Multilingual Articulatory Features

Interactive Systems Lab

Carnegie Mellon University, Pittsburgh, PA, USA
Universität Fridericiana zu Karlsruhe (TH), Germany

Diplomarbeit

Sebastian Stüker

Supervisors:
Dipl. Phys. Florian Metze
Dr. Tanja Schultz
Prof. Dr. Alex Waibel

April 2003

Hiermit erkläre ich, dass ich diese Diplomarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, den 14.04.2003

Sebastian Stüker

Abstract

In large vocabulary continuous speech recognition human speech is usually modelled as a sequence of phonemes or sub-phonemic units. Sometimes this model is called 'beads-on-a-string'. However, recent research indicates that phonemes are too coarse a model to capture the richness of human speech at the acoustic level which is necessary to deal with variability due to spontaneous speech, noise, reverberation and intra speaker variations. The use of monolingual articulatory features, such as place and manner of articulation, has been shown to improve the performance of speech recognition systems under different conditions and in different settings, especially when combining standard phoneme models with detectors for articulatory features. Models of articulatory features are more robust to noise and reverberation than phoneme models.

In this work I show that articulatory features are also robust to inter language variability. Using a global set of features derived from the GlobalPhone global unit set and the mapping of features to phonemes introduced by the International Phonetic Association, I trained binary monolingual and multilingual detectors for abstract feature classes. These detectors are able to detect articulatory features across languages. By pooling detectors from many languages it is also possible to achieve a better classification performance than with feature detectors from only one language.

By applying a flexible stream architecture that has been successfully used to combine detectors for articulatory features with phoneme based standard models I support an English and a Chinese HMM based speech recognition system with multilingual and crosslingual feature detectors. In doing so I compare two methods for the necessary selection of stream weights. In a first experiment I use a heuristic based on the classification accuracy of the feature detectors for selecting stream weights to show the potential in reducing the word error rate with cross- and multilingual feature detectors. For a second set of experiments I implemented a discriminative training method called 'Discriminative Model Combination' (DMC) to obtain more suitable sets of stream weights. Using DMC and crosslingual feature detectors from many languages I achieve a reduction in word error rate that matches the reductions when using monolingual feature detectors. On English I am able to reduce the word error rate by 12.4% relative.

Zusammenfassung

Bei der maschinellen Erkennung kontinuierlicher Sprache mit grossem Vokabular wird menschliche Sprache meistens als Folge von Phonemen oder subphonemischen Einheiten modelliert. Da hier die Phoneme zeitlich wie an einer Perlenschnur aufgereiht sind, wird manchmal der englische Fachausdruck 'beads-on-a-string' [Ost99] verwendet. In den letzten Jahren jedoch hat die Forschung auf dem Gebiet der akkustischen Modellierung [KFS00] [Kir98] [Eid01] [CGW01] [WGC01] [DS94] gezeigt, dass Phoneme ein zu grobes Modell sind, um die Reichhaltigkeit der Informationen des akkustischen Signals von menschlicher Sprache zu repräsentieren. Dies ist aber notwendig, um die Variabilität des akkustischen Signals, die bedingt ist durch spontansprachliche Effekte, Rauschen oder Hall, mit in das Modell einzubeziehen. Unter guten Bedingungen, wie z.B. kein Rauschen, Nahbesprechungsmikrofon und geplanter Sprache, erreichen moderne Spracherkennungssysteme akzeptable Wortfehlerraten. Für ungünstige Bedingungen, wie z.B. Rauschen oder Spontansprache, wie man sie in Anwendungen des Alltags häufig vorfindet, fällt die Erkennungsleistung dieser Systeme stark ab und bedarf dringend der Verbesserung.

Die Verwendung von monolingualen artikulatorischen Merkmalen, wie Art und Ort der Artikulation, hat zu deutlichen Verbesserungen der Erkennungsleistungen von maschinellen Spracherkennungssystemen geführt, insbesondere wenn standard Phonemmodelle mit Modellen für artikulatorische Merkmale kombiniert werden. Modelle für artikulatorische Merkmale sind robuster gegenüber Rauschen und Hall als Modelle für Phoneme.

In dieser Arbeit weise dich die Robustheit von artikulatorischen Merkmalen gegenüber Intersprachenvariabilitäten nach. Dafür definiere ich einen sprachenunabhängigen Satz von artikulatorischen Merkmalen, basierend auf dem GlobalPhone Phonemsatz und der von der International Phonetic Association (IPA) definierten Abbildung von artikulatorischen Merkmalsbündeln nach Phonemen. Dabei bin ich nicht daran interessiert, die numerisch exakten Werte bestimmter artikulatorischen Merkmale zu erkennen, z.B. die horizontale Position des Dorsum, sondern klassifiziere nach binären, abstrakten Klassen, die auf dem Konzept des artikulatorischen Ziels (im Englischen: artiuclatory target) von IPA basieren. Als Grundlage meiner Experimente habe ich Detektoren für diese artikulatorischen Merkmale für die fünf Sprachen Chinesisch, Deutsch, Englisch, Japanisch und Spanisch trainiert. Anhand der Klassifikationsgenauigkeit dieser Detektoren auf den Testmengen aller fünf Sprachen belege ich, dass artikulatorische Merkmale für eine Vielzahl von Sprachen erkannt werden können und insbesondere auch über Sprachengrenzen hinweg. Durch die Kombination von Detektoren aus mehreren Sprachen zeige ich, wie die Klassifikationsgenauigkeit gegenüber Detektoren, die nur auf der Testsprache trainiert wurden, verbessert werden kann. Ferner habe ich mit Hilfe des Verfahrens 'ML-Mix' [SW01] zur sprachenunabhängigen akkustischen Modellierung alle möglichen Kombinationen multilingualer Detektorensätze auf den fünf Sprachen trainiert.

Als nächsten Schritt integriere ich die trainierten Detektoren in Phonem basierte HMM Spracherkenner unter Verwendung einer flexiblen Architektur, die mit parallele Datenströmen arbeitet [MW02]. Ich vergleiche die Erkennungsleistungen eines englischen und

eines chinesischen Erkenners, wenn ich sie mit mono-, cross- und multilingualen artiuklatorischen Merkmalsdetektoren unterstütze. Bei der für den Modellkombinationsansatz notwendigen Auswahl von Datenstromgewichten vergleiche ich eine Heuristik mit dem diskriminativen Trainingsansatz 'Discriminative Model Combination' (DMC) [Bey00].

Die Anwendung der Heuristik zeigt die Machbarkeit der Verbesserung der Erkennungsleistung durch Kombination von Standardmodellen und Detektoren für artikulatorische Merkmale. Mit Hilfe von DMC kann zwar keine weitere Verbesserung für die Kombination mit monolingualen Detektoren gegenüber den Gewinnen der Heuristik erreicht werden. Allerdings habe ich eine weitere Verbesserung für die Kombination mit cross- und multilingualen Detektoren gegenüber der Anwendung der Heuristik erreicht, so dass die Reduktion der Fehlerraten durch die Hinzunahme der cross- und multilingualen Detektoren der Fehlerreduktion durch monolinguale Detektoren gleichkommt. Für Englisch habe ich die Wortfehlerraten um 12,4% relativ reduziert.

Acknowledgements

During the course of this research project I received support from many people. First I would like to thank Prof. Alex Waibel for giving me the opportunity to conduct my research at the Interactive Systems Lab at Carnegie Mellon University in Pittsburgh. Further my thanks go to Dr. Tanja Schultz and Florian Metze for their constant advice, guidance and support without which this project would not have been possible. I would also like to thank Hagen Soltau and Christian Fügen for their help with the Janus Recognition Toolkit. Further my thanks go to Michael Bett, Rob Malkin, and Victoria MacLaren for their help with the administration and technical infrastructure at CMU. Last but not least I would like to thank my parents for their continuous and ever present support during the years of my studies.

Contents

1 Introduction **1**
- 1.1 Automatic Speech Recognition . . . 2
  - 1.1.1 Acoustic Modelling with Articulatory Features . . . 2
- 1.2 The JANUS Recognition Toolkit . . . 3
- 1.3 Objective and Contributions of this Work . . . 4

2 Multilingual Speech Recognition **5**
- 2.1 The GlobalPhone Project . . . 5
- 2.2 Overview of the Database . . . 6
- 2.3 A Global Phoneme Set . . . 6
- 2.4 Share Factor . . . 7
- 2.5 Language Independent Acoustic Modelling . . . 9
  - 2.5.1 Mono-, Cross-, and Multilingual . . . 10
  - 2.5.2 ML-mix . . . 10

3 Articulatory Features **11**
- 3.1 Human Speech Production . . . 11
- 3.2 Phonetic Description of Speech . . . 12
  - 3.2.1 Consonants, Vowels, and Features . . . 12
- 3.3 Articulatory Features for ASR . . . 14
  - 3.3.1 The Conventional Approach . . . 15
  - 3.3.2 Advantages from Feature Detectors . . . 16
- 3.4 A Language Independent Set of Articulatory Features . . . 17
  - 3.4.1 Share Factor for AF . . . 17

4 Training Articulatory Feature Detectors **21**
- 4.1 Monolingual Detectors in Five Languages . . . 21
  - 4.1.1 Feature Extraction . . . 22
  - 4.1.2 Training . . . 22
  - 4.1.3 Evaluation . . . 22
- 4.2 Crosslingual AF Detection . . . 23
  - 4.2.1 Crosslingual Combination of AF Detectors . . . 24
- 4.3 Multilingual Classification . . . 28

4.3.1 Training . . . 28
4.3.2 Evaluation . . . 28

5 Stream Architecture for Including AF into ASR **31**
5.1 A Flexible Stream Architecture . . . 32
5.2 Adapting Stream Weights . . . 33
5.2.1 Educated Guess . . . 33
5.2.2 Weight Selection with DMC . . . 34

6 Decoding Experiments **37**
6.1 Initial Experiments . . . 37
6.2 Decoding without Articulatory Features . . . 38
6.2.1 Training . . . 38
6.2.2 Evaluation . . . 39
6.3 Decoding using AF streams and heuristic stream weights . . . 39
6.3.1 Monolingual case . . . 41
6.3.2 Crosslingual case . . . 42
6.3.3 Multilingual case . . . 43
6.3.4 Generalization . . . 43
6.3.5 Conclusion . . . 44
6.4 Decoding using AF and adapted stream weights . . . 44
6.4.1 Monolingual case . . . 46
6.4.2 Crosslingual case . . . 46
6.4.3 Multilingual case . . . 47
6.4.4 Complete Detector Set . . . 47
6.4.5 Learned Feature Weights . . . 47
6.4.6 Conclusion . . . 51

7 Conclusion **53**
7.1 Summary . . . 53
7.2 Future Work . . . 53

A Results of the Monolingual AF Detectors **55**
A.1 Chinese AF Detectors . . . 55
A.2 English AF Detectors . . . 56
A.3 German AF Detectors . . . 57
A.4 Japanese AF Detectors . . . 58
A.5 Spanish AF Detectors . . . 59

B Results of the Multilingual AF Detectors **61**
B.1 MM2 Detectors . . . 61
B.2 MM3 Detectors . . . 62
B.3 MM4 Detectors . . . 62

B.4 MM5 Detectors 62

List of Tables

2.1 Overview of the data used from the GlobalPhone corpus 7
2.2 Size of the GlobalPhone dictionaries 7
2.3 Overview of the Global Phoneme Set [SW01] 8

3.1 Table of the global feature set and the languages in which the features occur . . 19

4.1 Classification Accuracy of the AF detectors 23
4.2 Average classification accuracy of the AF detectors 24
4.3 Results of the crosslingual combination of feature detectors 25
4.4 Comparison between MM4 detectors that were not trained on the test language and MM5 detectors 29

6.1 Results of the BN+ESST acoustic on the GP EN development and evaluation set with standard language model parameters 38
6.2 Word error rates of the English and German baseline systems on their development and evaluation sets 39
6.3 WER when decoding the EN development set using AF detectors as additional knowledge source in a monolingual, crosslingual, and multilingual scenario. 40
6.4 WER when decoding the CH development set using AF detectors as additional knowledge source in a monolingual, crosslingual, and multilingual scenario. 40
6.5 WER when decoding the EN evaluation set using AF detectors as additional knowledge source in a monolingual, crosslingual, and multilingual scenario. 41
6.6 WER when decoding the CH evaluation set using AF detectors as additional knowledge source in a monolingual, crosslingual, and multilingual scenario. 41
6.7 WER when decoding the EN development set using AF detectors as additional knowledge source in a monolingual, crosslingual, and multilingual scenario and DMC adapted weights 45
6.8 WER when decoding the CH development set using AF detectors as additional knowledge source in a monolingual, crosslingual, and multilingual scenario and DMC adapted weights 45
6.9 Feature weighting as learned by the DMC on English 48

6.10 Feature selection and weighting as learned by the DMC on English when using the feature detectors from all languages 49
6.11 Feature weighting as learned by the DMC on Chinese 50

List of Figures

1.1 The components of a speech recognition system that makes additional use of articulatory features . . . 3

2.1 Average share factor and range for all possible subsets of the GlobalPhone languages [SW01] . . . 9
2.2 Language mixed acoustic modelling vs. language dependent [SW01] . . . 10

3.1 The consonant table from the IPA chart [Ass99] . . . 12
3.2 The vowel quadrilateral from the IPA chart [Ass99] . . . 13
3.3 Mid-sagittal plane of the human head [Ell97] . . . 14
3.4 Progress (word error rates) in speech recognition over the years[Rog01] . . . 15
3.5 Average Share Factor for the Five Selected Languages . . . 18

4.1 Results of the crosslingual evaluation of all feature detectors on the CH, EN, GE, and JA test set . . . 26
4.2 Results of the crosslingual evaluation of all feature detectors on the SP test set . . . 27
4.3 Performance overview of the MMn recognizers . . . 30

5.1 Stream setup with combined 'feature absent' and 'feature present' detectors . . . 32

Chapter 1

Introduction

Over the last decades the field of speech recognition has seen enormous progress. Speech recognition systems have been built for a wide variety of tasks such as spelling, single words, or continuous speech. Systems exist that deal with planned, read, or spontaneous speech as well as different kinds of environments and channels, e.g. close talking microphones and high fidelity recordings or noisy environments and telephone lines. Speech recognizers can work speaker independently or, if called for, can be tailored to a specific speaker. Speech recognition has also been studied in situations that involve one or many speakers, e.g. a meeting room scenario. Quite a number of applications for real life have resulted from this research. Different kinds of speech recognizers are now commercially available either as software packages for use with a personal computer or embedded in appliances such as cellular phones or car navigation systems.

However today's speech recognition systems are far from being perfect. Though speech recognition can be reliably done under very favorable conditions — e.g. quiet environment, close talking microphone, high fidelity sound recording, and planned speech — many tasks that relate to real life scenarios exist where the performance of the recognition systems is in clear need of improvement.

Many of today's recognition systems view human speech as a list of phonemes and consequently model it as such. However problems when dealing with adverse conditions, e.g. modelling pronunciation variants in spontaneous speech, have lead many researchers to the conclusion that phonemes are too coarse a model of human speech (see chapter 3). Therefore a change in paradigm is necessary. Recent research in the field of acoustic modelling indicates that speech is better modelled in terms of *articulatory features* (AF) than phonemes.

In this chapter I will give a very brief introduction to the general task of automatic speech recognition. I will further introduce the Janus Recognition Toolkit with which the experiments for this work have been performed. After having described the problems I am facing I will define the objective and motivation of my research.

1.1 Automatic Speech Recognition

The goal of *automatic speech recognition* (ASR) is to enable a machine, in my case a computer, to extract the words spoken by one or more persons from the resulting sound wave, which consists of a change in air pressure caused by the human speech production process. For this purpose the sound wave is usually recorded using a microphone and then digitalized with the use of electronic equipment. This recording process results in a digital representation of the wave form of the sound wave over time. The wave form is then transformed further by the preprocessing unit of the speech recognition system into a sequence of so called feature vectors.

It is then the task of the decoder to find the sequence of words W that yields the highest probability $P(W|X)$ given the observed sequence of feature vectors X and the internal model of the recognition system.

With the use of Bayes rule the calculation of this probability can be further decomposed into what is known as the *fundamental equation of speech recognition* .

$$P(W|X) = \frac{p(X|W) * P(W)}{p(X)} \tag{1.1}$$

$p(X)$ is the prior probability to observe the sequence of feature vectors X. $p(X|W)$ is the probability that, given the sequence of words W, the feature vectors X are observed. This part of the equation is commonly called the *acoustic model*. $P(W)$ is the prior probability of observing W independently of the feature vector X and is usually called the *language model*.

Thus the decoder now tries to find:

$$\begin{aligned} \widehat{W} &= \underset{W}{argmax}\, P(W|X) \\ &= \underset{W}{argmax}\, \frac{p(X|W) * P(W)}{p(X)} \\ &= \underset{W}{argmax}\, p(X|W) * P(W) \end{aligned} \tag{1.2}$$

Usually the search space is limited by a dictionary that defines the set of allowed words of which W can be composed. Figure 1.1 gives a schematic overview of the resulting speech recognition system. Note that the acoustic model in this picture is already depicted as one that makes additional use of articulatory features.

