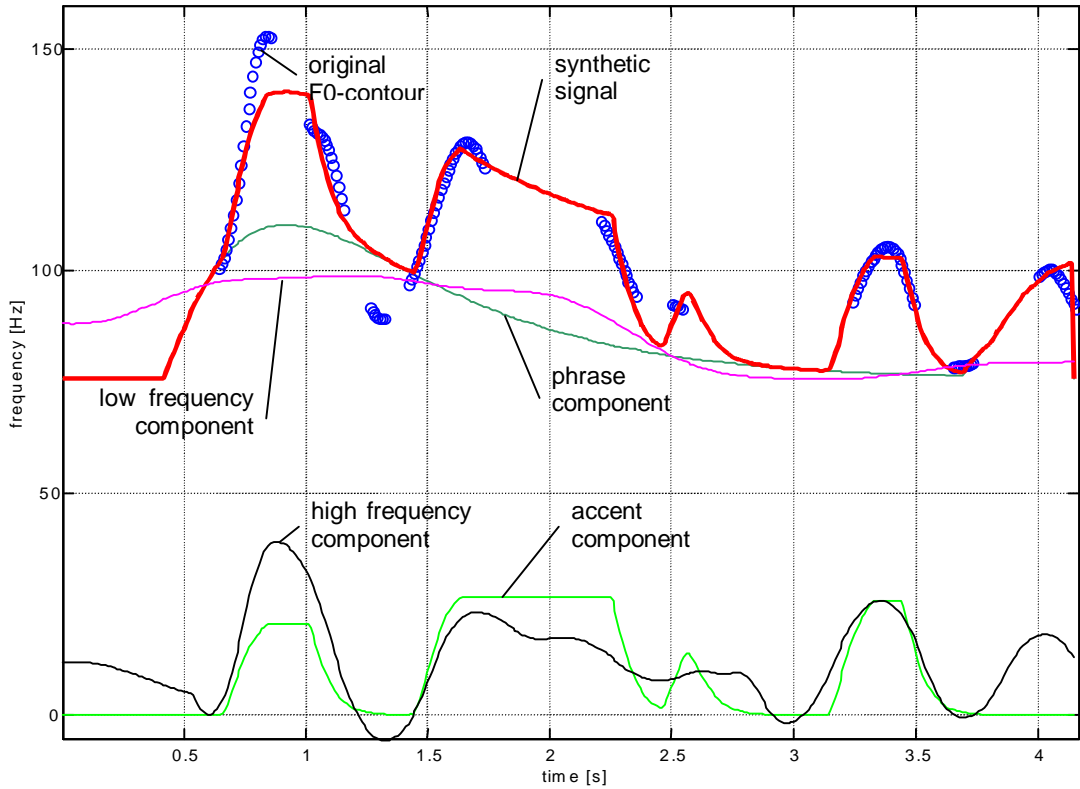


Figure 5.7: Components of Fujisaki Model

Several possible ranges for α and β were evaluated and finally $\alpha=2$ and $\beta=20$ yielded the best results. [Mixdorff1983] already suggested those values for German.

This leaves two variables Aa and Ap as features for the identification of the different dialect groups. Since there can be several accent and phrase components per utterance, representative values for each sentence were calculated. Best results offered a median computation so that there is one median Aa and Ap value for every sentence (Figure 5.8).

For some sentences, the Ap output of the autofuji-program was not calculated, which is obviously wrong. Additionally some abnormal program terminations have occurred. Those errors may be caused by the missing voicing degree, which was not available. However, the

⁹ Since ESPS-waves was not available, the missing degree of voicing parameter was substituted with a binary value. This might decrease the performance achieved.

involved datasets were still used, because Aa was accurately calculated. This was taken care of by setting the median value of the invalid Ap components to zero.

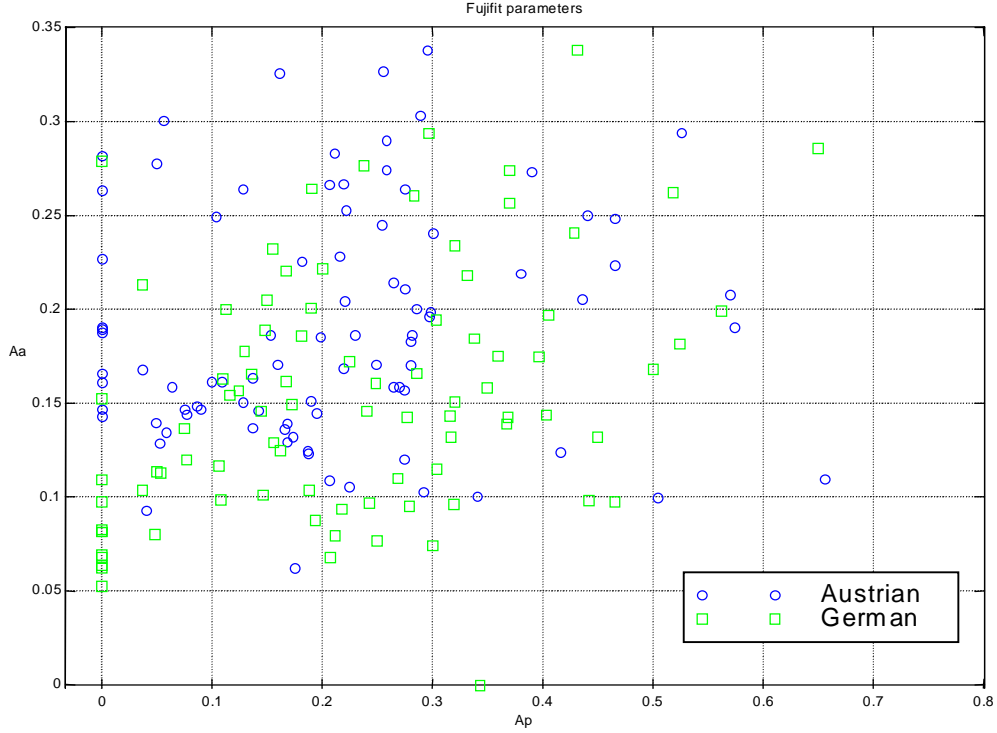


Figure 5.8: Fujisaki Parameters Ap vs. Aa

5.2.4 LPC-Coefficients

This approach is also rooted in intonation modeling for speech synthesis. [Mersdorf1999] proposed a system where LPC-Coefficients for speaker dependent modeling of intonation were used. The LPC-Intonation model consists of the following stages:

- F0-postprocessing: Outliners are removed
- Interpolation: For the LPC analysis a continuous, derivable representation of the F0 contour is recommended. They suggest a cubical spline interpolation assuming a ‘virtual F0’ in unvoiced segments (see Figure 5.9). The interpolation is motivated by the assumption that temporary switching into unvoiced excitation only interrupts a continuous speaker’s intonational gesture [Mersdorf1997].

