Lehrstuhl für Mensch-Maschine-Kommunikation
Technische Universität München

Rhythm Information for Automated Spoken Language Identification

Ekaterina Timoshenko

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung der akademisches Grades eines Doktors der Naturwissenschaften genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr.-Ing. Jörg Eberspächer
Prüfer der Dissertation:	1. Univ.-Prof. Dr.-Ing. habil. Gerhard Rigoll
	2. Univ.-Prof. Dr. techn. Stefan Kramer
	3. Hon. Prof. Dr. phil. nat. Harald Höge

Die Dissertation wurde am 01.02.2011 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 07.02.2012 angenommen

Abstract

Automatic Language IDentification (LID) is the task of the automatic language recognition from a spoken utterance. For instance, it can be used to select the language for any human-to-machine or human-to-human communication system, when initially it is not known which language should be used.

Language identification is performed using different types of information that can be extracted from a speech utterance. Most successful LID systems implement more than one speech component by the joint modeling of several information types or by combination of isolated components at various levels. Standard LID systems are based on the segmental information and they consider the essential differences among languages either by modeling distributions of spectral features directly (so-called acoustic LID systems), or by modeling strong language-specific frequencies and dependencies among individual speech units in an utterance (so-called phonotactic LID systems). According to the existing human LID perceptual experiments, prosodic information as pitch, intonation, and especially rhythm can be expected to be useful in automatic LID. However, LID systems based on prosody are relatively rare and speech rhythm is still less explored prosodic component.

This thesis presents the research results on the investigation of speech rhythm with the purpose to use it as the additional information type in order to improve the performance of automatic LID systems. The main idea is to explore the duration of neighboring syllable-like units as language discriminative feature. For an appropriate treatment of speech rhythm the algorithm for the segmentation of speech into suited rhythmic units is proposed and a language independent approach for rhythm modeling is developed. As the result, a rhythm-based LID system does not require transliterated training data and can be easily extended with new languages.

To explore the influence of rhythm features on the performance of LID systems utilizing other information types, one phonotactic and two acoustic LID systems are suggested and were implemented. These systems were taken as baseline. All proposed LID systems were trained and tested on the SpeechDat II database, a corpus designed to train commercial speech recognizers. Individual LID systems are firstly evaluated separately and then fused together in different combinations. Adding rhythm to individual baseline systems improves the identification abilities of the resulting system by more than 50 % relatively. This corresponds to the relative reduction of more than 20 % for the detection scenario that shows the reliability of the system decision. The performed experiments in general present the common tendency: Every next system brings less improvement. Therefore adding rhythm to the combination of three baseline systems can improve the performance of the resulting system by almost 3 % and 7 % respectively for identification and detection scenarios. This confirms that speech rhythm as defined in this thesis can be successfully used for any LID system.

Acknowledgments

It is a pleasure to thank many people who made this thesis possible and who helped and inspired me during my doctoral study.

Foremost, I would like to express my sincere gratitude to Professor Gerhard Rigoll for supervising my thesis and offering me the opportunity to complete my doctoral study at the Institute for Human-Machine Communication, Munich University of Technology, Department of Electrical Engineering and Information Technology.

I am deeply grateful to my supervisor, Professor Harald Höge for the continuous support during all the time of research and writing of this thesis, for motivation, enthusiasm, immense knowledge and excellent ideas.

My sincere thanks also goes to Dr. Björn Schuller for his detailed and constructive comments and excellent advice during the preparation of this thesis.

I especially want to thank Dr. Josef Bauer, who introduced me to the field of speech processing and whose wide knowledge and readiness to help have had a remarkable influence on my entire career in the area of human-machine communication research.

For appropriate working environment and convivial place to work, I am grateful to my former and present colleagues from Siemens AG (Professional Speech Processing department) and Svox Deutschland GmbH. In particular, I would like to thank

Ute Ziegenhein for essential assistance in reviewing the thesis from linguistic point of view, for valuable advice, and friendly help,

Dr. Georg Stemmer for providing helpful suggestions, for understanding and support during the completion of my thesis,

Dr. Stephan Grashey for supporting me on Gaussian mixture modeling, and

Joachim Hofer for his help on vowel segmentation used in this thesis and for the friendly and stimulating discussions.

This thesis would not have been possible without the financial support from Siemens AG, Professional Speech Processing department, and later from Svox Deutschland GmbH.

Last but not least, I would like to thank my husband for his never-ending patience and loving support during my studies. I also offer my regards to my friends for their encouragement and understanding, especially to Kris who is responsible for my English, and to Gio who is responsible for all precious moments we shared last years.

Ekaterina Timoshenko

Contents

List of Figures

List of Tables

1 Introduction

The automatic Language IDentification (LID) of spoken utterances attempts to automatically identify the language or dialect that is spoken by a human speaker. With respect to the current global need for multilingual human-computer interfaces [126], LID plays an essential role in providing speech applications. During the past three decades, LID has become a key technology in areas of spoken language processing such as multilingual speech recognition/understanding systems, spoken dialog systems, human-to-human communication systems, spoken document retrieval, and multimedia mining systems.

For example, telephone companies would like to quickly identify the language of foreign callers and route their calls to operators who can speak the language. A multi-language translation system dealing with more than two or three languages needs a language identification frontend that will route the speech to the appropriate translation system. And, of course, governments around the world have long been interested in spoken LID for monitoring purposes. Additional information about important applications of LID systems can be found in the following publications of Dai et al. [35], Ma et al. [76], Waibel et al. [142], and Zue et al. [153].

LID systems may have varying levels of computational complexity and different requirements for the training data, depending on the approach and the information type used to distinguish among languages.

There is a variety of cues that humans and machines use to discriminate one language from another. On the segmental level, languages can be distinguished from each other by the inventory of "phonemes"[1] and their acoustic realizations, i. e., "phones". Each language uses a subset of phonemes from the set of all possible speech sounds. Even though many languages share a common subset of phonemes, the phoneme frequencies and acoustic re-

[1]The term "phoneme" is used as a mental representation of a phonological unit (or segment) in a language (Fromkin et al. [44]).

alizations vary across languages. Thus, these variations can be used to differentiate among languages. Languages also differ in their phonotactic constraints, i. e., rules governing the allowed sequences of phonemes. This causes certain phoneme sequences to be more likely in one language than in another. At the word level, languages are distinct in their vocabularies and the rules by which words are formed. Furthermore, it is possible to distinguish languages by syntactic and semantic rules which govern the concatenation of words into sentence patterns. Using such features requires the existence of large vocabulary speech recognizers, which are language and domain dependent. Moreover, one needs to consider computational costs during training and testing, which can be critical for applications on portable and hand-held devices.

Beside information that is derived from the segmental level, each language has prosodic or supra-segmental[2] properties that allow to determine the meaning of an utterance. Such properties are the modulation of pitch, the shortening and lengthening of segment and syllable duration, and the variations of overall loudness. On the perceptual level, the variation of pitch provides some recognizable intonation and the speech timing, in accordance with some underlying pattern, gives rhythmic properties to speech. In an ideal case, all of the above described properties should be used for language identification. However, standard up-to-date LID systems mostly classify languages by the types of allowed phonemes and phonemes' combinations. Only very few systems use other approaches (e. g., prosodic cues) to further improve their performance. It is shown in detail in Section 2.3.4 that LID systems which use prosody as supplemental information to other components have great potential, even if the extraction and modeling of prosodic cues are not straightforward issues.

One of the basic and most promising prosodic cues is speech rhythm. This component is still less explored than pitch or other prosodic information. Results of perception studies on human language identification have shown that rhythm information carries a substantial part of the language identity that may be sufficient for humans to identify some languages (Section 2.1.3). However, using rhythm for automatic language identification is complicated in terms of providing a theoretical definition of rhythm and its automatic processing.

This thesis presents the investigation of speech rhythm with respect to the LID task. The main idea is to explore the duration of neighboring syllable-like units as language discriminative feature. This definition of speech rhythm does not require additional language-specific training data and therefore suits well for the automatic LID task. The proposed language-independent algorithm for the segmentation of a speech utterance into rhythm units and for rhythm modeling can extend any LID system with a rhythm component. The

[2]Supra-segmental properties are not confined to any one segment, but occur in some higher level of an utterance.

performed evaluation experiments have shown that the speech rhythm, as defined in this thesis, increases language identification accuracy of LID systems based on spectral and/or phonotactic information.

This thesis is organized as follows:
Chapter 2 gives an overview of fundamental notions for the language identification problem. Based on the results of humans' capability to differentiate among languages, the language identification task and corresponding general architecture of an automated LID system are presented. The chapter also describes the previous work and approaches used in current state-of-the-art LID systems according to the language discriminative information utilized there. Existing trends in development of LID systems and in combination of results from several different LID systems are introduced. Finally, the chapter considers a lack of investigation of prosodic information for LID (in particular, rhythm) and presents the goals of this thesis.

Chapter 3 covers the classification techniques that are used in this thesis for designing different LID systems.

Chapter 4 addresses the problem of modeling rhythm for the LID task. The main theories about rhythmic differences among languages and attempts to provide a general definition of rhythm are reviewed. Based on the analysis of current rhythm measurements, a definition of speech rhythm is proposed in the way it can be used for automated language identification.

Chapter 5 focuses on the architecture of the proposed LID systems that utilize different language-specific information. The technique evaluating the LID systems' performance and the fusion methods used to combine the results of individual systems are described in detail.

Chapter 6 contains the evaluation results. Performance of proposed LID systems as well as their combinations are presented and analyzed.

Chapter 7 summarizes the results of the work described in this thesis and presents some conclusions and directions for further investigations.

2

Fundamentals

This chapter starts with the fundamental notions about the language identification problem and at the end argues the motivation for this thesis. Section 2.1 first gives an overview of the basic concepts and linguistic requirements relevant for language identification from a human point of view and for automatic LID systems. The existing experiments on human language identification are summarized in order to show what perceptual cues are used by humans to identify a language. Based on the results of these experiments, a general architecture of an LID system is presented in Section 2.2.

In Section 2.3 current state-of-the-art LID approaches are discussed according to the different types of information they use to discriminate among languages. First, a formal statistical framework for automated LID systems is presented. The following sections provide descriptions of automated LID systems based on spectral information, historically called acoustic and phonotactic LID systems. The existing prosodic LID systems are reviewed in Section 2.3.4. The methods of combining the recognition results from different LID systems are covered in Section 2.3.5. The current trends in language recognition research are summarized in Section 2.3.6.

Finally, Section 2.4 describes the objective of this research and addresses the main goals.

2.1 Linguistic Preliminaries

2.1.1 Language Families

A language is defined as a system of symbols for encoding and decoding information. A human language is a language primarily intended for communication among humans. The exact number of human languages existing in the world is not known and is estimated to

be between 5 000 and 10 000, depending on the definition of the term "language".

Based on historical origins, languages are usually classified into groups related by descent from a common ancestor, so-called language families. There has been much debate about the categorization of a particular language belonging to a family. Complete information about languages and language families, as well as corresponding statistics over the number of speakers, location, dialects, and linguistic affiliations can be found in the most recent edition of the *Ethnologue* [70], a database describing all known living[1] languages, the last edition of which was released in 2009. According to the *Ethnologue*, there are 21 major language families such as *Afro-Asiatic*, *Australasian*, *Indo-European*, *Niger-Congo*, and *Sino-Tibetan*. Below a condensed overview of the *Indo-European* language family is given which is of particular interest for this thesis.

The Indo-European family includes most languages spoken in Europe and many languages spoken in India and the Middle East. It is subdivided into Baltic, Celtic, Germanic, Hellenic, Indo-Iranian, Italic, Romance, and Slavic branches. Each branch includes further subcategories. For example, the Germanic family consists of North Germanic languages (such as Danish, Icelandic, and Swedish) and West Germanic languages (such as Dutch, English, and German), the Romance branch (Italian and Spanish), and the Slavic branch (Eastern European languages, such as Czech, Polish, and Russian).

Languages, that are usually grouped under the same family, may have some similar properties. For example, Italian and Spanish share a large portion of their vocabulary and many grammatical features since they are both Romance languages derived from a common ancestor (Latin). On the other hand, the English and German languages, though both classified as Germanic languages, are very different in their grammatical characteristics.

Linguistic properties of any particular language carry more discriminative information and therefore are taken into account for automatic language identification.

2.1.2 Linguistic Concepts

Differences among languages exist at all linguistic levels. This thesis is concentrated on those linguistic characteristics of languages that are relevant for speech technology in general and, more specifically, for the language identification task.

The sound structure of a language can be described in terms of its phonetic, phonological, and prosodic properties:

[1] A language is called "living" if living native speakers of the language exist, i. e., those that acquire the language during childhood.

1. **Phonetics** deals with the descriptive analysis of physical properties of all speech sounds (*phones*), sound production, and sound perception without reference to language. Speech sounds are usually described in terms of place and manner of articulation. A widely known system for the phonetic transcription of sounds is the International Phonetic Alphabet (IPA), created by the International Phonetic Association.[2] IPA provides a unique symbol for every basic speech sound. The latest version of IPA (revised in 2005) is presented in Appendix A. A practical use of IPA is limited by the absence of a universal (or easy to use) ASCII representation. To compensate this disadvantage, a machine-readable phonetic alphabet called SAMPA[3] was developed. SAMPA, which is abbreviation for Speech Assessment Methods Phonetic Alphabet, basically consists of a mapping of symbols of the IPA onto ASCII and is used throughout this thesis. An example of SAMPA for the German language is displayed in Appendix B.

2. In contrast to phonetics, **phonology** analyzes how sounds function to encode meaning within a given language or across languages. Phonology studies phonemes as the smallest contrastive units in the sound system of a language. A phoneme is defined as a minimal unit that distinguishes meanings of words, i. e., "pin" versus "bin". The inventory of phonemes across languages may vary in complexity from simple systems consisting of eight consonants and five vowels (Hawaiian) to more complex systems with 17 vowels, three diphthongs, and 22 consonants (German).
   In addition to their phoneme inventories, languages can be distinguished by patterns of phoneme combinations, so-called **phonotactics**. Phonotactic constraints determine the syllable structure of a language. Languages such as Dutch, English, and German allow a fairly large number of consonants both at the beginning and at the end of a syllable, leading to a large number of possible syllable structures. By contrast, the Maori language, spoken in New Zealand, only allows syllables consisting of a vowel, two vowels, or a consonant plus a vowel. Other languages place constraints on the type of consonants and vowels that can be combined - for example, Spanish does not permit syllables beginning with "s" followed by consonant. Such a combination often appears in German, e. g., the word "Spiel" which is pronounced as */S p i : l/* (transcribed using SAMPA for the German language, see Appendix B).

3. **Prosody** is in general defined by Bagshaw as "modulation of acoustic characteristics of speech above the level of phonemic segments in order to convey linguistic and paralinguistic information" [5, p. 1]. Prosody may reflect various features of the speaker or the utterance:

[2] http://www.langsci.ucl.ac.uk/ipa/
[3] http://www.phon.ucl.ac.uk/home/sampa/

- the emotional state of a speaker;
- the type of sentence: whether an utterance is a statement, a question, or a command;
- the intent of the speaker: whether he is being ironic or sarcastic;
- the emphasis, the contrast, and the focus;
- other elements of language that may not be encoded by grammar or choice of vocabulary.

In terms of acoustics, prosody consists of rhythm, intonation, and vocal stress in speech, e. g., phenomena that stretch over more than one phonetic segment.

Intonation is defined as variation of pitch which corresponds to fundamental frequency (F_0). Some languages, called **tone languages**, use the pitch contour to distinguish the lexical meaning of a word. Different tone heights or contours (such as high, low, rising, or falling) give a different meaning to the word. This can be demonstrated by the following example from Mandarin Chinese (taken from [127]):

Word	*Tone*	*Gloss*
mā	high level	mother
má	high rising	hemp
mǎ	falling-rising	horse
mà	falling	to scold

In addition to lexical differences, tone can also signal grammatical features as in many African tone languages: In the Edo language, spoken in Nigeria, tone indicates tense (such as past versus present) and aspect (completed versus ongoing action).

Other languages (most languages in the Indo-European family), by contrast, use pitch contours to indicate the sentence type (e. g., question versus statement), phrase and sentence boundaries, and are also used for contrastive emphasis.

According to Hirst, a researcher who compared the intonation systems of twenty languages, "every language has intonation" [52]. Using intonation for discrimination among languages is not a straightforward issue since "intonation is paradoxically at the same time one of the most universal and one of the most language-specific features of human language" [52, p. 1]. Hirst's work has shown that languages differ greatly in functions of intonation but it is "extremely difficult to factor out the language-specific prosodic characteristics of a language from the theoretical assumptions and background of the author" [52, p. 42]. During the past two decades, several attempts to use intonation patterns for the automatic language identification task have been made. The description of these approaches is presented in Section 2.3.4.

Rhythm of speech, according to the definition proposed by Crystal, refers "to the perceived regularities of PROMINENT UNITS in speech. These regularities (or rhythmicality) may be stated in terms of patterns of STRESSED *v.* unstressed SYLLABLES, syllable LENGTH (long *v.* short) or PITCH (high *v.* low) — or some combination of these variables" [33, p. 400–401].

In terms of rhythm, different languages are broadly classified based on the so-called isochrony theory (where isochrony is defined as the property of speech to organize itself in portions of equal or equivalent durations):

- In **stress-timed languages**, like English and German, the intervals between two stressed syllables are said to be near-equal. Syllables that occur in between two stressed syllables are shortened to accommodate this property.
- In **syllable-timed languages**, such as French and Spanish, successive syllables are said to be of near-equal duration.
- In **mora-timed languages**, exampled by Japanese, the duration of successive morae are said to be near-equal. Mora "refers to a minimal unit of metrical time or weight" [33, p. 229] and consists of one short vowel and any preceding onset consonants.

Despite its popularity among linguists, the isochrony theory, also called *rhythm class hypothesis* and first introduced by Pike [103] and Abercrombie [1], is contradicted by several experiments that were carried out in search of empirical proof of isochrony properties of some languages. After Pike and Abercrombie, many researchers have been looking for an explanation for different rhythmical properties of world languages. Despite ongoing efforts, a well accepted definition of rhythm does not exist yet. Since rhythm of speech is in the main scope for this thesis, a detailed overview of existing rhythm theories is presented in Section 4.1.

2.1.3 Human Language Identification

Humans are able to recognize a language from extremely short audio samples, as long as they have a certain degree of familiarity with the language. Since the performance of human listeners on LID tasks is the target for an automatic LID system, a starting point for designing an LID system is to explore what perceptual cues are used by humans to identify a language.

Investigations on humans' capability for LID can be conventionally divided into two parts: those that use unmodified (real) speech and those which are performed with stimuli signals that were modified in different ways according to the specific task.

Various experiments on human language identification based on the unmodified speech were reported by Muthusamy et al. [88]. Utterances of spontaneous speech with durations from one to six seconds were recorded for ten languages: English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. Participants (ten native speakers of English and at least two native speakers of the remaining languages) had to identify the language after listening to the test samples. Several results, as summarized by the following, were obtained:

- Identification results varied significantly across the languages. As expected, English was identified with highest accuracy; French, German, and Spanish, languages that the listeners were most often exposed to, achieved relatively high identification rates; and accuracy in identifying Farsi, Korean, Tamil, and Vietnamese was very poor.

- A learning effect was detected: After every individual test sample, the correct answer was given to the participants, and the accuracy at the end of experiment improved in comparison with the results at the beginning.

- A positive correlation was observed between human accuracy and the duration of the test signal as well as the participants' general knowledge of foreign languages.

- As reported by participants, the following cues were taken into account: segmental information (manner and place of articulation, presence of typical sounds, syllables, and short words) and supra-segmental information (rhythm, intonation, tones).

Another study of human LID, reported by Nazzi [95], has demonstrated that French newborns are capable of discriminating among languages that belong to different rhythm classes as proposed by traditional rhythm class hypothesis. Infants in Nazzi's study could discriminate among languages with different rhythmic classifications but were not able to discriminate between two languages which belong to the same class. This indicates that the rhythm of speech plays an important role in the perception of language and can be used to discriminate groups of languages, but only if these groups are congruent with rhythmic classes.

In a similar study on modified speech, Ramus and Mehler [112] focused on the importance of various acoustic cues for discrimination of languages. The experiments were based on a speech re-synthesis technique that allows one to preserve or degrade phonotactics, rhythm, or intonation from natural utterances into different combinations. The original English and Japanese speech utterances were segmented into phonemes that were then replaced by French phonemes to exclude segmental cues. Four types of stimulus signals were created differing in the information they contained:

- broad phonotactics, rhythm, and intonation (all fricatives were replaced by /s/, stops by /t/, liquids by /l/, nasals by /n/, and vowels by /a/);
- rhythm and intonation (all consonants are replaced by /s/, and vowels by /a/);
- rhythm only (the same as the previous one but with no variation in fundamental frequency);
- intonation only (all segments are replaced by /a/).

Bilingual French adult test persons were told that the utterances are from acoustically modified sentences of two real and exotic languages and were asked to discriminate between them after passing a training session. The average accuracy in the discrimination experiments for the first three types of stimuli ranged between 65 % and 68 %. In the "intonation only" condition, test persons behaved in a way that looked like guessing (with recognition rate of 51 %). One of the main conclusions made by Ramus and Mehler from this study was that rhythm is "an excellent and possibly the best prosodic cue for the discrimination of languages that are said to differ in rhythm" [112].

Although more recent experiments, which will be discussed in Section 4.1, contradict the rhythm class hypothesis, these findings have demonstrated rhythm as relevant parameter for language discrimination.

In recent investigations, Navratil [92] showed the different importance of segmental and prosodic information. The experiment was based on five languages (Chinese, English, French, German, and Japanese) and consisted of three test sections:

- original signals without any modification;
- short segments of six seconds with shuffled syllables, so that the meaning of the original sentence as well as its prosodic contour were destroyed leaving only acoustic-phonotactic information;
- signals were filtered by an adaptive inverse Linear Predictive Coding (LPC) filter, thus flattening the spectral shape and removing the vocal tract information so that only prosodic information (F_0 and amplitude) were audible.

The overall LID rates (96 % for full speech, 73.9 % for syllables and 49.4 % for prosody) indicate that segmental parameters are the most important cues and prosodic parameters are the weakest cues when discriminating among languages. Nevertheless, the results of human perception experiments show prosody as one of the language discriminating components that should be investigated for automatic LID.

The importance of prosody for the LID task has been confirmed by different performance rates obtained by automatic LID systems which use prosodic features in combination with other language characteristics, such as acoustic-phonetic and/or phonotactic information [42, 49, 50, 73, 136, 149, 151]. Such systems, as well as other currently successful LID approaches, are described in more detail in the next sections.

2.2 Automatic Language Identification Task

Automatic language identification systems are based on the linguistic properties of languages extracted from input speech. Performance of such an LID system depends on the amount and the reliability of discriminative information and how efficient it is incorporated into the system.

Most language identification systems rely on the spectral information derived through short-time spectral analysis of the speech signal, such as acoustic properties of sound units (in literature usually referred as to acoustic-phonetics) and their sequences (referred to as phonotactics). In addition to spectral information, some LID systems incorporate prosodic information. A general architecture of an example LID system based on different discriminating cues is shown in Figure 2.1.

Every type of information used for LID has not only advantages but disadvantages as well. Spectral information, i. e., acoustic-phonetics and phonotactics, is easy to obtain and becomes a trade-off between computational complexity and performance. In the same time, the performance of LID systems based on spectral information usually degrades due to noise and unmatched acoustic condition. Prosodic properties are less affected by channel

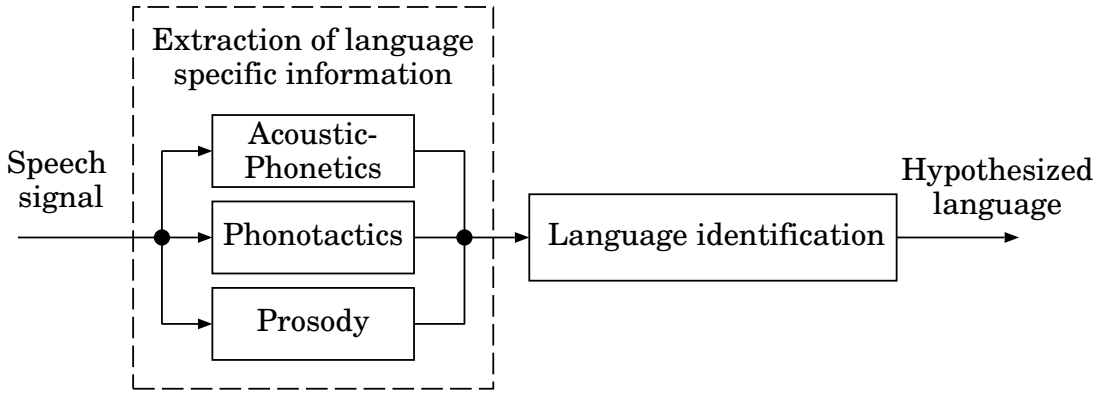


Figure 2.1: *General architecture of an LID system utilizing different discriminative information*

variation and noise [132] but are hard to extract reliably [55]. The prosody of an utterance is influenced by the speaker-specific characteristics such as voice type, speaking rate, emotional state, etc., by syntactic context of the utterance, e. g., statement, question, and so on. To successfully exploit prosodic information, the language-dependent characteristics should be separated from language-independent components.

In order to use different types of information for discriminating among languages, a trade-off between accuracy and efficiency must be considered. Design of an LID system depends on the concrete LID task, modeling methods, and the availability of a sufficient amount of training data. Some systems require only digitized speech and the corresponding identities of the languages being spoken. Other systems demand the existence of an orthographic transliteration (i. e., signal-phoneme correspondences) for each training utterance.

Ideally, the LID system should be highly accurate in identifying a language while also

- being computationally efficient,
- being robust against speaker, channel, environment, and vocabulary variability,
- requiring a minimum of language-specific information for the development of the system,
- allowing a new language to be included without much effort.

2.3 State-of-the-art LID

2.3.1 Statistical LID Framework

State-of-the-art language identification systems rely heavily on pattern classification theory [40, 119], one of the most challenging problems for machine learning.

Pattern classification aims to classify data (patterns) based on a-priori knowledge and a-posteriori information extracted from the patterns. The patterns to be classified are usually groups of measurements or observations that define points in an appropriate multidimensional space. The main parts of a pattern classification system are a "feature extractor" that computes numeric or symbolic information (features) from the input pattern and a "classifier" that evaluates the presented evidence and makes a final decision based on the extracted features.

Pattern classification systems work in two phases: training and recognition. During the training phase, the system learns characteristics of the classes from the training patterns.

For each class, the training patterns are analyzed and used to produce a model that is intended to represent the class-specific characteristics. The models are then taken for the second phase — recognition. During the recognition phase, the unknown test pattern is compared to each model and a measure of similarity (distance) between test pattern and reference model is computed. These distances (scores) are then used to choose which reference model matches the input test pattern in the best way.

For the LID task, a pattern is presented by language-specific information that characterizes a speech utterance and its classification label, i. e., language identity. The set of classes corresponds to the set of n languages $\mathcal{L} = \{L_1, \ldots, L_n\}$ to be identified from the speech signal. A general architecture of an LID system based on the pattern classification is presented in Figure 2.2.

The speech signal is represented by a sequence of vectors computed by the feature extraction component. Each individual feature vector represents relevant information from the signal for a particular time frame. The extraction of reliable features is one of the most important issues in all speech processing systems. In general, for a good recognizer the features should be independent (i. e., not correlated with each other) and salient (i. e., useful for subsequent acoustic matching). In the training mode, speech samples from a variety of languages are used to create a set of language-specific models. In the recognition mode (in this particular case — identification mode), the scores for all models are calculated and the model with the the best score is hypothesized.

Before designing any system, it is desirable to develop a strong theoretical framework on which the design can be based. State-of-the-art LID systems are based on the "formal probability framework" proposed by Hazen [49] and also used for this thesis.

Let $\mathcal{X} = \{\vec{x}_1, \ldots, \vec{x}_T\}$ be a sequence of T vectors which represents the spectral information

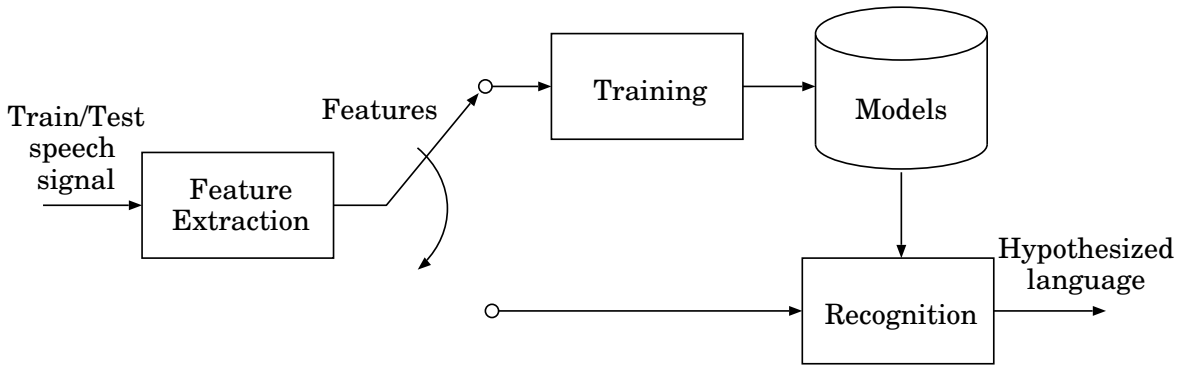


Figure 2.2: *General LID system based on the pattern classification paradigm*

of a spoken utterance and let $\mathcal{F} = \{\vec{f_1}, \ldots, \vec{f_T}\}$ be a sequence of T vectors which represents prosodic information. Then $P(L_i \mid \mathcal{X}, \mathcal{F})$ is the probability that an utterance described by $\mathcal{X}$ and $\mathcal{F}$ was spoken in language L_i.

The Maximum Likelihood (ML) approach to the problem is to choose the language which is most likely given $\mathcal{X}$ and $\mathcal{F}$. Viewed as maximization process, the ML decision rule can be expressed as follows:

$$L^* = \underset{i}{\operatorname{argmax}}\, P(L_i \mid \mathcal{X}, \mathcal{F}),\ i = 1, \ldots, n. \tag{2.1}$$

Let $\mathbf{C}$ denote a set of all possible phoneme sequences that can represent a spoken utterance where each sequence is of the form: $\mathcal{C} = \{c_1, \ldots, c_K\}$. The linguistic information can be involved into the theoretical framework of the LID task by rewriting Equation 2.1 into the following form:

$$L^* = \underset{i}{\operatorname{argmax}} \sum_{\mathbf{C}} P(L_i, \mathcal{C} \mid \mathcal{X}, \mathcal{F}). \tag{2.2}$$

The probability in Equation 2.2 can be transformed into several components that are easier to model. Using the definition of the conditional probability, Equation 2.2 is reworked as:

$$L^* = \underset{i}{\operatorname{argmax}} \sum_{\mathbf{C}} \frac{P(L_i, \mathcal{C}, \mathcal{X}, \mathcal{F})}{P(\mathcal{X}, \mathcal{F})}. \tag{2.3}$$

Since Equation 2.3 is the maximization process and $P(\mathcal{X}, \mathcal{F})$ does not depend on i, the above equation can be converted into the following:

$$L^* = \underset{i}{\operatorname{argmax}} \sum_{\mathbf{C}} P(L_i, \mathcal{C}, \mathcal{X}, \mathcal{F}). \tag{2.4}$$

In practical applications, a probability estimation for all possible phoneme sequences is not tractable. The required computations can be significantly reduced by a calculation of a single optimum phoneme hypothesis $\mathcal{C}^*$. If it is assumed that $\mathcal{C}^*$ can be found independent of the language, then the Equation 2.4 can be rewritten as

$$L^* = \underset{i}{\operatorname{argmax}}\, P(L_i, \mathcal{C}^*, \mathcal{X}, \mathcal{F}). \tag{2.5}$$

The expression $P(L_i, \mathcal{C}^*, \mathcal{X}, \mathcal{F})$ represents the probability of the occurrence of several de-

pendent events — $L_i, \mathcal{C}^*, \mathcal{X}$ and $\mathcal{F}$. Thus,

$$L^* = \underset{i}{\operatorname{argmax}}\, P(L_i)P(\mathcal{C}^* \mid L_i)P(\mathcal{X} \mid \mathcal{C}^*, L_i, \mathcal{F})P(\mathcal{F} \mid \mathcal{C}^*, L_i), \tag{2.6}$$

where	$P(L_i)$	is the a-priori language probability and can be ignored when only one database is used with the assumption that all languages in the set $\mathcal{L}$ are equally likely, (speech databases used to test the LID systems usually contain nearly equal amounts of data for each language);
	$P(\mathcal{C}^* \mid L_i)$	is called phonotactic model and describes the frequencies of phoneme co-occurrences;
	$P(\mathcal{X} \mid \mathcal{C}^*, L_i, \mathcal{F})$	is called acoustic[4] model which is usually considered independently from other components, i. e., as $P(\mathcal{X} \mid L_i)$ that gives pure acoustic probability;
	$P(\mathcal{F} \mid \mathcal{C}^*, L_i)$	represents the prosodic model.

Formulated in this way, Equation 2.6 expresses the formal set up of the LID problem where every component represents a distinct knowledge source used for language identification and can be modeled independently.

This framework is nowadays commonly and successfully applied for designing LID systems. The next sections give an overview of the approaches to automatic language identification from the aspect of the type of information used.

2.3.2 Acoustic Systems

Purely acoustic LID aims at capturing the essential differences among languages by modeling distributions of spectral features directly. This is typically done by extracting a language-independent set of spectral features from segments of speech and using a statistical classifier to identify the language-dependent patterns in such features.

The most preferred choice of spectral features is so-called **Mel-Frequency Cepstral Coefficients** (MFCC). MFCC features are computed during frontend analysis of input utterances using the following process: Segmented speech is transformed into frequency domain based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale. The lowest 13 cepstral coefficients and their first and second derivations form the cepstral feature vector. Additionally, noise reduction and channel normalization techniques

[4] In literature dealing with LID problems, a model that captures patterns of pronunciation contained in spectral features is usually referred to as an acoustic model and corresponding LID systems are called acoustic systems. This notation is also used throughout this thesis.

can be applied. MFCC features are used mainly in this thesis and a detailed description of their extraction is presented in Section 5.2.1.