1.1.1 Acoustic Modelling with Articulatory Features

Traditionally continuous speech recognition models speech as a sequence of phonemes or sub-phonemic units. Over time it has become clear that this is only a coarse model of the human speech production process. Thus the goal of recent research has been to model

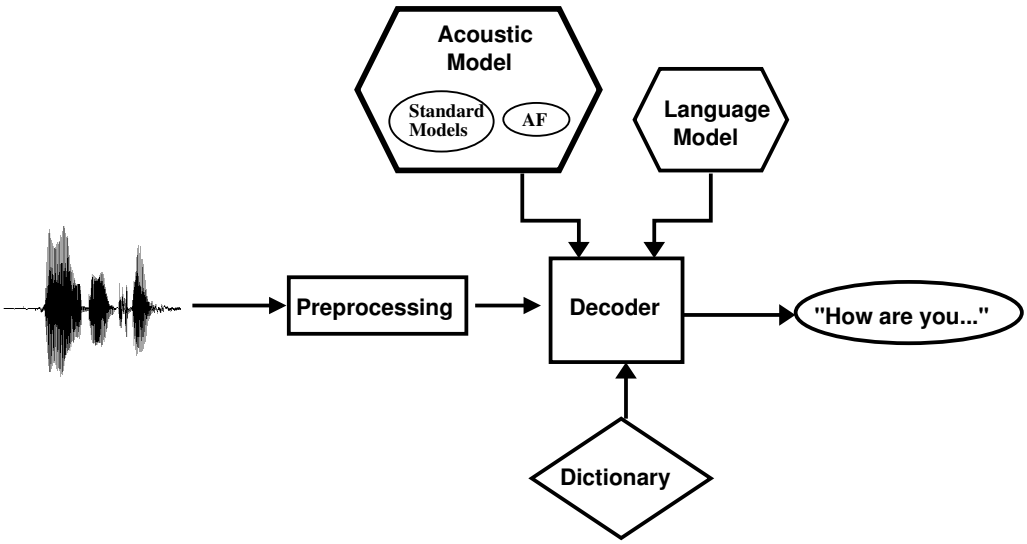


Figure 1.1: The components of a speech recognition system that makes additional use of articulatory features

speech with the help of articulatory features that describe the configuration of the human vocal tract.

In my research the articulatory features that I work with are the attributes that are used by the International Phonetic Association (IPA) to describe the way the human sounds are articulated. I am not concerned with the exact position of the articulators during the speech process but rather use abstract classes that focus on the main aspect of the articulators (e.g. whether lips are rounded or not, but not the actual degree of rounding).

In this work I am also not concerned with building a recognition system that is solely based on articulatory features but rather with supporting a recognizer based on standard sub-phonetic acoustic units with detectors for articulatory features. So the acoustic model for the AF enhanced recognition system will consist of phonemic models and models for articulatory features as shown in figure 1.1.

1.2 The JANUS Recognition Toolkit

The experiments for this research project were performed with the JANUS Recognition Toolkit (JRTk). The JRTk has been developed by the Interactive Systems Labs at Karlsruhe University and Carnegie Mellon University [FGH+97]. It is part of the JANUS speech-to-speech translation system [LWL+97].

The JRTk provides a flexible Tcl/Tk script based environment which enables researchers to build state-of-the-art speech recognizers and allows them to develop, implement, and evaluate new methods. It implements an object oriented approach that unlike other toolkits is not a set of libraries and precompiled modules but a programmable shell with transpar-

ent, yet efficient objects.

I used JRTk Version 5 which features the IBIS decoder [SMFW01]. The IBIS decoder is a one-pass decoder that is based on a re-entrant single pronunciation prefix tree and makes use of the concept of linguistic context polymorphism. It is therefore able to incorporate full linguistic knowledge at an early stage. It is possible to decode in one pass, using the same engine in combination with a statistical n-gram language model as well as context-free grammars. It is also possible to use the decoder to rescore lattices in a very efficient way. This results in a speed up compared to the decoder in previous versions of the JRTk which needed three passes to incorporate full linguistic knowledge.

1.3 Objective and Contributions of this Work

It has been shown in the monolingual case that articulatory feature detectors can be used as an additional knowledge source to support a conventional HMM recognizer, improving its performance significantly. Experiments demonstrate that articulatory features are more robust to variability due to noise and reverberation (see chapter 3).

In this work I first examine the robustness of articulatory features towards cross language variabilities. By building feature detectors in five languages I show that it is possible to detect articulatory features for a variety of languages and across languages. I further show that by combining feature detectors from different languages it is possible to improve the classification accuracy of feature detection on a given language, compared to when using only feature detectors that have been trained on the test language.

For the purpose of determining the potential of building language independent feature detectors I evaluate all possible combinations of feature detectors trained on two to five languages from a selected pool of five languages.

I further examine the potential of aiding speech recognition systems based on sub-phonemic standard models with the trained cross- and multilingual feature detectors. To do so I use the stream based approach to combine standard sub-phonemic acoustic models with articulatory feature detectors that has been developed by Metze [MW02], and do the necessary optimization of stream weights by applying a discriminative training technique called 'Discriminative Model Combination' (DMC) developed by Peyer Beyerlein [Bey00].

Using DMC I can show improvements for the crosslingual and multilingual scenarios compared to when stream weights are selected by the heuristic that has been used so far to select stream weights. The results can be found in chapter 6.

Chapter 2

Multilingual Speech Recognition

In this chapter I will give a short introduction into the field of multilingual speech recognition. The different sections cover core topics and introduce the essential terms and concepts that I use in my multilingual research. The concepts described here mainly apply to speech recognition based on phonemes as acoustic units. In chapter 3 I will show how the terminology and concepts can be extended for the use with articulatory features.

2.1 The GlobalPhone Project

Most of the experiments in this work were performed on the GlobalPhone corpus. The GlobalPhone corpus has been collected during the course of the GlobalPhone project [SWW97]. The purpose of this corpus is to support multilingual speech recognition research. In order to be able to focus on the differences among languages, uniformity of the data needed to be ensured so that these differences would not be superposed by mismatched conditions, e.g. in acoustic quality, between the different languages. Therefore the database had to fulfill the following requirements:

- The languages that are most important to speech recognition according to the number of their speakers and their economic and political relevance are covered.
- As many of the phones that are used by humans to communicate as possible are covered.
- The speakers are representative for the native speakers of their language. That includes attributes such as gender, and age.
- The transcribed material is large enough to train robust acoustic models.
- Large additional texts with millions of words are available for calculating statistical language models.
- The acoustic quality of the material is uniform so that language specific differences can be extracted from the results obtained from experiments.

- For all languages the same type of speech is collected (e.g. spontaneous, read or colloquial as a monologue or dialog).
- The data for all languages are similar with respect to their semantics.

Since the most time consuming and expensive task in building a database is the transcription of speech, only data read from electronically available texts were acquired. Therefore texts from international newspapers available on the World Wide Web with national, international political and economic topics were collected. This process has the advantage that it is possible to achieve a reasonable vocabulary coverage with acceptable OOV-rates. The database also contains cross lingual words such as proper names, products, and borrowings.

At the time that this report had been written the GlobalPhone corpus contained the following languages: Arabic(AR), Chinese(CH), Czech(CZ), German (GE), French (FR), Japanese (JA), Korean (KO), Croatian (CR), Portuguese (PO), Russian (RU), Spanish (SP), Swedish (SW), Tamil(TA), Turkish (TU). We further use the Wall Street Journal corpus *WSJ0* for our research in English (EN). Since the GlobalPhone corpus is modelled after the WSJ0 corpus this does not introduce a mismatch in conditions. For reasons of simplicity we will refer to these fifteen languages as the *GlobalPhone languages* throughout this work even though WSJ0 is not part of the GlobalPhone corpus.

2.2 Overview of the Database

For our experiments we used the five languages Chinese, English, German, Japanese, and Spanish. I selected these languages because they display a variety of different characteristics such as the set of sounds and features that they cover, or traits such as tonality. Table 2.1 gives an overview of the size of the training, development and evaluation sets for these five languages.

Table 2.2 shows the size of the dictionaries. Every word in our dictionary is tagged with the language it belongs to, so that it can be distinguished from words in other languages that might share the same orthography.

2.3 A Global Phoneme Set

The multilingual paradigm described in this chapter is based on the assumption that the articulatory representation of phonemes across different languages is so similar that phonemes can be seen as units independent of the underlying language. Thus the language specific sets of phonemes Υ_{L_i} of languages $L_i (i = 1 \ldots n)$ can be combined into a single language independent phoneme set $\Upsilon = \Upsilon_{L_1} \cup \Upsilon_{L_2} \cup \ldots \cup \Upsilon_{L_N}$. This concept had first been proposed by the International Phonetic Association (IPA) [Ass99]. Different international schemes for sharing phonemes across languages exist, such as Sampa [Wel89] or Worldbet [Hie93].

	#utterances (hours)				
Language	CH	EN	GE	JA	SP
Training	8,663 (26.9)	7,137 (15.0)	9,259 (16.9)	9,234 (23.9)	5,426 (17.6)
Development	250 (0.7)	144 (0.4)	199 (0.4)	250 (0.7)	250 (0.7)
Evaluation	240 (0.7)	152 (0.4)	250 (0.4)	250 (0.7)	250 (0.7)

Table 2.1: Overview of the data used from the GlobalPhone corpus

	#words				
Language	CH	EN	GE	JA	SP
Size Dict	13,340	9,461	24,000	15,420	18,510

Table 2.2: Size of the GlobalPhone dictionaries

In this research the definition of the global phoneme set is based on the IPA chart. In this global phoneme set sounds from different languages that share the same IPA symbol share the same unit. In accordance with this idea of language independent phonemes we distinguish between language independent *polyphonemes* Υ_{LI}, containing phonemes occurring in more than one language, and N remaining sets of language dependent *monophonemes* $\Upsilon_{LD_1}, \ldots, \Upsilon_{LD_{L_N}}$. One should be careful not to confuse the term polyphonemes with polyphones that is frequently used to describe phonemes in different contexts.

Currently the global phoneme set covers 162 symbols taken from twelve languages. 83 of them are polyphonemes and 79 are monophonemes. Table 2.3 gives an overview of the set, the involved languages, and the way the symbols are shared.

2.4 Share Factor

In order to be able to measure how well data from a set of languages Λ can be shared using the global phoneme set, [SW01] defined the *unit share factor* sf_Λ as the ratio between the sum of language specific phonemes and the size of the global phoneme set. The share factor can be interpreted as the average number of languages from Λ that share a phoneme of the global phoneme set.

$$sf_\Lambda = \frac{\sum_{L_i \in \Lambda} |\Upsilon_{L_i}|}{|\Upsilon|}, |\Upsilon| = |\Upsilon_{LI}| + \sum_{L_i \in \Lambda} |\Upsilon_{LD_{L_i}}| \tag{2.1}$$

In our case we have 485 language specific phonemes. So according to table 2.3 this

Shared by	#	Modeled Phonemes (IPA symbols)	
	83	Polyphonemes shared across ≥ 2 languages	
		Consonants	Vowels
All	4	m,n,s,l	-
11	7	p,b,t,d,k,g,f	-
10	3	-	i,u,e
9	6	ŋ,v,z,j	a,o
8	1	ʃ	-
7	3	r,h,ʧ	-
6	1	-	ɛ
5	9	ɲ,ʒ,x,ʦ,ʤ	iː,y,ə,ɔ
4	4	-	ɨ,ø,ɑ,ei
3	11	ʎ,w,ç	ɪ,uː,eː,œ,oː,æ,ai,aʊ
2	34	p^h,t^h,d^j,k^h,g^j,ʁ,$ɾ_r$, $θ$,ð,s^j,z^j,ʂ,ɻ,$ʦ^h$,$ʧ^j$	ˈi,yː,ɯ,ʊ,ˈe,ɛː,ɔøː,aː,ˈa,ɑː, ˈu,ˈo,aɪ,au,ia,io,eu,oi,oʊ
	79	Monophonemes belonging to *one* language	
		Consonants	Vowels
CH	15	tʂ,t^hʂ,cç,cç^h	iʊ,iɛ,ua,uɛ,uɔ,ya,yɛ, iao,uɛi,uai,ioʊ
EN	5	$ɾ_d$	ʌ,ɝ,ɔi,ɚ
FR	5	ɥ	ɛ̃,œ̃,ɑ̃,ɔ̃
GE	3	-	ɐ,ʏ,ɔʏ
JA	2	ʔ	ɯː
KO	14	p˺,p',t˺,t',k˺,k', s',c'^{h}	ie,iə,iu,ɨi,oa,uə
KR	1	ʤ^j	-
PO	8	-	ĩ,ũ,ẽ,õ,ɐ̃,ew,ow,aw
RU	15	p^j,b^j,t^j,m^j,r^j,v^j, $ʃ^j$,$ʒ^j$,l^j,ʃtʃ,ʃtʃ^j	ja,jɛ,jɔ,ju
SP	2	β,ɣ	-
SW	9	ʈ,ɖ,ɳ,ɭ,ks	œː,æː,ʉː,ɵ
TU	0	-	-
∑	162	Silence and noises shared across languages	

Table 2.3: Overview of the Global Phoneme Set [SW01]

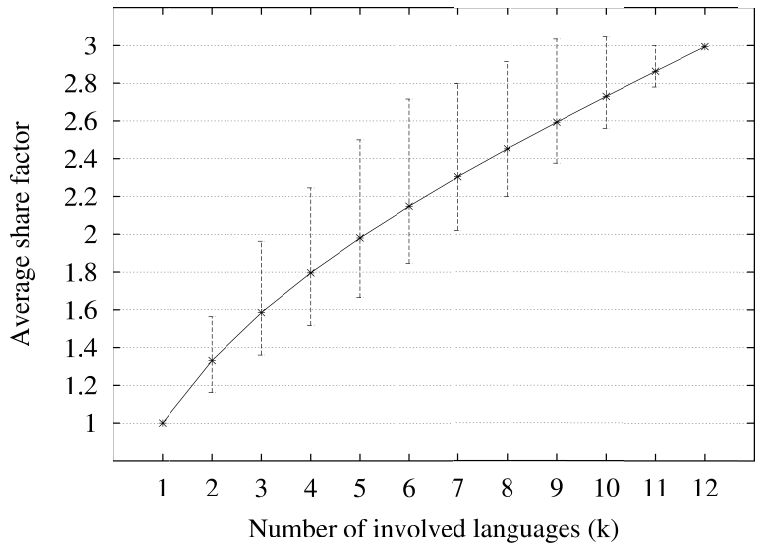


Figure 2.1: Average share factor and range for all possible subsets of the GlobalPhone languages [SW01]

results in a share factor for all twelve languages of

$$sf_{12} = \frac{|\Upsilon_{CH}| + |\Upsilon_{EN}| + \ldots + |\Upsilon_{TU}|}{\Upsilon} = \frac{485}{162} = 2.99$$

A plot of the average share factor as well as its range for all possible subsets of the twelve GlobalPhone languages can be found in figure 2.1. It would be desirable to have an almost linear growth of the share factor since this would mean that adding new languages does not increase the total number of sounds and thus provides a proportional growth in shareable training data for our acoustic models. But as we can see it does not show the desired linear growth. In chapter 3 I will introduce the same notion of share factor for articulatory features and will compare it to the phoneme share factor.

2.5 Language Independent Acoustic Modelling

The main motivation for multilingual speech recognition is the desire to be able to share acoustic training data across languages. Just as one wishes to be able to produce acoustic models that are independent of the individual speaker, speaking style or acoustic environment, language independent acoustic modeling tries to produce models that are independent of the language that a speaker uses.

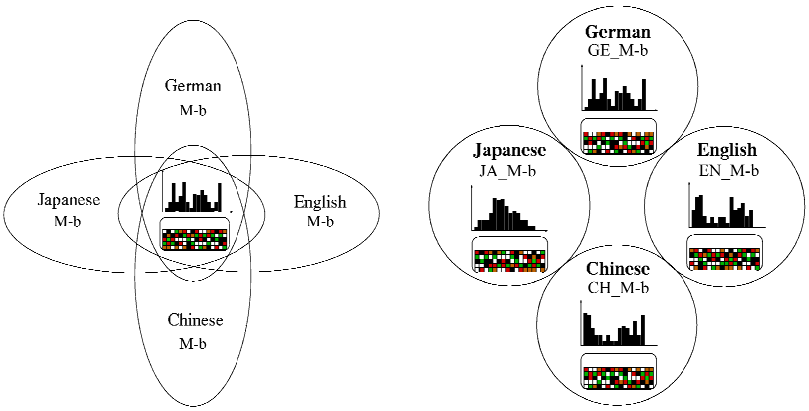


Figure 2.2: Language mixed acoustic modelling vs. language dependent [SW01]

2.5.1 Mono-, Cross-, and Multilingual

When I use the term *monolingual* in this work, e.g. monolingual feature detectors or monolingual setup, I refer to a scenario where acoustic models are applied to test data that is in the same language as the training data of the acoustic models.

On the other hand the term *crosslingual* is used when acoustic models are used on test data that is not in the same language as the training data of the models.

I call acoustic models that have been trained on more than one language *multilingual* . Note that multilingual acoustic models are thus used in a crosslingual way when they are applied to test data in a language that was not part of the languages of the training data.

2.5.2 ML-mix

I performed my multilingual experiments using a method for combining acoustic models across languages called ML-*mix* [SW01]. When training multilingual models with ML-mix, the models are common to more than one language and the training data of the models is shared across languages. The resulting models are assumed to be independent of the languages of the training data.

In the phoneme HMM based speech recognition systems used in this work the probability $p(x|s_i)$ to emit the feature vector x in state s_i is described by a weighted mixture of K_i Gaussians (compare to the acoustic model in the fundamental equation of speech recognition (1.1) in section 1.1): $p(x|s_i) = \sum_{k=1}^{K_i} w_{s_i k} N(x|\mu_{s_i k}, \sum_{s_i k})$. Figure 2.2 illustrates ML-*mix* in comparison to building individual recognizers for the involved languages. In this figure the mixture weights w are symbolized by bar graphs and the Gaussian components $N(x|\mu, \Sigma)$ by rounded boxes containing a gray scale map. In the ML-*mix* method the training data for the acoustic models of polyphonemes is fully shared across languages. For every symbol from the global phoneme set we initialize one mixture of Gaussians per state and train the model by sharing the data from all languages belonging to this symbol.