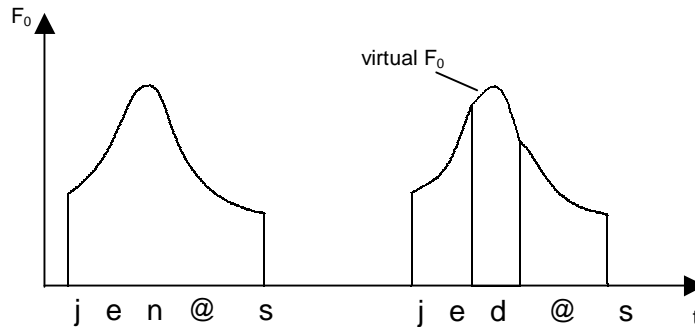


Figure 5.9: Example for virtual F0 (jenes, jedes); from [Mersdorf1997]

- Analysis: The analysis consists of an 8th order LPC of the interpolated contour over the whole sentence. For the whole speech material a single set of individual filter coefficients can be built by computing the arithmetical mean value for each coefficient.
- Approximation of command excitation and (re)synthesis are then used to generate an excitation signal for the resynthesis of a synthetic intonation contour. This can then be applied to synthetic speech using PSOLA or similar techniques.

In Figure 5.10 it is overt that there are quite significant differences in the impulse response of different speakers using 8th order LPC coefficients.

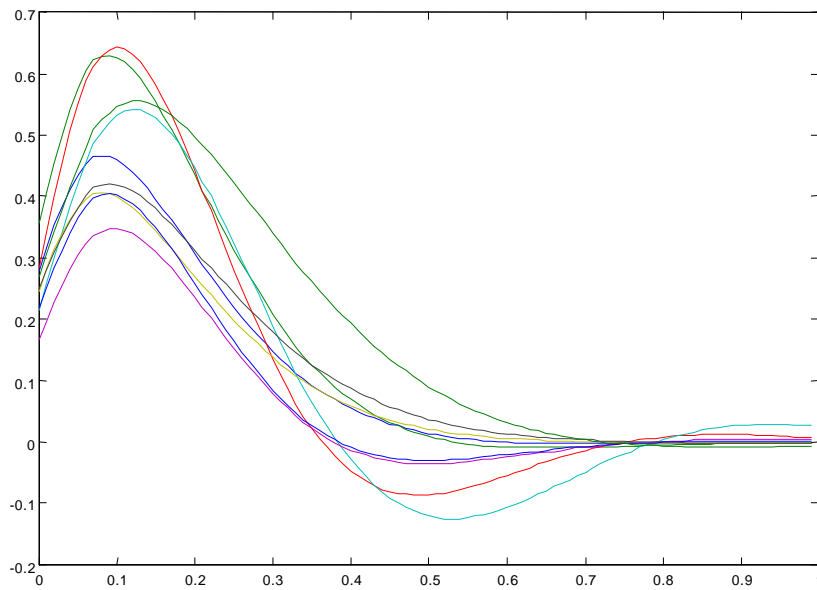


Figure 5.10: Speaker individual Impulse responses (Austrian speakers)

These are mean impulse responses for individual speakers. The idea proposed now is that significant differences between Austrian and German speakers can be expected. Instead of averaging over single speakers, mean filter parameters are calculated for all Austrian and all German speakers.

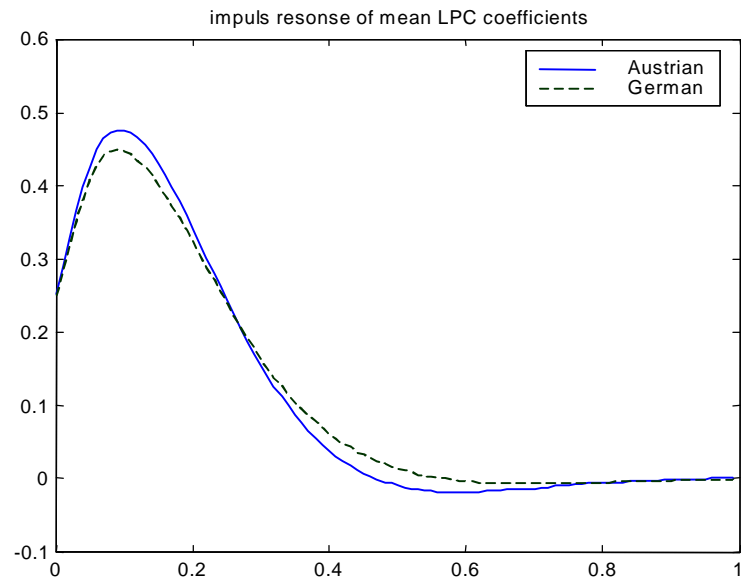


Figure 5.11: Average Austrian and German impulse response

Figure 5.11 shows that the differences are far less significant than with single speakers. The frequency response (Figure 5.12) does not show any differences either.

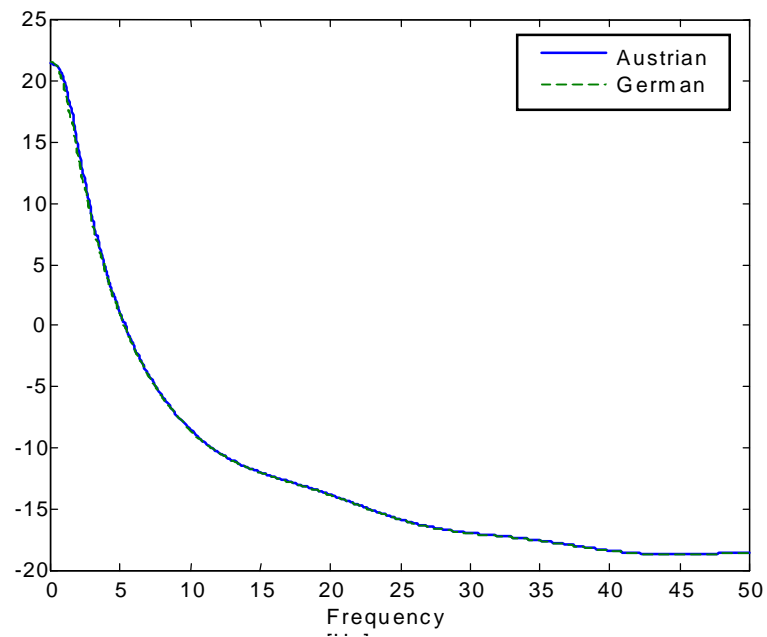


Figure 5.12: Frequency response of average LPC-Coefficients

Further analysis proved that LPC-Coefficients provide no useful features for the discerning of Austrian and German.

5.2.5 Peaks and Intervals

The most simple form of obtaining a data reduced representation of the pitch contour is to find minima and maxima. This was done in the log-domain; therefore intervals are given in semitones.

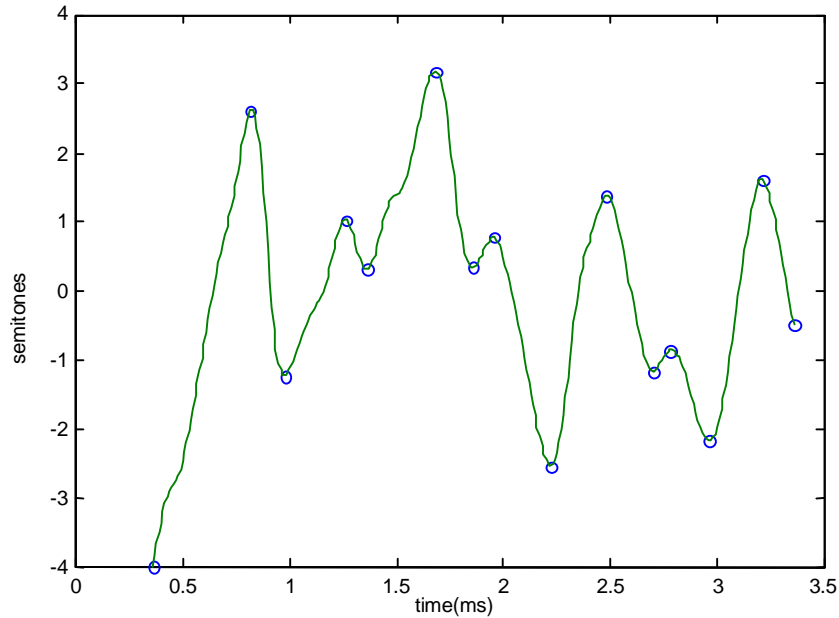


Figure 5.13: minima & maxima of F0-contour

The histogram of the mean intervals between a maximum and a minimum suggests a possible distinctive feature (Figure 5.14.).

