Acoustic Modelling for Under-Resourced Languages

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften
der Fakultät für Informatik
der Universität Fridericiana zu Karlsruhe (TH)

genehmigte

Dissertation

von

Sebastian Stüker

aus Detmold

Tag der mündlichen Prüfung:	23. Juli 2009
Erster Gutachter:	Prof. Dr. Alexander Waibel
Zweiter Gutachter:	Prof. Dr. Tanja Schultz

Abstract

Over the past decades research in the field of automatic speech recognition has lead to systems with a sufficiently high grade of maturity that makes them suitable for use in real-life applications. However, such recognition systems have been developed only for very few languages. Languages addressed are mainly those with a large population, a high economic power, or for which a high political interest exists. For the vast majority of the 4,000-7,000 languages in the world no well performing speech recognition systems exist.

Languages are dying at a rapid rate. Linguists estimate that up to 90% of today's languages may go extinct within a few generations. Often languages die, because their speakers abandon them in favor of a more wide-spread language, from which they expect more economic or cultural advantages. We believe that technology can play a role in stopping this trend, if it were to provide natural language processing technologies, including automatic speech recognition, for all languages in the world.

The traditional way of training speech recognition systems for a new language requires the collection of large amounts of transcribed audio recordings, text resources, and the creation of a pronunciation dictionary. Considering the vast number of languages in the world in combination with the fact that most of them are only spoken by comparatively few speakers leads to the conclusion that this approach is not feasible when wanting to address all languages in the world.

The pronunciation dictionary is a central component of a speech recognition system which is time-consuming and expensive to create. In this thesis we show that the use of graphemes instead of the traditionally used phonemes is a feasible approach to speech recognition for many languages in the world. The use of graphemes instead of phonemes eliminates the need for a pronunciation dictionary, and thus significantly eases the creation of a recognition system for a new language. Part of the knowledge previously encoded in the pronunciation dictionary now needs to be learned by the context cluster tree of the acoustic model. We therefore also examine the use of a more flexible cluster tree for grapheme based speech recognition.

In order to reduce the amount of transcribed audio data that is needed for the training of a speech recognition system, past research has developed methods for porting phoneme based speech recognition systems to new languages with the help of multilingual models. In this thesis we transfer this work to the notion of grapheme based recognition systems. We show that it is possible to train multilingual recognition systems using graphemes instead of phonemes. We further demonstrate that the multilingual systems can be used to initialize the acoustic model of a new language. Since the graphemes of the languages in the world are more diverse than the phonemes, we demonstrate two data driven approaches for applying a multilingual system to a new language.

Past research has shown that articulatory features can be reliably recognized across languages and that they can be modelled in a multilingual way. Past research has also developed ways of integrating models for articulatory features into an HMM based recognition system based on phoneme models. In this thesis we have used models for articulatory features for improving the performance when porting phoneme based recognition systems to new languages.

Linguists estimate that the vast majority of languages in the world is without a writing system. For the case that a speech recognition system in such a new language needs to be created, we examined the automatic discovery of word-like units in a new language. We treated the case that the speech recognition system is part of a speech translation system and that only an unsegmented, phonetic transcript of the training data in the new language is available. In our discovery algorithm we made use of all available knowledge, including an existing translation of the training material and compared it against a word discovery scheme which only makes use of the monolingual, unsegmented phoneme string. Taking into account the parallel data lead to clear improvements over the case that only the monolingual data was used.

Kurzfassung

In den letzen Jahrzehnten hat die Forschung auf dem Gebiet der automatischen Spracherkennung Systeme von hinreichender Güte für die Verwendung in der Praxis hervorgebracht. Jedoch wurden solche Spracherkennungssysteme nur für eine sehr beschränkte Anzahl von Sprachen entwickelt. Betrachtet wurden hauptsächlich Sprachen mit entweder einer hohen Anzahl an Sprechern, mit hoher Wirtschaftsleistung oder solche, die von politischer Relevanz sind. Für die große Mehrzahl der 4.000 bis 7.000 Sprachen in der Welt wurden bisher keine gut funktionierenden Spracherkennungssysteme entwickelt.

Sprachen sterben kontinuierlich aus, mit einer besorgniserregenden Geschwindigkeit. Linguisten schätzen, dass innerhalb weniger Generationen 90% der heutigen Sprachen ausgestorben sein werden. Sprachen sterben häufig, weil ihre Sprecher sie zu Gunsten einer anderen Sprache aufgeben, von der sie sich materielle oder kulturelle Vorteile erhoffen. Wir glauben, dass der Einsatz von Technik helfen kann, diesen Trend zu stoppen, wenn es gelingt, Sprachverarbeitungssysteme, einschließlich Systeme zur automatischen Spracherkennung, für alle Sprachen in der Welt zur Verfügung zu stellen.

Der traditionelle Ansatz zum Training von Spracherkennern beinhaltet das Sammeln großer Mengen transkribierter Audiodaten, sowie die Erstellung eines phonetischen Aussprachewörterbuchs für die Zielsprache. Bedenkt man die hohe Anzahl an Sprachen in der Welt, sowie die Tatsache, dass die meisten von ihnen nur über verhältnismäßig wenig Sprecher verfügen, so wird klar, dass dieser traditionelle Ansatz nicht geeignet ist, um Erkennungssysteme für alle Sprachen der Welt zu trainieren, da er zu zeit- und kostenintensiv ist. In dieser Arbeit haben wir daher Methoden untersucht, um den Aufwand zur Erstellung eines

Spracherkennungssystems in einer neuen Sprache signifikant zu reduzieren.

Graphembasierte Akustische Modellierung

Zentraler Bestandteil eines Spracherkennungssystems ist das phonetische Aussprachewörterbuch. Sein Entwurf ist sehr zeitintensiv und teuer, und erfordert häufig die Mithilfe eines Experten der Zielsprache. Daher zeigen wir in dieser Arbeit, dass die Verwendung von Graphemen, statt der normalerweise genutzten Phoneme, als Modellierungseinheiten in Spracherkennungssystemen eine brauchbare Alternative ist. Durch die Verwendung von Graphemen anstelle von Phonemen entfällt die Notwendigkeit eines Aussprachewörterbuchs und die Entwicklung eines Spracherkennungssystems in einer neuen Sprache wird deutlich vereinfacht.

Unsere Experimente zeigen, dass Spracherkennungssysteme, die auf Graphemen beruhen, eine ähnlich gute Erkennungsleistung erbringen, wie phonembasierte. Dabei hängt die Differenz der Leistung im Vergleich zu phonembasierten Erkennern von der betrachteten Sprache und deren Verhältnis von Schrift zu Aussprache ab. Das Wissen, das bei phonembasierten Systemen im Aussprachewörterbuch enthalten ist, muss bei graphembasierten Systemen durch das akustische Modell, inklusive des Kontextclusterbaums, gelernt werden. Deshalb haben wir in unseren Experimenten die Verwendung eines flexiblen Clusterbaums für die graphembasierte Spracherkennung untersucht, der das Verhältnis von Schrift zu Sprache besser erlernen kann, als es der sonst verwendete Baum kann. Mit Hilfe des flexiblen Baums konnten wir Gewinne für alle betrachteten Sprachen nachweisen.

Graphembasierte Multilinguale und Crosslinguale Akustische Modellierung

Um die Menge der transkribierten Audiodaten, die zum Training eines Spracherkennungssystems in einer neuen Sprache benötigt werden, zu reduzieren, wurden in der Vergangenheit Verfahren entwickelt, um phonembasierte Spracherkennungssystem mit Hilfe multilingualer akustischer Modelle schnell auf neue Sprachen zu portieren. In unserer Arbeit zeigen wir, dass es auch für graphembasierte Erkenner möglich ist, multilinguale akustische Modelle zu trainieren. Wir zeigen ferner, dass mit Hilfe dieser Modelle, die akustischen Modelle einer neuen Sprache initialisiert werden können. Durch die Verwendung geringer

Mengen an Adaptionsmaterial zeigen wir, wie ein initiales Erkennungssystem mit akzeptabler Erkennungsleistung mit Hife der multilingualen Modelle erstellt werden kann. Wegen der Bedeutung des Clusterbaums haben wir auch die Verwendung der bekannten Polyphone Decision Tree Specialization untersucht und sie mit einem Baumbeschneidungsverfahren kombiniert, und so die Portierungsqualität verbessert.

Crosslinguale Akustische Modellierung mit Artikulatorischen Merkmalen

In der Vergangenheit wurde gezeigt, dass artikulatorische Merkmale zuverlässig über Sprachen hinweg erkannt werden können, und dass es möglich ist, sie multilingual zu modellieren. In der Vergangenheit wurden auch Verfahren entwickelt, um Modelle für artikulatorische Merkmale in HMM basierte Spracherkenner, die auf Phonemen basieren, zu integrieren. In dieser Arbeit haben wir die Modelle für artikulatorische Merkmale dazu verwendet, um die Leistung bei der Portierung von multilingualen, phonembasierten Modellen auf neue Sprachen zu verbessern. Durch den Einsatz der artikulatorischen Merkmalen innerhalb einer streambasierten Architektur in Kombination mit einem diskriminativen Verfahren zur Bestimmung der Gewichte für die artikulatorischen Merkmale konnten wir die Wortfehlerrate bei der Portierung auf neue Sprachen in verschiedenen Szenarien verbessern.

Ungeschriebene Sprache: Entdeckung von Worteinheiten

Linguisten schätzen, dass die große Mehrzahl der Sprachen in der Welt über kein Schriftsystem verfügt. Für den Fall, dass für eine solche Sprache ein Spracherkennungssystem entworfen werden soll, haben wir die automatische Entdeckung von wortähnlichen Einheiten in neuen Sprachen untersucht. Wir haben dabei den Fall behandelt, dass das Spracherkennungssystem Teil eines Sprachübersetzungssystems ist, und dass nur eine unsegmentierte, phonetische Transkription des Trainingsmaterials in der neuen Sprache verfügbar ist. Unser Verfahren zur Entdeckung der Worteinheiten bezieht dabei auch die Übersetzung des Trainingsmaterials mit ein, die für das Training des Übersetzungssystems zur Verfügung steht. Wir haben dieses Verfahren verglichen mit einem Verfahren aus der Literatur, das nur auf den monolingualen, phonetischen Transkripten arbeitet. Die Verwendung der parallelen Daten hat dabei zu Verbesserungen gegenüber dem monolingualen Verfahren geführt. Zur Evaluation der Verfahren haben wir die nach unserem Wissen erste Ende-zu-Ende Evaluation durchgeführt, die auch Spracherkennung mit den automatisch gefundenen Worteinheiten beinhaltet.

Danksagung

Mein erster Dank gilt selbstverständlich meinem Doktorvater Professor Alexander Waibel, der mir die Möglichkeit zur Forschung an seinem Lehrstuhl, den Interactive Systems Laboratories, gab, beginned als wissenschaftliche Hilfskraft über Studien- und Diplomarbeit bis hin zu dieser Dissertation. Durch seinen immerwährenden Ideenreichtum und seinen stets visionären Blick konnter mich schon früh für das Feld der Mensch-Maschine-Interaktion gewinnen.

Genauso gilt mein Dank auch meiner Koreferrentin Professor Tanja Schultz, die meine Forschung bereits seit der Studienarbeit mitbetreut und unterstützt hat und stets Mentor und Tutor für mich war und meine Arbeit mit ihrem Fachwissen begleitet hat.

Das Forschen und Lehren bei den Interactive Systems Laboratories ist geprägt durch die sehr kollegiale Zusammenarbeit und das kontinuierliche Lernen von den langjährigen Mitarbeitern, sowie durch viele konstruktive Gespräche unter den Kollegen. Mein besonderer Dank gilt dabei insbesondere Professor Ivica Rogina, der mich schon sehr früh in meinem Studium für die Problematik der automatischen Spracherkennung begeistert hat und stets ein guter Lehrer und Tutor war. Auch Christian Fügen gilt ein besonderes Dankeschön für die vielen, hilfreichen Gespräche und seine Unterstützung beim Umgang mit dem Janus Recognition Toolkit; gleicher Dank gebührt dabei auch Florian Metze, der auch meine Diplomarbeit mitbetreut hat, sowie Hagen Soltau.

Mein Dank gilt selbstverständlich auch allen anderen aktuellen und ehemaligen Kollegen am Lehrstuhl in Karlsruhe und im Labor an der Carnegie Mellon University: Keni Bernardin, Ralf Biedert, Susanne Burger, Paisarn Charoenpornsawat, Michael Dambier, Hazim Kemal Ekenel, Petra Gieselmann, Teresa Herrmann, Silja Hildebrand, Hartwig Holzapfel, Roger Hsiao, Shajith Ikbal, Qin Jin, Szu-Chen Jou, Thilo Köhler, Muntsin Kolss, Florian Kraft, Kenichi Kumatani, Ian Lane, Kornel Laskowski, Rob Malkin, John McDonough, Kai Nickel, Jan Niehues, Kevin Kilgour, Thilo Köhler, Mohammed Noamany, Matthias Paulik, Jürgen Reichert, Rainer Stiefelhagen, Christian Saam, Rainer Saam, Yik-Cheung Tam, Yury Titov, Ashish Venugopal, Stephan Vogel, Matthias Wölfel, Hua Yu, Ying Zhang, Bing Zhao, and Andreas Zollmann.

Die Arbeit ist natürlich nur möglich durch die vielen, helfenden Hände, die die Abläufe am Lehrstuhl organisieren. Deshalb gilt mein großer Dank auch Silke Dannenmaier, Annette Römer und Margit Rödder für die stete Hilfe und sehr gute Zusammenarbeit.

Auch den technischen Admininistratoren am Lehrstuhl ist es zu verdanken, dass die Forschung überhaupt erst möglich wird. Deshalb möchte ich mich auch bedanken bei Norbert Berger, Klaus Joas und Dan Valsan.

Desweiteren gilt mein sehr großer Dank meinen Eltern, Edith und Fritz Stüker, die mich immer vorbehaltlos auf meinem Bildungsweg unterstützt, begeleitet und schon früh gefördert haben. Auch meinen Schwiegereltern, Marat und Raisa Mimer gebührt Dank, da sie meiner Frau und mir in vielen Belangen die Bewältigung des Alltags neben der Arbeit um vieles erleichtern.

Ein ganz großer Dank gilt aber meiner Frau, Borislava, die mich während meiner Promotion immer unterstützt hat und mir immer Motivation und Antrieb war. Sie hat sicherlich in den letzten Jahren die größte Geduld mit mir aufbringen müssen, wenn die Wochenende für die Durchführung der Dissertation verwendet wurden oder es wieder spät im Büro wurde.

Contents

Chapter 1

Introduction

During the last decades the world has been radically changed on the economic, technological, sociocultural, and political sector by a process commonly referred to as *globalization*. Though the term is often used solely in the context of economic globalization, it in fact covers much more areas than that. Globalization can thus be described as a process by which the people of the world are unified into a single society and function together [Wikd, Bar08].

The availability of affordable and fast inter-continental travel and the interconnection of all parts of the world by means of modern communication technology, such as telphone or the Internet, has brought the once isolated peoples of the world close together. The world now forms the *Global Village*, a term coined by Marshall McLuhan in 1964 [McL64a, McL64b].

While modern technology makes it possible for every part of the world to communicate with every other, while affordable long range transportation has brought unprecedented mobility to the world, and while political initiatives have lead to global world trade, the language barrier remains as one of the last obstacles for enabling true communication among the peoples of the world [Ste06]. For example, the European Commission recognizes the language barrier as the last remaining obstacle to free trade and to the free flow of information within the European Union [otEC05].

The extent of the problem is often underestimated. As we will show in Chapter 2, estimates place the number of languages in the world at 4,000-7,000. The vast number of languages is only spoken comparatively small communities. Many of these languages are threatened by extinction, unless measures are taken to preserve them. Just as Vivian Redding, European Commissioner for Information Society and Media, does, we see language as the paramount achievement of mankind [Ste06]. The language of any culture is the base upon which it is built and with which it is inseparably intertwined. We thus agree with many linguists that the upkeep of a linguistic diversity in the world is fundamental to keeping up a healthy cultural diversity which is essential to mankind's prosperity.

1.1 Maturity of Speech Recognition Technology

After several decades of research, speech recognition systems for continuous speech with large vocabulary have reached a sufficient grade of maturity, that they are being deployed in real life. Commercial products are available for transcribing dictated speech[1][2], or in embedded devices such as car navigation systems [3][4].

Speech recognition systems are also being deployed for transcribing continuous speech in special, limited domains, e.g. for medical documents[5], legal documents[6], or pick-to-voice systems[7].

Speech recognition systems are further used in combination with machine translation technology in order to form speech translation systems in as of now limited domains. For example, in preparation of the Olympic Games 2008 in China, a consortium of companies and universities, sponsored by the Chinese and Beijing city government through the company CapInfo[8], developed a prototype for a limited domain speech translation systems between the languages English, Chinese, and Spanish for a touristic and medical domain [SZR$^+$06]. NTT-DoCoMo in cooperation with ATR-TREK[9] offers server based speech translation services

[1] IBM ViaVoice: http://www.nuance.com/viavoice, Naturally Speaking: http://www.nuance.com/naturallyspeaking

[2] Windows Vista Speech Recognition: http://www.microsoft.com/enable/products/ windowsvista/speech.aspx

[3] IBM Embedded ViaVoice http://www-01.ibm.com/software/websphere/products/ mobilespeech

[4] , http://www.harmanbecker.com

[5] http://www.mmodal.com/products.jsp

[6] http://www.citrix.com/English/ps2/accessanswers/challenge.asp?contentID=25643

[7] Pick by Voice: http://www.ssi-schaefer.de/

[8] http://www.capinfo.com.cn

[9] http://www.atr-trek.co.jp/

for their 905i cell phone series between English and Japanese, as well as between Chinese and Japanese in a tourist domain[10][POY$^+$08]. Mobile Technologies LLC. provides speech translation on hand-held devices in limited domains, such as tourism or health care for travelers or humanitarian workers [11].

Lately, several research projects have made progress in speech translation technology for very large domains. From April 2004 until March 2007 the European Commission sponsored the project *Technology and Corpora for Speech to Speech Translation* (TC-STAR)[12], an effort to advance research in all core technologies for speech-to-speech translation—automatic speech recognition, spoken language translation, and speech synthesis. TC-STAR aimed at a breakthrough that significantly reduces the gap between human and machine translation performance. The project targeted translation of unrestricted conversational speech on large and unconstrained domains of discourse. The main task chosen was the translation of speeches delivered in the European Parliament. Progress was driven by annual, competitive evaluations.

On the military side, the United States' (US) *Defense Advanced Research Projects Agency* (DARPA) sponsors the project *Global Autonomous Language Exploitation* (GALE). GALE develops technologies to absorb, analyze, and interpret huge volumes of speech and text in multiple languages. In the program, speech technology to recognize huge amounts of foreign speech (e.g. Chinese and Arabic) is developed as well as technologies to translate this information into English.

In 2005 the International Center for Advanced Communication Technology (InterACT)[13] presented the first simultaneous translation system from English to German and Spanish. The system automatically recognizes English speech in real-time and simultaneously translates it using statistical machine translation [FWK07, WF08].

Progress in this field is still being made, and systems still commit errors depending on the task addressed and the recording conditions. However, experts are confident that speech recognition will achieve a 99% accuracy within the next twenty years [GP06].

[10] http://www.nttdocomo.com/pr/2008/001402.html

[11] http://www.mobytrans.com/

[12] http://www.tc-star.org

[13] http://isl.ira.uka.de/index.php

1.2 The Challenge of Addressing all Languages in the World

Out of the 7,000 languages of the world, only for few of them automatic speech recognition systems have been created so far (Dragon Naturally speaking, for example, is only available in six languages; Nuance Recognizer V9 by Nuance Communication is available in 54 lanugages). The languages which were addressed are mainly those with a large population, a high economic value, or with high political importance.

Especially the latter point is a very volatile one. Languages which were not of interest in the past, can suddenly become of interest due to policy changes, political events or natural disasters. After the terrorist attacks on the World Trade Center in New York City and the Pentagon in Washington, D.C., which are commonly referred to as 9/11, and the resulting war in Afghanistan, Pasthu became of interest to the Defense Advanced Research Agency of the United States in the project Transtac [DAR08]. This has also to be seen in the light of the tensions about nuclear armament with the Republic of Iran. Also, due to the events of 9/11, Arabic became a focus language in the DARPA project GALE. After the beginning of the second Gulf War and the invasion of Iraq, Iraqi also became of interest to the American military, and was started to be addressed in Transtac.

As we will argue in Chapter 2, political interest from major countries is not the only reason, why all languages in the world should be the target of automatic speech recognition or natural language processing technology in general. In order to keep up the language diversity in the world, technology can play a vital role by allowing the speakers of endangered languages to access information across languages using their native language only. If technology were only to address the major, economically important languages in the world, and occasionally some less resourced languages due to some temporary, volatile self-interest, it would widen the *digital divide* between languages for which language technology is available and those without one [Yu02]. If, for example, systems that allow users to access information in large databases, such as Informedia [HMC^{+}03] or View4You [KWW00], are only available to the major languages in the world, pressure on the endangered languages grows, and they will be more readily abandoned by their speakers.

But, if in return, technology in the form of automatic speech recognition and machine translation allows instant access to all information in the world, regardless of which language the information is provided in, and regardless of the mother tongue of the information seeker, this would be a significant contribution to keeping up the language diversity in the world. So, besides the self-interest of not knowing which language might become interesting in the future, helping to keep up the language diversity in the world, is an important reason to try to cover all languages in the world by technology, including automatic speech recognition.

1.3 The Need for Low-Resources ASR development

The traditional process of creating an automatic speech recognition system in a new language requires many resources. In order to train the models used in statistical ASR, large amounts of audio recordings in the target language are necessary—modern research systems use up to several thousands of hours of speech. These audio recordings are usually transcribed at the word level. The production of such transcriptions requires the help of native speakers of that language and are usually time intensive to create.

Also, automatic speech recognition systems usually use phonemes as their modeling units, and thus need a pronunciation dictionary that maps the orthography of a word to its phoneme sequence. These pronunciation dictionaries are difficult to create and need the help of a specialist in the target language. Some automatic methods for creating dictionaries exist, but they require large amounts of training material which also need to be created by a specialist. Often, such an expert that can create the mapping between the words of a language and their pronunciation is not easily available to the developers of an ASR system.

Research in the past addressed the creation of phoneme based, acoustic models in new languages with only few adaptation material in the new language. However, these methods do not eliminate the need for the creation of a pronunciation dictionary in the new language.

Further, many languages in the world are without a writing system. For these languages the approach of collecting either the full amount of training material or only even a small amount of adaptation data does not work. For many applications, such as speech translation, it is not necessary that the result of the automatic speech recognition corresponds to a transcript of the speech in an existing writing system. Here, it is sufficient that the result from the ASR component is suited for processing by the component that follows in the chain, e.g. the translation component. In this case the words of a new language can be automatically discovered from an unsegmented, phonetic transcript.

1.4 Outline and Contributions

The work in this thesis is organized as follows. Chapter 2 will cover the vast diversity of the languages in the world. In it we will argue for the importance of keeping this diversity alive. We will further point out the role that human language technology, e.g. in the form of automatic speech recognition, can play in achieving that goal.

In Chapter 3 we will give a brief introduction into the field of automatic speech recognition. This introduction is not intended as a complete reference to the field, but rather introduces the concepts and terminology that is necessary to understand the experiments that we conducted.

After these introductory chapters the remaining chapters will describe our experiments in porting recognition systems to new languages with only limited resources and in a rapid way with as little knowledge about the new language as possible.

Chapter 4 treats the case in which sufficient amounts of transcribed speech data in the target language are available, but no pronunciation dictionary for that language. We will show in this chapter that for this case the use of graphemes as modeling units instead of the traditional phonemes is a viable solution for many languages in the world. We will point out the heightened importance of the model cluster tree of an ASR system that utilizes graphemes as modeling units. To accommodate this, we apply a flexible cluster tree that better captures the complex relationship between graphemes and their pronunciation than our traditionally used tree does. We will show improvements in the recognition systems using this flexible tree.

In the past, techniques for constructing and porting multilingual, acoustic models to new languages have been studied for phoneme based speech recognition systems. In Chapter 5 we will extend this notion to the grapheme based systems that we introduced. We will discuss the peculiarities in constructing multilingual grapheme based systems, namely the problem of only a bad correspondence of the pronunciation of graphemes in different languages and the potentially low or non-existing overlap of graphemes across different languages. We will propose methods for finding automatic mappings between the graphemes in different languages, by applying two data driven mapping methods. Using the automatic mappings we will show that it is possible to port a grapheme based multilingual acoustic model to a new language that uses a completely different writing system than those of the languages in the multilingual model.

For monolingual, phoneme based ASR systems, the use of articulatory features has been studied in order to better model speech, especially spontaneous one. Past research has also shown that articulatory features can be recognized across languages. In Chapter 6 we will incorporate the use of articulatory feature detectors into the process of porting a multilingual, phoneme based acoustic model to a new language. We will consider different scenarios of porting acoustic models, and apply different methods for estimating the stream weights that are necessary for the approach used to integrated the feature models into the recognition process.

Many languages in the world are not written, either because no writing system for them exists, or because speakers of that language resort to a different language for written communication. In Chapter 7 we present a novel method for discovering new word units in an unwritten language from speech. Since these generic word units would be of little value on their own when being recognized, we study the use of an ASR system using them in the context of speech translation. We will show that the word units that we find are suitable for automatic speech recognition in combination with statistical machine translation in order to form a speech translation system. We will show that our approach gives significant improvements over the state of the art in literature. We will further carry out the first full speech translation evaluation in literature for this task including both, speech recognition and machine translation, with the newly found units.

Chapter 2

Global Language Diversity and its Role

The world boasts a wide variety of very different languages in high numbers. This diversity in languages is valuable to civilization, but at the same time in danger. Languages are dying at a rapid rate, and many will become extinct, unless countermeasures are being taken.

In this chapter we will described the diversity in languages in more detail and its value. We will discuss the current process of extinction and explore the reasons for it. The European Union will serve us as an example of a political body who has recognized the value of its language diversity and is taking steps to protect it. We will show that modern technology, including automatic speech recognition, can be part of the solution if it is developed in such a way that makes it suitable for this challenge.

2.1 Language Diversity

The number of languages in the world given by linguists surprises most ordinary readers. Since the definition of a what constitutes a language is ambiguous, and because of insufficient research in the field, linguists' estimates place the number of languages in today's world in the range from 4,000 to 7,000 languages. The fifteenth edition of the Ethnologue [GG05] lists 7,299 living languages. Table 2.1 shows the top 20 languages with the most first speakers according to Ethnologue, as summarized by Wikipedia [Wikc]. In this list the figures of Chinese and Arabic include the sum of all of their varieties, which are not necessary mutual intelligible. Though these languages all show a large population of speakers, this is not the case for most languages in the world. 347, that is about 5% of the world's languages, have at least one million speakers and account for 94% of the population. The remaining 95% of the world's languages are therefore only spoken by 6% of the world's population [GG05]. Considering that there are roughly 200 countries in the world, there are 25 to 30 times as many languages in the world as countries. But on the other side, 83% of all languages are only spoken in one country. That means that countries exist with a high diversity of languages within the country. Hand in hand with this relation between countries and languages goes the fact that many languages exist in a *diglossic* relationship [NR00]. Two languages being in a diglossic relationship means that a functional specialization between them has evolved: One is being used for interaction within home or personal domains, while the other is being used for higher functions, such as government, media, and education.

When it comes to the density of languages in terms of geographical region, in general countries around the tropics show the highest densities in languages, with the density declining when moving towards the poles [Net98]. In total, 70% of all languages exist in only twenty countries. In contrast, in Europe only 3% of all languages reside, and in China only 2.6% of all languages, though it contains 21.5% of the world's population and 8.6% of its land mass. Figure 2.1 illustrates this concentration of languages in the tropic regions of the world.

Neither the economic strength of a country nor its technological degree of development, though one might intuitively think so, seems to be a factor in the regional density of languages. In a statistical analysis by Sutherland, neither Gross Domestic Product nor number of television sets per 1,000 people were significantly related to the number of languages in a country [Sut03].

Given the high diversity in languages, the fact that many languages are spoken only by comparatively few speakers, and their location in often remote and inaccessible areas of the world, it comes to no surprise that the vast majority of the languages in the world is not well studied by linguists. Nettle and Romaine [NR00] claim that many of the languages that are known today only have been briefly looked at by the person that discovered them, often a missionary or explorer, but has never been studied by a linguist.

Nettle, Romaine, but also other linguists, such as David Crystal, are of the opinion that SIL International, the publisher of Ethnologue, is the most com-

Table 2.1: The 20 languages with the most number of first speakers according to Ethnologue, summarized by Wikipedia

Rank	Language	No. of Speakers in Mio.
1	Chinese	885
2	Spanish	322.3
3	English	309.4
4	Arabic	206
5	Hindi	180.8
6	Portuguese	177.5
7	Bengali	171
8	Russian	145
9	Japanese	122
10	Standard German	95.4
11	Javanese	75.5
12	Telugu	69.7
13	Marathi	68
14	Vietnamese	67.4
15	Korean	67
16	Tamil	66
17	French	64.9
18	Italian	61.5
19	Western Panjabi	60.8
20	Urdu	60.5

prehensive source on the existing languages in the world, especially the lesser known ones [NR00][Cry00]. Though ‘SIL’ originally stood for ‘Summer Institute of Linguistics’ it is in fact a Evangelical Christian non-profit organization and started out as summer training session in the U.S. in order to train missionaries. It is financed by the ‘Wycliffe Bible Translators’, an organization with the goal to provide Bible translations in all living languages in the world.

According to Nettle and Romaine the main effort of linguistic research has been focused on the few majority languages in the world with a large amount of speakers. This fact makes it more difficult to provide sound and comprehensive data on the language diversity in the world and the properties of the living languages. Therefore, numbers and descriptions given in literature are often approximations and different sources give different numbers.

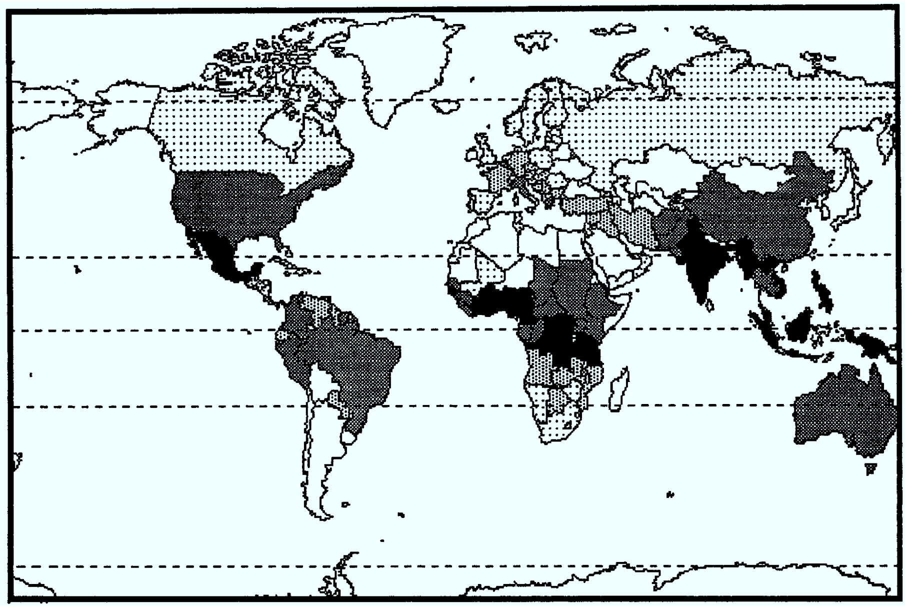


Figure 2.1: Geographic distribution of writing systems [from [Net98], courtesy of Academic Press]

2.2 Extinction of Languages

The diversity of languages described above is in danger. Languages are frequently disappearing. Nettle and Romaine claim that about half of the languages of the world have vanished in the last 500 years and are continuing to do so, as they show by many examples [NR00]. Even though languages are not living organisms, they are closely connected with humans, culture, and environment. For these reasons Nettle and Romain adopt the terminology of *language death* and talk about the *extinction* of languages.

2.2.1 Extent of the Problem

Classifying the size of the danger to the language diversity is difficult. The picture painted in [NR00] indicates a grave problem. Not only seems the rate of extinction to be increasing, but language death is also not a geographically isolated problem, but rather happens all over the world. Among linguists the estimate that at least half of the living languages today will become extinct in the next century seems to find consensus [NR00] [Jan02][Cry00]. The endangerment of language is a matter of degree. Ethnologue in its fifteenth edition lists 497 languages as nearly extinct. According to their definition that means that either a language has fewer than 50 speakers or is only spoken by a tiny fraction of its ethnic group [GG05]. If one assumes that languages with more than 100,000 speakers are considered safe, then there are only 600 safe languages in the world [Kra92]. Within Europe, Irish, Scottish Gaelic, Welsh, and Breton are considered to be endangered languages.

Sutherland classified the level of endangerment of the existing languages using the same scheme as being used for birds and mammals [Sut03]. His analysis showed that languages are even more threatened than birds and mammals, classifying 7.1% of all languages as critically endangered, as opposed to 1.9% of all birds and 4.1% of all mammals. He counted that since 1600 A.D. 306 languages have gone extinct, as opposed to 125 birds and 87 mammals.

Language death by itself is not an isolated event, but is very often the symptom of death for a whole culture. The world's linguistic diversity is a benchmark of its cultural diversity

A language is said to be extinct when there are no more living people who can fluently speak that language. If there are people living who know isolated phrases, but are not able to freely communicate with the help of that language, that does not count as a living language.

But inbetween the two poles of a living language and an extinct one there are different shades of endangerment to a language, similarly as it is done for species of animals or plants. In Crystal's eye a three class classification for the state of a language is common sense [Cry00]: *safe*, *endangered*, or *extinct*. Beyond that Crystal cites other classification schemes which are more fine grained. Krauss adds to this classification pattern the concept of *moribund* languages [Kra92]. This class refers to a language which is still living, i.e. spoken by a population, but which is almost certain to go extinct soon, since it is not being passed on to younger generations. In a philosophical sense, a language that is not being passed on to younger generations is as good as dead. When only one speaker of a language is left, that language by definition is not yet extinct. But is not actually living anymore, since it cannot be used for communication anymore, since the sole speaker is lacking a counterpart for communication

Then there are classification schemes which use an even finer grained resolution and distinguish between languages which are definitely safe and those which are lesser so, as done in [Kin91]:

viable languages: have population bases that are sufficiently large and thriving to mean that no threat to the long-term survival is likely

viable but small languages: have more than 1,000 speakers, and are spoken in communities that are isolated or with a strong internal organization, and aware of the way their language is a marker of identity

endangered languages: are spoken by enough people to make survival a possibility, but only in favorable circumstances and with a growth in community support

nearly extinct languages: are thought to be beyond the possibility of survival, usually because they are spoken by just a few elderly people

extinct languages: are those where the last fluent speaker has died, and there is no sign of any revival

As last example Crystal gives the classification scheme by Stephen Wurm which has an even finer grained resolution of the endangered languages, but which is not a full classification scheme because it excludes the safe languages [Wur98]:

potentially endangered languages: are socially and economically disadvantaged, under heavy pressure from a larger language, and beginning to loose child speakers

endangered languages: have few or no children learning the language, and the youngest good speakers are young adults

seriously endangered languages: have the youngest good speakers at age 50 or older

moribund languages: have only a handful of good speakers left, mostly very old

extinct languages: have no speakers left

Sutherland [Sut03] used the species extinction risk which is assessed by standard quantitative criteria based on population size, actual or suspected population decline, range size changes, and habitat fragmentation [IUC94]:

vulnerable languages: facing a high risk of extinction in the wild in the medium-term future

endangered languages: facing a very high risk of extinction in the wild in the near future

critically endangered languages: facing an extremely high risk of extinction in the wild in the immediate future

extinct languages: There is no reasonable doubt that the last speaker has died.

Using this scheme Sutherland has classified all languages documented into these categories, whereas 639 languages were data deficient and could not be classified. Table 2.2 lists the result. A total of 1,676 languages, roughly 25% of the living languages, are on this red list of languages.

Table 2.2: Red list of languages

Category	Extinct	Critical	Endangered	Vulnerable
No. of Languages	304	438	506	732

The definition of the different stages of endangerment already transfer the notion of the circumstances or symptoms that classify a languages as being in peril. As the reader can already guess from the varying classification schemes, different indicators are cited in literature in order to diagnose whether a language is at threat or not.

2.2.2 Conditions under which Languages are at Risk

Some scientists use the number of speakers as an indicator for whether a language is at danger or not. E.g. Krauss gives as a rough estimate that a language with at least 100,000 speakers is considered safe [Kra92]. Nettle and Romaine however point out that the sheer number of speakers of a language is not a sufficient indicator. They underline their opinion by giving a number of examples where this arbitrary threshold is not a good indicator. For example, Icelandic has only 100,000 thousand speakers but is far from being at risk. On the other hand, Breton as late as 1926 had over a million of speakers. But today Breton is considered an endagered language. Thus, being a strong language in the past does not guarantee survival. Even if a language is the official language in a country and is actually backed by the government of a country, this is not a guarantee for survival. Irish, for example, has a large group of speakers, is backed by the Irish government, but is still in grave danger of being replaced by English. On the other side, in Vanatu none of indigenous languages has more than 3,000 speakers, but most of them seem to continue living. In Micronesia the two languages with the highest risk of extinction are the largest one—Chamorro with 60,000 speakers—and the smallest one—Sonsorolese with approx. 300 speakers. So, instead of taking the number of speakers as an indicator, they argue that a language is safe, i.e. will continue living, when parents or caretakers in general pass their language on to their children. Where this is not the case, a language will eventually die.