Chapter 3

Articulatory Features

In this chapter I introduce the concept of articulatory features that I am using for my research. I start out by giving a brief introduction into the human speech production process, the role that the features play in describing it, and which kind of articulatory features I am actually modelling. I then show how the use of articulatory features can benefit automatic speech recognition by comparing the possibilities that they offer in modelling speech to the currently widely used phoneme based HMM approach. In order to do so I summarize previous research on the use of articulatory features in recognition systems.

Also as indicated before I am going to expand the multilingual concepts that were introduced in chapter 2 for phonemes to articulatory features by introducing a global articulatory feature set. I demonstrate that its potential for sharing training data across languages is superior to that of the global phoneme set.

3.1 Human Speech Production

The production of human speech is mainly based upon the modification of an egressive airstream by the articulators in the human vocal tract. The modifications by the articulators are usually a combination of the potential vibration of the vocal cords and a resonation in the remaining vocal tract depending on its current shape. The activity of the vocal organs in making a speech sound is called *articulation*. The necessary air stream is usually produced by the lungs and in some languages only these so called *pulmonic* sounds exist. However in many languages at least one of two additional mechanisms for producing the necessary air stream exists. First by closing the glottis, air that is trapped between the glottis and an additional constriction in the vocal tract can be used to produce an airflow that either flows out of the vocal tract or into it. By compressing the air it is forced to flow outwards creating a sound that is called an *ejective*. Expanding the trapped air leads to an inward air stream when the forward closure is released. This results in an *implosive* sound.

Second, when the back of the tongue against the soft palate is used instead of the glottis to create a little room of trapped air one gets sounds that are commonly known as *clicks*.

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Figure 3.1: The consonant table from the IPA chart [Ass99]

3.2 Phonetic Description of Speech

X-ray films of the speech organs in action show that they are in continuous motion during the act of speaking. The same can be seen when looking at the spectrogram representation of speech. The patterns are in a constant flow of motion and no boundaries between sounds seem to exist. However when linguistic knowledge of the underlying language is taken into account it is possible to segment speech by identifying points where linguistically significant changes can be made. The existence of such a segmentation is the base of current phonological analysis. Furthermore it is assumed that every segment can be assigned an *articulatory target*. The articulatory target describes the configuration of the vocal tract and organs that are representative for the described segment and sound respectively. Usually the involved articulators make a continuous movement from and to the target during the speech production. And in some instances the target might be held for a certain amount of time.

3.2.1 Consonants, Vowels, and Features

For the description of the above segments IPA heavily relies on the distinction between vowels and consonants. Speech involves consecutive widening and narrowing of the vocal tract. The openings are used to define syllables and act as the *nucleus* of the syllable. Segments that involve a narrow or closed vocal tract are called *consonants*. Sounds with a wide vocal tract in which the air flows largely uninhibited carry the terminus *vowel*. Because of this general difference between vowels and consonants IPA has decided to use different schemes to describe them. This results in an IPA chart for describing phonemes that has separate sections for vowels and consonants. For a detailed description of the IPA chart and the possibilities it offers for describing the sounds of human speech the reader may refer to [Ass99].

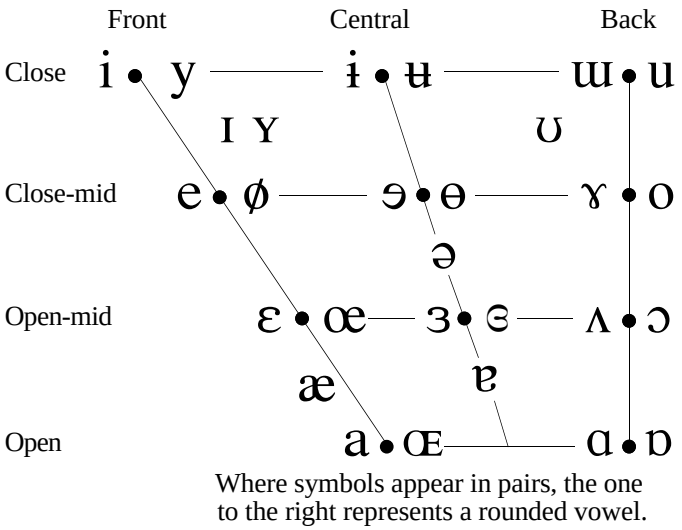


Figure 3.2: The vowel quadrilateral from the IPA chart [Ass99]

The generic classification into vowels and consonants as well as the different attributes used to describe the way the sounds from this classes are articulated is what we refer to as *articulatory features* (AF) in this work.

Consonants

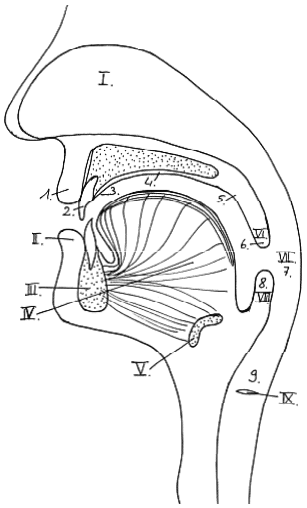
There are commonly three articulatory feature dimensions in which to describe consonants. First there is the *place of articulation* which tries to describe the position of the constriction of the vocal tract on the mid-sagittal plane. The different places are represented by the columns in the IPA consonant chart (see figure 3.1). Figure 3.3 shows the mid-sagittal plane of the human vocal tract and names possible places of articulation.

Second *manner of articulation* is used as another dimension. It describes the degree of the constriction of the vocal tract, the position of the velum, and some other attributes referring the behavior of the articulators such as vibration and redirection of the airstream from the middle to the side of the vocal tract.

The third dimension describes the vocal cord vibration by classifying consonants as either *voiced* (vocal cords vibrate) or *unvoiced* (no vibration). In the IPA table for consonants every cell is split into half. The left half always refers to the unvoiced version of a consonant and the right one to the voiced version.

Vowels

Because of the generally open character of vowels they cannot as easily be described by means of 'place of articulation' as consonants can. For vowels it is more appropriate to classify them by describing the horizontal and vertical position of the highest point of the



Articulators are marked by Roman numbers: I. nasal cavity, II. lower lip, III. mandible, IV. tongue, V. hyoid, VI. uvula, VII. pharynx, VIII. epiglottis, IX. glottis

Places of articulation are marked by arabic numbers: 1. Lips, 2. incisors, 3. teeth-ridge, 4. hard palate, 5. soft palate, 6. uvula, 7. pharynx, 8. epiglottis, 9. glottis

Figure 3.3: Mid-sagittal plane of the human head [Ell97]

tongue called the *dorsum*. The two dimensions of the dorsum position lead to the notion of an abstract vowel space that is usually visualized using the *Vowel Quadrilateral* depicted in figure 3.2. In order to incorporate the use of the lips unrounded vowels are placed to the left of the back or front line of the quadrilateral and rounded ones to the right. Also all vowels are classified as voiced sounds.

3.3 Articulatory Features for ASR

One of the major problems that is encountered when doing automatic speech recognition is the amount of variability in the acoustic signal. Not only does the same word sequence spoken by different speakers show a very different acoustical representation, e.g. when looking at the spectrogram. But also when the same utterance is repeatedly produced by the same speaker under the same circumstances the individual acoustic representations can be largely different. Our knowledge at the acoustic level is still insufficient to model and 'normalize' these variations that we encounter. This has lead to the fact that models of human speech heavily rely on automatic learning techniques often combined with statistical methods. With progress in understanding the laws governing the acoustical representation of human speech, as well as the ability to represent this knowledge in a form suitable for machines, the existing learning techniques will be refined and new ones developed that will make it possible to more accurately model the human speech production process.

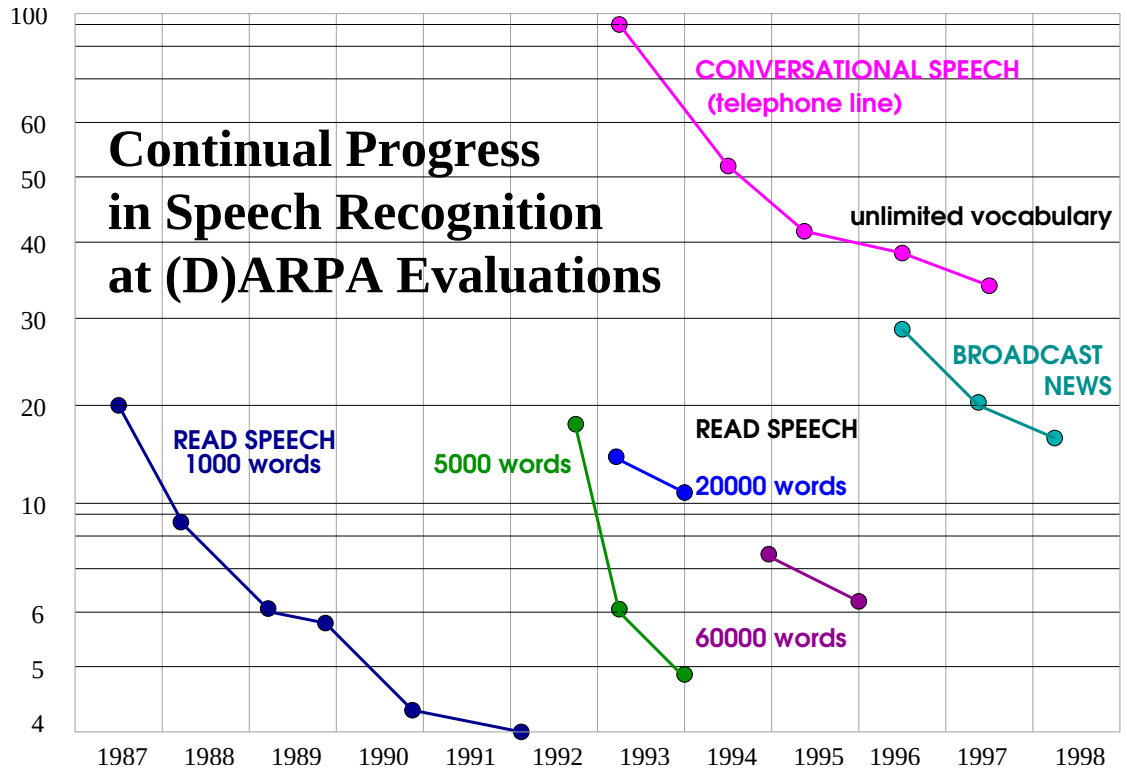


Figure 3.4: Progress (word error rates) in speech recognition over the years[Rog01]

3.3.1 The Conventional Approach

Over the last decades speech recognition research has worked under the assumption that words in speech are composed of a sequence of phonemes. Sometimes this model is called 'beads-on-a-string' [Ost99].

The application of this paradigm has produced well performing ASR systems on a wide variety of task, especially for very controlled conditions such as read or planned speech in noiseless environments. However when it comes to spontaneous speech or to more adverse environmental conditions, e.g. noisy or reverberant environments, the performance of today's speech recognition systems degrades rapidly. For example in the 1998 DARPA Broadcast News Benchmark Test the lowest word error rate that was achieved for planned speech under high fidelity, no noise conditions (F0 conditions) was 7.8% [PFG+99]. However in the 2001 NIST Hub-5 Evaluation the lowest reported word error rate on the Switchboard corpus, which consists of conversational speech, was 24.5%. As the word error rates drop with the development of better performing speech recognition systems this performance gap between speaking styles continues to exist. Figure 3.4 illustrates this fact by giving an overview of the progress of speech recognition over the years on different corpora with different speaking styles. It makes clear that read or planned speech is much easier to recognize than spontaneous one.

3.3.2 Advantages from Feature Detectors

Different explanations for the poor performance of HMM based recognizers on spontaneous speech as well as reasons why articulatory features might help in overcoming the encountered problems have been proposed by different researchers.

Ostendorf [Ost99], for example, argues that pronunciation variability in spontaneous speech is the main reason for the poor performance. She claims that though it is possible to model pronunciation variants using a phonetic representation of words the success of this approach has been limited. Ostendorf therefore assumes that pronunciation variants are only poorly described by means of phoneme substitution, deletion, and insertion. She also thinks that the use of linguistically motivated distinctive features could provide the necessary granularity to better deal with pronunciation variants by using context dependent rules that describe the value changes of features.

Kirchhoff [Kir00] also acknowledges that it is easier to model pronunciation variants with the help of articulatory features. She points out that articulatory features exhibit a dual nature because they have a relation to the speech signal as well as to higher-level linguistic units. Furthermore, since a feature often is common to multiple phonemes, training data is better shared for features than for phonemes. Also for AF detection fewer classes have to be distinguished (e.g. binary features). Therefore statistical models can be trained more robustly for articulatory features than for phonemes. Consequently feature recognition rates frequently outperform phoneme recognition rates.

Another reason for the poor performance of automatic speech recognition systems on spontaneous speech is the increased occurrence of coarticulation effects as compared to planned or read speech. In [Kir98] Kirchhoff makes the assumption that coarticulation can be modelled more robustly in the production based domain than in the acoustic one. She also assumes articulatory features are more robust towards cross speaker variation and signal distortions such as additive noise.

Eide [Eid01] argues that the direct modelling of phonemes from the waveform as it is usually done in the beads-on-a-string model disregards some of the phenomena of conversational speech such as the relaxation of the requirements on the production of certain distinctive features. She claims that variations in the pronunciation may cause big phonemic differences while in terms of articulatory features the difference may be considerably smaller because only few articulatory features actually change their value. Therefore she argues that the task of recovering a word sequence from a feature representation is more feasible than from a phonemic representation.

Wester, Chang, and Greenberg [CGW01][WGC01] believe that corpora are optimally annotated at the articulatory-acoustic feature level. They are of the opinion that the transformation from AF to phonetic segments does not transport sufficient detail and richness common to the speech signal at the phonetic level.

Deng [DS94] sees 'residual' variability in speech that is difficult to explain in terms of general properties as the main obstacle in achieving a high word recognition accuracy. He argues that today's speech recognition systems make use of statistical methods and automatic learning procedures in order to model speech at a detailed level because of

a lack of reliable speech knowledge. He proposes to use constellations of overlapping articulatory features as speech units that should be able to model these variations in speech incorporating all necessary contextual information. At the same time the number of units is small enough as not to demand too high an amount of training data.

3.4 A Language Independent Set of Articulatory Features

In chapter 2 I introduced the global phoneme set that Schultz used for language independent acoustic modelling. In order to be able to model articulatory features in a language independent way I expand the concept of language independent phonemes to features.

Since the sounds of the global phoneme set, that was derived from the IPA tables, are presumed to be independent of the underlying language, I can view the articulatory features, which IPA uses to describe these symbols, also as language independent. Thus I obtain a set of language independent articulatory features by reversing the mapping from bundles of features to phonemes introduced by IPA. This global set of features is depicted in table 3.1 as well as the languages in which the individual features occur. The feature set has been built on the five languages Chinese, English, German, Japanese, and Spanish to which I limited my research (see 2.2).

The table shows 37 different features. 21 of those features occur in all languages.

3.4.1 Share Factor for AF

Just as it has been done for the units of the global phoneme set I define the language dependent sets of features Φ_{L_i} containing all the features that are attributed to at least one sound from language L_i. Also let Φ_{LI} denote the set of language independent articulatory features occurring in more than one language, and let $\Phi_{LD_{L_i}}$ refer to the set of language dependent features only occurring in language L_i. So I define the feature share factor analogous to the unit share factor in 2.1 as the ratio between the sum of language specific articulatory features and the number of features for a global feature set given a set of languages Λ. The feature share factor can be interpreted as the average number of languages that share an articulatory feature, averaged over all features.

$$sf_\Lambda = \frac{\sum_{i\in\Lambda} \mid \Phi_{L_i} \mid}{\mid \Phi \mid}, \mid \Phi \mid = \mid \Phi_{LI} \mid + \sum_{i\in\Lambda} \mid \Phi_{LD_{L_i}} \mid \tag{3.1}$$

Figure 3.5 shows the average share factor and its range for the AF in comparison to the share factor of the units for all possible subsets of fixed size from our set of five selected languages. When comparing the share factor of the AF to the share factor of the global phonetic units one sees that the factor of the AF is always larger, that it grows almost linearly, and that the variation of the share factor for the sets of a fixed size is smaller. One can therefore expect that training the AF detectors in a multilingual way is going to

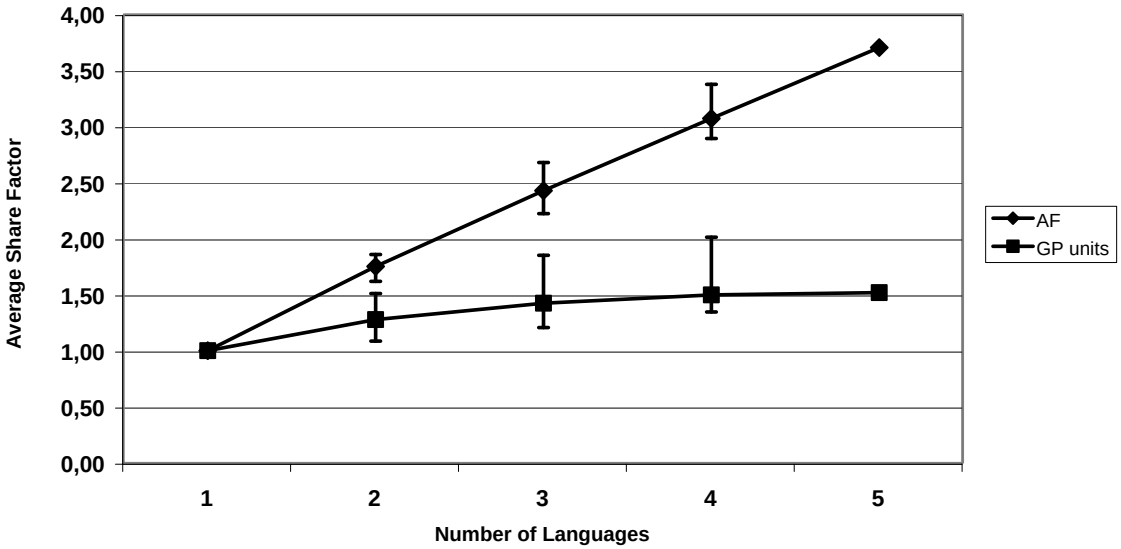


Figure 3.5: Average Share Factor for the Five Selected Languages

make better use of the training data from the different languages than the multilingual training of the phonetic units.