The core modeling principle of most LID systems based on acoustic-phonetic information is the same: A feature vector is assumed to be generated according to a probability density function which is chosen depending on the specifics of the application.

The most preferred choice in many state-of-the-art acoustic LID systems are **Gaussian Mixture Models** (GMM). Under the GMM assumption, the probability is modeled as a weighted sum of multi-variate normal density (Gaussian) functions with parameters estimated on the training data (explained in more detail in Section 3.1). The earliest GMM LID system based on MFCC features and a maximum-likelihood decision rule was proposed by Zissman and is described in [150]. GMM is computationally inexpensive, does not require phonetically labeled training speech, and is well suited for text-independent tasks, where there is no strong prior knowledge of the spoken text. On the other hand, LID system based on GMM performs only static classification in the sense that the feature vectors are assumed to be independent from each other and feature vector sequences are not utilized.

To overcome this disadvantage and to capture language-discriminative information that resides in the dynamic patterns of spectral features, other LID systems have used **Hidden Markov Models** (HMM). The HMM is a finite set of states (a Markov chain of first order), each of which is associated with a probability distribution (typically designed by GMM). The state structure of HMM takes into consideration the sequential time behavior of the modeled speech, while the observation probability functions capture temporal acoustic patterns (more information is given in Section 3.2). HMM-based language identification was first proposed by House and Neuburg [56] who used symbol sequences derived from known phonetic transcriptions of text for training and testing. Later experiments tried to overcome the main problems of HMM-based LID systems: requiring of phonetic transcriptions for training data and performing training on unlabeled data. The results, when compared with static classifiers, were ambiguous: For example, Zissman [150] found that HMM trained in this unsupervised order did not perform better than GMM while Nakagawa et al. [90] eventually obtained better performance for their HMM approach than for their static system. Despite these uncertain results, the language identification community has preferred to go in the direction of improving the performance of GMM-based LID systems in order to meet rigid specifications in development costs.

During the last two decades, powerful algorithms for GMM structures have been proposed.

An attempt to incorporate temporal information from speech data was made by Torres-Carrasquillo et al. [138] who proposed using the so-called **Shifted Delta Cepstra** (SDC)

as spectral features for GMM systems. SDC features are created by stacking delta cepstra, computed across multiple speech frames (the computation of SDC features is illustrated in Section 3.1), and have become an essential part of the acoustic LID systems.

Furthermore, Torres-Carrasquillo and colleagues successfully combined SDC features with high-order GMM [129] (2 048 densities versus the 512 used earlier). Increasing the number of mixtures and the dimension of the feature vectors improves the performance of GMM-based LID systems. In the same time, it is time consuming for both training and testing. These costs can be however reduced by applying an adaptation technique from a large, common for all languages, GMM called the Universal Background Model (UBM). The UBM was proposed by Reynolds [114] for speaker verification and first applied by Wong for LID [146]. Under this approach, a single background GMM is trained from the entire training data and language-dependent models are adapted from the resulting UBM using the language-specific portions of the data.

Further refinement of GMM was achieved by Matejka et al. [83] using discriminative training criteria called Maximum Mutual Information (MMI). Here, an initial set of models was trained under the conventional maximum likelihood framework, which aims to maximize the overall likelihood of training data given the transcriptions. Then these models were discriminatively re-trained with the MMI objective function that maximizes the posterior probability of correctly recognizing all training utterances. That there is a clear advantage of the MMI over the standard maximum likelihood training framework was shown by Burget [17].

In parallel to the improvements of the estimation of GMM parameters, additional methods were proposed for increasing the quality of frontend processing. Since LID performance can be highly affected by speaker and channel variability, several attempts were made in order to reduce this source of influence:

- The RASTA (RelAtive SpecTrAl) filtering of cepstral trajectories, proposed for LID by Zissman [151], is used to remove slowly varying, linear channel effects from raw feature vectors.
- Vocal-Tract Length Normalization (VTLN) performs simple speaker adaptation as it is used in speech recognition [31]. Nowadays, after Matejka et al. [83], VTLN is a commonly used normalization technique for the LID task.
- Factor analysis techniques (Latent Factor Analysis (LFA) and Nuisance Attribute Projection (NAP), proposed by Vair et al. in [139]) are used to remove undesired variation coming from a low-dimensional source. Application examples in the model and the feature domain can be found, e. g., in [27, 28, 29, 137].

- Eigen-channel adaptation (in the feature domain) is used to compensate features in channel mismatch. The technique was introduced by Kenny [64] for speaker recognition and then adopted by Burget et al. [18] and used by the same authors for LID [82]. Eigen-channel adaptation can also be used in the model domain [82]: For each utterance both target language-specific GMM and UBM are adapted to channel conditions of test conversation.

The approaches described above can be successfully combined as, for example, Burget et al. did in [82]: Mean parameters of target language GMM are adapted from a UBM using features first compensated using eigen-channel adaptation in the feature domain and further re-estimated using MMI criterion.

GMM also have another application in performing language identification:

- GMM can be a part of a phonotactic LID system where they are used for tokenizing the speech into a discrete sequences of indices that are then modeled by a phonotactic component. One such system has been presented by Torres-Carrasquillo et al. in [138]. Phonotactic LID systems using GMM are described in the next section.
- GMM can be used as speech activity detectors by segmenting utterances into speech/non-speech regions at the preprocessing step as was done by BenZeghiba [12]. Since non-speech segments contain no language-specific information, Zissman [151] has found it desirable to train and test LID systems only on active speech.

Another successful LID approach uses **Support Vector Machines (SVM)**. The essence of SVM lies in representing the class-separating boundaries in a high-dimensional space in terms of few crucial points obtained from the training sample, termed support vectors. The selection is carried out to attain a maximum-margin solution best separating the classes. The transition from an original input space to the high-dimensional space is achieved by applying a certain function called sequence kernel, or dynamic kernel.

A number of sequence kernels have been proposed by different authors. The kernels differ in their feature expansion mechanism (basis function selection), feature transformations and normalizations, and in other kernel-specific steps. Some approaches are shortly described below:

- Campbell et al. [19] proposed a Generalized Linear Discriminant Sequence (GLDS) kernel that was first successfully applied for speaker recognition and then similarly implemented for the LID task [24]. The GLDS kernel performs feature expansion with monomial bases, followed by averaging and variance normalization of the expanded vectors.

- The GMM super-vector kernel proposed by the same authors [21, 25] is based on distance metrics between GMM models: First, a so-called super-vector that maps a segment of a language to a high dimensional space is obtained by appending the adapted mean value of all the Gaussians of a GMM in a single stream, after appropriate rescaling; then computed in such a way that supervectors are used as samples for training linear SVM classifiers.

- Probabilistic Sequential Kernel (PSK) frontend, suggested by Lee et al. [68], performs nonlinear mapping of SDC features. The PSK frontend is built upon a collection of Gaussian densities that are trained to represent elementary speech sound units characterizing a wide variety of languages. By means of this set of Gaussian densities, the variable-length speech utterances are transformed into higher dimensional characteristic vectors. An SVM then operates on these vectors to make classification decisions.

Currently, most SVM-based LID systems adopt their ideas from GMM-based systems. They can use MFCC and SDC features, channel and speaker-variability compensation techniques. SVM-based LID systems show excellent classification accuracy comparable to GMM and phonotactic LID systems described in the next section [27, 29, 82, 131, 137].

To summarize the overview of acoustic LID approaches, the following common tendencies can be mentioned:

- the typical frontend consists of SDC and MFCC features;
- GMM and SVM are used for modeling with additional improvement techniques;
- new techniques are consistently proposed to reduce speaker and channel variability.

2.3.3 Phonotactic Systems

Phonotactic information is one of the most widely used sources in LID. It corresponds to the constraints on the relative frequencies of sound units and their sequences in speech. The phonotactic LID system is build on the property of each language to have specific frequencies and dependencies between individual phones. Phonotactic LID systems are mostly based on the so-called Phone Recognition followed by Language Modeling (PRLM) architecture described in earliest LID publications from Zissman [150, 152], Hazen [49, 50], and Yan [147]. One or several single-language phone recognizers serve as the frontend, and the backend performs modeling of spoken language type. The number of recognizers is limited only by the number of languages for which labeled training material is available.

One of the advantages of the PRLM approach is the possibility to use a multilingual recognizer as Hazen did in [50] as the frontend or recognizers that are trained on available data from any language (even those not included in the identification task).

Phonotactic LID is still an active research area with two main problems:

1. The high accuracy of statistical models usually increases the complexity of models in terms of the number of model parameters that have to be estimated. Furthermore additional training data is required.
2. The quality of the PRLM system is influenced by the recognition accuracy of the sound units from continuous speech.

Despite these shortcomings, phonotactic systems have high discrimination power, on the one hand, and relative simple implementation and low system development costs on the other.

In the following, the details of state-of-the-art phonotactic LID will be presented. They differ from each other by the implementations of the decoding frontend and language modeling backend.

The goal of the frontend in phonotactic LID systems is to represent the spoken utterances as sequences of discrete "tokens" that have to cover the phones across languages in an appropriate way. State-of-the-art tokenizers are based either on HMM with probability distributions modeled by GMM [8, 9, 45, 135], or on HMM/Artificial Neural Network (ANN) hybrids [113, 117, 128]. One of the most popular recognizers is implemented by Brno University of Technology [128] which is based on HMM/ANN approach and is trained on the data from the Hungarian language that has a large phoneme inventory. The tokenizer uses ANN to estimate posterior probabilities of phones from Mel filter bank log energies using split left and right contexts around current frame. The phonotactic LID subsystems based on Brno's recognizer are among the most successful ones [46, 82, 137]. Recently, powerful phoneme recognizer was developed at the Technical University of Munich [144]. Proposed in [144] system for continuous speech recognition uses phoneme predictions generated by a bidirectional Long Short-Term Memory recurrent neural network that are observed by an HMM, in addition to conventional speech features. This architecture has shown high word accuracy for large-vocabulary continuous speech recognition in a challenging spontaneous and noisy speech scenario that are especially important for the LID tasks.

The language modeling backend is often realized using so-called N-grams that model subsequences of N items from a given sequence, where $N = 1, 2, \ldots$. The models are built by collecting N-gram statistics (counts) and computing their conditional probabilities either from phoneme strings generated by phoneme recognizers (as was done in first phonotactic

systems [49, 85, 152]); or from the phoneme lattices that allow taking into account the alternative phoneme recognition paths, giving useful information and improving the performance [45]. Despite the fact that the complexity of N-gram models grows exponentially with its order, current phonotactic LID systems use up to 4-grams [82].

Recently, some approaches that improve the quality of backend models of phonotactic LID systems were implemented:

- Discriminative MMI training was proposed for bigrams ($N = 2$) as it was done by Kuo for speech recognition [65]: The bigram probabilities are iteratively re-estimated using a gradient descent algorithm to optimize the MMI objective function [82].
- Binary decision trees based on a single language independent tree and adapted to individual data were proposed by Navratil [92, 93, 94] as another method for modeling of language type in phonotactic systems [22, 82].
- The high order N-grams (their usage was limited due to the great requirements in training data) became possible with an adaptation scheme proposed for binary decision trees and adopted to the trigrams ($N = 3$) based on lattice counts [46, 82].
- In the case where the available training data is coming from different sources, the technique of multiple models was applied by different sites [46, 82, 137]: Instead of training one model per language, the data is clustered according to databases/dialects/etc., and several separate models are trained on these clusters.
- An SVM classifier trained on the N-gram lattice counts coming from phoneme recognizer was used by Campbell et al. [23] and Matejka et al. [82]. In order to use N-grams with higher order, Richardson [116] has modified this approach via an alternative filter-wrapper feature selection method. These so-called keyword SVM allowed the creation of a 4-gram SVM system that showed a significant improvement in performance over the initial trigram system [137].
- Another SVM approach, proposed by Campbell [20], is based on the so-called Term Frequency Log-Likelihood Ratio (TFLLR) kernel. Here the frequencies of N-grams computed within one sequence and normalized by the square root of corresponding frequencies in the whole training set are appended in a single vector to obtain a TFLLR kernel. A linear SVM model of a target language is trained by using the vectors computed for each segment of the target language as positive examples and the set of vectors of all the other language segments as negative examples [29].
- Vector Space modeling (VSM) as a backend was proposed by Li et al. [71]: High-dimensional feature vectors of phoneme N-gram probability attributes (known as

bag-of-sounds vectors) are created for each of the phoneme sequence coming from phoneme recognizers and then stacked to form a composite vector on which the SVM is applied.

In general, comparing the approaches described above, the following conclusions can be drawn:

- The phoneme recognizer from Brno University of Technology [128] is in the leading position among existing frontends used for phonotactic LID;
- N-gram models and binary trees as backends have very similar performance, but the trees are more efficient;
- SVM outperform standard N-grams;
- Multiple models per language are beneficial when training data represent different dialects, groups of speakers, and/or databases.

2.3.4 Prosodic Systems

Along with acoustic and phonotactic LID systems, the prosodic information can be used to discriminate languages. Despite the fact that infants (as Ramus has demonstrated in [112]) can use prosodic information to distinguish among languages, LID systems based solely on prosody are relative rare. Most prosody-based LID systems capture the duration, the pitch pattern, and the stress pattern in a language and are described below.

One of the first approaches utilizing only the prosodic information was proposed by Itahashi et al. [58, 59] and used fundamental frequency (F0) contours of speech. In this system, fundamental frequency and energy contours were extracted from the speech signal and approximated via a piecewise-linear function. First, the F0 contour was approximated by a set of polygonal lines so that the mean square error between the lines and F0 values was minimized; the optimum boundaries of the lines were determined using a dynamic programming procedure. The starting frequency, slope, and duration of each line were calculated. Seventeen parameters, including mean values and standard deviations of the above parameters, were used for the analysis. Then parameters derived from F0-patterns were analyzed using principal component analysis. The system was then evaluated on speech data from six languages. The resulting identification rates ranged between 70 and 100 %. However, test data for every language contains data from five speakers only.

Several important conclusions were drawn from the work of Thyme-Gobbel and Hutchins, reported in [132]. The paper [132] presents probably the most thorough study of prosodic

features for LID. The authors compared 220 prosodically motivated features that were calculated from syllable segments represented as histograms and used for classification with a likelihood ratio detector. The experiments involved a series of pair-wise language classification tasks and showed the following:

- prosodic features can be useful in LID;
- fundamental frequency (F_0) generally seems to be more useful than amplitude information;
- usefulness of individual features and their combinations varies strongly with specific language pairs.

Comparable results were reported by Cummins [34] who pursued a more data-driven approach for the extraction of prosodic features to determine the usefulness of individual features. The approach used Long Short-Term Memory model by applying differenced log-F0 and amplitude envelope information. It showed that automated selection of suitable prosodic features is feasible, but a purely prosodic set of features may lead to merely moderate results if used as the only information source.

In addition to the small number of purely prosodic LID systems, many researchers use certain prosodic features as a component in combination with other evidence, such as spectral or phonotactic information. For example, Hazen has proposed an approach [49, 50] where prosodic information was modeled assuming that fundamental frequency contours are independent of the phone durations and the phonetic string:

- probability distributions were created to model the number of frames within a segment for each phonetic class of each language, and
- F0 and delta F0 were measured for each language.

In most studies, the prosodic information is statically processed, and the frame or segment-based likelihood score is simply accumulated over an utterance. The precise dynamics of prosody is largely ignored. In contrast, Adami [2] proposed an approach that uses temporal trajectories of the fundamental frequency and short-time energy to segment the speech signal into a small set of discrete units. The segmented trajectories are then quantized and labeled into a set of classes that describes the dynamics of fundamental frequency and energy contours. Being an extension of the phonotactic approach, Adami's approach takes into account the inter-segmental dynamics.

Trying to consider the intra-segment dynamics, Obuchi introduced prosodic HMM [100] for classifying the prosodic segments dynamically. Feature vectors are formed using the

power, F0, and reliability of the F0 estimate, computed for each frame. Since it is difficult to obtain multi-language corpora with prosodic labels, a data-driven clustering technique was used to create language-dependent prosodic HMM by unsupervised training.

A similar approach, described by Nagarajan [89], proposed creating phonetic HMM of syllable-like units. Each utterance is segmented into syllable-like units using a group delay function of minimum phase signal. Similar syllable segments are then grouped together and syllable models are trained incrementally. The language-dependent syllable models are used then to perform language identification based on accumulated acoustic log-likelihoods in a framework similar to the parallel phoneme recognition approach.

Another prosody-based LID system that uses information extracted from pitch contours was introduced by Lin and Wang in [72]. This method utilizes the autocorrelation function to detect vocalic segments and find pitch candidates. A Viterbi algorithm [107] is used to find the most suitable contour path that later is approximated by a set of Legendre polynomial coefficients and modeled by a Gaussian mixture model. The system was tested on the pair-wise language identification task and further improved with a novel dynamic model in ergodic topology in order to take advantage of the temporal information across several pitch patterns [74].

Tong et al. [136], along with spectral and phonotactic features, have explored such prosodic features as phoneme duration and pitch. For a given utterance, 11 dimensional pitch features are extracted from each frame. GMM models for each target language are obtained by adaptation from a universal background model (as described in Section 2.3.2 for spectral GMM systems) that is trained on the feature vectors from all languages. The duration component is also modeled by GMM. Duration statistics are obtained for each phoneme resulting in a duration feature vector that has three elements representing the duration of three states in a phoneme.

Mary and Yegnanarayana [80] hypothesized that prosody is linked to linguistic units such as syllables and is manifested in terms of changes in such measurable parameters as fundamental frequency, duration and energy. The authors proposed segmenting continuous speech into syllable-like units by automatically locating the vowel onset points (VOP) from the Hilbert envelope of the Linear Predictive (LP) residual of the speech signal. The region between two successive VOPs is considered as a syllable-like region, and parameters are derived to represent duration, dynamics of F_0 contour and energy variations corresponding to each region. The direction of the F_0 change, either rising or falling, is used to model intonation characteristics. The syllabic rhythm is represented by the distance between successive VOPs and by the duration of voiced region within each syllable-like region. Stress is modeled using change in log energy corresponding to the voiced regions of a syllable, along with the F_0 contour and duration features. In order to preserve temporal dynamics of

prosodic parameters, the context of a syllable, i. e., characteristics of the preceding and the succeeding syllable along with present one, is taken as input for a multilayer feedforward neural network.

The investigations of speech rhythm for LID were started with hand-labeled data and followed by automatic modeling of rhythm features from pseudo-syllabic segments utilizing a GMM-based classifier: So-called pseudo-syllables are derived from the most frequent syllable structure in the world, the Consonant-Vowel (CV) structure; rhythmic units matching the CV-structure are extracted automatically using a vowel detection algorithm and then learned using Gaussian Mixtures. The experiments were first performed by Farinas et al. on read speech [41, 123] and then extended by their colleagues to spontaneous speech [124]. In more recent investigations, Rouas [121, 122, 125] modeled prosody dynamics by the separation of long-term and short-term components of prosody. The long-term component characterizes prosodic movements over several pseudo-syllables while the short-term component represents prosodic movements inside a pseudo-syllable. Unfortunately, the proposed approach, originally designed on read speech and showing promising results on this data, achieved poor performance on spontaneous data.

Summarizing the results of the above described prosody-based systems one can conclude that:

- most investigations are focused on the utilization of pitch information for the LID task;
- prosodic cues are very useful in automatic LID in combination with spectral and phonotactic information but still require further investigation.

2.3.5 Fusion of Different LID Systems

Most state-of-the-art language recognition systems usually consist of several individual subsystems. A challenge is to find a combination (or fusion) technique that allows the final system to efficiently use the complementary information of every subsystem. The main goal is to benefit from all available cues in order to improve system performance.

Currently, several different fusion approaches have produced remarkable improvements when compared to a single system. The basic idea behind these approaches involves applying weighting coefficients to the likelihood scores produced by individual LID systems. One of the most straightforward fusion techniques of this type is linear score weighting with the weights found as values with the highest performance on the development data set. An example of such linear combination of LID systems can be found in publications

from Wong and Sridharan [145] and Cernoky [30]. The scores in linear combination can also be equally weighted as in the LID system from Campbell et al. [23].

Most language recognition systems use fusion that is performed using a Gaussian classification backend (as in Singer et al. [129], Torres-Carrasquillo et al. [138], Campbell et al. [24]): The language-specific scores from all subsystems are processed through the GMM and then converted to log-likelihood ratios for final evaluation. More recent investigations can be found in [137].

Gutierrez et al. [48] calculated the performance confidence indexes derived from each LID system and applied theory of evidence to perform the fusion process.

Obuchi and Sato [100] have combined the scores of primary systems using linear discriminant analysis. The local identification results can also be fused using a Bayesian framework. Under this framework, proposed by Lin and Wang [73], local decisions, the associated false-alarm and miss probabilities are fused via a Bayesian formulation to make the final decision. Fusion suggested by Tong et al. [136] was carried out by multiplying the log-likelihood ratios from individual classifiers. Individual scores can also be combined by means of a linear SVM backend to produce the final score as introduced by White et al. [143] and later used by Castaldo et al. [29].

In contrast to the approaches described above, where the weighting coefficients are only different for the varied subsystems, Yin et al. [148] introduced a language-dependent fusion technique. With this technique varied weighting coefficients are applied to not only each subsystem but also to each language. Furthermore, weighting coefficients are calculated from the performance of all possible language pairs, which reflect the differences among the languages.

A new fusion tool called FoCal was recently proposed by Brummer [16] and consists of a collection of utilities that provides discriminative fusion with either

- a logistic regression objective which is stated in terms of a cost, which must be minimized, or
- generative Gaussian backends.

FoCal will be used as a fusion mechanism for the combination of LID systems described in this thesis and therefore will be presented in more detail in Section 5.6.

Regardless of the chosen fusion methods, the combination of LID systems which utilize different models or features shows better final evaluation results than individual systems and has become a standard strategy of the current state-of-the-art LID systems.

2.3.6 Summary of Current LID Trends

Due to the steadily increasing interest in automatic LID, the intense research efforts have resulted in significant progress over the last three decades. At the beginning, the field of spoken language ID suffered from the lack of a common, public-domain multilingual speech corpus that could be used to evaluate and compare different approaches. The first multilingual database for LID research was created by the Center for Spoken Language Understanding at the Oregon Graduate Institute (OGI)[5] and consisted of telephone speech in 11 languages. The release of the OGI database published by Muthusamy in 1992 [87] noticeably increased the volume of technical publications on LID and entailed first attempts to compare different approaches for language identification [86, 151]. These papers not only reported the current results on the research, development, and evaluation of LID systems but also have shown the need for a standardized, multilanguage speech corpora available for training and testing LID systems. Using such corpora under carefully controlled conditions should allow the comparison of LID systems and therefore enhance the LID research.

The Language Recognition Evaluation (LRE) campaign was first performed by the National Institute of Standards and Technology[6] (NIST) in 1996 and continued in 2003, 2005, 2007, and 2009. The evaluation data mostly comes from the CallFriend corpora provided by the Linguistic Data Consortium[7]. The goal of the NIST language recognition evaluations is to quantify performance of LID systems for conversational telephone speech using uniform evaluation procedures. NIST provides a common base for comparison of LID systems on a well-defined task: The participants are provided development and evaluation data and a plan that regulates the conditions of the evaluation. The results are examined by NIST in order to obtain a single-valued measure of the systems' accuracy.

According to the results of the latest NIST evaluations, the most successful LID systems consist of a combination of several spectral and phonotactic subsystems that use different approaches:

1. LID systems based on phonotactic information use the standard PRLM approach with one or multiple lattice decoders. The most popular one is the phoneme recognizer from Brno University of Technology [128]. The phoneme sequences are modeled then by trigrams (or up to 4-grams), support vector machines, or binary trees.

2. LID systems based on the pure spectral information that use Gaussian mixture models and support vector machines have achieved very high recognition accuracy and have become as effective as phonotactic ones.

[5] http://cslu.cse.ogi.edu/
[6] http://www.nist.gov/speech/tests/lang
[7] http://www.ldc.upenn.edu

It is generally believed that spectral and phonotactic features provide language cues complementary to each other [129]. They are both easy to obtain, but phonotactic features, which are more robust against effects such as speaker and channel, become a trade-off between computational complexity and performance. In addition, phonotactic features that cover higher level linguistic elements, such as syllables or words, carry more language discriminative information than a phoneme. However, practical application of phonotactic system is limited by the minimum duration of test utterances, which lies between five and ten seconds, and therefore is highly dependent on the concrete LID task.

NIST LRE also discovered a lack of prosody-based LID systems. Though the importance of prosodic information, such as duration and pitch, has long been acknowledged, hardly any system among the best ones was based on prosodic features. Hence, the question of whether prosody can be effectively used in LID systems is still open. A robust modeling of prosodic features for the language identification task remains a considerable challenge.

2.4 Objective of the Research

As shown in Section 2.1.2, there are three major kinds of linguistic cues that can be used by humans and machines to identify a language:

- acoustic properties of sound units (phonemes) and their frequencies of occurrence;
- phonotactic properties of phoneme sequences;
- prosodic properties such as rhythm and intonation.

Language recognition systems perform well in favorable acoustic conditions, but their performance may degrade due to noise and unmatched acoustic environment. Prosodic features, derived from pitch, energy, and duration are relatively less affected by channel variations and noise [132]. Though LID systems based on spectral features outperform the prosody-based systems, they can be combined to improve performance and gain the needed robustness.

Meanwhile, speech rhythm is actually one of the most promising prosodic features to be considered for LID since it may be sufficient for humans to perceptually identify some languages. As it follows from the overview of existed prosody-based LID systems given in Section 2.3.4, rhythm is currently the least explored information source for the LID task. The reason is that using rhythm is not a straightforward issue, both in terms of its theoretical definition and its automatic processing.

Speech rhythm has been under investigation for a while, resulting in several rhythm-oriented theories. These are given in Section 4.1. All these considerations emphasize the potential use of an efficient rhythm model and the difficulty in its creation.

Therefore, in this thesis the speech rhythm is investigated with the idea of using it for discriminating among languages. The definition of rhythm suitable for automatic processing within a language identification system is proposed. The importance of speech rhythm for discriminating among languages is investigated.

The basic goal of this thesis is to show how the performance of LID systems based on spectral and/or phonotactic information can be improved by using speech rhythm. It can be split into the following tasks:

- propose and implement LID systems based on spectral and phonotactic information that will be used as the baseline systems;
- find a suited model for using rhythm: define basic rhythmic units and develop a language independent approach for segmenting speech and modeling of rhythm;
- propose and implement an LID system based on rhythm;
- find a suitable method to merge individual LID systems;
- experimentally evaluate the performance of the different systems and their combinations with the speech rhythm system;
- explore the impact of speech rhythm on LID tasks.

3

Classification Methods

This chapter presents in detail the methods that are used to model language-specific characteristics and to classify the languages in the experiments described later in this thesis. Gaussian Mixture Models and Hidden Markov Models are applied to model spectral properties of languages. To design LID systems based on phonotactic and rhythm information, the N-gram models are used to model respectively sequences of phonemes and syllable-like units. Finally, this chapter provides the description of artificial neural networks that serve as additional post-classifiers for individual LID systems and as a fusion mechanism for combining the results from several systems.

3.1 Gaussian Mixture Models

Recent research indicates that the most successful acoustic LID systems are based on Gaussian Mixture Models (GMM) that classify languages using the spectral content of the speech signal. A GMM is a probabilistic model for density estimation using a mixture distribution and is defined as a weighted sum of multi-variate Gaussian densities:

$$p(\vec{x} \mid \Lambda) = \sum_{i=1}^{M} w_i N_i(\vec{x}), \tag{3.1}$$

where $\vec{x} = x_1, \ldots, x_D$ is an observation vector with dimensionality D,
Λ is a set of model parameters,
M is a number of mixture components,
w_i (with $i = 1, \ldots, M$) is a mixture weight constrained so that $\sum_{i=1}^{M} w_i = 1$,
$N_i(\vec{x})$ with $i = 1, \ldots, M$ are the multi-variate Gaussian densities defined by a mean $D \times 1$ vector $\vec{\mu}_i$ and a $D \times D$ covariance matrix Σ_i in the following way:

$$N_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)'(\Sigma_i^{-1})(\vec{x} - \vec{\mu}_i)\}. \tag{3.2}$$

The complete GMM is then parametrized by the mean vectors, covariance matrices, and mixture weights from all component densities. It is represented by the notation:

$$\Lambda = \{w_i, \vec{\mu}_i, \Sigma_i\}, i = 1, \ldots, M. \tag{3.3}$$

The GMM can have several different forms depending on the choice of covariance matrices. The following possibilities could be used:

- one covariance matrix per Gaussian component;
- one covariance matrix for all Gaussian components in the mixture (grand covariance);
- or a single covariance matrix shared by all models in a set (global covariance).

The covariance matrix can also be full or diagonal. The diagonal covariance matrices are widely used in speech processing systems due to the fact that diagonal covariance GMM are computationally more efficient for training and sometimes outperform full matrix GMM.

The whole set of GMM parameters Λ has to be estimated during training so that it best matches the distribution of training data. The most popular and well-established technique available to determine the parameters of a GMM is the Maximum Likelihood Estimation (MLE). The aim of an MLE is to find the model parameters which maximize the likelihood of the GMM, given training data. For a sequence of T training vectors $\mathcal{X} = \vec{x_1}, \ldots, \vec{x_T}$, the GMM likelihood can be found as

$$p(\mathcal{X} \mid \Lambda) = \prod_{t=1}^{T} p(\vec{x_t} \mid \Lambda), \tag{3.4}$$

where $p(\vec{x_t} \mid \Lambda)$ is computed using Equation 3.1. To compute the Maximum Likelihood (ML) estimate, the iterative Expectation-Maximization (EM) algorithm is used. The basic idea of the EM algorithm is to estimate a new model $\overline{\Lambda}$, such that $p(\mathcal{X} \mid \overline{\Lambda}) \geq p(\mathcal{X} \mid \Lambda)$, where Λ is an initial model. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached.

Each iteration of the EM algorithm consists of two processes:

1. In the expectation, or E-step, the expected value of the log-likelihood function is calculated given the observed data and current estimate of the model parameters.

2. The M-step computes the parameters which maximize the expected log-likelihood found on the E-step. These parameters are then used to determine the distribution of the latent variables in the next E-step until the algorithm has converged. Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration where the following re-estimation formulas are used:

   - Mixture weights:
   $$\overline{w}_i = \frac{1}{T}\sum_{t=1}^{T} p(i \mid \vec{x}_t, \Lambda); \tag{3.5}$$
   - Means:
   $$\vec{\mu}_i = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \Lambda)\vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \Lambda)}; \tag{3.6}$$
   - Variances:
   $$\vec{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \Lambda)\vec{x}_t^2}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \Lambda)} - \vec{\mu}_i^2. \tag{3.7}$$

   The a-posteriori probability is given by

   $$p(i \mid \vec{x}_t, \Lambda) = \frac{w_i N_i(\vec{x}_t)}{\sum_{k=1}^{M} w_k N_k(\vec{x}_t)}. \tag{3.8}$$

The critical factors in training a GMM are selecting the order M of the mixture and initializing the model parameters prior to the EM algorithm. Using mixture models with high number of components results in increasing of the system performance from one side and in increasing its complexity from the other side. Different initial parameters can lead to a convergence of the algorithm in different local maxima that may also have influence on the recognition performance of the resulting models. The most often used methods include:

- Initialization by randomly found parameters;
- Initialization of the GMM based on mean and variance of the feature distribution from training data;
- Initialization of the GMM with different clustering algorithms, e. g., hierarchical EM-clustering that is used in this thesis.

In most applications the best suitable approach is determined experimentally for a concrete task.

Current GMM-based LID systems utilize SDC features and show better performance than those based on the standard MFCC [138, 83]. Since this approach was taken to represent the pure spectral LID systems in this thesis, in the following, the SDC features are described.

SDC features

Using SDC features for the language identification is motivated by the idea of incorporating additional temporal information about the speech signal into the feature vectors as it is usually done in phonotactic approaches that naturally base their tokenization over multiple frames. As Torres-Carrasquillo et al. have shown in [138], the performance of GMM-based LID systems that use SDC features is comparable to the performance of the phonotactic LID systems. And since GMM-based systems are computationally more efficient, they have become one of the most popular approaches to language identification.

SDC features are obtained by stacking delta cepstra computed across multiple speech frames. The SDC features are specified by a set of four parameters N, d, P, k (usually denoted as N-d-P-k), where:

- N is the number of cepstral coefficients computed at each time frame,
- d is the time advance and delay for the delta computation,
- k is the number of blocks whose delta coefficients are concatenated to form the final feature vector, and
- P is the time shift between consecutive blocks.

For each frame of data, MFCC are calculated based on N including c_0 (i. e., the coefficients $c_0, c_1, \ldots, c_{N-1}$). The components of the SDC vector at time t are computed as follows:

$$\Delta c(t, i) = c(t + iP + d) - c(t + iP - d), \tag{3.9}$$

where $i = 0, \ldots, k - 1$. The computation of SDC features is illustrated in Figure 3.1.

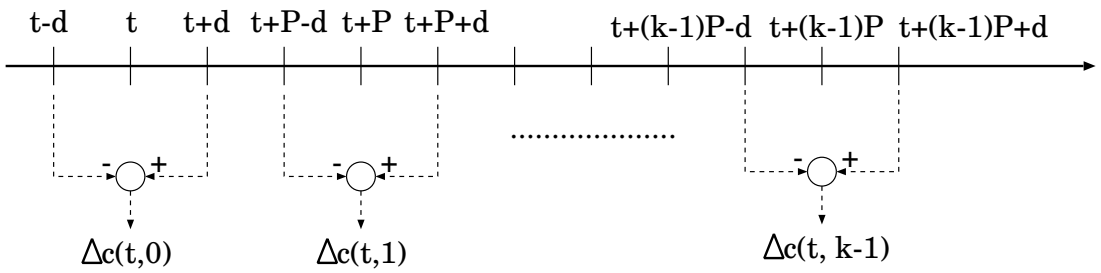


Figure 3.1: *Computation of an SDC feature vector for a given time t*

3.2 Hidden Markov Models

A Hidden Markov Model (HMM) is a statistical model which outputs a sequence of symbols or quantities. The HMM is a finite set of states, each state is associated with a (generally multidimensional) probability distribution. Transitions among the states are defined by a set of probabilities called transition probabilities. In a particular state, an outcome or observation can be generated according to the associated probability distribution. Only the emission, not the state, is visible to an external observer and therefore states are "hidden" from the outside.