In that sense the classification schemes of Wurm and Kincade are a good combination of these two factors, number of speakers and tradition of the language to the next generations.

These indicators for danger to a language do not yet give reasons for why this condition might arise. The reasons why parents refuse to pass their language on to their children are not uniform and need to be discussed separately.

2.2.3 Reasons for Extinction

Since a language dies when its speakers die, any event that impacts the physical well-being of a community of speakers can lead to the extinction of a language. While for languages with a large group of speakers, that are globally wide spread, such an event seems rather unlikely, for many of the hundreds and thousands of languages with only very few, locally concentrated speakers, this option is not an unlikely one. In history natural disasters such as hurricanes, floodings, volcano eruptions, Tsunamis, and earthquakes are known to have killed large portions of the population of the region in which the events occurred.

Crystal cites the example of a massive earthquake on 17 July 1998, off the coast of E. Saundaun Province, Papua New Guinea [Cry00]. The earthquake killed over 2,200 people and displaced over 10,000. Within the disaster area were four village, of whose population 30% were killed, the rest displaced and distributed to different medical and emergency relief centers. According to the SIL there were strong indications that each village had its own distinct language, with further research on that matter necessary. However, after the event of the earthquake and the displacement of the villagers, it seems unlikely to Crystal that these small communities and with them their language have survived.

In this event the destruction of the habitat lead to a reduction of the population and the displacement of the remainder. In other cases the habitat might not be destroyed but just unfavorable for survival. Often famines and drought, though leaving the habitat mainly intact, lead to the displacement of a population. For example the potatoe famine in Ireland between 1845 and 1851, leading to the death of 1 million people and lead to a long period of emigration. It is believed that this emigration is one benefactor to the threat of the Irish language, since the emigrants often adopted the language of the country to which they fled—often English as many emigrants travelled to the U.S. Nowadays economic factors can create the same kind of stress as famines do. Due to exploitation of resources more and more areas are prone to the effect of *dessertification* which in return leads to the displacement of people.

Especially in the case of indigenous people, imported diseases play a critical role in the extinction of peoples and the resulting loss of their languages. Within the 200 years of the arrival of the first Europeans in the Americas over 90% of the indigenous population was killed by diseases, mainly smallpox, imported by the explorers and conquerors [Duf53][Pea95][SS45]. Currently AIDS is believed to be one of the greatest threats to languages in terms of diseases. Areas affected mainly include Sub-Saharan Africa, South and South-east Asia, and Latin America. These areas contain over three-quarters of the world languages.

Also into the category of language extinction by physical harm to its speakers falls the case of genocide. Nettle and Romaine cite the example of the death of Yahi, the language of the Yahi Indians in the area of what is now California, who were murdered and driven into exile by white settlers.

Similarly, a language can begin to die when their people are political persecuted and the use of the language is severely punished, as for example in 1932 in El Salvador after an Indian uprise or in the 1970s when the Ubykh living near Sochi were scrutinized.

From all these examples it becomes already clear that the threat to a language is not simply identified by a threat to the life of its speakers. Rather the danger to a language results from the fact that the culture that is affiliated with that language is in distress. The displacement of a people would not be a factor in the death of a language, if the displacement would not mean at the same time an immense stress to the culture of the displaced people. Thus, in the words of Nettle and Romaine, the *thread to languages extens to a thread to cultures* [NR00]. So, whenever circumstances put a culture at stress, the underlying language is also automatically affected.

Physical threat to the health of its people are not the only cause for stress to a culture. Often the people of a culture continue living, but their culture fades away and with it their former language. Though Crystal makes a clear distinction between the factors that put people's safety at risk and factors that change, or in the end kill, a culture, we believe this distinction to be wrong. We rather believe that the threat to the physical safety is just one of the factors that might extinguish a culture. The stress on culture is in our eyes the superordinate factor for threat to people. With this notion we follow Nettle and Romaine who also see distress to culture as the ultimate reason for the death of a language [NR00]. So, Nettle and Romaine distinguish rather between the sudden or at least fast death of a language and its gradual death. For them a language always dies out “because an enduring social network to which people sought to belong somehow ceases to be ”. They give three major classes of reasons why this can be the case. Like in Crystal’s classification the first reason is that the people who speak a language cease to exist, usually by population loss, e.g. through disasters or diseases. The other two reasons are concerned with a shift to a different language. They distinguish between a forced shift, as their second reason for language shift, and a voluntary shift as the third reason.

The current, premier example of a language to which speakers of a dying language shift, a so called *killer language*, is English. For example, the speakers of Gaelic and Irish are currently switching to English and the two languages are gradually dying [Pri84]. This trend of a shift to English is a global one as can be seen by the fact that by 1996, 70% of the world’s mail and 60% of radio and TV broadcast were in English.

A forced shift of a language can be sought to be achieved by different means. In history, dominant groups have often sought to suppress a language by making the dominant language compulsory. However, this policy rarely seems to have worked, because the stigmatized language may gain in value, becoming a symbol of resistance. Methods that have worked in the past for forcing a dominant language upon a less dominant people, and that are also in use today, e.g.in the rain forest, are, according to Nettle and Romain: enslavement, forcing the people into a sub-ordinate rule, and seizing the land and resources upon which their communities are based.

The reasons for voluntary shift are according to Nettle and Romaine linked to the fact that a community or people perceive an advantage by switching to a different language. The advantage could be an economic one, where by adopting a different language, people might be able to obtain better jobs, improve their living circumstances and wealth. Or, the advantage could be a perceived increase in prestige and social status. They give as an example a study by Susan Gal who determined that the newly available status as ‘worker’ in Austria prompted a previously monolingual Hungarian community to start to speak German. Young women chose to speak German in order to find German-speaking marriage partners in order to adopt a new social identity, leaving the old one of a peasant behind [Gal79].

According to [NR00] a language that gradually dies will go through a period when it is not used for all functions anymore. Over time speakers will more and more resort to routine and formulaic speech and will not be able to readily create new utterances on the spot. Nettle and Romaine give the example of a study of Dyirbal, an Australian, Aboriginal language [Sch85]. That study showed that younger speakers of that language lost more specific words and replaced them with general ones. They were also prone to loose names referring to culture specific items relating to weather, geography, ceremonies, and kinship. Other words then had widened their meaning which now include objects introduced by whites into Dyirbal culture. Also, the grammatical complexity of the language will decrease over time.

Sutherland's statistical analysis also showed that as languages become rare they become less attractive, a self-reinforcing process [Sut03]. Abrams and Strogatz reach a similar conclusion. They developed a model to capture the dynamics of language death. Their model predicts that two languages cannot coexist stably, since one will eventually drive out the other [AS03]. In their analysis, the status of a language is the most relevant factor and could serve as good indicator for the threat to a language.

2.2.4 Reasons to Keep Up the Language Diversity

Given the fact that languages are dying at a rapid rate for the reasons discussed above, the question remains why we as as scientific community should actually worry about stopping this trend and to keep up the diversity of languages in the world. The extinction of languages due to natural disasters or through epidemic diseases are often a force majeure which cannot be prevented by human intervention. Other reasons, such as the switch to more prevalent languages, e.g. English, are often due to the fact that the abandoning of a language gives people a benefit, improving their outlook in live or their prosperity. Under the assumption of the benefit of a free market, why should mankind intervene and try to change conditions in order for the languages to survive? For some of the reasons it is self-evident that they should be removed, even be fought. The deliberate oppression of a people or culture, including the "murder" of their language, is against the acknowledged human rights which include the right to preserve and live ones own culture. This is even more true for the extreme of genocide, also one of the reasons for the death of a language, which is not to be tolerated by mankind. In these cases, it is not a question of preserving a language, but rather preventing crimes against humanities. As a byproduct, this prevention will also safe a language. But in these cases, the question remains, whether it should be worth the effort to try to keep a language alive which has been brought to the brink of extinction by political suppression or genocide, after the crimes have been stopped. Or should one just leave the languages on their own without interfering with their further development? Where is the benefit in reviving such a scrutinized language?

For linguists, such as Nettle and Romaine, the obvious answer is the scientific interest itself. Only by examining as many different languages as possible, they claim, linguists can extend and perfect their theory of languages. Understanding languages is important, because they are an expression of variation in people's mind and thinking. According to them new discoveries about languages are still being made by studying new languages. Since they also claim that many of the languages in the world have not been treated by scientists yet, the death of languages means the potential loss of a new discovery that would further our understanding of languages. They give as an example of a near miss the language Hixkaryana which at the writing of their book [NR00] had only 350 speakers and is thus at the brink of extinction. Hixkaryana has an unusual sentence structure in that sense that it puts the object of a sentence at its beginning. Had the languages not been discovered by chance recently, it might have died out without the scientific community ever knowing about its unique sentence structure. Nettle and Romaine claim that many of these unique properties of languages, and the insight into the human mind and the way it organizes thoughts and concepts, are mainly found in isolated, small languages, which are thus prone to extinction. The knowledge that can be gained from them cannot be gained from languages which are widely spoken. These major languages are being grammatically streamlined and are becoming more and more alike due to cultural exchange and intertranslation. Thus, a loss in language diversity also leads to a loss in complexity of expression. Crystal, also a linguist, shares this opinion of course and also cites scientific insight as one of the reasons.

Beyond purely scientific curiosity, Nettle and Romaine also cite as a reason that language is an uniquely human invention and key to culture, technology, and accumulated wisdom. Vivian Redding, European Commissioner for Information Society and Media, sees language as the paramount achievement of mankind [Ste06]. Unlike technology, to Nettle and Romaine languages are not interchangeable, since every language has an individual window to the world. A loss of a language is a loss in this diversity and therefore a loss to all of us.

Both, Crystal, as well as Nettle and Domain, show that languages contain knowledge which is unique to them and would be lost, should they die. From the process of a gradual language death described above it becomes easily clear how the knowledge encoded in specific vocabulary of a language is lost during its gradual decline. Much indigenous knowledge is encoded in languages, such as medical knowledge of plants in the rain forests, climate behavior in the arctic or marine resource management in Polynesia. When the language dies that transports this knowledge, the knowledge itself vanishes. As an impressive example Nettle and Romaine cite the example of a Paluan traditional fisherman, born in 1894 and interviewed by marine biologist R.E. Johannes [Joh81]. The fisherman had names for 300 different species of fish, and knew the lunar spawning cycles of several times as many species of fish as has been described in scientific literature for the entire world. Similar knowledge about sustainable management of ecological resources or environmental knowledge is encoded in the languages of many other indigenous people. In the case of the Paluan fisherman, with the switch of the younger generations to other languages, they also loose the knowledge encoded in the language of their parents or grandparents. In that sense, Nettle and Romaine see many similarities between the loss of linguistic diversity and biodiversity. They call this combination today's biolinguistic diversity crisis.

Besides this knowledge encoded in languages, including historic knowledge, Crystal also stresses the more abstract knowledge and insight that can be gained by multilingualism. To him multilingualism is the normal state of human mind, proven to him by the fact, that humans are easily capable of learning multiple languages. He also refers to the multitude of artistic achievements achieved through language, e.g. by literature, and doubts that works by Shakespear would have been possible if English had been displaced by a different language.

Both, Nettle/Romaine and Crystal, see linguistic diversity as an important factor by itself. Both draw parallels to biological diversity which is vital to a healthy ecosphere. For them, only in diversity lies the guarantee for a healthy linguistic environment. Just as cultural and biological diversity is a desirable state, so should be linguistic diversity.

2.2.5 Measures to Prevent the Mass Extinction of Languages

Having shown the extent of the problem of languages dying and having discussed the conditions and reasons that lead to it, the question remains what can be done to stop this trend. Nettle and Romaine claim that traditionally linguists have tried to preserve languages by capturing their grammar and dictionaries [NR00]. However, actually the preservation of a language should mean to maintain a group who speaks it. They distinguish two broad approaches to reaching this goal, both being in line with the aim of a sustainable future. The first approach is the initiation and support of bottom-up approaches. They include the organization of local teaching of the language to the younger generations of speakers, e.g. at local schools, by local non-governmental organizations, or groups of parents. They cite that the absence of schooling in an endangered language makes maintenance difficult. In a study of 46 linguistic minorities in 14 European countries, a clear link between language and schooling emerged [All79]. A minority language which is not taught tends to decline.

In the opinion of Nettle and Romaine one should in the process of preserving language diversity accept and seek a state of bilingualism. Not every of the small languages in the world will be able to survive as the major language of its region.

Often a minority language's only chance of survival will lie within bilingualism and existing in a state of diglossia. But, as Abrams and Strogatz showed, under the current circumstances two languages in the same country cannot coexist without altering the status of a language [AS03]. In their article they postulate therefore policy-making, education, and advertisement as measures to change the status of a language and to preserve a bilingual environment.

The second class of solutions that Nettle and Romaine suggest are top-down strategies. Interestingly, they think that too much attention focused on official policy statements can be counterproductive, especially in the absence of low-level activities. Instead, they propagate to make language preservation part of general activism on behalf of the environment. Throughout their book they claim that environmental preservation receives more attention than language preservation. Since the solution to both problems are in their eyes very similar, a coupling of both approaches could be beneficial. Also, they see the area of Human Rights as a possible field in which the preservation of languages can be propagated. The top-down approaches should in their opinion establish language policies on local, regional, and international level as part of overall political planning and resource management.

Crystal postulates six measures to keep up the language diversity, which in part overlap with the approaches above [Cry00]:

1. To increase the prestige of a language within the dominant community

2. To increase the speakers' wealth relative to the dominant community

3. To increase the legitimate power of the speakers in the eyes of the dominant community

4. To have a strong presence of the speakers in the educational system

5. To enable the speakers to write down their language

6. To enable speakers to make use of electronic technology

For the last point, Crystal mainly has the Internet in mind as a medium for affordable distribution of a language. However, we believe that language technology, especially in the form of automatic speech recognition and machine translation systems, can make a significant contribution to the upkeep of the world's language diversity. We will argue this point in more detail in Section 2.4

2.3 The Example of the European Union

The European Commission (EC) has acknowledged in a communication to the European council, the European Parliament, the European Economic and Social Committee, and the Committee of the Regions, that multilingualism is essential for the proper functioning of the European Union [otEC05]. In addition to 23 official languages, the EC counts 60 other indigenous languages and scores of non-indigenous languages spoken by migrant communities [Nel96]. Table 2.3 gives an overview of the most commonly used languages within the European Union. About half of the Union's citizens state that they can hold a conversation in at least one language other than their mother tongue. Figures were taken from the Eurobarometer Report [eur05].

To the commission, language is the most direct expression of culture. This is reflected by Article 22 of the Charter of Fundamental Rights of the European Union, which states that the Union shall respect cultural, religious and linguistic diversity. Article 21 explicitly prohibits discrimination on the grounds of language, among other reasons.

The EC defines *multilingualism* as both, a person's ability to use several languages and the co-existence of different language communities in one geographical area. The EC's multilingualism policy has three aims: a) to encourage language learning and promote linguistic diversity in society, b) to promote a healthy multilingual economy, and c) to give citizens access to European Union legislation, procedures and information in their own language.

In its communication, the EC recognizes that European businesses need skills in the languages of the EU as well as of the trading partners around the globe, in order to be successful in trade. The EC also states that its citizens should be able to communicate with the institutions of the EU and read EU law in their own national language and take part in the European project without encountering any language barriers. It thus also recognizes that the language barrier in the world is a potentially risk to successful international trade as well as the functioning of a global governance. The EU spends currently 1.05% of its total budget, or 2.28 € per citizen per year, on translations. In total the EU spends 1.1 billion € per year on their translation and interpretation services [Ste06].

In order to achieve the goal of a multilingual European Union, the commission sponsors and supports the study of foreign languages by its citizens. It is the goal that eventually every citizen of the EU speaks at least two foreign languages.

The European Union also sponsors and executes programs for research and development in multilingualism. It specifically targets fully automatic translation systems, as of now low-to-medium quality, and automatic speech recognition and synthesis. In the eyes of the Commission a multilingual information society requires applications for all languages of the Union, including the less widely used ones. Using technology it also aims at keeping the costs for translations within the institutions of the Union in balance with the goals.

Table 2.3: Languages most commonly used in the EU

Language	Mother tongue	Foreign Language	Total
English	13%	34%	47%
German	18%	12%	30%
French	12%	11%	23%
Italian	13%	2%	15%
Spanish	9%	5%	6%
Polish	9%	1%	10%
Dutch	5%	1%	6%
Russian	1%	5%	5%

2.4 The Role of Technology in a Linguistically Diverse World and its Challenges

Just as language is still a barrier to trade within the European Union [Ste06] it is a barrier to global trade and communication as well. At the same time, as discussed in Section 2.2.4, it is desirable to keep up the high language diversity in the world—instead of following the trend of extinction of languages towards the establishment of only one or a few major languages. We believe that modern technology in the form of *human language technologies* and *speech-to-speech translation* systems can play an important role in reaching the goal of living in a multilingual global village.

As the discussion in Section 2.2.5 has shown, many reasons that lead people to abandon a language are related to the perceived social status of a language or to the economic costs associated with keeping up a minority language. By providing language technologies, such as automatic speech recognition systems, for these endangered languages, the prestige and usefulness of that language is heightened. It becomes more useful since thoughts expressed and information provided in it become available to a larger audience. At the same time time and money saving techniques that are based on speech recognition systems (e.g.

pick-to-voice or dictation systems) become available for that language. The availability of this technology, that only existed for a small number of languages in the past, to a new language raises its status at the same time, removing a discriminating factor that devalauted that language from others in the past.

As the example of the immensely high translation costs within the institutions of the European Union show, providing affordable translation services for all languages in the world by conventional means is not feasible. Here, automatic speech translation systems are an alternative, providing cheap translation services where expensive human translations are not affordable. The EU has recognized the potential in that technology for reducing their own translation costs, but the advantages are available to the whole world, not just the EU.

In the case that by automatic translation and speech translation service people world wide are enabled to access information across languages and to freely communicate across languages—or in other words to easily jump the language barrier—pressure is taken from them to abandon their language in favor of a more widely spoken language. In case of global translation, all languages become equal in terms of reach and distribution.

In this way technology can provide to the speakers of a minority language "new channels for the use of their language, and so to strengthen it" as Nicholas Ostler puts it [Ost01]. In Ostler's view the speakers of the language would then "find that the world is their oyster and available to them on something like their own terms". This emphasizes the fact that the diverse people of the world could then interact with each other and act globally while at the same time preserving their own culture.

But in order to for this vision to become true, language technology needs to address all languages in the world, not only the handful of languages treated so far. Languages addressed until now are mainly those with either a large population of speakers, with sufficient economic funding, or with high political impact [DAR08]. The fact that applications using ASR only address a small fraction of the world's languages bears the danger of creating a digital divide between those languages for which ASR systems exist and those without one. Also, one of the criteria that have prompted the treatment of specific language so far—political impact—can be a very volatile one. Due to conflicts, natural disasters, or otherwise changing conflicts, recognition and translation systems might suddenly be needed for a new language with only very little time for development.

Given the vast number of languages that need to be addressed, it becomes clear that for the technology of automatic speech recognition the traditional way of creating such systems does not scale to its application to the complete set of languages. Though many of the techniques utilized for automatic speech recognition systems are language independent (see Chapter 3) they at least require the collection of large amount of language specific resources, such as transcribed audio recordings and large text collections. This collection as well as other design decisions also often require the help of an expert in the language addressed. As we have seen in Section 2.1 many languages have too small of a population, as that this procedure would be feasible.

The challenge to technology, including automatic speech recognition, is to be able to address all languages in the world at affordable costs and in a timely manner. To achieve this, new methods and paradigms need to be developed.

2.5 Conclusion

Roughly 7,000 languages exist in the world, many with only comparatively few speakers. However, this high diversity in languages is the base of our cultural diversity and of great value to mankind, instead of being a burden. Languages are currently vanishing at an alarming rate. Estimates predict that up to 90% of today's languages will be extinct within a few generations. The main reasons for that is a frequent shift of speakers of minority languages to more dominant languages. Political institutions, such as the European Union, have recognized the value of a multilingual world and multilingual societies, and are putting up measures and programs to preserve the multitude of languages in the world.

Modern technology, including automatic speech recognition, can play a role in solving this challenge of upkeeping all languages in the world. However, new methods and paradigms need to be developed to address all languages in the world at affordable costs and in a timely manner.

CHAPTER 3

Basic Concepts of Automatic Speech Recognition

In this chapter we describe the fundamentals of *automatic speech recognition* (ASR) on which the experiments, that we performed to port speech recognition systems to new languages, build. This chapter is only intendend as a brief introduction and overview, in order to declare the terminology that we will use in the later chapters.

Automatic speech recognition, as we use the term in this work, is the recovery of the word sequence uttered by a person in an audio recording. The audio recording is produced with the help of a microphone and then digitized in order to be processed by a computer. In this book we focus on *large vocabulary continuous speech recognition* (LVCSR), that is the recognition of continuous utterances with a comparatively high number of different words, e.g. several tens of thousands. Current state of the art LVCSR systems make use of several techniques from the machine learning world which we will briefly describe below. In order to be able to recognize human speech, it is also necessary to have a at least cursory understanding of the human speech production process and human language.

3.1 Human Speech Production

We give here only a short introduction into the topic which covers briefly the fundamentals necessary for understanding the experiments in this book. A more detailed coverage of the topic is widely available in literature and can for example be found in [Can05, Ass99, CY95].

Humans produce speech by pressing air from the lungs through the mouth and nasal cavity and out of the mouth and/or nostrils. This egressive air stream coming from the lungs is modified by various *articulators* on its way, in order to produce the desired sounds.

The sound emitted leads to a change in air pressure over time, a *sound wave*, which then can be either pickep up by the human ear of the listener or which can be measured by a microphone.

The main articulators involved in creating this sound wave, as given by [HAH01], are depicted in Figure 3.1:

Lungs: produce the airstream passing through the articulatory apparatus and emanating from mouth and nostrils.

Vocal cords: are located in the larynx. By holding them close, they can be brought into oscillation and modulate the airstream from the lungs. The part of speech in which the vocal cords vibrate are called *voiced.* If the cords rest, the speech is called *unvoiced.*

Velum: can be open or closed and thus either allow or forbid the passage of air into the naval cavity and further out of the nostrils.

Hard palate: a hard surface at the roof of the inside mouth. The tongue can be put against it in order to inhibit the passage of air.

Tongue: can be used to either constrict the passage of air or to allow different kinds of resonation inside the mouth cavity.

Teeth: like the hard palate another immobile place against which the tongue can be placed

Lips: can be used to temporarily or for a longer time shut off the emission of air through the mouth. By rounding or keeping them flat they can also modulate the emanating air stream.

The modifications to the egressive airstream by the articulators above are usually a combination of the potential vibration of the vocal cords and a resonance

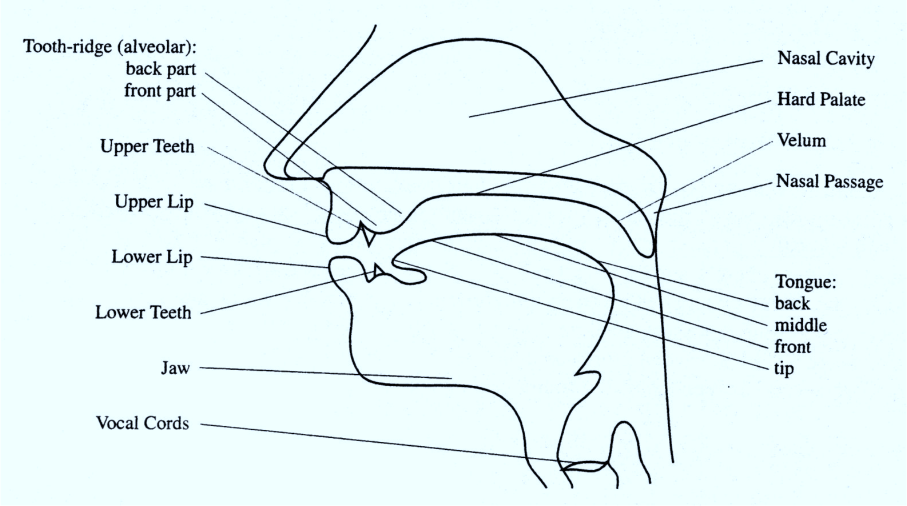


Figure 3.1: The articulatory apparatus [from [HAH01]]

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 mostly 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*.

During the speech process the articulators are in constant motion and are changing repeatedly between a relative open and a relative closed configuration. Sounds which are produced by a rather open configuration are called *vowels* while sounds that are produced by a closed configuration are called *consonants*.

Vowels are mainly characterized by the position of the tongue in the mouth cavity and whether the lips are rounded or not. The most distinctive point of the tongue in the case of vowels is the horizontal and vertical position of its highest point, called the *dorsum linguae*. Vowels are allways voiced.

3.2 ASR as Pattern Recognition Problem

The goal of automatic speech recognition is to convert the sound wave of human speech to the sequence of words spoken by the producer of the sound wave. For this purpose the sound wave is usually recorded using a microphone and then digitized 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 a *preprocessing* unit of the speech recognition system into a sequence of *feature vectors*.

Automatic speech recognition, as it is treated by state of the art LVCSR systems, is essentially a pattern recognition problem. The pattern to be classified is the sequence of feature vectors. The class that the feature vector is supposed to be assigned to is the correct word sequence that belongs to the pattern, selected from the set of all possible word sequence.

The act of finding this class is often either called *decoding*, because if decodes the feature vectors sequence into a word sequence, or *search* because it searches for the correct word sequences among all possible word sequences.

Human speech is highly variable. Recordings of the sound wave of the same sequence of words will always look different. Variations can occur for a wide variety of reasons: different speakers, different speaking rates, different acoustic environments, different microphones, different emotional states of the speaker, different speaking styles, etc. But even if the same speaker utters the same word sequence under exactly the same circumstance, the sound waves will look different. For that reason speech recognition systems use statistical methods for recognizing speech [Jel97].

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

$$P(W|X) = \frac{p(X|W) * P(W)}{p(X)} \tag{3.1}$$

$p(X)$ 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*.

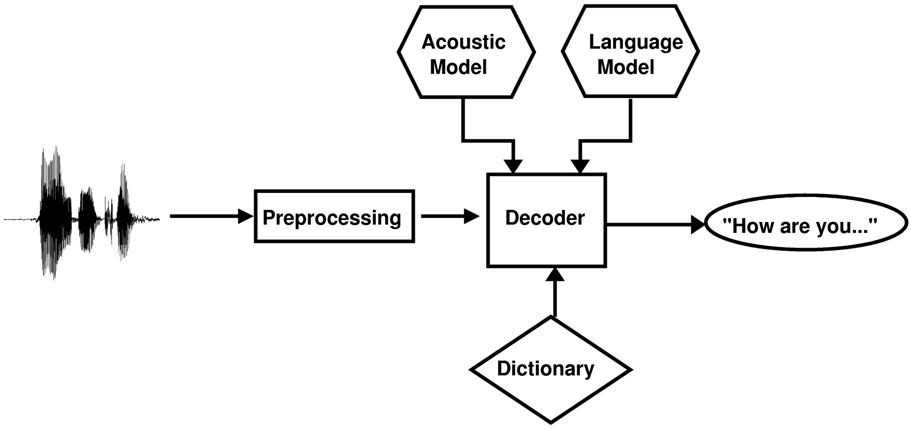


Figure 3.2: System overview of a statistical speech recognition system

Thus the decoder now tries to find:

$$\begin{aligned}\widehat{W} &= \underset{W}{argmax}\, P(W|X)\\ &= \underset{W}{argmax}\, \frac{p(X|W) * P(W)}{p(X)}\\ &= \underset{W}{argmax}\, p(X|W) * P(W) \qquad (3.2)\end{aligned}$$

Usually the search space is limited by a dictionary that defines the set of allowed words of which W can be composed. Figure 3.2 gives a schematic overview of the resulting speech recognition system.

3.2.1 Preprocessing

Goal of the preprocessing unit is to transform the recording of the speech signal into a series of feature vectors which are suitable for performing speech recognition. After the *digitization* of the signal—for ASR purposes a resolution of 16bit and a sampling frequency of 16kHz is common—the relevant information in the signal is emphasized while useless information is discarded.

A typical preprocessing used in LVCSR might look like this: The signal is processed by a short-time fourier analysis executed on overlapping windows of 16-20ms length, and a window shift of 10ms. Often Hamming or Hanning windows are used for this. From the fourier analysis only the power spectrum is used. The power spectrum is then scaled by a Mel-Filterbank and transformed into the cepstral domain by a cosine transform. Several consecutive vectors are then concatenated to a higher dimensional vector whose dimension is reduced again by a linear discriminant analysis transformation.

Additional methods for normalization such as *vocal tract length normalization* (VTLN) [ZW97] or *feature space constraint maximum likelihood linear regression* (cMLLR) [Gal97] can be applied.

Since the acoustic front-end is mostly independent of the language addressed, it is not of interest in this work. A more detailed discussion on the topic of pre-processing audio recordings for speech recognition can be found in [HAH01, RS78].

3.2.2 Acoustic Models based on HMMs

Current state-of-the art systems for automatic speech recognition [LGA+07, SAB+07, SFKW07, WSK07] utilize *Hidden Markov Models* (HMMs) [Rab89] for their acoustic models. An HMM models a sequence of states as a two-fold stochastic process in which only the output from the states are observable but not the states that emit them. So, an HMM models a process which runs through a sequence of states $Q = q_1, q_2, \ldots, q_t, q_{t+1}, \ldots, q_T$ When entering a state, that state emits a symbol with a probability that is dependent on the current state. The proability to enter a state at time $t+1$ only depends on the state at time t. An HMM is formally defined as a quintuple consisting of:

1. The output alphabet V. V can either contain M discrete elements, $V = \{v_1, \ldots v_M\}$, or can be a continuous space, leading to either *discrete* or *continuous* HMMs. In LVCSR V is usually an n-dimensional space of real numbers $\mathbb{R}^n$. Here n is the dimension of the feature vectors that are the result of the pre-processing.

2. A set S of N states $S = \{S_1, \ldots, S_n\}$. In ASR these will be the atomic models for human speech. The models used in LVCSR will be explained in more detail below.

3. A state transition probability $A = \{a_{ij}\}$, with $a_{ij} = P(q_{t+1} = S_j | q_t = S_i), 1 \leq i, j \leq N$. This is the probability of making a transition from S_i to S_j in a discrete time step.

4. An emission probability distribution b for every state i: $b_i(k) = P(v_k \; at \; t | q_t = S_i), 1 \leq i \leq N, \; 1 \leq k \leq M$. This is the probability that when in

state j the symbol v_k is emitted. In the case of a continuous HMM the emission probability distribution becomes a probability density function: $b_i(v) = p(v\ at\ t|q_t = S_i), v \in V$

5. Initial state distribution $\pi = \{\pi_i\}$ with $\pi_i = P(q_1 = S_i)$. This is the probability of being in state π_i at the beginning of the process.

The set of states S and the transition probability A define the *topology* of the HMM. When transitions are allowed from every state into every state we get an *ergodic* HMM, and $a_{ij} > 0\ \forall a_{ij} \in A$. If the number of permissible transitions is limited, as it is the case for many HMMs, some of the transition probabilities become 0: $\exists a_{ij} \in A : a_{ij} = 0$.

In LVCSR mainly continuous HMMs are being used. For the emission probability distributions mainly *Gaussian Mixture Models* (GMMs) are used. A GMM is a linear combination of several Gaussian densities $\mathcal{N}$. For the case that every state has its own GMM with w Gaussian densities we get $b_i(v) = \sum_{j=1}^{W} w_{ij}\mathcal{N}(v, \Sigma_{ij}, \mu_{ij}) = \sum_{j=1}^{W} w_j \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(v-\mu)^T\Sigma^{-1}(v-\mu))$. Sometimes the states in an HMM do not have a complete GMM of their own. Rather a global pool of Gaussian densities $\mathcal{N}_1, \ldots \mathcal{N}_L$ exists that is shared by all states. Only the weights w_i are specific to the state. This kind of HMM is called *semi-continuous* HMM and b becomes: $b_i(v) = \sum_{j=1}^{W} w_{ij}\mathcal{N}(v, \Sigma_j, \mu_j)$.

3.2.2.1 From a Word Sequence to the HMM

In order to model $P(X|W)$ using HMMs the word sequence W needs to be transformed into a sequence of HMM states. To do so, the word sequence is first transformed into single words which are connected from left-to-right in a linear way. Words are then decomposed into phoneme sequences which are also connected from left-to-right in a linear way. The mapping from a word to its phoneme sequence is performed with the help of a *pronunciation dictionary* that lists the phoneme sequences of all words.

Phonemes are further subdivided into smaller units. This subdivision is reasonable in order to account for the fact that the articulators are in constant motion during the pronunciation of a phoneme. Thus the speech signal of a phoneme changes accordingly over time. In state-of-the art systems phonemes are usually sub-divided into three states, a begin, a middle, and an end state.

Figure 3.3 shows this process of transforming a word sequence into an HMM.

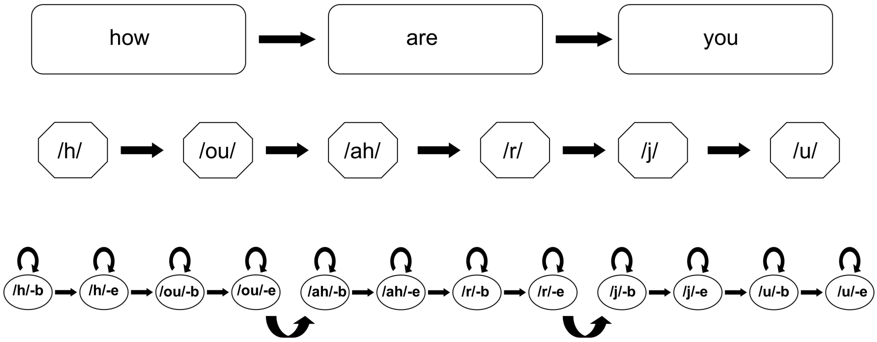


Figure 3.3: Constructing an HMM from a word sequence

3.2.2.2 From Context Independent to Context Dependent Models

As we already described in 3.2.2 the emission proabilities of the states in these HMMs are usually modelled by Gaussian Mixture model. For continuous HMMs every state—so in our case every subphoneme—receives its own mixture model. In the case that the HMMs are constructed from the words as described in the previous section, subphonemes are modelled independently of their context. Every state belonging to the same sub-grapheme uses the same Gaussian mixture model independently of the context that it occurs in.

This way of modeling speech is inaccurate. In reality the acoustic manifestation of a phoneme is highly dependent on the context in which it is spoken. This is due to the inertia and physiological constraints of the articulators. Depending on the phoneme sequence spoken, the articulatory targets of the different phonemes are only reached to varying degrees. Features are either less well articulated, changed, or sometimes even omitted.

In LVCSR this phenomena is modelled by the introduction of context dependent phone models, called *polyphones.* A polyphone is a phoneme in a specific context. So different phonemes or subphonemes are modelled by separate states depending on the context in which they occur. LVCSR usually considers context of either one or two phonemes to the left and the right of the phoneme that is being modeled. Polyphones for the former case are called triphones, those for the latter case quinphones.

The size of a typical phoneme set used in LVCSR is in the range of 50 different phonemes. If using triphones this would lead to a total of 125,000 possible triphones. In order to robustly estimate the parameters of a Gaussian mixture model for a triphone, one needs to observe roughly one hundred, better several hundreds, of samples of that triphone. So it becomes clear that under normal circumstances it is impossible to collect sufficient amounts of training data to train a model that consists purely of polyphones.

Therefore, the models of the polyphones are usually tied in LVCSR [YW93]. For our recognizers we use a top-down clustering approach. For all sets of polyphones that have the same center phoneme a separate *classification and regression tree* is grown. The tree gets as an input a polyphone whose model we are looking for. The leaves of the tree correspond to the final models of a polyphone. In the nodes of the tree questions are asked about the context of the polyphone. The questions are linguistically motivated and ask about the phonetic properties of the phonemes in the polyphone context, e.g. whether a phoneme is voiced or not. Since separate trees are grown for polyphones with different center phonemes, only polyphones with the same center phoneme will potentially share the same model. For growing the tree we us a entropy gain as distance measure between two clusters [Lee88]. The clustering process is stopped when a minimum number of training samples per model or a maximum total number of models is reached.

3.2.2.3 Parameter Estimation

The parameters of an HMM in LVCSR are usually trained on large amounts of speech recordings—tens to thousands of hours of speech—that are transcribed at the word level. The topology and the parameters of an HMM cannot be simultaneously optimized [Rab89]. Instead, the topology of the HMM is predefined by the designer of an ASR systems, usually as described in the previous paragraph, and the parameters for the transition probabilities as well as the emission probabilities are estimated from the training data.