	Feature	**Languages**
CONSONANT		CH GE EN JA SP
	VOICED	CH GE EN JA SP
	UNVOICED	CH GE EN JA SP
	ASPIRATED	CH EN
	PLOSIVE	CH GE EN JA SP
	NASAL	CH GE EN JA SP
Manner	TRILL	GE SP
	FLAP	EN SP
	FRICATIVE	CH GE EN JA SP
	AFFRICATE	CH GE EN JA SP
	APPROXIMANT	CH GE EN JA SP
	LATERAL-APPROXIMANT	CH GE EN JA SP
	BILABIAL	CH GE EN JA SP
	LABIODENTAL	CH GE EN JA SP
	DENTAL	EN SP
	ALVEOLAR	CH GE EN JA SP
Place	POSTALVEOLAR	GE EN JA SP
	RETROFLEX	CH EN
	PALATAL	CH GE EN JA SP
	VELAR	CH GE EN JA SP
	UVULAR	JA
	GLOTTAL	GE EN JA
VOWEL		CH GE EN JA SP
	ROUND	CH GE EN JA SP
	UNROUND	CH GE EN JA SP
	TONAL1-5	CH
Tongue Position		
	CLOSE	CH GE EN JA SP
Vertical	CLOSE-MID	GE EN JA SP
	OPEN	CH GE EN JA SP
	OPEN-MID	CH GE EN
	FRONT	CH GE EN JA SP
Horizontal	CENTRAL	GE EN
	BACK	CH GE EN JA SP

Table 3.1: Table of the global feature set and the languages in which the features occur

Chapter 4

Training Articulatory Feature Detectors

The aim of this research is to incorporate the concept of articulatory features into a speech recognition system. A first step in that direction is to build dedicated detectors for these features in order to examine whether it is possible to reliably extract the feature information from the acoustic signal for different languages.

I therefore built articulatory feature detectors for a set of five languages. These feature detectors were then evaluated on their individual languages as well as on the other four languages in order to investigate the potential of detecting articulatory features across languages.

Using the technique ML-mix (see 2.5) for language independent acoustic modelling I trained and evaluated a set of multilingual detectors, using all possible combinations of the five selected languages.

4.1 Monolingual Detectors in Five Languages

In [MW02] Metze built articulatory feature detectors by regarding articulatory features as an abstract description of a speaker's phonological intention. So the same modelling techniques as for words or phonemes can be used. In his setup Metze estimated Gaussian mixture models on mel-frequency scaled coefficients for 76 binary phonological features. For every feature two models were estimated, one for detecting the presence of the feature, and one for detecting its absence.

For my research I used the same modelling techniques to build detectors for articulatory features. The selection of articulatory features in this work however differs slightly from the one Metze used. He took the linguistically motivated questions that are used during the construction of a decision tree for context-dependent modelling as a set of features. I on the other hand modelled the articulatory features defined by IPA in its phoneme charts to describe the sounds of human speech (see 3.2). These features are just as in Metze's case binary features that are either present or absent.

As a first experiment I built feature detectors on the five selected languages in order to test my setup and to see whether the modelling techniques from [MW02] can be applied to the GlobalPhone corpus.

4.1.1 Feature Extraction

Every 'feature present' and 'feature absent' detector was modelled by a mixture of 256 Gaussians. The input vectors for the mixtures were obtained from 13 dimensional mel frequency scaled cepstral coefficients (MFCC) combined with their deltas and delta-deltas, the zero crossing rate of the signal, its power, and the first and second derivative of the power. The resulting 43 dimensional feature vector was then reduced to 32 dimensions using an LDA transformation. It is possible to use a comparatively high number of Gaussians to model the articulatory features, because more training data is available per model than, for example, for phonemes.

4.1.2 Training

Recognizers based on context dependent sub-phonetic units already existed for the five languages that I chose for my research. In those recognizers every phoneme is modelled by three states (begin, middle, end). Using this recognizers I produced transcriptions of the training and test data on a sub-phonetic level by means of a forced alignment.

The first step in training the feature detectors was the calculation of the LDA transformation with the context independent sub-phonetic units as classes. Then the models for the feature detectors were initialized using the k-means algorithm and trained with four iterations of label training. The mapping of the sub-phonetic transcription to the features was done using the IPA table that describes phonemes in terms of articulatory features (see 3.2.1). For example the phoneme \ɘ\ is attributed with the features CENTRAL, CLOSE-MID, and UNROUND. So feature vectors that according to the transcription belong to \ɘ\ were used to train the present models for CENTRAL, CLOSE-MID, and UNROUND, as well as the absent models of all the other features. The feature detectors were only trained with acoustic material that belonged to sub-phonetic middle states. This was done because articulatory features are not static but rather change dynamically. Since I only model abstract classes of articulatory features, I assume that the acoustic data that belongs to middle states is the most representative data for the respective classes.

In addition to the acoustic models for the detectors I also estimated prior probabilities for the occurrence of the individual features by counting the number of training vectors each model got.

4.1.3 Evaluation

Using the acoustic models for the features and the calculated prior probabilities I evaluated the feature detectors by determining their classification accuracy on the development set of their language.

	Test Set				
CA	CH	EN	GE	JA	SP
	93.52%	93.83%	92.94%	95.22%	93.46%

Table 4.1: Classification Accuracy of the AF detectors

Just as the training, the evaluation was only done for the acoustic vectors that according to the transcription belong to sub-phonetic middle states. For each test vector every feature was classified into either present or absent. To do so the likelihood score of the absent model was subtracted from the score of the present model and an offset was added that was the difference between the score of the feature present prior probability and the score of the absent prior probability. If the resulting value was below or equal zero the frame was classified as feature present, otherwise as feature absent.

The resulting classification accuracies averaged over all features are shown in table 4.1. Detailed results for every single feature can be found in appendix A.

One can see that I get a similarly high average classification accuracy for all languages. This is consistent with the expectation mentioned in 3.3.2 that statistical models for binary features can be estimated very robustly. From appendix A one can see that within a language the classification of the individual features lies roughly in the range from 80% to 99%. The only exception is Japanese. Here the lowest classification accuracy is still 89.92%.

4.2 Crosslingual AF Detection

With my next experiment I wanted to find out whether articulatory feature detection is robust to inter language variabilities. For this purpose I tested each monolingual feature detector on the other four languages that it was not trained on. For this crosslingual classification I used the prior probabilities that were estimated on the language that the classificators were trained on.

Table 4.2 shows the results of this evaluation. Every row gives the results of the detectors trained on one language (AF LID) when tested on each of the five languages. The results are averaged over the classification accuracy of the detectors for the individual features. However since not all features of the test set language might be covered by the detectors from the language that is being tested, the classification accuracies could only be averaged over the detectors for features that exist in both, the test and training language. So for example, when I tested the Japanese feature detectors on Spanish, I could not determine the classification accuracy for the features TRILL, DENTAL, and FLAP. These features are attributed to some Spanish phonemes, however no Japanese phonemes with these features exist, and thus no Japanese feature detectors for them. At the same time there are Japanese feature detectors for GLOTTAL and UVULAR. However I could not test them on the Spanish test set, because these features do not occur in the

	Test Set				
AF LID	CH	EN	GE	JA	SP
CH	**93.52%**	87.42%	88.23%	86.45%	83.22%
EN	87.74%	**93.83%**	89.17%	88.41%	87.90%
GE	88.57%	87.90%	**92.94%**	86.46%	82.68%
JA	87.11%	87.65%	86.77%	**95.22%**	87.39%
SP	84.76%	86.36%	83.31%	87.76%	**93.46%**

Table 4.2: Average classification accuracy of the AF detectors

Spanish phonemes. The diagonal of the result matrix naturally gives the monolingual results mentioned earlier. The detailed results for the individual feature detectors from all languages tested on all languages can be found in the appendix A.

As one can see the highest relative drop in average classification accuracy is 11.53%, and occurs when decoding Spanish with Chinese features. The least loss occurs when using English feature detectors to classify the German data. For this constellation the average classification accuracy drops only 4.1% relative.

However these numbers only paint a very rough picture because they only concern subsets of the features in the test languages. More details can be drawn from the results of the individual feature detectors. The largest relative drop in performance, 11.54%, for a single feature detector occurs for the German detector for the feature ALVEOLAR on the Japanese test set. On the other hand, for every test set there are detectors from languages other than the test language that show a relative increase in performance. The highest relative gain, 13.50%, is seen when the Japanese feature detector for ALVEOLAR is tested on the German data.

The next subsection takes a closer look at this phenomena of possible gains.

4.2.1 Crosslingual Combination of AF Detectors

As mentioned above for individual features of a language l it can happen that feature detectors from languages other than l perform better, in terms of classification accuracy, than the detector trained on l. Lets take the feature LABIODENTAL for Chinese as an example. As we can see from appendix A.1 the Chinese feature detector for this feature achieves a classification accuracy of 98.46%. However the Spanish feature detectors reaches a classification accuracy of 98.84% for Chinese (see A.5), and the Japanese feature detector classifies with an accuracy of 99.23% (see A.4). Therefore both outperform the Chinese AF detector for LABIODENTAL. Figures 4.1 and 4.2 illustrate this effect for the feature detectors of all five languages. Every graph shows the results of all five sets of feature detectors on one of the five test sets. The solid line connects the results of the feature detectors that were trained on the language of the test set. Whenever a dot appears above this line that means that a feature detector trained on a different language than that of the test set outperforms the feature detector that was trained on the test language.

	Test Set				
	CH	EN	GE	JA	SP
monolingual	93.52%	93.83%	92.94%	95.22%	93.46%
incl. test set	95.04%	96.13%	96.12%	96.26%	96.36%
monolingual subset	94.36%	93.83%	92.94%	95.18%	93.46%
excl. test set	95.67%	95.61%	95.88%	95.58%	96.12%

Table 4.3: Results of the crosslingual combination of feature detectors

For every test language we can now combine the best feature detectors from all languages to form a new set of feature detectors for a given test set and thus improve the average classification accuracy as compared to the monolingual results in 4.1.3. If we combine the articulatory feature detectors from languages other than the test set only, we should see a gain compared to the crosslingual results in table 4.2.

The average classification accuracies on all five test sets for this two ways of combination are shown in table 4.3. The top half of the table compares the average classification accuracy of the corresponding monolingual feature detectors ('monolingual') with the average classification accuracy, when the best feature detectors from all five languages, including the test set language, are combined ('incl. test set').

The lower half of the table shows the results of the second method of combination. Given a language l, the other four languages might not completely cover the complete feature set of l. Therefore we first give the average classification accuracy of the monolingual feature detectors on the subset of features in l that is indeed covered by the detectors from the other four languages. These numbers can be found in the row labeled 'monolingual subset'. We can then compare this number to the average classification accuracy of the detectors from the four languages other than l on this subset ('excl. test set').

As we can see for every of those scenarios the cross lingual combination brings a performance gain. Especially for German and Spanish there are many feature detectors from other languages that outperform the 'original' detectors. This leads to the largest relative gains in both scenarios. For German the classification accuracy improves 3.4% relative when selecting from all feature detectors, and 3.2% relative when selecting from the subset.

It is interesting to note that in figures 4.1 and 4.2 there are quite a number of features in certain languages where the feature detector that has been trained on the test language is the one that performs the worst. Examples for such feature detectors are the Chinese detector for AFFRICATE, the English detector for APPROXIMANT, the German detector for LABIODENTAL, the JAPANESE detector for GLOTTAL, and the Spanish detector for POSTALVEOLAR.

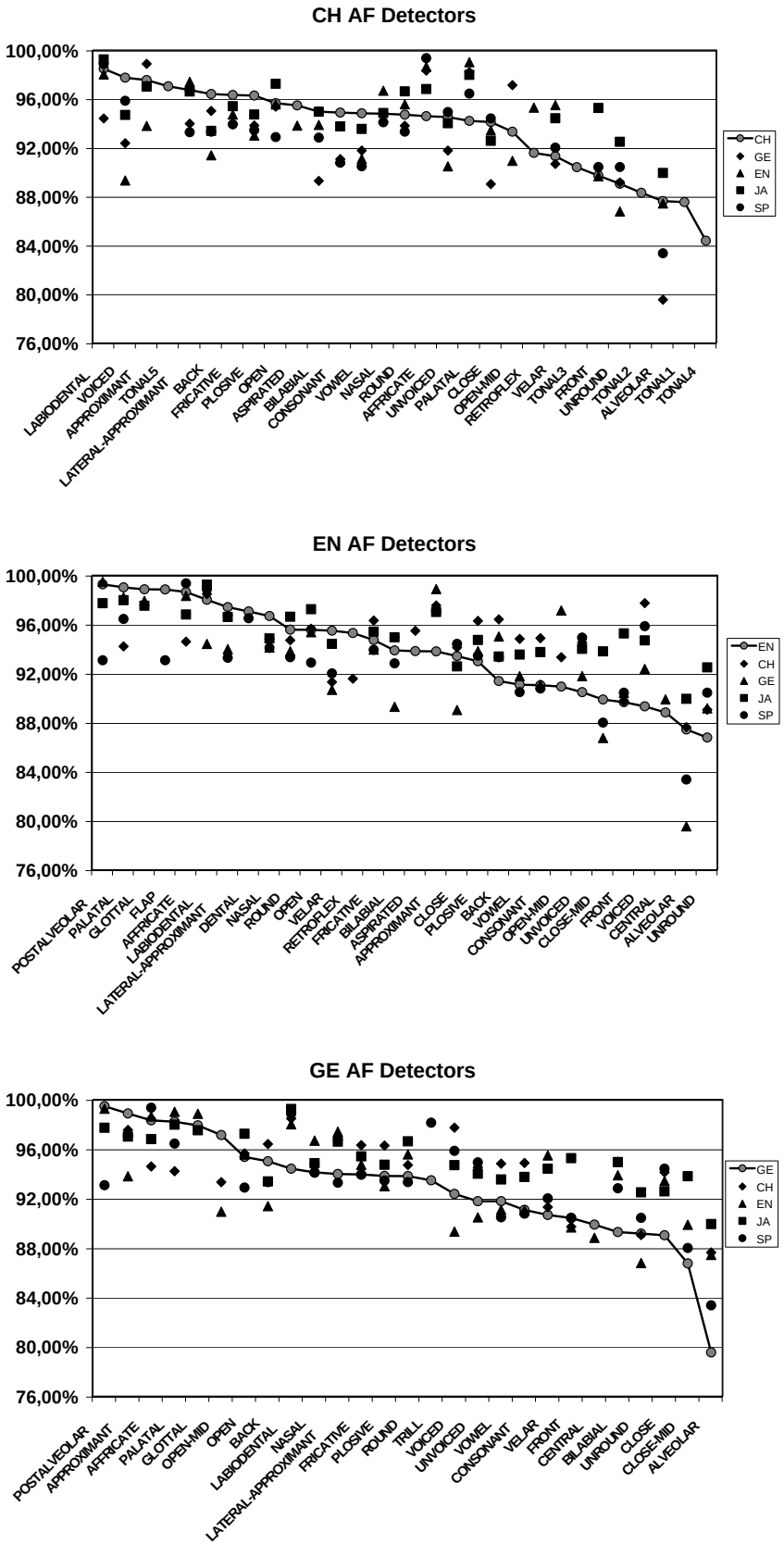


Figure 4.1: Results of the crosslingual evaluation of all feature detectors on the CH, EN, GE, and JA test set

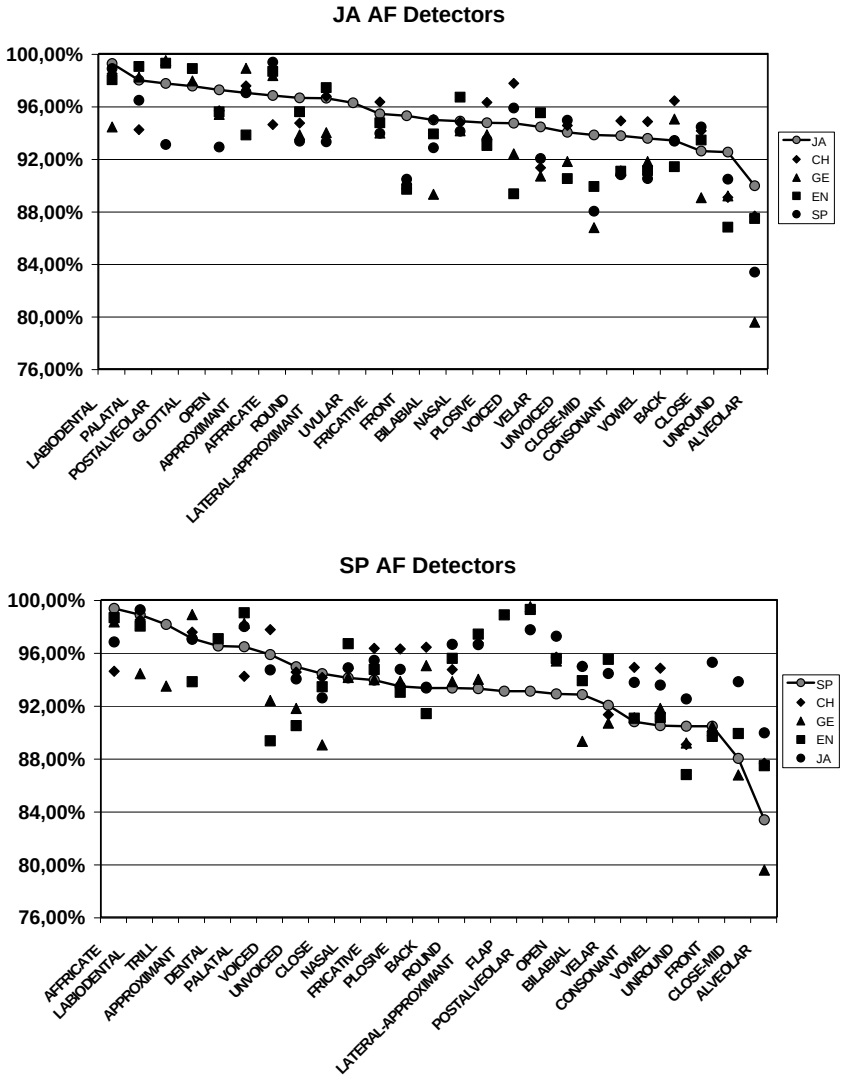


Figure 4.2: Results of the crosslingual evaluation of all feature detectors on the SP test set

4.3 Multilingual Classification

I trained multilingual AF detectors by sharing the training data from n languages to train detectors that are no longer language specific but can be used to detect features in many languages. Since I used the training method 'Multilingual Mixed' (see 4.3.1) I call a set of feature detectors trained on n languages MMn. If I refer to a set of specific languages that the detectors were trained on I will do so by simply combining the training language identifiers with underscores. E.g. MM3 feature detectors trained on the languages English, German, and Japanese will be called EN_GE_JA detectors.