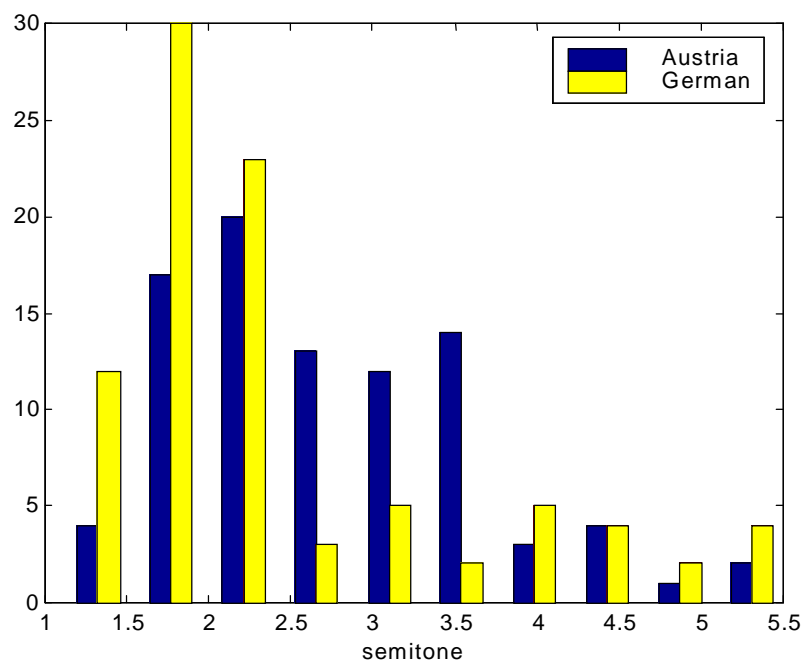


Figure 5.14: Histogram of mean Intervals

5.3 Intensity

Additionally, an intensity contour was calculated in MatLab with the following algorithm (RMS):

$$X[n] = \sqrt{\frac{1}{N} \sum_{i=n-79}^{n+80} x_i^2} \quad \text{Equation 5-4}$$

The values of the sound wave are squared and then summed over a window length of 160 samples (20ms). The hopsize used was 10 ms (80samples).

From the linear intensity contour the logarithm was taken $\{20 \cdot \log_{10}(X[n])\}$. For normalization the maximum of each file was set to 80dB. It was then stored as a Praat Intensity-Tier file.

5.4 Statistical Features

Very little useful information is available from intonation research, which could lead to more knowledge-based features. Statistical features might be a possible approach.

Percentiles (P10, 25, 50, 75, 90), standard deviation, skewness, kurtosis were calculated for the following signals:

- **F0-Contour:** The logarithm was taken, converted to MIDI-numbers and to detrend the data (see Section 2.2.1) the first regression line was subtracted.
- **Delta F0-Contour:** It is said that a lot of prosodic information lies in slopes and intervals (see Section 2.2.1), so delta-F0 could provide some useful information.
- **Intensity-Contour** (framesizes=160/20ms): Intensity is said not to be very useful according to [Thymé-Gobbel+1996], but there is still some information encoded in the intensity contour. Because of background noise the file is much longer than the actual speech. This fact was taken into account by assuming that the actual speech-length is from the first to the last voiced frame. This segment was used to analyze the intensity-contour.
- **Delta Intensity-Contour** (framesizes=160/20ms): The length was calculated as above and the difference was computed.
- **AutoCorrelation(F0):** Auto-correlation provides spectral information about the F0 contour. This function was computed on the logarithmical and detrended F0-contour.
- **AutoCorrelation(Intensity):** The auto-correlation function was computed on the logarithmical and detrended Intensity-contour

- **CrossCorrelation(F0,Intensity (lin/log)):** Cross-correlation shows dependencies between F0 and intensity contour. It is expected that Austrian and German show different patterns in interaction between F0 and intensity.
- **F0*Intensity-Contour (lin/log):** This was computed as just an additional feature that might show dependencies between F0 and intensity.
- **Voiced Ratio:** One widely acknowledged difference between German and Austrian pronunciation is the voicing of consonants [see Section 2.4.2]. Austrians rarely use voiced consonants such as [z, b, d,...], but substitute them with their voiceless counterpart. Even though this is not a clearly prosodic feature, it is assumed that in German speech the voiced rate must be higher than in Austrian utterances. Two different approaches were made. First the *Praat* voicing decision was utilized by computing the ratio of the number of voiced frames to the number of all frames of an utterance. An alternative way was the computation of the zero-crossing rate (see Section 3.1)

5.5 Summary

This chapter explored how to get features, which are relevant for classification of Austrian and German using prosody. Two acoustical characteristics are explored, fundamental frequency and intensity. Two different parameterizations of fundamental frequency are applied, Intofit and Fujisaki. Both offer interesting results.

For both, F0 and intensity several signals such as delta, correlation, etc are calculated and then statistically evaluated, using the t-test.

Processing time of the most important calculations is shown in Table 5-2. The system used was an Intel Pentium III – 500 MHz with 192 MB RAM. Values are in % of real-time. It has to be noted, that the statistical features from Section 5.4 are calculated in MatLab, which is rather slow. Optimizations could decrease the processing time.

Task	processing time % of real-time
Pitch Tracking	33%
Intofit	10%
Fujisaki	40%
Statistical Features (for each signal)	<5%

Table 5-2: Processing time

See a summary of the feature extraction in Figure 5.15.