In order to define an HMM completely, the following elements are needed [107]:

- S, the number of states in a model;
- $\mathcal{Q} = \{q_1, \ldots, q_S\}$, a set of states;
- R, the number of observation symbols in the alphabet;
- $\mathcal{V} = \{v_r\}$, $r = 1, \ldots, R$, a discrete set of possible symbol observations;
- $\mathcal{A} = \{a_{s,s'}\}$, a set of state transition probabilities, where $a_{s,s'} = P(q_{s'} \mid q_s)$, $q_{s'}$ is a state at time $t+1$ and q_s is a state at time t;
- $\mathcal{B} = \{b_s(r)\}$, a set of emission probability distributions, where $b_s(r) = P(v_r \mid q_s)$;
- $\pi = \{\pi_s\}$, a set of initial state distributions, where $\pi_s = P(q_s)$ at time $t = 1$;
- T, the length of the observation sequence.

The emission and transition probabilities should satisfy the following stochastic constraints:

- $a_{s,s'} \geq 0,\ 1 \leq s, s' \leq S$ and $\sum_{s'=1}^{S} a_{s,s'} = 1,\ 1 \leq s \leq S$;
- $b_s(r) \geq 0,\ 1 \leq s \leq S,\ 1 \leq r \leq R$ and $\sum_{r=1}^{R} b_s(r) = 1,\ 1 \leq s \leq S$.

Usually, continuous observations are used: Each observation v_r is represented by the feature vector $\vec{x}_t$. Instead of a set of discrete emission probabilities, a continuous probability density function is used. For every state s, it is modeled by a GMM as introduced in Equation 3.1:

$$b_s(\vec{x}) = \sum_{i=1}^{M_s} w_{s,i} \cdot N_i(\vec{x} \mid s), \tag{3.10}$$

where $\vec{x}$ is a D-dimensional feature vector;
M_s is the number of mixture components for state s;
$w_{s,i}$ is a mixture weight coefficient that satisfies the following:

- $w_{s,i} \geq 0,\ 1 \leq i \leq M_s,\ 0 \leq s \leq S$,
- $\sum_{i=1}^{M_s} w_{s,i} = 1,\ 0 \leq s \leq S$;

To simplify the computation, the Equation 3.10 is approximated using the best mixture component assumption:

$$b_s(\vec{x}) \approx \max_i w_{s,i} \cdot N_i(\vec{x}|s). \tag{3.11}$$

The Gauss probability density function $\mathcal{N}_i$ for the mixture i is represented by a mean vector $\vec{\mu}_i$, where all vector components are distributed equally and independently from each other with one globally defined variance σ:

$$N_i(\vec{x}, \vec{\mu}_{s,i}) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^D} \prod_{d=1}^{D} \exp\left(-\frac{(x_d - \mu_{s,i,d})^2}{2\sigma^2}\right), \tag{3.12}$$

where $\vec{\mu}_{s,i} = (\mu_{s,i,1}, \mu_{s,i,2}, \ldots, \mu_{s,i,D})$.

With the notations introduced above, an HMM with continuous densities is described by a parameter set Λ:

$$\Lambda = \{\pi_s, a_{s,s'}, w_{s,i}, \vec{\mu}_{s,i}\}. \tag{3.13}$$

Then, the probability given optimal path Θ of $\mathcal{X}$ in Λ is defined in the following way:

$$P(\mathcal{X}, \Theta \mid \Lambda) = \pi_{\theta_1} b_{\theta_1}(\vec{x}_1) \cdot \prod_{t=1}^{T-1} a_{\theta_t,\theta_{t+1}} b_{\theta_{t+1}}(\vec{x}_{t+1}), \tag{3.14}$$

where $\Theta = \{\theta_1, \theta_2, \ldots, \theta_T\}$ is an optimal path (with maximum probability);
θ_t is a state at time t according to Θ.

The emission and transition probabilities always have to be multiplied. For the sake of computation simplicity all probabilities are handled on logarithmic scale so that every multiplication simplifies to an addition (which can be calculated faster). The negative logarithmic counterpart for a probability is called score (or penalty). To make the computation of emission probabilities easier, the logarithm is also multiplied by $2\sigma^2$. Thus, the probability of the optimal path is calculated as so called negative log-likelihood score (neglog-likelihood):

$$g(\mathcal{X}, \Lambda) = -2\sigma^2 \log P(\mathcal{X}, \Theta \mid \Lambda). \tag{3.15}$$

Using the definition of the probability from Equation 3.14, the neglog-likelihood score is defined in the following way:

$$g(\mathcal{X}, \Lambda) = -2\sigma^2 \log \pi_{\theta_1} + \sum_{t=1}^{T-1}(-2\sigma^2 \log a_{\theta_t,\theta_{t+1}}) + \sum_{t=1}^{T-1}(-2\sigma^2 \log b_{\theta_t}(\vec{x}_t)). \tag{3.16}$$

The expressions $-2\sigma^2 \log \pi_{\theta_1}$ and $-2\sigma^2 \log a_{\theta_t,\theta_{t+1}}$ are called initialization and transition penalties, respectively. The emission score is computed in the following way:

$$\begin{aligned}
-2\sigma^2 \log b_s(\vec{x}) &\approx -2\sigma^2 \log \left(\max_i c_{s,i} \cdot N_i(\vec{x}, \vec{\mu}_{s,i}) \right) \\
&= \min_i \left\{ -2\sigma^2 \log \left(c_{s,i} \frac{1}{\left(\sigma\sqrt{2\pi}\right)^D} \prod_{d=1}^{D} \exp\left(-\frac{(x_d - \mu_{s,i,d})^2}{2\sigma^2} \right) \right) \right\} \\
&= \min_i \left\{ -2\sigma^2 \log \frac{c_{s,i}}{\left(\sigma\sqrt{2\pi}\right)^D} + \sum_{d=1}^{D} (x_d - \mu_{s,i,d})^2 \right\} \\
&= \min_i \left\{ -2\sigma^2 \log c_{s,i} + \mid \vec{x} - \vec{\mu}_{s,i} \mid^2 + 2\sigma^2 D \log \sigma\sqrt{2\pi} \right\}.
\end{aligned}$$

The expression $2\sigma^2 D \log \sigma\sqrt{2\pi}$, which is a constant value, and the Gaussian weight penalties $-2\sigma^2 \log c_{s,i}$ are usually pre-calculated. Therefore only the distances between feature vectors and mean vectors must be computed.

In order to use such probabilistic models for the recognition process, it is necessary to specify the parameters of the probability density functions so that the phonemes will be identified correctly. These model specific parameters, typically mean vectors, can be

defined using an initialization process followed by some training procedure.

The initialization and supervised training procedures require the transliterated and labeled speech training data as well as the phonetic lexicon for every language. Orthographic transliteration means that the signal-phoneme correspondences of the utterances are known. Labeled data means that every sample speech file containing an utterance is accompanied by a label file describing its segmentation on phoneme level (i.e. time information).

The initialization starts from the definition of the model topology (the number of states, allowable transitions). A phoneme is represented by a sequence of states. Every state has a variable emission probability defined by a density function and constant transition probabilities. Three-state left-to-right HMM are used to model normal phonemes and one state is used to model silence. The first and the third state represent the transitions to the neighboring phonemes and the middle state represent the center of a phoneme. Figure 3.2 gives an example of this kind of model, which has a good capability of modeling co-articulation effects.

Any phoneme sequence represented by a corresponding sequence of feature vectors is constructed by concatenating phoneme models. Starting from the left most state, the first feature vector is inserted into the corresponding probability density function to calculate the emission probability. Then, according to the possible transitions to the next state, some new paths are opened. For every path, the state emission probability is multiplied with the corresponding transition probability. In this way, a lattice of pronunciation possibilities for the corresponding phoneme is generated. To produce pronunciation variations of whole phoneme sequences it is possible to chain the states of phoneme models together. A silence model precedes and follows each sequence.

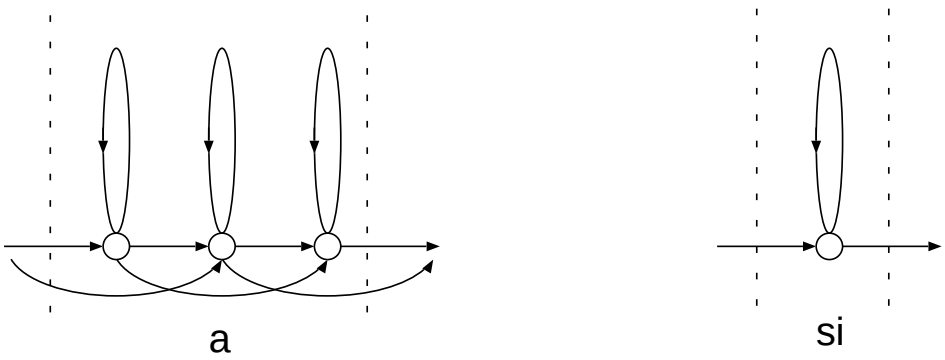


Figure 3.2: *Structure of the phoneme model "a" and the silence model "si"*

The initialization of model parameters starts from initial state probabilities that are defined as:

$$\pi_1 = 1,\ \pi_s = 0,\ (2 \leq s \leq S).$$

The estimation of transition probabilities $a_{s,s'}$ is usually performed according to the statistical analysis of the training utterances. Thus, the $a_{s,s'}$ are set to fixed values for all models.

The main issue for the initialization process is the estimation of probability density functions parameters - mixture weights $c_{s,m}$ and mean vectors $\vec{\mu}_{s,m}$. Normally the variance σ should be re-estimated too, but since in this thesis the Gauss probability density functions use one globally defined variance, σ is set to a constant value for all models. Every segment of a phoneme has its own probability density function, for which a set of parameters must be defined.

The goal of the initialization process is to find an initial set of such parameters for every segment. The required labeling is performed automatically using the recognition algorithm called Forced Viterbi algorithm [108, 107] and a phonetic lexicon. From a large amount of labeled training data, it is possible to derive large collections of prototype feature vectors for every phoneme. With these phoneme specific prototype collections, the initialization algorithm based on on-line-clustering [7] is able to estimate the parameter sets for every segment.

Initialized in such a way, parameters can be updated further by Forced Viterbi Training on the same labeled training data. The training procedure utilizes the Maximum Likelihood (ML) method to find the parameters that make the observed data most likely. Formally, if Λ, defined in Equation 3.13, presents the model parameters, the outcome of ML estimation can be expressed as follows:

$$\Lambda_{ML} = \underset{\Lambda}{\operatorname{argmax}}\, P(\mathcal{X} \mid \Lambda), \tag{3.17}$$

where the objective function $P(\mathcal{X} \mid \Lambda)$ is the likelihood of the parameters Λ given the training pattern $\mathcal{X}$.

To find the maximum likelihood estimates of parameters Λ, the Forced Viterbi algorithm is used. The algorithm uses the "best path" approach to determine the neglog-likelihood of the model generating the observation sequence. The likelihood measure is based on the assumption that an HMM generates the observation sequence $\mathcal{X} = \{\vec{x}_1, \ldots, \vec{x}_T\}$ by using the best of possible sequences of states $\Theta = \{\theta_1, \theta_2, \ldots, \theta_T\}$ of the model Λ.

To describe the Viterbi algorithm the following notations are used:

$s,\ s'$	state indices,
$1 \leq t \leq T$	frame index (time step),
$\delta_t(s)$	local score,
$g(\mathcal{X}, \Lambda)$	cumulative Viterbi score,
$A_{s,s'} = -2\sigma^2 \log a_{s,s'}$	transition penalty from state s to s',
$\Pi_s = -2\sigma^2 \log \pi_{\theta_1}$	initialization penalty for state s,
$B_s(\vec{x}_t) = -2\sigma^2 \log b_s(\vec{x}_t)$	emission penalty,
$\psi_t(s)$	predecessor state index of state s at time step t.

The complete state sequence determination procedure using the Viterbi algorithm can be stated in the following steps:

1. Initialization for t=1 and $1 \leq s \leq S$

$$\begin{aligned} \delta_1(s) &= \Pi_s + B_s(\vec{x}_1) \\ \psi_1(s) &= 0 \end{aligned}$$

2. Recursion for $2 \leq t \leq T$ and $1 \leq s \leq S$

$$\begin{aligned} \delta_t(s') &= \min_{1 \leq s \leq S} \{\delta_{t-1}(s) + A_{s,s'}\} + B'_s(\vec{x}_t) \\ \psi_t(s') &= \operatorname*{argmin}_{1 \leq s \leq S} \{\delta_{t-1}(s) + A_{s,s'}\} \end{aligned}$$

3. Termination

$$\begin{aligned} g(\mathcal{X}, \Lambda) &= \min_{1 \leq s \leq S} \{\delta_T(s)\} \\ \theta_T &= \operatorname*{argmin}_{1 \leq s \leq S} \{\delta_T(s)\} \end{aligned}$$

4. Path Backtracking (information for a path history)

$$\theta_t = \psi_{t+1}(\theta_{t+1}), \quad t = T-1, T-2, \ldots, 1.$$

Once the optimal path is known, each observation vector is assigned to the state on the optimal path that produces it by examining the backtracking information. The re-estimation of the parameters is performed by taking an average value in the following way:

$$\begin{aligned} c_{s,i} &= \frac{Nr_{s,i}}{Nr_s}, \\ \vec{\mu}_{s,i} &= \frac{1}{Nr_{s,i}} \sum_{t=1}^{Nr_{s,i}} \vec{x}_t \mid_{\vec{x}_t \sim s,i}, \end{aligned} \tag{3.18}$$

where Nr_s is the number of observation vectors being assigned at state s,
$Nr_{s,i}$ is the number of observation vectors being assigned to mixture i
of state s ($\vec{x}_t \sim s, t$).

The Forced Viterbi algorithm as a simplified version of the Expectation-Maximization (EM) algorithm has been proven to increase the objective function with every iteration until it converges [104].

3.3 N-gram Models

N-grams are probabilistic models that formalize the idea of word prediction (in the case of the language identification problem — phoneme prediction). An N-gram model predicts the next phoneme from previous $N-1$ phonemes. Such statistical models of phoneme sequences are called language models (LMs). In the case of an LID, computing the probability of the next phoneme will turn out to be closely related to computing the probability of a sequence of words.

More formally, the language model $P(C \mid L_i)$ is used to incorporate the restrictions by which the phonetic elements $\{c_1, \ldots, c_K\}$ can be concatenated to form the whole sequence C. Using the definition of conditional probability, the following decomposition is obtained:

$$P(c_1, \ldots, c_K \mid L_i) = \prod_{k=1}^{K} P(c_k \mid c_1, \ldots, c_{k-1}, L_i). \tag{3.19}$$

For continuous speech, these conditional probabilities are typically used in the following way [61]:

1. The dependence of the conditional probability of observing an element c_k at a position k is modeled using restriction to its immediate $(N-1)$ predecessor elements.
2. The resulting model is referred to as N-gram model (N is the dimensionality of the model).

In this thesis the so-called bigram ($N=2$) and trigram ($N=3$) models are used. In the case of the bigram model, the probability for element c_k depends on its predecessor c_{k-1}:

$$P(c_k \mid c_1, \ldots, c_{k-1}) = P(c_k \mid c_{k-1}). \tag{3.20}$$

According to this assumption, Equation 3.19 can be rewritten to represent a bigram lan-

guage model:

$$P(c_1, \ldots, c_K \mid L_i) = P(c_1 \mid L_i) \prod_{k=1}^{K} P(c_k \mid c_{k-1}, L_i). \tag{3.21}$$

For a trigram model, the probability for element c_k depends on two preceding elements c_{k-2} and c_{k-1}:

$$P(c_k \mid c_1, \ldots, c_{k-1}) = P(c_k \mid c_{k-2}, c_{k-1}). \tag{3.22}$$

Then the corresponding probability for the whole sequence is defined as:

$$P(c_1, \ldots, c_K \mid L_i) = P(c_1 \mid L_i) \prod_{k=1}^{K} P(c_k \mid c_{k-2}, c_{k-1}, L_i). \tag{3.23}$$

The N-gram language model for language L_i is obtained by computing the statistics of a large amount of phoneme sequences. The number of occurrences of every N-gram (sequence of N phonemes) is computed. The result is a set of N-gram histograms, one per language, under the assumption that they are different for every language. Then the probability for every N-gram is computed:

- for bigram

$$P(c_k \mid c_{k-1}) = \frac{Nr(c_{k-1}, c_k)}{Nr(c_{k-1})}, \tag{3.24}$$

- for trigram

$$P(c_k \mid c_{k-2}, c_{k-1}) = \frac{Nr(c_{k-2}, c_{k-1}, c_k)}{Nr(c_{k-2}, c_{k-1})}, \tag{3.25}$$

where $Nr(c_{k-1}, c_k)$ is the number of observed bigrams $c_{k-1}c_k$;
$Nr(c_{k-1})$ is the number of observed elements c_{k-1},
$Nr(c_{k-2}, c_{k-1}, c_k)$ is the number of observed trigrams $c_{k-2}c_{k-1}c_k$;
$Nr(c_{k-2}, c_{k-1})$ is the number of observed bigrams $c_{k-2}c_{k-1}$.

The probabilities for the language models are estimated from a speech corpus during a training phase. However, due to experimental conditions, there is always a problem of availability of training data. Most of the possible events, in our case phoneme pairs and triples, are never seen in training [115]. As a result, the probability estimated for each unseen event is zero and the phoneme sequence that contain these unseen events cannot possibly be hypothesized during the identification process. To overcome these shortcomings, some sort of "smoothing" has to be applied to make sure that each probability estimate is larger than zero [97]. The easiest ways of "smoothing" are initializing each histogram with

an arbitrarily chosen minimum value and modeling of not observed bigrams (or trigrams) with the help of unigram models [98].

The influence of using the language model on the performance of the LID system is very much dependent upon the amount of available training material, upon the number of phonetic classes that represent the elements of C, and also upon how accurately C represents the underlying string of phonetic elements.

3.4 Artificial Neural Networks

An artificial neural network (ANN) is an information processing system and configured for a specific application (such as pattern recognition or data classification) through an automated learning process. An appropriately trained neural network can be thought of as an "expert" in the category of information it has been given to analyze and can be very useful for classification or identification type tasks on unfamiliar data.

An artificial neural network (or connectionist model) [14, 81, 118] is an interconnected group of artificial neurons (simple, non-linear, computational elements) that uses a mathematical or computational model for information processing. The artificial neuron (also called "node") receives one or more inputs, sums these, and produces an output after passing the sum through a (usually) non-linear function known as an activation or transfer function. The general structure of the artificial neuron is presented in Figure 3.3. It consists of the N inputs, labeled $x_1, x_2, \ldots, x_N$, which are summed with weights $w_1, w_2, \ldots, w_N$, thresholded, and then compressed to give the output y, defined as:

$$y = f\left(\sum_{i=1}^{N} w_i x_i - \phi\right), \tag{3.26}$$

where ϕ is an internal threshold or offset and
f is an activation function.

To implement an ANN, one should define the network topology. The so-called feedforward networks have an appropriate structure for classification tasks. A feedforward neural network, which is one of the most common neural network types, is composed of a set of nodes and connections. The nodes are arranged in layers. The connections are typically formed by connecting each of the nodes in a given layer to all neurons in the next layer. In this way, every node in a given layer is connected to every node in the next layer.

As the most popular structure for a neural network classifier, a so-called Multi-layer perceptron (MLP) is used. It has three layers - an input layer, a hidden layer, and an output

layer. The input layer does not perform any processing - it is simply where the data vector is fed into the network. The input layer then feeds into the hidden layer. The hidden layer, in turn, feeds into the output layer. The actual processing in the network occurs in the nodes of the hidden and output layer. Such an MLP is presented in Figure 3.4.

To completely specify an ANN using SENN, the following decisions are required:

- choice of the activation functions for all layers;
- error function for output layer;
- choice of the learning algorithm.

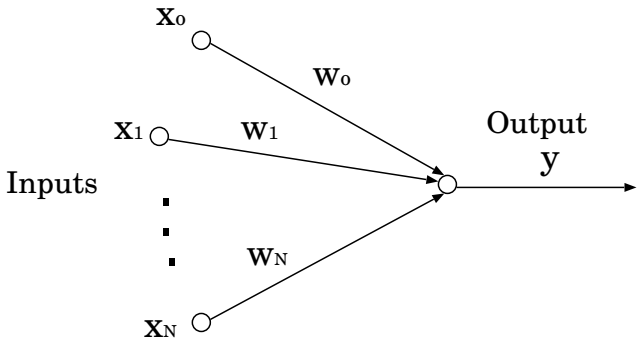


Figure 3.3: *A simple computation element of a neural network*

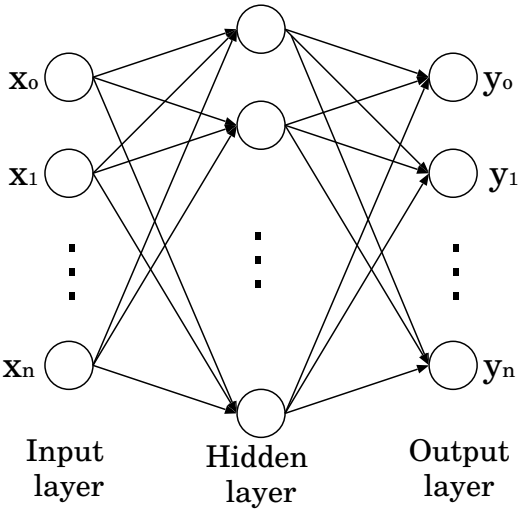


Figure 3.4: *Three-layer perceptron*

The activation function f from Equation 3.26 can be of different types depending on the particular task for which the ANN is designed. In particular, the logistic function is used in this thesis:

$$1/(1+e^{-x}).$$

The error function presents an objective function for the maximum likelihood method that is used to adjust the parameters of an ANN (its weights) so that the error between the desired output and the actual output is reduced. In this thesis a square function is used. The square function is the sum of the squared deviations of the corresponding output and target values:

$$\sum_i (output_i - target_i)^2.$$

The ANN has to be trained using a set of training data. In this mode, the actual output of a neural network is compared to the desired (target) output to compute the value of some predefined error-function. The error value is then fed back through the network. Using this information, the learning algorithm adjusts the weights of each connection in order to reduce the value of the error function by some small amount. Formally, beginning with randomly chosen starting weight vector w^0, the sequence of weight vectors is constructed by determining a search vector d^i (a unit vector) and a step-length η^i (a real number) at each point w^i and computing the next iteration according to

$$w^{i+1} = w^i + \eta^i \cdot d^i. \tag{3.27}$$

The learning procedure tries to minimize the observed objective function for all processing elements. This global error reduction is created over time by continuously modifying the input weights until an acceptable network accuracy is reached. In this case it is possible to say that the network has learned a certain target function.

The learning procedures usually differ in the choice of parameters d^i and η^i and use the gradient of the error function to determine the search direction d^i. The gradient of the error function E with weight vector w^i of length L is defined as vector of partial derivatives: $\nabla E = \left(\frac{\partial E}{\partial w_1}, \ldots, \frac{\partial E}{\partial w_L}\right)$ and points in the direction of steepest ascent of the error function. In this thesis the so called on-line Back Propagation method is used. It takes as search direction the approximation of the negative gradient of the error function. More information about this learning technique can be found in [14, 81].

4

Speech Rhythm

The main focus of this chapter is to address the question of rhythm variation in different languages. First, a number of previous studies is presented in order to find a promising measure of speech rhythm suitable for automated processing. Since the existing theories lead mostly to controversial results, Section 4.2 comes up with a new definition of rhythm. According to the definition, two different algorithms for segmentation of speech utterances into rhythmic units are proposed, resulting in several types of rhythmic features. Finally, Section 4.4 provides a description of modeling techniques that are used to incorporate rhythm features into language identification systems.

4.1 Rhythm Theories

4.1.1 Rhythm Perception

Speech is perceived as a sequence of events, and the expression "rhythm" is used to refer to the way these events are distributed in time. Rhythm can be defined as a systematic organization of prominent and less prominent speech units in time. Prominence of these units is expressed by higher fundamental frequency, higher duration and/or higher intensity.

The perception of speech rhythm has been a subject of interest among linguists for decades. This interest comes from the observation that languages have different types of rhythm. For example, Lloyd James [60] has detected that the English language has a rhythm similar to "Morse code": It can be divided into more or less equal intervals of time, each of which begins with a stressed syllable. In contrast, languages such as French have so-called "machine-gun" rhythm: Individual syllables are perceived to be of nearly equal duration and therefore occurring at regular time intervals. In addition, since French does not have lexical word stress, there are no notable fluctuations of pitch or amplitude to

lend prominence to individual syllables. Thus, according to MacCarthy [77], "continuous French spoken fluently by native speakers conveys the general auditory impression that syllables in each group ... are being uttered at a very regular rate."

As it was shown in Section 2.1.3, human listeners are able to categorically perceive different types of rhythm. The next sections give an overview of the studies in speech rhythm. The aim of these studies is to find measurable regularities or properties of the speech signal that can predict listeners' classification of rhythm types.

4.1.2 Rhythm Class Hypothesis

The observations of the rhythmic organization of the world's languages were summarized by Pike [103] and Abercrombie [1]. They proposed so-called rhythm class hypotheses. Pike [103] suggested that two types of speech rhythm exist: One is due to the recurrence of stresses, while the other one is due to the recurrence of syllables, giving the terminology "stress-timed" and "syllable-timed". The languages with stress-timed rhythm show patterns of equal duration between stressed (prominent) syllables, whereas syllable-timed languages have syllables of equal duration.

Abercrombie [1] generalized this assumption and further claimed that all languages can be classified into one of these rhythmic classes: syllable-timed languages (such as English, and Russian) and stress-timed languages (such as Arabic, French, Telugu, and Yoroba). Additionally, the hypothesis says that rhythmical structure is based on the isochrony of the corresponding rhythmical units, that is, the isochrony of stresses for the former category and the isochrony of syllables for the latter. The theory also claims that the two rhythm categories are mutually exclusive and that every language is characterized by either one or the other of these two types of rhythm.

According to Abercrombie, the rhythm of a language is related to the physiology of speech production and can be defined as a combination of chest pulses and stress pulses. Based on these notions, the rhythm classes are characterized as follows:

stress-timed languages:

- stress pulses are equally spaced — chest pulses are not;
- no isochrony between inter-stress intervals can be measured;

syllable-timed languages:

- chest pulses are equally spaced — stress pulses are not;
- no isochrony between syllable durations can be measured.

Ladefoged [66] later proposed a third rhythmic class, called **mora-timed**, for languages such as Japanese and Tamil, where rhythm is determined by units smaller than syllables, known as morae. Traditionally, morae are sub-units of syllables consisting of one short vowel and any preceding onset consonants. In mora-timed languages, successive morae are said to be nearly equal in duration, which makes these languages more similar to syllable-timed than to stress-timed ones [47].

4.1.3 Problems of the Isochrony Theory

Since the 1960s, phonetic researchers have been trying to find experimental evidence of the isochrony theory proposed by Pike and Abercrombie. The experiments [15, 37, 120] have shown that the classification of the languages on the basis of rhythm class hypothesis into syllable-timed, stress-timed, and mora-timed is not easy: The measurements in the speech signal have failed to confirm the existence of different types of isochronous intervals in spoken language. In stress-timed languages, inter-stress intervals are far from equal, and inter-stress intervals do not behave more regularly in stress-timed than in syllable-timed languages. Additionally, Bolinger [15] showed that the duration of inter-stress intervals depends on the specific types of syllables they contain as well as on the position of the interval within the utterance. Inter-stress intervals in stress-timed languages do not seem to have a constant duration.

Trying to find evidence that the languages classified as stress-timed exhibit any more variability of syllable duration than the languages classified as syllable-timed, Roach [120] has established an experimental test based on two claims made by Abercrombie [1, p. 98]:

> **(i)** "there is considerable variation in syllable length in a language spoken with stress-timed rhythm whereas in a language spoken with a syllable-timed rhythm the syllables tend to be equal in length";
>
> **(ii)** "in syllable-timed languages, stress pulses are unevenly spaced".

For the test, examples from the six languages classified by Abercrombie were recorded: French, Telugu, and Yoruba as syllable-timed representatives and Arabic, English, and Russian as stress-timed. The languages were examined to see if it is possible to assign languages to one of the two categories based on the above claims.

First, the standard deviation of syllable durations for all languages was measured. The results did not support claim (i): Syllable variation is not significantly different for stress-timed and syllable-timed languages. To verify claim (ii), the duration of inter-stress inter-

vals (from the onset of each syllable which appeared to be stressed until the onset of the next one within the same intonation unit) was measured. This duration was expressed as a percentage of the duration of the whole intonation unit to compensate for any possible effects of change of tempo. The results of this, contrary to what would be predicted by the typological distinction, showed greater variability in the duration of the inter-stress intervals for the so-called stress-timed languages (especially for English) than for the so-called syllable-timed languages. There was, furthermore, no evidence that the duration of inter-stress intervals was any less correlated with the number of syllables which they contained for the stress-timed languages than for the syllable-timed languages.

Similar work was done by Dauer [37] on English (stress-timed) and Greek, Italian, and Spanish (syllable-timed). Dauer found that inter-stress intervals were no more regular in stress-timed language than in syllable-timed one and that the mean duration of inter-stress intervals for all languages analyzed is proportional to the number of syllables in the interval.

Isochrony in mora-timed languages was investigated by Port and colleagues [105, 106]. They provided some preliminary support for the mora as a constant time unit. While investigating segmental timing in Japanese, the authors have demonstrated that words with an increasing number of morae increase in duration by nearly constant increments. By stretching or compressing the duration of neighboring segments and adjacent morae it was found that the duration of a word stays very close to a target duration that depends on the number of morae in it. These, however, contradict the results of other researchers [10, 53] that questioned the acoustic basis for mora-timing. Beckman [10] examined a preliminary corpus of utterances used in the development of synthesis rules for Japanese but did not reveal the tendency toward moraic isochrony. Hoequist [53] performed a comparative study of duration characteristics in Spanish (syllable-timed) and Japanese (mora-timed) that confirm the absence of a strict isochronous rhythm but yield evidence fitting a less strict hypothesis of rhythm categories.

Despite the contradictory conclusions of these experiments, many linguists agree that the principle of isochrony can underlie the rhythm of a language even if it is not demonstrated experimentally from a phonetic point of view. Couper-Kuhlen [32] and Lehiste [69] have tried to regard isochrony primarily as a perceptual phenomenon. The perception of isochrony on either syllabic or stress level is dependent upon the human cognitive tendency to impose rhythm, upon things occurring with some resemblance of regularity, such as a clock ticking, the sound of footsteps, or the motion of wind-shield wipers. It was pointed out that because differences in duration between stress or syllables are well below the threshold of perception, humans still perceive the unit to be recurring isochronously even though the principle cannot be proven quantitatively.

Other attempts were made by Beckman [11] and Laver [67] which retreated from “sub-

jective" to "objective" isochrony. These researchers described the physical regularity of isochrony as a tendency: True isochrony is assumed to be an underlying constraint, and the surface realizations of isochronous units are perturbed by the phonetic, phonological and grammatical characteristics of the language. A scalar model of speech rhythm, proposed by Laver [67] with two (hypothetical) languages at the two extremes of the rhythm scale, should be able to account better for the observable facts than the traditional dichotomous distinction. Such a model should make it possible to find the place of a given language on the rhythm scale with reference to, for example, English or French.

4.1.4 Other Views of Speech Rhythm

A new proposal for rhythm classification was suggested by Dasher and Bolinger [36]. According to them, the impression of different types of rhythm is the result of the coexistence of specific phonological phenomena such as variety of syllable types, the presence or absence of phonological vowel length distinction, and vowel reduction. Along this line of research, Dauer [37], analyzing stress-timed and syllable-timed languages, has emphasized a number of different distinctive properties among them:

- Syllable structure: Stress-timed languages have a greater variety of syllable types than syllable-timed languages. As a result, they tend to have more complex syllables. In addition, this feature is correlated with the fact that in stress-timed languages, stress most often falls on the heaviest syllables, while in syllable-timed languages stress and syllable weight tend to be independent.
- Vowel reduction: In stress-timed languages, unstressed syllables usually have a reduced vocalic system (sometimes reduced to just one vowel, schwa), and unstressed vowels are consistently shorter, or even absent.

In Dauer's view, the different properties mentioned above could be independent and cumulative: All languages are more or less stress-based. The more typical stress-timed language properties a language presents, the more it is stress-timed, and the less it is syllable-timed. Dauer thus suggested a continuous unidimensional model of rhythm, with typical stress-timed and syllable-timed languages at either end of the continuum.

In a more recent publication [38], Dauer defined rhythm as "the grouping of elements into larger units", where "elements that are grouped are syllables". In this case there should be an instrument that serves to mark off one group of syllables from another. Dauer hypothesized that linguistic accent can be a basis of rhythmic grouping. However, since "all languages have rhythmic grouping, but that not all necessarily have accent", rhythm is then defined as a total effect which involves a number of components, namely:

- syllable and vowel duration;
- syllable structure and quantity as major factors responsible for length;
- intonation and tone as means of achieving pitch distinctions;
- vowel and consonant quality;
- function of linguistic accent.

In [38], a rating system was developed, according to which each component is broken down into "features", and each feature is assigned a plus or a minus value (or sometimes zero). In this way, a relative rhythm "score" for a given language is obtained: The more pluses a language has, when assessed in terms of the above components, the more likely it is that the language has "strong stress" and that it is stress-timed.