4.3.1 Training

When training acoustic models with the method 'Multilingual Mixed', combining n languages by simply using the training material from all n languages would mean that the available training material would roughly increase n fold. Therefore, in order to ensure that the observed effects do not just occur because of an increase in training material, I only took a fraction of the training material of each involved language depending on how many languages were involved. E.g. for MM AF detectors trained with German and English data I used half of the German training utterances and half of the English.

Since I am working on five languages I can build MM2, MM3, MM4, and MM5 feature detectors. When training on n out of five languages there are $\binom{n}{5}$ possible combinations of languages. In order to explore the multilingual possibilities I trained all possibilities for combining two to five languages.

4.3.2 Evaluation

Figure 4.3 gives an overview over the performance of the MMn detectors. For every MMn detector the corresponding chart shows the range of the performance of all possible MMn detector sets on all possible test languages compared to the performance of the monolingual AF detectors that were trained on the test language. The performance averaged over the individual AF detectors for all possible combinations training data can be found in appendix A. We can see that if we choose the right combination of languages for a given test set the performance of the MMn detectors is only slightly worse than that of the corresponding monolingual ones.

In order to see whether using all available training data instead of just a fraction for training the multilingual detectors would improve their performance, I trained the MM5 detector on the complete training data of the five languages. However the evaluation only showed very little absolute improvements of 0.75% on the Chinese test set, 0.24% on English, and 0.22% on Japanese. On the German and Spanish set the performance suffered slightly by just 0.09% and 0.08%. So given the number of parameters of the feature detectors the fraction of training material from the individual languages seems to be sufficient to learn the language dependent properties of the features. This might be an indication that the acoustic manifestation of articulatory features is indeed very similar for

	Test LID				
	CH	EN	GE	JA	SP
MM5	90.56%	90.40%	88.94%	90.90%	88.71%
MM4	89.51%	88.27%	88.04%	88.02%	87.06%
rel. loss	1.6%	2.4%	1.0%	3.2%	1.9%

Table 4.4: Comparison between MM4 detectors that were not trained on the test language and MM5 detectors

different languages, so that there are only few language dependent characteristics in the acoustic signal.

Given the five languages it is also of interest which influence the presence of the test language among the training languages has. Table 4.4 compares the performance of the MM4 detectors that were trained on all four languages except the test language with the performance of the the detectors trained on all five languages (MM5 detectors), thus including the test language. Again there is the problem that not all features of the test language might be covered by the MM4 feature detectors. Therefore the classification accuracy of the MM5 detectors is only averaged over the features of the test language that are also covered by the corresponding MM4 detectors.

As is to be expected the MM5 detectors always outperform the MM4 detectors, since the test language has been seen during training. The highest relative loss in classification accuracy, 3.2%, occurs for Japanese, while the lowest relative loss, 1.0%, occurs for German.

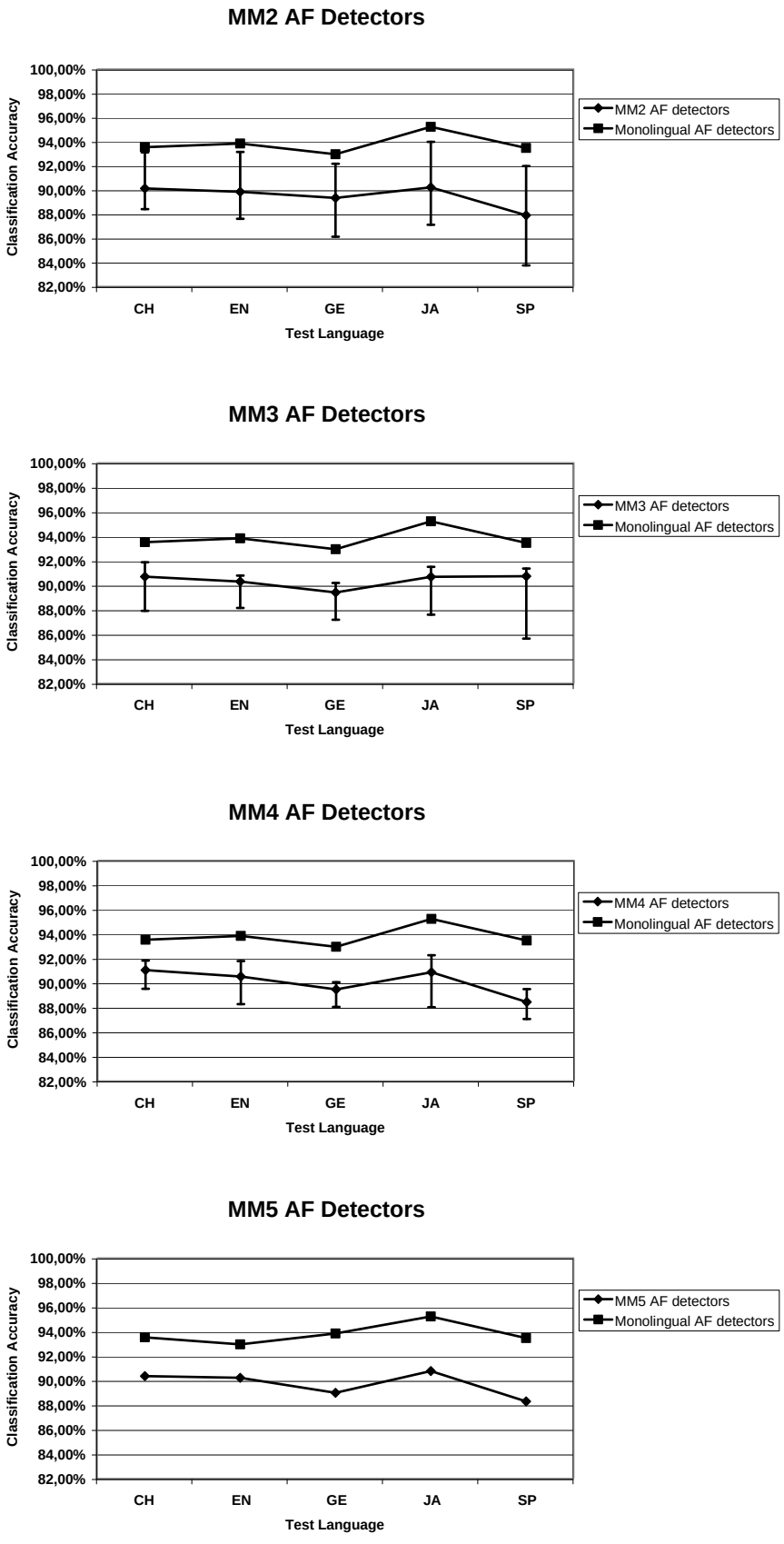


Figure 4.3: Performance overview of the MMn recognizers

Chapter 5

Stream Architecture for Including AF into ASR

The previous chapter has introduced a method to build dedicated detectors for articulatory features using Gaussian mixture models. This chapter will take the step from just detecting features to actively using them in the task of recognizing speech.

Different systems and ways have been proposed that accomplish this task. A speech recognition frame work that makes sole use of articulatory features using the concept of overlapping features has been proposed in [DS94]. AF detectors have also been used to improve robustness towards noise and reverberation. [Kir98] reports speech recognition experiments for this conditions by comparing individual acoustic, articulatory, and combined speech recognition systems that are hybrid ANN/HMM recognizers. She shows that under very noisy conditions a recognition system that is solely based on articulatory features perform better than one that uses phonemic models. A system that combines the output from phoneme models and feature models even performs better under clean, reverberant and low noise conditions. Other recent work [Eid01] makes use of articulatory information by including the output of AF classifiers into the front-end of otherwise standard low-resource recognizers. In that work a system that integrates articulatory feature detectors with phoneme based models also shows an improved performance under noisy conditions.

The goal of the research in this work is not to build a recognition system solely based on articulatory features. Instead, I concentrate on supporting an existing HMM based recognizer with models for articulatory features as an additional source of information. My focus is on the question whether articulatory features can be used in multilingual and crosslingual settings in order to improve the recognition accuracy of the base HMM recognizer. To do so I make use of a flexible approach that integrates dedicated detectors for articulatory features with conventional context-dependent sub-phone models, using a stream architecture [MW02].

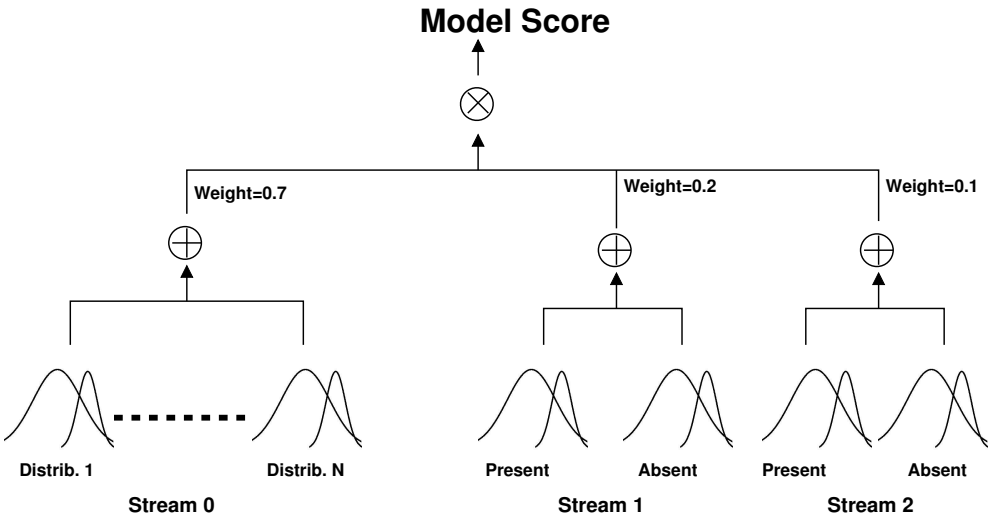


Figure 5.1: Stream setup with combined 'feature absent' and 'feature present' detectors

5.1 A Flexible Stream Architecture

In [MW02] Metze regarded feature detectors, such as the ones from chapter 4, as phonologically distinctive properties of speech sounds that can be used to support conventional acoustic models.

If one wants to combine feature models with standard models, Kirchhoff has shown in [Kir99] that the most promising approach is the combination of scores at the log-likelihood level. The conventional models that Metze used and that I use in this research are context dependent sub-phonetic units that are modelled as a mixture of Gaussians. Because of that, and because of the design of our feature detectors as described in chapter 4, the acoustic *score* (negative log probability) for a model is now computed as the weighted sum of Gaussian mixtures models, representing the standard models, and 'feature' probability distribution functions. The result is a flexible stream based architecture which is illustrated in figure 5.1. The 0th stream consists of the context dependent standard models. For every articulatory feature that I use I add an additional stream that contains the 'present' and 'absent' models for this feature as described in the last chapter. When the decoder now computes the score of a model m given a feature Vector X it adds the score of the corresponding context dependent model from the 0th stream with the scores from either the absent or present models from the other streams, depending whether m is attributed with the respective feature or not. The mapping to determine whether a context dependent model is attributed with a feature or not is done according to our global feature set.

For the first experiments the single weights were hand selected by relying on empirical experience and common sense. As described below I also tried to learn the optimal combination of weights by using a discriminative training approach.

5.2 Adapting Stream Weights

The weighted combination of the scores from the HMM based models and the articulatory feature detectors as described above requires the selection of an appropriate set of weights. The weights control the influence that the individual detectors have on calculating the score and thus have a great impact on the search for the best hypothesis. The task is to find an optimal set of weights $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ that minimizes the word error rate of the recognition system.

5.2.1 Educated Guess

[MW02] used heuristical methods to impose an order on the features detectors. The feature detectors were then added in this order. Every detector was assigned the same constant weight. Depending on the number of features added, the 0th model stream then got the remaining weight mass in order to normalize the sum of the weights to 1.0. So for example, if one would decide to use two feature streams in addition to standard models and a feature stream weight of 0.05, one would get $\Lambda = (0.9, 0.05, 0.05)$

Three heuristics were developed and tested:

FRAME-CR Add the feature detectors in the order of their classification accuracy, starting with the best.

DECODE Evaluate all possible systems consisting of the standard model stream and one additional feature stream. Add the features according to which gave the best improvements in the two streams setup.

TREE A divisive clustering tree on a generic speech model was created, using the data driven strategy for generating context dependent models. The features were used as splitting questions in the tree. The clustering algorithms generates splits according to greatest likelihood gain and features are being added in this order.

All three heuristics lead to similar reductions in word error rate; no selection method seems to be superior. Even though the heuristics might be helpful in preselecting features, they do not give any hints on what weight to assign to the features. The use of a constant weight that is the same for all features seems counter intuitive, as one would expect that some features should have a higher influence on the scoring process than others. This influence should also vary depending on the other features involved in calculating the score.

Since the IBIS decoder implements a beam search, and because the scores from the feature detectors might have a different magnitude than the scores from the conventional models, one has to keep an eye on the total score of the decoded utterances. Because the decoder works with absolute beams, the simple down scaling of the acoustic score, e.g. leaving the feature stream weights at 0.0 and scaling the 0th stream to 0.8, would mean effectively widening the beams. For this reason, one has to make sure that the total sum

of the scores in all experiments is always greater or equal to the sum of the scores for the baseline system against which I compare my results.

5.2.2 Weight Selection with DMC

'Guessing' the weights for the feature streams is naturally unsatisfying since it will most likely provide a solution that is far from optimal. Also the fact that none of the heuristic feature selection methods introduced above seemed to be superior to the others, gives the impression that more improvements can be reached by better ways of selecting the stream weights. So far it does not seem to be feasible to apply rules, e.g. obtained from linguistic knowledge, in order to find an optimal set of weights — that is the combination of weights that gives the lowest word error rate. It is therefore desirable to have a data-driven machine learning method that finds a good, if not optimal, weighting of the feature streams.

For this purpose Florian Metze and I implemented the iterative version of the *Discriminative Model Combination (DMC)* developed by Peter Beyerlein [Bey98] [Bey00].

DMC is an approach that can be used to integrate multiple acoustic and/or language models into one log-linear posterior probability distribution. In this approach the different models are combined in a weighted sum at the log likelihood level. The weights of the sum are then optimized using a discriminative method.

So given a hypothesis k, a weight vector Λ and the feature vector x the posterior probability $p_\Lambda(k|x)$ is:

$$p_\Lambda(k|x) = C(\Lambda, x) exp \left\{ \sum_{j=1}^{M} \lambda_j \log p_j(k|x) \right\} \tag{5.1}$$

$C(\Lambda, x)$ is a constant necessary for normalization so that $p_\Lambda(k|x)$ really is a probability distribution. However since we are only interested in finding the hypothesis k with the highest probability, we will ignore C for the sake of simplicity, since it does not depend on k.

In our special case, with the combination of a standard model stream and the feature detector streams as described above, $p_0(k|x)$ is the posterior probability of k as given by the standard models, while the $p_1, \ldots, p_M$ are the posterior probabilities from the M feature detectors. This combination as a weighted sum at the log likelihood level is exactly how the stream based approach for integrating the feature streams works.

From the different methods that Beyerlein developed we implemented his iterative approach, called *Minimization of the Smoothed Word Error Rate (MWE)* that is based on the *Generalized Probabilistic Descent (GPD)* [JCL95].