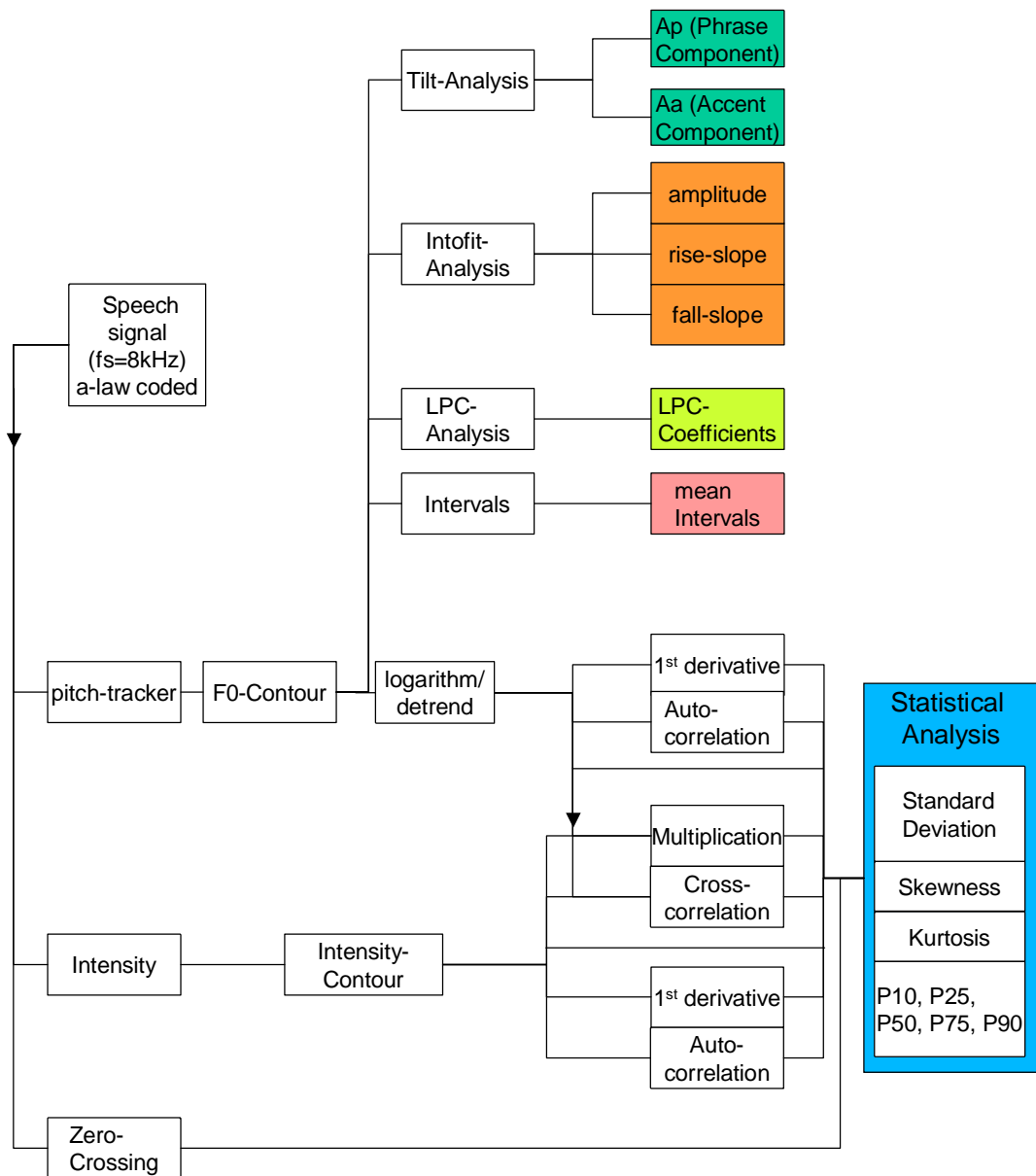


Figure 5.15: Summary of feature extraction

6. EVALUATION

After extracting features from the signal, there has to be made a decision which features are used for discerning the language groups, to achieve the best performance. Not all features mentioned in the previous chapter can improve the classification task.

6.1 Classification algorithm

There are many algorithms to classify signals around, but since this would have exceeded the scope of this thesis I only used one simple classifier, keeping in mind of course, that a better-suited algorithm would have lead to increased performance. For comparison a standard MatLab Multi-Layer-Perceptron is used for the best results.

The main purpose of a classifier is to decide which class a specific data sample with a certain feature vector belongs to (see Figure 6.1).

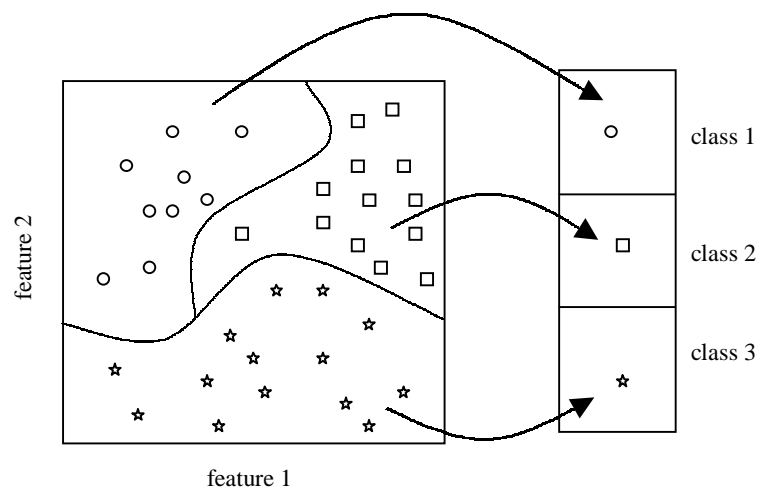


Figure 6.1: principle of classification from [Korl1999]

The algorithm used was to split the whole range of data into sections of equal size or equal percentiles. For each section the number of samples of a class (Austrian/German) from the training data determined which class the section belonged to. Then the test was performed on another set of data (Figure 6.2). Samples that lie in the according area are counted as correct recognition. The class distribution has only limited impact on the class boundaries, because only squares are used. Placing the boundaries in a way that data distribution is considered could lead to better performance.

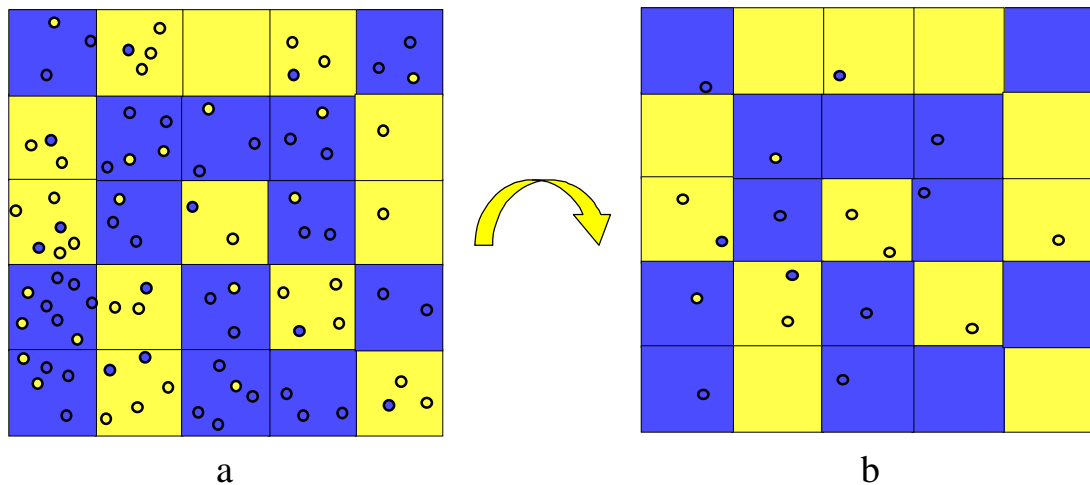


Figure 6.2: a) Determine class boundaries using training data b) Applying test data

6.2 Feature Evaluation

6.2.1 T-Test on all Statistical Features

Since there are many statistical features, the t-test [Hartung1998] is a possible way to determine which of those are worth taking a further look at. To find out which of the features calculated contain information that can be used to determine whether two samples from a normal distribution (in this case Austrian and German) could have the same mean when the standard deviations are unknown but assumed equal. The t-test assumes normal distribution of the data, which we suppose, applies to our feature set. The calculated value *significance* is the probability that the observed value of T could be as large or even larger by chance under the null hypothesis that the mean of x is equal to the mean of y. Small values of *significance* cast doubt on the validity of the null hypothesis. That indicates that the mean could be different.