For estimating these parameters most commonly the *Expectation Maximization* (EM) algorithm is used [Dem70]. EM is performed with the help of the *Baum-Welch algorithm* which uses the *forward-backward algorithm* [BPSW70, Jel90]. The Baum Welch algorithm collects sufficient statistics on the training data and then modifies the parameters of the Gaussian mixture models and the transition probabilities using the Baum-Welch update rules in such a way that the probability of the models on the training data is maximized. The EM algorithm is an iterative approach. It can be shown that the algorithm converges towards a local maximum with every iteration. If the parameters estimated do not change after an iteration, a local maximum has been reached.

In addition to the EM algorithm further discriminative training methods exist, that try to maximize the posterior probability. These methods, such as the *Maximum Mutual Information* (MMI) training or *Minimum Phone Error* (MPE) training are normally applied after EM training [Pov04, Sch00a].

3.2.3 Language Model

In LVCSR systems the probabilities for the language model $P(W)$ in equation (3.1) are normally modelled with the help of statistical n-gram language models. These models are based on the decomposition of $P(W)$ into a product of probabilities of smaller word sequences [Jel90]:

$$P(W) = \prod_{i=1}^{n} P(w_i|w_1, \ldots, w_{i-1}) \tag{3.3}$$

$w_1, \ldots, w_{i-1}$ is usually called the history of word w_i. This decomposition is useful for ASR applications since it refelects that the problem of finding the correct word sequence is linear in time, and provides good intermediate results and quality estimates for a partially decoded word sequence.

The probabilities $P(w_i|w_1, \ldots, w_{i-1})$ are estimated by occurrence counts in large collections of texts. However, even for comparatively small numbers of i the number of possible word histories becomes very large, and it is not possible to collect sufficient amounts of text data in order to get reliable estimates for all possible combinations of words and word histories. Therefore, the word histories are clustered into equivalence classes. N-gram language models form these clusters by limiting the length of the history to k words:

$$P(W) = \prod_{i=1}^{n} P(w_i|w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{n} P(w_i|w_{i-k}, \ldots, w_{i-1}) \tag{3.4}$$

Language models that use $k = 1$ are called bigram language models, those that use $k = 2$ trigram, etc. Modern systems use trigram to 5-gram models for their language models, depending on the task and available training data.

Even for trigrams it frequently happens that not a all combinations of a word and all possible histories are observed in training. These combinations would thus be assigned a probability of 0 and would be impossible to recognize. This is undesirable, since an ASR system should also be able to recognize word sequences which have not been seen during training. To circumvent this problem, a back-off technique is used which falls back to shorter word histories, in order to estimate the probability of a word and its context which have not been seen in training. In order to avoid numerical problems with this scheme, the back-off probabilities are estimated differently than for the full ngrams [KN95].

3.3 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 Universität Karlsruhe (TH) 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 transparent, yet efficient objects.

We used version 5 of the JRTK 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.

CHAPTER 4

Graphemic Acoustic Models

4.1 The Role of Pronunciation Dictionaries

As described in Chapter 3 the acoustic models of current state-of-the-art speech recognition system usually use phonemes or sub-phonemic units as states in their HMMs. In order to build an HMM given the written representation of a word or word sequence, a pronunciation dictionary is used which maps the orthography of a word to its corresponding phoneme sequence. This makes the pronunciation dictionary a central component of an automatic speech recognition system. Often the mapping between a word and its phoneme sequence is not unique. For that case it is possible to add several alternative pronunciation variants for a word to the dictionary.

Several methods exist in order to create a pronunciation dictionary. The most common methods currently used are [ADL06]:

- manually defining the mapping for every word
- manually defining a set of rules that generates this mapping, and possible manual post editing to compensate for deficiencies in the rules
- using machine learning techniques to automatically learn the mapping from a set of manually annotated training material, either in a batch or

incremental mode; again manual post-editing may be used to compensate for deficiencies in the rules learned

All these methods require an expert in the target language either for defining the rules, annotating the training material, or doing the post editing and error correction of dictionary entries. Often manual and automatic generation methods are combined, in order to generate an effective dictionary building process [SMT04]. In order to learn the pronunciations in a semi-automatic way, an iterative approach can be used, to first label the 200-500 most frequent words, then learn letter-to-sound rules on them, and then predict the pronunciation of new chunks of 100 or so words. After every chunk the newly produced pronunciations need to be manually checked and corrected. Using the extended dictionary, the letter-to-sound rules are learned again and the process is iterated with the next chunk of new words. Once a sufficient accuracy in predicting the pronunication of new words is reached, the learning is terminated and the letter-to-sound rules are kept fix and used for processing the remaining words [Bla06].

Considering that dictionaries of modern large vocabulary ASR systems can contain tenth of thousands to up to several houndreds of thousands words, it becomes clear that dictionary creation can become a very time consuming process. Also, the large amount of manual labor of a dedicated expert is very costly in terms of money. In order to facilitate the dictionary creation process, some tools try to simplify the process of manually checking the dictionary by the expert in a language in such a way, that the expert does not need to be a trained phonetician, but that an interested native speaker of the language can perform this task under the supervision of a technology expert [SBB+07, DB04b, DB04a]. But, when pressed for time and faced with a less prevalent language it can be quite difficult for the creators of an ASR system to even find an available expert in the target language, even if any interested native speaker can take that role.

Thus, eliminating the need for a pronunciation dictionary can greatly simplify and reduce the cost of the creation of an ASR system for a new language, as well as speed up the creation process.

4.2 Acoustic Modelling for Graphmes vs. Phonemes

One way to eliminate the need for a pronunciation dictionary or to make its creation a trivial task, is to substitute the phonemes as modeling units for the HMM. An alternative to modeling phonemes is to directly model the *graphemes* of the words to be recognized. A *grapheme* is defined as the fundamental unit in written language [Wikb]. E.g. the letter of the English alphabet are considered graphemes, or the characters of the Chinese script. The term grapheme is modelled after the terms phoneme or morpheme which refer to significant units of sound and meaning respectively. However, [DB96] points out that writing and language bear so many fundamental differences that the usual meaning of the suffix *-eme* does not apply in the term grapheme. Many attributes of phonemes do not find any parallel in graphemes and vice versa.

When using graphemes or subgraphemes as modeling units instead of phonemes the creation of the HMM from the written representation of the word becomes a trivial task, eliminating the need for a pronunciation dictionary as mapping function in order to find the correct HMM state sequence.

4.2.1 Requirements for Modeling Units in ASR

While using graphemes as units after which to construct the HMM corresponding to a word eliminates the costly creation of a pronunciation dictionary, it raises the question of the suitability of graphemes as models for the HMM, especially considering the remarks on the fundamental differences between graphemes and phonemes. In order to judge the suitability of graphemes as modeling units, we first take a look at the reasons for using phonemes as modeling units in ASR. The use of phonemes is generally motivated by two reasons:

The first reason is that phonemes are closely related to the acoustic manifestations of the part of speech that they represent. Human speech is generally described as a sequence of phonemes by phoneticians and linguists. Though phonemes can consists of several allophones, their pronunciation, that is the acoustic manifestation as measured by the microphone, is in general very similar. Therefore, it is assumed that one model per phoneme for recognition purposes is capable of learning the acoustic pattern that it produces and to discriminate it against the acoustic pattern of the other phonemes in that language. Where that is not the case, pronunciation variants can be used to cover several different phone variants for a phoneme. Variations in the pronunciation of phonemes, which are introduced by such effects as coarticulation, are treated by using context-dependent models. They model a phoneme that occurs in a specific context.

The second reason, why phonemes are used as modeling units, is the fact that their number per language is rather limited as compared to the number of words or even syllables per language. In order to be able to robustly estimate the parameters of the models, it is necessary to collect sufficient amounts of training examples per model, preferably from several different speakers. For example, for words in large vocabulary speech recognition it is often not feasible to collect sufficient numbers of training examples in order to use them as modeling units. Using the set of phonemes of a language, however, it is possible to compose all words in that language from them. At the same time, they are occurring so frequently that by collecting a reasonable amount of acoustic training data, enough training examples per phoneme are observed in order to train a model for them.

In order to apply the modeling techniques that have been used for phonemes in speech recognition so far to graphemes, graphemes need to fulfill the same two properties as phonemes. They need to bear a sufficiently close relation to their pronunciation and which should to a certain degree be consistent per grapheme. Further, the number of graphemes per language needs to be limited to a number that is small enough so that it is possible to collect sufficient amounts of training data per grapheme for robust model parameter estimation.

As we will see, this is not the case for all writing systems in the world, but for a large number of them. For these languages substituting phonemes by graphemes as modeling units is a suitable approach.

4.2.2 Writing Systems of the World

In [DB96] Daniels defines a writing systems as “a system of more or less permanent marks used to represent an utterance in such a way that it can be recovered more or less exactly without the intervention of the utterer”. In that way a writing system preserves speech over time and distance. The basic units of a writing system are called graphemes. The collection of all graphemes of a language is called its script. Graphemes refer to minimally significant elements. In that way they are similar to phonemes. Different classifications of writing systems exist in literature. Writing systems are generally classified with respect to the concepts which their graphemes represent.

4.2.2.1 Typology of Writing Systems

In literature a wide variety of typologies of writing systems have been proposed. Often, writing systems are classified as either logographic, syllabic, or alphabetic. Dating from 1883 when it was probably first laid out by Isaac Taylor, it has been and still is the most popular one [DB96]. However it leads to problems when applied to certain languages leading to unlikely suggestions about their script. Thus many alternative typologies have been proposed. One typology that can be often found is presented by Crystal in [Cry87]. He divides writing systems into those which have a clear relationship between the symbols in the script and the phones in the language, and those were this not the case. As an example for the latter Crystal lists writing systems using pictograms. However when following the definition from Daniels, pictograms are not considered a writing system, but rather a forerunner to writing systems, because writing is bound up with language by that definition. Pictography cannot capture abstract notions, many verbs, grammatical inflections, particles, and names. In order to be able to do this, a script must represent the sounds of a language.

Since in this work we are dealing with the science of automatic speech recognition, only writing systems are of interest to us that allow the recording of the sounds of a language. We thus follow the typology of writing systems by Daniels, rathern than by Crystal.

According to Daniels a writing systems can be of one of the following six types:

- Logosyllabary: Characters represent individual words or a particular syllable.
- Syllabary: Characters represent particular syllables.
- Abjad or Consonantory: Characters denote consonants. The name abjad is derived from the first letters of the most common example, the Arabic script.
- Alphabet: Characters denote vowels or consonants.
- Abugida: Each character denotes a consonant accompanied by a specific vowel. Other vowels are represented by a consistent modification of the consonant symbol. The word abugida is Ethiopic from the first four consonants and the first four vowels of its script.
- Featural: The shapes of the characters correlate with distinctive features of the segment of the language that they represent.

Chinese is a prominent example of a language with a logosyllabary writing system. The Japanese hiragana and katakana are examples of syllabary scripts. Arabic has already been cited as a prominent abjad, also the Hebrew script falls into this category. The European languages mostly use alphabets, often based on the Latin alphabet. The Indic script, used for many languages in South and Southeast Asia is an abugida. Korean is a language that uses a featural writing system.

4.2.2.2 Distribution of Writing Systems

As we will discuss below, not every type of writing system is equally suited for grapheme based ASR. Therefore, it is interesting to get an overview of the distribution of the different types of writing systems in the world. Omniglot [Omn] lists 3 currently used abjads, 20 alphabets, 28 abugidas, 19 syllabaries, and 2 logosyllabaries. In the Omniglot classification the featural Korean script is classified as an alphabet. Table 4.1 shows the writing systems as listed by Omniglot, the native name of the script written in its own characters, as well as the languages known to use that script.

Table 4.1: Writing systems of the worlds and the languages that use them as listed by Omniglot[Omn]

Logosyllabaries		
Chinese (Zhngwn)	中文	Modern Standard Chinese, Cantonese, Japanese (kanji), Korean (hanja), Vietnamese(ch-nm)
Syllabaries		
Cherokee (Tsalagi)	ᏣᎳᎩ	Cherokee
Cree (Nhiyaw)	ᓀᐦᐃᔭᐤ	Cree
Hiragana (Japanese)	ひらがな	Japanese
Inuktitut	ᑎᑎᕋᐅᓯᖅ ᓄᑕᐊᖅ	Inuktitut
Katakana (Japanese)	カタカナ	Japanese
Mende		Mende
Naskapi (Innu Aimun)	ᓇᔅᑲᐱ	Naskapi
Ndjuk		Ndjuk
Ojibwe (Anishinaabe)	ᐊᓂᔑᓇᐯ	Ojibwe
Yi (Lolo)	ꆈꌠ	Yi
Abjads		
Syriac	ܠܫܢܐ ܣܘܪܝܝܐ	Syriac
Hebrew	עברית	Hebrew, Judeo-Arabic, Ladino, Yiddish
Arabic	العربية	Arabic, Azeri, Baluchi, Bosnian, Dari, Hausa, Kabyle, Kashmiri, Kazakh, Kurdish, Kyrghyz, Malay, Morisco, Pashto, Persian/Farsi, Punjabi, Sindhi, Siraiki, Tatar, Urdu, Uyghur

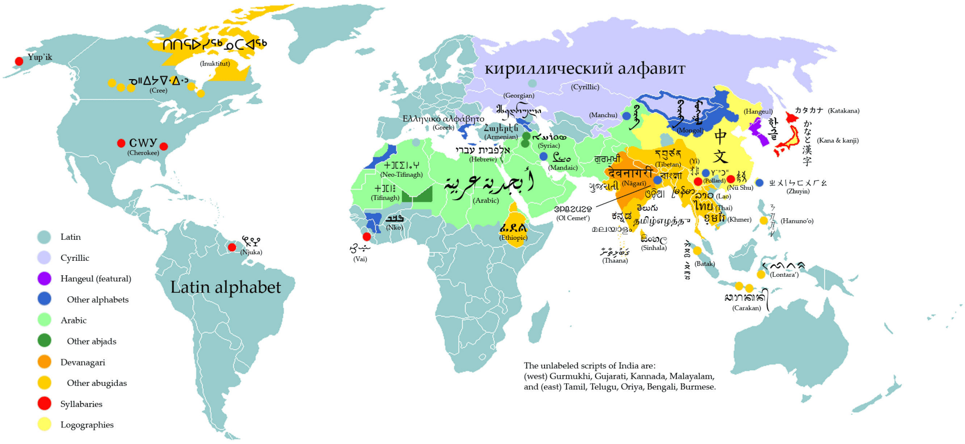
Table 4.1: (continued)

Alphabets		
Armenian	Հայերէն	Armenian
Bassa (Vah)	𖫢𖫧𖫳𖫒	Bassa
Cyrillic	Кириллица	*see Table 4.3*
Fraser	ꓑꓒꓐꓩꓪ	Lisu
Georgian (Mkhedruli)	მხედრული	Georgian, Laz, Svan
Greek	Ελληνικά	Greek
Kayah Li	ꤊꤢ꤬ꤛꤢ꤭ ꤜꤟꤤ꤬	Kayah Li
Korean (hangl)	한국어	Korean
Latin/Roman	Latin/Roman	*see Table 4.2*
Manchu	ᠮᠠᠨᠵᡠ	Manchu
Mongolian	ᠮᠣᠩᠭᠣᠯ	Mongolian
N'Ko	ߒߞߏ	Malinke, Bambara, Dyula
Ol Cemet'/Ol Chiki (Santali)	ᱚᱞ ᱪᱤᱠᱤ	Santali
Oirat Clear Script	ᡐᡆᡑᡆ	Kalmyk
Pollard Miao	𖼀𖼰𖼺𖽑	A-Hmao
Tai Dam	ꪼꪕꪒꪾ	Tai Dam
Thaana	ދިވެހި ބަސް	Dhivehi (Maldivian)
Tifinagh	ⵜⵉⴼⵉⵏⴰⵖ	Kabyle, Tamazight
Abugidas		
Bengali	বাংলা	Bengali, Assamese
Buhid	ᝊᝓᝑᝒ	Buhid
Burmese/Myanmar	မြန်မာစာ	Burmese/Myanmar
Cham	ꨌꩌ	Cham
Dehong Dai	ᥖᥭᥰᥖᥬᥳᥑᥨᥒᥰ	Dehong Dai
Devangar	देवनागरी लिपि	Hindi, Marathi, Nepali, Pali, Sanskrit, Sindhi
Ge'ez (Ethiopic)	ፊደል	Amharic, Ge'ez, Tigrinya
Gujart	ગુજરાતી	Gujart, Kachchi
Gurmukhi (Punjabi)	ਗੁਰਮੁਖੀ	Panjabi
Hanuno	ᜱᜨᜳᜨᜳᜢ	Hanuno
Hmong	𖬌𖬣𖬵 𖬖𖬲𖬞 𖬅𖬮	Hmong
Kannada	ಕನ್ನಡ	Kannada
Khmer	ភាសាខ្មែរ	Khmer
Lanna	ᨲᩫ᩠ᩅᨾᩮᩥᩬᨦ	Northern Thai (Kam Mu'ang), Tai Lue, Khn
Lao	ພາສາລາວ	Lao

Table 4.1: (continued)

Lepcha (Rng-Rng)	ᰛᰩᰵᰛᰧᰵ	Lepcha (Rng-Rng)
Limbu	ᤛᤡᤖᤡᤈᤨᤅ	Limbu
Malayalam	മലയാളം	Malayalam
Manipuri	ꯃꯤꯇꯩ ꯃꯌꯦꯛ	Manipuri
New Tai Lue	ᦟᦲᧅ ᦷᦎ ᦺᦑ	Lue
Oriya	ଓଡ଼ିଆ	Oriya
Sinhala	සිංහල	Sinhala, also for Pali and Sanskrit in Sri Lanka
Sorang Sompeng	𑃐𑃚𑃝𑃗 𑃐𑃚𑃝𑃗	Sora
Tamil	தமிழ் எழுத்து	Tamil
Telugu	తెలుగు	Telugu
Thai	ภาษาไทย	Thai
Tibetan	བོད་སྐད།	Tibetan, Dzongkha (Bhutanese)
Varang Kshiti	𑣁𑢼𑣜 𑣏𑣂𑣕𑣂	Ho

Figure 4.1 illustrates the geographic distribution of the different writing systems in the world. With regard to geographic distribution alphabets clearly dominate. Most parts of North and South America, Europe, the parts of Asia that belonged to the former Soviet Union, and the Southern half of Africa use some form of alphabet mostly Latin/Roman or Cyrillic based. Arabic dominates large parts of Northern Africa. Large parts of Asia use logosyllabaries, syllabaries, and abugidas. For graphemes their relation to their pronunciation is not that well defined. Depending on the language in question the relationship between graphemes and corresponding phonemes can vary widely. Some languages show a very close relationship of their graphemes to the corresponding pronunciation. For other languages their is a very complex correspondence of the written form and the pronunciation. In Chinese, for example, the mapping of the characters to the pronunciation require extensive knowledge of the context of the characters and the resulting semantic meaning of the word sequence to be spoken. Chinese is also an example for a semanto-phonetic writing system, called that way because graphemes are used to transport pronunciation and meaning. This results in a very large number of graphemes for these languages. In Chinese, for example, there are over 20,000 different characters with 10,000 in use today. When using graphemes as modeling units for these kind of languages, the same problem arise as when using words or syllables as modeling units, that is the problem of finding sufficient amounts of training examples for estimating the model's parameters. In this case the use of graphemes as modeling units is an inappropriate and infeasible approach.



The unlabeled scripts of India are:
(west) Gurmukhi, Gujarati, Kannada, Malayalam,
and (east) Tamil, Telugu, Oriya, Bengali, Burmese.

Figure 4.1: Geographic distribution of writing systems [Wika]

Many writing systems in the world, however, make use of alphabets with a limited set of graphemes.In our tables we listed 18 writing systems that are alphabets. These 18 alphabets are used by 292 languages in the world.

Languages using the Roman/Latin alphabet
Abenaki, Afaan Oromo, Afar, Afrikaans, Ainu, Akan, Alabama, Albanian, Aleut, Alsatian, Apache, Aragonese, Aranese, Arapaho, Aromanian, Arrernte, Asturian, Aymara, Azeri, Bambara, Basque, Belarusian, Bemba, Bikol, Bislama, Breton, Burushaski, Catalan, Cayuga, Cebuano, Chamorro, Chavacano, Chechen, Cheyenne, Cimbrian, Chichewa, Chickasaw, Choctaw, Comanche, Cornish, Corsican, Cape Verdean Creole, Croatian, Czech, Danish, Dawan, Delaware, Dholuo, Dinka, Drehu, Duala, Dutch, English, Esperanto, Estonian, Ewe, Ewondo, Faroese, Fijian, Filipino, Finnish, Folkspraak, French, Frisian, Friulian, Ga, Gagauz, Galician, Ganda, Genoese, German, Gooniyandi, Greenlandic, Guadeloupean Creole, Guarani, Gugadja/Kukatja, Gwichin, Haida, Haitian Creole, Hn, Hausa, Hawaiian, Herero, Hiligaynon, Hixkaryana, Hopi, Hotck, Hungarian, Icelandic, Ido, Igbo, Ilocano, Indonesian, Interglossa, Interlingua, Iupiaq, Irish, Italian, Jamaican Creole, Jrriais, Kabyle, Kaingang, Kala Lagaw Ya, Kapampangan, Karakalpak, Karelian, Kashubian, Kinyarwanda, Kiribati, Kirundi, Klallam, Klamath, Kurdish, Kwakiutl, Lingala, Latin, Latvian, Lingua Franca Nova, Lithuanian, Livonian, Lojban, Lombard, Low Saxon, Luxembourgish, Maasai, Makhuwa, Malagasy, Malay, Maltese, Manx, Mori, Marshallese, Meriam Mir, Mi'kmaq, Mirandese, Mohawk, Montagnais, Murrinh-Patha, Nagamese, Nahuatl, Nama, Naskapi, Navajo, Naxi, Neapolitan, Ngiyambaa, Noongar, Norwegian, Novial, Occidental, Occitan, Okinawan, O'odham, Old Norse, Ossetian, Papiamento, Piedmontese, Pirah, Pitjantjatjara, Polish, Portuguese, Potawatomi, Quechua, Rarotongan, Rotokas, Romanian, Romansh, Romany, Rotuman, Saami/Sami, Saanich, Samoan, Sango, Sardinian, Scots, Scottish Gaelic, Shavante, Shawnee, Shona, Sicilian, Silesian, Sioux, Slovak, Slovene, Slovio, Somali, Sorbian, Southern Sotho, Spanish, Swahili, Swedish, Tagalog, Tahitian, Tatar, Taiwanese, Tetum, Tlingit, Tok Pisin, Tongan, Turkish, Turkmen, Tuvaluan, Tuvan, Twi, Uyghur, Venetian, Vietnamese, Volapk, Vro, Walloon, Warlpiri, Waray-Waray, Wayuu, Welsh, Wik-Mungkan, Wiradjuri, Wolof, Xhosa, Yapese, Yindjibarndi, Yolngu, Yoruba, Zhuang, Zulu

Table 4.2: Languages using the Latin Alphabet

Languages using the Cyrillic alphabet
Abaza, Abkhaz, Adyghe, Avar, Azeri, Balkar, Bashkir, Belarusian, Bulgarian, Buryat, Chechen, Chukchi, Chuvash, Crimean Tatar, Dargwa, Dungan, Erzya, Even, Evenki, Gagauz, Ingush, Kabardian, Kalmyk, Karakalpak, Kazakh, Khanty, Kildin Sami, Komi, Koryak, Kumyk, Kurdish, Kyrghyz, Lak, Lezgi, Lingua Franca Nova, Macedonian, Mansi, Mari, Moksha, Moldovan, Mongolian, Nanai, Nenets, Nivkh, Old Church Slavonic, Ossetian, Russian, Ruthenian, Serbian, Slovio, Tabassaran, Tajik, Tatar, Tsez, Turkmen, Tuvan, Ubykh, Udmurt, Ukrainian, Uyghur, Uzbek, Votic, Yakut, Yukaghir, Yupik

Table 4.3: Languages using the Cyrillic Alphabet

4.2.3 Suitability of Graphemes as Modeling Units

A good indication for the suitability of graphemes as modeling units is when they fulfill the same two properties that we listed as motivation for the use of phonemes in Subsection 4.2.1. Here, the requirement that the number of graphemes per language is small enough so that it is possible to collect sufficient amounts of training data per grapheme for robust model parameter estimation is easier to check for a writing system than the first property, the relation between grapheme and acoustic manifestation of the corresponding sound.

Logosyllabary writing systems are normally not suited as modeling units in ASR systems because they usually contain an extensively large number of graphemes. Due to the fact that graphemes mostly represent whole words, their number per language is too large to collect sufficient training material per model. For example for Chinese there is no theoretical upper limit to the number of characters. In newspapers approximately 3,000 characters are sufficient for reading. For Chinese literature or technical documents approximately 6,000 characters are necessary.

Alphabets and abjads on the other hands are well formed in that respect for grapheme based ASR. The graphemes in abjads and alphabets represent normally vowels and consonants of their language and thus the number of graphemes in these writing systems is in the range of the number of phonemes of the corresponding language. Therefore, for these writing systems it is possible to collect a sufficient number of training sampels per model.

With respect to the correspondence between the graphemes in abjads and alphabets and their acoustic manifestation as measured by the microphone a general judgment is not possible, but dependent on the language in question. Since abjads only write consonants but not vowels this is clearly a disadvantage. Vowels need to modelled implicitely in the consonants' models. Thus the models must be able to discriminate not only the acoustics of the different consonants. Instead, each model must learn the acoustics of all vowels that can come with the consonant in question. Also, all combinations of a consonant and the vowels that can theoretically come with it must be observed during acoustic model training. Therefore, more training material must be collected than if vowels would be written explicitely as in alphabets. Thus alphabets have an advantage over abjads in that regard.

Alphabets and abjads also show a common difficulty with regard to the relation between their graphemes and their pronunciation. Though per definition the graphemes in alphabets represent the vowels and consonants and abjads only the consonants of their language, this clear relation was usually only given at the time of their creation. However, as language evolves over time, it often happend that the pronunciation of a language evolved differently from its writing system. The pronunciation and writing of English, for example, have evolved in such a way that, today, they show a very complicated relationship between them. So, it is very difficult to predict the pronunciation of a word from its spelling and vice versa. Other languages show a very strict mapping from graphemes to phonemes, using so called phonemic alphabets. Finnish, Russian, Serbo-Croation or Spanish are good examples for such alphabets.

Abugidas are very similar to alphabets and abjads, as they also denote consonants and vowels by modifying the graphemes for consonants, e.g. by using diacritics or by modifying the shape of the consonant grapheme. Here the advantage over abjads is, that different vowels are marked differently. When they are marked with diacritics it is possible to treat the diacritics as separate modeling units for the ASR system, so that the abugida then shows the same properties with respect to use in an ASR system as alphabets do. In the case where the shape of the grapheme is modified abugidas suffer from one of the problems that abjads have, that is to collect sufficient amounts of training material per (modified) grapheme in order to capture all permissible consonant-vowel combinations.

Syllabaries generally show a very good relationship between their graphemes and their pronunciation, since they directly represent syllables. The number of syllables in a language can be rather large and the syllable structure rather complex. So, depending on the language, the number of graphemes can be significantly larger than in an alphabet. For example, for English with its very complex syllable structure and ample syllables, using a syllabary would be very cumbersome. This is one of the reason why in ASR phonemes have become the preferred modeling units and not syllables. So, syllabaries also show the problem that collecting sufficient training material for all possibly syllables might be difficult. However, often syllabaries are used for writing languages with comparatively few syllables, so that grapheme based ASR is also feasible for them.

Fortunately the writing systems for many languages in the world, Omniglot lists 292, make use of alphabets as writing systems, so that for them the grapheme based approach is promissing. Abjads seem less suitable for ASR but they are currently only in use for 27 languages. For syllabaries, grapheme based ASR might be feasible, depending on the language. Currently ominglot only lists 7 languages that still use syllabaries. Abugidas, that seem more suitable for grapheme based ASR than syllabaries, make up a large portion of the writing systems in the world. Omniglot lists their use for 28 languages. The real number, however, must be higher, as Omniglot does not list all Indic scripts.

4.3 Related Work

Past research has demonstrated for several languages that the use of graphemes as modeling units, instead of phonemes, can be a suitable approach to ASR.

In [STNE$^+$93] Schukat-Talamzzini et al. demonstrated the use of graphemes as modeling units for German on a train scheduling task. They utilized context-dependent models and used a back-off scheme for models that were not seen during training.

Schillo et al. [SFK00] also experimented with German grapheme based models, targeting an isolated word recognition task of German city names. They used context-independent grapheme models and trigrapheme models for their experiments. A back-off scheme was not necessary, since only trigraphemes that were seen during training, were used during recognition.

Kanthak and Ney [KN02] experimented with context-dependent grapheme based acoustic models on Dutch, Italian, German, and English. In contrast to previous work they used context-dependent models with decision tree based HMM state-tying as described in Chapter 3. Decision trees for HMM state-tying require a set of questions to ask about the context of the models, traditionally about the phonetic properties of the context. The properties of the phonetic context are used because of their influence on the pronunciation of the phoneme that is modelled in this context. In the case of graphemes is not clear what properties of the context of a grapheme that is to be modelled context-dependently do really influence its pronunciation. Especially it is not clear how to derive such properties from the graphemes without any knowledge about the relation between the graphemes and their pronunciation. Therefore, [KN02] compared the use of manually derived phonetic questions for graphemes to automatically generated ones and detected a slight increase in word error rate for the automatically generated questions.

In [KN03] Kanthak and Ney expanded their work to multilingual grapheme based models, showing improvements for German with the multilingual acoustic models. They commented that the decision tree used for HMM state tying, also captures in part the grapheme to phoneme relation of a language.

At the same time Killer et al. [KSS03] experimented with context-dependent, grapheme based models for the languages English, German, and Spanish on a large vocabulary dictation task. They also examined different types of questions in the decision tree for the HMM state-tying, and found that simply asking for the identities of the graphemes in the context of the polygraphemes works better than phonetically motivated or automatically derived questions.

Charoenpornsawat et al. [CHS06] demonstrated the use of grapheme based acoustic models for Thai grapheme based ASR. In order to get a maximum of performance they applied a text normalization scheme to the Thai graphemes that makes use of detailed knowledge of the grapheme-to-phoneme relation in Thai.

Sung et.al. examined the behavior of grapheme based ASR system when large amounts of training data are available [SHBS09]. They worked on the GOOG-411 task, which uses ASR and web search to help people call businesses. They found that for English with increasing amounts of training data the gap between a phoneme based recognizer and a grapheme based one closes. In that way they wanted to show that the pronunciation modeling knowledge that is usually encoded in the pronunciation dictionary can be learned by the acoustic model. They found the grapheme system especially beneficial because it allowed for easy addition of the many out of vocabulary words that they encountered on their task, such as proper names of businesses and people.

4.4 Monolingual Grapheme Based Recognizers

As basis for our further work we trained monolingual grapheme based recognizers in the five languages English (EN), German (GE), Russian (RU), Spanish (SP), and Thai (TH). For these experiments we assumed as given as little knowledge as possible about the target language and its relation between the graphemes and pronunciation of the words. For example, unlike as in [CHS06] we did not perform any preprocessing on the graphemes of the words in the vocabulary of our recognizers; though this technique is known to boost the performance of grapheme based recognition systems by improving, i.e. homogenizing and simplifying, the relation between the written representation of the words and their pronunciations. However, the necessary knowledge about the languages' writing systems in order to define these kind of pre-processing rules we do not want to take as given.

In selecting the five languages we cover three Latin based alphabets—English, German, and Spanish—, one Cyrillic based—Russian—, and one Abugida—Thai. As discussed above, these languages therefore cover the most promissing kinds of writing systems for use in grapheme based ASR systems.

4.4.1 Corpus and Task

We conducted our experiments on a selection of languages from the GlobalPhone [Sch02] corpus. GlobalPhone is an ongoing data collection effort that now provides transcribed speech data in 18 languages. The corpus has been designed for research in multilingual speech recognition, rapid deployment of speech processing systems in new languages, language and speaker identification tasks, monolingual speech recognition in a large variety of languages, as well as comparisons across major languages based on text and speech data. To achieve this, data collection in all languages covered is done in an uniform way under equal acoustic conditions. The corpus contains read speech by native speakers collected with close talking microphones. The texts read are newspaper articles covering national politics, international politics, and economics. In that way it is modeled after the Wall Street Journal 0 (WSJ0) corpus. The audio was recorded with head-mounted, close-talking microphones using a sampling frequency of 16kHz and a resolution of 16bit. Since English is not part of GlobalPhone, the WSJ0 corpus, which matches the other data, was used for the experiments on English.

For every language three data sets are available: one for acoustic model training (train), one for development work (dev), such as finding the correct language model weight, and one for evaluation (eval). All three sets are speaker disjunct. For our experiments in Chapter 5, German and Thai will receive the role of previously unseen languages to which we will port ASR systems. For these experiments we then only assume a small adaptation (adapt) set of roughly fifteen minutes as given for these two languages.

Table 4.4 shows the size of the individual data sets for the five languages in terms of length in time, number of utterances, and number of speakers.

4.4.2 Preprocessing

The 16kHz, 16 bit audio data was preprocessed by calculating 30 log-mel scaled cepstral coefficients, liftering to 13 coefficients, and concatenation of 6 neighboring feature vectors. The resulting 91 dimensional vector was reduced to 32 dimensions with the use of *linear discriminant analysis* (LDA) [HUN92]. The mean of the cepstral coefficients was subtracted and their variance normalized on a per utterance basis. During decoding *incremental feature space constrained MLLR* (cMLLR) [Gal97] and incremental cepstral mean subtraction and variance normalization on a per utterance basis were performed.

Table 4.4: Size of the data sets for the different languages in hours, number of utterances, and number of speakers

		EN	GE	RU	SP	TH
train	hours	15.0	16.0	17.0	17.6	24.5
	#utt	7,137	9,259	8.170	5,426	12,260
	#spkrs	83	65	84	82	80
dev	hours	0.4	0.4	1.3	2.1	1.3
	#utt	144	199	898	680	613
	#spkrs	10	6	6	10	4
eval	hours	0.4	0.4	1.6	1.7	1.1
	#utt	152	250	1,029	564	568
	#spkrs	10	6	6	8	4
adapt	hours	—	0.25	—	—	0.25
	#utt	—	101	—	—	140
	#spkrs	—	1	—	—	1

4.4.3 Acoustic Model Training

Based on the setup from [KSS03] and [MSS04] the systems were trained from scratch. For every language we used the graphemes of that language as base for our modelling units. Table 4.5 shows that number of graphemes per language that we have used. For every grapheme an HMM with three states was trained, as it is the case for the phoneme based acoustic models.

Initial alignments between the training samples and the HMM states were obtained by uniformly distributing the training samples of an utterances over the states of the HMM of that utterance. From this initial alignments models were initialized with the use of k-means. The resulting models were used to obtain a first forced alignment of the training data.

In a next step, *context-independent* (CI) models were trained. Starting from the first forced alignments several training cycles were performed. Thereby, each training cycle is composed of the following steps:

1. Estimation of LDA transformation matrix

2. Initialization of parameters using the K-Means algorithm

3. Six iterations of label training along the forced alignments

4. Four iterations of EM training starting from the parameters resulting from the label training

5. Calculation of new forced alignments using the newly trained weights

From the resulting CI systems, *context-dependent* (CD) systems were created by a divisive clustering of trigraphemes with the help of a decision tree. The best partition of the set of trigraphemes in each node of the decision tree is determined by the entropy gain criterion. Since it is computationally not feasible to consider all possible partitions in a node during the tree growing process, the set of possible partitions to be searched is given by a set of binary questions that can be applied to the trigraphemes in every node of the cluster tree. [KSS03] found that just asking for the identity of the grapheme in the left or right context of the trigrapheme gives better results than automatic ways of finding good questions. These kind of questions are called *singleton questions*. The use of phonetically motivated questions is not permissible in our case, since we assume that no phonetic knowledge about the target language and its grapheme-to-phoneme relation is available. We therefore used the singelton questions for growing the trigrapheme decision tree, since they give a good performance and do not require any knowledge about the relationship between the writing system and the pronunciation of the corresponding language.

Starting from the forced alignments of the CI systems, the CD models were trained by two iterations of the same training cycle as the CI systems.

In the Thai script, words are normally not separated by white space. For the experiments in this work we therefore worked with automatically segmented data that was provided by [SCB$^+$05].

Table 4.5: Number of graphemes that are modelled per language

	EN	GE	RU	SP	TH
#graphemes	26	29	33	35	69

4.4.4 Language Models

As language models we used statistical trigram models that were trained on in-domain data for the corresponding language, mainly newspaper articles, mostly collected from the web or coming from respective distributions from newspaper publishers on CDs or DVDs.

For English we used the official 64K trigram language model t95+60k1 with 8,814,128 trigrams and 7,454,368 back-off bigrams as it was provided for the official Wallstreet Journal evaluation. This language model was built on a text corpus of more than 300 million words [Rog97].

The German language model was trained on 40 million words of texts coming from the View4You project [Kem99]. The texts include data from the website of the Bayrischer Rundfunk 5 radio station, texts from the Frankfurter Allgemeine Zeitung, and transcriptions of broadcasts of the German news show "Tagesschau".