MWE implements a gradient descent on a numerically estimated and smoothed word error rate function that is dependent on the weight vector Λ for the combination of the models. The estimation of the error function is necessary because the real error function over Λ is not known. And even if the error function would be given, since it maps the weight vector Λ, that is defined in $\mathbb{R}^n$, to the number of errors which is defined in $\mathbb{N}$, the derivative of the function for any Λ would be either not defined or zero. Therefore it is

necessary to smooth the empirical approximation of the error function. The smoothed approximation of the error function that is used for MWE is:

$$E_{MWE}(\Lambda) = \frac{1}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \sum_{k \neq k_n} \mathcal{L}(k, k_n) S(k, n, \Lambda) \tag{5.2}$$

In this equation the $k \neq k_n$ are all possible hypotheses, while the k_n $(n = 1 \dots N)$ are the N given training references for the discriminative training. $\mathcal{L}(k, k_n)$ is the Levenshtein-distance. $S(k, n, \Lambda)$ is an indicator function that is used for smoothing the Levenshtein-distance. If no smoothing is done, then S would be 1 if k is the hypothesis from the decoder, and 0 otherwise. In order to get a differentiable error function E_{MWE}, S is now set to be:

$$S(k, n, \Lambda) = \frac{p_\Lambda(k|x_n)^\eta}{\sum_{k'} p_\Lambda(k'|x_n)^\eta} \tag{5.3}$$

$p_\Lambda(k|x_n)$ is the posterior probability of hypothesis k, given the set of weights Λ and the internal model of the recognizer, for the feature vector x_n of the nth training utterance. η determines the amount of smoothing that is done by S. The higher η is the more accurately S describes the decision of the recognizer, and thereby the real error function. However η should not be chosen to be too large, in order to be able to numerically compute S. For my experiments I used $\eta = 3$.

For the estimation of E_{MWE}, equation 5.2 and 5.3 take into account all possible hypotheses k. This is clearly not feasible for the numerical computation of E_{MWE}. Therefore the set of hypotheses is limited to the most likeliest ones. In my experiments I used the hypotheses from an n-best list, where n was set to 150, that resulted from a lattice rescoring. The derivative of E_{MWE} is now:

$$\frac{\partial E_{MWE}(\Lambda)}{\partial \lambda_i} = \frac{\eta}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \sum_{k \neq k_n} S(k, n, \Lambda) \tilde{\mathcal{L}}(k, n, \Lambda) \log \frac{p_i(k|x_n)}{p_i(k_n|x_n)}$$

where

$$\tilde{\mathcal{L}}(k, n, \Lambda) = \mathcal{L}(k, k_n) - \sum_{k' \neq k_n} S(k', n, \Lambda) \mathcal{L}(k', k_n) \tag{5.4}$$

With this partial derivative one can construct a gradient descent:

$$\lambda_j^{(I+1)} = \lambda_j^{(I)} - \frac{\epsilon \eta}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \sum_{k \neq k_n} S(k, n, \Lambda^{(I)}) \tilde{\mathcal{L}}(k, n, \Lambda^{(I)}) \log \frac{p_j(k|x_n)}{p_j(k_n|x_n)} \tag{5.5}$$

Here ϵ is the learning rate, and has to be chosen carefully in order to adjust the change in the weights per iteration.

Also in my research I approximated the posterior probabilities with the likelihoods of the hypotheses that were returned by the decoder. Since in the case of the likelihoods the classification rule stays the same as with the posterior probabilities this does not change the update rules for the gradient descent.

Chapter 6

Decoding Experiments

With the methods for integrating the trained feature detectors with HMM based recognition systems and finding stream weights described in the previous chapter I performed a series of experiments. I did not only examine the monolingual case but also worked with multilingual and crosslingual scenarios showing that using AF detectors in such ways leads to comparable reductions in word error rate.

6.1 Initial Experiments

The speech recognition system used for the experiments in [MW02] had been trained on roughly 65 hours of original Broadcast News (BN) data and 35 hours of the English Verbmobil (ESST) data. The Broadcast News corpus contains the recordings of news broadcasts over radio and television networks, e.g. ABC, CNN, and NPR [Gra96]. ESST consists of spontaneous dialogues in the travel and scheduling domain collected during the Verbmobil project [Wah00]. The test data consists of 17 minutes of original BN texts read under clean conditions (ReadBN).

In order to be able to train and test articulatory feature detectors on the GlobalPhone corpus I ran an initial experiment on the English GlobalPhone test data using the acoustic models, the language model, and feature detectors from [MW02]. In a first step I decoded the GlobalPhone evaluation and cross validation data with default language model parameters and then added up to eight feature detectors in the order "OBSTRUENT SONORANT SIBILANT HIGH-VOW NASAL VLS-FR MH-DIP RETROFLEX Y-GLIDE BF-DIP" taken from the method "TREE" in [MW02]. Table 6.1 shows the word error rates (WER) for this system with zero, two, four, six, and eight of the above AF detectors. Using four AF detectors resulted in a relative reduction in WER of 8.7% on the development set (dev) compared to the system without feature detectors, while on the evaluation set (eval) the WER dropped by 4.3% relative when using two feature detectors.

These results are consistent with the ones reported in [MW02]. The reductions that I got are not quite as big as the ones Metze got on the ReadBN test set, however in my case I still have a mismatch between the training material of the acoustic models and the

#AF	dev set	eval set
baseline	19.6	20.8
2	18.4	19.9
4	17.9	20.0
6	18.1	20.4
8	18.1	20.8
best rel. reduction	8.7%	4.3%

Table 6.1: Results of the BN+ESST acoustic on the GP EN development and evaluation set with standard language model parameters

feature detectors on the one side and the test set on the other side. But the results prove the feasibility of applying the methods from [MW02] to the GlobalPhone corpus.

6.2 Decoding without Articulatory Features

At Carnegie Mellon University and Karlsruhe University phoneme HMM based large vocabulary continuous speech recognition systems have been developed for ten languages of the GlobalPhone corpus. The five languages that I chose for this research are included in these languages. So I used the recognizers for Chinese and English as a baseline against which to compare the new systems that resulted from enhancing those recognizers with articulatory features as described in chapter 5.

6.2.1 Training

The recognizers from which I took the baselines have all been developed solely on the GlobalPhone corpus. In order to facilitate the training and keep the development time within reasonable limits the training process was conducted in a largely automatized fashion.

The dictionaries were generated using letter-to-sound mapping tools and the language models were calculated on text resources fully automatically acquired from the World Wide Web. The acoustic models were initialized using a fast and efficient bootstrapping algorithm with the help of a four-lingual phoneme pool [SW97].

The acoustic models for each language consist of a fully continuous HMM system with 3000 sub-triphone and sub-quintphone models respectively. A sub-polyphone here refers to the begin, middle or end state of a divided polyphone. Each sub-polyphone is modelled by a mixture of 32 Gaussians, each Gaussian being 32-dimensional.

The feature vector is made up of 13 Mel-scale cepstral coefficients plus approximations of the first and second order derivatives as well as power and zero crossing rate. After cepstral mean subtraction the feature vector is reduced to 32 dimensions by a linear discriminant analysis (LDA). Note that this is the same feature extraction that we used for the training of the articulatory feature detectors (see chapter 4).

LID	CH	EN
dev	22.6%	13.1%
eval	28.8%	16.1%

Table 6.2: Word error rates of the English and German baseline systems on their development and evaluation sets

The sub-polyphone models were created with the use of a decision tree clustering procedure that uses an entropy gain based distance measure defined over the mixture weights of the Gaussians. The set of available questions consists of linguistically motivated questions about the phonetic context of a model [FR97].

The models were trained with several iterations of a label training. New labels (forced alignments) were always written after four training iterations.

6.2.2 Evaluation

First the language model parameters used for decoding were optimized on the development sets. Using these parameters, the final evaluation of the recognizers was done on the corresponding evaluation set. Table 6.2 shows the word error rate (WER) for the Chinese and English recognizers with the optimized language model parameters on their development (dev) and evaluation set (eval). These two recognizers are the baselines for the experiments in which I use feature detectors in addition to the standard models.

6.3 Decoding using AF streams and heuristic stream weights

In order to explore the possibilities of enhancing a standard HMM recognizer with streams of articulatory feature detectors as an additional knowledge source, I examined a monolingual, two crosslingual and one multilingual scenario for two languages. The experiments were performed on the English and Chinese development sets and their results are summarized in tables 6.3 and 6.4. In order to see whether the number of feature detectors that lead to the best results on the development set generalize, I also decoded the evaluation set using the best combination of feature detectors found on the development set. These results can be found in tables 6.5 and 6.6. As baselines serve the performances of the English and the Chinese recognizer when no AF detectors are used.

	AF LID			
#AF detectors	EN	GE	MM4	MM5
0	13.1%			
1	12.9%	**12.2%**	13.0%	12.8%
2	12.7%	12.3%	12.8%	12.8%
3	12.7%	13.0%	13.1%	12.8%
4	12.5%	20.0%	13.2%	12.7%
5	12.3%	36.1%	12.8%	12.3%
6	12.2%	43.8%	**12.6%**	**12.1%**
7	11.9%	85.1%	12.9%	12.2%
8	11.8%	94.3%	12.8%	12.2%
9	**11.7%**	98.1%	12.7%	12.3%
10	12.0%	99.5%	13.6%	12.4%
best rel. red.	10.8%	6.9%	3.8%	7.6%

Table 6.3: WER when decoding the EN development set using AF detectors as additional knowledge source in a monolingual, crosslingual, and multilingual scenario.

	AF LID			
#AF detectors	CH	JA	MM4	MM5
0	22.6%			
1	22.2%	22.3%	22.2%	22.2%
2	22.0%	22.0%	22.2%	22.1%
3	21.5%	21.8%	21.7%	21.7%
4	**21.3%**	**21.5%**	**21.6%**	**21.5%**
5	21.6%	21.8%	22.1%	21.8%
6	21.6%	21.9%	23.1%	22.2%
7	21.8%	22.0%	24.4%	23.0%
8	22.0%	22.5%	28.9%	24.9%
9	22.0%	22.8%	40.1%	27.7%
10	22.5%	23.3%	49.2%	32.5%
best rel. red.	5.8%	4.9%	4.4%	4.9%

Table 6.4: WER when decoding the CH development set using AF detectors as additional knowledge source in a monolingual, crosslingual, and multilingual scenario.

	AF LID			
#AF detectors	EN	GE	MM4	MM5
baseline	16.1%			
AF	14.1%	15.5%	15.4%	14.7%
best rel. red.	12.4%	3.7%	4.3%	8.7%

Table 6.5: WER when decoding the EN evaluation set using AF detectors as additional knowledge source in a monolingual, crosslingual, and multilingual scenario.

	AF LID			
#AF detectors	CH	JA	MM4	MM5
baseline	28.8%			
AF	28.2%	28.7%	28.2%	28.1%
best rel. red.	2.1%	0.3%	2.1%	2.4%

Table 6.6: WER when decoding the CH evaluation set using AF detectors as additional knowledge source in a monolingual, crosslingual, and multilingual scenario.

6.3.1 Monolingual case

English

For the English monolingual case I added English AF detectors to the recognizer to decode the English development set. The detectors were added in the order of their classification accuracy (see 4.1.3 and table A.2 in appendix A): POSTALVEOLAR, PALATAL, GLOTTAL, AFFRICATE, LABIODENTAL, LATERAL-APPROXIMANT, NASAL, ROUND, OPEN, VELAR.

The articulatory feature detectors for FLAP and DENTAL have been left out because these features are not attributed to any German phoneme. Therefore no German feature detectors exist for them. For the monolingual scenario this is not a problem, but by leaving out these two features the obtained results can be better compared to the crosslingual results reported below.

As we can see from the results, shown in the column 'EN' in table 6.3, the word error rate starts to drop as we add the AF detectors. It comes to a low when adding nine detectors. After that the error rate starts to rise again. Adding nine feature detectors yields a word error rate of 11.7%. Compared to the baseline this a reduction in WER of 10.8% relative.

Chinese

The column labeled 'CH' of table 6.4 shows the word error rates when adding Chinese feature detectors to the Chinese standard models for decoding the Chinese development

set.

As for English, I added the detectors in the order of their classification accuracy: LABIODENTAL, VOICED, APPROXIMANT, LATERAL-APPROXIMANT, BACK, FRICATIVE, PLOSIVE, OPEN, BILABIAL, CONSONANT (see 4.1.3 and table A.1 in appendix A). I left out the detectors for TONAL5 and ASPIRATED, because these features do not exist for Japanese. Again, leaving them out ensures the comparability of the monolingual results and the crosslingual results reported below.

Similar to the English case the word error rate starts to drop as I add more feature detectors. It reaches a minimum when adding four detectors, reducing the error rate by 5.8% relative.

6.3.2 Crosslingual case

English

For the crosslingual case I decided to run two experiments. For the first experiment I used the German feature detectors as additional streams, because the German detectors show the best average classification accuracy on English. Since in the monolingual case we left out the features that are not covered by German we can use the same order in which to add the feature detectors as in the English case.

The results in the column 'GE' of table 6.3 show that the word error rate starts to drop significantly right away and comes to a low after adding only one feature detector. After that the word error rate starts to rise again, significantly faster than in the monolingual case. So adding the feature detector for POSTALVEOLAR leads to a word error rate of 12.2%, a reduction of 6.9% relative to the baseline.

As a second crosslingual experiment I combined the English standard models with the MM4 feature detectors that were trained on the four languages without English. The MM4 detectors are trained using the language independent acoustic modeling technique 'Multilingual Mixed' (see 2.5). But since English, the language of the test set, is not part of the training material, the decoding experiments itself fall into the crosslingual category. Again I add the same feature detectors in the same order as in the monolingual case.

As we can see from the results in the column 'MM4' in table 6.3 the word error rate again drops as we add feature detectors. It comes to a low when adding six feature detectors, reducing the word error rate by 3.8% relative. When adding more features the error rate starts to rise again. The reduction in WER that we see here falls clearly short of what is possible in the monolingual case. But as we will see later, selecting the features and weights with DMC leads to an increase in performance that comes close to the monolingual gains.

Chinese

For the first crosslingual experiment involving the Chinese standard models on the Chinese development set I decided to add the Japanese feature detectors, because they show the

highest average classification accuracy on the Chinese test set besides the Chinese AF detectors. As in the monlingual case adding four feature detectors yields the lowest word error rate. It is reduced by 4.9% relative to 21.5%. When decoding with the aid of the MM4 detectors that where trained on all the selected languages except Chinese the minimum word error rate again is reached after adding four detectors. This time the relative reduction is 4.4%, the lowest reduction when it comes to fixed stream weights.

6.3.3 Multilingual case

For my experiments in decoding with the help of multilingual feature detectors I decided to use the feature detectors that were trained on all available languages. Since the language pool contains the five languages mentioned earlier, this leads to the use of the MM5 multilingual mixed feature detectors trained on the languages Chinese, English, German, Japanese, and Spanish. Again the experiments were performed with the English and the Chinese standard models on their respective development sets. The feature detectors were added in the same order as in the monolingual scenarios.

English

The results on the English development set can be seen in table 6.3 in the column labelled 'MM5'. As in the crosslingual case when using the MM4 detectors the highest reduction in word error rate can be achieved when using six feature detectors. But this time the relative reduction is 7.6%, and therefore higher than in the crosslingual case, but still not as good as in the monolingual scenario.

Chinese

The Chinese results are shown in table 6.4 in the column 'MM5'. Just as in the monolingual and crosslingual cases adding four feature detectors yields the lowest word error rate, giving a reduction of 4.9% relative to the baseline. This is just as good as if using the Japanese detectors.

6.3.4 Generalization

When comparing the best relative word error rate reductions on the Chinese and English development sets with the reductions on the evaluation sets, one can see that for the English case the number of AF detectors to add generalizes pretty well. For the monolingual case, for the MM4, and for the MM5 detectors the relative WER reduction is even better on the evaluation set than on the development set. Only when adding the German feature detector that gave the best performance on the development set the relative error rate reduction on the evaluation set is considerably less than on the development set.

For Chinese the number of feature detectors does not generalize that well. When adding the Japanese feature detectors almost no reduction in the word error rate can be seen (0.3%

relative). When adding the other detectors, the observed relative reduction is only roughly half as large as on the development set.

6.3.5 Conclusion

These first experiments show that the word error rate of a large vocabulary continuous speech recognition systems can be significantly reduced by using cross lingual and multilingual articulatory feature detectors. The weights with which the feature detectors and the HMM recognizers were combined were just educated guesses. The same is true for the selection of the set of features used. Though selecting feature detectors by means of their classification accuracy is probably a reasonable heuristic for our first experiments, the experiments in [MW02] indicate that also sets that contain feature detectors that rank among the worst when it comes to classification accuracy can bring the same reductions in word error rate.

So there should naturally be room for further improvement by finding a more suitable set of weights and features. For example, while for Chinese the number of feature detectors that gives the best performance improvement is always the same for the monolingual, crosslingual, and multilingual scenarios, in the English case this number varies from scenario to scenario. When combining the English standard models with the German feature detectors the minimum error rate is already reached after only adding one feature detector. An analysis of the scores produced by the German models on the English test data revealed that the average score of the German feature detectors was considerably higher than that of the English detectors. So using the same stream weights for the German detectors as for the English detectors produced hypotheses with a considerably higher average score than when using the English detectors. Since the IBIS decoder implements a beam search with constant beams this means that ceteris paribus the beams are effectively narrowed. This effect can explain that adding more than one feature detector produces the bad recognition results that one can see in table 6.3.

6.4 Decoding using AF and adapted stream weights

With the use of MWE, the iterative version of the Discriminative Model Combination (see 5.2.2), I calculated stream weights for the same scenarios as described in the last section. MWE was presented with all possible features that exist in both the language of the standard model stream and that of the articulatory feature streams. For the calculation of the smoothed word error rate function E_{MWE} the hypotheses from an n-best list are used. The n-best list was 150 hypotheses long and was obtained from a lattice rescoring. The smoothing factor η was set to 3.0. When using a higher η I got overflows during the calculation of the gradient.