Another view of speech rhythm that differs from Dauer's continuous system was offered by Nespor [96]. Nespor questions the dichotomy between syllable-times and stress-timed languages by presenting languages that share phonological properties of both types. E. g., Polish has been classified as stress-timed but does not exhibit vowel reduction at normal speech rates, and at the same time has a great variety of syllable types and high syllabic complexity, like stress-timed languages. Catalan that has the same syllabic structure as Spanish, and thus should be syllable-timed, also has vowel reduction like stress-timed languages. For such languages, one would like to know whether they can be discriminated from syllable-timed, stress-timed, both, or neither. The rhythm class hypothesis, in its current formulation, would hold only if they clustered along with one or the other language group.

The results of the discussion about rhythm classes at the beginning of the 1990s can be summarized by following statements:

- "numerous experiments have shown that a language cannot be assigned to one or the other category on the basis of instrumental measurements of inter-stress intervals or syllable durations" [38];
- rhythm is a mere perceptual phenomenon.

4.1.5 Recent Rhythm Measurements

Despite the failure of experimental attempts to demonstrate quantifiable isochrony in stress-timed and syllable-timed languages, some experiments on the human language discrimination (a review can be found in Section 2.1.3) brought the rhythm class hypothesis back into discussion.

Recent studies [47, 111] have demonstrated that there are quantitative rhythmic differences between stress- and syllable-timed languages. These studies have focused on the role of vowel duration in the two types of languages. They hypothesized that since vowels are primarily responsible for the length of a syllable, an increased variability in vowel length would result in greater variability in syllable length, whereas a language with little variation in vowel length would have little overall variation in syllable length.

The first study by Ramus et al. [111] presents instrumental measurements based on consonant/vowel segmentation for eight languages (Catalan, Dutch, English, French, Italian, Japanese, Polish, and Spanish). Using recordings of five sentences spoken by four speakers for each language, sentences were segmented into "vocalic intervals" and "consonantal intervals", defined as portions of the speech signal containing sequences of only vowels or only consonants. Then the following parameters, each taking one value per sentence, were calculated:

- the sum of the durations of the vocalic intervals expressed as a percentage of the total duration of the sentence — $\%V$;
- the standard deviation of the consonantal intervals within each sentence — ΔC;
- the standard deviation of the vocalic intervals — ΔV.

Analyzing these parameters and their combinations for different languages, Ramus and colleagues showed that the combination of $\%V$ and ΔC provides the best acoustic correlate of rhythm classes. Therefore the theory can be used to separate the three rhythmic classes. Figure 4.1 represents the findings of Ramus et al. in symbolical way.

According to the results presented in [111], ΔC and $\%V$ not only support the notion of rhythm classes, but are also directly related to the syllabic structure of a language. Having more syllable types in language means more variability in the number of consonants, more variability in their overall duration in the syllable, and thus a higher ΔC. This in turn implies a greater consonant/vowel ratio in average, i. e., a lower $\%V$. This assumption is supported by the placement of different languages on the ($\%V$, ΔC) scales: English, Dutch, and Polish are at one end of the scales and have more than 15 syllable types, and Japanese is at the other end with 4 syllable types.

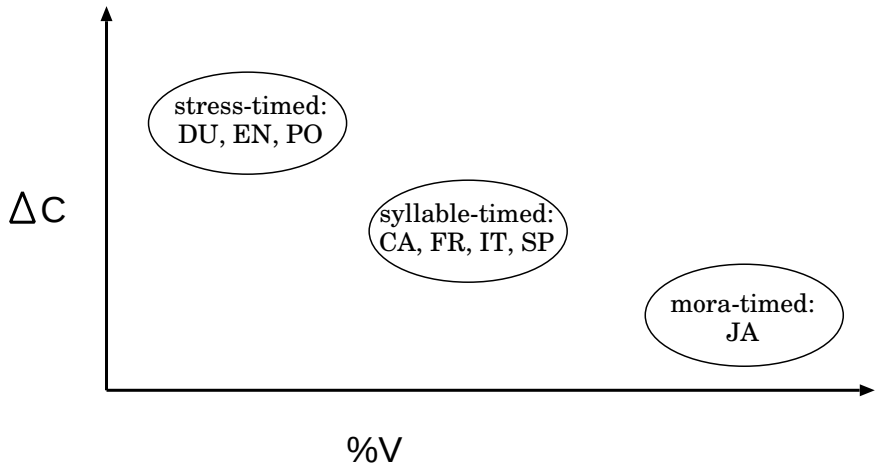


Figure 4.1: *Duration of vocalic intervals as percentage of total duration (%V) and standard deviation of consonant intervals (ΔC) for Catalan (CA), Dutch (DU), English (EN), French (FR), Italian (IT), Japanese (JA), Polish (PO), and Spanish (SP) reproduced symbolically from Ramus et al. [111]*

Investigations of ΔV have shown that its value reflects the sum of several phonological factors that influence the variability of vocalic intervals:

- vowel reduction (as in Catalan, Dutch, and English);
- contrastive vowel length (Dutch and Japanese);
- vowel lengthening in specific contexts (Italian);
- existence of certain vowels that are significantly longer as other (English and French).

Thus, ΔV provides some information about the phonology of languages. At the same time, the ΔV scale shows no relation to the usual rhythm classes but suggests that Polish (but not Catalan) in some aspects is very different from the other stress-timed languages. This is consistent with finding of Nespor [96]. New experiments are needed to clarify whether ΔV plays a role in rhythm perception.

The perceptual experiments that directly test the notion of rhythm classes, the simulations predictions, and the question of intermediate languages were performed by Ramus et al. later and presented in [110]. Human language discrimination between Catalan and Polish and reference languages like English and Spanish was performed using a speech re-synthesis technique [112] to ensure that only rhythmical cues are available to the subjects. The results were compatible with the rhythm class hypothesis and Catalan was identified as

syllable-timed. Polish, however, seems to be different from any other language studied and thus constitutes a new rhythm class.

In a related study, Grabe and Low [47] tried to find the relationship between speech timing and rhythmic classifications of languages that fell under the traditional categories of stress-timed and syllable-timed. Departing from the search for isochrony, the authors measured the durations of vowels and the duration of intervals between vowels (excluding pauses) in a passage of speech. To estimate the amount of duration variability in a language, the authors proposed a so-called "Pairwise Variability Index" (PVI) which does take the sequential variability into consideration by averaging the duration difference between consecutive vowel or consonant intervals. The raw Pairwise Variability Index (rPVI) was used for consonantal intervals and is defined as:

$$rPVI = \left[\sum_{k=1}^{m-1} |d_k - d_{k+1}|/(m-1)\right], \tag{4.1}$$

where m is the number of intervals and d_k is the duration of the k-th interval.

For vocalic intervals the rPVI was normalized to correct the changes in speaking rate. The normalized PVI (nPVI) is calculated in the following way:

$$nPVI = 100 \times \left[\sum_{k=1}^{m-1} \left|\frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2}\right| /(m-1)\right]. \tag{4.2}$$

Calculated in this way, the pairwise variability index expresses the level of variability in successive measurements. More intuitively, a sequence where neighboring measurements tend to have a larger contrast in duration would have a larger PVI. A sequence of measurements with low contrast in duration would have a lower PVI. Thus, the stress-timed languages would exhibit high vocalic nPVI and high intervocalic rPVI values, and syllable-timed languages would have low PVI values.

Grabe and Low used PVI to provide evidence for rhythmic classification of eighteen languages (one speaker per language) traditionally classified as stress-timed, syllable-timed, mora-timed, mixed, or unclassified. It was determined that the duration variability of vowels is in fact quantitatively greater in a stress-timed language than in a syllable-timed one. This can be explained by the compression and expansion of syllables in stress-timed languages that lead to the higher average difference in duration (in comparison with syllable-timed languages).

The results obtained (presented in Figure 4.2) agree with the classification of Dutch, English, and German as stress-timed and French and Spanish as syllable-timed. Values for

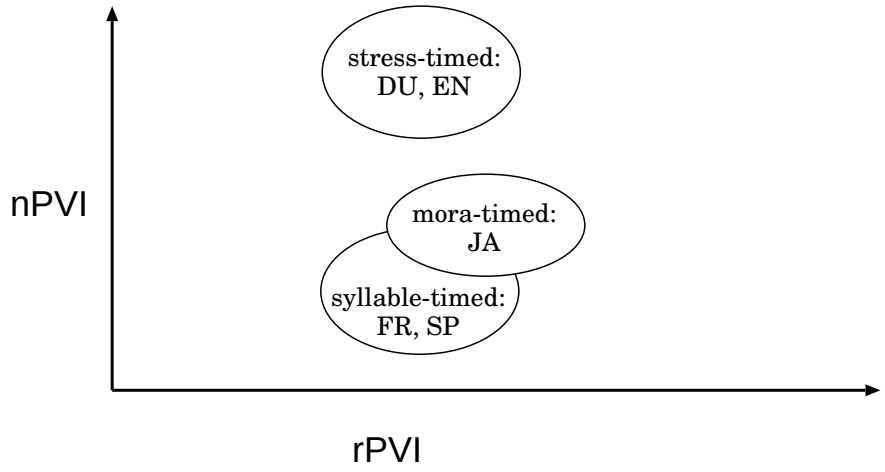


Figure 4.2: *PVI profiles from prototypical languages Dutch (DU), English (EN), French (FR), Japanese (JA), and Spanish (SP), reproduced symbolically from Grabe and Low [47]*

Japanese, a mora-timed language, were similar to those from syllable-timed languages. Previously unclassified languages (e. g., Greek, Mandarin, Romanian, Welsh) did not fit into any of the three rhythm classes — their values overlapped with the margins of both the stress- and syllable-timed group. In addition, the results illustrated a continuum within the categories of stress-timed and syllable-timed languages. For example, French and Spanish are on the low end of the pairwise variability index range for syllable-timed languages, while Singapore English and Tamil are on the high end; thus French and Spanish may be said to be more strongly syllable-timed than Singapore English and Tamil. However, Grabe and Low's results should be considered as preliminary. The conclusions made on data from only one speaker per language have to be verified using more data from different speakers.

Grabe and Low also tried to replicate the findings of Ramus et al. [112] on their data and came to significantly different results that do not support the cluster hypothesis introduced in [112]. In a later work, Ramus [109] suspected that, among other factors (e. g., speaker typical influence), the not well controlled speech rate could be the reason for contradictory results.

A number of proposals for rhythm metrics that tried either to verify the proposals of Ramus et al. and Grabe and Low or to improve their measures have been published. The following is a quick summary of some of these approaches:

- Extension of the PVI measure, that takes the consonant and vowel interval together, thus capturing the varying complexity of consonantal and vowel groupings in se-

quence, was suggested by Barry et al. [6].

- An attempt to use the coefficient of variation of vocalic and consonantal intervals rather than the standard deviation in Ramus et al. measures was made by Dellwo and Wagner [141].

- Applying the PVI on the level of the foot as well as on the level of the syllable has been proposed by Asu and Nolan [4].

- Control/Compensation Index which gives a relative measure of the PVI to the number of segments composing each consonantal or vocalic interval was proposed by Bertinetto and Bertini [13].

A series of experiments were designed in order to investigate the influence of speech rate on the rhythm of different languages. They have shown that different rhythm measures are to a great degree dependent on the overall speech rate of utterances:

- Barry et al. [6] have shown that ΔC and ΔV measures decrease with an increase in speech rate and nPVI does not normalize for speech rate;

- Dellwo and Wagner [39] came to a similar conclusion regarding the behaviour of ΔC and found that $\%V$ is constant over all speech rates.

The usefulness of these various measures seems to be dependent on the task for which they are employed. They give only a crude intuition into the way in which rhythmic structures are realized in different languages. In [51], Hirst has shown that some of the rhythm measures are more sensitive to the rhythm of the text than to the rhythm of the utterance itself. The main conclusion of his work was the need for more detailed studies using large corpora in order to develop more sophisticated models.

Such an attempt was made recently by Loukina et al. [75] who applied already published rhythm measures to a large corpus of data to test whether they can reliably separate languages. To avoid inconsistencies introduced by manual segmentation of data, a simple automatic segmentation into consonant-like and vowel-like regions was applied. The authors tested different combinations of 15 rhythm measures building classifiers that are based on single measures, on two or three measures, and multidimensional classifiers (up to 15 rhythm measures). The following general conclusions were made:

- some rhythm measures perform better than others and their efficiency depends on the languages they have to separate;

- within-language variation of the rhythm measures is large and comparable to the observed between-language variations;
- different published measures capture different aspects of rhythm;
- rhythm appears to be described sufficiently by two dimensions (significant improvement from using more than two different rhythm measures has not been observed) and results of different pairs of rhythm measures were comparable;
- investigation of the speech rate has shown that it cannot separate languages on its own, but it is definitely one of the variables in the 'rhythm equation' and should be included in any model of rhythm.

4.2 Proposal of Rhythm Definition

This thesis deals with rhythm and the modeling of rhythm from the point of its application in a language identification task. As it appears from the previous section, a satisfying definition of speech rhythm, as well as a perceptual evidence for rhythm class hypothesis, has not yet been found.

Nevertheless, it is commonly agreed that rhythm is related to the duration of some speech units. Most linguistic studies of speech rhythm support an assumption that a rhythmic unit corresponds to the syllable combined with an optional stress pattern. The idea of using a syllable as an appropriate rhythm unit has also been supported by Nooteboom in a reiterative study [99] that has shown how rhythm of speech utterances can be imitated by a sequence of identical nonsense syllables. At the same time, Dauer [38] has mentioned: "Neither "syllable" nor "stress" have general phonetic definitions, which from the start makes a purely phonetic definition of language rhythm impossible. All instrumental studies as well as all phonological studies have had to decide in advance where the stresses (if any) fall and what a syllable is in the language under investigation in order to proceed".

Another question, which still has to be clearly identified, is related to the actual role of the syllable: whether the important feature is the syllable itself (as a linguistic unit) or its boundaries (as milestones for the segmentation process). This thesis follows the second assumption.

In order to use speech rhythm for the language identification task, the following is proposed:

- to investigate syllables as the most intuitive rhythmical units,
- to model speech rhythm by the durations of two successive syllables.

A theoretical confirmation of this idea can be found in a recently proposed method for visualization of rhythmical structures in speech (Wagner, [140]). Wagner has shown that rhythmic events can be described by two-dimensional scatter plots spanned by the duration of successive syllables obtained by manual segmentation of speech. This method is able to show the clear distinctions between stress-timed and syllable-timed languages while pointing out inter- and intra-group timing differences at different prosodic levels.

Besides the problems related to the efficient definition of rhythm, there are unanswered questions concerning an appropriate treatment of speech rhythm for language identification. These can be articulated as follows:

- how to automatically segment speech into suited rhythmic units, and
- how to develop a language independent approach for rhythm modeling.

The segmentation of speech into syllables suited for a language independent approach means an automatic extraction of syllables (in particular, boundary detection). The main problem here is that syllable boundaries cannot be detected, because they are often found between consonant clusters, where automatic segmentation is difficult. Furthermore, segmenting a speech signal into syllables is a language-specific task that requires a set of syllables to be provided for each new language investigated.

To overcome this problem, a syllable is replaced by the so-called syllable-like unit that is presented in the next section.

4.3 Extraction of Rhythm Features

In order to create a language independent algorithm for segmentation of speech into rhythmic units, the notion of syllable (more precisely its duration) is substituted by an abstract syllable-like event that can be defined in two different ways:

1. A syllable is replaced by a so-called "pseudo-syllable" and the duration of a pseudo-syllable is taken to represent a rhythmical unit. The notion of a pseudo-syllable was first introduced by Farinas et al. [41] and is explained in Section 4.3.1.

2. A syllable is specified as a time interval between two successive vowels found using an algorithm for automatic vowel detection as described in Section 4.3.2.

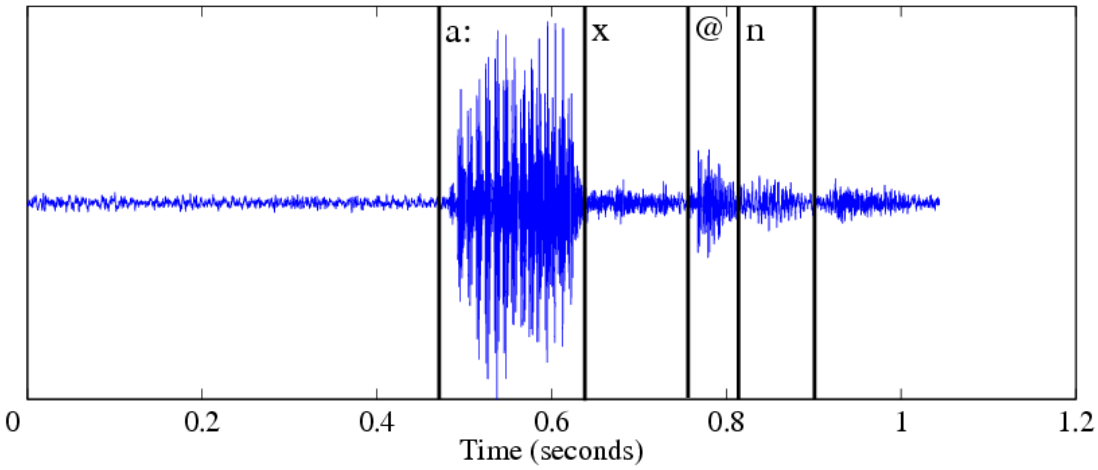


Figure 4.3: *Consonant/Vowel segmentation of the German word "Aachen"*

4.3.1 Pseudo-syllable Segmentation

The notion of pseudo-syllable proposed by Farinas et al. [41] was derived from the most frequent syllable structure over the world's languages, namely the CV structure, where C denotes a consonant and V denotes a vowel. The pseudo-syllable is defined as a pattern C^nV with n being an integer that can be zero. The segmentation into pseudo-syllables using the German city name "Aachen" (presented graphically in Figure 4.3), as an example, looks like the following:

$$(a\colon .x.@.n) \Longrightarrow V.CV.C \Longrightarrow V.CV$$

According to [41], the rhythm unit that corresponds to a pseudo-syllable is defined as a triple consisting of duration of consonantal part, the duration of vowel part, and the number of consonantal segments.

Unlike [41] where the C^nV structure of pseudo-syllable was used as a rhythm feature, in this thesis the potential of the duration of the pseudo-syllable itself is explored as a feature for LID and is defined as the total duration of its consonant and vowel parts.

Applying pseudo-syllables, the automatic language independent algorithm for rhythm feature extraction can be easily derived using a consonant-vowel segmentation mechanism. This approach needs no language-specific knowledge concerning syllables. Segment boundaries are always at vowels, which can be detected more reliably.

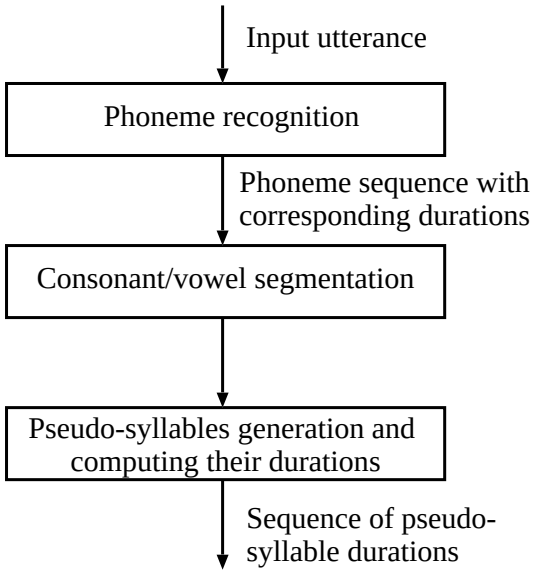


Figure 4.4: *Extraction of pseudo-syllable durations*

For segmentation, a HMM-based phoneme recognizer which outputs not only the sequence of recognized phonemes but also the timing of the state sequence given during the Viterbi decoding process is used. More information about HMM based phoneme recognizers can be found in Section 3.2. The temporal ending points of the vowels are determined by the timing of the last state of each vowel. In order to have a language independent approach, a multilingual HMM was created for a united set of phonemes from all languages in the task.

The rhythm feature extraction procedure is performed in several steps (presented in Figure 4.4):

- The input signal is transformed into a sequence of phonemes with corresponding phoneme durations using a language independent phoneme recognizer realized by a multilingual HMM.
- The phoneme sequence is converted into a consonant-vowel sequence.
- According to the notion of the pseudo-syllable, the consonant-vowel sequence is parsed into patterns matching the structure C^nV. For the resulting sequence of pseudo-syllables the corresponding durations d_s are computed.

4.3.2 Vowel Detection

The basic idea of this method is to measure the time elapsed between two successive vowels and interpret it as the duration of the syllable-like unit.

Over the last several decades, plenty of effort has still been directed to the problem of automatic detection of vowels, showing that it is still far away from being solved. Since the vowel detection task itself is outside the scope of this thesis, the corresponding details are omitted here. A survey of the existing techniques can be found in [57].

The vowel detection approach is also used as automatic rhythm extraction for a rhythm LID system as proposed in this thesis and will be compared later with the algorithm based on pseudo-syllables from the previous section.

Positions of vowels in an utterance for this thesis are found using a signal based method where the short time energy is extracted from speech and energy peaks are used as detection points. This method, which is a combination of the approaches published by Pfitzinger [102] and Narayanan [91], was originally created by Hofer [54] for duration modeling in Mandarin connected digit recognition [3]. As presented in Figure 4.5, the method utilizes the short time energy in the following way:

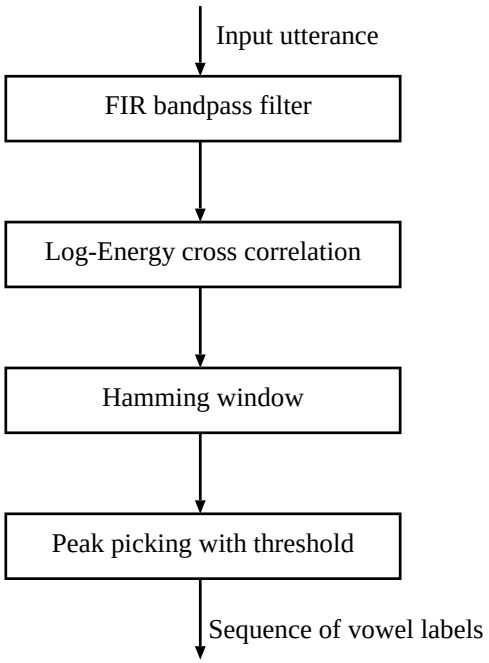


Figure 4.5: *Vowel detection algorithm*

- The input signal is filtered using a Finite Impulse Response (FIR) bandpass filter in the frequency range from 50 Hz to 2000 Hz.

- Cross correlation in the time domain is based on the idea presented in [91] and is performed every 10 ms on frames with a duration of 50 ms :
  For $x(t), x(t+1), \ldots, x(t+K-1)$ being a frame of the input speech with length $K = 50\,\text{ms}$ the new trajectory $y(t)$ is computed as:

$$y(t) = \log\left(\frac{1}{K(K-1)}\sum_{j=0}^{K-2}\left(\sum_{p=j+1}^{K-1} x(t+j)\cdot x(t+p)\right)^2\right).$$

- In order to suppress possible ripples that do not contain reliable information, a Hamming window with a length of 50 ms is convolved with the function to achieve the necessary smoothing.

- Local maxima that correspond to vowels are found using a threshold based algorithm as in [54]. The algorithm picks only those maxima where the difference in value to the surrounding local minima surpasses a certain threshold: If $y(t)$ is the smoothed energy function and a and b are local maxima, then the following conditions must hold for some threshold T:

$$y(a) - \min_{a<t<b}(y(t)) > T \times [\max_{\forall t}(y(t)) - \min_{\forall t}(y(t))]$$

  and

$$y(b) - \min_{a<t<b}(y(t)) > T \times [\max_{\forall t}(y(t)) - \min_{\forall t}(y(t))]$$

  The threshold is optimized on a development data and is set to 0.01, i. e., a peak should have at least one percent of the possible maxima.

An example of the estimated energy function with found local maxima marked by circles is presented in Figure 4.6. The intervals between successive local maxima are taken as durations of corresponding syllable-like units.

4.3.3 Speech Rate

As was already mentioned in Section 4.1, some studies of speech rhythm [6, 39] have detected an interaction between speech rate and language rhythm. Moreover, Dellwo [39] has found that some languages (for example, English and German) tend to vary in rhythm as a function of speech rate.

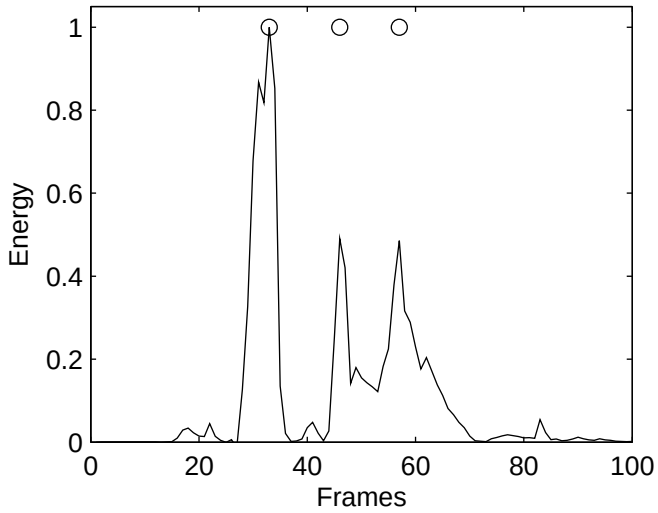


Figure 4.6: *Estimated energy function and the detected positions of vowels*

To verify an influence of speech rate on the rhythm properties of a language, speech rate is defined as a number of syllables (in this particular case — syllable-like units) per second. Speech rate is used in two different ways:

- for normalization of durations of syllable-like units (in order to exclude the influence of the pronouncing speed on rhythmic properties);
- as a self-sufficient rhythm feature.

Speech rate is computed according to the definition proposed by Pfizinger [101]. An utterance is presented as a sequence of syllable-like events $S = s_1 \dots s_j \dots s_M$, where s_j is a boundary of a syllable-like event j. The distances between subsequent segmentation marks falling in a window of 410 ms are accumulated and divided by their number, giving the measure for speech rate between two points L and R:

$$rate_{LR} = \frac{\frac{S_{l+1}-w_L}{S_{l+1}-S_l} + \frac{w_R-S_r}{S_{r+1}-S_r} + r - l + 1}{S_{l+1} - w_L + w_R - S_r + \sum_{i=l+1}^{r-1}(S_{i+1} - S_i)}, \tag{4.3}$$

where w_L and w_R are the left and the right window boundaries, S_l and S_r are the left and the right syllable-like events covered only partly by the window and that is why have to be partially weighted. To obtain smoothing speech rate, the Hanning window is applied as was done in [101]:

- Normalized to one in its total surface, the Hanning window is defined as

$$\int_a^b (1 - \cos x)dx = [x - \sin x]_a^b = b - a + \sin a - \sin b; \tag{4.4}$$

- then a weighting function for a syllable-like event that starts from a and ends at b is defined as:

$$H(a, b) = 1 + \frac{\sin a - \sin b}{b - a}, 0 \leq a, b \leq 2\pi; \tag{4.5}$$

- all boundary values should be mapped to the range from 0 to 2π resulting in the weighting function:

$$W(x, y) = H(2\pi \frac{x - w_L}{w_R - w_L}, 2\pi \frac{y - w_L}{w_R - w_L}) \tag{4.6}$$

Introducing the weighting function W defined in Equation 4.6 into the definition given in Equation 4.3, the smoothed speech rate is found as:

$$rate_{LR} = \frac{\frac{S_{l+1} - w_L}{S_{l+1} - S_l} W(w_L, S_{l+1}) + \frac{w_R - S_r}{S_{r+1} - S_r} W(S_R, w_R) + \sum_{i=l+1}^{r-1} W(S_i, S_{i+1})}{S_{l+1} - w_L + w_R - S_r + \sum_{i=l+1}^{r-1} (S_{i+1} - S_i)}. \tag{4.7}$$

4.4 Modeling Rhythm

For the language identification task, a speech utterance is modeled as a sequence of syllable-like units $S = s_1 s_2 \dots s_M$. The syllable-like units are found by one of the previously presented methods: either using the notion of pseudo-syllable or using vowel detection method.

The following possibilities are tested:

1. **Language rhythm is modeled via durations of successive syllable-like events.** Assigning to each syllable-like event s_j its duration

$$d_j = s_{j+1} - s_j, \tag{4.8}$$

the utterance is represented as a sequence of durations $D = d_1, d_2, \dots, d_M$.

According to Equation 2.6, the most likely language with respect to prosodic information can be found as follows:

$$L^* = \underset{i}{\operatorname{argmax}}\, P(D \mid L_i). \tag{4.9}$$

For each language L_i $(i = 1, 2, \ldots, n)$, the likelihood $P(D \mid L_i)$, which measures how likely it is that underlying model of language L_i generates the data sample defined by D, is calculated using Bigram approximation:

$$\begin{aligned} P(D \mid L_i) &= P(d_1, d_2, \ldots, d_M \mid L_i) \\ &= P(d_1 \mid L_i) \prod_{j=2}^{M} P(d_j \mid d_{j-1}, \ldots, d_M, L_i) \\ &\approx P(d_1 \mid L_i) \prod_{j=2}^{M} P(d_j \mid d_{j-1}, L_i). \end{aligned} \tag{4.10}$$

Using the definition of conditional probability:

$$P(d_j \mid d_{j-1}, L_i) = \frac{P(d_j, d_{j-1} \mid L_i)}{P(d_{j-1} \mid L_i)}, \tag{4.11}$$

Equation 4.10 can be further transformed in the following way:

$$\begin{aligned} P(D \mid L_i) &\approx P(d_1 \mid L_i) \prod_{j=2}^{M} \frac{P(d_j, d_{j-1} \mid L_i)}{P(d_{j-1} \mid L_i)} \\ &\approx P(d_1 \mid L_i) \frac{\prod_{j=2}^{M} P(d_j, d_{j-1} \mid L_i)}{P(d_1 \mid L_i) \prod_{j=3}^{M} P(d_{j-1})} \\ &\approx \frac{\prod_{j=2}^{M} P(d_j, d_{j-1} \mid L_i)}{\prod_{j=3}^{M} P(d_{j-1} \mid L_i)} \end{aligned} \tag{4.12}$$

During preliminary experiments it was found that the durations of syllable-like units themselves do not contain language discrimination information. In this case the denominator $\prod_{j=3}^{M} P(d_{j-1} \mid L_i)$ can be omitted, and the rhythm likelihood can be defined using the joint probability of a duration pair for neighboring syllable-like units, i. e.

$$P(D \mid L_i) = \prod_{j=1}^{M-1} P(d_j, d_{j+1} \mid L_i). \tag{4.13}$$

Rhythm systems based on conditional and joint probabilities, as defined in Equations 4.10 and 4.13 respectively, were trained and evaluated. Although conditional probabilities model the whole timing structure of the syllable-like sequence better, it was discovered that the system based on the joint probabilities yields better re-

sults. Therefore is was decided to use joint probabilities and to model rhythm via the durations of two successive syllable-like units:

$$P(d_j, d_{j+1} \mid L_i).$$

The segmentation into syllable-like events is performed with a language independent analysis of the signal, which provides a single estimated sequence D for each utterance. Thus the resulting probability score $s_R(D \mid L_i)$ does not need to be normalized when language-specific scores derived from the same utterance are compared.

To simplify the computation, the neglog-likelihood score $s_R(D \mid L_i)$ is computed. This allows approximating multiplication by summation:

$$\begin{aligned} s_R(D \mid L_i) &= -\log(P(D \mid L_i)) \\ &= -\log(\prod_{j=1}^{M-1} P(d_j, d_{j+1} \mid L_i)) \\ &\approx -\sum_{j=1}^{M-1} \log P(d_j, d_{j+1} \mid L_i)) \end{aligned} \tag{4.14}$$

The language is classified with a pseudo maximum likelihood method:

$$\hat{L} = \underset{L_i}{\operatorname{argmin}}\, s_R(D \mid L_i), \tag{4.15}$$

where $s_R(D \mid L_i)$ is defined by Equation 4.14.

2. **Language rhythm is modeled via normalized durations of successive syllable-like events.**
For every syllable-like unit from the sequence $S = s_1 s_2 \dots s_M$, the corresponding duration d_j is computed according to Equation 4.8. The speech rate $r_{j,j+1}$ is estimated as the mean value of speech rates for two neighboring events

$$r_{j,j+1} = (r_{j+1} - r_j)/2,$$

where r_{j+1} and r_j are computed using Equation 4.7.
Then the durations are normalized as follows:

$$d_j^{norm} = d_j * r_{j,j+1}. \tag{4.16}$$

Language-specific rhythm scores $s_R^{norm}(D^{norm} \mid L_i)$ for the sequence

$D^{norm} = d_1^{norm}, d_2^{norm}, \ldots, d_M^{norm}$ of normalized durations of syllable-like events are computed according to Equation 4.14, and a language is hypothesized as:

$$\hat{L} = \underset{L_i}{\operatorname{argmin}}\, s_R^{norm}(D^{norm} \mid L_i). \tag{4.17}$$

3. **Language rhythm is modeled via speech rate.**
Each syllable-like unit from sequence $S = s_1 s_2 \ldots s_M$ is associated with speech rate $r_{j,j+1}$, computed as described above and giving a sequence $R = r_{1,2}, r_{2,3}, \ldots, r_{M-1,M}$. Then a simple probability model $P(r_{j,j+1} \mid L_i)$ for each language L_i $(i = 1, 2, \ldots, n)$ is built. The language-specific score $s_R^{rate}(R \mid L_i)$ is computed as

$$s_R^{rate}(R \mid L_i) = -\sum_{j=1}^{M-1} \log P(r_{j,j+1} \mid L_i) \tag{4.18}$$

and the recognized language is found according to the rule:

$$\hat{L} = \underset{L_i}{\operatorname{argmin}}\, s_R^{rate}(R \mid L_i). \tag{4.19}$$

5

Proposed LID systems

This chapter describes the design of all LID systems presented in this thesis. First, a general architecture of LID systems is presented based on the structure illustrated in Figure 2.2. Then, the subsequent sections provide descriptions of different LID systems based on the different types of information extracted from the speech signal. Section 5.5 introduces the measures of the recognition performance used to compare LID systems. Finally, Section 5.6 presents the scheme for combination of the results from different systems.