For Russian the trigram language model was trained on 19 million words of newspaper texts collected from the online editions of six newspapers. The articles are from the period of 1997 to 2004.

The Spanish language model was trained on 62 million words of the Spanish Language News Corpus produced by the Linguistic Data Consortium (LDC).

The trigram language model that was used for Thai was created with the help of the SRI Language Model Toolkit [Sto02] and is an interpolation of a trigram model trained on 3.3 million words of newspaper texts and a trigram model trained on the transcriptions of the training data, amounting to 12 thousand words [Stü08a]. The interpolation weight was chosen by minimizing the perplexity of the language model on the development set.

Table 4.6 summarizes the sizes and perplexities of the language moels for the different languages.

Table 4.6: Size of the corpora for language modeling in number of words, size of the language models in number of ngrams, and perplexity on the development and evaluation set for the different languages

	EN	GE	RU	SP	TH
# words	300 mio.	40 mio.	19 mio.	62 mio	3.3 mio
# 3-grams	6.507.987	1.679.444	4.548.890	14.117.393	360.845
# 2-grams	3.662.939	714.103	1.365.851	2.726.540	454.524
# unigrams	9.222	24.152	24.968	23.074	7.546
PPL on dev	160	424	1243	227	111
PPL on eval	117	443	1098	219	111

4.4.5 Results

4.4.5.1 Grapheme Based Systems

We tested the resulting models on the development and evaluation sets of their respective language. The development set was used as a cross-validation set in order to determine the optimal language model weight and word penalty. Figure 4.2 shows the word error rates of the context-independent models on their respective language, while Figure 4.3 shows the error rates of the context-dependent models.

For the context-independent models the word error rates on the evaluation set range from 28.7% for Thai to 55.8% for Russian. The word error rates for the context-dependent models are in the range of 14.0% for Thai to 39.3% for Russian.

The reason for the generally high WER for Russian is due to the very high perplexity of the language model that results from the highly inflective nature of Russian and its very loose word order [SS04].

When comparing the relative difference in performance between the context-independent and context-dependent models for the individual languages one can observe some differences between the languages. Figure 4.4 shows the relative reductions in WER for the five different languages. The WER for English and German is reduced the most when going from context-independent to context-dependent acoustic modeling. This reflects the fact that the pronunciation of a grapheme in these two alphabets is highly influenced by its graphemic context. Russian on the other side shows by most the least reduction in WER. The reason for that lies in the fact that the relation between the Russian graphemes and their phonemes, pronunciation respectively, is quite straight forward. Only very few exceptions and rules in Russian exist that alter the pronunciation of a grapheme depending on its letter context. Thai and Spanish lie somewhat in the middle between English and German on the one side and Russian on the other side, indicating that the relation between their graphemes and their pronunciation lies somewhat in between.

In general, the gain from context-dependent modeling in the grapheme based cases turns out to be higher than when using phonemes. This is in line with our comments on the suitability of graphemes as modeling units in Subsection 4.2.3 that already commented on the more complex relation between graphemes and their acoustic manifestation than for phonemes. However, the decision tree and the singleton questions used for the context-dependent models seem to be able to a certain degree to implicitly learn the rules for the relation between the graphemes and pronunciation from the training data. The rules that were previously contained in the listing in the pronunciation dictionary are now encoded in the cluster tree and the Gaussian mixture models.

For our experiments in Chapter 6 we also trained speech recognition systems on the same languages that use phonemes as modelling unit instead of graphemes. These systems use the same training data and the same training procedure as the grapheme systems that we described in this section.

Figure 4.5 compares the performance of the grapheme based systems with that of the phoneme based systems on the evaluation sets of the respective languages. As one can see, except for Spanish, the phoneme based systems perform better than the grapheme based systems. For English and Thai the drop in performance is the largest. For English the reason in the large drops lies within the fact, that though English uses an alphabet script, the pronunciation of English has developed over time away from its written form. Therefore today's English shows a rather complex graphem-to-phoneme relation which makes grapheme based speech recognition harder. Thai uses an abugida as a script. It therefore is more difficult to build grapheme based recognition systems for it, since it contains more graphemes than languages with an alphabetic script. Also, Thai has some complex rules that map its writing to the pronunciation. Certain consonants are either pronunced differently or are reversed in their order, depending on the graphemic context. Spanish on the other hand is very well suited for grapheme based ASR. Here, the grapheme based system even outperforms the phoneme based system, showing that Spanish has a very simple correspondence between its writing and its pronunciation.

Though the grapheme based systems often perform worse than the phoneme based systems the drop is small enough, so that the performance of the resulting recognition systems is still good enough for real-world application. The real benefit in grapheme based systems now lies within the fact that they do not require any pronunciation dictionary. So, when creating a speech recognition system for a new language, the process becomes much simpler and faster, because the time and cost intensive creation of the pronunication dicitionary is not necessary any more.

In the following section we will improve the performance of the grapheme based systems by considering the fact, that the knowledge that is encoded in the pronunciation dictionary of a phoneme based system, now needs to be automatically learned by the acoustic model of the recognition system.

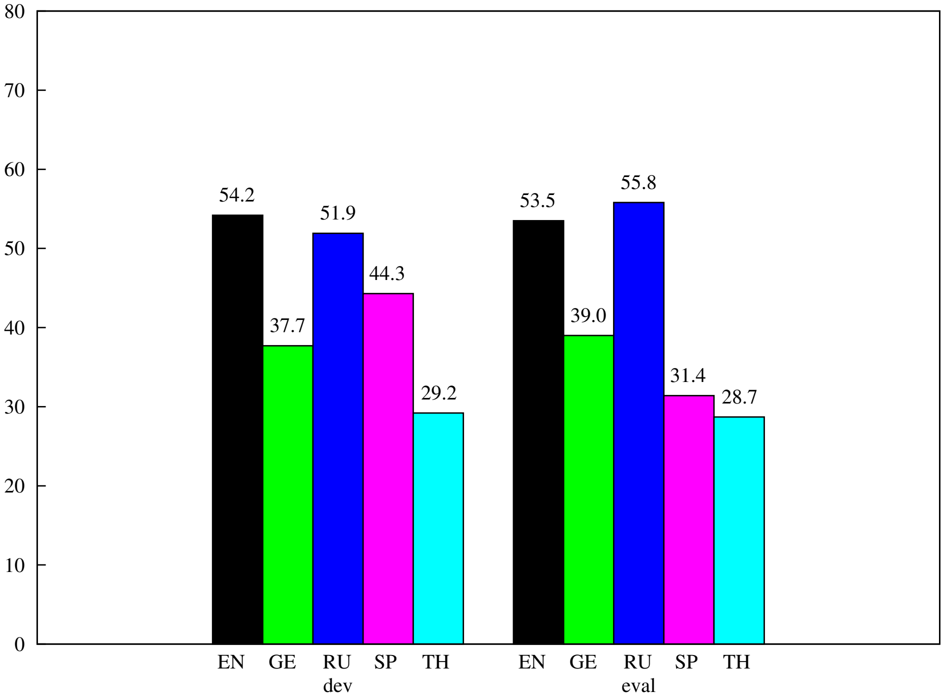


Figure 4.2: WER of the CI monolingual grapheme based ASR systems on their training language

4.5 Flexible Decision Trees for Grapheme Based ASR

As we have seen in the previous section the cluster tree used for tying the polygraphemes for context-dependent modeling plays an important role in the performance of grapheme-based ASR systems, since it implicitly learns the relation between the graphemes and their pronunciation, i.e. the acoustic manifestation as measured by the microphone, which for phoneme based systems is contained in the pronunciation dictionary.

Traditionally, as described in Chapter 3, the decision trees in phoneme based speech recognition systems consists of several sub-trees—one tree for every possible center-phone for all polyphones. In that way it is not possible to share parameters between polyphones with different center-phones. Since phonemes show a close relationship to their pronunciation, this is sensible, since it can be expected that the pronunciations of polyphones with different centerphones are

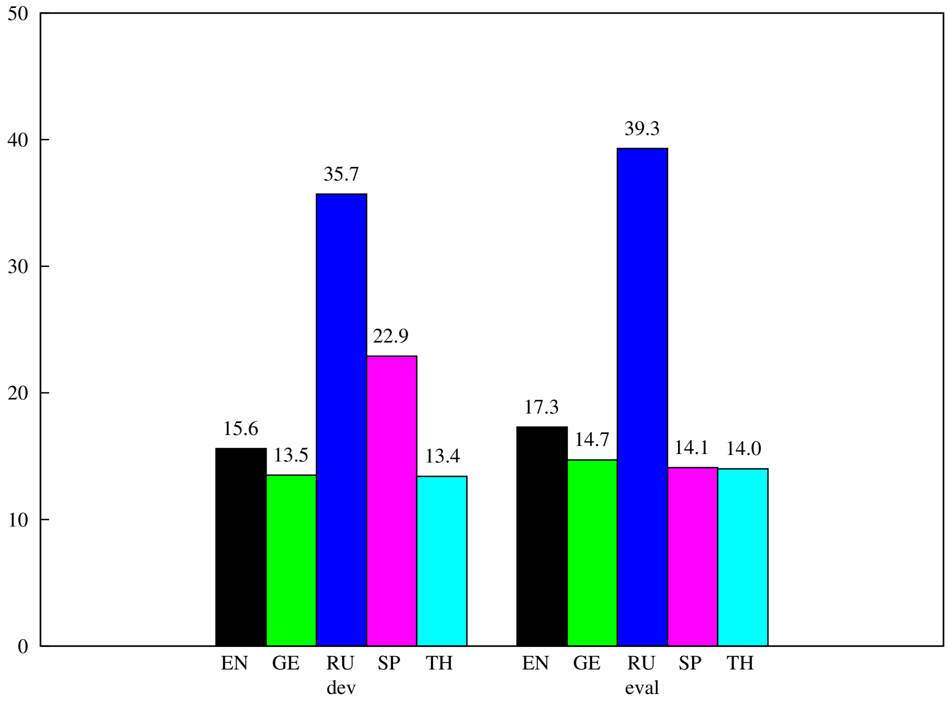


Figure 4.3: WER of the CD monolingual graphemebased ASR systems on their training language

so different from each other that no sharing of parameters is desirable. This kind of manual intervention of keeping the models with differing center-phones separate is necessary, because the entropy gain criterion used in growing the polyphone decision is not optimal with respect to word error rate.

A similar role plays the use of phonetically motivated questions when growing the decision tree. Its main purpose is to limit the search space for finding the best partition of a node according to the entropy-gain criterion in the decision tree in order to reduce the run-time of the clustering algorithm. At the same time it also limits the set of possible partitions to a sensible set, and thus again it avoids finding unsuitable trees that are optimal in term of entropy-gain but not word error rate.

For grapheme based acoustic models the case is different. Here two effects, especially in read and planned speech, are much more prominent than it is the case when using phonemes instead:

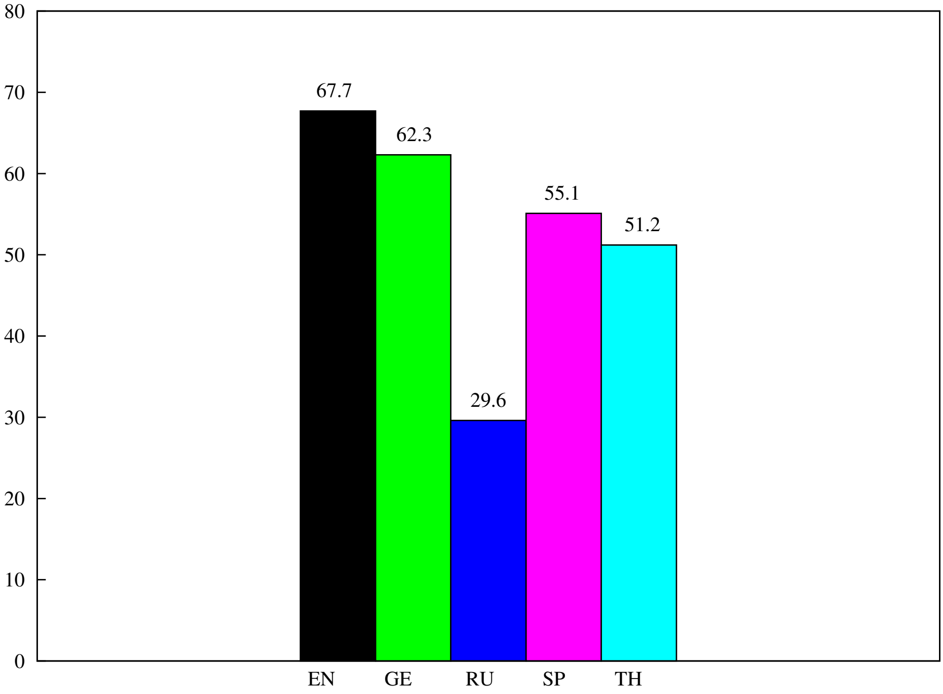


Figure 4.4: Relative Reduction in WER when going from context-independent to context-dependent grapheme based models

a) The same grapheme might be pronounced in different ways depending on its graphemic context.

b) Different graphemes might be pronounced the same way depending on the graphemic context.

The traditional clustering procedure is able to deal with the first effect, but not with the second effect. A modified tree clustering is needed in order to be able to capture the ramifications of the second effect.

4.5.1 Flexible Cluster Trees

Similar effects as described above are encountered when recognizing casual speech. In sloppy speech people do not differentiate phonemes as much as they do in read speech. Different phonemes might be pronounced very similar.

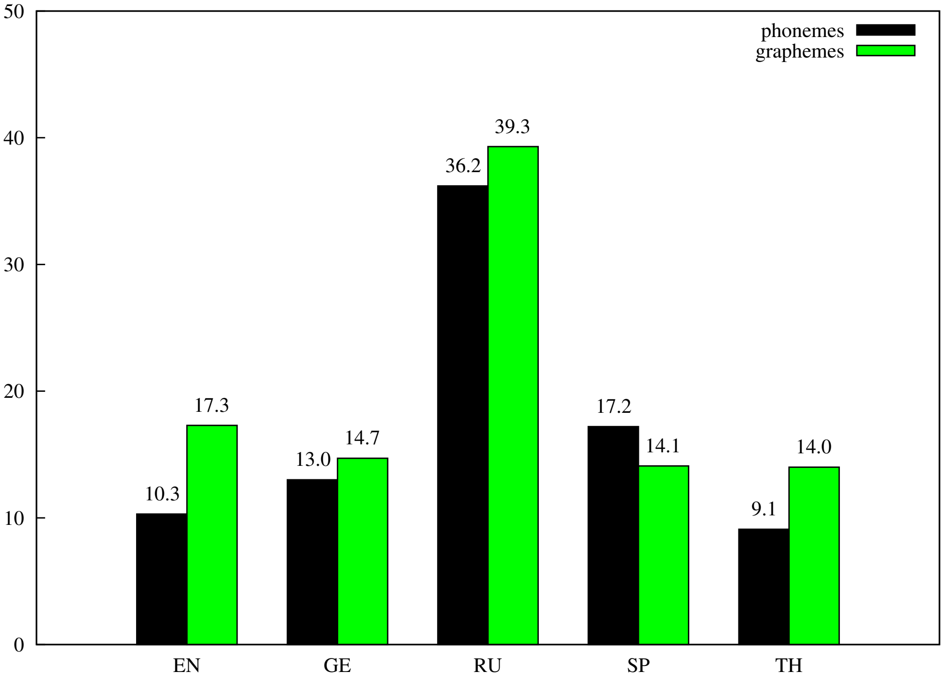


Figure 4.5: Comparison of the WER of the grapheme based ASR systems and the phoneme based systems on the evaluation set of the respective language

[YS03] presented a new tree clustering approach that lifted the limitations imposed by the growing of separate decision trees for different phonemes with the aim to address these effects in spontaneous speech. In contrast to the traditional decision tree based state tying, the enhanced tree clustering allows flexible parameter sharing across phonemes. With the enhanced tree clustering one single decision tree is constructed for all the sub-states of all phonemes. The clustering procedure starts with all polyphones at the root. The decision tree can ask questions regarding the identity and phonetic properties of the center phoneme.

This procedure is also suited to address the problem of different graphemes being pronounced similarly depending on the context, since it allows the sharing of parameters across poly-graphemes with different center-grapheme [Mim04, MSS04, SS04].

When performing this flexible tree clustering approach some design choices have to be made, and some parameter settings have to be modified from the traditional approach in addition to the different clustering set-up. For the traditional clustering, a semi-continuous model for all polygraphemes is being trained whereas all the polygraphemes with the same center-grapheme share the same codebook. Or in other words the semi-continuous grapheme models contain one codebook for every sub-tree to be grown during the clustering process.

Since in the extreme case, the flexible cluster tree contains only one tree that is being grown, it only contains one codebook that is trained on the training data from all the polygraphemes. It thus needs to capture a much wider acoustic variety, but is trained with much more training data. Therefore, the number of Gaussian components in the Gaussian mixture model of the codebook needs to be raised in order to train a good model for calculating the entropy gain criterion.

Also, our experiments in [Mim04] and [MSS04] have shown that it is sub-optimal to only grow one big tree for all sub-grapheme models. Instead it is better to keep the models for the begin, middle, and end states of a grapheme separate. Also, sometimes it can be beneficial to introduce another separation criterion, in order to keep specific models apart for which one expects no benefit from the sharing of parameters. For phonemes a separation into vowels and consonants is sensible, since vowels and consonants are distinguish themselves notably in their articulation. For graphemes this separation also seems reasonable. Especially in the case of alphabets, where graphemes either represent vowels or consonants, this separation is easily done. In the case of abugidas or syllabaries this kind of separation is not that easily achieved, and impossible for abjads, since only consonants are written in them.

Figure 4.6 shows the concept of a flexible tree for the sub-tree of the middle states of the grapheme HMM models. In this case the sub-tree again consists of two sub-trees for consonants and vowels. In comparison, the traditional way of clustering separate trees for phonemes is depicted in Figure 4.7.

4.5.2 Experiments

In our experiments with flexible cluster trees on the monolingual grapheme based ASR systems we always grew separate trees for the begin, middle, and end states of the grapheme HMMs. However, we examined two separate set-ups—one with one common tree for all polygraphemes, the other one with separate trees for vowels and consonants as depicted in Figure 4.6

In order to find the optimal number of Gaussian components for the semi-continuous model used for calculating the entropy gain, we empirically tested a range of number of Gaussians on the development sets and then applied the resulting number of parameters to the evaluation sets.

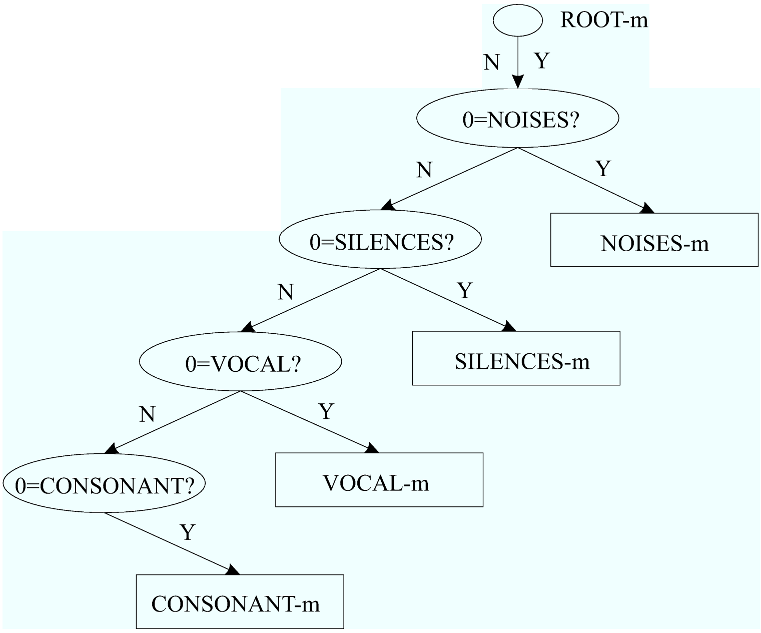


Figure 4.6: Flexible cluster tree for the middle states of the grapheme HMM models with two separate sub-trees for vowels and consonants

4.5.2.1 Optimizing the Codebook Size

Training the ASR systems required the training of a semi-continuous system with one model per polygrapheme and either one codebook per begin, middle, and end states of the grapheme HMMs—a total of three codebooks—, or with one codebook per begin, middle, and end state and separated according to whether the center-grapheme of the polygrapheme is a vowel or consonant—thus a total of six codebooks. Since neither silence nor noises are modeled context-dependently in our system, data assigned to them were not of relevance for the training and clustering.

After training the semi-continuous models for every polygrapheme, the polygraphemes belonging to every codebook were clustered to a maximum of 3,000 models using the divisive clustering approach described in Chapter 3.

The context-dependent models obtained this way were trained with two iterations of the training procedure described in Section 4.4.3 starting from the forced alignments obtained from the context-independent models. Therefore,

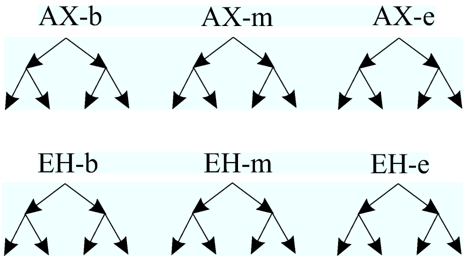


Figure 4.7: Traditional cluster tree with separate trees for the begin, middle, and end states of the grapheme HMMs and separate trees for polygraphemes with different center-graphemes

the models obtained that way can be directly compared to the models of the monolingual recognizers with the traditional cluster tree from Section 4.4.5.1.

As discussed above, the number of Gaussian components in the Gaussian mixture models of the codebooks needs to be enlarged in order to compensate for the increase in training material and the fact that the codebooks are now assigned to a much wider variety of polygraphemes. The resulting context-dependent models then again used 32 Gaussian components per mixture.

Table 4.7 shows the word error rates of for different codebook sizes on the English development set for both cases,growing only one tree for all polygraphemes (Single Tree) in which all polygraphemes share one codebook and keeping the models for polygraphemes separate according to whether the center-grapheme is a vowel or consonant (V&C Tree).

When compared against the WER of 15.6% that the models with the traditional tree from Section 4.4.5.1 achieve, the single tree models were not able to outperform the baseline. However, the models with separate trees for vowels and consonants show improvements over the baseline. As Table 4.7 shows, the best performance on the English development set was reached with 1,536 Gaussian components for the semi-continuous models used for clustering the polygraphemes and yielded a word error rate of 14.5%. This is a relative reduction in word error rate of 7.1% over the traditional tree.

4.5.3 Flexible Cluster Trees for All Languages

Using the experiences collected with the flexible cluster tree on the English development, we also trained flexible cluster trees for the other four languages.

Table 4.7: WERs for different codebook sizes for the semi-continuous models used for clustering for English

#Gaussians	Single Tree	V&C Tree
32	16.3	15.7
256	15.8	15.0
512	16.9	15.0
768	18.4	15.5
1024	16.1	15.2
1280	16.4	15.2
1536	16.4	**14.5**
1792	16.1	15.6
2048	17.0	15.4
2304	16.0	15.2

This time we only grew separate trees for vowels and consonants as center-graphemes of the polygraphemes. Table 4.8 shows the word error rates of the resulting trees for different codebook sizes used during clustering for all languages on the development set of their respective language. As we can see for English, German, and Russian using such a flexible cluster tree lead to improvements over the regular tree, while for Spanish and Thai no improvements could be achieved.

Figure 4.8 shows the performance of the traditional tree versus the V&C tree on all five languages on the evaluation set using the tree with the best performance on the respective development sets. Table 4.9 shows the relative reduction in WER on the evaluation set of the respective language when using the flexible tree that was best on its development set. For four of the five languages we can see large gains in performance when applying the flexible cluster tree. We can see that the biggest relative gain could be achieved for German with a reduction in WER of 6.1% relative. Here the flexible cluster trees seems to capture the implicit variations in pronouncing the German alphabet the best. For Spanish we can see a slight reduction in word error rate on the development set, but on the evaluation set the performance stays the same. So, the flexible cluster tree does not give any benefits in this case, but also does not hurt the performance.

The results confirm our assumption that the cluster tree for the context-dependent models is of high importance in grapheme based ASR. By using a flexible tree it is possible to learn the knowledge normally encoded in the pronunciation dictionary automatically on the training data.

Table 4.8: WERs for different codebook sizes for the semi-continuous models used for clustering V&C Trees for all languages

#Gaussians	EN	GE	RU	SP	TH
regular tree	15.6	13.5	35.7	22.9	12.7
32	15.7	13.1	35.2	23.2	12.4
256	15.0	13.2	35.3	23.3	12.8
512	15.0	13.1	35.3	23.1	**12.3**
768	15.5	13.1	35.2	22.8	12.5
1024	15.2	13.1	35.2	23.1	12.7
1280	15.2	**12.9**	35.5	23.1	12.4
1536	**14.5**	13.2	35.4	23.0	12.4
1792	15.6	13.2	35.0	23.2	12.4
2048	15.4	13.0	35.1	23.0	12.8
2304	15.2	13.2	**34.7**	**22.7**	12.6

Table 4.9: Relative Reduction WER when using a flexible V&C tree instead of the regular one on the development sets of the respective language

#Gaussians	EN	GE	RU	SP	TH
rel. reduction in WER	2.3	6.1	2.3	0.0	2.9

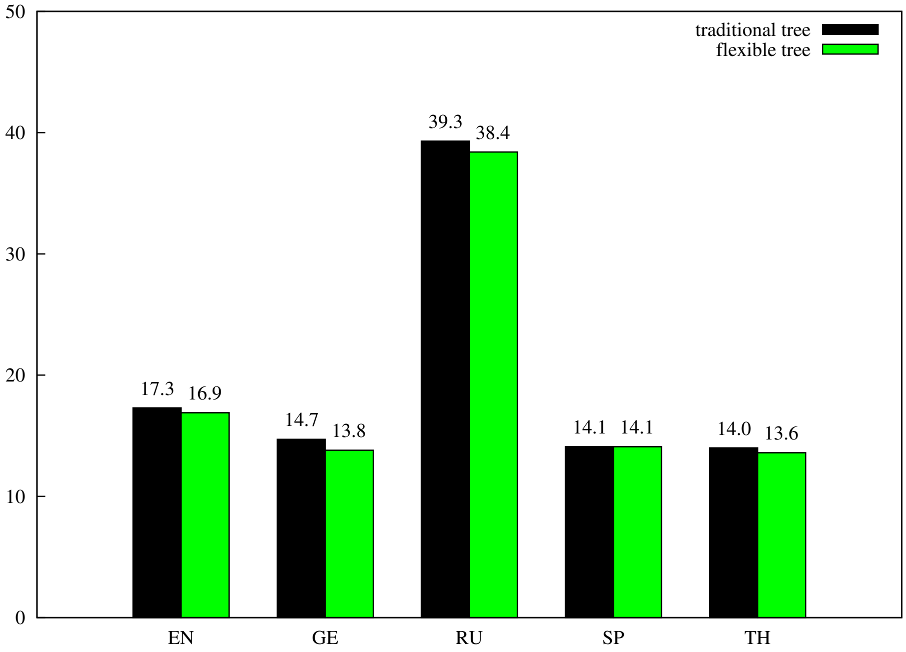


Figure 4.8: Comparison of word error rates between traditional and flexible V&C tree on the evaluation set of the different languages

Chapter 5

Multilingual and Crosslingual Graphemic Modelling

The experiments in the previous chapter were aimed at simplifying the creation of an ASR system in a new language by eliminating the need for a phoneme based pronunciation dictionary which is costly to create. By substituting graphemes for phonemes as modeling units the creation of the pronunciation dictionary became a trivial task. As discussed and shown in the previous chapter this approach is feasible for a wide variety and the majority of languages in the world.

However, the training of the grapheme based ASR systems still requires large amounts of transcribed audio data for training their acoustic models, just as it is needed for phoneme based ASR systems. The task of transcribing a novel language demands the use of native-speakers or at least very fluent-speakers of the language involved. Furthermore, the data that needs to be collected, has to be spoken by a wide variety of native speakers as well. For languages with a large group of speakers, audio resources might be readily available, e.g. in the form of radio and TV broadcasts, publications on the Internet, or in archives that include audio recordings. But for less resourced languages with only few speakers or little economic strengths this is not the case. Here, the audio recordings that can be transcribed and then be used for training, first need to be recorded. Since in practice it is impossible to find a sufficient number of speakers of the language that are willing to donate their speech for system development at the site at which the ASR system is being trained, this means that collection has to be done in a field expedition style of action in the countries and regions in which the language is spoken. Recent research has worked on using technology in order to help bridging the gap between language technology experts and native speakers of under-ressourced languages. The project *Speech Processing — Interactive Creation and Evaluation* (SPICE) provides interactive tools that enable native speakers, untrained in language technology, to develop speech processing models. The users can collect and upload audio files, and in an interactive fashion compose and rapidly boot-strap the components for speech recognition and speech synthesis systems[SBB$^+$07].

Thus, collecting and annotating acoustic training data for acoustic model training in the traditional way for all languages in the world, especially for the under-resourced and less prevalent ones, is an impossible task for which the necessary resources cannot be allocated.

One way to circumvent this problem is the attempt to train language independent acoustic models on the available training data from multiple languages and to either directly apply them to a new language or to port them to a new language using methods that require only little data in the target language [WKAM94, Köh98, SW98c].

5.1 Related Work

Work in porting acoustic models to new languages is closely related with the field of *multilingual speech recognition.* In 5.2 we will describe in more detail the view of multilingual speech recognition that we adopt in this work by following [SW01].

In multilingual speech recognition, which is also called *language independent* ASR, acoustic models are trained by sharing the training data from many languages to train one unified set of models. Different approaches to finding such a unified model set exist. [Sch00b] classifies the approaches into:

Heuristic combinations: Either phonetic or articulatory classification schemes [DA92, CDG$^+$97, WRN$^+$98, WBNS97] or reference schemes such as IPA [Köh98, Köh99, SW98a] and Sampa [AAB$^+$96, AAB$^+$97, USN98] are used for finding common models.

Data-drive model combination: Different criteria are used to identify phonemes in different languages that can share the same model: confusion matrices [ADB93, PD94, BI99], combinations of different distance measures [BGM97, MPF99], likelihood distances [AD97, Köh99], and a-posteriori distances [CAGADL97]

Hierarchical Combination of heuristics and data-driven methods: First, the phonemes are grouped into classes by a heuristic, then a data-driven clustering is performed[Köh99, Köh96, WBNS97, CDG^{+}97, WRN^{+}98, SW98c, SW98b].

[Sch00b] gives a comprehensive overview over the different combination methods and ways to find a common model set.

Once a common phoneme set has been found it is usually trained on a number of languages. In Section 5.2 we describe the methods ML-Mix and ML-Tag [SW01, SW98a] which we have used for our experiments. The resulting language-independent acoustic models are then able to recognize speech from all the training languages. ML-Mix can also be applied to new languages not seen during training. When it is applied to a new language the models have to be mapped to the phonemes in the target language. This task can be achieved the same way that the common model set for the training languages was found. If phonemes in the target language exist that are not covered by the multilingual model, they have to be mapped to the closest covered phoneme.

In this chapter we examine the case where either no or only very little transcribed acoustic training material is available in the target language. Other works deal with the case that plenty, but untranscribed audio material in either the target language or target domain is available. If already an initial model exists, which is then refined on untranscribed training material, the terms *unsupervised training* [ZSCB98, KW99, Ram05, GN08] and *unsupervised adaptation* are used [WPG96]. Training and adaptation is performed by first using the existing model to recognize the untranscribed training or adaptation data, and then to train on that data by taking the automatic transcription as reference transcription. Different methods were introduced by the before-mentioned publications in order to carefully select suitable portions of the training material. This selection is necessary, since the automatic transcriptions contain errors and thus might taint the models when trained on them. [Ram05] found that, when already a good, initial model exists, i.e. trained on 200h of transcribed speech, a multifold of untranscribed data is necessary in order to see significant gains in performance.

Also, research has begun to address the case in which only untranscribed audio data in a new language is available [PG08]. This work is however still at its beginning and only exploratory experiments have been conducted so far.

Also, some research has addressed the scenario in which acoustic models that are based on phonemes are ported to grapheme based models. [ŽKD+05, ŽK06] used a Slovenian phoneme model and a linear combination of phonemes in that model to initialize grapheme models for Slovenian. After initialization regular training was performed on large amounts of available data.

5.2 Multilingual and Language Independent Acoustic Modeling

When using the term *multilingual Automatic Speech Recognition* (ML-ASR) we follow [SW01] which defines multilingual recognition systems as systems that are capable of simultaneously recognizing languages which have been presented during training. These multilingual models are often not just a simple combination of the language dependent models of several monolingual recognizers but try to capture synergetic effects by sharing models for several languages so that the combination of models is smaller in size than the sum of the individual models and so that the single models receive more training data than in their original languages.

The acoustic model obtained from this kind of multilingual models can be used for porting ASR systems to new languages. Here the hope is that, if sufficient training material from a wide variety of languages is seen during training, the acoustic manifestation of a new language is already covered by the multilingual model so that no new acoustic model needs to be trained. In the optimal case the multilingual acoustic model would turn into a language independent model that can recognize any language in the world. This goal has of course not been reached yet, and it can be doubted that it will ever be achieved with the same level of performance as specialized and highly optimized monolingual models.

However, even if the multilingual language model does not cover the new language completely, and the acoustics learned by the multilingual model distinguishes itself from the acoustics of a new language, one can still use it as a starting point for the fast training of an acoustic model for that new language, especially if only a very limited amount of training data in the target language is available [Sch00b].

For our experiments in multilingual modeling we used two techniques from [SW01]: ML-Mix and ML-Tag. For our experiments in cross-lingual acoustic modeling we only used the technique ML-Mix.

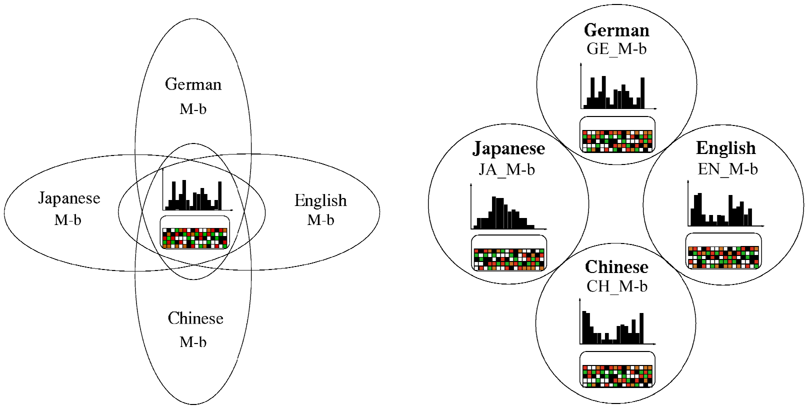


Figure 5.1: Language mixed acoustic modeling vs. language dependent, from [SW01]

5.2.1 ML-Mix

In the method ML-Mix, models—in our case either sub-phonemes or sub-graphemes—that are common to one language share the same model and are treated as identical in the rest of the system, e.g. in the cluster tree for the context-dependent models. All information about which language a model originally belonged to, is discarded in the system. Instead, models common to several languages share all training data from that languages.

Figure 5.1 illustrates this concept for the languages Chinese, English, German, and Japanese and the model of the middle state of the phoneme /M/. For the monolingual case depicted on the right side of the figure, every language has its own model for that HMM state of /M/. Only training material from the language that that model belongs to is used to train it. The left side shows the multilingual case. Here the training material from all languages is used to train one single model that can be used for all the languages that the training material comes from.

5.2.2 ML-Tag

While ML-Mix discards all information on which language the training data belongs to that is used to estimate the multilingual model, ML-Tag preserves some language information. While in ML-Tag for codebooks, that is the Gaussian components of a GMM model, the same sharing of the parameters and training material takes place as for ML-Mix, the mixture weights are trained language dependently. So, every model in every language gets its own model. But models in every language that refer to the same phoneme share the same set of Gaussian densities.

For context-dependent models, the polyphone clustering tree is additionally allowed to ask for the identity of the language to which the polyphone belongs that it models. Therefore, if the entropy gain decision criteria deems it appropriate, the codebooks of the polyphone models can be separated by language as well, if they are too dissimilar.

5.3 Multilingual Grapheme Based Speech Recognition Using ML-Mix and ML-Tag

Multilingual models are based on the fact, that phonemes in different languages are pronounced the same, or at least very similar. For graphemes this assumption clearly does not hold. Different graphemes in different languages can be pronounced very differently. However, especially in the case of consonants, they can also be very similar. However, we can expect that a multilingual model based on graphemes will not perform as well as a multilingual model based on phonemes. But in the case, where we cannot use phonemes as modelling units, due to a lack of pronunications dictionaries, the question remains whether multilingual grapheme models can still be utilized to recognize several languages with only one acoustic model, and whether it is possible to use them as a base for creating an acoustic model in a new language.

The use of Gaussian mixture models as emission probabilities of the HMM states should make it possible for one model to learn the different acoustic manifestations of the graphemes in the different languages. Further, the use of context-dependent models should also make it possible for the context-dependent, multilingual models to learn the different pronunciations of the graphemes in the different languages.