The step width ϵ for the gradient descent was selected so that the maximum change of a single stream weight equaled a constant δ. For the monolingual case δ was initially set to 0.01. For the cross- and multilingual case δ was chosen to be 0.005. The smaller

	AF LID					
	EN		GE		MM4	
	dev	eval	dev	eval	dev	eval
baseline	13.1%	16.1%	13.1%	16.1%	13.1%	16.1%
DMC adapted weights	11.7%	14.4%	11.9%	15.1%	11.8%	14.8%
best rel. red.	10.8%	10.6%	9.2%	6.2%	9.9%	8.1%
	MM5		All			
	dev	eval	dev	eval		
baseline	13.1%	16.1%	13.1%	16.1%		
DMC adapted weights	11.9%	14.5%	11.5%	14.1%		
	9.2%	9.9%	12.2%	12.4%		

Table 6.7: WER when decoding the EN development set using AF detectors as additional knowledge source in a monolingual, crosslingual, and multilingual scenario and DMC adapted weights

	AF LID							
	CH		JA		MM4		MM5	
	dev	eval	dev	eval	dev	eval	dev	eval
baseline	22.6%	28.8%	22.6%	28.8%	22.6%	28.8%	22.6%	28.8%
DMC adapted weights	21.4%	28.3%	21.4%	28.6%	21.4%	27.9%	21.4%	28.3
best rel. red.	5.3%	1.7%	5.3%	0.7%	5.3%	3.9%	5.3%	1.7%

Table 6.8: WER when decoding the CH development set using AF detectors as additional knowledge source in a monolingual, crosslingual, and multilingual scenario and DMC adapted weights

δ was necessary to compensate for the higher average scores that the feature detectors gave when used across languages. After several iterations, when it seemed as if a local minimum was found, δ was decreased, and further iterations were calculated until no further improvements were seen.

During the search for weights it often happened that after several iterations the total score of the found hypotheses was significantly higher than that of the baseline. As explained before this effectively narrows the search beams. Therefore the weight for the standard model streams was lowered by hand in such cases, in order to decrease the scores of the hypotheses that the decoder found.

The utterances from the development set served as training set for the DMC. In order to see how well the weights found for the development set generalize, I decoded the evaluation set using the calculated stream weights. The results for the MWE adapted weights on the development and evaluation set are summarized in tables 6.7 and 6.8.

6.4.1 Monolingual case

For the monolingual case I could not find weights that yield a better performance on the development set than the ones that were found with the heuristic described in 6.3. For English I was able to exactly achieve the same word error rate as with the fixed stream weights. For Chinese the word error rate of the DMC adapted weights are 0.1% absolute higher than that of the fixed weights.

6.4.2 Crosslingual case

In the crosslingual case it was possible to find weights that give a higher reduction in word error rate than with the fixed stream weights.

English

Combining English standard models with German feature streams leads to a word error rate of 11.9% on the development set, a relative reduction of 9.2% compared to the baseline. Using fixed stream weights the error rate could only be reduced by 6.9% relative. Using the MM4 feature detectors the word error rate was reduced to 11.8%, a relative reduction of 9.9%. While in the case of fixed stream weights the relative reduction of 3.8%, that was achieved with the MM4 detectors, was considerably smaller than that of the German feature detectors, now, with the DMC adapted weights, the MM4 feature detectors match and even outperform the German detectors.

Chinese

Adapting the stream weights with MWE, the combination of Chinese standard models with Japanese or MM4 feature streams leads to a word error rate of 21.4%, a relative reduction of 5.3%. This also is an improvement over the fixed stream weights, that showed a relative WER reduction of 4.9% for the Japanese detectors, and 4.4% for the MM4 detectors.

6.4.3 Multilingual case

Adapting the weights for the MM5 streams with DMC also showed improvements compared to the fixed stream weights. On the English development set the word error rate was reduced to 11.9%, a relative reduction of 9.2%. With the fixed stream weights I only achieved a relative reduction of 7.6%.

When combining the Chinese standard models with the MM5 AF detectors and adapting the stream weights with DMC, the error rate dropped by 5.3% relative to 21.4%. With the fixed stream weights the error rate was reduced by 4.9% relative.

6.4.4 Complete Detector Set

In 4.2.1 I showed that combining the feature detectors from many languages can improve the average classification accuracy.

In order to see whether it is possible to utilize this effect for the combination of the standard models with the feature detectors I presented the feature detectors from all languages and the standard models from the English recognizer to the DMC.

The result is shown in the column 'All' in table 6.7. After several iterations of DMC we got hypotheses which average score was approximately 30% higher than that of the baseline. We therefore decided that it is justified to widen the search beam by 15% and still compare the results to the original baseline. As we can see it is possible to get a relative reduction in WER of 12.2%. This is the best reduction that we were able to achieve so far.

6.4.5 Learned Feature Weights

Tables 6.9, 6.10, and 6.11 show the feature weights as learned by the DMC for the different combinations of standard models and feature detectors. In these tables only features with a weight greater or equal than 0.00001 are shown.

For English as well as Chinese, DMC generally selects more feature detectors than the heuristic but assigns them smaller weights. For Chinese only approximately half as many features are selected as for English.

It is also remarkable that only Chinese and Spanish detectors are chosen, when the English standard models and the feature detectors from all languages are presented to the DMC. Neither the English detectors, which show the best classification accuracy on English, nor the German detectors, which show the best crosslingual performance on English, are selected.

The DMC also often selects the same features independent on which language(s) they have been trained. From the 24 features that were selected when combining English standard models and English feature detectors, 18 are also among the selected German detectors, 17 among the MM4, and also 17 among the selected MM5 detectors. When combining Chinese standard models with Chinese detectors, nine features are being selected. Of these nine features, seven are also among the selected Japanese detectors, again seven among the selected MM4 detectors, and six among the MM5 detectors. This overlap and the

AF LID

EN		GE	
Feature	Weight	Feature	Weight
standard models	0.60000	standard models	0.90165
AFFRICATE	0.02061	AFFRICATE	0.00811
APPROXIMANT	0.01613	ALVEOLAR	0.00003
BACK	0.02765	APPROXIMANT	0.00561
BILABIAL	0.03270	ASPIRATED	0.00011
CENTRAL	0.01757	BACK	0.00391
CLOSE	0.00058	BILABIAL	0.00020
CLOSE-MID	0.00879	CLOSE	0.00704
CONSONANT	0.00391	CLOSE-MID	0.00067
DENTAL	0.04785	CONSONANT	0.01118
FLAP	0.02847	DENTAL	0.00407
GLOTTAL	0.05009	FLAP	0.00304
LABIODENTAL	0.01890	FRICATIVE	0.00320
LATERAL-APPROXIMANT	0.01549	FRONT	0.00001
NASAL	0.00191	GLOTTAL	0.01057
OPEN	0.02349	LABIODENTAL	0.02340
OPEN-MID	0.02227	LATERAL-APPROXIMANT	0.00011
PALATAL	0.03478	NASAL	0.00015
PLOSIVE	0.03056	OPEN-MID	0.00445
POSTALVEOLAR	0.06919	PALATAL	0.00139
ROUND	0.02823	PLOSIVE	0.00086
UNVOICED	0.05961	POSTALVEOLAR	0.00233
VELAR	0.03079	RETROFLEX	0.00470
VOICED	0.02356	ROUND	0.01235
VOWEL	0.02314	VELAR	0.00539

MM4		MM5	
Feature	Weight	Feature	Weight
standard models	0.65009	standard models	0.68000
AFFRICATE	0.05515	AFFRICATE	0.02780
ALVEOLAR	0.00145	ALVEOLAR	0.00062
APPROXIMANT	0.01678	APPROXIMANT	0.01645
BILABIAL	0.01435	BILABIAL	0.01719
CENTRAL	0.00004	CLOSE	0.00773
CLOSE	0.00812	CLOSE-MID	0.00496
CLOSE-MID	0.00700	DENTAL	0.00007
DENTAL	0.00318	FLAP	0.01933
FLAP	0.03879	FRONT	0.00811
FRONT	0.00737	GLOTTAL	0.03064
GLOTTAL	0.02548	LABIODENTAL	0.04350
LABIODENTAL	0.03969	LATERAL-APPROXIMANT	0.00726
LATERAL-APPROXIMANT	0.00715	OPEN	0.00031
OPEN-MID	0.01898	OPEN-MID	0.00925
PALATAL	0.03780	PALATAL	0.02197
PLOSIVE	0.01157	PLOSIVE	0.00574
POSTALVEOLAR	0.03209	POSTALVEOLAR	0.02024
RETROFLEX	0.00525	ROUND	0.02471
ROUND	0.02358	VELAR	0.01071
VELAR	0.00153		

Table 6.9: Feature weighting as learned by the DMC on English

Feature	Weight
standard	0.76281
AFFRICATE_CH	0.00764
ALVEOLAR_CH	0.00614
APPROXIMANT_CH	0.00491
ASPIRATED_CH	0.00655
BACK_CH	0.00927
BILABIAL_CH	0.00778
CLOSE_CH	0.00794
CONSONANT_CH	0.00537
FRICATIVE_CH	0.00625
FRONT_CH	0.00325
LABIODENTAL_CH	0.00537
LATERAL-APPROXIMANT_CH	0.00969
NASAL_CH	0.00527
OPEN_CH	0.01075
OPEN-MID_CH	0.00655
PALATAL_CH	0.00577
PLOSIVE_CH	0.00451
RETROFLEX_CH	0.00920
ROUND_CH	0.00560
UNROUND_CH	0.00442
UNVOICED_CH	0.00666
VELAR_CH	0.00751
VOICED_CH	0.00224
VOWEL_CH	0.00556
AFFRICATE_SP	0.01316
ALVEOLAR_SP	0.01398
APPROXIMANT_SP	0.01101
BACK_SP	0.01465
BILABIAL_SP	0.01249
CLOSE_SP	0.01073
CLOSE-MID_SP	0.01253
CONSONANT_SP	0.01093
DENTAL_SP	0.01463
FLAP_SP	0.01329
FRICATIVE_SP	0.01267
FRONT_SP	0.00788
LABIODENTAL_SP	0.01273
LATERAL-APPROXIMANT_SP	0.01523
NASAL_SP	0.00649
OPEN_SP	0.01343
PALATAL_SP	0.01258
PLOSIVE_SP	0.01150
POSTALVEOLAR_SP	0.01284
ROUND_SP	0.01233
UNROUND_SP	0.00787
UNVOICED_SP	0.01568
VELAR_SP	0.01425
VOICED_SP	0.00958
VOWEL_SP	0.01105

Table 6.10: Feature selection and weighting as learned by the DMC on English when using the feature detectors from all languages

AF LID

CH		JA	
Feature	Weight	Feature	Weight
standard models	0.88000	standard models	0.84000
AFFRICATE	0.01348	AFFRICATE	0.01166
ALVEOLAR	0.00823	ALVEOLAR	0.00683
APPROXIMANT	0.00327	APPROXIMANT	0.00258
ASPIRATED	0.00891	BACK	0.02348
BACK	0.00853	BILABIAL	0.00079
LABIODENTAL	0.00887	CLOSE	0.00707
NASAL	0.00215	CONSONANT	0.00242
PLOSIVE	0.02756	FRICATIVE	0.00642
RETROFLEX	0.00449	LABIODENTAL	0.00259
		NASAL	0.00141
		PALATAL	0.00497
		PLOSIVE	0.00206
		ROUND	0.02914
		UNVOICED	0.00770
		VOICED	0.00630
		VOWEL	0.00634

MM4		MM5	
Feature	Weight	Feature	Weight
standard models	0.89000	standard models	0.86750
AFFRICATE	0.00273	AFFRICATE	0.00361
ALVEOLAR	0.00088	ALVEOLAR	0.00149
APPROXIMANT	0.00605	APPROXIMANT	0.01024
ASPIRATED	0.00733	ASPIRATED	0.01173
BACK	0.01415	BACK	0.01803
CLOSE	0.00662	CLOSE	0.00028
FRICATIVE	0.00413	FRICATIVE	0.00651
LABIODENTAL	0.00085	RETROFLEX	0.03466
RETROFLEX	0.01339	ROUND	0.02784
ROUND	0.01817	TONAL3	0.00140
UNVOICED	0.00200	UNVOICED	0.00361

Table 6.11: Feature weighting as learned by the DMC on Chinese

fact, that only a portion of the available feature detectors is actually used, is a strong indication that articulatory features contain specific information important to decoding human speech. For example all seven of the nine Chinese feature detectors that DMC combined with the Chinese standard models and that also occur in Japanese are chosen by the DMC when Chinese models are combined with Japanese detectors. Therefore these seven features seem to contain useful information for the speech recognizer that does not get lost when the training language is changed.

6.4.6 Conclusion

Using DMC it is possible to find suitable weights for the flexible stream based approach described in 5.1 in a data-driven way. In the monolingual case the found weights do not perform better than the ones that were chosen by the heuristic described in 5.2.1. For English the same reduction in WER can be achieved, in the case of Chinese the performance of the DMC adapted weights on the development set is 0.1% absolute worse than that of the fixed stream weights.

However for the cross- and multilingual scenarios further reductions in the word error rate over the fixed stream weights were seen for both English and Chinese. When testing the combination of Chinese standard models and cross- or multilingual feature detectors on the development set, the performance of the DMC adapted stream weights is only 0.1% absolute worse than the best known combination of Chinese standard models and Chinese feature detectors.

For English the found stream weights generalize well. However for Chinese the generalization is very poor. During the course of the experiments I found out that the Chinese development and evaluation sets contain serious transcription errors where the audio file and corresponding transcription do not match at all. However it was not possible to remove all errors during the course of this research project due to time constraints and a lack of expert knowledge in Chinese. Errors in the transcription are very bad for a discriminative training. I therefore suspect that the transcription errors are the reason why the stream weights for Chinese generalize so badly.

Chapter 7

Conclusion

7.1 Summary

Making use of the GlobalPhone unit set and the mapping of articulatory features to phonemes introduced by the International Phonetic Association I trained monolingual and all combinations of multilingual articulatory feature detectors on five languages. Using these detectors I showed that articulatory features can be reliably detected for a variety of languages. More important, I demonstrated that models for articulatory features are robust to inter language variability. I successfully detected features across languages. I further showed that combining feature detectors from many languages outperforms monolingual detectors in terms of classification accuracy.

I made use of a flexible stream architecture to combine standard models with feature detectors for Chinese and English. In a first experiment I used a heuristic to select the necessary stream weights and to show the potential of reducing the word error rate when using cross- and multilingual feature detectors. For a second set of experiments I implemented a discriminative training technique called 'Discriminative Model Combination' to find suitable sets of stream weights. For the combination of standard models with cross- and multilingual detectors DMC showed improvements over the heuristical stream weight selection. For the monolingual case DMC only matched the heuristical method. But with cross- and multilingual detectors and DMC adapted stream weights I was able to achieve the same reductions in word error rate as in the monolingual case. For English the word error rate was reduced by up to 12.4% relative.

7.2 Future Work

The results of the experiments in this work are encouraging that with the use of multilingual articulatory features it is possible to better address the problem of rapid deployment of speech recognition systems in new target languages, to improve multilingual acoustic modelling, or to apply articulatory features to speaker adaptation.

The monolingual results of the two stream weight selection methods that I used in this

work show however that there is still a need for a better method for selecting stream weights. With such a method it will be possible to take the step to context dependent stream weights for articulatory features. In that way articulatory features will really enable us to leave the 'beads-on-a-string' model of speech. This will lead to a more accurate model that will be better able to capture the acoustic variations encountered in spontaneous speech, and that will be more robust to adverse environmental conditions.