5.1 General Architecture

LID systems presented in this thesis are designed for the language identification of a spoken utterance from a given set of languages. This means that during testing a system is able to identify a language that was contained in the training set. Thus, using the system for the set of languages of interest requires the existence of available training material (depending on the approach transliterated or not) for every language from the set.

Every LID system operates in two phases: training and recognition. During the ***training*** phase, the language-specific models are created for all languages in the task according to the particular approach. During the ***recognition*** phase, the following is performed:

1. The system receives as input the speech waveform which consists of sample data taken at a rate of 8 kHz.

2. The digitized acoustic signal is converted into a compact representation (sequence of feature vectors) that captures the particular characteristics of the speech signal.

3. The sequence of feature vectors proceeds through all language-specific models. This produces probabilities that show how likely the models giving the input utterance

presented by these features are.

4. The system decision is made either according to the maximum likelihood rule or by using an ANN classifier.

Using the **maximum likelihood rule** as defined in Section 2.3.1 means that a language with maximum likelihood produced by language-specific models is taken as the LID result (answer of the LID system). When the models produce scores (neglog-likelihoods), the language with minimum score is taken as the identification result.

The likelihoods are created by independently trained models that result in a loss of their discriminative power. This disadvantage is overcome by using an ANN optimized to approximate the probability distribution over output classes conditioned on the input (i. e., a-posteriori probabilities). This approach was already proven to be well-suited for language identification in previous work [8, 9, 133, 134].

An **ANN classifier** is implemented as three layer perceptron with ten hidden nodes. The number of input and output nodes is the number of considered languages. The ANN is trained on the z-normalized scores (or probabilities) produced by processing the training material through all models and uses sigmoid function for activation and Backpropagation as a learning algorithm. The output node that corresponds to the spoken language is a binary value (one or zero). By iterating the learning procedure, the ANN learns the relations between the scores. During classification, the ANN takes normalized language scores (or probabilities) as input and aims to produce a-posteriori probabilities for every language. The language with maximum probability is hypothesized.

5.2 Spectral LID Systems

5.2.1 Extraction of MFCC

Language-specific information for spectral LID systems is presented with MFCC features (first introduced in Section 2.3.2). MFCC features are based on a short-time spectral analysis of the speech signal and are extracted as follows:

1. The speech signal is segmented into successive time sections called frames in an overlapping manner (acoustic properties of the speech signal are assumed to be constant throughout one frame). The length of a frame is 20 ms with a frame shift every 10 ms.

2. After pre-emphasis, which is done using a first-order high pass filter with the coefficient 0.95, every frame is multiplied with a Hamming window function to fade out the signal values in direction to the frame margins.

3. Using Fast Fourier Transformation (FFT), the data is translated to the spectral scale and the frame power spectrum that provides intensity values for the discrete frequencies of the speech signal is calculated.

4. The power spectrum is filtered with a band pass to extract the specific frequency range (frequencies below 250 Hz and above 3400 Hz are ignored).

5. The filtered FFT output is passed through a set of triangular filters that are arranged in Mel-frequency spacing which is related to the human ear. The logarithm is applied to the filter output energies.

6. Channel compensation then eliminates the influence of acoustic transfer properties of varying input channels.

7. The final feature vector is formed using different components:

   - logarithmic filter energies,
   - filtered total logarithmic frame energy,
   - first derivations to time calculated as differences to one of the previous frames, and
   - second derivations calculated as differences between first derivations.

   The first and second derivations are used in order to include information about how the acoustic parameters change with time.

8. A time series of the baseline feature vectors is aligned to a super-vector and is then linearly transformed using a transformation based on Linear Discriminant Analysis (LDA). On the one hand, the LDA based transformation reduces the correlation between feature components. This is crucial for the applied Gaussian density modeling using no covariance matrices. On the other hand, the LDA transformation rotates the feature space in such a way that the lower feature vector components contain a maximum on discriminant information. Therefore, the resulting feature vector can be reduced in size without major loss of discriminant information.

The feature vectors processed in this way separate the phonetic segments of the input speech and are used further for modeling or recognition according to the particular LID system's design.

5.2.2 GMM-based System

The LID system based on GMM is used in this thesis as an example of pure spectral approach to the language identification problem. Following Matejka et al. [83], the 56-dimensional feature vectors are taken to present a speech utterance in this thesis: seven MFCC coefficients (including c_0 coefficient) are stacked with the set of SDC features computed by applying a 7-1-3-7 SDC scheme.

Under the GMM framework, language-specific models are realized as the 2 048-order GMM. The complete GMM-based LID system consists of a feature extraction preprocessor, a GMM for every target language, and a backend classifier.

The architecture of the system is presented in Figure 5.1:

- A test utterance is transformed into a sequence of MFCC feature vectors (first seven coefficients are taken for further processing).
- SDC features are calculated as described in Section 3.1.
- 56-dimensional feature vectors are formed to represent the test utterance.
- Each GMM presented by a set of parameters λ_i receives a sequence of feature vectors

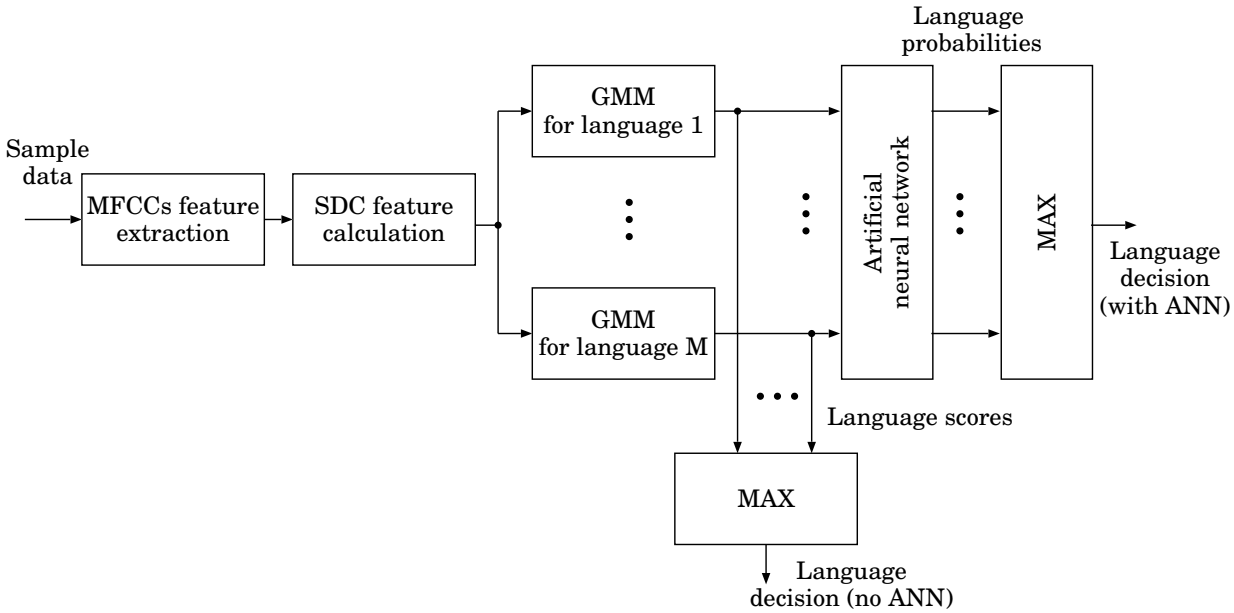


Figure 5.1: *Example of GMM-based LID system for M languages*

$\mathcal{X} = \{x_1, \ldots, x_T\}$ and produces the likelihood:

$$g_{GMM}(\mathcal{X} \mid \lambda_i) = \log p(\mathcal{X} \mid \lambda_i) = \sum_{t=1}^{T} \log p(x_t \mid \lambda_i), \tag{5.1}$$

where $p(x_t \mid \lambda_i)$ is computed as in Equation 3.1.

- The system decision is made either by taking a maximum of the log-likelihood language scores or by using an artificial neural network. The ANN takes language scores as an input and produces the a-posteriori probabilities for every language. The language with maximum probability is hypothesized as the answer of the system.

5.2.3 HMM-based System

The proposed HMM LID system is based on the evaluation of likelihood scores provided by language-specific phoneme recognizers. The phoneme recognizers are realized as three-state monophone HMM with fixed transition penalties. Every HMM uses a set of 2 048 Gaussian densities with diagonal covariance matrices and only one global variance parameter. In order to incorporate phonotactic constraints into the system, phoneme recognizers use integrated language-specific bigram models during the Viterbi decoding process. As a result, phoneme recognizers produce the joint spectral-phonotactic likelihoods that correspond to the most likely phoneme sequences that are optimal with respect to the combination of both spectral and phonotactics information.

The HMM-based LID system functions in the following way:

- A test utterance is transformed into a sequence of MFCC feature vectors.
- Each phoneme recognizer achieves MFCC features computed from the test utterance as input.
- Viterbi decoding is performed once for each recognizer.
- For each frame x_t, each phoneme recognizer L_i delivers a likelihood score $-\log P(x_t \mid Q_l, L_i)$, where Q_l denotes the HMM state of the optimal path found by the Viterbi search. For the sequence of frames $\mathcal{X} = x_1, \ldots, x_T$, the spectral language score is defined as

$$g_{HMM}(\mathcal{X} \mid L_i) = -\sum_{t=1}^{T} \log P(x_t \mid Q_L, L_i)). \tag{5.2}$$

- Language-specific scores are normalized. Due to the fact that the denominator in Equation 2.3 was removed (as shown in Section 2.3.1), the score produced by the phoneme recognizer has to be normalized to compensate it. Since the score is the logarithmic counterpart of the probability, the logarithmic value of $P(\mathcal{X})$ should be subtracted from the score: The value of $P(\mathcal{X})$, defined as the sum of several weighted density functions, is approximated by the maximum component. Then using 'neglog' transformation, the component with minimum score is subtracted from the total score produced by the phoneme recognizer. This common normalization technique is described in detail in [62, 63]. Finally, the score is normalized with respect to time (i.e., is divided by the number of frames).
- The system decision is then made either by taking a minimum of the normalized language scores, or by using the respectively trained ANN as it was done for the GMM-based system (described in Section 5.2.2).

The architecture of the proposed LID system is presented in Figure 5.2.

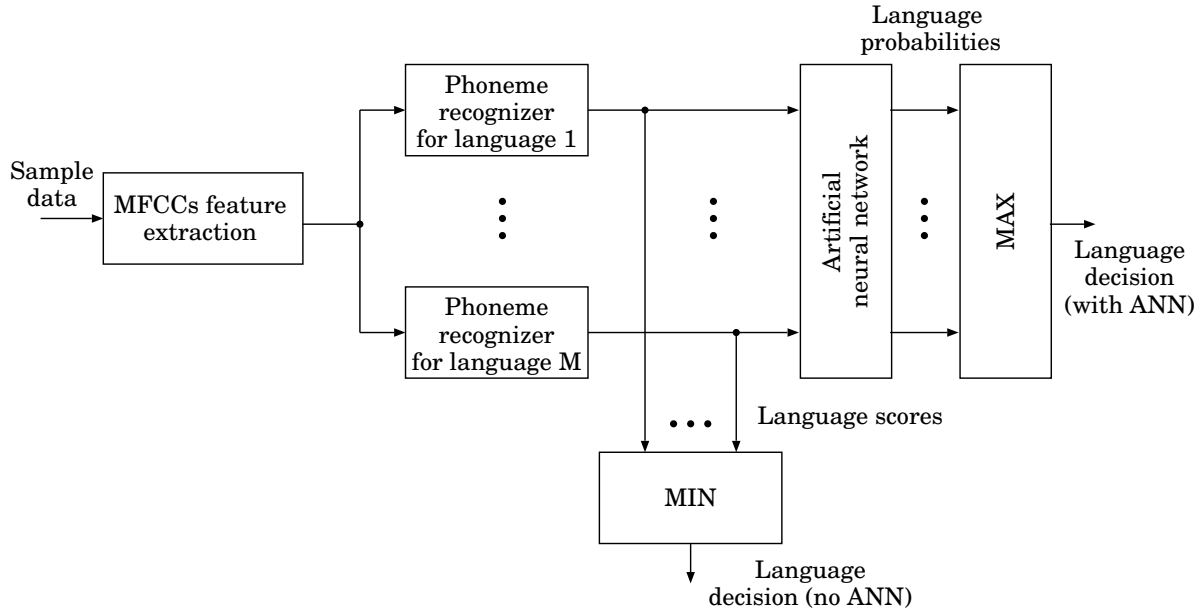


Figure 5.2: *Example of HMM-based LID system for M languages*

5.3 Phonotactic LID system

The phonotactic LID system is implemented using parallel Phone Recognition followed by the Language Modeling (PRLM) approach already discussed in Section 2.3.3. In these systems, several language dependent phone recognizers are used in parallel to tokenize the input utterances. The phone sequences produced by the recognizer are analyzed by applying a statistical language model for each language in the set. Since language models are estimated from the output of the language dependent phone recognizers, the labeled training data does not need to be available for the languages to be identified. The number of recognizers may vary according to the number of languages for which labeled training material is available.

For the PRLM LID system presented in this thesis, the language-specific HMM trained for the spectral LID system (described in Section 5.2.3) are used to tokenize the input utterances. The language dependent scores produced by phoneme recognizers are ignored and only corresponding phoneme sequences are used for further processing through language models that are implemented as N-grams with $N = 3$ (trigrams).

The phonotactic LID system operates in the following way:

- MFCC features are computed from a test utterance as was done for the HMM-based LID system.
- The features are passed through all phoneme recognizers in parallel, producing one phoneme sequence per recognizer.
- For every phoneme sequence $C = c_0, \ldots, c_K$ and every trigram language model that corresponds to language L_i, the negative log-likelihood score is computed as

$$g_{PRLM}(C \mid L_i) = -\sum_{k=1}^{K} \log P(c_k \mid c_{k-2}, c_{k-1}, L_i). \tag{5.3}$$

- The scores are normalized by the number of phonemes in the sequences.
- The decision of the system can be produced using one of the following possibilities:
  1. The pseudo maximum-likelihood decision rule is to hypothesize a language with the minimum score value.
  2. The pseudo maximum-likelihood decision rule is to hypothesize a language with the minimal transformed score. According to the proposed simple score transformation, the resulting language score is computed as a sum over the scores

for the corresponding language models coming from every phoneme recognizer. This score transformation (as will be shown experimentally) can improve the system performance by more than two times the performance with a simple minimum decision rule.

3. The scores are transformed into probabilities using a respectively trained ANN as it was done for the other systems presented in this thesis.

A block diagram of the proposed parallel PRLM system is presented in Figure 5.3.

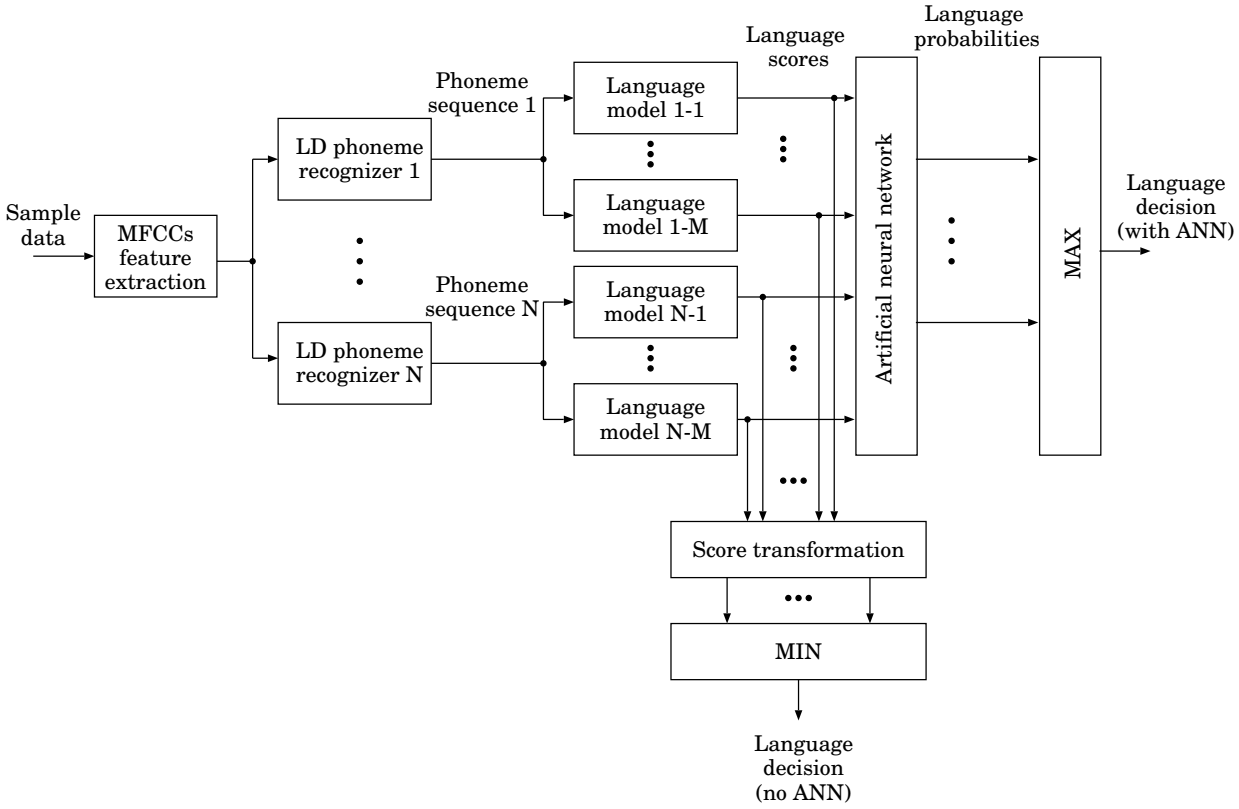


Figure 5.3: *Parallel PRLM block diagram for a LID system with N phoneme recognizers and an M languages task*

5.4 Rhythm LID system

Depending on the method used for the segmentation of speech into rhythm units, the rhythm LID system can work in various operating modes. As it was proposed in Section 4.3, the segmentation of the speech utterances into sequences of syllable-like units can be performed:

- either by segmentation into so-called pseudo-syllables as displayed in Figure 4.4, or
- by detecting the vowels that are used as milestones for computing syllable-like units using the algorithm illustrated in Figure 4.5.

Accumulating the above possibilities, extraction of the rhythm features is expanded to the scheme presented in Figure 5.4.

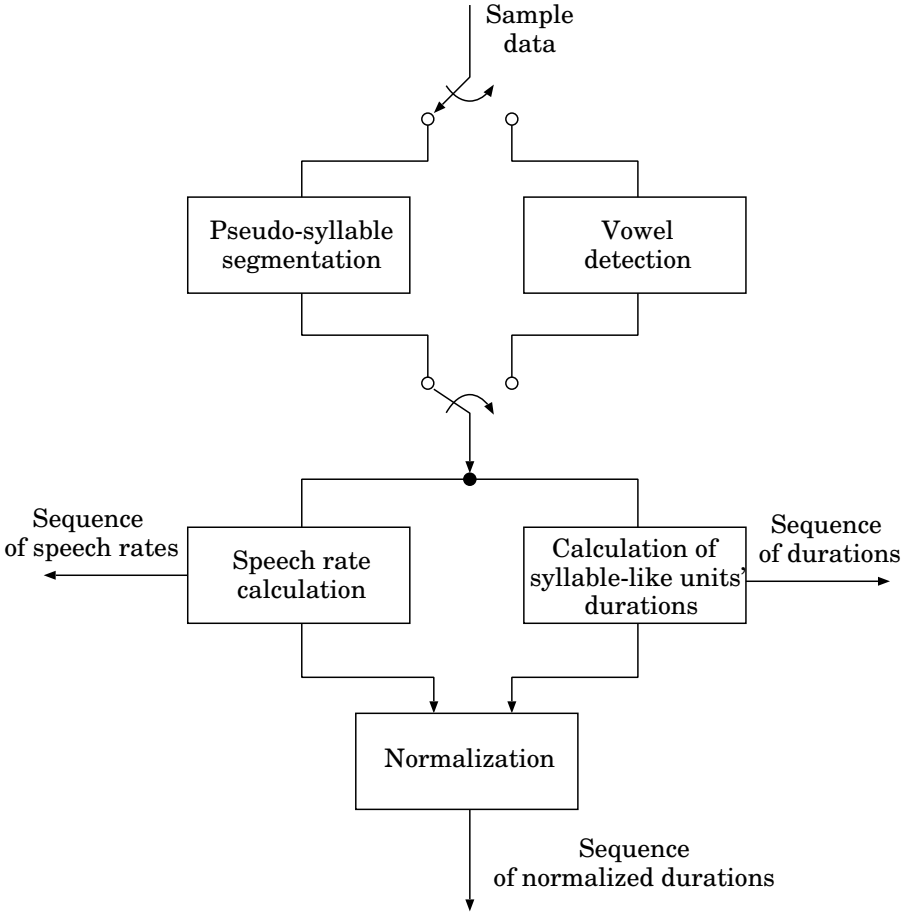


Figure 5.4: *Extraction of different rhythm features*

Then the following can be used as rhythm features:

1. durations of syllable-like units;
2. durations of syllable-like units normalized by speech rate;
3. speech rates.

The LID system works in the following way:

1. A test utterance is presented as a sequence of syllable-like units.
2. Rhythm features are computed according to Figure 5.4.
3. Language-specific rhythm scores for all types of rhythm features are computed as described in Section 4.4.
4. The system makes a decision either with the pseudo maximum likelihood rule or with the help of the appropriate trained ANN as described in the previous sections.

The general architecture of the rhythm LID system is presented in Figure 5.5.

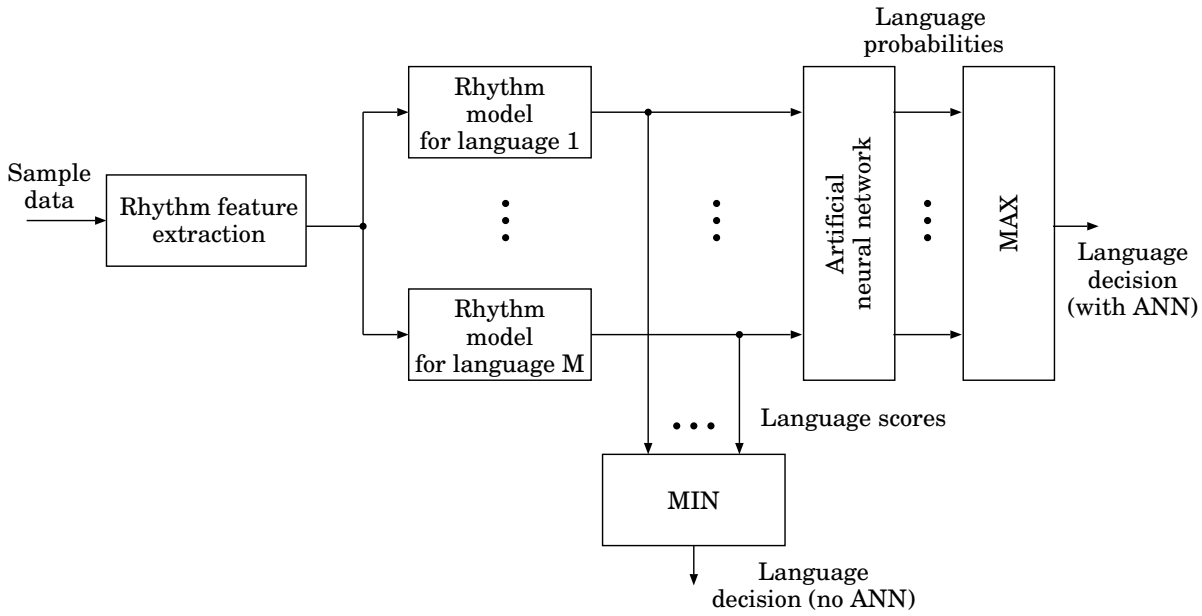


Figure 5.5: *Example of a rhythm LID system for M languages. Rhythm models depend on the type of rhythm features they use*

5.5 LID System Evaluation

This section focuses on the problem in the evaluation of different LID systems. The system performance is estimated by running the system on a large set of test data whose correct outputs are known. Since the work is performed with the NIST Language Recognition Evaluation in mind, two different application scenarios are investigated: Identification and Detection.

5.5.1 Identification Scenario

In the identification scenario, a set of possible languages $L_1, \ldots, L_N$ is considered to each utterance. The set can either contain the explicitly specified language classes (so-called *closed-set* task), or include an additional class that denotes the *none-of-the-above*, or the *out-of-set* language hypothesis (*open-set* task). This thesis is concentrated on the *closed-set* task.

The **recognition question** for the identification scenario is: Which of the N classes does the input utterance belong to?

The LID system delivers one language out of the set of possible languages as the classification result. This result is either correct or incorrect.

The most widely used measure is the so-called error rate (ER) defined as the ratio of classification errors and the number of tested utterances:

$$ER = \frac{Nr_{Erorrs}}{Nr_{Utterances}}. \tag{5.4}$$

To compare the performance of the LID systems with different settings, the relative reduction of ER will be used. For two error rates ER1 and ER2 (ER2 is smaller than ER1), the relative reduction of error rate is defined as

$$100\% - \frac{ER2}{ER1} \cdot 100\%. \tag{5.5}$$

For this case, one can say that there is an improvement of the system performance with relative reduction of the ER computed in such a way.

5.5.2 Detection Scenario

Detection of a target class can include both *closed-set* and *open-set* identification problems. In the detection scenario, a so-called confidence measure is assigned to all classification results. A confidence measure is used to indicate the likelihood that the classification is correct.

The detection scenario is considered for all LID systems presented in this thesis. First, the confidence measures are calculated in one of the following ways:

1. If the system's decision is based on the minimum of the language scores produced by corresponding models, then the confidence measure of the result is equal to the exponential function applied to the negative score.

2. If the system's decision is based on the maximum of the language probabilities produced by an ANN, then the confidence measure of the result is the corresponding output of the ANN.

The confidence measure is compared to a certain threshold. If it is below this threshold, rejection takes place. If it is above the threshold, acceptance takes place. For every utterance, N classification systems are applied (N being the number of considered languages). Each system is based on a single language L_i, $i \in \{1, \ldots, n\}$ with the following **recognition question**: Does the input utterance belong to class L_i? There are four possible cases for each of the N classification systems:

- Correct acceptance occurs when the spoken language of the utterance is identical to the considered language L_i of the system and no rejection takes place;

- Correct rejection occurs when the spoken language of the utterance is different from the considered language L_i of the system and rejection takes place;

- False rejection (or miss) occurs when the spoken language of the utterance is identical to the considered language L_i of the system but rejection takes place;

- False acceptance occurs when the spoken language of the utterance is different from the considered language L_i of the system but no rejection takes place.

It is obvious that the LID system performance depends on how accurately the languages are identified. The values of false acceptance and false rejection give additionally a measure of system performance. In speech applications, Martin et al. [78] have found it useful to

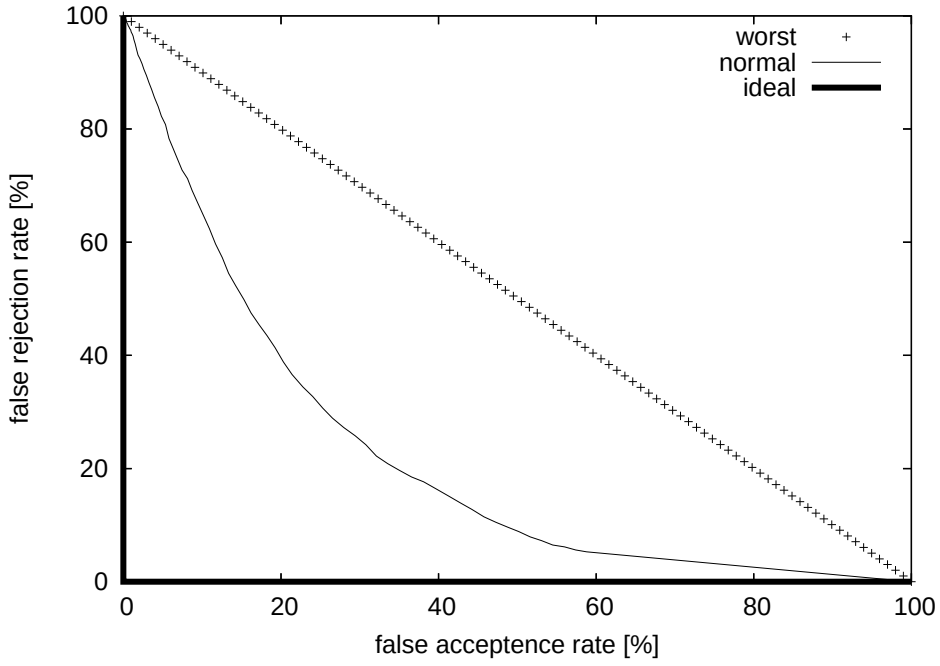


Figure 5.6: *Examples of DET curves*

evaluate the system performance on detection tasks as a trade-off between these two error types. This characteristic is called Detection Error Trade-off (DET).

An examination of the DET can help to determine the reliability of the system scoring mechanism. The DET reveals how the system performance is affected by the introduction of a rejection region (so-called threshold). The DET is calculated by setting a threshold on the system's top-choice score and rejecting all utterances which fall below that threshold. The threshold is varied from 0 % to 100 % rejection to examine the system's accuracy.

These results are presented graphically using DET curves. The DET curve is a plot of false acceptance rate (as percentage) versus false rejection rate (also as percentage) and thus gives equal emphasis to both types of errors. Figure 5.6 illustrates the examples of the DET curves that present the system with worst, ideal, and normal detection characteristics. The more closed the curve to the point $[0, 0]$, the better characteristics of the system.

The DET curve also provides various operating points that allow for the evaluation of the system's performance. The most important point is the so-called equal error rate (EER) of a particular test. The EER is the point at which the false acceptance rate is equal to the false rejection rate. The EER can be found as a point where the DET curve intersects with the "false rejection rate = false acceptance rate" line. Consequently, the smaller the EER, the higher the accuracy of a system. The EER indicates the point where the cost of

each type of error is equal and therefore is an important characteristic of the LID system.

To evaluate the performance of different LID systems developed in this thesis, along with the ER and EER, the so-called cost function is used. The cost function was proposed by NIST 2007 LRE[1] to evaluate the performance of the detection LID system characterized by its miss and false alarm probabilities. To define the cost function, the following is computed for each test:

- Miss Probability for each target language - $P_{Miss}(L_T)$,
- False alarm probability for each target/non-target language pair - $P_{FA}(L_T, L_N)$.

These probabilities are combined into a single number according to an application-motivated cost model:

$$C(L_T, L_N) = C_{Miss} \cdot P_{Target} \cdot P_{Miss}(L_T) + C_{FA} \cdot (1 - P_{Target}) \cdot P_{FA}(L_T, L_N),$$

where the parameters C_{Miss} and C_{FA} represent the relative costs of a miss and a false alarm respectively and P_{Target} is the a-priori probability of the target language. As used in NIST 2007 LRE:

$$C_{Miss} = C_{FA} = 1 \text{ and } P_{Target} = 0.5.$$

The average cost for the whole system with a closed-set task is computed as follows:

$$C_{avg} = \frac{1}{N_L} \cdot \sum_{L_T} \left\{ \begin{array}{l} C_{Miss} \cdot P_{Target} \cdot P_{Miss}(L_T) \\ + \sum_{L_N} C_{FA} \cdot P_{Non-Target} \cdot P_{FA}(L_T, L_N) \end{array} \right\} \tag{5.6}$$

where N_L is the number of languages in the closed-set test;
$P_{Non-Target} = (1 - P_{Target})/(N_L - 1)$.

Defined in this way, C_{avg} represents the expected cost of making a detection decision and is used in this thesis to compare the performance of different LID systems.

[1] http://www.itl.nist.gov/iad/mig/tests/lre/2007/LRE07EvalPlan-v8b.pdf

5.6 Fused LID System

For a language identification system that consists of several individual subsystems, a challenge is to find a combination (or fusion) technique that allows the final system to efficiently use the complementary information of every subsystem in order to improve the performance. This section describes two different approaches to the fusion problem: namely, using a neural network as combination technique and using the recently introduced FoCal Multi-Class [16] Toolkit.

5.6.1 ANN Fusion

In the case of ANN fusion, the resulting language scores coming from different LID systems are used as inputs for an ANN as presented in Figure 5.7.

The ANN is implemented in the same way as it was done for post-classification of the results of individual LID systems. The only difference is the number of input nodes which is equal to the number of LID systems to be combined multiplied by the amount of languages to be identified. The ANN receives the complete set of normalized language scores from the LID systems and extracts the corresponding a-posteriori probabilities for every language in the task. The language with the maximum a-posteriori probability is hypothesized as the identification result.

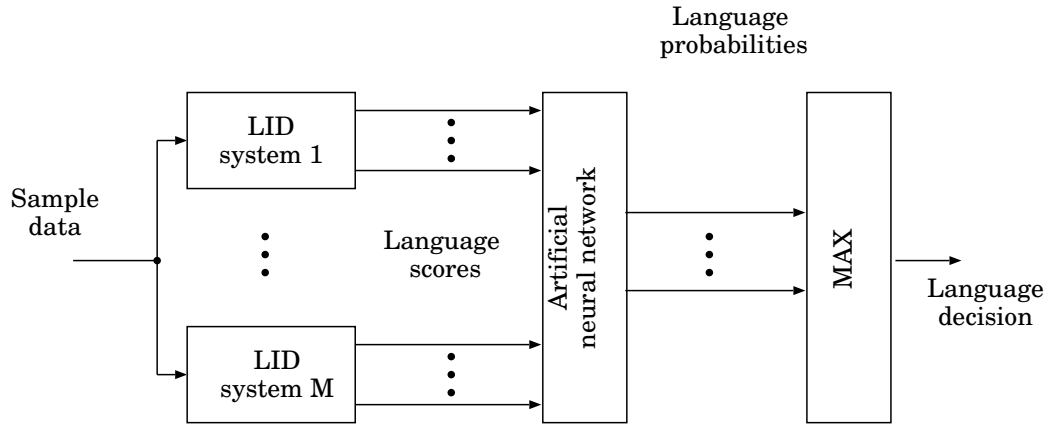


Figure 5.7: *ANN fusion of M different LID systems*

5.6.2 FoCal Fusion

In the recent 2007 NIST LRE[2], the FoCal Multi-Class [16] designed by Brümmer was presented as a free, open-source MATLAB Toolkit for evaluation, calibration, fusion of, and decision-making with, multi-class statistical pattern recognition scores. FoCal was successfully approved[3] on the language recognizers of seven different research teams participating in NIST 2007 LRE and, because of this, was used as an alternative fusion mechanism for the combination of proposed LID systems.