The calculating the t-test yielded the following significance level:

Evaluation

	Features	P10	P25	Median	P75	P90	StdDev	Skew	Kurtosis
01	F0	0.0422	0.0997	0.2261	0.0837	0.3661	0.2218	0.7473	0.0696
02	Int	0.8797	0.1796	0.8024	0.8392	0.2732	0.7824	0.6258	0.4663
03	Int1	0.5747	0.1464	0.6763	0.4276	0.1432	0.4587	0.3186	0.0488
04	DeltaF0	0.0487	0.0046	0.5412	0.0192	0.0092	0.1267	0.5029	0.7360
05	Delta Int	0.9231	0.6649	0.6735	0.8878	0.8246	0.3762	0.7891	0.4114
06	Delta Int1	0.6134	0.7776	0.0427	0.0926	0.7281	0.7301	0.9748	0.1905
07	Acf F0	0.1510	0.1709	0.1346	0.3012	0.7132	0.3015	0.0767	0.1534
08	Acf Int	0.6989	0.3665	0.9128	0.2909	0.6047	0.7951	0.5501	0.7564
09	Xcorr	0.9688	0.7286	0.4313	0.1742	0.6728	0.8012	0.5330	0.0105
10	Xcorr log	0.9932	0.5527	0.3145	0.1626	0.6220	0.7150	0.6061	0.0057
11	Xmult	0.3412	0.4502	0.1605	0.2567	0.9533	0.4289	0.3278	0.5973
12	Xmult lin	0.9020	0.5197	0.8719	0.5901	0.8462	0.6223	0.9055	0.6260
13	Xmult log	0.1729	0.6506	0.1331	0.3574	0.5946	0.2370	0.6695	0.8594
14	ZeroXing	0.8409	0.3859	0.2307	0.1611	0.5284	0.0395	0.1418	0.2472

Table 6-1: Significance of the statistical features

First conclusions can be drawn from the significance of the features.

The **F0 contour** as expected offers some significant differences between the two national variants. The **intensity contour** provides very little differences apart from the kurtosis of the lesser-smoothed contour. **Delta F0** seems to be a very interesting signal providing very significant percentiles, however for **delta intensity** only one feature is significant (again the lesser smoothed contour). The **auto-correlation of F0** and the **cross-correlation** both provide only one significant features. Multiplication of the signals does not lead to useful differences, so this approach is left out. The **zero-crossing** rate offers only one significant feature as well.

As mentioned above [Thymé-Gobbel+1996] found pitch and delta-pitch the most useful features, which comply with our results, where features from those signals perform best compared to other signals (see scatter-plot in Figure 6.3).

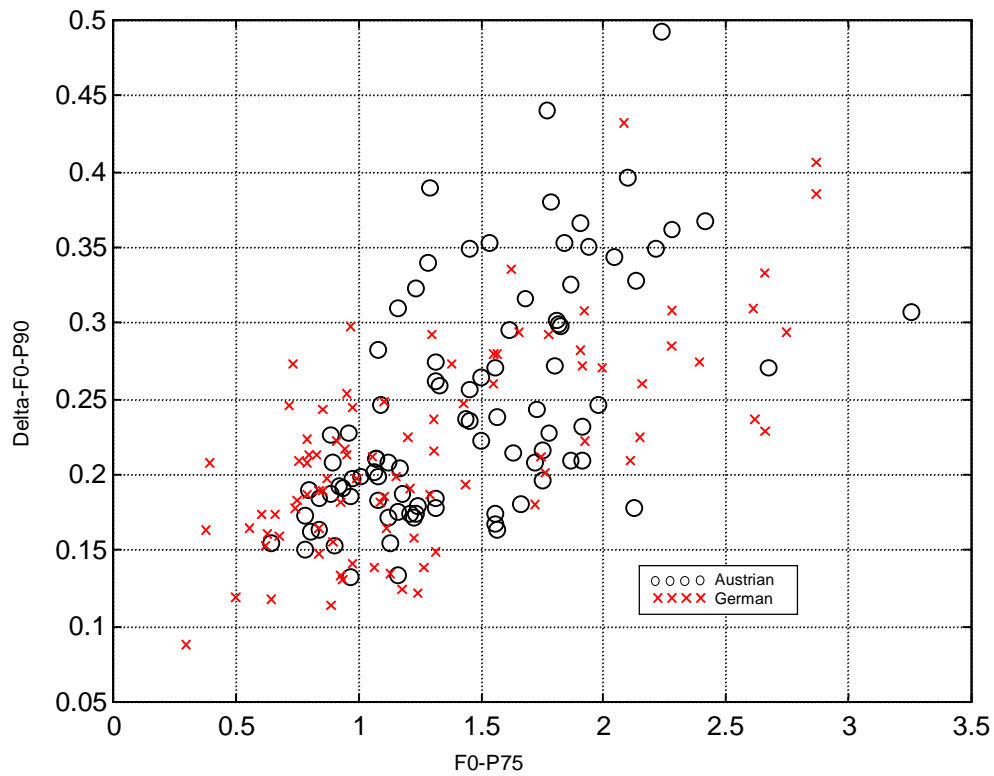


Figure 6.3: Statistical features F0-P75 vs. deltaF0-P90

6.2.2 Feature combination

In order to choose the features, the result of the t-test in section 6.2.1, scatter plots and histograms were used. The most potential features found were then used as input for the classifier, as explained above.

For statistical features, the t-test gave a first decision criterion, which limited the number of useful features. Further analysis showed that only those signals carrying more than just one significant feature in the t-test were able to provide information for the discrimination of the regional variants. The scatter-plots of the parameterized features already suggested potential for possible distinctive features.

The potential features were then used in a pair wise combination test. The classification algorithm used 90% of the data set to train the classifier and 10% of the data for testing. Due to the small size of the database, the whole procedure was repeated 50 times with each time different randomly chosen data samples for testing. Different numbers of division of the data-range (with equal size and using percentiles) were tested; finally, five sections with equal size per feature yielded the best results. See Table 6-2 for averaged results.

	Fujisaki		Intofit			F0		DeltaF0	
	Aa	Ap	Amplitude	rising	falling	P10	P75	P25	P90
mean Interval	63	55	61	57	54	60	64	64	57
Fujisaki Aa	-	65	66	57	52	62	62	53	65
Fujisaki Ap	-	-	52	58	53	64	62	68	56
Intofit Amplitude	-	-	-	61	62	60	62	59	57
Intofit rising	-	-	-	-	62	62	54	57	56
Intofit falling	-	-	-	-	-	51	54	54	50
F0 P10	-	-	-	-	-	-	65	52	56
F0 P75	-	-	-	-	-	-	-	63	66
dF0 P25	-	-	-	-	-	-	-	-	62

Table 6-2: Pair wise feature combination: Recognition rates in %

Figure 6.4 shows the classification using the two Fujisaki parameters; Figure 6.5 shows Fujisaki Ap versus deltaF0 P25, as examples for well classified features.