Appendix A

Results of the Monolingual AF Detectors

A.1 Chinese AF Detectors

AF	CH	EN	GE	JA	SP
LABIODENTAL	98.46%	96.77%	93.09%	96.36%	98.02%
VOICED	97.73%	83.16%	85.66%	89.51%	84.79%
APPROXIMANT	97.53%	91.80%	95.12%	92.46%	93.02%
TONAL5	97.02%	—	—	—	—
LATERAL-APPROXIMANT	96.72%	92.39%	92.90%	92.42%	88.44%
BACK	96.39%	90.12%	95.28%	72.63%	90.87%
FRICATIVE	96.31%	88.83%	90.95%	93.07%	85.30%
PLOSIVE	96.27%	88.04%	90.75%	89.63%	84.89%
OPEN	95.64%	95.45%	92.96%	90.53%	89.88%
ASPIRATED	95.46%	90.79%	—	—	—
BILABIAL	94.95%	91.05%	86.65%	88.63%	90.90%
CONSONANT	94.87%	85.03%	87.23%	85.23%	71.20%
VOWEL	94.81%	84.65%	87.51%	84.83%	70.42%
NASAL	94.78%	91.53%	90.27%	87.37%	79.65%
ROUND	94.70%	93.48%	93.53%	87.85%	90.19%
AFFRICATE	94.58%	88.47%	92.49%	91.19%	86.94%
UNVOICED	94.51%	80.66%	83.26%	85.53%	76.32%
PALATAL	94.19%	87.48%	91.35%	86.79%	87.77%
CLOSE	94.10%	92.88%	89.16%	84.40%	91.65%
OPEN-MID	93.31%	84.94%	88.29%	—	—
RETROFLEX	91.57%	83.69%	—	—	—
VELAR	91.29%	88.48%	82.84%	82.79%	82.70%
TONAL3	90.39%	—	—	—	—
FRONT	89.71%	78.98%	80.98%	79.67%	70.23%
UNROUND	89.04%	78.52%	79.06%	76.29%	69.26%
TONAL2	88.30%	—	—	—	—
ALVEOLAR	87.62%	70.96%	71.83%	78.35%	65.23%
TONAL1	87.54%	—	—	—	—
TONAL4	84.37%	—	—	—	—

A.2 English AF Detectors

AF	EN	CH	GE	JA	SP
POSTALVEOLAR	99.25%	—	98.67%	96.38%	94.92%
PALATAL	99.00%	89.90%	96.13%	97.16%	96.85%
GLOTTAL	98.84%	—	97.22%	96.64%	—
FLAP	98.84%	—	—	—	94.50%
AFFRICATE	98.63%	91.01%	96.74%	96.21%	99.44%
LABIODENTAL	97.99%	98.98%	94.26%	98.08%	97.95%
LATERAL-APPROXIMANT	97.39%	91.74%	91.06%	90.74%	88.90%
DENTAL	97.04%	—	—	—	91.65%
NASAL	96.66%	90.82%	94.49%	93.19%	91.76%
ROUND	95.55%	90.72%	91.09%	85.59%	88.70%
OPEN	95.54%	92.87%	94.94%	89.69%	88.07%
VELAR	95.48%	87.86%	91.27%	91.59%	92.18%
RETROFLEX	95.28%	90.14%	—	—	—
FRICATIVE	94.71%	91.88%	89.12%	91.93%	90.59%
BILABIAL	93.86%	94.81%	89.41%	91.63%	92.08%
ASPIRATED	93.81%	90.99%	—	—	—
APPROXIMANT	93.79%	92.22%	92.08%	93.88%	93.69%
CLOSE	93.40%	87.57%	88.07%	85.02%	88.48%
PLOSIVE	92.99%	89.00%	89.74%	89.29%	84.09%
BACK	91.38%	85.65%	85.36%	76.49%	86.64%
VOWEL	91.08%	86.24%	87.34%	86.76%	78.81%
CONSONANT	91.03%	85.64%	87.44%	85.47%	78.34%
OPEN-MID	90.92%	87.32%	87.59%	—	—
UNVOICED	90.46%	83.69%	83.88%	86.35%	84.72%
CLOSE-MID	89.87%	—	83.14%	76.59%	80.45%
FRONT	89.65%	78.35%	82.99%	85.19%	77.67%
VOICED	89.31%	81.36%	83.52%	86.23%	84.54%
CENTRAL	88.81%	—	86.50%	—	—
ALVEOLAR	87.43%	70.59%	72.14%	71.78%	73.97%
UNROUND	86.76%	76.29%	84.10%	79.84%	78.49%

A.3 German AF Detectors

AF	GE	CH	EN	JA	SP
POSTALVEOLAR	99.46%	—	98.20%	96.60%	94.29%
APPROXIMANT	98.86%	96.87%	93.02%	95.60%	95.97%
AFFRICATE	98.31%	90.33%	97.45%	95.59%	98.80%
PALATAL	98.20%	90.98%	97.83%	94.52%	95.13%
GLOTTAL	97.90%	—	96.06%	94.72%	—
OPEN-MID	97.11%	89.50%	92.17%	—	—
OPEN	95.36%	88.64%	93.77%	88.40%	85.68%
BACK	95.00%	93.99%	89.69%	75.11%	90.35%
LABIODENTAL	94.39%	97.92%	95.13%	94.95%	96.23%
NASAL	94.11%	89.54%	91.14%	89.90%	86.22%
LATERAL-APPROXIMANT	93.97%	94.71%	95.60%	90.23%	89.01%
FRICATIVE	93.94%	90.75%	82.82%	92.09%	85.99%
PLOSIVE	93.81%	92.40%	87.67%	87.69%	78.22%
ROUND	93.79%	92.26%	94.29%	88.25%	90.08%
TRILL	93.46%	—	—	—	85.13%
VOICED	92.36%	83.02%	75.46%	83.75%	75.73%
UNVOICED	91.77%	84.79%	73.99%	82.81%	74.79%
VOWEL	91.77%	86.98%	75.09%	77.23%	63.79%
CONSONANT	91.06%	85.75%	73.07%	77.86%	65.53%
VELAR	90.66%	87.05%	91.74%	87.20%	84.90%
FRONT	90.41%	77.31%	81.27%	83.85%	70.62%
CENTRAL	89.88%	—	91.63%	—	—
BILABIAL	89.27%	95.43%	93.15%	92.52%	91.32%
UNROUND	89.15%	77.20%	77.70%	76.73%	70.00%
CLOSE	89.01%	87.29%	88.78%	81.21%	83.10%
CLOSE-MID	86.74%	—	90.98%	78.46%	76.15%
ALVEOLAR	79.54%	75.84%	67.63%	69.89%	57.36%

A.4 Japanese AF Detectors

AF	JA	CH	EN	GE	SP
LABIODENTAL	99.23%	98.70%	95.88%	94.50%	98.23%
PALATAL	97.96%	89.47%	97.23%	94.09%	95.03%
POSTALVEOLAR	97.71%	—	95.00%	96.11%	91.15%
GLOTTAL	97.50%	—	96.15%	91.85%	—
OPEN	97.23%	87.19%	86.99%	91.12%	91.66%
APPROXIMANT	96.99%	94.68%	91.69%	94.43%	94.89%
AFFRICATE	96.80%	93.88%	94.50%	96.60%	97.51%
ROUND	96.61%	88.25%	87.74%	90.00%	91.32%
LATERAL-APPROXIMANT	96.59%	96.02%	92.83%	92.26%	91.40%
UVULAR	96.24%	—	—	—	—
FRICATIVE	95.40%	91.96%	89.00%	90.41%	87.17%
FRONT	95.25%	77.47%	85.31%	79.78%	76.62%
BILABIAL	94.94%	96.61%	92.70%	91.80%	93.07%
NASAL	94.84%	90.63%	93.35%	92.74%	89.70%
PLOSIVE	94.72%	92.63%	87.44%	87.65%	90.82%
VOICED	94.68%	84.94%	83.86%	83.96%	87.82%
VELAR	94.40%	85.78%	91.35%	88.71%	90.92%
UNVOICED	94.00%	85.81%	83.89%	84.35%	87.58%
CLOSE-MID	93.78%	—	83.61%	82.46%	83.44%
CONSONANT	93.73%	85.48%	81.48%	80.37%	81.68%
VOWEL	93.53%	83.95%	81.81%	79.50%	80.53%
BACK	93.37%	68.05%	74.97%	71.58%	76.77%
CLOSE	92.56%	82.74%	88.92%	82.17%	82.53%
UNROUND	92.48%	73.12%	77.01%	77.02%	77.10%
ALVEOLAR	89.92%	81.86%	70.95%	69.08%	72.97%

A.5 Spanish AF Detectors

AF	SP	CH	EN	GE	JA
AFFRICATE	99.33%	91.45%	95.56%	96.76%	95.05%
LABIODENTAL	98.84%	98.78%	96.35%	94.14%	97.67%
TRILL	98.12%	—	—	91.43%	—
APPROXIMANT	97.05%	89.84%	90.41%	92.69%	95.26%
DENTAL	96.49%	—	96.68%	—	—
PALATAL	96.43%	84.35%	94.56%	91.63%	94.04%
VOICED	95.84%	87.00%	83.76%	78.70%	90.18%
UNVOICED	94.92%	86.04%	83.28%	77.77%	88.08%
CLOSE	94.39%	88.64%	91.52%	87.79%	81.48%
NASAL	94.07%	87.59%	91.04%	87.75%	91.45%
FRICATIVE	93.91%	88.71%	87.07%	83.03%	89.26%
PLOSIVE	93.43%	86.39%	84.43%	77.19%	83.12%
BACK	93.31%	88.31%	85.43%	86.87%	80.85%
ROUND	93.31%	87.13%	88.28%	85.99%	92.89%
LATERAL-APPROXIMANT	93.26%	89.81%	89.94%	88.63%	92.49%
FLAP	93.07%	—	87.90%	—	—
POSTALVEOLAR	93.07%	—	86.59%	86.19%	90.79%
OPEN	92.87%	86.90%	83.30%	89.93%	95.42%
BILABIAL	92.81%	90.89%	90.06%	84.81%	84.64%
VELAR	92.00%	81.29%	85.86%	81.00%	82.86%
CONSONANT	90.76%	73.45%	76.43%	70.06%	82.68%
VOWEL	90.47%	70.13%	77.53%	67.37%	82.75%
UNROUND	90.42%	72.13%	79.05%	74.11%	84.17%
FRONT	90.42%	71.88%	75.88%	77.97%	80.94%
CLOSE-MID	87.98%	—	82.91%	78.22%	86.65%
ALVEOLAR	83.34%	79.18%	75.33%	69.41%	75.71%

Appendix B

Results of the Multilingual AF Detectors

B.1 MM2 Detectors

	Test Set					
AF LID	CH	DE	EN	JA	SP	TRAIN
CH_DE	92,92%	91,23%	89,14%	87,10%	83,74%	92,75%
CH_EN	93,09%	89,44%	91,61%	87,88%	86,93%	93,39%
CH_JA	92,36%	88,16%	87,71%	92,35%	86,13%	93,05%
CH_SP	92,34%	87,16%	87,60%	87,31%	88,85%	91,48%
DE_EN	89,17%	92,16%	92,76%	88,25%	86,45%	92,90%
DE_JA	89,43%	91,25%	87,85%	93,30%	86,28%	92,72%
DE_SP	88,49%	91,40%	87,75%	88,46%	89,76%	90,56%
EN_JA	89,17%	88,66%	92,17%	93,95%	87,70%	93,87%
EN_SP	87,37%	87,61%	93,13%	89,38%	91,96%	92,37%
JA_SP	86,75%	86,11%	88,59%	93,97%	91,13%	93,46%

B.2 MM3 Detectors

	Test Set					
	CH	DE	EN	JA	SP	Train
CH_DE_EN	92,83%	90,87%	91,32%	87,60%	86,54%	93,11%
CH_DE_JA	92,20%	89,92%	88,51%	91,31%	85,66%	92,39%
CH_DE_SP	91,87%	89,89%	89,17%	88,54%	89,79%	92,14%
CH_EN_JA	92,34%	88,74%	91,07%	92,05%	87,37%	93,10%
CH_EN_SP	91,82%	88,43%	90,60%	88,73%	90,53%	92,47%
CH_JA_SP	91,14%	87,19%	88,93%	91,49%	89,86%	92,38%
DE_EN_JA	89,88%	90,87%	91,82%	92,77%	87,36%	92,91%
DE_EN_SP	88,72%	90,83%	91,88%	89,37%	91,03%	92,03%
DE_JA_SP	88,32%	89,96%	88,15%	92,37%	90,83%	92,06%
EN_JA_SP	87,91%	87,49%	91,66%	92,73%	91,21%	93,04%

B.3 MM4 Detectors

	Test Set					
AF LID	CH	DE	EN	JA	SP	Train
CH_DE_EN_JA	91,82%	90,05%	91,04%	91,45%	87,06%	92,66%
CH_DE_EN_SP	91,56%	89,87%	90,68%	88,02%	88,48%	91,99%
CH_DE_JA_SP	91,07%	89,30%	88,27%	90,99%	88,16%	91,93%
CH_EN_JA_SP	91,28%	88,04%	90,81%	91,65%	89,05%	92,58%
DE_EN_JA_SP	89,51%	90,05%	91,78%	92,25%	89,49%	92,37%

B.4 MM5 Detectors

	Test Set					
AF LID	CH	DE	EN	JA	SP	Train
CH_DE_EN_JA_SP	90,36%	89,00%	90,22%	90,77%	88,29%	91,32%

Bibliography

[Ass99] International Phonetic Association. *Handbook of the International Phonetic Association*. Cambridge University Press, 1999. vii, 6, 12, 13

[Bey98] Peter Beyerlein. Discriminative Model Combination. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 481–484, Seattle, Washington, USA, May 1998. 34

[Bey00] Peter Beyerlein. *Diskriminative Modellkombination in Spracherkennungssystemen mit gross em Wortschatz*. PhD thesis, Fakultät fur Mathematik, Informatik und Naturwissenschaften der Rheinisch-Westfälischen Technischen Hochschule zu Aachen, Aachen, Germany, October 2000. d, 4, 34

[CGW01] Shuangyu Chang, Steven Greenberg, and Mirjam Webster. An Elitist Approach to Articulatory-Acoustic Feature Classification. In *Proceedings of the Seventh European Conference on Speech Communication and Technology*, pages 1725–1728, Aalborg, Denmark, September 2001. c, 16

[DS94] Li Deng and Don X. Sun. A Statistical Approach to Automatic Speech Recognition Using the Atomic Speech Units Constructed from Overlapping Articulatory Features. *Journal of the Acoustical Society of America*, 95(5):2702–2719, May 1994. c, 16, 31

[Eid01] Ellen Eide. Distinctive Features for Use in an Automatic Speech Recognition System. In *Proceedings of the Seventh European Conference on Speech Communication and Technology*, pages 1613–1617, Aalborg, Denmark, September 2001. c, 16, 31

[Ell97] T. Ellbogen. Phonetik seminar. Internet, http://www.phonetik.uni-muenchen.de/MUSE/Seminare/PHON_Einf/anatomie, 1997. vii, 14

[FGH+97] Michael Finke, Petra Geutner, Hermann Hild, Thomas Kemp, Klaus Ries, and Martin Westphal. The Karlsruhe-VERBMOBIL Speech Recognition Engine. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 83–86, Munich, Germany, 1997. 3

[FR97] Michael Finke and Ivica Rogina. Wide Context Acoustic Modeling in Read vs. Spontaneous Speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1743–1746, Munich, Germany, 1997. 39

[Gra96] David Graff. The 1996 Broadcast News Speech and Language-Model Corpus. In *Proceedings of the 1997 DARPA Speech Recognition Workshop*, Westfields International Conference Center, Chantilly, Virginia, February 1996. DARPA. 37

[Hie93] J. L. Hieronymus. ASCII Phonetic Symbols for the World's Languages: Worldbet. *Journal of the International Phonetics Association*, (23), 1993. 6

[JCL95] B. H. Juang, W. Chou, and C.H. Lee. *Statistical and Discriminative Methods for Speech Recognition and Coding - New Advances and Trends.* Springer Verlag, Berlin-Heidelberg, 1995. 34

[KFS00] K. Kirchhoff, G. A. Fink, and G. Sagerer. Conversational Speech Recognition Using Acoustic and Articulatory Input. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, 2000. IEEE. c

[Kir98] Katrin Kirchhoff. Combining Articulatory and Acoustic Information for Speech Recognition in Noisy and Reverberant Environments. In *Proceedings of the Fifth International Conference on Spoken Language Processing*, pages 891–894, December 1998. c, 16, 31

[Kir99] Katrin Kirchhoff. *Robust Speech Recognition Using Articulatory Information.* PhD thesis, Technische Fakultät der Universität Bielefeld, Bielefeld, Germany, June 1999. 32

[Kir00] Katrin Kirchhoff. Integrating Articulatory Features into Acoustic Models for Speech Recognition. In *Proceedings of the Workshop on Phonetics and Phonology in ASR. Parameters and Features, and their Implications*, Saarbrücken, Germany, March 1-3 2000. 16

[LWL+97] Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan. JANUS III: Speech-to-Speech Translation in Multiple Languages. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, 1997. 3

[MW02] Florian Metze and Alex Waibel. A Flexible Stream Architecture for ASR Using Articulatory Features. In *Proceedings of the 7th International Conference On Spoken Language Processing*, pages 2133–2136, Denver, Colorado, USA, September 2002. c, 4, 21, 22, 31, 32, 33, 37, 38, 44

[Ost99] M. Ostendorf. Moving Beyond the 'Beads-On-A-String' Model of Speech. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, volume 1, page 79, Keystone, Colorado, USA, December 12-15 1999. c, 15, 16

[PFG+99] David S. Pallet, Jonathan G. Fiscus, John S. Garofolo, Alvin Martin, and Mark Przybocki. 1998 Broadcast News Benchmark Test Results. In *Proceedings of the DARPA Broadcast News Workshop*, Herndon, Virginia, Feburary 28 - March 3 1999. 15

[Rog01] Ivica Rogina. Vorlesung: Sprachliche mensch-maschine-kommunikation. http://isl.ira.uka.de/sprachVorlesung, 2001. vii, 15

[SMFW01] Hagen Soltau, Florian Metze, Christian Fügen, and Alex Waibel. A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Italy, December 2001. 4

[SW97] Tanja Schultz and Alex Waibel. Fast Bootstrapping of LVSCR Systems with Multilingual Phoneme Sets. In *Proceedings of the Fifth European Conference on Speech Communication and Technology*, volume 1, pages 371–374, Rhodes, Greece, September 22-25 1997. 38

[SW01] Tanja Schultz and Alex Waibel. Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition. *Speech Communication*, 35(1-2):31–51, August 2001. c, v, vii, 7, 8, 9, 10

[SWW97] Tanja Schultz, Martin Westphal, and Alex Waibel. The GlobalPhone Project: Multilingual LVCSR with JANUS-3. In *Proceedings of the 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, Pilzen, Czech Republic, 1997. 5

[Wah00] W. Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translation.* Artificial Intelligence. Springer, Heidelberg, 2000. 37

[Wel89] C. J. Wells. Computer-Coded Phonemic Notation of Individual Languages of the European Community. *Journal of the International Phonetic Association*, 19:32–54, 1989. 6

[WGC01] Mirjam Wester, Steven Greenberg, and Shuangyu Chang. A Dutch Treatment of an Elitist Approach to Articulatory-Acoustic Feature Classification. In *Proceedings of the Seventh European Conference on Speech Communication and Technology*, pages 1729–1732, Aalborg, Denmark, September 2001. c, 16