The FoCal tool proposes a discriminative fusion of different LID systems whose results should be presented in log-likelihood form. In the following, the basic principles of FoCal are explained in more detail.

Let there be N classes with $H_1, H_2, \ldots, H_N$ class hypotheses, K input recognizers and T trials, where, for every trial t:

- x_t is the input speech segment;
- $c(t) \in 1, 2, \ldots, N$ denotes the true class;
- $\overrightarrow{l_k}(x_t)$ is the log-likelihood vector of scores produced by the kth evaluee-recognizer for every trail t.

Then the fusion transformation is performed to obtain the resulting log-likelihood vector:

$$\overrightarrow{l'}(x_t) = \sum_{k=1}^{K} \alpha_k \overrightarrow{l_k}(x_t) + \overrightarrow{\beta},$$

where α_k is a positive scalar and $\overrightarrow{\beta}$ is an N-vector.

The fusion parameters $\Lambda = (\alpha_1, \alpha_2, \ldots, \alpha_K, \overrightarrow{\beta})$ are constant for all trials of an evaluation and can be found using an objective cost function called *multi-class* C_{llr}:

$$\Lambda = \operatorname{argmax} C'_{llr}$$

C'_{llr} is calculated for the fused $\overrightarrow{l'}()$ over a supervised training databases as:

$$C'_{llr} = -\frac{1}{T} \sum_{t=1}^{T} w_t \log_2 P'_t,$$

[2]http://www.itl.nist.gov/iad/mig/tests/lre/2007/index.html
[3]http://www.fit.vutbr.cz/research/groups/speech/servite/2009/20070226NBrummer.pdf

where P'_t is the posterior probability for the true class of trial t, as calculated from the given log-likelihoods and the flat prior, $P(H_j) = P_{flat} = \frac{1}{N}$:

$$P'_t = P(H_{c(t)}|\overrightarrow{l'}(x_t)) = \frac{\exp(l'_{c(t)}(x_t))}{\sum_{j=1}^{N}\exp(l'_j(x_t))}$$

and where w_t is a weighting factor to normalize the class proportions in the evaluation trials:

$$w_t = \frac{P_{flat}}{Q_{c(t)}}, \quad Q_j = \frac{Nr.\ of\ trials\ of\ class\ H_j}{T}$$

If there are an equal number of trials of each class, then $w_t = 1$.

In order to use FoCal to combine individual LID systems proposed for this thesis, all recognition scores are transformed as required to the log-likelihood vector. The fusion transformation of the log-likelihood vectors is performed as described above, applying fusion parameters estimated over training data.

6

Experiments and Results

This chapter provides the results of all experiments performed in the scope of this thesis. All proposed LID systems are implemented in the spirit of the NIST evaluation paradigm (explained in more details in Section 2.3). However, to test the systems an own evaluation data set is used. This set includes languages from the widely used SpeechDat database family [130] designed to train commercial speech recognizers as introduced in the next section. The choice of the SpeechDat databases for this thesis is motivated by the availability of great amount of training and evaluation data for every language provided and existence of transliterations for all utterances that were necessary for some investigations described later in this chapter.

The performance of all LID systems is first presented separately and then used together in different combination to demonstrate how different types of information contained in a speech signal influence the results. In order to show that the proposed LID systems are on the level with other state-of-the-art systems, data from the NIST 2005 Language Recognition Evaluation[1] was used to compare the performance of some LID systems designed for this thesis with analogous ones reported by other researchers. These results will be presented in Section 6.7.

6.1 Evaluation Data

For testing and evaluating the LID systems described in this thesis, the SpeechDat [130] corpus was used. SpeechDat is a series of speech data collection projects funded by the European Union. The aim of SpeechDat is to establish speech databases for the development of voice operated teleservices and speech interfaces. Due to its design, the databases are very well suited to investigate LID systems. The SpeechDat family covers a wide range of

[1] http://www.itl.nist.gov/iad/mig/tests/lre/2005/

Set's characteristic	Training set	Development set	Test set
Mean utt. length (sec)	7.04	6.88	6.90
Data amount (hours)	823.14	26.44	7.98

Table 6.1: *Amounts of speech data used in this thesis*

languages and is continuously extending (for example LILA project). It allows evaluating LID approaches for various language sets and studying language-specific properties.

The main objective of SpeechDat was to produce speech databases with a large coverage of languages and applications.

The main features of these databases can be described as follows:

- coverage of different applications (application-oriented words, phonetically rich sentences, spontaneous utterances);
- coverage of speaking styles (commands, carefully pronounced words and spontaneous speech);
- coverage of environmental influences (mobile and fixed telephone networks);
- suitability to develop, train, and test speech processing systems.

In particular, for all experiments, the following seven databases from the fixed telephone network of SpeechDat II project were used: Dutch (NL), English (EN), German (DE), French (FR), Italian (IT), Polish (PL), and Spanish (ES).

The data for each language was divided into several subsets using utterances from speakers defined in SpeechDat II for training, for development, and for testing so that they do not overlap and do not contain utterances with common wordings as it was done in [26]. The amounts of data per language average to 30 hours of speech for training and 3.5 hours for development and fusion. For each language, about 600 test utterances were used. Every utterance is about 6–7.5 seconds long, which corresponds to about 12–20 syllable-like units. All sets contain only phonetically rich sentences. Table 6.1 contains the summary information on the amount of data taken.

6.2 GMM-based system

For the GMM-based LID system, the GMM were trained under the ML framework using SDC features as described in Section 5.2.2 for all seven languages (DE, EN, ES, FR, IT, NL, and PL) using the training data. For each input speech utterance, the GMM produces scores which have log-likelihood nature: Most positive score favors the corresponding language. To explore the ability of neural networks to classify the languages, an ANN for the corresponding setup was created. To train the ANN, the LID experiment was performed for the development set of data. The resulting language scores were used as the input for training procedure to adjust the weights of the ANN.

The LID decision is made either directly by taking the language with maximum score or by processing the scores through the respectively trained ANN and taking the maximum of output probabilities. Error rates for each language separately for both cases are shown in Table 6.2. Despite the fact that the ANN has negative influence on English and Dutch languages, the language-specific error rates throughout the whole language set are more well-balanced.

To compare the GMM systems without and with ANN for different application scenarios, the following performance measures for the whole language set were calculated and are summarized in Table 6.3: mean error rates, equal error rates, and system's costs. Unfortunately, only a small improvement in mean ER is achieved while using ANN for identification scenario. The detection abilities of the GMM system are presented by EERs and cost measures C_{avg} computed as described in Section 5.5. The last column in Table 6.3 shows that EER for the system with ANN is significantly (at level 0.001)[2] better than the

[2]More information about statistical significance can be found in [43]

Language	ER in %	
	without ANN	with ANN
DE	16.78	15.77
EN	20.31	18.24
ES	9.53	10.43
FR	13.07	10.40
IT	22.67	22.48
NL	13.37	15.34
PL	23.03	18.34

Table 6.2: *Performance of the GMM LID system for the SpeechDat II database with different post-classifiers: language-specific error rates by identification scenario*

results without ANN. The improvements of mean ER and C_{avg} are not significant.

Performance measure [%]	without ANN	with ANN	significance
Mean ER	16.97	15.86	not significant
EER	10.79	7.90	significant at 0.001
C_{avg}	9.90	9.25	not significant

Table 6.3: *Comparison of different performance measures for the GMM LID system trained and tested on the SpeechDat II database. The last column shows the results of the significance test while comparing the performance of the system without and with ANN*

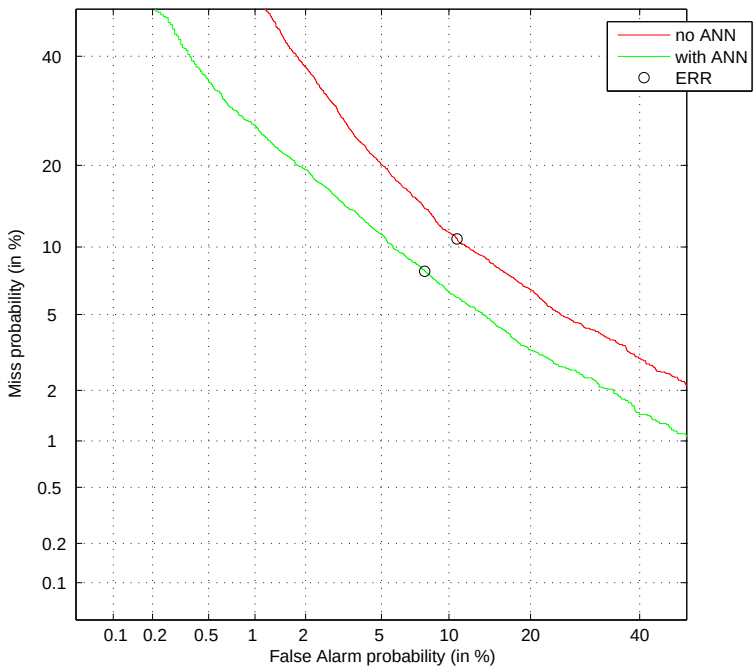


Figure 6.1: *DET curves for the GMM-based LID system trained and tested on the SpeechDat II database*

6.3 HMM-based system

For the HMM-based system described in Section 5.2.3, the language-specific HMM and corresponding bigram models were trained for all languages. The system decision was made based on the minimum of language-specific scores (neglog-likelihoods) produced by phoneme recognizers (without an ANN) for the whole test set. Additionally, an ANN was trained as it was done for the GMM LID system. The ANN was used to classify the languages based on the normalized scores produced by phoneme recognizers.

The results for both experiments (without and with the ANN) are shown in Tables 6.4 and 6.5 and in Figure 6.2. The positive influence of the ANN is demonstrated for all types of performance measures: mean ER, EER, and C_{avg} are reduced respectively by 23 %, 86 %, and 23 %. All performance measures for the HMM system with ANN are significantly better than the results without ANN.

Language	ER in %	
	without ANN	with ANN
DE	7.89	6.71
EN	15.49	7.57
ES	10.79	7.37
FR	4.00	4.00
IT	16.27	11.88
NL	5.62	6.23
PL	8.03	8.57

Table 6.4: *Performance of the HMM LID system for the SpeechDat II database with different post-classifiers: language-specific error rates by identification scenario*

Performance measure [%]	without ANN	with ANN	significance
Mean ER	9.73	7.48	significant at 0.001
EER	26.2	3.56	significant at 0.001
C_{avg}	5.67	4.36	significant at 0.005

Table 6.5: *Comparison of different performance measures for the HMM LID system trained and tested on the SpeechDat II database. The last column shows the results of the significance test while comparing the performance of the system without and with ANN*

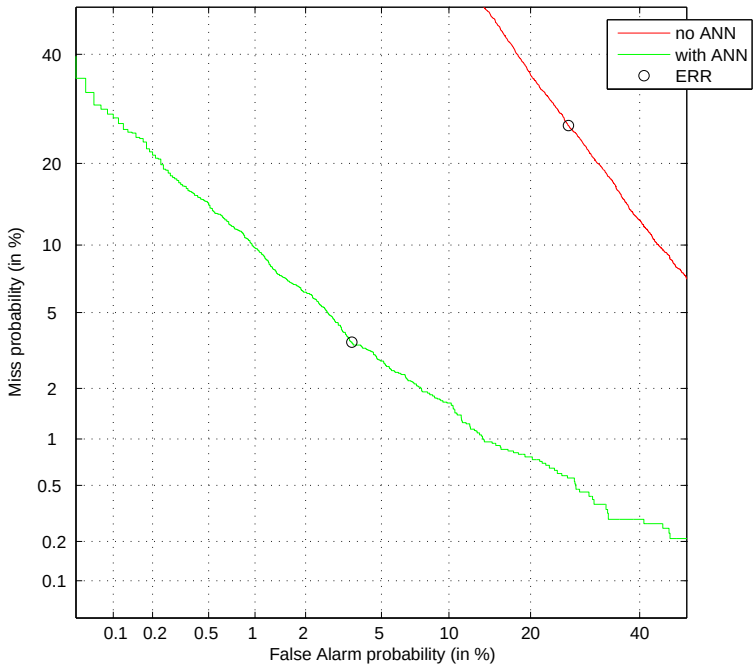


Figure 6.2: *DET curves for the HMM-based LID system trained and tested on the SpeechDat II database*

6.4 Phonotactic system

The phonotactic LID subsystem is realized using seven language-specific HMM (one for every language in the task) for phoneme recognition and 49 (7x7) language-dependent trigrams for backend language modeling. HMM (with integrated 0-gram language models) and trigram models are created using the training data set. The system decision can be made in different ways:

- based on the minimum over all 49 language-specific scores produced by backend trigrams (later marked as *overall Min*);
- based on the minimum over language-specific scores transformed as described in Section 5.3 and producing seven scores (one for each target language) out of 49 initial scores; this case will be denoted as *transf. Min*;
- based on the maximum of language-specific probabilities produced by the ANN, trained specially for this task using development data (*ANN Max*).

Table 6.6 presents resulting ER for separate test lists and for all three possibilities. Different performance measures over the whole languages set are presented in Table 6.7. Relative

Language	ER in %		
	overall Min	transf. Min	ANN Max
DE	32.89	11.07	13.09
EN	45.61	25.47	10.84
ES	17.99	5.22	7.19
FR	28.80	6.67	8.00
IT	44.06	28.34	13.89
NL	41.79	10.94	8.66
PL	22.36	5.49	4.69

Table 6.6: *Performance of the phonotactic LID system for the SpeechDat II database with different post-classifiers: language-specific error rates by identification scenario*

Performance measure [%]	overall Min	transf. Min	ANN Max	significance
Mean ER	33.36	13.31	9.48	significant at 0.001
EER	52.45	67.67	5.92	significant at 0.001
C_{avg}	7.63	7.77	5.53	significant at 0.001

Table 6.7: *Comparison of different performance measures for the phonotactic LID system trained and tested on the SpeechDat II database. The last column shows the results of the significance test while comparing the performance of the system with overall Min and ANN Max*

reduction of the mean ER after applying a simple score transformation is 60 %. As expected, the system with the ANN-classifier outperforms these results, i. e., it provides about 70 % relative reduction. The system with the ANN classifier also gives significantly better results for detection scenario (additionally shown in Figure 6.3).

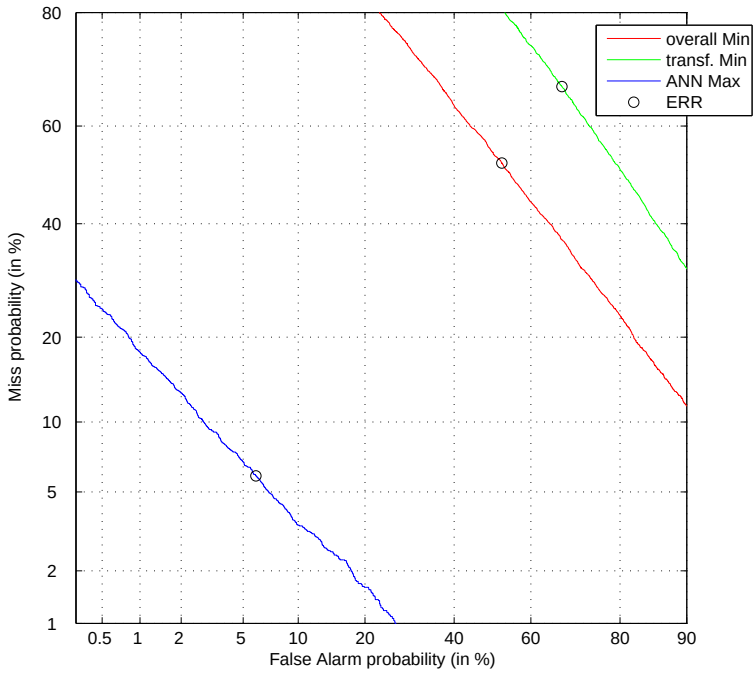


Figure 6.3: *DET curves for the phonotactic LID system trained and tested on the SpeechDat II database*

6.5 Rhythm system

6.5.1 Using pseudo-syllables

For the rhythm LID system based on the pseudo-syllable extraction, a multilingual HMM was estimated from the training data available for all languages together. The corresponding phoneme set consists of combined phonemes from all languages in the task. All phonemes are labeled to differentiate vowels and consonants. The resulting phoneme set presented in SAMPA notation is found in Table 6.8.

Multilingual HMM produces a sequence of phonemes that are then mapped into vowel and consonant classes according to Table 6.8. The phoneme sequence is segmented into the pseudo-syllables and their durations are computed using the phoneme durations produced by the HMM as described in Section 5.4. For the resulting sequence of durations, three types of rhythm features are computed.

The rhythm models were created by computing the histogram statistics for every pair of pseudo-syllables durations provided by training data. The probability distributions of duration are given for discrete values, which are determined by the numbers of frames

Vowels	Consonants	Vowels	Consonants
{	?	*eI*	*n‘*
2	*?‘*	*i*	*N*
2 :	@	*i :*	*p*
3	*b*	*I*	*r*
6	*C*	*I@*	*rr*
9	*d*	*o*	*s*
9	*d‘*	*o*	*s‘*
9y	*D*	*o :*	*S*
a	*D‘*	*O*	*t*
a	*f*	*OI*	*t‘*
a :	*g*	*u*	*T*
A	*h*	*u :*	*v*
aI	*j*	*@U*	*w*
aU	*J*	*U*	*x*
e	*k*	*V*	*z*
e	*l*	*y*	*z‘*
e :	*L*	*y :*	*Z*
e@	*m*	*Y*	
E	*n*		

Table 6.8: *Multilingual phoneme set*

regarded. The discrete distribution values building a histogram are not smoothed. For unseen durations, a fixed floor value is used. In the same way the models for speech rate features and durations of pseudo-syllables normalized by speech rate are created.

The rhythm scores for all languages are calculated and the language with minimal score is hypothesized. Like all LID systems presented in this thesis, the rhythm system also has a post-processing ANN as an additional, optional classifier. The ANN is trained on the rhythm scores from development data and works as described in the previous sections.

The rhythm systems with and without a post-processing ANN were evaluated for different rhythm features as described in Section 4.4:

- durations of pseudo-syllables (corresponding complete results as it was done for other LID systems are presented in Appendix C in Tables C.1 and C.2 and in Figure C.1);
- normalized durations of pseudo-syllables (Tables C.3 and C.4 and Figure C.2);
- speech rates computed using durations of pseudo-syllables (Tables C.5 and C.6 and Figure C.3).

Different performance measures for all rhythm LID systems based on pseudo-syllable segmentation are summarized in Table 6.9. Again the results show that the ANN has a positive impact relative to the rhythm system behavior. The ER and EER are decreased each by 6.6 % relatively for the system based on the durations of pseudo-syllables, by 5.6 % for the system based on the normalized durations, and by 8 % for the system based on the speech rate. All improvements from using ANN as additional classifier are significant in order to be used further.

Direct comparison of the results from Table 6.9 shows that the best performance has the rhythm system based on the non-normalized durations of pseudo-syllables. Introducing speech rate as a feature into the rhythm LID system does not lead to the expected improvement. Possible reasons could be that either the chosen languages do not vary much in speech rate or the method for modeling speech rate is not appropriate for the LID purpose.

To verify the first assumption, the speech rate statistics for different languages is plotted in Figure 6.4. The distributions of speech rate for different languages are similar which can explain poor recognition ability of corresponding rhythm systems.

The second hypothesis is checked in the next section, where speech is segmented into the rhythm units using a vowel detection algorithm.

Performance measure [%]	without ANN	with ANN	significance
Using durations of pseudo-syllables			
Mean ER	72.09	67.33	significant at 0.001
EER	49.29	34.82	significant at 0.001
C_{avg}	42.06	39.28	significant at 0.010
Using normalized durations			
Mean ER	79.89	75.36	significant at 0.001
EER	49.83	39.23	significant at 0.001
C_{avg}	46.61	43.96	significant at 0.010
Using speech rates			
Mean ER	80.39	73.91	significant at 0.001
EER	49.48	38.92	significant at 0.001
C_{avg}	46.89	43.11	significant at 0.001

Table 6.9: *Comparison of different performance measures for the rhythm LID systems based on pseudo-syllable segmentation: trained and tested on the SpeechDat II database. The last column shows the results of the significance test while comparing the performance of the system without and with ANN*

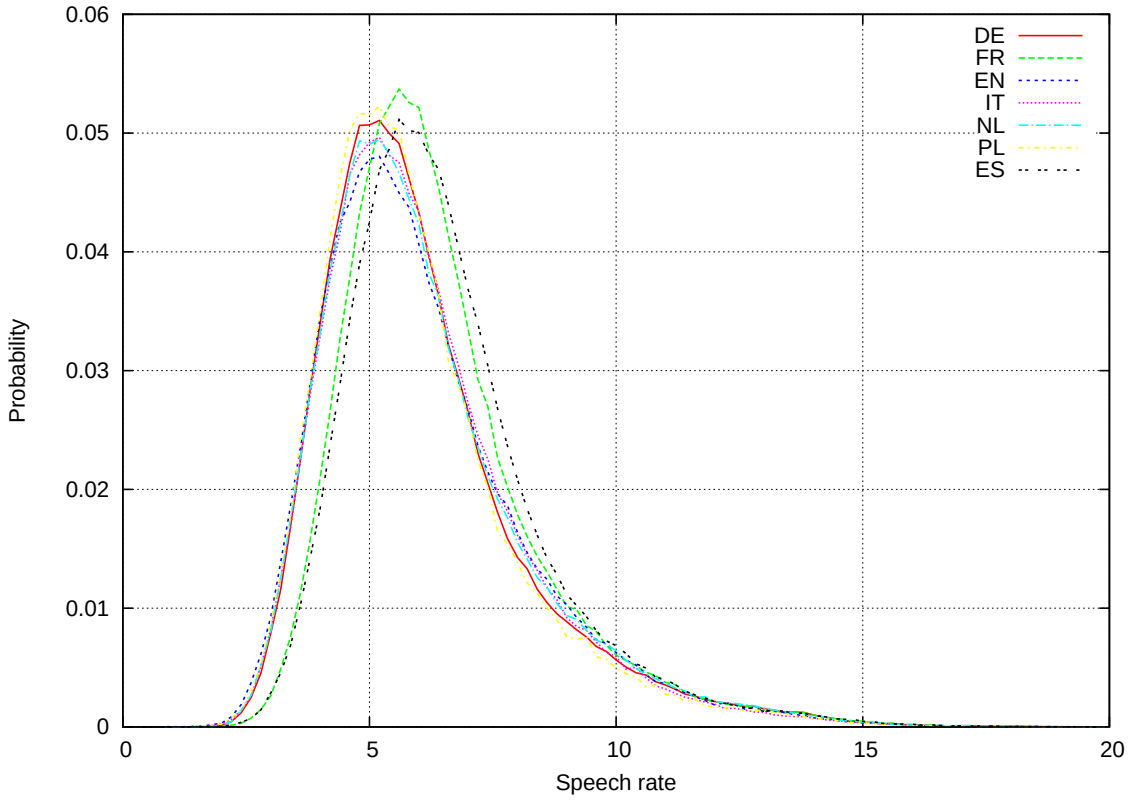


Figure 6.4: *Distribution of the speech rate computed using pseudo-syllables for different languages*

6.5.2 Using vowel detection

For this version of the rhythm LID system, the durations of syllable-like units are computed as the intervals between two successive vowels using the algorithm proposed in Section 4.3.2. The systems based on the durations of syllable-like units, on the normalized durations, and on the speech rate are trained and tested in the same way as the corresponding systems from the previous section. The evaluation results according to the identification scenario are presented in Appendix C in Tables C.7, C.9, and C.11. The corresponding performance measures for all systems are shown respectively in Tables C.8, C.10, C.12 and Figures C.4, C.5, and C.6. Different performance measures are summarized in Table 6.10 and can be compared directly.

Along with the positive impact of the ANN on the systems performance, one can see that results of all rhythm systems based on the vowel detection are even slightly worse than those based on the pseudo-syllable segmentation. The relative difference is about 4 % for the systems using duration of syllable-like units, 1.4 % for systems using normalized

Performance measure [%]	without ANN	with ANN	significance
Using durations of intervals between vowels			
Mean ER	76.05	70.29	significant at 0.001
EER	49.62	37.21	significant at 0.002
C_{avg}	44.36	41.00	significant at 0.010
Using normalized durations			
Mean ER	85.52	76.45	significant at 0.001
EER	49.93	42.23	significant at 0.001
C_{avg}	49.89	44.59	significant at 0.010
Using speech rates			
Mean ER	80.90	79.99	not significant
EER	49.89	47.19	significant at 0.001
C_{avg}	47.19	44.91	significant at 0.050

Table 6.10: *Comparison of different performance measures for the rhythm LID systems based on vowel detection: trained and tested on the SpeechDat II database. The last column shows the results of the significance test while comparing the performance of the system without and with ANN*

durations, and 8.6 % for systems based on the speech rate.

The distributions of speech rates illustrated in Figure 6.5 again do not show significant differences among the languages.

According to the comparison of all possible rhythm LID systems presented above, the system based on the non-normalized durations of pseudo-syllables performed the best and will be further investigated.

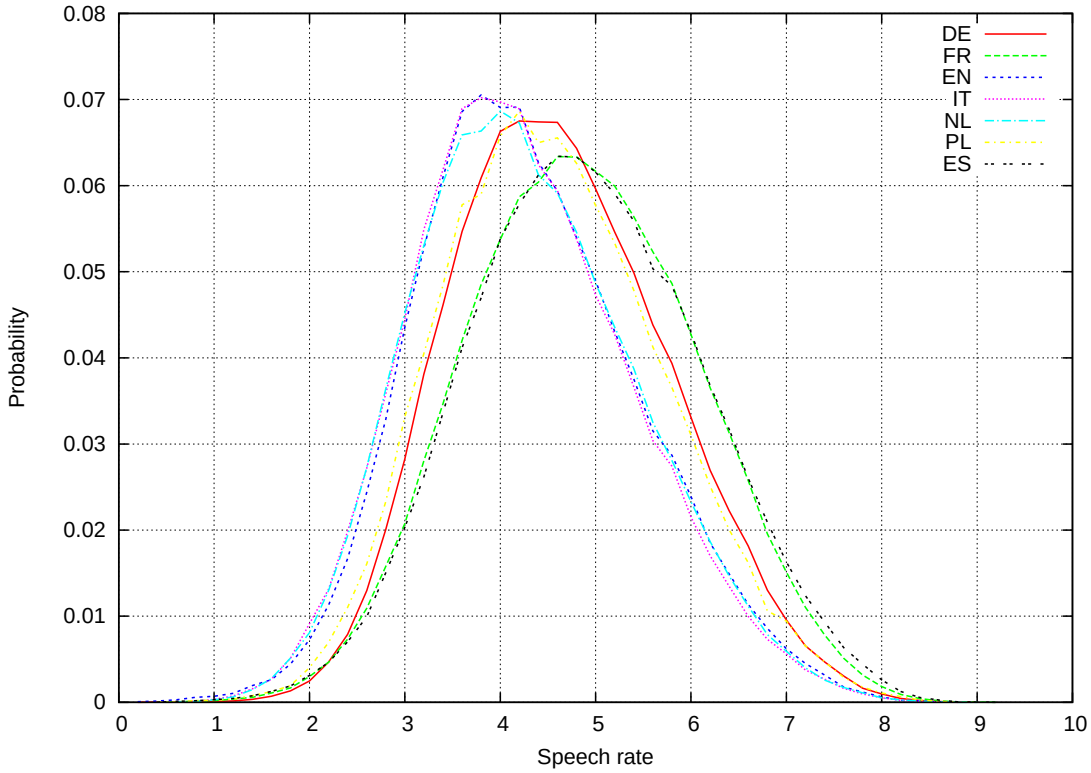


Figure 6.5: *Distribution of the speech rate computed using vowel detection for different languages*

6.5.3 "Cheating" Experiment

The quality of the rhythm LID systems based on the durations of pseudo-syllables highly depends on the correctness of the extraction of pseudo-syllables. The correctness of pseudo-syllables' extraction is, in turn, limited by the accuracy degree of the consonant-vowel segmentation algorithm. In this particular case (for rhythm LID system based on the pseudo-syllables) consonant-vowel segmentation depends on the quality of phoneme recognition. To exclude the influence of the multilingual phoneme recognizer to the recognition ability of rhythm models, a cheating experiment is performed using a forced Viterbi algorithm for the same data but with orthographical transcriptions that give the actual durations of phonemes.

To illustrate the differences in rhythm models, the corresponding distributions on the three dimensional space are plotted. The x-axis presents the duration of a pseudo-syllable i in frames, the y-axis presents the duration of subsequent pseudo-syllable $i+1$, and the z-axis presents the probability for that pair. As an example of such plots, Figure 6.6 presents the

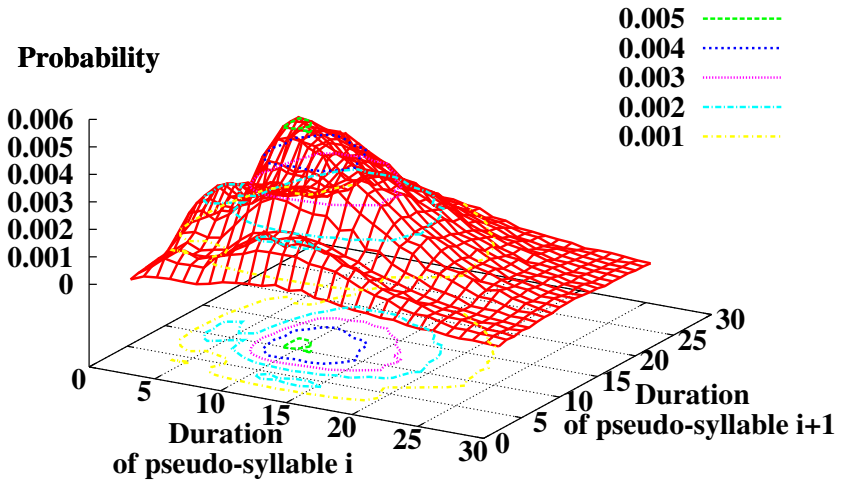


Figure 6.6: *Probability distribution for the German language obtained by a multilingual phoneme recognizer*

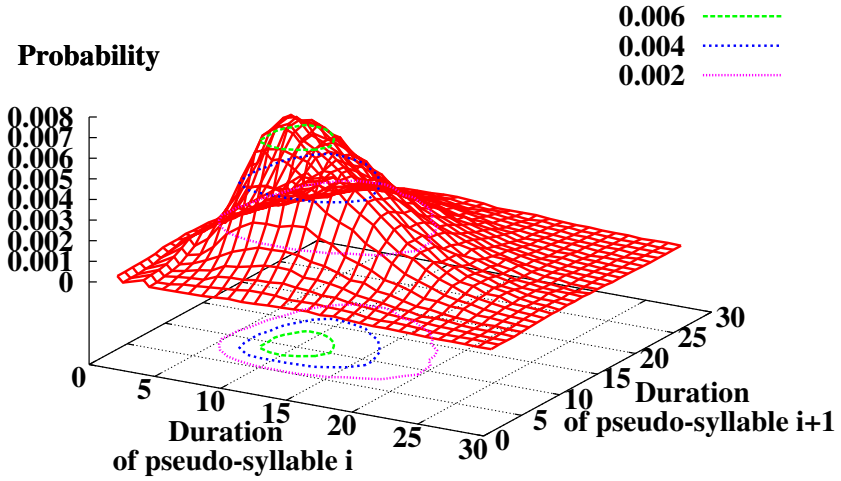


Figure 6.7: *Probability distribution for the German language obtained by a forced Viterbi algorithm*

distribution for the German language evaluated with a multilingual phoneme recognizer.

Here the curves on the x-y surface show the contours for the different probability values and graphically present the German rhythm model. The curves corresponding to the rhythm model obtained by cheating is displayed in Figure 6.7.

The differences between the two plots presented in Figures 6.6 and 6.7 can be explained by the relatively low phoneme recognition rate of the multilingual HMM, which is in the range of 22 % only. Additionally, the accuracy of segmenting speech during acoustic processing can have negative influence on the phoneme recognition performance.

In order to show the influence of phoneme recognition quality on the system's performance, the recognition test is made using cheating data. In this case cheating data means that the consonant-vowel segmentation of the utterance is known and corresponding durations are computed using the Viterbi algorithm. The results are presented in Tables 6.11, 6.12, and in Figure 6.8.