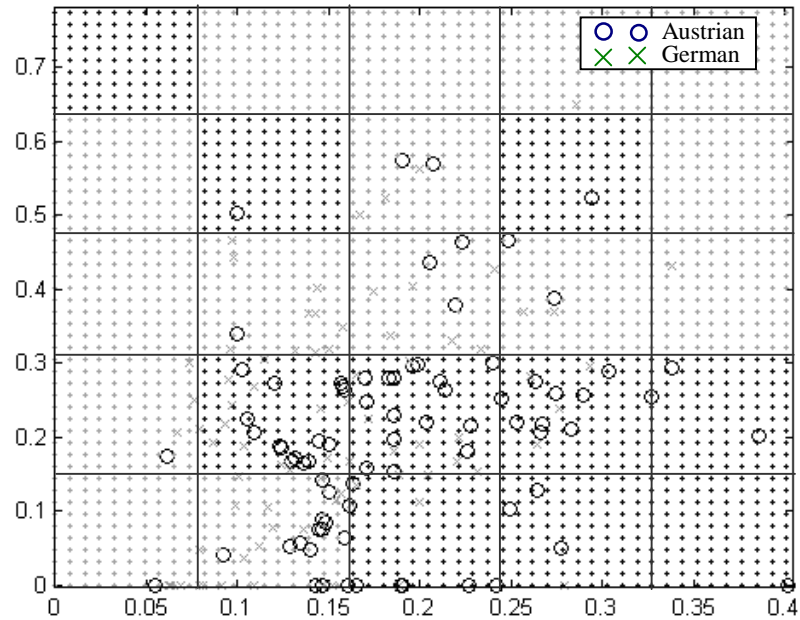


Figure 6.4: Classification of Fujisaki AP vs. Fujisaki AA: Dots show classification

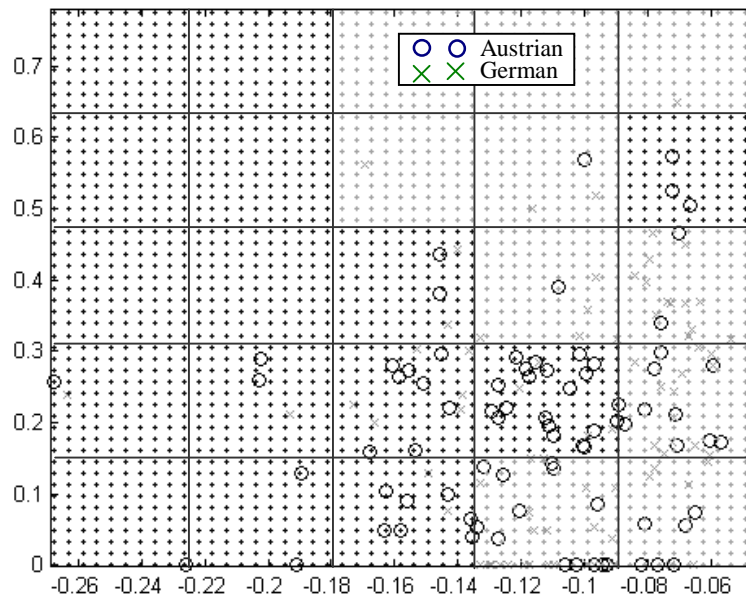


Figure 6.5: Classification of Fujisaki Ap vs. deltaF0 P25 Parameter

To the best combinations from above a third feature was added using only 4 sections per feature this time. Again, linear division yielded slightly better results than the percentile division. See Table 6-3 for results.

	mean Interval	Fujisaki		Intofit			F0		DeltaF0	
		Aa	Ap	Ampl	rising	falling	P10	P75	P25	P90
Fujisaki Aa Fujisaki Ap	71	-	-	59	54	60	71	72	69	54
Fujisaki Aa Intofit Amplitude	64	-	-	-	64	60	72	68	62	63
Fujisaki Aa deltaF0 P90	68	-	-	-	51	51	66	64	61	-
Fujiskai Ap deltaF0 P25	59	-	-	57	57	60	57	57	-	54
F0 P10 F0 P75	59	65	55	60	54	57	-	-	58	58
F0 P75 deltaF0 P90	59	-	60	62	59	56	-	-	56	-

Table 6-3: Triple feature combination: Recognition rates in %

Combinations with more than three features didn't improve the recognition performance. This is expected, because then we either get a sparse matrix (e.g. consider four feature

combinations: If using four sections per feature the matrix contains 256 elements which is more than there are sentences in our data set) or in case of using fewer divisions per feature, differentiation gets very poor.

6.2.3 Alternative Evaluation with MLP

For comparison, a standard MatLab Multi-Layer Perceptron (MLP) algorithm was used to get alternative results.

The MLP had one hidden layer with 40 elements. Testing was done as above with 90% of the data used for training and 10% for testing. 50 different test-data sets were randomly choose and the result is an average over all 50 classifications. Those results are compared with the best ones those from above.

Features	Simple Classification	MLP
Fujisaki Aa, Ap, mean Interval	71 %	68%
Fujisaki Aa, Ap, F0 P10	71 %	64 %
Fujisaki Aa, Ap, F0 P75	72 %	69 %
Fujisaki Aa, Intofit Ampl, F0 P10	72 %	70%

Table 6-4: Comparison of results

The MLP yields lower accuracy rates, probably the small data-set was not enough to train the MLP. However, distinction above chance level can be observed.

6.3 Discussion

This chapter introduced a simple classifier, which was used for the final feature-evaluation. Combinations of two and three features were evaluated. Combinations with Fujisaki features are superior to other groups. With either the mean interval feature or F0 percentiles they reach recognition rates above 70%. Overall processing time for the used features remains within real-time.

Considering prosodic approaches in the past, the current results seem to be very promising. [Thymé-Gobbel+1996] scored best with Mandarin versus Spanish reaching 86%. It is to mention, that Mandarin and Spanish are linguistically very different languages, so distinguishing is easier than identifying similar languages or variants of one language. This is the reason [Cummins+1999] included German to the language set already used by [Thymé-Gobbel+1996] to compare prosodically similar languages. They scored 55.7% for English versus German, which is practically chance level.

[Itahashi+1999] and [Hazen+1997] used prosodic features for identification of 10/11-languages and scored 28% / 20,9% correct recognition.

The reasonably good results of my examinations have to be handled carefully. The major shortcoming of this research is the size of the data-set. Because of the amount of information that lies in prosody, e.g. syntax, semantics, speakers intention, emotions, etc. (see Table 2-1) a large database would be necessary to validate the results of my research.

7. SUMMARY AND DISCUSSION

7.1 Discussion and Outlook

Where do we go from here? In Section 6.3 it was mentioned that the database was not sufficient for a reliable result. During the research some shortcomings of the database arose:

- The most important improvement would be to have much more speakers. 14 males and 6 females are statistically not representative for approximately 100 million German-speaking people. The problem is that speech corpora are very expensive.
- The German speakers were all living in Austria. There is no information how long they have been living there for. This leaves the question how this influenced their prosody. This point is emphasized because the main reason for omitting speakers was that Germans sounded rather Austrian.
- Another critical point of the database was that it all was read speech [Batliner1995]. Spontaneous speech is expected to have the most characteristic prosody concerning Austrian or German. However, having an application such as a phone data access system in mind, there will not be spontaneous sentences as well.

Further research should be done on a big dataset covering a wide variety of speakers from all of Germany and Austria.

For real life applications, the national variant of Switzerland would have to be included as well. This would require speech corpus for German, including all national and regional variants. Currently there is no such database available.

Even though recognition rates were quite promising compared to previous work, it is still clear that using prosodic features alone cannot be a reliable system. I don't expect prosodic features alone to offer reliable cues, due to the multitude of information that is transported via suprasegmentals (Table 2-1). Nevertheless, it can be used to improve phoneme-based systems.