We now trained two multilingual models, one using ML-Mix and one using ML-Tag. The ML-Mix model will be later used to apply it to a new language, making use of the fact that the identity of the training languages is of no relevance to the models. The ML-Tag model is not suited for application to a new language, since it uses language dependent models which cannot be applied to a new language.

In our experimental set up we assume that the languages English, Russian, and Spanish are well known languages for which large amounts of training material are available. German and Thai receive the role of languages for which we want to create new acoustic models and for which only a small adaptation set of transcribed data is available.

For creating the ML-Mix model, we first trained a context-independent ML-Mix recognizer (*ML3-Mix-CI*) on the languages that we assume as given. Then a polygrapheme decision tree with three thousand models was clustered and trained on these languages (*ML3-Mix-CD*). The same was also done for th ML-Tag model, first training an context-independent model (*ML3-Tag-CI*) and then clustering a context-dependent model (*ML3-Tag-CD*).

Just as it was done in [SW01] for phonemes, we allow the sharing of training data and the creation of a global model set, in our case based on the identity of the grapheme, not the identity of the phoneme, in different languages.

Since Russian uses Cyrillic script instead of a Latin based one, as the other two languages involved in the ML-Mix model do, the Cyrillic graphemes were mapped to a romanized representation in order to allow data sharing with the other languages. Table 5.1 shows this mapping. Since the ASR systems for the other three languages use lowercase representations for their graphemes, all Russian graphemes that are composed of more than one letter or contain an uppercase letter are only common to Russian but not the other languages, while the other Russian graphemes are shared with the other languages. Since Russian belongs to the pool of languages that we assume to be well known and well studied this extra knowledge of a suitable romanization is permissible.

Figure 5.2 gives the word error rates of the resulting context-independent models on the dev and eval sets of the individual languages that were used for training, Figure 5.3 the results for the context-dependent models. One can see from the results that for the languages English and Russian there is a clearly visible performance degradation compared to the monolingual recognizers. The degradation for English is larger than for Russian which is to be expected, since English has a more complex grapheme-to-phoneme relation than Russian. Also, Russian contains many graphemes that are not common to the other two languages, so that their models are not broadened by the training material comming from the other languages. One needs to consider that the multilingual model uses a total of three thousand models for recognizing all languages while the monolingual recognizers uses three thousand models per language, that is a total of nine thousand models. So, as a contrastive experiment, we also clustered an ML3-Mix model with nine thousand models. The resulting WERs of these models are depicted in Figure 5.4. As we can see, the performance improves for all languages, but still lacks behind the monolingual models. As to be expected

Table 5.1: The Cyrillic graphemes and their romanized form

Graphemes	Romanized	Graphemes	Romanized
а	a	р	r
б	b	с	s
в	w	т	t
г	g	у	u
д	d	ф	f
е	ye	х	h
ё	yo	ц	tS
ж	jscH	ч	scH
з	z	ш	sch
и	i	щ	schTsch
й	j	ъ	Q
к	k	ы	i2
л	l	ь	~
м	m	э	e
н	n	ю	yu
о	o	я	ya
п	p		

from the fact that language information is preserved in the model, the performance of the ML3-Tag models on their training languages is better than that of the ML3-Mix models. Figure 5.5 shows the word error rates of the context-independent ML3-Tag model on its training languages, while Figure 5.6 shows the performance of the context-dependent models with three thousand models. Just as for the ML3-Mix model we also trained a context-dependent ML3-Tag model with nine thousand models. The results of this system are plotted in Figure 5.7.

As we can see the performance of the ML3-Tag models is in general higher than that of the ML3-Mix model. Also, the context-dependent model with nine thousand models outperforms that with three thousand models. Table 5.2 summarizes the results and compares the different multilingual models against the performance of the monolingual recognizers on the three languages.

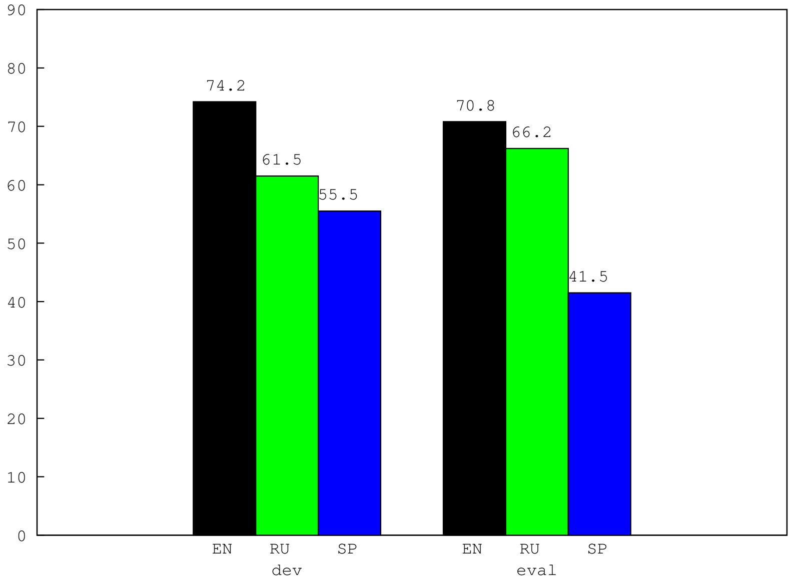


Figure 5.2: WER of the CI ML3-Mix graphemebased ASR systems on their training language

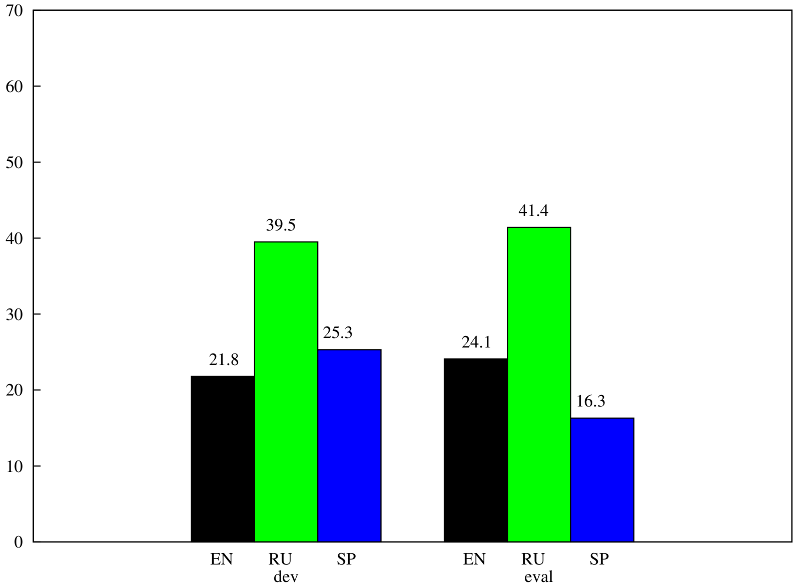


Figure 5.3: WER of the CD ML3-Mix graphemebased ASR systems on their training language with a total of 3,000 models.

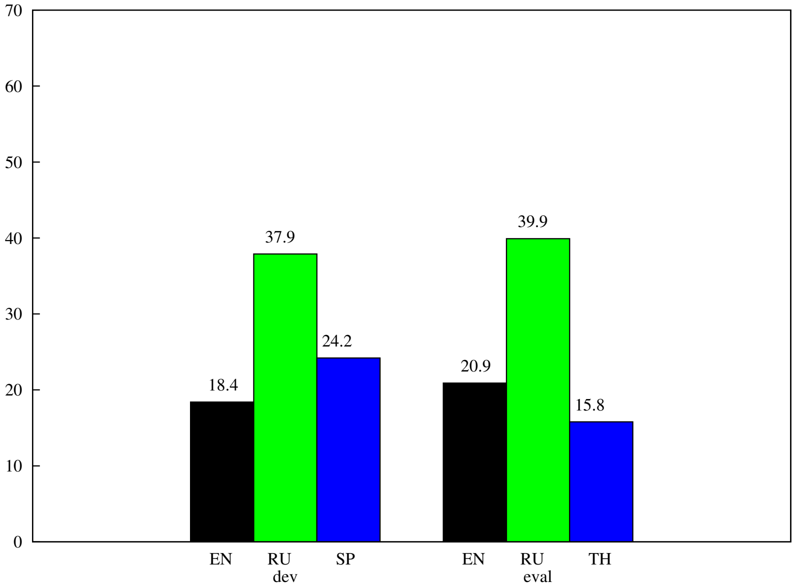


Figure 5.4: WER of the CD ML3-Mix graphemebased ASR systems on their training language with a total of 9,000 models.

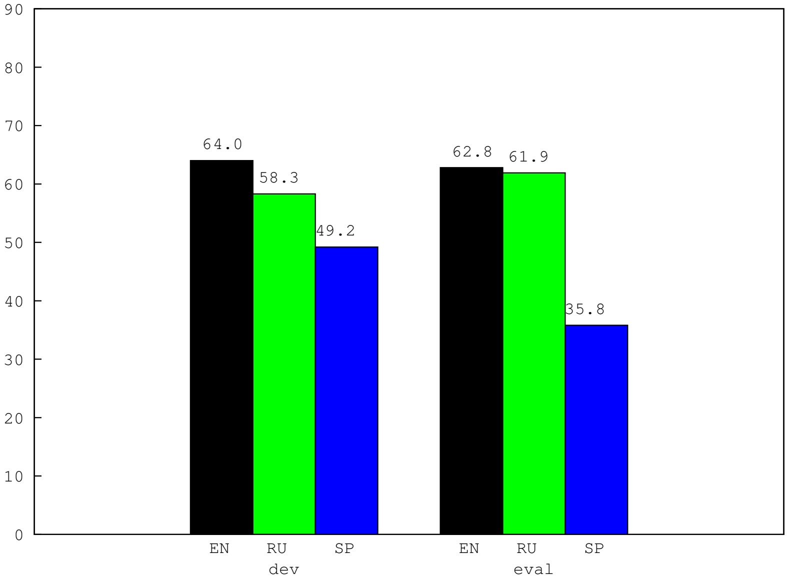


Figure 5.5: WER of the CI ML3-Tag graphemebased ASR systems on their training language with 3,000 models

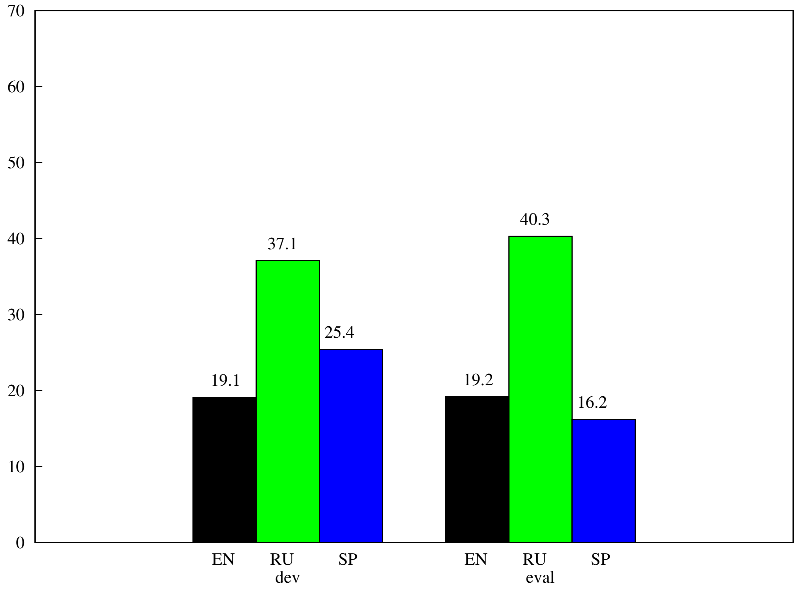


Figure 5.6: WER of the CD ML3-Tag graphemebased ASR systems on their training language with 3,000 models

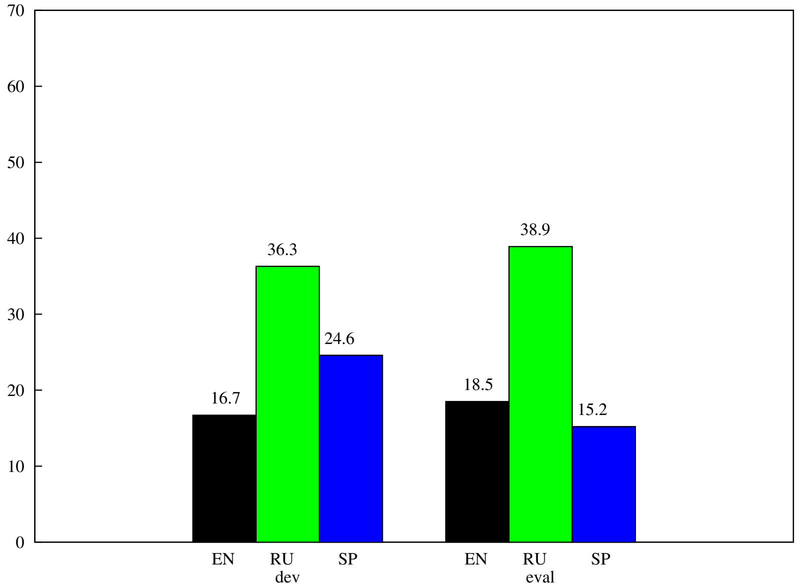


Figure 5.7: WER of the CD ML3-Tag graphemebased ASR systems on their training language with 9,000 models

Table 5.2: WERs of the different multilingual models in comparison to the monolingual models

System	EN		RU		SP	
	dev	eval	dev	eval	dev	eval
CI						
monolingual	54.2%	53.5%	51.9%	55.8%	44.3%	31.4%
ML-Mix	74.2%	70.8%	61.5%	66.2%	55.5%	41.5%
ML-Tag	64.0%	62.8%	58.3%	61.9%	49.2%	35.8%
CD						
monolingual	15.6%	17.3%	35.7%	39.3%	22.9%	14.1%
ML-Mix 3000	21.8%	24.1%	39.5%	41.4%	25.3%	16.3%
ML-Mix 9000	18.4%	20.9%	37.9%	39.9%	24.2%	15.8%
ML-Tag 3000	19.1%	19.2%	37.1%	40.3%	25.4%	16.2%
ML-Tag 9000	16.7%	18.5%	36.3%	38.9%	24.6%	15.2%

5.4 Porting from a Grapheme Based ML-Mix Model

When applying multilingual grapheme based ASR systems across languages one of the problems is that, unlike with phonemes, the overlap between the graphemes of the multilingual system and the target language can vary greatly—from a large overlap to no overlap at all. In the former case the multilingual models can be applied to the target language based on the grapheme identity. In the latter case this is not possible. For this case we experimented with data driven mapping methods that only require a minimal amount of training data in the target language.

Starting from the multilingual model trained on the languages English, Russian, and Spanish from the previous section we examined two basic porting scenarios. For the first scenario we investigated porting the multilingual model to German whose graphemes have a very large overlap with the multilingual model. For this scenario we worked with grapheme identity based mapping of the models as well as with data driven methods. In the second scenario we investigated porting our multilingual model to Thai whose graphemes do not have any overlap at all. Here, only a data driven mapping can be used to map the multilingual models to Thai.

5.4.1 Influence of the Multilingual LDA Transformation

In [SW00] it was shown for phoneme based models that an LDA matrix that has been trained on many languages performs either equally well or only slightly worse than an LDA matrix exclusively trained on data from the test language. In order to verify this result for grapheme based models the monolingual ASR systems for the languages in the training set of the ML3-Mix models—English, Russian, Spanish—as well as the systems for the languages to which we want to port to—German and Thai—were retrained, this time using the LDA matrix from the ML3-Mix models. The re-training was always performed on the full training set.

Figure 5.8 compares the performance of the CI monolingual models with the monolingual LDA of the respective language vs. the performance of the models trained with the multilingual LDA. Figure 5.9 does the same for the context-dependent models. In general, the same behavior for the grapheme based systems as for the phoneme based systems in [SW00]—that is no or only a slight degradation—can be observed. In many cases, e.g. for the Spanish, Russian, and German context-dependent models on the evaluation set, the multilingual LDA transform actually slightly outperforms the monolingual LDA transform.

We can also see that when using the LDA matrix trained on English, Russian, and Spanish for the German ASR system, the recognition performance improves slightly, while for Thai it stays essentially the same. This means that the multilingual LDA matrix is suited for porting grapheme based ASR systems to new languages. We can expect that the transformation learned on languages other than the one tested on does not introduce any degradation in performance over a purlely monolingual LDA transform trained on large amounts of data in the test language.

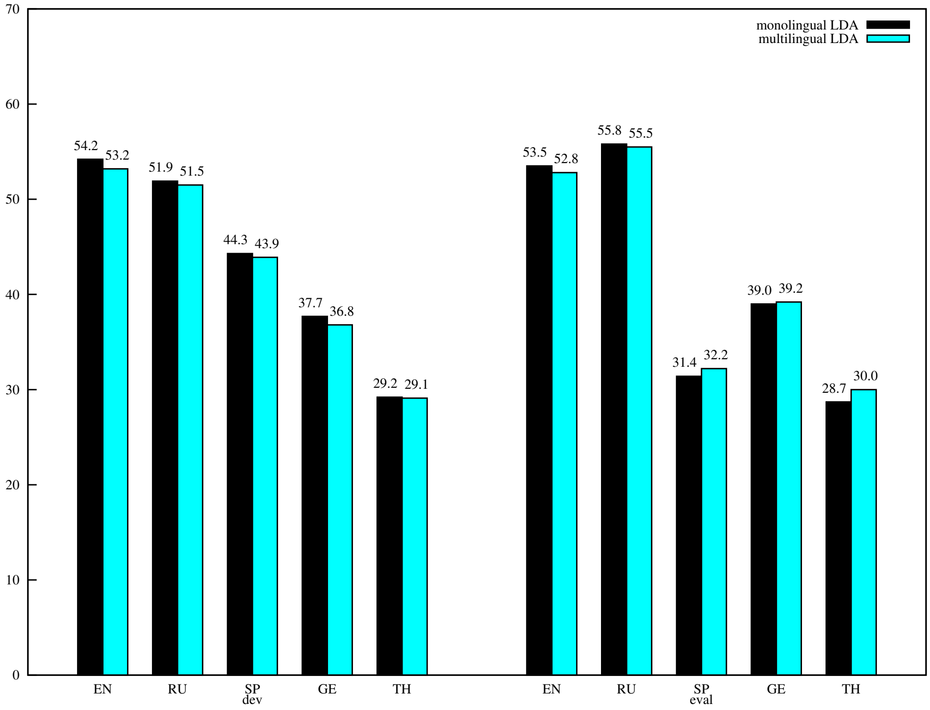


Figure 5.8: WER of the CI monolingual graphemebased ASR systems with monolingual LDA vs. multilingual LDA

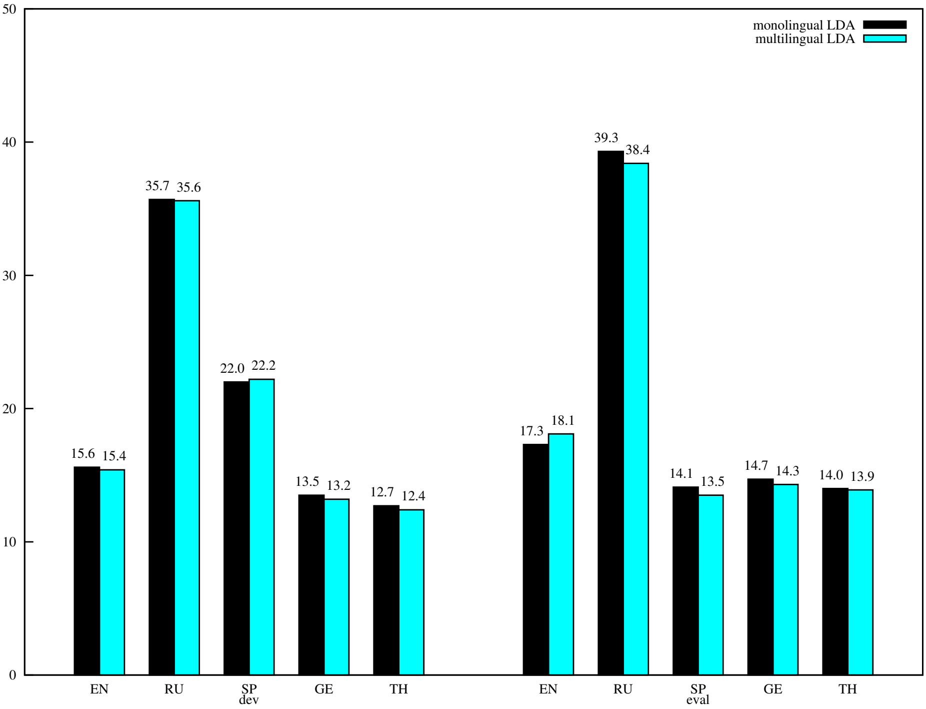


Figure 5.9: WER of the CD monolingual graphemebased ASR systems with monolingual LDA vs. multilingual LDA

5.4.2 Grapheme Identity Driven Cross Language Transfer of ML-Mix to German

The graphemes of the combined alphabets of the three languages that were used to train the ML3-Mix model in the previous subsection—the Russian alphabet in its romanized form—cover almost all graphemes in the German alphabet. Thus it is easy to apply the ML3-Mix model directly to German, by mapping the graphemes from the ML3-Mix models to their identic, German counterpart. Only the four German graphemes ä, ö, ü, and ß are not covered by the graphemes in the ML3-Mix model. For them, a manual mapping to the graphemes of the ML3-Mix model has to be found. We decided to map these four graphemes to their canonic grapheme sequence when using only the English alphabet—'ae', 'oe', 'ue', and 'ss'.

Using this mapping we recognized the German test data with the ML3-Mix model. Since we can expect the pronunciation of different graphemes to be quite different among the various languages that use them, one can expect that the ML3-Mix model only fit poorly to the German data and thus give a very low performance for this approach.

Figure 5.10 compares the performance of the ML3-Mix model on the German test data with the performance of the monolingual German recognizer that has been trained on the German training data. Indeed the performance suffers drastically when using the multilingual model that has not seen any German training data. When porting the context-independent ML3-Mix models to German with the identity driven approach, the WER increases by 136% relative on the dev set and 131% on the eval set. For the context-dependent models the WER rises even more by 485% relative on the dev set and 437% relative on the eval set. The, in comparison to the context-independent models, almost four times higher loss in performance for the context-dependent models, suggests that, besides the mismatch in pronunciation of the same graphemes in different languages, one of the major sources for the WER increase is the multilingual polygrapheme decision tree that only seems to poorly fit the German test data.

Also, one should keep in mind that this approach can only be applied for porting the ML3-Mix models to German but not to Thai, since the ML3-Mix languages and Thai do not have a single grapheme in common. Here, an extensive, manual mapping would be necessary which we assume as not feasible for our scenario, since it requires extensive phonetic and linguistic knowledge in the languages in the ML3-Mix model and for Thai. We therefore examined two data-driven approaches in Section 5.6 that can be applied without any phonetic or linguistic knowledge.

5.4.3 EM Adaptation

As we have seen in the previous subsection, directly applying the ML3-Mix models to German only gives unsatisfactory results. Some sort of adaptation of the ML3-Mix models on as little adaptation data as possible is necessary in order to improve the performance of the cross-language transfer.

For our experiments we assumed as given a small amount of available German adaptation data and refined the ML3-Mix model with it. Since this adaptation data should be as easy as possible to collect it only contains few material in terms of length, but also only from one single speaker. Normally, training data should contain as many speakers as possible in order to train speaker independent recognition systems. But collecting—even small amounts of—data from many speakers is rather costly and should be avoided for our purposes.

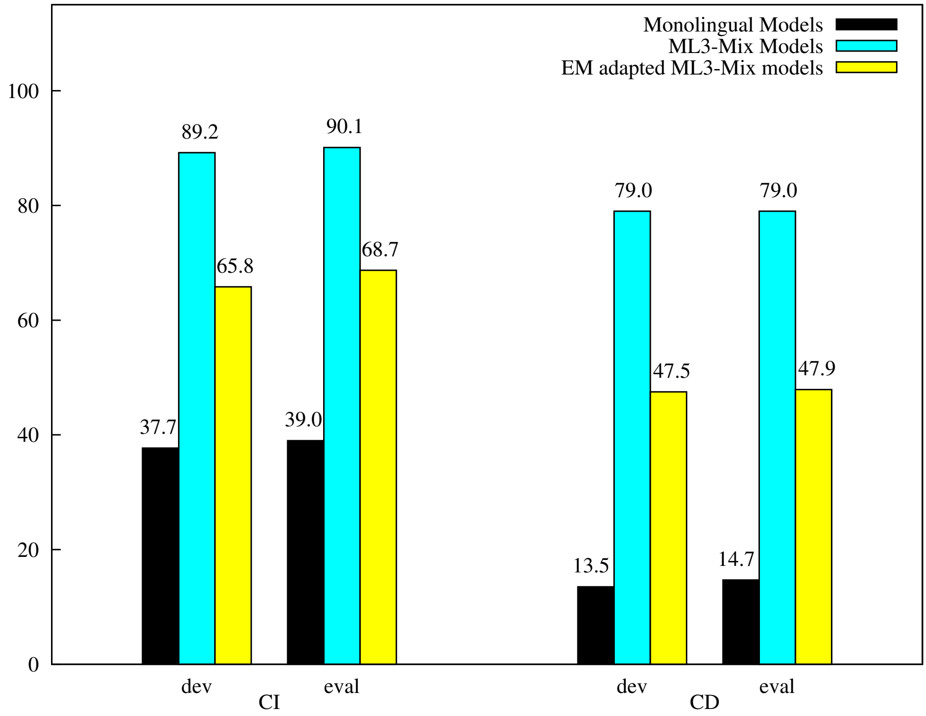


Figure 5.10: WER of the CD monolingual German graphemebased ASR systems on German in comparison with the ML3-Mix models and the EM adapted ML3-Mix models

In order to adapt the ML3-Mix models with the available fifteen minutes of German data we applied two iterations of EM training to the ML3-Mix-CI model and one iteration of EM training to the ML3-Mix-CD model. The resulting word error rates on the German development and evaluation set are shown in Figure 5.10 in comparison to the monolingual models and the unadapted ML3-Mix models. One can see that on the evaluation set the EM adaptation reduces the WERs of the models significantly by 23.8% relative for the context-independent models and 39.4% relative for the context-dependent models compared to the unadapted models.

Even though EM adaptation on the context-dependent models gives higher gains than for the context-independent models, the relative loss in performance measured against the monolingual models is still higher for the context-dependent models—226%—than for the context-independent models—76% relative. This again indicates that the ML3-Mix cluster tree is only ill fitted for the German data and needs to be adapted in addition to the emission probabilities of the

HMM state which were improved by the EM training.

5.4.4 Influence of the Polypgrapheme Decision Tree on Porting Performance

As described above the unbalanced loss in performance between context-independent and context-dependent ML3-Mix models—regardless whether EM adapted or not—hints at an inappropriateness of the ML3-Mix model cluster tree for the German data.

In order to exactly determine the influence of the polygrapheme decision tree on the porting performance, we retrained the monolingual, context-dependent ASR systems, this time using the multilingual LDA matrix and the multilingual polygrapheme decision tree from the ML3-Mix system. Should the German recognizer show a large drop in performance over the recognizer using a cluster tree trained on monolingual data, this can this time only be due to the cluster tree, and not e.g. due to the smaller amount of training material, since the cluster tree is the only component that has been exchanged compared to the systems from Section 5.4.1.

Figure 5.11 compares the word error rates of the resulting systems against the systems that use their monolingual tree. Table 5.3 shows the relative increase in word error rate for the systems with the ML3-Mix tree. For English, Russian, and Spanish there is only a moderate relative increase in WER on the development set—between 5.8% for English and 8.6% for Spanish. On the evaluation set the increase for these three language is likewise moderate. Only Spanish sticks somewhat out with its somewhat higher relative increase of 15.6%. However, for German the increase is much higher than for the other languages. With 35.6% on the development set and 26.6% on the evaluation set it is roughly double that of the other three languages.

This disproportionally high increase is solely due to the multilingual polygrapheme decision tree which was trained on English, Russian, and Spanish and which only poorly fits the German data. Thus, adapting the ML3-Mix cluster tree to German using the available German adaptation data bears a significant potential for improving the performance when applying the ML3-Mix models to the German data.

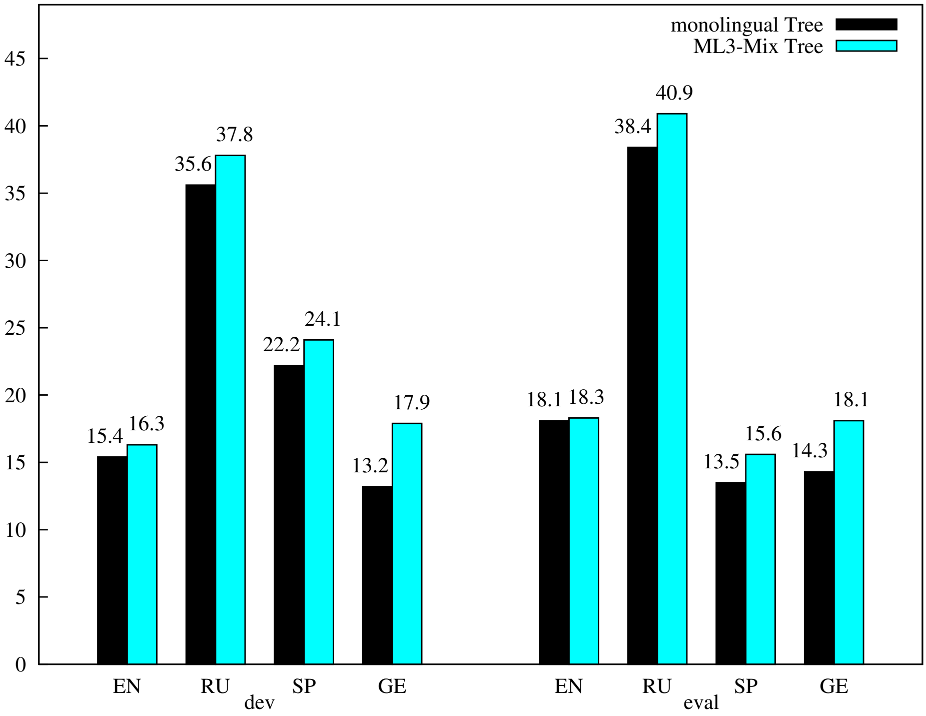


Figure 5.11: WER of the grapheme based ASR systems when using their monolingual cluster trees vs. the ML3-Mix cluster tree

5.5 Adapting the Polygrapheme Decision Tree

As seen in the previous section, the polygrapheme decision tree of the ML3-Mix recognizer only poorly fits the German data. This can be accounted to three basic problems:

- Polygraphemes frequently occurring in German might not be modeled detailed enough in the ML3-Mix tree in order to give good performance, because they are not observed often enough in the training data of the ML3-Mix models.
- Models in the ML3-Mix tree might exist that are too detailed for the German data, because they occur frequently in the ML3-Mix training data but not in the German data thus giving only a poor performance.
- The partition of the polygraphemes in the decision tree is not suited for Germans. Polygraphemes whose pronunciation is contradictory and not

Table 5.3: Relative increase in WER when using the ML3-Mix tree instead of the monolingual tree

Language	EN	RU	SP	GE
dev	5.8%	6.2%	8.6%	35.6%
eval	1.1%	6.5%	15.6%	26.6%

suited to be modelled by a common probability distribution are joined in one model.

In Subsection 5.4.3 we used the 15 minutes of available German adaptation data to adapt the GMMs of the output probabilities of the HMM states. We now used the same adaptation data to also adapt the polygrapheme decision tree. For this we first apply the technique *Polyphone Decision Tree Specialization* (PDTS) [Sch00b] with some modifications and then improved it by combining it with a decision tree pruning scheme [Wol99].

5.5.1 PDTS

In order to improve the porting performance, we adapted the multilingual polygrapheme decision tree using the 15 minutes of German data. [Sch00b] introduced *Polyhone Decision Tree Specialization* (PDTS) as an approach for adapting a multilingual polyphone tree to new languages. PDTS uses the fact that some of the leafs in the multilingual decision tree are not specialized enough for the new language. To do this, the following steps are performed:

- Contexts that are not seen in the target language, are completely removed from the tree.
- The clustering procedure is restarted on the adaptation material in the new language, leading to new, finer grained models that fit the target language better.

[Sch00b] does not describe the way, the new found models are being trained. [WS03] reports on using MAP to train the models, but also gives no further details on the training procedure for the new models.

For our experiments using PDTS we followed our own approach to initialize the newly clustered models [Stü08b] before applying the EM training for adaptation to German. This approach consists of two parts. In the first part we trained the models in the newly clustered tree on English, Russian, and Spanish using the LDA Matrix from the ML3-Mix models using the following procedure:

- Random samples from the English, Russian, and Spanish data were extracted using the existing forced alignments.
- The models were initialized using k-means.
- The standard training procedure as for the ML3-Mix model was performed.

The second part consists of initializing the models only on the German adaptation data by:

- Obtaining forced alignments on the German adaptation data using the EM adapted ML3-Mix model from Section 5.4.3
- Extracting random samples from the German adaptation data
- Performing k-means in order to initialize the codebooks of the newly found models

It can now happen that a model in the specialized decision tree, that has been trained on the English, Russian, and Spanish data and was refined on German, has not seen enough training data from English, Spanish, and Russian because its context was not observed often enough. Therefore models that saw fewer training material on English, Russian, and Spanish than on German were substituted with the models that were estimated by the k-means algorithm on the German adaptation data.

After this substitution, the models in the refined tree are initialized and we applied again one iteration of EM training on the German adaptation data as in Section 5.4.3.

When applying this procedure, the resulting system achieves a word error rate of 46.9% on the German dev data. Over the baseline without PDTS this is an improvement of 1.3% relative. As we will see in the next subsection, applying our modified version of PDTS, improves the WER even further.

5.5.2 Modified PDTS

In order to improve the performance of the PDTS we first applied a pruning scheme to the decision tree before applying PDTS. The pruning scheme removes leaves in the decision tree that are underrepresented in the German adaptation data, similar to the procedure described in [Wol99]. In the PDTS as described so far, only leaves are removed from the tree that were not observed at all in the German adaptation data. We now loosened that constraint by removing leaves that have been see fewer times than a certain threshold. In that way it is possible to trim sub-trees that are too specialized for German back to a level that has a grade of detail that is appropriate for German. Depending on the amount of adaptation data available it can also make sense to trim back the tree even further, leading to intermediate models that are coarser and more general than actually necessary. But then the following restart of the cluster procedure has the possibility to refine these coarser models again leading to a partition of the polygraphemes that is different from the one in the original ML3-Mix tree and that is more suitable for the distribution of polygraphemes in German speech and their relation to their underlying pronunciation.

In order to perform the pruning of under-represented models in the ML3-Mix tree, we needed an estimate for the frequency with which the sub-polygraphemes covered by the models occur. One possibility is to calculate forced alignments on the adaptation data and then do a framewise count of the occurrences. However, we opted for a different procedure that only uses the text data of the adaptation data without a frame wise assignment of the HMM states to the adaptation data. The motivation behind that is the fact that we do not want to rely on the rather bad forced alignment of the adaptation data that can be expected even from the EM adapted ML3-Mix models. But it is the best model that could be used for obtaining the forced alignments. The second reason is that the tree pruning scheme that is solely based on text material, can even be performed if no acoustic adaptation data were available.

So, we counted the occurrences of polygraphemes in the transcriptions of the adaptation material and determined the leave in the decision tree to which each polygrapheme belongs. In that way one gets a rough estimate on how much training data a model in the decision tree receives. The pruning was done in an iterative way:

- Determine the leave in the tree with the lowest count in the adaptation material
- Cut its trunk from the tree
- Distribute the polygraphemes that it covered over the remaining tree

Table 5.4: WER of naive adaptation, PDTS, and modified PDTS with varying pruning thresholds on GE dev and eval

Method	Threshold	dev	eval
—	—	47.5	47.9
PDTS	—	46.9	—
mod. PDTS	5	47.1	—
	10	47.5	—
	15	47.2	-
	20	46.1	—
	25	45.9	—
	30	45.8	46.2
	35	46.9	—
	40	48.3	—
	45	47.9	—
	50	47.8	—

- Iterate until the leave with the lowest mincount reaches a predetermined threshold

We determined the optimal count threshold empirically by trying out a series of thresholds on the German development set.

Pruning the ML3-Mix tree was followed by applying PDTS to it, as described in the previous section, and the new tree and its models were trained as before.

The results of the complete, modified PDTS for different pruning thresholds are shown in Table 5.4. They are compared against the case of the traditional PDTS and no PDTS. As the table shows, the best performance is reached with the modified PDTS, using a pruning threshold of 30. With this set-up the WER drops to 45.8% on the development set. On the evaluation set this leads to a WER of 46.2%, a relative reduction of 3.5% compared to the case of no adaptation of the decision tree.

5.5.3 Influence of Adaptation Data Size

In order to see whether and how the proposed method scales to larger adaptation set sizes, we repeated the experiments with more adaptation data. So, in addition to the fifteen minutes of adaptation data, we also ported the context-

Table 5.5: WER on the German dev and eval sets, when porting the ML3-Mix models using identity-driven cross-language transfer and modified PDTS

Adaptation data	dev	eval
15 min	45.8	46.2
30 min	30.2	34.6
60 min	28.0	30.9
90 min	25.1	25.3
90 min + mod.PDTS	21.7	23.6

dependent ML3-Mix models using thirty, sixty, and ninety minutes of adaptation data.