Language	ER in %	
	without ANN	with ANN
DE	70.64	70.47
EN	67.64	50.60
ES	22.66	30.22
FR	78.40	77.33
IT	74.41	62.16
NL	41.79	41.79
PL	47.79	29.32

Table 6.11: *Performance of the rhythm LID system for the cheating experiment using the SpeechDat II database with different post-classifiers: language-specific error rates by identification scenario*

Performance measure [%]	without ANN	with ANN	significance
Mean ER	57.62	51.70	significant at 0.001
EER	46.26	24.27	significant at 0.001
C_{avg}	33.61	30.16	significant at 0.001

Table 6.12: *Comparison of different performance measures for the rhythm LID system for the cheating experiment: trained and tested on the SpeechDat II database. The last column shows the results of the significance test while comparing the performance of the system without and with ANN*

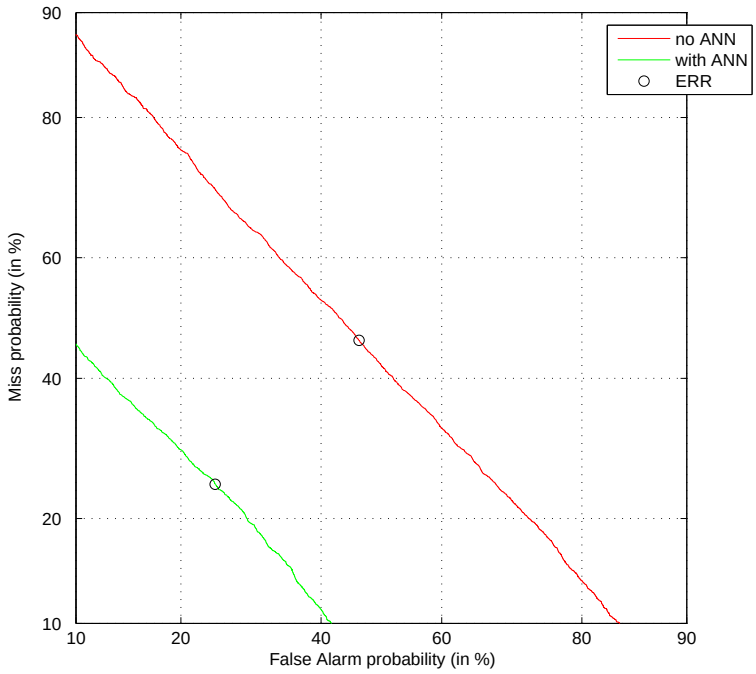


Figure 6.8: *DET curves for the rhythm LID system for the cheating experiment: trained and tested on the SpeechDat II database*

Comparing Table 6.12 with the corresponding results from the first part of Table 6.9, one can see that due to the phoneme recognition mistakes, the rhythm system loses relatively about 30 % for EER and 23 % in ER and in system's cost. This difference also means that the performance of the rhythm system can potentially be improved by using a phoneme recognizer close to "ideal".

Comparing the rhythm results in contrast with other systems, one can see that the discriminating abilities of rhythm as well as all prosody-based LID systems are not high enough to be used separately. The rhythm can be effective in combination with spectral and/or phonotactic systems as will be shown in the next section.

6.6 Combination of individual LID systems

The combination of different LID systems presented in this thesis can be performed using one of the following techniques:

1. Using an ANN trained as discussed in Section 5.6.1.
   To train the ANN, recognition tests are performed on the development data for the systems, which have to be combined. The resulting language-specific scores together with their true identities are used to estimate corresponding ANN parameters.

2. Using the FoCal tool [16] described in more detail in Section 5.6.2.
   The FoCal tool requires as input the scores that have log-likelihood nature, i. e., the most positive score favors the corresponding language. For fusion, the results produced by ANN are taken. In order to suit the FoCal input format, the probabilities of belonging to the particular language (i. e., values from 0 to 1) produced by ANN are first logarithmized. Then the training of fusion parameters is performed on a supervised development set of such logarithmic scores.

In order to choose one of these methods, they are compared on the example of combination of two systems: spectral system based on the GMM (referred to as GMM) and phonotactic system (referred to as PRLM). Table 6.13 presents combination results for GMM and PRLM systems.

Fusion with FoCal gives over 50 % better results than the fusion with ANN and therefore is used to perform the combination of different systems in this thesis.

Table 6.14 displays the results for all possible combinations of the LID systems (including the rhythm system from the cheating experiment in order to see the best possible improvement). The table presents the ER as a system performance measure for identification scenario and costs (C_{avg}) as the quality of detection abilities of the system. Last columns

System	Error Rate (%)
GMM	16.97
PRLM	13.31
Fused by ANN	8.60
Fused by FoCal	4.13

Table 6.13: *Comparison of ANN and FoCal fusion: trained and tested on the SpeechDat II database*

System	ER (%)	C_{avg} (%)	Relative improvement (%)	
			ER	C_{avg}
rhythm	67.33	39.28		
rhythm (cheat)	51.70	30.16	23	23
HMM	7.48	4.36		
HMM+rhythm	3.44	3.35	54	23
HMM+rhythm(cheat)	2.71	2.73	64	37
GMM	15.86	9.25		
GMM+rhythm	6.91	6.86	56	26
GMM+rhythm(cheat)	5.86	5.85	63	37
PRLM	9.48	5.53		
PRLM+rhythm	3.65	6.67	61	-17
PRLM+rhythm(cheat)	3.44	5.73	64	-3.5
GMM+PRLM	4.73	4.6		
GMM+PRLM+rhythm	4.50	4.47	4.9	2.8
GMM+PRLM+rhythm(cheat)	3.95	3.95	16	14
HMM+PRLM	2.92	2.81		
HMM+PRLM+rhythm	2.77	2.75	5.1	2.1
HMM+PRLM+rhythm(cheat)	2.30	2.31	21	17.8
HMM+GMM+pPRLM	2.47	2.45		
HMM+GMM+PRLM+rhythm	2.40	2.28	2.8	6.9
HMM+GMM+PRLM+rhythm(cheat)	2.34	2.25	5.3	8.2

Table 6.14: *Fusion of individual LID systems for the SpeechDat II database*

of the table show relative improvements in comparison with the baseline for both ER and C_{avg}.

The fist two lines show the comparison of the rhythm systems: first — with multilingual HMM as phoneme recognizer, and second — for the cheating models. Here the increasing of the phoneme recognition quality could lead to significant (at level 0.001) improvement of the system performance of up to 23 % relatively for both identification and detection scenarios.

The rest of the table is divided into several groups; each has one line demonstrating the results for the baseline system (HMM, GMM, PRLM, combination of GMM and PRLM, combination of HMM and PRLM, and combination of HMM, GMM, and PRLM) and two lines for the results of its combination with both rhythm systems.

One can see that the combination of individual baseline systems with rhythm decreases the ERs by more than 50 % relatively. Further improving of phoneme recognition for the rhythm system can gain additionally about 10 %. The detection abilities (additionally shown as DET curves in Appendix D in Figures D.1, D.2, and D.3) of these system can be increased by more than 20 % using rhythm systems and by more than 30 % using the cheated rhythm information. The improvement from using rhythm is significant (at the level 0.001) for all performance measures. The only exception is C_{avg} for combination of the phonotactic and rhythm systems which has a slightly increase in the overall cost. This is explained by relatively high false alarm error rate that has direct influence on the EER and systems cost.

The fusion results in general have a common tendency: Every next system leads to less improvement. Therefore adding rhythm to the combination of three baseline systems can improve the performance of the resulting system by almost 3 % for ER and 7 % for the systems cost. Even the cheating gives only additional 2.5 % and 1.3 %, relatively which is not significant. The corresponding DET plots are presented in Appendix D in Figure D.4.

Table 6.14 clearly shows that the rhythm as investigated in this thesis can be successfully used in combination with different baseline systems in order to increase their language recognition abilities.

6.7 State-of-the-art Test

In order to check, if the LID systems proposed in this thesis, are on the level of state-of-the-art systems, an experiment is performed using the data from the NIST 2005 Language Recognition Evaluation. This data mostly comes from CallFriend corpora provided by the Linguistic Data Consortium[3] and from data collected by Oregon Health and Science University (Beaverton, Oregon). The evaluation data consists of speech from the following languages and dialects: English (American and Indian), Hindi, Japanese, Korean, Mandarin (Mainland and Taiwan), Spanish (Mexican), and Tamil. The corpus consists of a series of speech segments containing approximately three seconds, 10 seconds, and 30 seconds of conversational telephone speech for a total of 44.2 hours. For the current thesis, only utterances with the length of 30 seconds are used.

NIST data is widely used to measure the performance of different LID systems and therefore it is the most appropriate for the comparison of different LID systems. The organizers of the NIST Language Recognition Evaluations usually publish the results of several systems that have the best performance. The summary of the results for the year 2005 can be

[3] http://www.ldc.upenn.edu

found in [79]. Published results correspond to the complex LID systems, i. e., combination of several individual subsystems which utilize different sources of information. This does not allow to compare the individual LID systems directly.

One of the participants of the NIST 2005 Language Recognition Evaluation, the Brno University of Technology, published their results of all used in the LRE subsystems separately [83]. According to [79], the LID system from Brno is among the best LID systems. It takes into account the scores from four components, namely, one GMM system and three phonotactic systems based on different phoneme recognizers. Moreover, the paper presents the results for other GMM systems. One of these systems is trained under the ML framework (2 048 components) with SDC features and RASTA filtering and carries an analogy to the GMM system presented in this thesis. This makes the direct comparison possible.

For comparison purpose a GMM system was trained and evaluated in compliance of the NIST 2005 LRE requirements. Table 6.15 presents the costs for both systems before and after using the FoCal tool. Here the value of the Brno cost function after FoCal is taken from the results published in [83]. The costs as the result of pure GMM scores comes from the personal communications with the developers of Brno's system[4]. The performance of the pure GMM system is almost the same. The difference in the results after FoCal may be caused by using additional training data for Brno system, i. e., OGI multilanguage and OGI 22 languages corpora [84].

As the results show, the GMM-system presented here is on the level of similar state-of-the-art LID systems.

System	C_{avg} (%)	C_{avg} (%) after calibration
GMM to compare	17.46	14.91
Brno GMM	16.21	11.60

Table 6.15: *Performance on the NIST 2005 task*

[4]The author would like to thank Marcel Kockmann for providing the results of pure GMM system

7 Conclusions

This thesis provides research results on the investigation of the language identification problem using different types of information extracted from the speech signal. In particular, this thesis is concentrated on the exploration of the speech rhythm for LID.

The main challenge of using speech rhythm for the LID task is the absence of the general definition of the rhythm and the algorithm for incorporation of the rhythm information into the LID system. Various studies on the language rhythm provide controversial results and are not suitable for automatic speech processing. Therefore, in this thesis the new definition of speech rhythm is proposed. A syllable-like unit is taken as appropriate rhythm entity and the durations of such units are explored as language discriminative features. Two different language-independent algorithms are suggested for the segmentation of the speech utterance into syllable-like units. The first one is based on the notion of pseudo-syllable and the second one defines the syllable-like unit as a distance between two successive vowels found by an automatic language-independent vowel detection.

Then various possibilities are proposed to model rhythm of a speech utterance: by the durations of two successive syllable-like units, by the durations of two successive syllable-like units normalized in order to take into account the speech rate, and the speech rate itself. For every possible rhythm feature the corresponding rhythm LID system was created.

It is well known that LID systems based on prosodic information exclusively do not have sufficient discriminative power and should be used in addition to a system based on the segmental information. To investigate this statement, the following baseline LID systems are suggested and implemented: two system utilizing spectral information (based on HMM with MFCC features and based on GMM with SDC features) and one phonotactic system designed as phoneme recognition followed by language modeling (PRLM). In order to be able to get advantages of using several discrimination features, an appropriate algorithm for fusion of different LID systems should be found. From the two proposed methods

(one is based on an ANN and another one on the FoCal tool), the FoCal tool is taken for combination experiments since it has produced results 50 % better in comparison with the ANN.

This thesis is performed in the spirit of the NIST evaluation paradigm but it uses its own evaluation data set based on the languages of the widely used SpeechDat database family. Due to their design, these databases are very well suited to investigate LID systems, to allow for the evaluation of LID approaches for various language sets, and to study language-specific properties. All LID systems were evaluated on a set consisting of seven languages: Dutch, English, German, French, Italian, Polish, and Spanish. About 30 hours of speech were used to train models for each language and about 600 utterances from each language with length of 6–7.5 seconds were tested.

All proposed LID systems were first evaluated separately and then fused together to measure the influence of different features types on the performance of the resulting system. The HMM-based system has shown the best performance of 4.36 % (against other baseline systems: 9.25 % for the GMM system and 5.53 % for the PRLM system). This can be explained by the fact that the HMM system uses bigram language models as integral parts of the phoneme recognizers and that is why it produces the score which is optimal with respect to the combination of both spectral and phonotactics information.

Comparison of the rhythm systems with different proposed features allows the following conclusions to be drawn:

- the notion of pseudo-syllable along with the corresponding algorithm for segmentation of a speech utterance into the sequence of syllable-like events is more suitable for the modeling of rhythmic units;
- despite the existing relation of language rhythm and speech rate, the normalization of rhythmic units by speech rate does not provide the expected improvement in system's performance (results of the rhythm system based on the normalized durations are about 10 % worse);
- since the distributions of speech rates for different languages are not considerably different, the speech rate itself should not be used as discriminative feature for LID.

Based on the conclusions given above, the system based on the duration of successive pseudo-syllable is taken as the resulting rhythm system. Since the performance of this design depends on the quality of segmentation into pseudo-syllables (that in this particular case means on the quality of phoneme recognition), an experiment with known orthographical transcriptions for test data is performed. The results of this cheating experiment have

shown that the performance of rhythm LID can potentially be improved by about 20 % relatively. Therefore, in the future it seems to be reasonable to improve the accuracy of pseudo-syllable segmentation by increasing the quality of the phoneme recognizer.

Results on the combination of the rhythm system with different baseline systems have shown that speech rhythm can be successfully integrated into the LID task. Adding rhythm to each of the baseline systems improves the performance of the resulting system by more then 50 % relatively. Fusing of rhythm with two baseline systems provides from 16 % to 21 % of relative improvement. The combination of all baseline systems with rhythm improves the performance by about 5 % resulting in error rate of 2.34 %. Despite the significant improvement of rhythm results achieved by using best possible phoneme segmentation, the combination with the cheated rhythm system provides a decrease of error rate of about 10 % additionally in case of the two-system fusion and only 1.3 % in case of the four-system fusion.

In order to show that proposed LID systems are on level of other state-of-the-art LID approaches, additional evaluation tests were performed. GMM-based LID system was trained and tested on data from the NIST LRE 2005. The resulting performance is comparable with the results of an analogous system that was among the best participants of NIST LRE 2005.

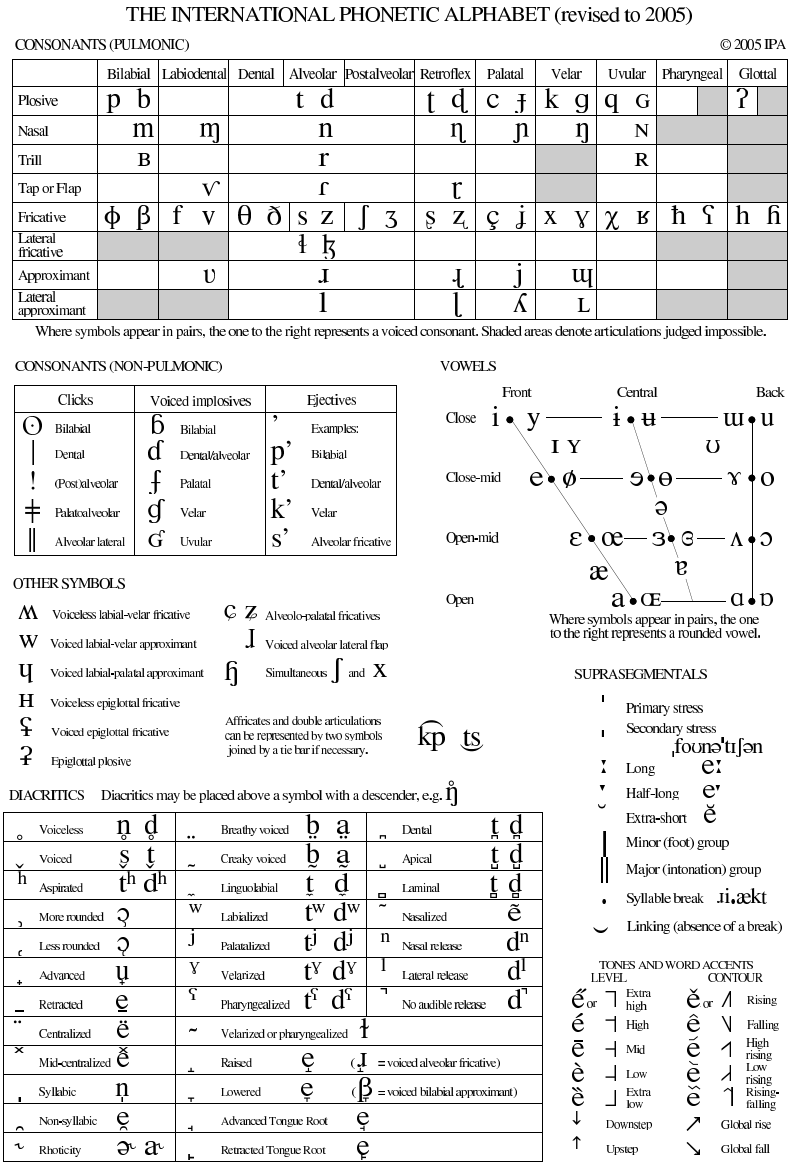
As the result of the investigation of speech rhythm, this thesis comes with the formal definition of speech rhythm suitable for using it as an additional source of information for the LID task. The experimental results show that the speech rhythm modeled by a pair of durations of syllable-like units can be successfully used for LID.

At the same time, there are some potentials to improve the achieved results, which should be further explored. First of all, there is a possibility for the increase of the accuracy of the phoneme recognition that has direct influence on the performance of the rhythm system presented in this thesis. Another potential improvement can be realized by another modeling of the probabilities for the pair of durations. This can be achieved either by applying a Gaussian Mixture Model or by using some continuous function for the approximation. Instead of combing the recognition results of different systems, a common feature vector that contains baseline and rhythm components can be used to represent the speech utterance.

Appendices

A

The International Phonetic Alphabet



Reproduction from the official IPA web-site:
http://www.langsci.ucl.ac.uk/ipa/IPA_chart_(C)2005.pdf

B

German SAMPA

Plosives:		
Symbol	Example	Transcription
p	Pein	paIn
b	Bein	baIn
t	Teich	taIC
d	Deich	daIC
k	Kunst	kUnst
g	Gunst	gUnst
Affricates:		
Symbol	Example	Transcription
pf	Pfahl	pfa:l
ts	Zahl	tsa:l
tS	deutsch	dOYtS
dZ (often replaced by tS)	Dschungel	"dZUN=l [1]
Fricatives:		
Symbol	Example	Transcription
f	fast	fast
v	was	vas
s	Tasse	"tas@
z	Hase	"ha:z@
S	waschen	"vaS=n
Z	Genie	Ze"ni:
C	sicher	"zIC6
j	Jahr	ja:6
x	Buch	bu:x
h	Hand	hant

[1] " denotes primary stress

Sonorants:		
Symbol	Example	Transcription
m	mein	maIn
n	nein	naIn
N	Ding	dIN
l	Leim	laIm
R	Reim	RaIm

Checked (short) vowels:		
Symbol	Example	Transcription
I	Sitz	zIts
E	Gesetz	g@”zEts
a	Satz	zats
O	Trotz	trOts
U	Schutz	SUts
Y	hübsch	hYpS
9	plötzlich	”pl9tslIC

Free (long) vowels and diphthongs:		
Symbol	Example	Transcription
i:	Lied	li:t
e:	Beet	be:t
E:	spt	SpE:t
a:	Tat	ta:t
o:	rot	ro:t
u:	Blut	blu:t
y:	süß	zy:s
2:	blöd	bl2:t
aI	Eis	aIs
aU	Haus	haUs
OY	Kreuz	krOYts

Unstressed ”schwa” vowel:		
Symbol	Example	Transcription
@	bitte	”bIt@
6 (as allophone before r in combination with vowel)	besser	”bEs6

C

Experimental Results for Different Rhythm LID Systems

Rhythm LID system based on the durations of pseudo-syllables

Language	ER in %	
	without ANN	with ANN
DE	75.67	63.42
EN	65.23	59.90
ES	42.99	45.32
FR	76.53	90.67
IT	74.04	77.70
NL	89.21	66.57
PL	80.99	67.74

Table C.1: *Performance of the rhythm LID system using durations of pseudo-syllables as features for the SpeechDat II database with different post-classifiers: language-specific error rates by identification scenario*

Performance measure [%]	without ANN	with ANN	significance
Mean ER	72.09	67.33	significant at 0.001
EER	49.29	34.82	significant at 0.001
C_{avg}	42.06	39.28	significant at 0.010

Table C.2: *Comparison of different performance measures for the rhythm LID system using durations of pseudo-syllables as features: trained and tested on the SpeechDat II database. The last column shows the results of the significance test while comparing the performance of the system without and with ANN*

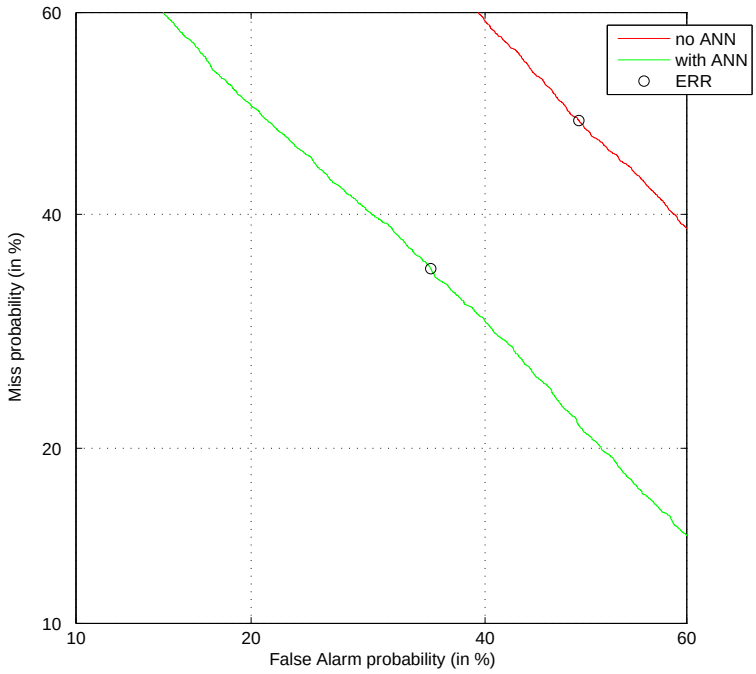


Figure C.1: *DET curves for the rhythm LID system using durations of pseudo-syllables as features: trained and tested on the SpeechDat II database*

Rhythm LID system based on the normalized durations of pseudo-syllables

Language	ER in %	
	without ANN	with ANN
DE	94.30	62.58
EN	63.51	73.15
ES	63.85	77.16
FR	75.47	94.67
IT	89.95	74.04
NL	82.37	67.63
PL	89.83	78.31

Table C.3: *Performance of the rhythm LID system using normalized durations of pseudo-syllables as features for the SpeechDat II database with different post-classifiers: language-specific error rates by identification scenario*

Performance measure [%]	without ANN	with ANN	significance
Mean ER	79.89	75.36	significant at 0.001
EER	49.83	39.23	significant at 0.001
C_{avg}	46.61	43.96	significant at 0.010

Table C.4: *Comparison of different performance measures for the rhythm LID system using normalized durations of pseudo-syllables as features: trained and tested on the SpeechDat II database. The last column shows the results of the significance test while comparing the performance of the system without and with ANN*

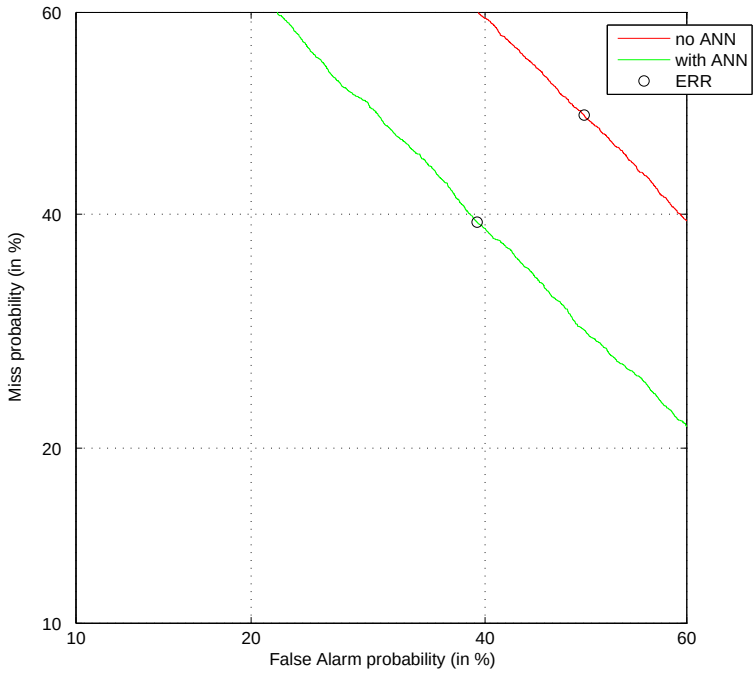


Figure C.2: *DET curves for the rhythm LID system using normalized durations of pseudo-syllables as features: trained and tested on the SpeechDat II database*

Rhythm LID system based on the speech rates computed using durations of pseudo-syllables

Language	ER in %	
	without ANN	with ANN
DE	95.47	62.08
EN	83.65	75.73
ES	52.52	50.18
FR	75.20	91.20
IT	97.26	91.77
NL	95.29	92.86
PL	63.32	53.55

Table C.5: *Performance of the rhythm LID system utilizing speech rates computed using durations of pseudo-syllables for the SpeechDat II database with different post-classifiers: language-specific error rates by identification scenario*

Performance measure [%]	without ANN	with ANN	significance
Mean ER	80.39	73.91	significant at 0.001
EER	49.48	38.92	significant at 0.001
C_{avg}	46.89	43.11	significant at 0.001

Table C.6: *Comparison of different performance measures for the rhythm LID system utilizing speech rates computed using durations of pseudo-syllables: trained and tested on the SpeechDat II database. The last column shows the results of the significance test while comparing the performance of the system without and with ANN*

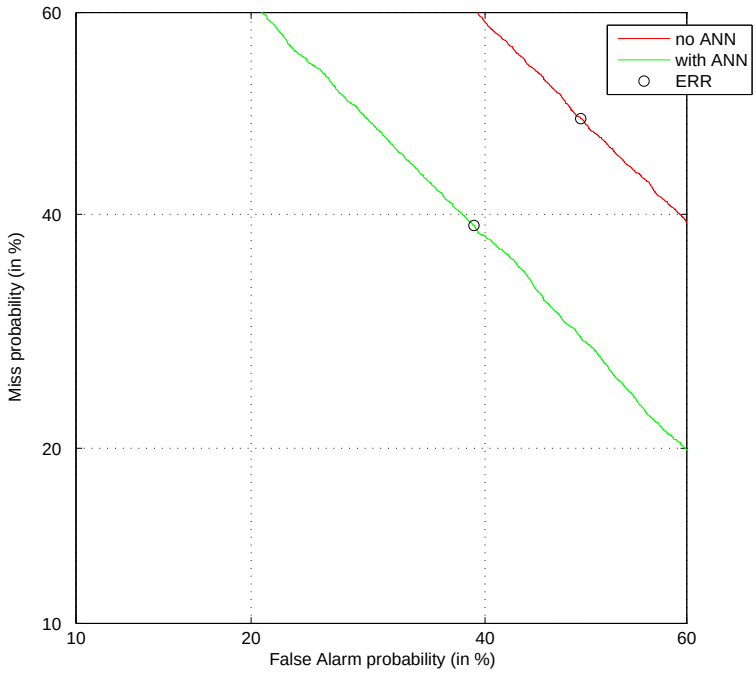


Figure C.3: *DET curves for the rhythm LID system utilizing speech rates computed using durations of pseudo-syllables: trained and tested on the SpeechDat II database*

Rhythm LID system based on durations computed as the intervals between vowels

Language	ER in %	
	without ANN	with ANN
DE	88.93	60.40
EN	62.99	58.86
ES	50.00	57.37
FR	78.40	84.27
IT	80.26	86.47
NL	89.06	81.61
PL	82.73	63.05

Table C.7: *Performance of the rhythm LID system using a pair of intervals between vowels as rhythm feature for the SpeechDat II database with different post-classifiers: language-specific error rates by identification scenario*

Performance measure [%]	without ANN	with ANN	significance
Mean ER	76.05	70.29	significant at 0.001
EER	49.62	37.21	significant at 0.001
C_{avg}	44.36	41.00	significant at 0.002

Table C.8: *Performance of the rhythm LID system using a pair of intervals between vowels as rhythm feature: trained and tested on the SpeechDat II database. The last column shows the results of the significance test while comparing the performance of the system without and with ANN*

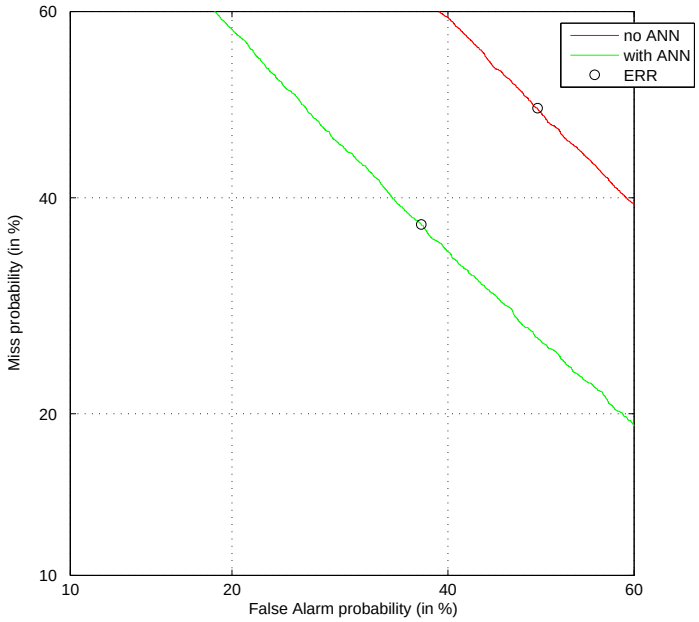


Figure C.4: *DET curves for the rhythm LID system using a pair of intervals between vowels as rhythm feature: trained and tested on the SpeechDat II database*

Rhythm LID system based on normalized durations computed as the intervals between vowels

Language	ER in %	
	without ANN	with ANN
DE	92.62	59.23
EN	79.86	63.68
ES	91.01	99.82
FR	91.47	99.99
IT	88.85	78.24
NL	63.53	74.32
PL	91.30	59.84

Table C.9: *Performance of the rhythm LID system using normalized durations between successive vowels as rhythm feature for the SpeechDat II database with different post-classifiers: language-specific error rates by identification scenario*

Performance measure [%]	without ANN	with ANN	significance
Mean ER	85.52	76.45	significant at 0.001
EER	49.93	42.23	significant at 0.001
C_{avg}	49.89	44.59	significant at 0.001

Table C.10: *Performance of the rhythm LID system using normalized durations between successive vowels as rhythm feature: trained and tested on the SpeechDat II database. The last column shows the results of the significance test while comparing the performance of the system without and with ANN*

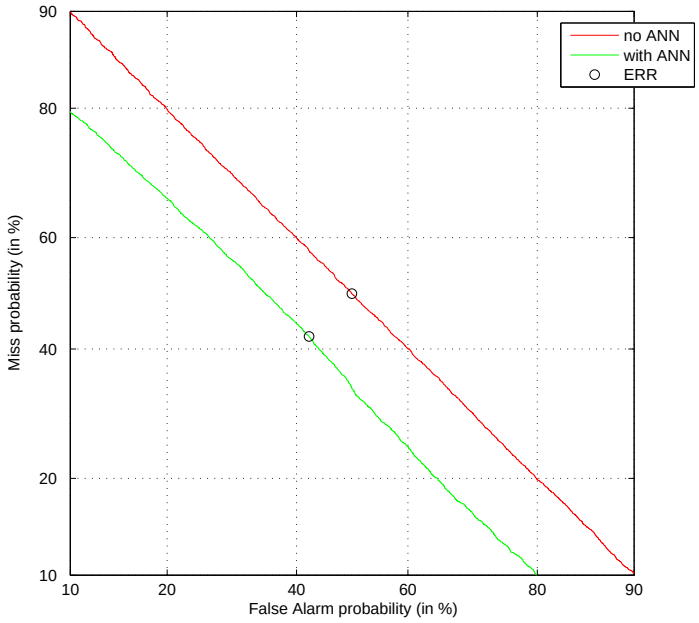


Figure C.5: *DET curves for the rhythm LID system using normalized durations between successive vowels as rhythm feature: trained and tested on the SpeechDat II database*

Rhythm LID system based on the speech rate computed as the intervals between vowels

Language	ER in %	
	without ANN	with ANN
DE	73.99	89.93
EN	92.77	32.87
ES	70.14	59.35
FR	81.33	93.6
IT	62.16	95.43
NL	89.82	98.94
PL	96.12	68.81
Mean ER	80.90	79.99
EER in %	49.89	47.19

Table C.11: *Performance of the rhythm LID system based on the speech rate (computed using syllable-like units defined as intervals between vowels as feature) for the SpeechDat II database with different post-classifiers: language-specific error rates by identification scenario*

Performance measure [%]	without ANN	with ANN	significance
Mean ER	80.90	79.99	not significant
EER	49.23	41.52	significant at 0.001
C_{avg}	47.19	44.91	significant at 0.050

Table C.12: *Performance of the rhythm LID system based on the speech rate (computed using syllable-like units defined as intervals between vowels as feature): trained and tested on the SpeechDat II database. The last column shows the results of the significance test while comparing the performance of the system without and with ANN*

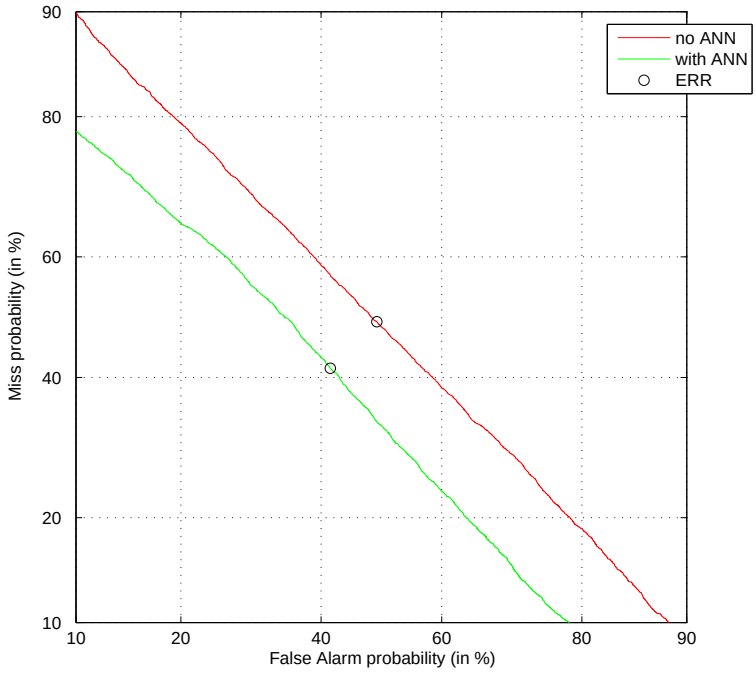


Figure C.6: *DET curves for the rhythm LID system based on speech rate (computed using syllable-like units defined as intervals between vowels as feature: trained and tested on the SpeechDat II database)*