Best performance is expected, when including lexical information and using a much more sophisticated prosodic analysis, as seen in [Nöth+1997]. Then syntactic and maybe semantic information can be considered. However, the computational cost would rise considerably, because all possible national variants had to be considered for phoneme recognition. Whereas now, all calculations can easily be done in real-time.

Of course, additional work has to be done in finding an optimal classifier that improves the recognition rates above.

7.2 Summary

It is useful for ASR applications to distinguish between an Austrian and a German speaker to improve performance. Because of possible degradation of information on phoneme level, an approach was chosen, which is less sensitive to disturbances on the transmission channel. Using suprasegmental features seems to be a possible method.

The speech fundamental frequency (F0) and speech intensity were calculated. Various possibilities to parameterize the F0-contour, such as *Tilt*, *Intofit*, *Fujisaki*, *LPC-coefficients* and *F0-peaks* were investigated

Additionally different signals derived from F0 and Intensity, such as delta, auto-correlation, cross-correlation, multiplication were calculated and then used for statistical analysis (standard deviation, skewness, kurtosis, percentiles). Those features were then evaluated using a standard t-test. Percentiles of Pitch and delta Pitch proved to be the most useful features.

Along with Intofit, Fujisaki and mean-intervals those features were used for pair wise classification. Combinations with three features left the Fujisaki features, Intofit Amplitude, mean Intervals and the F0 percentiles (P10, P75) as the most potential features reaching recognitions rate of above 70%.

Figure 7.1 gives an overview of the most useful features.

Some improvements of the research were mentioned, the most important of all the size of the speech corpus.

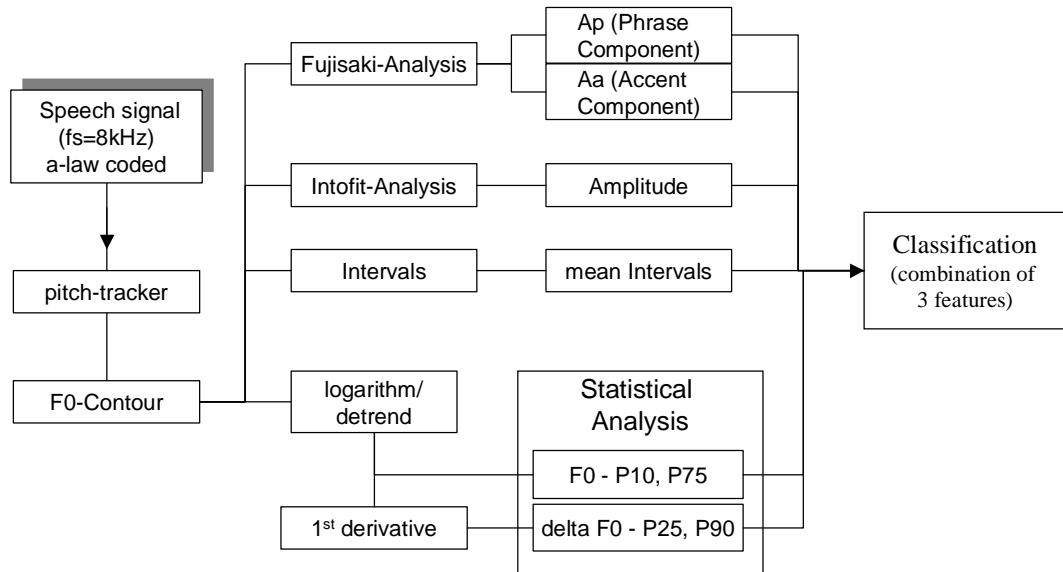


Figure 7.1: Overview of useful features

REFERENCES

- [Auer+1999], Peter Auer, P. Gilles, J. Peters, M. Selting (1999): „Intonation regionaler Varietäten des Deutschen“, *Vorstellung eines Forschungsprojekts*, Freiburg, Potsdam.
- [Batliner+1995], A. Batliner, R. Kompe, A. Kießling, E. Nöth, H. Niemann: “Can you tell apart spontaneous and read speech if you just look at prosody?” in Rubio-Ayuso, Lopez-Soler: 'Speech Recognition and Coding - New Advances and Trends' Springer 1995 pp. 101-104
- [Baum+2000], Micha Baum, Gregor Erbach, Gernot Kubin (2000): “SpeechDat-AT: A telephone speech database for Austrian German”, *Proc. of LREC 2000, Athens*
- [Boersma1993], Paul Boersma (1993): “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound”, *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam Vol.17*: 97-110
- [Buckow+1999], Jan Buckow, V. Warnke, R. Huber, A. Batliner, E. Noeth, H. Niemann (1999): Fast and robust features for prosodic classification, *Proc. TSD Marienbad*, pp. 193-198
- [Burger+1998], Susanne Burger, Florian Schiel: “RGV1 – A database for regional variants of contemporary German”, in Proceedings of the First International Conference on Language Resources And Evaluation 1998, Granada, Spain
- [Carey+1996], Michael J Carey, E S Parris, H Lloyd-Thomas, S Bennett (1996): “Robust Prosodic Features for speaker identification”, *Proc. ICSLP 96*, pp.1800-1803
- [Caseiro+1998] D Caseiro, I M Trancoso: "Identification of Spoken European Languages", in *Proceedings X European Signal Processing Conference (Eusipco-98)*, Rhodes, Greece, September 1998
- [Corredor-Ardoy+1997], C Corredor-Ardoy, J L Gauvain, M Adda-Decker, L Lamel: “Language Identification with Language-Independent Acoustic Models”, in

References

- Proceedings of the European Conference on Speech Technology, EuroSpeech*, Rhodes, Greece, September 1997
- [Cummins+1999], F Cummins, F Gers, J Schmidhuber, "Language Identification from Prosody without Explicit Features", in *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech 99)*, Budapest, Hungary, September 1999.
- [Dutoit1997], Thierry Dutoit: "High-quality text-to-speech synthesis: an overview." *Journal of Electrical and Electronics Engineering, Australia*, vol.17, no.1, March 1997 p25-36
- [Foil1986], J.T.Foil: "Language identification using noisy speech", in *Proc. ICASSP '86*, Apr. 1986, Vol. 2:861-864
- [Fujisaki1983], Hiroya Fujisaki (1983): "Dynamic Characteristics of Voiced Fundamental Frequency in Speech and Singing", in P.F. Mac-Neilage (ed.): "The production of speech", Berlin: Springer 1983, pp. 39-55
- [Gibbon1997], Dafydd Gibbon (1997): „Intonation in German“, from <http://coral.lili.uni-bielefeld.de/~gibbon/Hirst96/german96>
- [Grice+1995], Martine Grice, Ralf Benz Müller(1995): 'Transcription of german intonation using ToBI-Tones - the Saarbrücken System', in *Phonuns 1*, Institute of Phonetics, University of Saarland, 1995, pp.53-64
- [Hansen+1995], John H L Hansen, Levent M Arslan: "Foreign accent classification using source generator based prosodic features", in *Proc ICASSP'95 Vol.1*, pp. 836-839
- [Hartung1998], Joachim Hartung: "Statistik, Lehr- und Handbuch der angewandten Statistik“, *Wien, München: Oldenburg, 1998*, 11.Auflage
- [Hazen+1997], Timothy J Hazen, Victor W Zue: "Segment based automatic language identification", in *Journal of the Acoustical Society of America*, Vol.101 (4) pp.2323-2331, April 1997
- [Heuft+1995], B. Heuft; T. Portele; F. Höfer; J. Krämer; H. Meyer; M. Rauth; G. Sonntag (1995): "Parametric description of F0-contours in a prosodic database“, *Proc. ICPHS 95 Stockholm*, Vol.2:378-381
- [Itahashi+1999], Shuichi Itahashi, T. Kiuchi, M. Yamamoto (1999): "Spoken Language Identification Utilizing Fundamental Frequency and Cepstra" *Proc. Eurospeech 99 Budapest*, Vol. 1:383-386
- [Korl1999], Sascha Korl: „Automatische Klassifizierung zur Anwendung in Hörgeräten“, Diploma-Thesis, University of Technology Graz, 1999
- [Kumpf+1996], Karsten Kumpf, Robin W King: "Automatic accent classification of foreign accented Australian English speech" in *Proc. of ICSLP 96*