For the case of ninety minutes of adaptation data we also used the modified polyphone decision tree specialization in order to adapt the context-dependent models. The resulting word error rates are depicted in Table 5.5.

As one can see, with additional training data the word error rates start to drop further, reaching 21.7% on the development set and 23.6% on the evaluation set when using 90 minutes of adaptation data and modified PDTS.

5.6 Data Driven Model Mapping for Cross-Language Transfer

While for phonemes the assumption that the same phonemes in different languages have a very similar acoustic manifestation is a reasonable one, for graphemes this assumption clearly does not hold for all cases. The performance of the ML3-Mix models suffers from the mismatch of the pronunciation of the same grapheme in different languages. By training the models for the graphemes on the data from all languages, it is to a certain degree possible for the models to learn the different pronunciations of the graphemes in the different languages. However, as seen above, when applying the models to a new language, the ML3-Mix models often only poorly fit the new language.

Parts of the relation between the graphemes of a language and their pronunciation is captured in the rules learned by the polygrapheme-cluster tree. In the previous section we have shown that adapting the tree using modified PDTS can compensate for some of the mismatch that occurs when the cluster tree is learned on languages different from the target language. However, modified PDTS can only compensate for mismatches that occur from the graphemic context of the polygrapheme models. It cannot compensate for mismatches that originate from the fact that the center grapheme of a polygrapheme has a different pronunciation in the target language than learned from the training languages. Here, the underlying problem originates from the grapheme identity based mapping that we apply for the mapping of the ML3-Mix models to the new language.

Another problem arises if there is not a sufficient overlap between the graphemes in the ML-Mix model and the target language. In that case a grapheme identity driven mapping cannot be applied to the ML-Mix model to the new language. For example, it is not possible to apply our ML3-Mix model to the Thai data based on the grapheme identities, since Thai uses a completely different script than the ML3-Mix model. For the Cyrillic graphemes in the Russian recognizer we used a romanization. That is permissible, since in our setup Russian takes the role of a well studied language. However, Thai in our experiments takes the role of a new, previously unseen language, about which only very little knowledge is available. Thus, it is not permissible to assume a romanization of the Thai script as given. Here, a different way than grapheme identity based mapping for applying the ML-3Mix model to Thai has to be found.

Therefore, we have examined the use of two kinds of data driven mappings between the multilingual models and the models in the target language. We work under the assumption that permutations between the pronunciations of the graphemes in different languages can happen, which can be detected in an automatic way. Since we intended to perform the detection of theses permutations in a data-driven way we utilized the adaptation data assumed to be available in the target language.

The first method, based on model distances, utilizes auxiliary Gaussian models for the different graphemes in the multilingual recognizer and in the target language. It relies on the existence of a forced alignment of the adaptation data. For German as the target language, we can obtain such an alignment from the identity-driven cross-language transfer. For Thai this is not possible. Here we assume that we have a manual alignment given that identifies the start and end times of the Thai graphemes in the audio signal. To simulate such an alignment we use the forced alignment from the full monolingual Thai recognizer trained in Section 4.4.

However, such a manual mapping is difficult to obtain, much more difficult than for phonemes who have a closer relationship to the acoustic signal. Therefore, in our second approach we do not assume such an alignment as given. Instead we use the fact that we know the number of graphemes in the sentences of the adaptation data and that we want to find a one-to-one mapping of the graphemes in the multilingual model and the Thai model. We then find the mapping of the multilingual models to the Thai models with the help of a Viterbi path through an HMM that offers all multilingual graphemes as alternatives.

5.6.1 Model Distance Based Mapping

The general procedure in this approach is to first train auxillary models in the target language and for the ML-Mix recognizer, and then to establish a mapping between the members of the two sets of auxiliary models using model based distance measures. We then used the resulting mapping to apply the ML3-Mix model to the target language and perform the same steps as for the grapheme identity driven mapping, that is EM adaptation and modified PDTS.

The mapping between the target models and the multilingual models can happen at two levels. First, one can try to establish a mapping between the graphemes of the target language and the multilingual recognizer. From that the sub-grapheme models of the ML3-Mix model can be mapped accordingly by taking over the grapheme mapping for the sub-grapheme models. Second, one can try to establish a data driven mapping directly between the subgrapheme models. In the first case one needs to train an auxiliary model for every grapheme in the target language, and one for every grapheme in the ML-Mix model. In the second case the auxiliary model needs to be trained for every subgrapheme in the ML-Mix model and the target language.

For the first case, that is a grapheme level mapping, we assume that it is reasonable to obtain a manual segmentation of the small amount of adaptation data in the target language. A manual segmentation of the adaptation data on a sub-grapheme level is however not reasonable to assume. Therefore, in order to be able to establish the data driven mapping on the subgrapheme level, one needs to have a reasonably good, automatic alignment. For Thai this is not the case, since up to here we have not been able to port the ML3-Mix model to Thai. For German, however, we can use the alignment obtained from the EM adapted system in Subsection 5.4.3.

When establishing a mapping on the subgrapheme level, it is only possible to apply the context independent ML3-Mix models to the target language, since the mapping of the sub-graphemic models does not give a rule for transferring the grapheme decision tree for the context dependent ML3-Mix model. The reason for that is the fact, that the singelton questions in the cluster tree ask for the identity of the graphemes in the context of the polygrapheme, not for the identity of the sub-graphemes. From the mapping of the sub-graphemes it is not possible to obtain a unique mapping on the grapheme level, since different sub-grapheme models of the same grapheme can potentially be mapped to sub-grapheme models that belong to different graphemes. Thus, it is not clear which grapheme to choose as the target.

But when establishing a mapping at the grapheme level, it is easily possible to apply the context dependent ML3-Mix model to the target language. Now the identities of the graphemes in the singleton questions of the cluster tree can be mapped to respective identities in the target language, thus allowing to transfer the ML3-Mix-CD grapheme decision tree to the target language.

5.6.1.1 Auxiliary Models

In order to calculate the distances between either the grapheme or the sub-grapheme models, we trained auxiliary models in the respective target language and for the ML3-Mix models. In contrast to the GMMs of the full models each model in the auxiliary model only has one Gaussian component per model. There are two reasons for choosing only one Gaussian density instead of a full grown GMM. First, the model resulting from that will be more robust given the rather limited amount of training data in the target language. But since we are not interested in recognition but only identifying models that seem similar, the information from the single Gaussian component should be sufficient for that task. Second, as we will see, most distance measures between Gaussian distributions are only defined for single Gaussian distributions, but not for Gaussian Mixture Models. The distance measures would then have to be extended to GMMs. It may however been doubted whether the known extensions for GMMs are well suited for describing the similarity between models.

For the ML3-Mix model we trained one auxiliary model on the subgrapheme level and one on the grapheme level. The necessary alignments of the training data came from the forced alignments obtained with the best ML3-Mix model.

For German we trained an auxiliary model on the subgrapheme level and on the grapheme level. For training the models on the grapheme level, the training data alignment were supposed to be manual. We simulated this by taking the forced alignment from the best monolingual German system from Section 4.4. For training the auxiliary model on the subgrapheme level we used the forced alignments obtained from the EM ML3-Mix adapted system from Subsection 5.4.3.

For the reasons mentioned above, for Thai we only trained an auxiliary model on the grapheme level, again assuming manual alignments which we simulated by using the forced alignments from the best Thai recognizer from Section 4.4.

5.6.1.2 Distance Measures

Using the Gaussian auxiliary models we established a mapping between the ML3-Mix models and the models of the target language by finding the closest pairs in the two model sets using distance measures on the Gaussian of the respective models. From literature several distance measures between Gaussians are known. For our experiments we compared the performance of four different distance measures: the Euclidean distance, the extended Mahalanobis distance, the Kullback-Leibler distance, and the Bhattacharya distance.

Euclidean Distance Given two Gaussian distributions

$$\Gamma_1(x) = \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_1\right|}} \exp^{-\frac{1}{2}(x-\mu_1)\Sigma_1(x-\mu_1)}$$

and

$$\Gamma_2(x) = \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_2\right|}} \exp^{-\frac{1}{2}(x-\mu_2)\Sigma_2(x-\mu_2)}$$

where d is the dimension of the input vector x, μ_1 and μ_2 are the means of the Gaussian distributions, and Σ_1 and Σ_2 their covariance matrices, one can take the Euclidean distance between their two mean vectors μ_1 and μ_2 as distance measure between Γ_1 and Γ_2:

$$d_{eucl}(\Gamma_1, \Gamma_2) = \sqrt{(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T} \tag{5.1}$$

This distance measure ignores the covariance matrices of the two distributions and solely relies on the mean values of the distributions. It therefore makes sense to apply this measure in situations where no or only little information about the similarity of two distributions is expected to be contained in the covariance matrices.

Extended Mahalanobis Distance The Mahalanobis distance can be used to measure the distance of a vector x to a set of samples that are distributed with a mean of μ and a covariance of Σ:

$$d_{Mhn}(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \tag{5.2}$$

The Mahalanobis distance can be extended to a distance measure between two distributions by combining the covariance matrices of the distributions:

$$d_{extMhn}(\Gamma_1, \Gamma_2) = \sqrt{(\mu_1 - \mu_2)^T \left(\Sigma_1 + \Sigma_2\right)^{-1} (\mu_1 - \mu_2)} \tag{5.3}$$

Compared to the Euclidean distance the Extended Mahalanobis distance has the advantage that it also considers the covariance matrices of the distributions in addition to the mean vectors. However, the covariances of both models involved are first combined into one, thus not using the full distance information available in the two covariance matrices.

Kullback-Leibler Distance The Kullback-Leibler divergence between two probability functions P_1 and P_2 is defined as [Run07]:

$$d_{kl}(P_1, P_2) = \int P_1(x) \log \frac{P_1(x)}{P_2(x)} \tag{5.4}$$

The Kullback-Leibler divergence can bee seen as a dissimilarity measure between two probability functions. However, it is not symmetric and does not obey the triangle inequality and is thus not a true metric. In order to be able to use it as a distance function, one can make it symmetric by averaging the Kullback-Leibler divergence between P_1 and P_2 with the divergence between P_2 and P_1:

$$d_{kl-sym}(\Gamma_1, \Gamma_2) = d_{kl}(\Gamma_1, \Gamma_2) + d_{kl}(\Gamma_2, \Gamma_1) \tag{5.5}$$

For the case that P_1 and P_2 are Gauss distributions with diagonal covariance matrices, the symmetric Kullback-Leibler divergence takes the following form:

$$\begin{aligned} &d_{kl-sym}(\Gamma_1, \Gamma_2) \\ &\qquad = \frac{1}{2} \sum_{i=1}^{d} \frac{\sigma_{1,i}^2}{\sigma_{2,i}^2} + \frac{\sigma_{2,i}^2}{\sigma_{1,i}^2} - 2 + \left(\frac{1}{\sigma_{1,i}^2} + \frac{1}{\sigma_{2,i}^2} \right) (\mu_{1,i} - \mu_{2,i})^2 \end{aligned} \tag{5.6}$$

where μ_1, μ_2 are mean values of Γ_1 and Γ_2, while $\sigma_{1,i}$ and $\sigma_{2,i}$ are the ith element of the diagonal of covariance matrix Σ_1 and Σ_2, respectively.

Bhattacharya Distance When working in a two class scenario often the Bhattacharya distance is used [LT00]:

$$d_{bhatt}(P_1, P_2) = -\ln\left(\int_x \sqrt{P_1(x)P_2(x)}\right) \tag{5.7}$$

The Bhattacharya distance is symmetric but does not necessarily obey the triangle equation. For the case that Gaussian distributions with diagonal matrices are used it takes the form:

$$d_{bhatt}(\Gamma_1, \Gamma_2) = \frac{1}{2}\sum_{i=1}^{d} \ln\left(\frac{\sigma_{1,i}^2 + \sigma_{2,i}^2}{2\sqrt{\sigma_{1,i}^2\sigma_{2,i}^2}}\right) + \frac{|\mu_{1,i} - \mu_{2,i}|^2}{2\left(\sigma_{1,i}^2 + \sigma_{2,i}^2\right)} \tag{5.8}$$

5.6.1.3 Subgrapheme Level Mapping

With the help of these distance measures the best matching, context independent ML3-Mix subgrapheme models were found for the German sub-graphemes. Using this mapping, the ML3-Mix-CI models were applied to the German data.

Table 5.6 shows the word error rates for the different distance measures and compares them against the WER of the identity driven mapping approach from Section 5.4.2. From the word error rates we can see that the Euclidean distance and the extended Mahalanobis distance clearly lack behind the Bhattacharya and the Kullback-Leibler distance. This indicates that the covariance matrix of the models contains valuable information for finding a good mapping.

With a WER of 86.9% The Kullback-Leibler distance achieves the best performance on the German development set. On the eval set this mapping achieves a WER of 87.9%. Compared to the grapheme identity driven mapping this a relative reduction in WER of 2.4%. When adapting the resulting models on the German adaptation data using one iteration of EM training just as in Section 5.4.3 the word error rate drops to 62.1% on the dev set and 64.7% on the eval set. Compared to the case of the EM adapted grapheme identity mapped models this constitutes a relative reduction in WER of 6.9%.

Table 5.6: WER on the German test data for the different distance measures when establishing a mapping on the subgrapheme level

on German	dev	eval
grapheme identity	89.2%	90.1%
Euclidean	88.8%	—
Ext. Mahalanobis	89.3%	—
Bhattacharya	87.1%	—
Kullback-Leibler	86.9%	87.9%
grapheme identity + EM	65.8%	68.7%
Kullback-Leibler + EM	62.1%	64.7%

5.6.1.4 Grapheme Level Mapping

Using the distance measures above and the auxiliary models on the grapheme level, we established the best mapping between the ML3-Mix graphemes and the German graphemes, and the Thai graphemes respectively, again for the four different distance measures. Thus, this time it is also possible to port the ML3-Mix model to the Thai data, even though Thai has a completely different alphabet from the languages used to train the ML3-Mix model. Furthermore, since we are not mapping the models in the ML3-Mix model, but rather the graphemes, it is also possible to apply the context dependent ML3-Mix model to the German and Thai data.

Table 5.7 shows the resulting word error rates for the German system while Table 5.8 shows the same numbers for the corresponding Thai system.

For German we can again see that the Euclidean distance and the extended Mahalanobis distance lacks behind the Bhattacharya and the Kullback-Leibler distance just as for the sub-grapheme level mapping.

Finding a grapheme mapping based on either the Bhattacharya distance or the Kullback-Leibler distance performs best on the German development set. In the context independent case it leads to a WER of 88.2% on the development set, and a WER of 88.9% for the Bhattacharya distance, 88.7% for the Kullback-Leibler distance respectively, on the evaluation set. Compared to the grapheme identity based mapping this is a reduction in WER of 1.1% relative, far less than when automatically mapping on a subgrapheme level.

Unlike in the case when mapping on the subgrapheme level, this time we can also port the context dependent models. Again the Bhattacharya distance and the Kullback-Leibler distance perform best. They achieve a WER of 82.9% on the development set, and 84.4% on the evaluation set. Compared to the grapheme identity based mapping this means a loss in performance. Also, when applying modified PDTS to the EM adapted Kullback Leibler distance based mapping models, the WER is with 50.1% on the development set and 50.7% on the evaluation set still significantly higher than when using a grapheme identity based mapping in combination with PDTS. Therefore, even though we can only transfer the context-independent ML3-Mix models to German for the grapheme identity driven mapping at the subgrapheme level, it still performs better than the grapheme level mapping including the transfer of the context-dependent models. On Thai the Bhattacharya and the extended Mahanalobis Distance give the worst results. This behavior is in contrast to the results seen on German, where those two measures gave the best performance. Apparently when mapping the ML3-Mix models to Thai the information contained in the covariance matrix of the auxiliary models is misleading rather than helping.

Table 5.7: WER on the German test data for the different distance measures when establishing a mapping on the grapheme level

	Context Independent		Context Dependent	
on German	dev	eval	dev	eval
grapheme identity	89.2%	90.1%	79.0%	79.0%
+ mod. PDTS	—	—	46.4%	46.4%
Euclidean	91.6%	—	87.8%	—
Ext. Mahalanobis	92.4%	—	90.0%	—
Bhattacharya	88.2%	88.9%	82.9%	84.4%
Kullback-Leibler	88.2%	88.7%	82.9%	84.4%
Bhattacharya + EM	69.7%	72.3%	53.4%	54.3%
Kullback-Leibler + EM	69.7%	72.3%	53.4%	54.3%
+ mod. PDTS	—	—	50.1%	50.7%

So, for Thai, using the Extended Mahalanobis distance leads to the best results. For the context independent models it achieves a WER of 84.7% on the development set and 84.3% on the evaluation set. The context dependent models perform worse than the context independent ones, yielding a performance of 88.4% on the development set and 88.7% on the evaluation set. After adapting the mapped, context-independent models using EM training, the error rate drops significantly to 70.4% on the development set and 75.4% on the evaluation set. Adapting the context dependent, mapped models using EM improves the WER even further, bringing it down to 68.8% on the development set and 72.1% on the eval set. So, unlike prior to EM adaptation, the context-dependent models now perform better than the context-independent ones. When applying modified PDTS to the context dependent models, the WER is further reduced to 60.5% on the development set and 67.3% on the eval set.

Since a grapheme identity based mapping for the Thai graphemes is not possible this is the lowest error rate that could be achieved on Thai when porting the ML3-Mix model to Thai and exploiting the available 15 minutes of Thai adaptation data.

Table 5.8: WER on the Thai test data for the different distance measures when establishing a mapping on the grapheme level

on Thai	Context Independent		Context Dependent	
	dev	eval	dev	eval
Euclidean	86.6%	—	89.2%	—
Ext. Mahalanobis	84.7%	84.3%	88.4%	88.7%
Bhattacharya	89.1%	—	91.6%	—
Kullback-Leibler	89.6%	—	92.0%	—
Ext. Mahalanobis + EM	70.4%	75.4%	68.8%	72.1%
+ mod. PDTS	—	—	60.5%	67.3%

5.6.2 Maximum Likelihood Based Mapping

As mentioned before, the manual labeling at the grapheme level, that we assume as given for the Thai adaptation data, is very difficult to obtain by human annotators, because the relation between graphemes and the acoustic signal might be very loose for certain languages, e.g. when clusters of graphemes are pronounced as one sound, as for example the sequence of letters ‘th’ in English.

One common way to find a mapping in a data-driven way is the use of confusion matrices, e.g. see [ŽKD$^+$05, ŽK06]. For example, one could decode the adaptation data using the multilingual model and then calculate a frame-wise confusion matrix between the multilingual models and the Thai graphemes. However, when doing this, one does not use all the available information. We know that we want to have a one-to-one mapping of the graphemes, that is to every Thai grapheme we want to find only one multilingual grapheme. When only doing decoding with a grapheme recognizer, one has only limited control over the number of graphemes produced for an utterance in the adaptation data, though we exactly now the number of graphemes that we want to obtain.

We therefore did not use a phoneme decoder for finding the confusion matrix. Instead, for every utterance of the adaptation data, we constructed an HMM consisting of as many graphemes in sequence as in the transcription of the adaptation data. For every grapheme position in the HMM we allowed as alternatives all the graphemes in the multilingual models. At the beginning and end of the HMM and inbetween words, as given by the transcript of the adaptation data, we allowed the silence model as an optional state in the HMM. Figure 5.12 shows the structure of this HMM.

Using the HMM and the corresponding Thai audio recording we calculated the most likely path through the HMM using the Viterbi algorithm. This path then gives us directly the confusion with the Thai graphemes. By performing this kind of HMM building and alignment for all utterances in the adaptation data, we can built a confusion matrix.

With the mapping from the confusion matrix we then initialized the models of a context-independent Thai grapheme based recognition system. Table 5.9 shows the word error rates of the Thai recognizers initialized this way. Compared to the word error rates that were obtained by applying the ML3-Mix models to German using the grapheme identity driven approach, these word error rates are higher. They are also worse than those when porting the ML3-Mix model to Thai using the data-driven mapping described before. This was to be expected, because for this approach we do not assume a labeling of the adaptation data at the grapheme level as given.

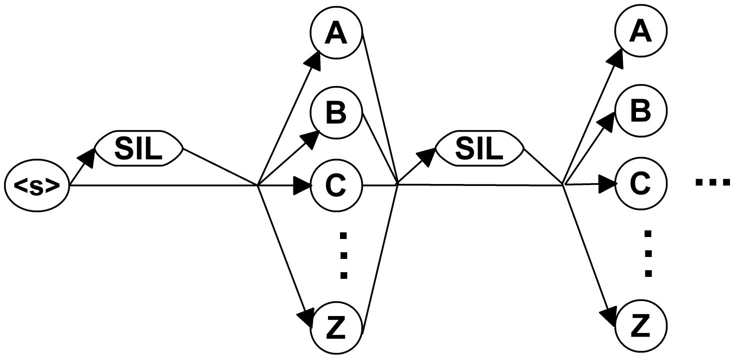


Figure 5.12: Structure of the HMM used for finding the most likely sequence of multilingual graphemes, given a Thai utterance

So, in the next step we start to adapt the initialized models using the adaptation data. Some of the Thai graphemes only occur very seldomly in the adaptation data. Even in the 90 minutes of adaptation data still two of the Thai graphemes do not occur at all. We therefore decided to use the largest available amount of adaptation data, i.e. the ninety minutes of Thai adaptation data. After

Table 5.9: WER of the Thai recognizers initialized using the maximum likelihood based mapping

adaptation data	dev
15 min	94.0%
30 min	94.2%
60 min	93.9%
90 min	92.9%

adapting the context-independent, acoustic models on the data, we made the step to context-dependent models.

Table 5.10 shows the resulting error rates. For the context-independent models the Thai recognizer reaches a WER of 41.2% on the development set and 42.5% on the evaluation set. The context-dependent models reach a word error rate of 34.5% on the development set and 35.5% on the evaluation set.

Due to the Thai writing system which is very different from the writing systems in the multilingual model from which we ported, the performance lacks behind that when applying the multilingual models to German. This due to the fact that the knowledge from the multilingual model cannot be exploited as efficiently in this case. We assume that one of the limiting factors is also the fact that the writing systems in the multilingual model are all alphabets, while Thai uses an abugida.

Table 5.10: WER of the context-independent and context-dependent Thai recognizers adapted on the ninety minutes of Thai adaptation data

adaptation data	dev	eval
CI	41.2%	42.5%
CD	34.5%	35.5%

Chapter 6

Crosslingual Acoustic Models with Articulatory Features

While in previous chapters we described our work in using graphemes as models for creating speech recognition systems and porting them to new languages, this chapter describes our work in enhancing the state-of-the art in porting traditional, phoneme based ASR systems. For that we make use of detectors for *articulatory features* (AF), such as place and manner of articulation. These detectors were shown in the past to give improvements in monolingual scenarios. In this chapter we extend their use to porting ASR systems to new languages.

6.1 Articulatory Features

Current state-of-the-art ASR systems usually model speech with Hidden Markov Models whose states correspond to phonemic or sub-phonemic units. Sometimes this model is called 'beads-on-a-string model' [Ost99]. Phonemes are a shorthand notation of articulatory positions, often the targets of the movement of specific articulators, that are characteristic for the sound that a phoneme is supposed to describe. The *International Phonetic Association* (IPA) has defined a set of phonemes for practically all languages in the world: the *International Phonetic Alphabet* also abbreviated as IPA [Ass99]. In that alphabet phonemes are separated into vowels and consonants. Vowels are described by the position of the highest point of the *dorsum linguae*—the upperside of the tongue—and whether the lips are rounded or not. Consonants are described as a combination of *voicing*, *place of articulation*, and *manner of articulation*. Voicing describes whether the vocal cords are vibrating during articulation or not. Sounds with vibrating cords are called *voiced* sounds, all others *unvoiced*. The place of articulation refers to the position of the greatest constriction in the vocal tract during articulation. For example, a sound can be dental or alveolar, depending on whether it is most constricted at the teeth or at the alveoli. The manner of articulation refers to the extend or kind of constriction at the place of articulation. For example, a sound for which the constriction is very narrow but still allows the flow of air is called a *fricative*. If the flow of air is temporarily cut-off and then suddenly released we call the sound a *plosive*. Figure 6.1 shows the complete alphabet for consonants and vowels and all manners and places of articulation classified by it.

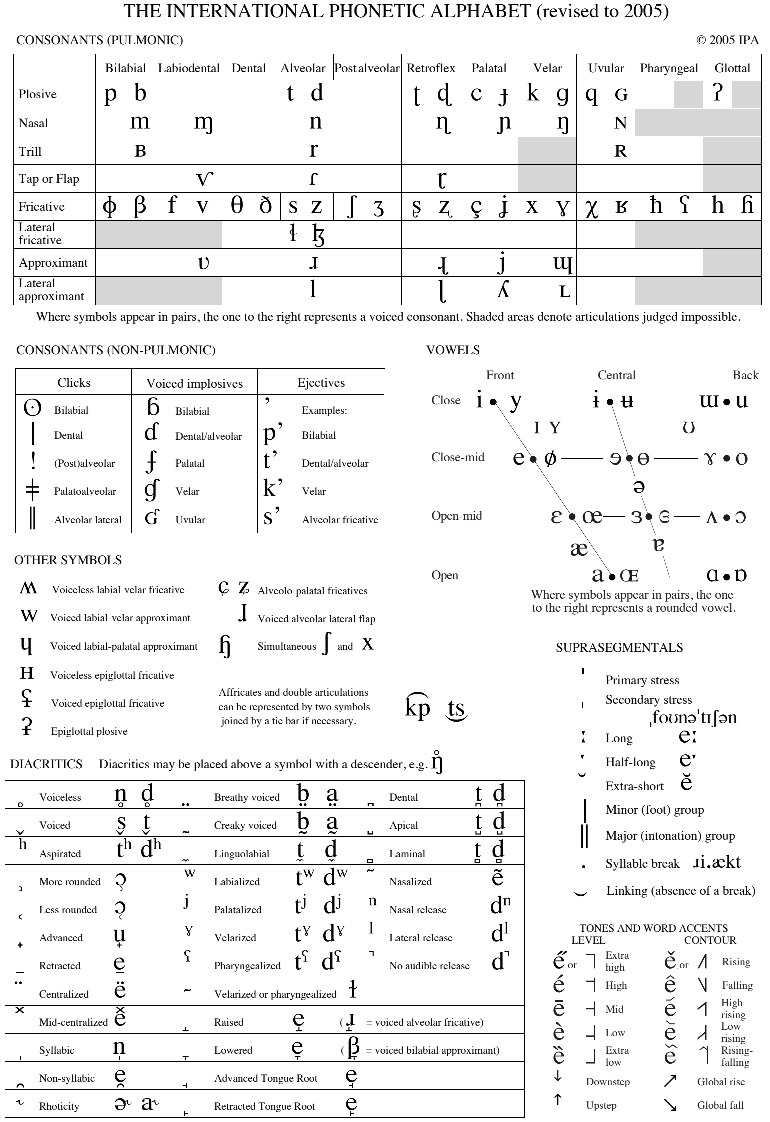
The articulatory properties, as for example IPA uses them to label phonemes, is what we call articulatory features in this work. We want to use this description of sound for modeling in order to improve existing techniques for porting speech recognition systems to new languages.

The use of phonemes as modeling units in ASR ignores the fact that the human articulators are in constant motion. Transitions among them are asynchronous and articulatory targets might be reached to differing degrees, e.g. depending on the phonetic context. The use of sub-phoneme models and context dependent phoneme models—polyphones—can compensate for this deficiency to a certain degree. However, polyphones suffer from the problem of accumulating sufficient training material in order to train robust models for all possible polyphones of a language. Therefore, they are clustered into generic models that share training data from multiple polyphones.

For these reasons, articulatory features seem to be better suited for modeling speech than phonemes and have been studied lately for use in ASR systems.

6.2 Related Work

In the past several researchers have worked on articulatory features in order to improve monolingual ASR systems. These works were mainly motivated by the fact that today's ASR systems still lack in performance compared to human capabilities in recognizing speech. Especially when switching from recognizing read or planned speech, which is very cleanly articulated, to spontaneous speech



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

© 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ	▨	ʔ ▨
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ			r				▨	ʀ		▨
Tap or Flap		ⱱ		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.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	ʼ	Examples:
ǀ	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
ǃ	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

VOWELS

	Front		Central		Back
Close	i • y		ɨ • ʉ		ɯ • u
		ɪ ʏ		ʊ	
Close-mid	e • ø		ɘ • ɵ		ɤ • o
			ə		
Open-mid	ɛ • œ		ɜ • ɞ		ʌ • ɔ
		æ	ɐ		
Open	a • ɶ				ɑ • ɒ

Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative

w Voiced labial-velar approximant

ɥ Voiced labial-palatal approximant

ʜ Voiceless epiglottal fricative

ʢ Voiced epiglottal fricative

ʡ Epiglottal plosive

ɕ ʑ Alveolo-palatal fricatives

ɺ Voiced alveolar lateral flap

ɧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. k͡p t͜s

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ŋ̊

̥	Voiceless	n̥ d̥	̤	Breathy voiced	b̤ a̤	̪	Dental	t̪ d̪
̬	Voiced	s̬ t̬	̰	Creaky voiced	b̰ a̰	̺	Apical	t̺ d̺
ʰ	Aspirated	tʰ dʰ	̼	Linguolabial	t̼ d̼	̻	Laminal	t̻ d̻
̹	More rounded	ɔ̹	ʷ	Labialized	tʷ dʷ	̃	Nasalized	ẽ
̜	Less rounded	ɔ̜	ʲ	Palatalized	tʲ dʲ	ⁿ	Nasal release	dⁿ
̟	Advanced	u̟	ˠ	Velarized	tˠ dˠ	ˡ	Lateral release	dˡ
̠	Retracted	e̠	ˤ	Pharyngealized	tˤ dˤ	̚	No audible release	d̚
̈	Centralized	ë	̴	Velarized or pharyngealized	ɫ			
̽	Mid-centralized	e̽	̝	Raised	e̝		(ɹ̝ = voiced alveolar fricative)	
̩	Syllabic	n̩	̞	Lowered	e̞		(β̞ = voiced bilabial approximant)	
̯	Non-syllabic	e̯	̘	Advanced Tongue Root	e̘			
˞	Rhoticity	ɚ a˞	̙	Retracted Tongue Root	e̙			

SUPRASEGMENTALS

ˈ Primary stress

ˌ Secondary stress ˌfoʊnəˈtɪʃən

ː Long eː

ˑ Half-long eˑ

̆ Extra-short ĕ

| Minor (foot) group

‖ Major (intonation) group

. Syllable break ɹi.ækt

‿ Linking (absence of a break)

TONES AND WORD ACCENTS

LEVEL			CONTOUR		
e̋ or	˥	Extra high	ě or	˩˥	Rising
é	˦	High	ê	˥˩	Falling
ē	˧	Mid	e᷄	˧˥	High rising
è	˨	Low	e᷅	˩˧	Low rising
ȅ	˩	Extra low	e᷈	˧˦˧	Rising-falling
ꜜ		Downstep	↗		Global rise
ꜛ		Upstep	↘		Global fall

Figure 6.1: The International Phonetic Alphabet reproduced by permission of the International Phonetic Association (Department of Theoretical and Applied Linguistics, School of English, Aristotle University of Thessaloniki, Thessaloniki 54124, GREECE)

a considerable drop in recognition performance can be observed with today's phoneme based models.

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.

In [Ost99] Ostendorf argues that pronunciation variability in spontaneous speech is the main reason for this drop in 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 [Kir98, Kir00, KFS00, Kir99] 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. For her, 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. She makes the assumption that coarticulation can be modelled more robustly in the production based domain than in the acoustic one. She also assumes that articulatory features are more robust towards cross speaker variation and signal distortions such as additive noise. Kirchhoff worked with *artificial neural networks* (ANNs) for classification of features and combined them in a hybrid HMM/ANN setup to obtain a recognition system for a small vocabulary recognition task. She later extended that work to a large vocabulary, continuous recognition task and showed improvements when combining articulatory feature based and phoneme based HMM recognition systems.

Eide [Eid01] argues that the direct modeling 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. In her experiments she augments the feature vector of a conventional HMM based recognition systems with the output from the classificators for a subset of features that seemed to give the best discriminative capabilities. The feature vector was not just augmented by the likelihoods of the corresponding features, but rather by a likelihood-ratio between models for the presence and the absence of a feature.

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. In their work they trained *multi-layer perceptrons* (MLPs) as detectors for features on a selection of training frames to boost classification performance. Their immediate goal in that is to have a high classification accuracy for features, but do not integrate them into a full ASR system. They also demonstrate that articulatory features can be recognized across languages by applying feature detectors trained on English to Dutch.

Livescu et.al. [LGB03] used *Dynamic Bayesian Networks* to model features such as voicing, manner, and velum position. They factored all possible feature states into a feasible number of clusters in order to counter the data sparseness problem. They have shown improvements with some features over a phoneme baseline on the Aurora task. In [LG04a] they applied their approach to pronunciation modeling. In [LG04b] they incorporated an approach for inter-feature asynchrony modeling.

In 2006, a research group at the Johns Hopkins summer workshop examined several different aspects of using articulatory features in speech recognition [LCHJ$^+$07, LCHJ$^+$06]. For them the most promising results were in using AF models within the tandem framework [HES00, CKK$^+$07]. Also, the hybrid HMM/ANN approach, though lacking behind other models in terms of classification accuracy, is promising for multilingual approaches, since it requires little training material. They also examined the use of articulatory features for audio-visual speech recognition [HJLLS07].

In [MW02] Metze and Waibel have enhanced monolingual, phoneme based recognizers with GMMs for articulatory feature in order to improve recognition performance. To do so, [MW02] introduced binary detectors for the presence and absence of a feature, e.g. whether a sound is voiced or not. Continuous features, such as the horizontal dorsum position for vowels, are modeled by multiple binary AF detectors for discrete positions, e.g for front, middle, and back. The binary detectors are modeled by GMMs, one GMM for detecting the presence of the feature, and one for detecting its absence. A flexible stream architecture is used to integrate the articulatory feature detectors into the recognition process.

In [SSMW03, SMSW03] we have demonstrated that the detectors for features as used in [MW02] can recognize features across languages. We have shown that for a total of five languages (Chinese, German, English, Japanese, and Spanish), multilingual feature detectors can be trained on multiple languages. We integrated them with monolingual phoneme based recognizers in cross- and multilingual ways showing that improvements for monolingual recognizers can be achieved by integrating feature detectors from multiple languages.

In [Met07, Met05] Metze then finds the stream weights for the feature detectors in a discriminative way and shows that it is possible to use the feature weights for speaker adaptation in a monolingual setup.

6.3 Articulatory Feature Detectors

For our experiments we used the same detectors for articulatory features as in [MW02, SSMW03, SMSW03, Stü04]. For every articulatory feature f two GMMs were trained. One GMM calculates the probability $P(x_i|f)$, i.e. that a sample x_i belongs to a sound with that feature. The other GMM calculates the probability $P(x_i|\bar{f})$; that is the probability, that the sample x_i belongs to a sound without that feature. Every GMM has 128 Gaussian components.

The labels for the training data were created with the help of forced alignments obtained from the phoneme based ASR systems. Since we assume that an articulatory feature is most stable in the middle of a phoneme, we trained the feature GMMs only on the middle states of the phonemes using 4 iterations of label training as in [MW02]. The preprocessing for the articulatory feature detectors was the same as for the phoneme based recognizers from which the forced alignments were obtained.

We evaluated the classification accuracy of the feature detectors on the development data of the respective languages. Evaluation was performed with the help of a naive Bayes classifier on a per frame basis. That is for every frame

x_i and for every AF f we evaluated the probability of the feature being present $P(x_i|f)$ versus the probability of the feature being absent $P(x_i|\bar{f})$. We then decide for the case with the higher probability:

$$P(f|x_i) \overset{?}{>} P(\bar{f}|x_i) \tag{6.1}$$

$$\frac{P(x_i|f) * P(f)}{p(x_i)} \overset{?}{>} \frac{P(x_i|\bar{f}) * P(\bar{f})}{p(x_i)} \tag{6.2}$$

$$P(x_i|f) * P(f) \overset{?}{>} P(x_i|\bar{f}) * P(\bar{f}) \tag{6.3}$$

Similarly as in the fundamental formula of speech recognition (3.1), we can decompose $P(f|x_i)$ and $P(\bar{f}|x_i)$ using the Bayes theorem and can then omit the term $P(x_i)$. Comparing the results of this classifier we then determined for every articulatory feature the frame-wise classification accuracy.

6.3.1 Multilingual Articulatory Features

In [SSMW03, SMSW03] we have shown that articulatory features can be reliably recognized across several languages. So, for example, AF detectors trained on English can be used to reliably detect the features of German speech. In that work it was also shown that AF can be modeled in a multilingual way. The share factor, that measures the overlap between different languages, was also shown to be larger for AF than for phonemes. The higher the share factor the more phonemes, features respectively, several languages have in come. So, a high sharefactor for AF indicates that AF might be very suitable for multilingual modeling and porting ASR systems to new languages. It was further demonstrated that in a monolingual scenario, in which the phoneme models were trained on the same language as the test set, performance can be improved by multilingual and crosslingual AF detectors, when combining them with phonemes based acoustic models.