D

DET Curves for Combination of Different LID Systems

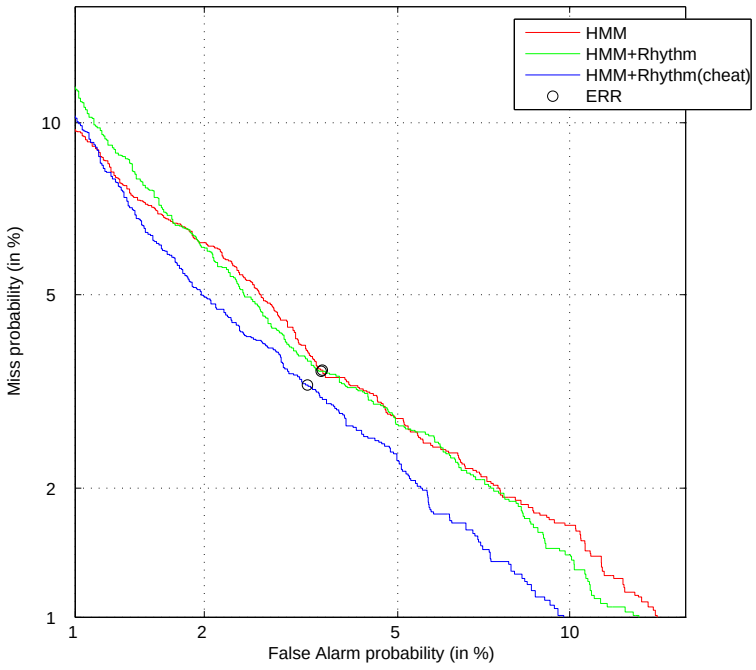


Figure D.1: *DET curves for combination of HMM and rhythm LID systems: trained and tested on the SpeechDat II database*

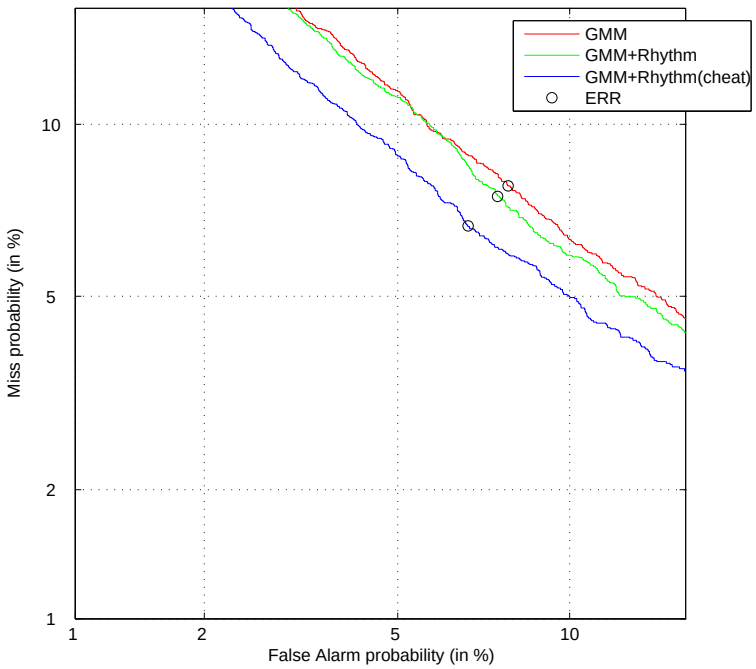


Figure D.2: *DET curves for combination of GMM and rhythm LID systems: trained and tested on the SpeechDat II database*

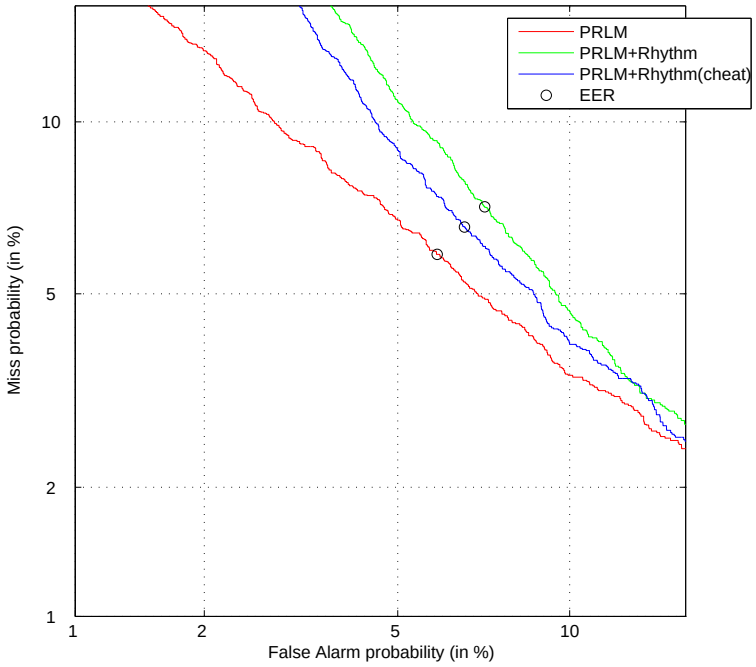


Figure D.3: *DET curves for combination of PRLM and rhythm LID systems: trained and tested on the SpeechDat II database*

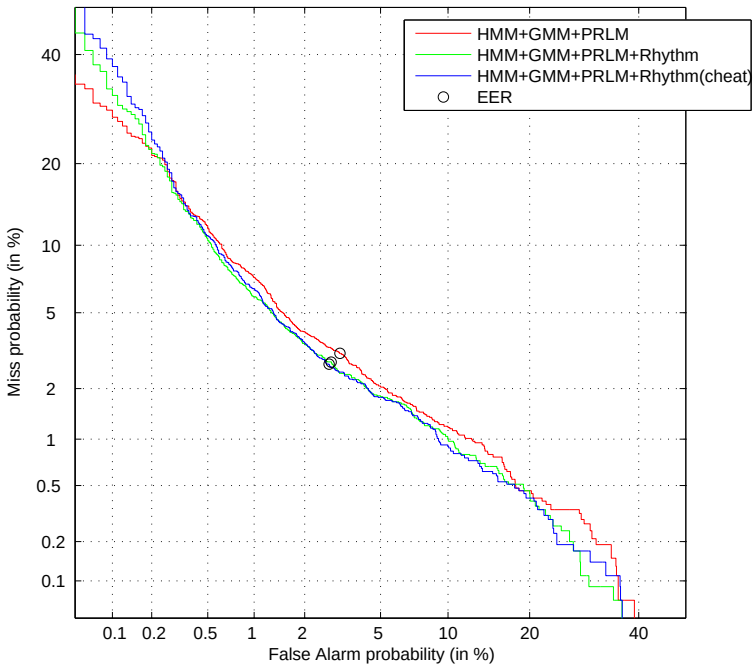


Figure D.4: *DET curves for combination of HMM, GMM, PRLM, and rhythm LID systems: trained and tested on the SpeechDat II database*

Index

Notations

$a_{s,s'}$	Transition probability
$\mathcal{A}$	State transition matrix
$A_{s,s'}$	Transition penalty from state s to s'
b_s	State emission probability
$\mathcal{B}$	State emission matrix
$B_s(\vec{x}_t)$	Emission penalty
$\mathbf{C}$	Set of all possible phoneme sequences
$\mathcal{C} = \{c_1, \ldots, c_K\}$	Phoneme sequence
C	Consonant
$\mathcal{D} = d_1, d_2, \ldots, d_M$	Sequence of the durations of syllable-like events
d_j	Duration of syllable-like event j
$\mathcal{F} = \{\vec{f}_1, \ldots, \vec{f}_T\}$	Sequence of feature vectors representing prosodic information
F_0	Fundamental frequency
g	Score
$\mathcal{L} = \{L_1, \ldots, L_n\}$	Set of languages to be identified
P	Probability
Λ	Set of model parameters
$\vec{\mu}$	Mixture mean vector
n	Number of languages to be identified
π	Set of initial state distributions
q_s	State of an HMM
$\mathcal{Q}$	Set of HMM states
R	Number of observation symbols
r_j	Speech rate on the interval that corresponds to syllable-like event j
$\mathcal{S} = s_1 \ldots s_j \ldots s_M$	Sequence of syllable-like events
s_j	Boundary of syllable-like event j
s, s'	State index
S	Number of states in an HMM
σ	Global variance
Σ	Covariance matrix
T	Length of observation sequence
$\mathcal{V}$	Set of observations
V	Vowel
w	Weight
$\vec{x} = x_1, \ldots, x_D$	Observation vector with dimensionality D
$\mathcal{X} = \{\vec{x}_1, \ldots, \vec{x}_T\}$	Sequence of feature vectors representing spectral information

Bibliography

[1] D. Abercrombie. *Elements of general phonetics*. Edinburgh University Press, Edinburgh, 1967.

[2] A. G. Adami and H. Hermansky. Segmentation of speech for speaker and language recognition. In *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pages 841–844, Geneva, September 2003.

[3] S. Astrov, J. Hofer, and H. Höge. Use of syllable center detection for improved duration modeling in chinese mandarin connected digits recognition. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 1793–1796, Antwerp, 2007.

[4] E. L. Asu and F. Nolan. Estonian and english rhythm: a two-dimensional quantification based on syllables and feet. In *Speech Prosody*, pages 249–253, Dresden, May 2006.

[5] P. Bagshaw. Automatic prosodic analysis for computer aided pronunciation teaching. Phd thesis, University of Edinburgh, UK, 1994.

[6] W. Barry, B. Andreeva, M. Russo, S. Dimitrova, and T. Kostadinova. Do rhythm measures tell us anything about language type? In *Proceedings of the 15th ICPhS*, pages 2693–2696, Barcelona, 2003.

[7] J. G. Bauer. *Diskriminative Methoden zur automatischen Spracherkennung für Telefon–Anwendungen*. Dissertation, Lehrstuhl für Mensch–Maschine–Kommunikation der Technischen Universität München, 2001.

[8] J. G. Bauer, B. Andrassy, and E. Timoshenko. Discriminative optimization of language adapted hmms for a language identification system based on parallel phoneme recognizers. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 166–169, Antwerp, August 2007.

[9] J. G. Bauer and E. Timoshenko. Minimum classification error training of hidden markov models for acoustic language identification. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 405–408, Pittsburgh, Pennsylvania, USA, September 2006.

[10] M. E. Beckman. Segment duration and the 'mora' in japanese. In *Phonetica*, volume 39, pages 113–135.

[11] M. E. Beckman. Evidence for speech rhythms across languages. In Y. Tohura, E. Vatikiotis-Bateson, and Y. Sagisaka, editors, *Speech Perception, Production and Linguistic structure*. Tokyo: Omsha and Amsterdam: IOS Press, 1992.

[12] M. F. BenZeghiba, J.-L. Gauvain, and L. Lamel. Context-dependent phone models and models adaptation for phonotactic language recognition. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 313–316, Brisbane, Australia, September 2008.

[13] P. M. Bertinetto and C. Bertini. On modeling the rhythm of natural languages. In *Speech Prosody*, pages 427–430, Campinas, 2008.

[14] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

[15] D. Bolinger. Pitch accent and sentence rhythm. in forms of english: Accent, morpheme, order. Cambridge MA: Harvard University Press, 1965.

[16] N. Brümmer. Focal multi-class: Toolkit for evaluation, calibration, fusion and decision-making with multi-class statistical pattern recognition scores. `http://niko.brummer.googlepages.com/focalmulticlass`.

[17] L. Burget, P. Matejka, and J. Cernoky. Discriminative training techniques for acoustic language identification. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 20, page 210229, April 2006.

[18] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernoky. Analysis of feature extraction and channel compensation in gmm speaker recognition system. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 15, pages 1979–1986, 2007.

[19] W. M. Campbell. Generalized linear discriminant sequence kernels for speaker recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 161–164, Orlando, Florida, May 2002.

[20] W. M. Campbell, J. P. Campbell, T. P. Gleason, D. A. Reynolds, and W. Shen. Speaker verification using support vector machines and high-level features. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 15, pages 2085–2094, 2007.

[21] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. Torres-Carrasquillo. Support vector machines for speaker and language recognition. In *Computer Speech and Language*, volume 20, pages 210–229, 2006.

[22] W. M. Campbell, T. Gleason, J. Navratil, D. A. Reynolds, W. Shen, E. Singer, and P. Torres-Carrasquillo. Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recognition evaluation. In *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, San Juan, 2006.

[23] W. M. Campbell, R. Richardson, and D. A. Reynolds. Language recognition with word lattices and support vector machines. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 989–992, Honolulu, April 2007.

[24] W. M. Campbell, E. Singer, P. Torres-Carrasquillo, and D. A. Reynolds. Language recognition with support vector machines. In *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, pages 41–44, Toledo, Spain, May 2004.

[25] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 97–100, Toulouse, France, 2006.

[26] D. Caseiro and I. M. Trancoso. Spoken language identification using the SpeechDat corpus. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 3197–3200, Sydney, Australia, December 1998.

[27] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair. Acoustic language identification using fast discriminative training. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 346–349, Antwerp, Belgium, August 2007.

[28] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair. Compensation of nuisance factors for speaker and language recognition. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 15, pages 1969–1978, 2007.

[29] F. Castaldo, E. Dalmasso, P. Laface, D. Colibro, and C. Vair. Politicnico di Torino system for the 2007 nist language recognition evaluation. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 227–230, Brisbane, Australia, September 2008.

[30] J. Cernoky, P. Matejka, L. Burget, and P. Schwarz. Automatic language identification system. In *Special seminar on new technologies in radiocommunications*, pages 1–6, Brno, CZ, UNOB, January 2006.

[31] J. Cohen, T. Kamm, and A. G. Andreou. Vocal tract normalization in speech recognition: Compensating for systematic speaker variability. In *Proc. of the Journal of the Acoustical Society of America*, number 97, pages 31–44, 1995.

[32] E. Couper-Kuhlen. English speech rhythm. form and function in everyday verbal interaction. Amsterdam, 1993. John Benjamins Publishing Company.

[33] D. Crystal. *A dictionary of Linguistics and Phonetics.* Blackwell Publishing Ltd. (5th edition), 2003.

[34] F. Cummins, F. Gers, and J. Schmidhuber. Language identification from prosody without explicit features. In *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pages 371–374, Budapest, Hungary, September 1999.

[35] P. Dai, U. Iurel, and G. Rigol. A novel feature combination approach for spoken document classification with support vector machines. In *Proc. Multimedia Information Retrieval Workshop*, pages 1–5, Toronto, Canada, August 2003.

[36] R. Dasher and D. Bolinger. On pre-accentual lengthening. In *Journal of the International Phonetic Association*, volume 12, pages 58–69, 1982.

[37] R. M. Dauer. Stress-timing and syllable-timing reanalyzed. In *Journal of Phonetics*, volume 11, pages 51–62, 1983.

[38] R. M. Dauer. Phonetic and phonological components of language rhythm. In *Proceedings of the XIth International Congress of Phonetic Sciences*, volume 5, pages 447–450, Tallinn, Estonia, 1987.

[39] V. Dellwo and P. Wagner. Relations between language rhythm and speech rate. In *Proceedings of the 15th ICPhS*, pages 471–474, 2003.

[40] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley & Sons, 2nd edition, 2001.

[41] J. Farinas and F. Pellegrino. Automatic rhythm modeling for language identification. In *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pages 2539–2542, Aalborg, Denmark, September 2001.

[42] J. Farinas, F. Pellegrino, J.-L. Rouas, and R. Andre-Obrecht. Merging segmental and rhythmic features for automatic language identification. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 753–756, Orlando, Florida, April 2002.

[43] J. E. Freund and B. M. Perles. *Modern Elementary Statistics.* Prentice-Hall, 12th edition, 2007.

[44] V. Fromkin, R. Rodman, and N. Hyams. *An Introdustion to Language.* Boston, MA: Thomson Wadsworth, 2007.

[45] J.-L. Gauvain, A. Messaoudi, and H. Schwenk. Language recognition using phone lattices. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 1283–1286, Jeju Island, Korea, October 2004.

[46] O. Glembek, P. Matejka, and T. Mikolov. Advances in phonotactic language recognition. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 743–746, Brisbane, Australia, September 2008.

[47] E. Grabe and E.L. Low. Duration variability in speech and the rhythm class hypothesis. volume 7, pages 515–546. Papers in Laboratory Phonology, 2002.

[48] J. Gutierrez, J.L. Rouas, and R. Andre-Obrecht. Fusing language identification systems using performance confidence indexes. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 385–388, Montreal, Canada, May 2004.

[49] T. J. Hazen and V. W. Zue. Automatic language identification using a segment-based approach. In *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pages 1303–1306, Berlin, Germany, September 1993.

[50] T. J. Hazen and V. W. Zue. Recent improvements in an approach to segment-based automatic language identification. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1883–1886, Adelaide, Australia, April 1994.

[51] D. Hirst. The rhythm of text and the rhythm of utterances: from metrics to models. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 1519–1522, Brighton, UK, September 2009.

[52] D. Hirst and A. Di Cristo. Intonation systems: a survey of twenty languages. Cambridge University Press, Cambridge, 1998.

[53] C. J. Hoequist. Durational correlates of linguistic rhythm categories. In *Phonetica*, volume 40, pages 19–31.

[54] J. Hofer. Syllable center detection based on phoneme recognition. In *AST Workshop*, Maribor, Slowenia, June 2007.

[55] H. Höge. Basic parameters in speech processing. The need for evaluation. In *Archives of Acoustics*, volume 32, 2007.

[56] A. S. House and E. P. Neuburg. Toward automatic identification of the language of an utterance. i. preliminary methodological considerations. In *Proc. of the Journal of Acoustic Society of America*, volume 62(3), pages 708–713, September 1977.

[57] A.-W. Howitt. Automatic syllable detection for vowel landmarks. Phd thesis, Massachusetts Institute of Technology, 2000.

[58] S. Itahashi and D. Liang. Language identification based on speech fundamental frequency. In *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pages 1359–1362, Madrid, Spain, September 1995.

[59] S. Itahashi, J. Zhou, and K. Tanaka. Spoken language discrimination using speech fundamental frequency. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 1899–1902, Yokohama, Japan, September 1994.

[60] A. Lloyd James. Speech signals in telephony. Pitman & Sons, London, 1940.

[61] F. Jelinek. Self-organized language modeling for speech recognition. In A. Weibel and K.-F. Lee, editors, *Readings in speech recognition*, pages 450–506. Morgan Kaufmann Publishers, San Mateo, CA, 1991.

[62] J. Junkawitsch. *Detektion von Schlüsselwörtern in fließender Sprache*. Shaker Verlag, Aachen, 2000.

[63] J. Junkawitsch, G. Ruske, and H. Höge. Efficient methods for detecting keywords in continuous speech. In *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pages 259–262, Rhodes, Greece, September 1997.

[64] P. Kenny and P. Dumouchel. Disentangling speaker and channel effects in speaker verification. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 37–40, Montreal, Canada, May 2004.

[65] H.-K. J. Kuo, E. Fosle-Lussier, H. Jiang, and C.-H. Lee. Discriminative training of language models for speech recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 325–328, Orlando, Florida, May 2002.

[66] P. Ladefoged. *A Course in Phonetics*. Harcourt Brace Jovanovich Inc., New York, 1975.

[67] J. Laver. *Principles of Phonetics*. Cambridge University Press, Cambridge, 1994.

[68] K. A. Lee, C. You, and H. Li. Spoken language recognition using support vector machines with generative front-end. pages 4153–4156, Las Vegas, Nevada, April 2008. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

[69] I. Leiste. Isochrony reconsidered. In *Journal of Phonetics*, volume 5, pages 253–263, 1977.

[70] M. P. Lewis, editor. *Ethnologue: Languages of the World*. Dallas, Tex.: SIL International, 16th edition, 2009.

[71] H. Li, B. Ma, and C.-H. Lee. A vector space modeling approach to spoken language identification. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 15, pages 271–284, 2007.

[72] C.-Y. Lin and H.-C. Wang. Language identification using pitch contour information. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 601–604, Philadelphia, PA, March 2005.

[73] C.-Y. Lin and H.-C. Wang. Fusion of phonotactic and prosodic knowledge for language identification. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 425–428, Pittsburgh, PA, September 2006.

[74] C.-Y. Lin and H.-C. Wang. Language identification using pitch contour information in the ergodic markov model. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 193–196, Toulouse, France, May 2006.

[75] A. Loukina, G. Kochanski, C. Shin, E. Keane, and I. Watson. Rhythm measures with language-independent segmentation. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 1531–1534, Brighton, UK, September 2009.

[76] B. Ma, C. Guan, H. Li, and C.-H. Lee. Multilingual speech recognition with language identification. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 505–508, Denver, Colorado, September 2002.

[77] P. MacCarthy. The pronunciation of French. Oxford University Press, London, 1975.

[78] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, volume 4, pages 1895–1898, Rhodes, Greece, September 1997.

[79] A. F. Martin and A. N. Le. The current state of language recognition: NIST 2005 evaluation results. In *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, pages 1–6, San Juan, June 2006.

[80] L. Mary and B. Yegnanarayana. Extraction and representation of prosodic features for language and speaker recognition. In *Speech Communication*, volume 50, pages 782–796, 2008.

[81] T. Masters. *Signal and Image Processing with Neural Networks*. John Wiley & Sons, New York, 1994.

[82] P. Matejka, L. Burget, O. Glembek, P. Schwarz, V. Hubejka, M. Fapso, T. Mikolov, O. Plchot, and J. Cernoky. BUT language recognition system for NIST 2007 evaluations. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 739–742, Brisbane, Australia, 2008.

[83] P. Matejka, L. Burget, P. Schwarz, and J. Cernoky. BRNO university of technology system for NIST 2005 language recognition evaluation. In *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, San Juan.

[84] OGI multilanguage telephone speech. *http://www.cslu.ogi.edu/corpora/mlts/.*

[85] Y. K. Muthusamy, E. Barnard, and R. A. Cole. Reviewing automatic language identification. In *IEEE Signal Processing Magazine*, volume 11, pages 33–41, 1994.

[86] Y. K. Muthusamy, K. Berkling, T. Arai, R. A. Cole, and E. Barnard. A comparison of approaches to automatic language identification using telephone speech. In *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pages 1307–1310, Geneva, Switzerland, September 1993.

[87] Y. K. Muthusamy, R. A. Cole, and B. Oshika. The OGI multi-language telephone speech corpus. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 895–898, Banff, Alberta, Canada, October 1992.

[88] Y. K. Muthusamy, N. Jan, and R. A. Cole. Perceptual benchmarks for automatic language identification. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 333–336, Adelaide, Australia, April 1994.

[89] T. Nagarajan and H. A. Murthy. Language identification using acoustic log-likelihoods of syllable-like units. In *Speech Communication*, volume 48, pages 913–926, 2006.

[90] S. Nakagawa, T. Seino, and Y. Ueda. Spoken language identification by ergodic hmms and its state sequences. In *Electron. Commun. Japan, Pt. 3*, volume 77, pages 70–79, Februar 1994.

[91] S. Narayanan and D. Wang. Speech rate estimation via temporal correlation and selected sub-band correlation. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 413–416, Philadelphia, PA, March 2005.

[92] J. Navratil. Spoken language recognition - a step towards multilinguality in speech processing. In *IEEE Transactions on Speech and Audio Processing*, volume 9, pages 9(6):678–685, September 2001.

[93] J. Navratil. Recent advances in phonotactic language recognition using binary-decision trees. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 421–424, Pittsburgh, PA, September 2006.

[94] J. Navratil, Q. Jin, W. Andrews, and J. P. Campbell. Phonetic speaker recognition using maximum-likelihood binary-decision tree models. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 796–799, 2003.

[95] T. Nazzi, J. Bertoncini, and J. Mehler. Language discrimination by newborns: Toward an understanding of the role of rhythm. In *Journal of Experimental Psychology: Human Perception and Performance*, volume 24(3), pages 756–766, 1998.

[96] M. Nespor. On the rhythm parameter in phonology. In I. Roca, editor, *In Logical issues in language acquisition*, pages 157–175. Foris Publications, Dordrecht, 1990.

[97] H. Ney and U. Essen. On smoothing techniques for bigram-based natural language modeling. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 825–828, Toronto, Canada, May 1991.

[98] H. Ney, D. Mergel, A. Noll, and A. Paeseler. Continuous speech recognition using a stochastic language model. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 719–722, Glasgow, Scotland, May 1989.

[99] S. Nooteboom. The prosody of speech: melody and rhythm. In W. J. Hardcastle and J. Laver, editors, *The handbook of phonetic science*. Blackwell, Oxford, 1997.

[100] Y. Obuchi and N. Sato. Language identification using phonetic and prosodic hmms with feature normalization. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 569–572, Philadelphia, PA, March 2005.

[101] H. R. Pfitzinger. Phonetische analyse der sprechgeschwindigkeit. Phd thesis, Ludwig-Maximilians-University of Munich, Germany, 2001.

[102] H. R. Pfitzinger, S. Burger, and S. Heid. Syllable detection in read and spontaneous speech. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 1261 – 1264, Philadelphia, PA, October 1996.

[103] K. L. Pike. *The Intonation of American English*. University of Migigan Press, 1945.

[104] B. Plannerer. *Erkennung fließender Sprache mit integrierten Suchmethoden*. Dissertation, Lehrstuhl für Datenverarbeitung der Technischen Universität München, 1995.

[105] R. F. Port, S. Al Ani, and S. Maeda. Temporal compensation and universal phonetics. In *Phonetica*, volume 37, pages 235–252.

[106] R. F. Port, J. Dalby, and M. O'Dell. Evidence for mora-timing in japanese. In *Proc. of the Journal of the Acoustical Society of America*, volume 81, pages 1574–1585.

[107] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of IEEE*, 77(2), pages 257–286, February 1989.

[108] L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Signal Processing Series. Prentice Hall, Englewood Cliffs, NJ, 1993.

[109] F. Ramus. Acoustic correlates of linguistic rhythm: Perspectives. In B. Bel and I. Marlin, editors, *Proceedings of Speech Prosody*, 2002.

[110] F. Ramus, E. Dupoux, and J. Mehler. The psychological reality of rhythm classes: perceptual studies. In *The 15th International Congress of Phonetic Sciences*, pages 337–340, Barcelona, Spain, 2003.

[111] F. Ramus, M. Nespor, and J. Mehler. Correlates of linguistic rhythm in the speech signal. In *Cognition*, volume 73(3), pages 265–292, 1999.

[112] R. Ramus and J. Mehler. Language identification with suprasegmental cues: A study based on speech resynthesis. In *Ptoc. of the Journal of Acoustic Society of America*, volume 5, 1999.

[113] Loquendo ASR Recognizer. http://www.loquendo.com/en/technology/asr.htm.

[114] D. A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pages 963–966, Rhodes, Greece, September 1997.

[115] G. Riccardi, E. Bocchieri, and R. Perraccini. Non deterministic stochastic language models for speech recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 237–240, Detroit, Michigan, May 1995.

[116] F. S. Richardson and W. M. Campbell. Language recognition with discriminative keyword selection. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4145–4148, Las Vegas, Nevada, April 2008.

[117] G. Rigoll. Maximum mutual information neural networks for hybrid connectionist-hmm speech recognition systems. In *IEEE Transactions on Speech and Audio Processing*, volume 2, pages 175–184, January 1994.

[118] G. Rigoll. *Neuronale Netze: eine Einfhrung fr Ingenieure, Informatiker und Naturwissenschaftler*. Expert-Verlag, Renningen-Malmsheim, 1994.

[119] G. Rigoll. *Lecture Script Pattern Recognition*. Lehrstuhl für Mensch–Maschine–Kommunikation der Technischen Universität München, 2006.

[120] P. Roach. On the distinction between "stress-timed" and "syllable-timed" languages. In D. Crystal, editor, *Linguistic controversies*. London: Edward Arnold, 1982.

[121] J.-L. Rouas. Modeling long and short-term prosody for language identification. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 2257–2260, Lisboa, Portugal, September 2005.

[122] J.-L. Rouas. Automatic prosodic variations modeling for language and dialect discrimination. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 15, pages 1904–1911, 2007.

[123] J.-L. Rouas, J. Farinas, and F. Pellegrino. Automatic modeling of rhythm and intonation for language identification. In *Proc. International Congress of Phonetic Sciences*, pages 567–570, Barcelona, Spain, 2003.

[124] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. Andre-Obrecht. Modeling prosody for language identification on read and spontaneous speech. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 40–43, Hong Kong, April 2003.

[125] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. Andre-Obrecht. Rhythmic unit extraction and modelling for automatic language identification. In *Speech Communication*, volume 47, pages 436–456, 2005.

[126] B. Schuller, M. Ablaßmeier, R. Müller, S. Reifinger, T. Poitschke, and G. Rigoll. Chapter: Speech communication and multimodal interfaces. In K.-F. Kraiss, editor, *Advanced Man-Machine Interaction*, pages 141–190. Springer Verlag Berlin Heidelberg New York, 2006.

[127] T. Schultz and Ed. K. Kirchhoff. Multilingual speech processing. Academic Press, Inc., 2006.

[128] P. Schwarz, P. Matejka, and J. Cernoky. Hierarchical structures of neural networks for phoneme recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[129] E. Singer, P. Torres-Carrasquillo, T. Gleason, W. Campbell, and D.A. Reynolds. Acoustic, phonetic and discriminative approaches to automatic language identification. In *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pages 1345–1348, Geneva, Switzerland, September 2003.

[130] SpeechDat Internet Site. *http://www.speechdat.org*.

[131] A. Solomonoff, C. Quillen, and W. M. Campbell. Channel compensation for SVM speaker recognition. In *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, pages 57–62, Toledo, Spain, 2004.

[132] A. E. Thyme-Gobbel and S. E. Hutchins. On using prosodic cues in automatic language identification. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 1768–1771, Philadelphia, PA, October 1996.

[133] E. Timoshenko. Classifiers for spoken language identification with varying amounts of knowledge sources. Masters thesis, TU Dresden, Germany, 2005.

[134] E. Timoshenko and J. G. Bauer. Unsupervised adaptation for acoustic language identification. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 409–412, Pittsburgh, Pennsylvania, USA, September 2006.

[135] E. Timoshenko and H. Höge. Using speech rhythm for acoustic language identification. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 182–185, Antwerp, Belgium, August 2007.

[136] R. Tong, B. Ma, D. Zhu, H. Li, and E. S. Chng. Integrating acoustic, prosodic and phonotactic features for spoken language identification. pages 205–208, Toulouse, France, May 2006. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

[137] P. Torres-Carrasquillo, E. Singer, W. Campbell, T. Gleason, A. McCree, D. A. Reynolds, F. Richardson, W. Shen, and D. Sturim. The MITLL NIST LRE 2007 language recognition system. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 719–722, Brisbane, Australia, September 2008.

[138] P. Torres-Carrasquillo, E. Singer, M.A. Kohler, R. Green, D.A. Reynolds, and J.R. Deller Jr. Approaches to language identification using gaussian mixture models and shifted delta cepstral features. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, pages 719–722, Denver, Colorado, September 2002.

[139] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface. Channel factors compensation in model and feature domain for speaker recognition. In *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, pages 1–6, San Juan.

[140] P. Wagner. Visualization of speech rhythm. In *Archives of Acoustics*, volume 32, 2007.

[141] P. Wagner and V. Dellwo. Inrtoducing YARD (yet another rhythm determination) and re-introducing isochronny to rhythm research. In *Proceedings of the Second International Conference on Speech Prosody*, pages 227–230, Nara, Japan, March 2004.

[142] A. Waibel, P. Geutner, L.M. Tomokiyo, T. Schultz, and M. Woszczyna. Multilinguality in speech and spoken language systems. In *Proc. of the IEEE*, volume 88, pages 1181–1190, 2000.

[143] C. White, I. Shafran, and J.-L. Guavain. Discriminative classifiers for language recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 213–216, Toulouse, France, May 2006.

[144] M. Wöllmer, F. Eyben, B. Schuller, S. Steidl, and G. Rigoll. Recognition of spontaneous conversational speech using long short-term memory phoneme predictions. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, page 19461949, Makuhari, Japan, September 2010.

[145] E. Wong and S. Sridharan. Fusion of output scores on language identification system. In *Workshop on Multilingual Speech and Language Processing*, Aalborg, Denmark, 2001.

[146] E. Wong and S. Sridharan. Methods to improve gaussian mixture model based language identification system. In *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, page Fusion of output scores on language identification system, Denver, Colorado, September 2002.

[147] Y. Yan and E.Barnard. An approach to automatic language identification based on language-dependent phone recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3511–3514, Detroit, Michigan, May 1995.

[148] B. Yin, E. Ambikairajah, and F. Chen. Language-dependent fusion for language identification. In *Proc. of thee 11th Australian International Conference on Speech Science and Technology*, pages 52–57, Auckland, Australia, 2006.

[149] B. Yin, E. Ambikairajah, and F. Chen. Hierarchical language identification based on automatic language clustering. pages 178–181, Antwerp, Belgium, August 2007. Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP).

[150] M. A. Zissman. Automatic language identification using gaussian mixture and hidden markov models. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 399–402, Minneapolis, Minnesota, April 1993.

[151] M. A. Zissman. Comparison of four approaches to automatic language identification of telephone speech. In *IEEE Transactions on Speech and Audio Processing*, volume 4, pages 31–44, January 1996.

[152] M. A. Zissman and E. Singer. Automatic language identification of telephone speech messages using phone recognition and n-gram modeling. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 305–308, Adelaide, Australia, April 1994.

[153] V. W. Zue and J. R. Grass. Conversational interfaces: Advances and challenges. In *Proc. of the IEEE*, volume 88, pages 1166–1180, 2000.