References

- [Lamel+1994], L.F. Lamel, J.L. Gauvain: "Language Identification Using Phoneme-base Acoustic Likelihoods", in *Proc. ICASS 1994, Vol. 1:293-296*
- [Mersdorf+1997], Joachim J. Mersdorf, T. Domhöver (1997): "A perceptual study for modeling speaker-dependent intonation in TTS and dialog systems." *Proc. Eurospeech 97 (Rhodos)*
- [Mersdorf+1999], Joachim J. Mersdorf, Kai U. Schmidt, Stefanie Köster (1999): „Linear prediction coding of individual pitch accent shapes", in *Proc. Eurospeech 99, (Budapest)*
- [Mixdorff1997], Hansjörg Mixdorff: "Intonation Patterns of German – Model-based Quantitative Analysis and Synthesis of F₀ contours", *Dissertation at the Technical University of Dresden, 1997*
- [Mixdorff1999], Hansjörg Mixdorff: "Program for Estimating Fujisaki-Parameters (Autofuji)", Program description obtained from the author
- [Morgan+1991], David P Morgan, Christopher L Scofield, "Neural Networks and Speech Processing", *Kluwer Academic Publishers, Dordrecht 1991*
- [Muhr2000], Rudolf Muhr: "Österreichisches Sprachdiplom Deutsch - Lernzielkataloge", öbv+hpt, Wien 2000, CD-ROM1
- [Muthusamy+1992], Y K Muthusamy, R A Cole, B T Oshika: "The OGI multi-language telephone speech corpus", in *Proceedings of the International Conference on Spoken Language Processing (ICSLP 92), Alberta, October 1992.*
- [Muthusamy+1994], Y. K. Muthusamy, E. Barnard, R. A. Cole, "Reviewing Automatic Language Identification", in *IEEE Signal Processing Magazine, October 1994.*
- [Muthusamy+1994a], Y. K. Muthusamy, Neena Jain, R. A. Cole, "Perceptual benchmarks for automatic language identification", in *Proceedings IEEE ICASSP 94, Adelaide, Australia, April 1994*
- [Navrátil 1999], Jiří Navrátil: Untersuchungen zur automatischen Sprachen-Identifikation auf Basis der Phonotaktik, Akustik und Prosodie", *Dissertation at the University of Technology Ilmenau, 1999*
- [Neppert+1992], Joachim Neppert, Magnús Peterson: Elemente einer akustischen Phonetik, 3. Auflage, Hamburg: Buske, 1992
- [Nöth+1997], Elmar Nöth, et al. (1997): Prosodische Information: Begriffsbestimmung und Nutzen für das Sprachverstehen. in *Paulus Wahl: Mustererkennung 1997, Informatik aktuell, Springer Heidelberg 1997, pp.37-52*
- [Oppenheim+1995], Alan V. Oppenheim, Roland W. Schaffer: „Zeitdiskrete Signalverarbeitung“, *München, Wien: R. Oldenbourg Verlag, 1995*
- [O'Shaughnessy2000], Douglas O'Shaughnessy: "Speech Communications, Human and Machine", *New York: IEEE Press, 2000*

References

- [ÖWB1979]: Österreichisches Wörterbuch, Österreichischer Bundesverlag, Wien, 1979
- [Portele+1995], T. Portele, J. Krämer, B. Heuft, G. Sonntag: „Parametrisierung von Grundfrequenzkonturen“ in *Fortschritte der Akustik, DAGA '95*, Bad Honnef
- [Putz+1998], R. Putz, R. Pabst (editors): „Sobotta, Atlas der Anatomie des Menschen“, 20.Auflage, Urban & Schwarzenberg, CD-Rom Version 1.5
- [Rabiner1989], Lawrence R. Rabiner: “A tutorial on Hidden Markov Models and Selected Applications in Speech Recognitions”, in *Proc. Of the IEEE, Vol. 77, No. 2, February 1989*
- [Schaeffler+1999], Felix Schaeffler, Robert Summers (1999): Recognizing German Dialects by prosodic features alone, *Proc. ICPhS 99 San Francisco*, pp. 2311-2314
- [Schiel+1999], Florian Schiel, Christoph Draxler (1999): “Bavarian Archive for Speech Signals (BAS) - Status Report 1995-1998”; internal report (from: <http://www.phonetik.uni-muenchen.de/Bas/BasLiteratur.html>)
- [Strom1995], Volker Strom: “Detection of Accent, Phrase Boundaries and Sentence Modality in German with Prosodic Features”, in *Proc. of European Conference on Speech Communication and Technology*, Vol.3:2039-2041, Madrid, 1995
- [Takahashi1996]: Hideaki Takahashi: „Die richtige Aussprache des Deutsch in Deutschland, Österreich und der Schweiz nach Maßgabe der kodifizierten Normen“, Duisburger Arbeiten zur Sprach- und Kulturwissenschaft, Bd, 27, Frankfurt: Peter Lang, 1996
- [Taylor1995], Paul Taylor (1995): “The rise/fall/connection model of intonation”, *Speech Communications*, 15:169-186
- [Taylor2000], Paul Taylor (2000): “Analysis and Synthesis of Intonation using the Tilt Model”, *Journal of the Acoustical Society of America*, Vol.107 (3) p.1697-1714
- [Teixeira+1996], C. Teixeira, I. Trancoso, A. Serralheiro: “Accent Identification”, in *Proc. of ICSLP '96*, Philadelphia, Vol. 3:1784-1787
- [Thymé-Gobbel+1996], A. Thymé-Gobbel, S.E. Huchins (1996): ”On using prosodic cues in automatic language identification”, *Proc. ICSLP'96*, Philadelphia, Vol. 3:1768-1771
- [Walster1996], W. Walster: „Presseerklärung zum Verbmobil-Forschungsprototypen“, 1996, <http://www.dfki.de/verbmobil>
- [Waibel+1996], Alex Waibel, F. Dellaert, T. Polzin: “Recognizing emotions in speech” *Proc. of ICSLP '96*
- [Weiss+1999], Andreas Weiss, Gerlinde Weiss: “Das österreichische Deutsch“ *Publikationen des Landes-Europabüros Nr. 12, Amt der Salzburger Landesregierung*, pp.3-12
- [Zissmann1996], Marc A Zissmann: “Comparison of Four Approaches to Automatic Language Identification of Telephone Speech” in *IEEE Transactions on Speech and Audio Processing Vol. 4 No1 pp. 31-44*, January 1996