6.3.2 Integrating AF Detectors into ASR

In order to integrate the AF detectors described above, Metze developed a flexible, stream based set-up [MW02, Met05]. In this set-up depicted in Figure 6.3 the likelihoods from the articulatory feature detectors are combined with the likelihoods from the phoneme models at the state level. The emission probability of a state in the HMM is calculated as a linear combination of probabilities in the log domain. This combination sums the likelihood from the corresponding sub-phoneme model with the corresponding feature present and feature absent

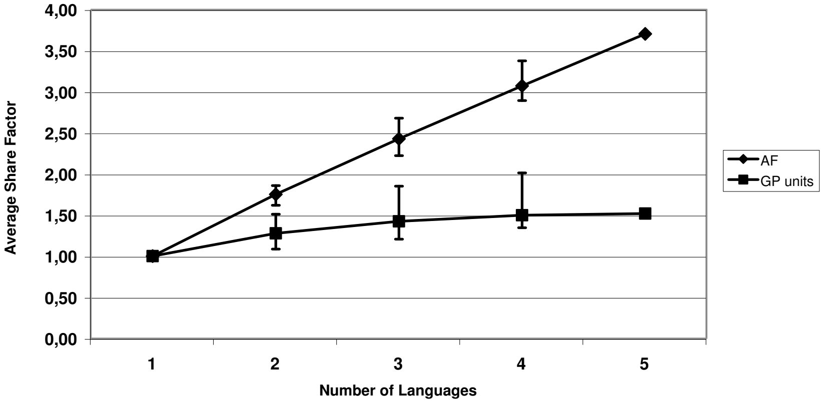


Figure 6.2: Average share factor for Chinese, German, English, Japanese, and Spanish for GlobalPhone phonemes and articulatory features (from [Stü04])

models that belong to the phoneme that the state stands for. So, for example, the likelihood of a voiced palatal phoneme, that has no other feature associated with it, is calculated as the sum of the likelihood of the sub-phoneme model, the likelihood of the voiced feature detector, the likelihood of the palatal detector, and the likelihoods of all the absent detectors for all other AF.

6.4 Selecting Stream Weights

The combination of AF detectors and phoneme based models in the stream based architecture described in Section 6.1 requires the selection of a suitable set of stream 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.

For the past monolingual experiments we used two different approaches to select appropriate weights. The first approach is a simple heuristic based on the classification accuracy of the feature detectors, the second approach implements a discriminative training scheme which tries to select weights that minimize the word error rate of the resulting recognizer.

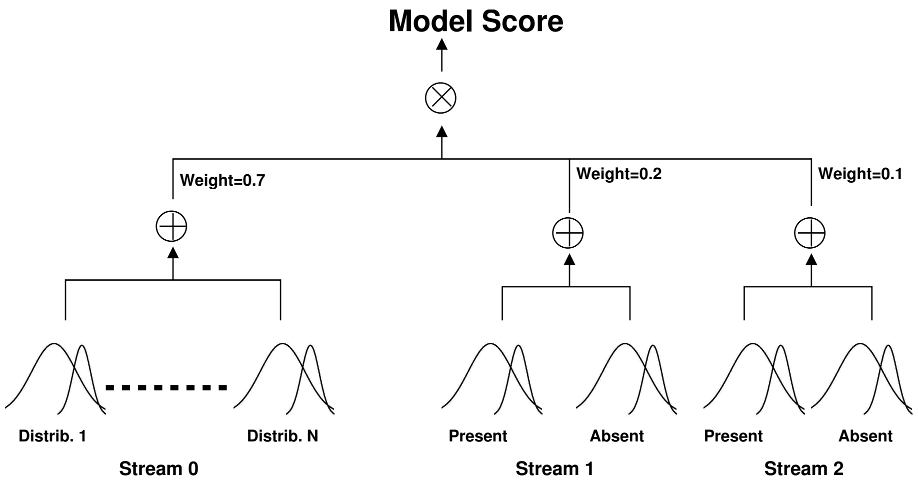


Figure 6.3: Stream based architecture for integrating the articulatory feature models (from [Met05])

6.4.1 Heuristic Weight Selection

The heuristic approach selects features based on their classification accuracy. To do so, the weight, which the feature detectors shall receive, is preselected as a fixed value—the same for all detectors added. In our case, a weight of 0.05 turned out to give good results. Then, feature detectors are successively added in the order of their classification accuracy on the development set of their training language. The weight of the phoneme HMM is chosen in such a way that all weights sum up to 1.0. After every addition of a new detector the WER of the resulting recognizers is measured on the development set. Usually the error rate starts to drop when adding detectors and reaches a minimum after adding a certain number of detectors. After that, the word error rate starts to rise again. In this way, the best number of feature detectors to add is determined.

6.4.2 Discriminative Model Combination

For a more refined training of the feature weights than the heuristic above, in the past we implemented the iterative approach of the *Discriminative Model Combination* (DMC), developed by Peter Beyerlein [Bey98], which is called *Minimum Word Error Rate* (MWE). MWE is based on the *Generalized Probabilistic De-*

scent (GPD) [JCL95].

DMC can be used to integrate multiple acoustic models into one log-linear posterior probability distribution, combining the different scores in a weighted sum at the log likelihood level. This is just as it is done in the approach of incorporating the feature detectors into the speech recognition system that we use.

So, given a hypothesis k, a weight vector Λ and the feature vector x the posterior probability of a hypothesis given an acoustic observation and a weight vector for combining the different streams is $p_\Lambda(k|x)$:

$$p_\Lambda(k|x) = C(\Lambda, x) exp \left\{ \sum_{j=1}^{M} \lambda_j \log p_j(k|x) \right\} \tag{6.4}$$

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.

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 smoothed approximation of the error function E_{MWE} that is used for MWE is:

$$E_{MWE}(\Lambda) = \frac{1}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \sum_{k \neq k_n} \mathcal{L}(k, k_n) S(k, n, \Lambda) \tag{6.5}$$

In this equation the k_n $(n = 1 \ldots N)$ are the N given training references for the discriminative training, while the $k \neq k_n$ are all other possible hypotheses. L_n is the length of the nth training utterance, $\mathcal{L}(k, k_n)$ the Levenshtein-distance. $S(k, n, \Lambda)$ is an indicator function that is used for smoothing the Levenshtein-distance. In order to get a differentiable error function E_{MWE}, S is set to be:

$$S(k, n, \Lambda) = \frac{p_\Lambda(k|x_n)^\eta}{\sum_{k'} p_\Lambda(k'|x_n)^\eta} \tag{6.6}$$

$p_\Lambda(k|x_n)$ 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 our experiments we used $\eta = 3$ and also approximated the posterior probabilities of the hypotheses by their acoustic likelihood.

For the estimation of E_{MWE}, equation 6.5 and 6.6 take into account all possible hypotheses k. This is of course not feasible for the numerical computation of E_{MWE}. Therefore, the set of hypotheses is limited to the most likeliest ones. In our experiments we used the hypotheses from an n-best list, where n was set to 150. The n-best list was obtained by rescoring the word lattice that resulted from the decoding process.

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

$$\tilde{\mathcal{L}}(k,n,\Lambda) = \mathcal{L}(k,k_n) - \sum_{k' \neq k_n} S(k',n,\Lambda)\mathcal{L}(k',k_n) \qquad (6.7)$$

With this partial derivative one can construct a gradient descent:

$$\lambda_j^{(I+1)} = \lambda_j^{(I)} - \frac{\epsilon\eta}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \sum_{k \neq k_n} S(k,n,\Lambda^{(I)}) \tilde{\mathcal{L}}(k,n,\Lambda^{(I)}) \log \frac{p_j(k|x_n)}{p_j(k_n|x_n)} \qquad (6.8)$$

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 our research we 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. Also, for the use of similar, discriminative training schemes, such as *Maximum Mutual Information* (MMI), is has turned out that the use of a weak language model is of advantage, in order to strengthen the influence of the acoustic model during training [Pov04].

6.5 Experiments

In order to test whether AF models can help when porting ASR systems to new languages, we examined several different scenarios. In all scenarios German takes the role of the new, previously unseen language, to which we want to port ASR models. We ran experiments for porting monolingual, English phoneme models enhanced by monolingual and multilingual articulatory feature detectors and for porting multilingual phoneme models enhanced by monolingual and multilingual articulatory feature detectors to German.

For the selection of suitable stream weights we compare the performance of the heuristic described in 6.4.1 against the performance of weights determined by the DMC as described in 6.4.2. Since for both approaches the features and weights are determined on a development set, which is not necessarily available in the target language, we also examined whether we can use the parameters determined on a development set in a different language for porting the systems to German.

6.5.1 Baseline Systems

As a baseline for our experiments serves the performance of monolingual phoneme based speech recognition systems tested on their training language. The acoustic models of the recognizers are left-to-right continuous HMMs with three states per phoneme. Training was done with the help of forced alignments obtained from previous systems. For training the acoustic models, first the LDA matrix was estimated, after that random samples for every model were extracted in order to initialize the models with the help of the k-means algorithm. Then these models were refined by six iterations of label training along the forced alignments and four iterations of *expectation maximization* (EM) training. The resulting models were used to obtain new forced alignments and the training procedure was iterated until a minimal *word error rate* (WER) on the development set was reached. *Context-independent* (CI) as well as *context-dependent* (CD) models were trained in this way. Table 6.1 shows the word error rates of the context-independent and context-dependent models for every language on their respective development and evaluation sets. The trigram language models used for English, Russian, and Spanish were unchanged from previous experiments, e.g. in [SW01].

We further trained a multilingual model using the technique ML-Mix on the languages English, Russian, and Spanish. Table 6.2 shows the word error rates of this model on the individual training languages. As expected we can see that

Language		EN	GE	RU	SP
CI	dev	19.5%	23.4%	51.8%	40.2%
	eval	20.2%	28.1%	54.8%	28.7%
CD	dev	9.0%	11.7%	33.9%	25.2%
	eval	10.3%	13.0%	36.2%	17.2%

Table 6.1: WER of the monolingual phoneme based ASR systems on the dev and eval sets of their respective language

the word error rates go up for the multilingual model in all cases. This is due to the fact that sounds with the same IPA symbol are still pronounced slightly differently in the various languages. Therefore, the models are broadened for the different model classes and do not fit the individual languages as well as when trained exclusively on one of them.

Language		EN	RU	SP
CI	dev	24.4%	56.5%	45.7%
	eval	25.8%	59.6%	32.8%
CD	dev	12.4%	38.8%	27.8%
	eval	14.1%	40.7%	20.2%

Table 6.2: WER of the ML-Mix ASR system on the dev and eval sets of its training languages

6.5.2 Articulatory Feature Detectors

Using forced alignments obtained from the phoneme based ASR systems we trained models for the articulatory features as described in Section 6.1. We also trained multilingual detectors, as described above and in [SSMW03], on the languages English, German, and Spanish, just as for the phoneme based ML-Mix recognizer.

6.5.3 Porting Across Languages

For our porting experiments we examined two principal scenarios. In the first scenario we used an English recognizer which we applied to the German test data, in the second scenario we used an ML-Mix model trained on the languages English, Russian, and Spanish which we applied to the German data.

6.5.3.1 Porting Monolingual Recognizers Across Languages

In order to apply the English recognizer to German, the German phonemes in the German pronunciation dictionary that were not covered by the English model, were manually mapped to their closest English phoneme. As shown in Table 6.3, applying the English acoustic model in this way leads to a WER of 73.4% on the German development set, and 76.4% on the evaluation set.

Adding the English AF models to the phoneme based recognizer using the heuristic described in 6.4.1 reduces the WER to 68.7.% on the German development set. On the evaluation set the WER goes slightly up to 76.6%. This increase in WER on the evaluation set is a phenomena which we have observed before. It means that the weights found with the heuristic often do not generalize very well to unseen data. The optimal number of detectors was determined on the German development set for this experiment.

When calculating the weights for the AF detectors using DMC as described in 6.4.2 the word error rate drops to 68.4% on the development set. A slightly better reduction than with the heuristic. This time, the word error rate also drops on the evaluation set. This time the DMC weights were optimized on the English development set, not the German one, in order to apply as little German data and knowledge as possible. So, DMC has generalized well from the English development set to the German development and evaluation set. In other words the weights found by the DMC also generalize very well across languages, not only different test sets.

In the past, it was also shown to be beneficial to combine monolingual phoneme models with feature detectors from different languages [SMSW03]. We therefore also combined the English phonemes with the English, Russian, and Spanish feature detectors. Since the number of feature detectors becomes large and it is not clear whether the absolute classification error rates of the feature detectors are comparable across languages, for this experiment we only used the DMC for finding stream weights, but not the heuristic. Again, DMC was performed on the English development set. Using the detectors from all languages, the word error rate reaches 71.8% on the German development set and 75.3% on the evaluation set. An improvement compared to the phoneme baseline but not as good as if only using English feature detectors.

It is remarkable in the DMC experiments, that though the stream weights have been determined on the English development set, the weights that were found generalize very well to German and still lead to good improvements. When selecting weights for the AF detectors from all languages, however, this works not quite as well, as when just using English AF detectors. Here the mismatch between the English development set for weight optimization and a mixture of AF detectors from many languages seems to be too high when also switching the test language.

EN to GE	dev		eval	
	heuristic	DMC	heuristic	DMC
Phon.	73.4%		76.4%	
Phon. + EN AF	68.7%	68.4%	76.6%	73.0%
Phon. + all AF	—	71.8%	—	75.3%

Table 6.3: WER when applying the English recognizer to the German test data, without and with Articulatory Feature models

6.5.3.2 Porting the Multilingual Recognizer to German

For the multilingual scenario we first applied the ML-Mix model to the German test data without the use of AF detectors. This, like in the English case, serves as our baseline. As Table 6.4 shows, this leads to a WER rate of 65.0% on the German development set and 70.4% on the German evaluation set. As to be expected from earlier work these WERs are lower than when using only the English models. The multilingual models gain from the fact that the phoneme models have seen more diverse training data and more of the German phonemes are covered by the models from the ML-Mix model.

Next, we added the English models to the ML-Mix models as done before for the English phoneme models. When adding them using the heuristic, the WER drops slightly to 64.6% on the development set and 69.7% on the evaluation set. So, this time the weights determined by the heuristic generalize from the German development set to the English one. Applying DMC instead of the heuristic gives no improvements however. Apparently in this case the weights found by the DMC on the English development set do not generalize very well to German. This might be due to the mismatch between the multilingual phoneme model and the English only AF models.

When using the ML-Mix AF detectors instead of the English ones and adding them using the heuristic, the WER on the development drops down to 64.4%. On the evaluation set a WER of 69.6% is reached. The DMC, however, fails to find suitable feature weights in this case, assigning all feature streams a weight of 0 and thus leading to no improvement.

When adding the monolingual feature detectors from all languages, as it was done for English, the WER drops further down to 64.2% on the development set and 69.5% on the evaluation set, a relative reduction in WER of 1.3%. This time, the DMC was performed on the joint development sets of the ML-Mix training languages, English, Russian, and Spanish. The use of the monolingual AF detectors from all languages gave the best gain in performance. Again, it is remarkable that the weights found on the English, Russian, and Spanish development sets generalize very well to the German development and evaluation set.

ML-Mix to GE	dev		eval	
	heuristic	DMC	heuristic	DMC
Phon.	65.0%		70.4%	
Phon. + EN AF	64.6%	65.0%	69.7%	70.3%
Phon. + ML AF	64.4%	—	69.6%	—
Phon. + all AF	—	64.2%	—	69.5%

Table 6.4: WER when applying the ML-Mix recognizer to the German test data, with and without Articulatory Feature models

6.5.4 Porting the EM adapted Multilingual Recognizer to German

Like done in in our experiments for porting grapheme based recognition systems to new languages (see Chapter 5), we assume a small set of German adaptation data of 15 minutes length as given, in order to further improve the porting performance of the multilingual recognizer. In order to adapt the ML-Mix recognizer we use two iterations of EM training on the context-independent models and one iteration of EM training on the context-dependent models, just as before.

This adaptation without the use of the AF detectors brings the WER of the context-independent models down to 46.0% on the development set and 49.0% on the evaluation set. The WER of the context-dependent models falls to 42.7% on the development set and 44.8% on the evaluation set.

When now adding all monolingual AF detectors to the adapted, context-dependent models using DMC the WER drops further down to 42.1% on the development set and reaches 44.4% on the evaluation set.

ML-Mix to German	dev	eval
Phonemes CI	46.0%	49.0%
Phonemes CD	42.7%	44.8%
Phonemes CD + all AF	42.1%	44.4%

Table 6.5: WER when applying the EM adapted ML-Mix recognizer to the German test data, with and without Articulatory Feature models

6.5.5 DMC on the German Development Set

So far, when applying DMC, we have estimated the stream weights of the AF detectors on the dev sets of the training languages of the ML-Mix model: English, Russian, and Spanish. We expect that the weights estimated in that way are not optimal for German. In our last experiments we therefore estimated the stream weights on the German development set. Table 6.6 shows that this reduces the WER for the unadapted, context-dependent phonemes to 63.6% on the development set and 69.4% on the evaluation set. This is a relative reduction in WER of 2.2% on the dev set and 1.4% on the evaluation sets. Both reductions are higher than when estimating the DMC weights on the development sets of the training languages of the AF detectors.

For the adapted phoneme models the word error rate is reduced to 41.4% on the development set and 44.2% on the evaluation set. Especially for the adapted models in combination with all AF on the German development set the gains are much higher than when finding the weights on the dev sets of the AF training languages. These are the best word error rates that can be achieved with the 15 minutes of German adaptation data.

6.5.5.1 Scalability to More Adaptation Data

Just as we have done in the case of porting grapheme based systems to new languages before, we also examined the gains, that can be achieved from more adaptation data, for this scenario. For that we repeated the previous experiment with the ninety minutes adaptation data set. The results are shown in Table 6.6.

ML-Mix to German	dev	eval
phonemes	65.0%	70.4%
phonemes + all AF	63.6%	69.4%
adapt.phonemes 15min.	42.7%	44.8%
adapt.phonemes 15min. + all AF	41.4%	44.2%
adapt phonemes 90min.	24.6%	27.1%
adapt phonemes 90min. + all AF	24.0%	26.1%

Table 6.6: WER when applying the unadapted and EM adapted ML-Mix recognizer to the German test data, with and without Articulatory Feature models using DMC weights estimated on the German development set

One can see that going from fifteen to ninety minutes of adaptation data reduces the word error rate significantly. Without the articulatory feature detectors the word error rates drops to 24.6% on the development set and 27.1% on the evaluation set. Again, adding the articulatory feature detectors reduces the word error rate further, to 24.0% on the development set and 26.1% on the evaluation set.

6.6 Conclusion

In this work we examined the use of articulatory feature detectors in porting the acoustic model of a speech recognition system to a new language. For this, we combined monolingual and multilingual phoneme models with monolingual and multilingual articulatory feature detectors in a stream based setup. In all cases the word error rate could be lowered by the use of articulatory feature detectors. In more badly matched conditions, such as when porting an English recognizer to German, or unadapted ML-Mix models to German, the gains were higher—up to 4.5% relative—than in better matched conditions, such as porting an EM adapted ML-Mix model to German.

The stream weights that are necessary for our approach were either found with the help of a heuristic or by applying DMC. The latter showed better generalization behavior than the heuristic. Also, the weights that were estimated with the help of DMC on the languages other than the final test language generalized well to the new, unseen language.

Chapter 7

Unwritten Languages: Discovering Word Units

In the previous chapters we have examined scenarios in which we ported automatic speech recognition systems to new languages for which a writing system existed. For the case in which no pronunciation dictionary for the language in question existed we resorted to using graphemes as modeling units instead of phonemes.

In this chapter we will now examine the case that either no writing system for the language in which we want to create an ASR system exists, or even if one exists it is not commonly used. We will examine the case of creating an ASR system for a speech-to-speech translation system between a widely spoken language, lets say English, and a new, unknown language. We assume that currently human interpreters translate between these two languages, and that we can observe the human translator and the persons that he translates between.

To the best of our knowledge no concrete numbers of how many languages in the world are without a writing system exist in literature. But linguists estimate that the vast majority of languages is without a writing system [NR00]. Omniglot attributes their list of writing systems to only 685 languages [Omn]. So, if one wants to address all languages in the world, one has to prepare for encountering many languages without a writing system.

7.1 Scenario Description

In the scenario, that we want to examine in this chapter, we assume that we can observe the action of a human interpreter translating between a well known language, let us say English, and a new language for which we want to create an ASR system. The resulting ASR system is not supposed to work as a stand-alone transcription system for the new language. Rather, its output should be suitable for translating from that language into the well known language.

When communication with speakers of a less resourced language, maybe even one without a written representation, becomes necessary, it is often achieved with the help of bilingual human interpreters, a very costly resource. For example, English speaking doctors in a remote disaster area might communicate with their patients with the help of a human translator. Our goal is to exploit the translations of the human interpreter, in order to gather the material needed for training ASR and translation systems. In our experiments we examine the feasibility of automatically learning word units in the unknown language and their pronunciation by aligning the English word sequences, that are being translated by the human interpreter, with the phonetic output from the translator's speech. We assume that we have no knowledge about a potential writing system in the target language nor about possible word units. Thus we are only able to work with the phonetic representation of the interpreter's speech.

It is now our intention to exploit the observable actions for automatically discovering word units in the new language. The available knowledge sources in this scenario are:

- The utterances of the speaker of the well known language that is being translated. By referring to language as 'well-known' we intend to say that automatic language processing technologies such as well performing ASR systems exist. We thus assume that we can automatically and reliably recognize the speech from that speaker.
- The utterances from the speaker of the new languages. Since this language is not known, we assume that we do not have any NLP technologies for that language. In order to be able to exploit the data we assume that we can only obtain a phonetic transcription of that language without any word boundaries. That phonetic transcription can either be obtained manually, e.g. by a phonetician, or it is conceivable that this transcriptions are obtained automatically or semi-automatically by a language independent phoneme recognizer.
- The speech from the human translator. Here, the utterances in the well-

know language can be treated in the same manner as in the first bullet, while the speech in the new language needs to be treated the same way as described in the second bullet.

7.2 Related Work

In the past much work has been committed to discovering word units in an unsegmented phoneme sequence. These algorithms work without supervision and also do not take into account the parallel data that is available in our case. [RP02], [DeM96], and [Bre96] proposed algorithms for word discovery from raw data. [Gol01] and [CL05] describe unsupervised learning of morphology for highly-inflected languages. Similar approaches can also be found in genomics literature [Bre04].

7.2.1 Monolingual Word Discovery

Besacier et.al. combined several of the ideas and approaches presented in literature and proposed in [BZG06] to train speech translation systems on data that contains English words on the one side and phonemes on the other side. They conducted experiments on English words and Iraqi phonemes. In order to achieve good translation performance [BZG06] first ran an unsupervised word discovery algorithm on the Iraqi phonemes without considering the corresponding English word sequence and then trained the translation system on the discovered word like units. The phoneme sequence, on which they ran the word discovery algorithm, resulted from replacing the words in ASR output by their phoneme sequence as given in the pronunciation dictionary of the ASR system. As expected, translation performance dropped somewhat when using the automatically found word units instead of the regular ones, but the drop in performance was small enough, as to proof that the automatically found units can be used instead of words in speech-translation systems.

The algorithm from [BZG06] combines the following ideas:

1. Use the predictability of phonemes: the basic idea here, first suggested by [Har55], is that the number of distinct phonemes that are possible successors of the preceding string reduces rapidly with the length of that string unless a morph boundary is crossed. A slightly different way to implement this same idea is to compute the *mutual information* (MI) between all successive phonemes of an utterance, and to detect a morph

boundary when MI reaches a local minimum which is, at the same time, below a certain threshold.

2. Use word boundaries that are already available before (respectively after) phone sequences commonly seen at the beginning (respectively the end) of sentences.
3. Use word frequencies: after a first segmentation, discovered words with high frequency counts are probably real words while words with low counts may result from badly placed word boundaries
4. Use the strength of Viterbi decoding.

With these ideas, the iterative algorithm consists of four steps:

1. **Initialization:** perform a first word segmentation of the foreign training corpus using the MI criterion only.
2. **Vocabulary and segment language model training:** build a vocabulary of the 1,000 most frequent words found in the last segmented corpus; put word boundary marks in the unsegmented corpus according to this 1,000 word vocabulary and train a n-gram LM from this data.
3. **Decoding:** for each unsegmented utterance, infer the most likely segmentation (location of segment boundaries) using the language model obtained in step 2
4. **Iterate:** Go back to step 1 until a fixed number of iterations is reached.

7.3 Word Discovery from Parallel Data

In contrast to previous work, we now perform the word discovery by utilizing the knowledge that can be gained from automatically aligning the English word sequences with the foreign phoneme sequences. We feel that the English word sequence which is known to correspond to the foreign phonemes should give additional information that can be used for the word discovery.

The goal of our experiments is to automatically exploit all the data that is generated in the human interpreter scenario described above. We assume that one of the languages involved is a well known language that has been examined already for NLP, meaning that for example ASR systems for this language exist. English is such a language that is often used in scenarios as described here. For the other language it is only assumed that an unsegmented phonetic transcript of the words articulated by the translator is available. In a real-world application scenario this transcript has to be obtained manually by a skilled phonetician, or even better in an automatic way by a language independent phoneme recognition system. The construction of such systems is an area of research by itself (e.g. [Köh96], [SW01]). For the experiments in this thesis we chose to work with a reference phoneme transcription of the target speech, instead of automatic ones. In this way we want to exclude effects introduced by errors in the phoneme recognition of the target language and concentrate on the techniques for exploiting the parallel data.

7.3.1 Word Alignment

In order to segment the phoneme string of the target language into appropriate word units we propose to exploit the original English speech by establishing word-to-phoneme alignments between the individual English words and chunks from the phoneme sequence. The science of establishing word-to-word alignments for bilingual sentences has been well studied in the field of *machine translation* (MT). The alignment between a given source string with J words $s_1^J = s_1, s_2, ..., s_J$ and a target string with I words $t_1^I = t_1, t_2, ..., t_I$ is defined as a subset of the Cartesian product between the word positions of the two strings [ON03], [BPPM93], and [Koe09]:

$$A \subseteq \{(i, j) : j = 1, ...J; i = 1, ..., I\} \tag{7.1}$$

Usually the alignments are constrained in such a way that each source word is assigned exactly one target word; so for every word position j in the source sentence a word position $i = a_j$ in the target sentence is assigned and we can write the alignments as $a_1^J = a_1, ..., a_J$.

One solution to automatically finding such alignments between two sentences is the use of *statistical alignment models* and *statistical translation models* from *statistical machine translation* (SMT) [ON03]. One part of SMT tries to model the translation probability $P(s_1^J|t_1^I)$ which describes the relationship between a source language string s_1^J and a target language string t_1^I. Now, given the alignment a_1^J between s_1^J and t_1^I, that maps the source word at position j to the target word at position a_j, a statistical alignment model is defined as $P(s_1^J, a_1^J|t_1^I)$, and $P(s_1^J|t_1^I)$ can be expressed as:

$$P(s_1^J|t_1^I) = \sum_{a_1^J} P(s_1^J, a_1^J|t_1^I) \tag{7.2}$$

The statistical models in general depend on a set of parameters Θ: $P(s_1^J, a_1^J|t_1^I) = P_\Theta(s_1^J, a_1^J|t_1^I)$. The best parameters $\bar{\Theta}$ are found on a set S of parallel training sentences in such a way that they maximize the probability of the training set. One way to do this is to use Expectation Maximization (EM) training which in general will only find a local maximum for $\bar{\Theta}$. Given a sentence pair (s_1^J, t_1^I) the best alignment, that is the most probable alignment, between the two sentences can be found with the help of the trained parameters:

$$\bar{a}_1^J = \underset{a_1^j}{\operatorname{argmax}} P_{\bar{\Theta}}(s_1^J, a_1^J | t_1^I) \tag{7.3}$$

Different models, with different sets of parameters exist in literature, such as HMM models [VNT96] and the IBM 1-5 models [BPPM93]. For our experiments we use the IBM-4 model to generate the sentence alignments.

7.3.2 Alignment Error Rate

For assessing the quality of the found alignments between two sentences, [ON03] defines the *alignment error rate* (AER). For calculating the AER a set of manually annotated reference alignments is created. Due to the complexity and ambiguity of creating a reference alignment, the alignments a_j are labeled as either belonging to sure (S) alignments or possible (P) alignments, which are used for ambiguous alignments. Every sure alignment is also considered to be a possible alignment ($S \subseteq P$). The quality of the alignment found is then measured by appropriately defined precision and recall measures:

$$recall = \frac{|A \cap S|}{|S|}, precision = \frac{|A \cap P|}{|A|} \tag{7.4}$$

Thus a recall error only occurs if a sure alignment has not been found, while a precision error occurs if a found alignment is not even possible. The alignment error rate (AER) is derived from the well known F-measure [van79]:

$$AER(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \tag{7.5}$$

7.4 Combination of the Approaches

In order to combine the two approaches from Section 7.2.1 and Section 7.3 we took the most frequent words from the monolingual word discovery and replaced the corresponding phoneme sequences in the training data. The reasoning behind this approach is that the very frequently occurring words from the monolingual word discovery algorithm likely correspond to real words. By substituting them in the phoneme sequences, it should be easier for the alignment process to align the remaining phonemes in the training material to the English words. It turned out that taken the frequent words after the initial mutual information segmentation worked best for that purpose.

7.5 Data

Our experiments were conducted with the help of the English portion of the *Basic Travel Expression Corpus* (BTEC) [KSTY03] and a Spanish translation of it. BTEC consists of travel expressions taken from phrase books in order to cover every potential subject in travel conversations. Our version of BTEC with the corresponding Spanish translation of it consists of 155K parallel sentences. The size of the English vocabulary is 12K while that of the Spanish one is 20K.

In our experiments English plays the role of the well-studied language while Spanish takes the role of the under-resourced language about which little to nothing is known. In reality Spanish is of course a well known language with existing resources and systems. However, by pretending that it is unknown to us, we can simulate our approaches on existing data and can easily evaluate them. As mentioned above, for our exploratory experiments we use a perfect phoneme transcription of the Spanish sentences which we obtained by transforming the words in the Spanish corpus with the help of a dictionary that was generated by a rule based system that was used in the past for generating Spanish pronunciation dictionaries for ASR systems.

In preparation for the later experiments, sentence pairs from the training corpus were removed that were longer than 50 words or phonemes respectively and that exceeded a sentence length ration of 9-1. We then divided the corpus into three sets: a training set containing 142,810 sentence pairs, a development set with 2,000 sentence pairs, and a test set with 2,000 sentences. The development set is used for parameter tuning for the translation system, while the performance of the resulting translation systems is measured on the test set. All three sets are completely disjunct and do not contain any doublets.

7.6 Experiments

In order to evaluate the suitability of the word segmentations that we obtained from the different approaches, we trained Spanish-to-English translation systems on the respective corpora resulting from the segmentations. For training and testing we used the Moses toolkit [mos]. We performed the standard training and decoding procedure as described on the Moses homepage.

In addition to the systems trained on the automatically found word units, we also trained a translation system with the original Spanish, word based corpus. The result serves as our gold standard, whose performance provides us with an upper bound for the performance of the other systems. This standard system, as one would train it, if a word segmentation were available, achieves a BLEU score of 0.56 on our test set. Naturally, we expect the automatically found word units to perform more or less worse than this gold standard.

The results of all approaches are summarized in Table 7.2 and are explained in more detail in the following.

7.6.1 Monolingual Word Discovery

We applied the algorithm from [BZG06] as described in Section 7.2.1 to the phonemes of the Spanish training data on a per sentence level. The initial MI threshold was set to 1.0. Three iterations of language model training and segmentation were performed. The language model from the last iteration was then used to segment the Spanish development and test sets into word-like units as well. It is important to note that in step 2 of the algorithm, a segment boundary is made a priori more likely by using a bias factor in order to perform a more aggressive segmentation. The reason for this is that a false detection (put an incorrect word segment) may be not too critical for the training of the phrase table, while a false rejection (do not segment multiple words) may freeze some bad sequences before the MT training.

The translation system trained on this corpus reaches a BLEU score of 0.34 when translating from Spanish to English. This is a drastically higher drop in performance from the gold standard than reported in [BZG06]. One of the reasons for this can be the different corpus (Spanish-English BTEC instead of Iraqi-English). Another reason could be the fact that we are not working with the phoneme sequence produced by a word based speech recognition system, but rather the phoneme sequence as given by the reference and a pronunciation dictionary (i.e. the output from a speech recognition system with 0% word error rate).

7.6.1.1 Modifications to the Algorithm

We suspected that one of the problems with this approach could be that the segmentation of the development and test set with the language model introduces a lot of unknown words which were not seen in the training corpus. We therefore modified the segmentation procedure of the corpus in the following way:

From the word units found on the training data we extracted a dictionary with occurrence counts for the words. Using this dictionary we substituted the phonemes in the training, development, and test set by recursively substituting the longest matching phoneme sequence.

When we now train a translation system on the resulting training corpora and test it on the resulting test corpus, the BLEU score goes up by 5 points to 0.39.

7.6.2 Word Discovery from Parallel Data

7.6.2.1 Word-to-Phoneme Alignment

In the word to phoneme alignment we want to assign every English word a sequence of Spanish phonemes. For finding the word alignments we used the GIZA++ [ON00] toolkit and the Pharaoh training script [Koe04]. One result of the GIZA++ training besides the learned translation models is a word alignment for the sentences in the training set. Since the alignments have the restriction that each source word is assigned exactly one target word, English is the target language and Spanish the source language.

The quality of the resulting alignment can be measured with the help of the alignment error rate described in Section 7.3.2. In order to have a baseline number for the error rate of the alignments from the training, we also performed the IBM-4 model training for the word based Spanish corpus, instead of the phoneme based one.

Table 7.1 shows the precision, recall, and alignment error rate for the training on the bilingual corpus using Spanish phonemes and the bilingual corpus using Spanish words as a comparison. The alignment error rate for the alignment between the English words and the Spanish phonemes is, as was to be expected, higher than for the alignment with the Spanish words. This is due to the more complex task of aligning words with phonemes, instead of words. However, the numbers also show that the task is feasible and can be done with the existing alignment techniques. In order to get an impression of the alignments found by the training Figure 7.1 shows three sample alignments between the English words (top) and the Spanish phonemes (middle). Below the Spanish phonemes the figure shows the Spanish word transcription together with the word to phoneme mapping as given by our dictionary. The alignment a) in this figure is an example for a perfect alignment in a rather simple case, where the number of English words matches the number of Spanish words. b) is an example of a more complex alignment where the English word 'please' needs to be aligned to two Spanish words. Again the alignment found is correct. c) Shows an example of an even more complicated alignment. Here the alignment also needs to do a word reordering, the words 'hot' and 'milk' need to be swapped. And the English words 'I'd' and 'like' both need to be mapped to the Spanish word 'querria'. While the swap of 'hot' and 'milk' is done correctly, the alignment found for 'I'd' and 'like' is clearly wrong. Due to its constraints the IBM-4 model cannot find the correct alignment.

Table 7.1: Precision, recall, and AER for the alignments between English words and Spanish phonemes, words respectively

	Spanish phonemes	Spanish words
precision	83.5%	88.8%
recall	66.9%	75.3%
AER	25.4%	18.1%

7.6.2.2 Dictionary Extraction

From the found alignments it is now easily possible to extract dictionary entries. Every English word that is aligned to Spanish phonemes is a potential entry in the Spanish dictionary, with the English word serving as a generic word id in the Spanish dictionary. Different English words that were mapped to the same phoneme sequence were not combined into one word, so that homophones were generated. One special case, when extracting the words, needs to be considered. It can happen that an English word is aligned to a phoneme sequence that is not continuous in its phonemes' positions, but that has got holes or reorderings in its sequence. These sequences have to be split into its continuous subsequences, each subsequence corresponding to one Spanish word. Each subsequence then receives its own word identifier based on the English word to which it was aligned.

In a second step the resulting dictionary is filtered. Pronunciation variants to

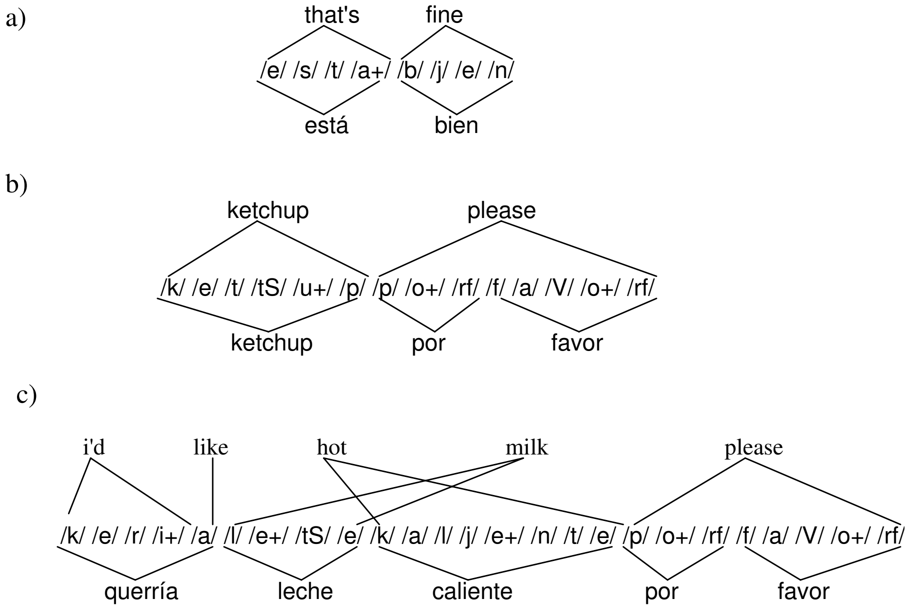

a)
that's
fine
/e/ /s/ /t/ /a+/ /b/ /j/ /e/ /n/
está
bien
b)
ketchup
please
/k/ /e/ /t/ /tS/ /u+/ /p/ /p/ /o+/ /rf/ /f/ /a/ /V/ /o+/ /rf/
ketchup
por
favor
c)
i'd
like
hot
milk
please
/k/ /e/ /r/ /i+/ /a/ /l/ /e+/ /tS/ /e/ /k/ /a/ /l/ /j/ /e+/ /n/ /t/ /e/ /p/ /o+/ /rf/ /f/ /a/ /V/ /o+/ /rf/
querría
leche
caliente
por
favor

Figure 7.1: Samples of alignments found by GIZA++

a word that occur less than 100 times in the training text are removed from the dictionary. This step is taken in order to eliminate pronunciation variants that were created due to erroneous word-to-phoneme alignments. The dictionary constructed in this way contains 15K words. 3,172 words in the original Spanish dictionary have an exact phonetic match in the dictionary constructed this way.

7.6.2.3 Evaluation

Using the extracted dictionary the phoneme sequences in the training, development, and test set were replaced by the words in the dictionary. Replacement was done by recursively replacing the longest matching phoneme sequence in a sentence with the corresponding word from the dictionary. In case of multiple, matching words the most frequent one was chosen.

We now performed the same training as before for the monolingual approach on the newly found corpora. On the test corpus the translation system reaches a BLEU score of 0.50. This is considerably higher than for the monolingual approach. So, the additional information given by the parallel English sentences could be exploited to find word units better suited for translation, now only lacking six BLEU points behind the gold standard.

7.6.3 Combination Results

When we combine the two approaches as described in Section 7.4, we can improve on that performance by one BLEU point reaching a score of 0.51. The number of words to take from the initial MI segmentation was empirically determined on the development set and set to 40. When using the most frequent words from the complete segmentation procedure, the gain from combination is smaller, only about 0.5 BLEU points.

Table 7.2: Results in BLEU Score of the Spanish-to-English translation with the different word discovery approaches

Word Segmentation Approach	BLEU
Gold Standard	0.56
Monolingual	0.34
Monolingual modified	0.39
Parallel Data	0.50
Combination	0.51

7.6.4 Suitability for ASR

So far, the experiments have only examined the suitability of the automatically found word units for machine translation purposes. However, we are interested in a full speech translation system which also includes ASR. Whether the word units are also suitable for ASR can be determined by looking at the language model perplexities of the different approaches. In order to measure this we trained 3-gram language models on the training sets that were segmented by the different approaches and measured their perplexities on the respective development and test sets.

Table 7.3 gives an overview over the perplexities of the language models estimated with the different kinds of word units on the development and test set.

Table 7.3: Language model perplexities for the Spanish dev and test set for the two monolingual word discovery approaches, the word discovery from parallel data, and the combination of the two approaches

Word Segmentation Approach	dev	test	vocab size
Gold Standard	45.6	45.1	16,838
Monolingual	52.7	51.5	22,883
Monolingual modified	104.9	104.6	21,004
Parallel Data	47.6	47.2	11,713
Combination	53.1	51.6	11,872

7.6.5 End-to-End Evaluation

In order to test the suitability of the discovered word units for real speech translation, we carried out a full end-to-end speech translation evaluation.

For this end-to-end evaluation we used a database of 1,000 Spanish BTEC sentences read by 12 speakers. The database and the acoustic model for the Spanish recognition system were taken from [PFS$^+$05, PSF$^+$05].

Using this system, the language models from Section 7.6.4, and the dictionaries constructed in Sections 7.2.1 and 7.6.2.2 we recognized the 1,000 read BTEC sentences. The vocabulary of the recognizer was derived from the training data, by taking all words that occur in the training data of the translation system. The word based system, that again is the gold standard for this experiment, reaches a word error rate of 16.3%. When translating this ASR output into English, the system reaches a BLEU score of 0.36.

Table 7.4 shows the BLEU scores when translating the ASR output corresponding to the monolingual word discovery approach and the approach exploiting the parallel data. As we can see, taking into account the parallel data as supervision improves the performance of the resulting speech translation system by three BLEU points.

Table 7.4: BLEU score when translating the ASR output obtained with the automatically discovered word units

Word Segmentation Approach	BLEU
Gold Standard	0.36
Monolingual	0.26
Parallel Data	0.29

7.7 Conclusion and Outlook

In this chapter we have examined the scenario of acquiring the data resources necessary for training a speech-to-speech translation system by observing the actions of a human interpreter. We have focused on discovering word like units in a previously unknown language for which we only have a phonetic transcript, but no word segmentation. We have introduced an approach to segmenting the phoneme sequence in the foreign language into words that makes use of all the available knowledge sources, including the parallel English sentences. The word discovery approach also produces a pronunciation dictionary which is necessary for the creation of a speech recognition system.

We conducted the first true end-to-end evaluation reported in literature by measuring the translation performance on the output of ASR systems that use the automatically found word units.

We have compared our new approach with the latest existing approach in literature that only works on the monolingual, foreign data. Our approach significantly outperforms this approach, showing that we can effectively exploit the parallel data.

When deploying such a system in real life, one would typically first try to learn short and simple sentences, or even only single words, mostly nouns. Similar to the way that a person would try to learn and discover a language when faced with it for the first time without a teacher at hand. In order to show the potential for this procedure in combination with our technique we conducted an oracle experiment. We assumed that we were able to learn all the nouns in the new languages, e.g. by pointing and recording the speech from the native speakers. We thus substituted all nouns in the unsegmented phoneme string of the target language by the correct word and then repeated our word discovery algorithm. When evaluating this approach on the perfect phoneme sequence of the test data, as in Section 7.6.2.3 the BLEU score improved from 0.50 to 0.53. Thus, learning a language in the right order will greatly improve the performance of the systems created in this way.

CHAPTER 8

Conclusion

The languages of the world show a high diversity. Roughly 4,000–7,000 languages exist in the world, many of which are only spoken by comparatively few speakers. At the same time many languages are threatened by extinction. Often speakers switch from their native language to a new language which seems to offer them better opportunities in a world in which information flows freely around the globe. At the same time, linguists argue that the upkeeping of a high linguistic diversity in the world is essential to a sound environment and the foundation of the many cultures in the world. By developing natural language processing systems, including automatic speech recognition systems, for all languages in the world, we believe that technology can help to stop this trend.

However, in order to be able to handle the high number of languages in the world, techniques have to be devised in order to develop natural language processing systems in a fast and affordable way.

In this thesis we have presented our work in providing and refining techniques for porting speech recognition systems to new languages. For the case that no dictionary in the target language is available or can be created, we have shown that the use of graphemes as modeling units is a suitable alternative. Their use eliminates the need for a pronunciation dictionary which is expensive and time-intensive to create.

We transferred work in porting phoneme based acoustic models to porting grapheme based models. Since for graphemes the overlap of models between the different languages might either be low or even no overlap might exist, we applied two data driven mapping techniques for porting grapheme based models to new languages.

In order to improve the performance when porting phoneme based acoustic models to new languages we incorporated articulatory feature detectors into the porting process, showing improvements over the case that no articulatory features are used.

Finally, we examined the scenario that the target language does not have any writing system. For this case we showed how by data-driven means, word units in the new language can be detected. By conduction a full end-to-end evaluation, we showed that these generically found word units are suitable for the use in a speech translation system.

8.1 Graphemes as Modeling Units

The pronunciation dictionary of a phoneme based ASR system, is time and labor intensive in its creation. The process usually requires either intensive manually labor of an expert or large amounts of training material or both.

The use of graphemes instead of phonemes as modeling units eliminates the need for such a pronunciation dictionary. In this work we have demonstrated that this approach is a feasible solution for many languages in the world. Due to the often complex relation between graphemes and phonemes, the model cluster tree of a grapheme based ASR system is of heightened importance. We accounted for this importance by replacing our traditional cluster tree with a flexible version. Using the flexible tree we could show improvements in word error rate for all languages examined.

8.2 Porting Grapheme based ASR to New Languages

Past research has addressed the problem of porting phoneme based acoustic models to new languages. Techniques for creating a common phoneme set for all languages were examined and language independent acoustic models were trained. For phoneme based models, finding a common model set is very simple and can be easily done by existing reference schemes such as IPA. Also, the application of such a model to a new language is very uncomplicated due to the language independent notation introduced by such reference schemes.

For grapheme based models the challenges are higher. Unlike phonemes, graphemes are pronounced very differently across languages. Also, many languages use different writing systems, so that no overlap or only a very little one exists between the languages.

In this thesis we transferred the method ML-Mix for creating and porting language-independent acoustic models to new languages to the case that models are based on graphemes. For the case that only little or no overlap between the graphemes of the multilingual model and those of the target language exists, we examined the use of two data-driven mapping methods for finding appropriate correspondences between the multilingual model and the target language.

Since the cluster tree for the context-dependent models is of heightened importance in grapheme based ASR, we used the technique of polyphone decision tree specialization to adapt the tree to the target language. We enhanced the specialization procedure by combining it with a pruning scheme which is applied prior to specialization. For the case that graphemes are used as modeling units we could show gains over the pure form of specialization.

8.3 Porting with the Help of Articulatory Features

In the past, models of articulatory features were successfully integrated into monolingual, phoneme based speech recognition systems. They improved the recognition performance, especially for conversational speech. Past research demonstrated also that articulatory features can be modelled in a multilingual way and can be recognized across languages.

In our work we used models for articulatory features to port speech recognition systems to a new language. We showed improvements when porting monolingual and multilingual ASR systems to a new language supported by articulatory features.

The architecture that we used for integrating the AF models into our recognition systems requires the selection of suitable stream weights. We were able to show that the weights learned by a discriminative method on a known language successfully generalize to the new target language. Thus, the use of AF does not necessarily require additional training data in the target language for selecting the stream weights.

8.4 Discovering Word Units in New Languages

Many languages in the world do not have a writing system. For this case we examined the scenario that the ASR system for theses languages is part of a speech-translation system. We have shown that word-like units can be automatically discovered in the unsegmented phoneme string of the new language by exploiting the bilingual data available for training the translation system.

In our experiments we conducted the first true end-to-end evaluation in literature, by using the automatically generated word units for real speech recognition followed by machine translation performed on the ASR output.

We compared our word discovery algorithm against the most recent, monolingual word discovery algorithm known from literature. We found that the exploitation of the additional information in the form of parallel data, leads to automatically generated word units that are better suited for automatic speech recognition and translation.

8.5 Outlook

In order to be able to provide automatic speech recognition systems for all languages in the world, we have presented in this thesis techniques which speed up and facilitate the process of porting existing ASR systems to new languages and which reduce the labor in this, e.g. by eliminating the need for a carefully constructed pronunciation dictionary.

But, especially when only very little adaptation data in the target language is available, current methods of porting ASR systems to new languages do not yet produce systems that are equal in quality to systems that have been studied thoroughly, in part over decades. But even though the systems resulting from our methods still lack in performance when compared to the well studied languages, they can serve as good initial systems for either unsupervised or semi-supervised learning. In statistical ASR the performance of systems is closely linked to the available amount of training material. For the major languages in the world time and money were invested over the past in order to produce these resources. For the majority of the languages in the world, the less-resourced ones, these resources will never be available.

In order to still produce systems that are of use in real-life applications, the old paradigm of first collecting sufficient amounts of training material, and then estimating the parameters of the models on them, has to be abandoned. Instead speech recognition systems will have to learn autonomously, when observing real-life interactions. In Chapter 7 we have already hinted at how these future systems might look like.

The techniques presented in this paper are suitable for incorporation into a framework of observing human interaction and learning from them in an unsupervised way. Automatic speech recognition and machine translation can form self-enhancing parts that can gain from each other [PFS+05, PSF+05, PSF06, PSF+07, PW08]. These systems then refine themselves during the interactions that are taking place anyway, and their training thus will not incur high additional costs.

When developing this notion further by not imposing any limits anymore on the type of media that the observing systems use for learning, these systems will turn into learning omnivores that improve their models from whatever piece of information will present itself to them during their deployment in real-life.

Bibliography

[AAB+96] U. Ackermann, B. Angelini, Fabio Brugnara, Marcello Federico, D. Giuliani, R. Gretter, G. Lazzari, and Heinrich Niemann. Speedata: Multilingual spoken data entry. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, pages 2211–2214, Philadelphia, PA, USA, October 1996. ISCA.

[AAB+97] Ulla Ackermann, Bianca Angelini, Fabio Brugnara, Marcello Federico, Diego Giuliani, Roberto Gretter, and Heinrich Niemann. Speedata: A prototype for multilingual spoken data-entry. In *Proceedings the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, pages 1807–1810, Rhodes, Greece, September 1997. ISCA.

[AD97] Ove Andersen and Paul Dalsgaard. Language-identification based on cross-language acoustic models and optimised information combination. In *Proceedings the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, pages 67–70, Rhodes, Greece, September 1997. ISCA.

[ADB93] Ove Andersen, Paul Dalsgaard, and William Barry. Data-driven identification of poly- and mono-phonemes for four european languages. In *Proceedings of the 3rd European Conference on Speech Communication and Technology EUROSPEECH'93*, pages 759–762, Berlin, Germany, September 1993. ISCA.

[ADL06] Martine Adda-Decker and Lori Lamel. *Multilingual Dictionaries*. Academic Press, Burlington, MA, 2006.

[All79] Erik Allardt. Implications of the ethnic revival in modern industrialized society: A comparative study of the linguistic minorities in western europe. In *Commentationes Scientarium Socialium*, volume 12. Commentationes Scientarium Socialium, Helsinki, 1979.

[AS03] Daniel M. Abrams and Steven H. Strogatz. Modelling the dynamics of language death. *Nature*, 424(21):900, August 2003.

[Ass99] International Phonetic Association. *Handbook of the International Phonetic Association.* Cambridge University Press, Cambridge, England, 1999.

[Bar08] Christopher Barbara. *International legal personality: Panacea or pandemonium? Theorizing about the individual and the state in the era of globalization.* VDM Verlag Dr. Müller, first edition, 2008.

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

[BGM97] Patrizia Bonaventura, Filippo Gallocchio, and Giorgio Micca. Multilingual speech recognition for flexible vocabularies. In *Proceedings the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, pages 355–358, Rhodes, Greece, September 1997. ISCA.

[BI99] Bogomir Horvat Bojan Imperl. The clustering algorithm for the definition of multilingual set of context dependent speech models. In *Proceedings of the Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99)*, pages 887–890, Budapest, Hungary, September 1999. ISCA.

[Bla06] Alan W. Black. *Multilingual Speech Synthesis.* Academic Press, Burlington, MA, 2006.

[BPPM93] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[BPSW70] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical statis*, (1):164–171, 1970.

[Bre96] Michael R. Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1-3):71–105, February 1996.

[Bre04] Michael R. Brent. Recent advances in gene structure prediction. *Current Opinion in Structural Biology*, 14(3):264–272, May 2004.

[BZG06] Laurent Besacier, Bowen Zhou, and Yuqing Gao. Towards speech translation of non written languages. In *Spoken Language Technology Workshop, 2006*, pages 222–225, Aruba, December 2006. IEEE.

[CAGADL97] Cristobal Corredor-Ardoy, Jean Luc Gauvain, Martine Adda-Decker, and Lori Lamel. Language identification with language-independent acoustic models. In *Proceedings the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, pages 55–58, Rhodes, Greece, September 1997. ISCA.

[Can05] Luciano Canepari. *A Handbook of Phonetics*. Lincom Europa, Munich, Germany, 2005.

[CDG$^+$97] P. Cohen, S. Dharanipragada, J. Gros, M. Monkowski, C. Neti, S. Roukos, and T. Ward. Towards a universal speech recognizer for multiple languages. In *Proceedings the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 591–598, Santa Barbara, CA, USA, December 1997. IEEE.

[CGW01] S. Chang, S. Greenberg, and M. Webster. An elitist approach to articulatory-acoustic feature classification. In *Proceedings of the Seventh European Conference on Speech Communication and Technology EUROSPEECH 2001 Scandinavia*, pages 1725–1728, Aalborg, Denmark, September 2001.

[CHS06] Paisarn Charoenpornsawat, Sanjika Hewavitharana, and Tanja Schultz. Thai grapheme-based speech recognition. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 17–20, New York, NY, USA, June 2006. Association for Computational Linguistics.

[CKK$^+$07] O. Cetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel, and K. Livescu. An articulatory feature-based tandem approach and factored observation modeling. In *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, April 2007. IEEE.

[CL05] M. Creutz and K. Lagus. Induction of a simple morphology for highly-inflecting languages. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51, Barcelona, July 2005.

[Cry87] David Crystal. *The Cambridge Encyclopedia of Language*. Cambridge University Press, Cambridge / New York / Melbourne, 1987.

[Cry00] David Crystal. *Language Death*. Cambridge University Press, Cambridge, UK, 2000.

[CY95] John Clark and Colin Yallop. *An Introduction to Phonetics and Phonology*. Blackwell Publishers, 2nd edition, 1995.

[DA92] Paul Dalsgaard and Ove Andersen. Identification of mono- and poly-phonemes using acoustic-phonetic features derived by a self-organising neural network. In *Proceedings of the Second International Conference on Spoken Language Processing (ICSLP'92)*, pages 547–550, Banff, Alberta, Canada, October 1992. ISCA.

[DAR08] Translation technology. In *DARPA: 50 Years of Bridging the Gap*. Faircount Media Group, Tampa, FL, USA, 2008.

[DB96] Peter T. Daniels and William Bright, editors. *The World's Writing Systems*. Oxford University Press, 198 Madison Avenue, New York, NY 10016, USA, 1996.

[DB04a] Marelie Davel and Etienne Barnard. The efficient generation of pronunciation dictionaries: Human factors during bootstrapping. In *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004 — ICSLP)*, pages 2796–2800, Jeju Island, Korea, October 2004. ISCA.

[DB04b] Marelie Davel and Etienne Barnard. The efficient generation of pronunciation dictionaries: Machine learning factors during bootstrapping. In *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004 — ICSLP)*, pages 2781–2784, Jeju Island, Korea, October 2004. ISCA.

[Dem70] A.P. Dempster. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1970.

[DeM96] C. DeMarcken. *Unsupervised Language Acquisition*. PhD thesis, Massachusetts Institute of Technology, September 1996.

[DS94] L. Deng and D. 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.

[Duf53] J. Duffy. *The Effect of Smallpox on the Destiny of the Amerindian.* Louisiana State University Press, Baton Rouge, LA, U.S., 1953.

[Eid01] E. Eide. Distinctive features for use in an automatic speech recognition system. In *Proceedings of the Seventh European Conference on Speech Communication and Technology EUROSPEECH 2001 Scandinavia*, pages 1613–1617, Aalborg, Denmark, September 2001.

[eur05] Eurobarometer 63, 2005.

[FGH$^+$97] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal. The karlsruhe-verbmobil speech recognition engine. In *Proceedings the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 83–86, Munich, Germany, April 1997. IEEE.

[FWK07] Christian Fügen, Alex Waibel, and Muntsin Kolss. Simultaneous translation of lectures and speeches. *Machine Translation*, 21:209–252, 2007.

[Gal79] Susan Gal. *Language Shift: Social Determinants of Linguistic Change in Bilingual Austria.* Academic Press, New York, 1979.

[Gal97] M.J.F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. Technical report, Cambridge University, Engineering Department, May 1997.

[GG05] Raymond G. Gordon, Jr. and Barbara F. Grimes, editors. *Ethnologue: Languages of the World.* SIL International, Dallas, Texas, USA, 2005.

[GN08] Christian Gollan and Hermann Ney. Towards automatic learning in lvcsr: Radip development of a persian broadcast transcription system. In *Proceedings of the 9th Interspeech*, pages 1441–1444, Brisbrane, Australia, September 2008. ISCA.

[Gol01] John Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, June 2001.

[GP06] Marina Gorbis and David Pescovitz. Bursting tech bubbles before they balloon. *IEEE Spectrum*, pages 42–47, September 2006.

[HAH01] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall, Upper Saddle River, New Jersey, 2001.

[Har55] Zellig S. Harris. From phoneme to morpheme. *Language*, 31(2):190–222, April-June 1955.

[HES00] H. Hermansky, D.P.W. Ellis, and S.Sharma. Tandem connectionist feature extraction for conventional hmmsystems. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, volume 3, pages 1635–1638, Istanbul, Turkey, June 2000. IEEE.

[HJLLS07] M. Hasegawa-Johnson, K. Livescu, P. Lal, and K. Saenko. Audiovisual speech recognition with articulator positions as hidden variables. In *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, August 2007.

[HMC$^+$03] A. Hauptmann, N. Moraveji, M-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Baron, W-H. Lin, J. Yang, T.D. Ng, N. Papernick, C.G.M. Snoek, G. Tzanetakis, H. Wactlar, R. Yan, and R. Jin. Informedia at trecvid 2003: Analyzing and searching broadcast news video. In *Proceedings of (VIDEO) TREC 2003 (Twelfth Text Retrieval Conference)*, Gaithersburg, MD,USA, November 2003.

[HUN92] R. Haeb-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proceedings the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 13–16, San Francisco, CA, USA, March 1992. IEEE.

[IUC94] IUCN. Iucn red list categories and criteria version 2.3. Prepared by the IUCN Species Survival Commission, 1994.

[Jan02] Tore Janson. *Speak — A Short History of Languages.* Oxford University Press, Oxford, UK, 2002.

[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.

[Jel90] F. Jelinek. Self-organized language models for speech recognition. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*. Morgan Kaufmann Publishers, Inc., San Mateo, California, 1990.

[Jel97] Fredrick Jelinek, editor. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, Massachusetts, London, England, 1997.

[Joh81] R.E. Johannes. *World of the Lagoon: Fishing and Marine Lore in the Palau district of Micronesia*. University of California Press, Berkley, 1981.

[Kem99] Thomas Kemp. *Ein automatisches Indexierungssystem für Fernsehnachrichtensendungen*. PhD thesis, Universiät Karlsruhe (TH), November 1999.

[KFS00] K. Kirchhoff, G. A. Fink, and G. Sagerer. Conversational speech recognition using acoustic and articulatory input. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, volume 3, pages 1435–1438, Istanbul, Turkey, June 2000. IEEE.

[Kin91] Dale M. Kincade. The decline of native languages in canada. In R.H. Robins and Eugenius M. Uhlenbeck, editors, *Endangered Languages*. Berg Pub Ltd, Oxford and New York, 1991.

[Kir98] K. 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 (ICSLP'98)*, pages 891–894. ISCA, December 1998.

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

[Kir00] K. 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 2000.

[KN95] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *Proceedings the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, Detroit, MI, USA, May 1995. IEEE.

[KN02] S. Kanthak and H. Ney. Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition. In *Proceedings the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, volume 1, pages 845–848, Orlando, Florida, USA, 2002. IEEE.

[KN03] S. Kanthak and H. Ney. Multilingual acoustic modeling using graphems. In *Proceedings of the 8th European Conference on Speech Communication and Technology EUROSPEECH'03*, pages 1145–1148, Geneva, Switzerland, September 2003. ISCA.

[Koe04] Philipp Koehn. Pharaoh: a beam search decoder for phrase-based statistical machine translation models, 2004. http://www.isi.edu/licensed-sw/pharaoh.

[Koe09] Philipp Koehn. *Statistical Machine Transation.* Cambridge University Press, Cambridge, 2009.

[Köh96] Joachim Köhler. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, pages 2195–2198, Philadelphia, PA, USA, October 1996. ISCA.

[Köh98] J. Köhler. Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks. In *Proceedings the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 417–420, Seattle, Washington, USA, May 1998. IEEE.

[Köh99] Joachim Köhler. Comparing three methods to create multilingual phone models for vocabulary independent speech recognition tasks. In *Proceedings of the Workshop on Multi-lingual Interoperability in Speech Technology (MIST)*, pages 79–84, Leusden, Netherlands, September 1999. ISCA.

[Kra92] Michael Krauss. The world's languages in crisis. *Language*, 68(1):4–10, 1992.

[KSS03] Mirijam Killer, Sebastian Stüker, and Tanja Schultz. Grapheme based speech recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology EUROSPEECH'03*, pages 3141–3144, Geneva, Switzerland, September 2003. ISCA.

[KSTY03] Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. Creating corpora for speech-to-speech translation. In *Proceedings of the 8th European Conference on Speech*

Communication and Technology EUROSPEECH'03, pages 381–384, Geneve, Switzerland, September 2003. ISCA.

[KW99] Thomas Kemp and Alex Waibel. Unsupervised training of a speech recognizer. In *Proceedings of the Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99)*, pages 2725–2728, Budapest, Hungary, September 1999. ISCA.

[KWW00] Thomas Kemp, Manfred Weber, and Alex Waibel. End to end evaluation of the isl view4you broadcast news transcription system. In *In the Proceedings of Content-Based Multimedia Information Access 2000, RIAO 2000*, Paris, France, April 2000.

[LCHJ$^+$06] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, and B. Woods. Articulatory feature-based methods for acoustic and audio-visual speech recognition: 2006 jhu summer workshop final report. Technical report, Center for Language and Speech Processing, Johns Hopkins University, February 2006.

[LCHJ$^+$07] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, B. Woods, J. Frankel, M. Magimai-Doss, and K. Saenko. Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 jhu summer workshop. In *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, April 2007. IEEE.

[Lee88] Kai-Fu Lee. *Large-Vocabulary Speaker Independent Continuous Speech Recognition: The SPHINX System.* PhD thesis, Carnegie Mellon University, 1988.

[LG04a] Karen Livescu and James Glass. Feature-based pronunciation modeling for speech recognition. In *Proceedings of Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL 2004)*, pages 81–84, Boston, USA, May 2004. ACL.

[LG04b] Karen Livescu and James Glass. Feature-based pronunciation modeling with trainable asyncrony probabilities. In *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004 — ICSLP)*, pages 667–680, Jeju Island, Korea, October 2004. ISCA.

[LGA$^+$07] L. Lamel, J.-L. Gauvain, G. Adda, C. Barras, E. Bilinski, O.Galibert, A. Pujol, H. Schwenk, and X. Zhu. The limsi 2006 tc-star epps transcription systems. In *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, April 2007.

[LGB03] Karen Livescu, James Glass, and Jeff Blimes. Hidden feature models for speech recognition using dynamic bayesian networks. In *Proceedings of the 8th European Conference on Speech Communication and Technology EUROSPEECH'03*, pages 2529–2532, Geneve, Switzerland, September 2003. ISCA.

[LT00] Spyros Liapis and Georgios Tziritas. Image retrieval by colour and texture using chromaticity histograms and wavelet frames. In *Advances in Visual Information Systems*, volume 1929 of *Springer Lecture Notes in Computer Science*. Springer, Heidelberg, 2000.

[LWL$^+$97] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zhan. Janus iii: Speech-to-speech translation in multiple languages. In *Proceedings the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, April 1997. IEEE.

[McL64a] Marshall McLuhan. *The Gutenberg Galaxy: the making of typographic man.* University of Toronto Press, Toronto, Canada, 1964.

[McL64b] Marshall McLuhan. *Understanding Media: The Extensions of Man.* McGraw-Hill, New York, USA, 1964.

[Met05] Florian Metze. *Articulatory features for conversational speech recognition.* PhD thesis, Universiät Karlsruhe (TH), December 2005.

[Met07] Florian Metze. Discriminative speaker adaptation using articulatory features. *Speech Communication*, 49(5):348–360, May 2007.

[Mim04] Borislava Mimer. Flexible ballungsverfahren für graphembasierte spracherkennung. Studienarbeit, Universität Karlsruhe TH, 2004.

[mos] Moses, a factored phrase-based beam-search decoder for machine translation. http://www.statmt.org/moses.

[MPF99] Giorgio Micca, Enrico Palme, and Alessandra Frasca. Multilingual vocabularies in automatic speech recognition. In *Proceedings* *of the Workshop on Multi-lingual Interoperability in Speech Technology (MIST)*, Leusden, Netherlands, September 1999. ISCA.

[MSS04] Borislava Mimer, Sebastian Stüker, and Tanja Schultz. Flexible decision trees for grapheme based speech recognition. In *Elektronische Sprachsignalverarbeitung Tagungsband der 15. Konferenz*, number 30 in Studientexte zur Sprachkommunikation, pages 79–86, Cottbus, Germany, September 2004. w.e.b. Universitätsverlag.

[MW02] Florian Metze and Alex Waibel. A flexible stream architecture for asr using articulatory features. In *Proceedings the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, volume 1, pages I–709–I–712, Orlando, Florida, USA, 2002. IEEE.

[Nel96] Peter H. Nelde, editor. *Euromosaic: The Production and Reproduction of the Minority Language Groups in the European Union (Education, training, youth)*. European Communities, Luxemburg, 1996.

[Net98] Daniel Nettle. Explaining global patterns of language diversity. *Journal of Anthropological Archaeology*, 17:354–374, 1998.

[NR00] Daniel Nettle and Suzanne Romaine. *Vanishing Voices*. Oxford University Press Inc., New York, NY, USA, 2000.

[Omn] Omniglot writing systems and languages of the world. http://www.omniglot.com.

[ON00] F. J. Och and H. Ney. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447, Hongkong, China, October 2000. Association for Computational Linguistics Morristown, NJ, USA.

[ON03] Franz J. Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March 2003.

[Ost99] M. Ostendorf. Moving beyond the 'beads-on-a-string' model of speech. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 79–84, Keystone, Colorado, USA, December 1999. IEEE.

[Ost01] Nicholas Ostler. What is this technology ever going to do for minority languages. *elsnews - The Newsletter of the European Network in Human Language Technologies*, 10(1):6–7, 2001.

[otEC05] Commission of the European Communities. A new framework strategy for multilingualism. Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, November 2005. COM(2005) 596 final.

[PD94] Ove Andersen Paul Dalsgaard. Application of inter-language phoneme similarities for language identification. In *Proceedings the Third International Conference on Spoken Language Processing (ICSLP 94)*, pages 1903–1906, Yokohama, Japan, September 1994. ISCA.

[Pea95] David F. Peat. *Blackfoot Physics.* Fourth Estate, London, U.K., 1995.

[PFS$^+$05] Matthias Paulik, Christian Fügen, Sebastian Stüker, Tanja Schultz, Thomas Schaaf, and Alex Waibel. Document driven machine translation enhanced asr. In *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH-2005)*, Lisbon, Portugal, September 2005. ISCA.

[PG08] Alex S. Park and James R. Glass. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:186–197, 2008.

[Pov04] Daniel Povey. *Discriminative Training for Large Vocabulary Speech Recognition.* PhD thesis, University of Cambridge, July 2004.

[POY$^+$08] Michael Paul, Hideo Okuma, Hirofumi Yamamoto, Eiichiro Sumita, Shigeki Matsuda, Tohru Shimizu, and Satoshi Nakamura. Multilingual mobile-phone translation services for world travelers. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 165–168, Manchester, U.K, August 2008.

[Pri84] Glanville Price. *The Languages of Britain.* Edward Arnold, London, U.K., 1984.

[PSF$^+$05] Matthias Paulik, Sebastian Stüker, Christian Fügen, Tanja Schultz, Thomas Schaaf, and Alex Waibel. Speech translation enhanced automatic speech recognition. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '05)*, San Juan, Puerto Rico, December 2005. IEEE.

[PSF06] Matthias Paulik, Sebastian Stüker, and Christian Fügen. Speech recognition in human mediated translation scenarios. In *Proceedings of the 2006 IEEE Mediterranean Electrotechnical Conference*, pages 1232–1235, Malaga, Spain, May 2006. IEEE.

[PSF$^+$07] Matthias Paulik, Sebastian Stüker, Christian Fügen, Tanja Schultz, and Alex Waibel. Translating language with technology's help. *IEEE Potentials*, 26(3):30–35, May-June 2007.

[PW08] Matthias Paulik and Alex Waibel. Extracting clues from human interpreter speech for spoken language translation. In *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5097–5100, Las Vegas, NV, USA, April 2008. IEEE.

[Rab89] L.R. Rabiner. A tutorial on hidden markov models and selected applications inspeech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[Ram05] Bhuvana Ramabhadran. Exploiting large quantities of spontaneous speech for unsupervised training of acoustic models. In *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH-2005)*, pages 1617–1620, Lisbon, Portugal, September 2005. ISCA.

[Rog97] Ivica Rogina. *Parameterraumoptimierung Für Diktiersysteme Mit Unbeschränktem Vokabular.* PhD thesis, Universiät Karlsruhe (TH), November 1997.

[RP02] D. K. Roy and A. Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26(1):113–146, January-February 2002.

[RS78] Lawrence Rabiner and Ronald W. Schafer. *Digital Processing of Speech Signals.* Prentice Hall, 1978.

[Run07] Andrew R. Runnalls. A kullback-leibler approach to gaussian mixture reduction. *IEEE Transactions on Aerospace and Electronic Systems*, 43(3):989–999, July 2007.

[SAB$^+$07] Andreas Stolcke, Xavier Anguera, Kofi Boakye, Özgür Cetin, Adam Janin, Mathew Magimai-Doss, Chuck Wooters, and Jing Zheng. The sri-icsi spring 2007 meeting and lecture recognition system. In *Proceedings of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop (RT-07)*, Baltimore, MD, USA, May 2007.

[SBB$^+$07] Tanja Schultz, Alan W. Black, Sameer Badaskar, Matthew Hornyak, and John Kominek. Spice: Web-based tools for rapid language adaptation in speech processing systems. In *Proceedings of the 10th European Conference on Speech Communication and Technology (INTERSPEECH 2007)*, pages 2125–2128, Antwerp, Belgium, August 2007.

[SCB$^+$05] S. Suebvisai, P. Charoenpornsawat, A.W. Black, M. Woszczyna, and T. Schultz. Thai automatic speech recognition. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, Philadelphia, PA, USA, March 2005. IEEE.

[Sch85] Annette Schmidt. *Young People's Dyirbal: A Case of Language Death from Australia.* Cambridge University Press, Cambridge, 1985.

[Sch00a] Ralf Schlüter. *Investigations on Discriminative Training Criteria.* PhD thesis, Rheinisch Westflische Technische Hochschule Aachen, September 2000.

[Sch00b] Tanja Schultz. *Multilinguale Spracherkennung — Kombination akustischer Modelle zur Portierung auf neue Sprachen.* PhD thesis, Universiät Karlsruhe (TH), July 2000.

[Sch02] Tanja Schultz. Globalphone: A multilingual speech and text database developed at karlsruhe university. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 345–348, Denver, CO, USA, September 2002. ISCA.

[SFK00] Christoph Schillo, Gernot A. Fink, and Franz Kummert. Grapheme based speech recognition for large vocabularies. In *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000)*, pages 584–587, Beijing, China, October 2000. ISCA.

[SFKW07] Sebastian Stüker, Christian Fügen, Florian Kraft, and Matthias Wölfel. The isl 2007 english speech transcription system for european parliament speeches. In *Proceedings of the 10th European Conference on Speech Communication and Technology (INTERSPEECH 2007)*, pages 2609–2612, Antwerp, Belgium, August 2007.

[SHBS09] Yun-Hsuan Sung, Thad Hughes, Francoise Beaufays, and Brian Strope. Revisiting graphemes with increasing amounts of data. In *Proceedings of the 2008 IEEE International Conference on*

Acoustics, Speech, and Signal Processing, Taipei, Taiwan, April 2009. IEEE.

[SMFW01] H. Soltau, F. Metze, C. Fügen, and A. Waibel. A one pass-decoder based on polymorphic linguistic context assignment. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '01)*, pages 214–217, Madonna di Campiglio Trento, Italy, December 2001.

[SMSW03] S. Stüker, F. Metze, T. Schultz, and A. Waibel. Integrating multilingual articulatory features into speech recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology EUROSPEECH'03*, pages 1033–1036, Geneve, Switzerland, September 2003. ISCA.

[SMT04] A. Black S. Maskey and L. Mayfield Tomokiyo. Boostrapping phonetic lexicons for new languages. In *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004 — ICSLP)*, pages 69–72, Jeju Island, Korea, October 2004. ISCA.

[SS45] E.W. Stearn and A.E. Stearn. *The effect of smallpox on the destiny of the Amerindian.* Humphries, Boston, 1945.

[SS04] Sebastian Stüker and Tanja Schultz. A grapheme based speech recognition system for russian. In *Proceedings of the 9th International Conference "Speech And Computer" SPECOM'2004*, pages 297–303, Saint-Petersburg, Russia, September 2004. Anatolya.

[SSMW03] S. Stüker, T. Schultz, F. Metze, and A. Waibel. Multilingual articulatory features. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, volume 1, pages 144–147, Hong Kong, April 2003. IEEE.

[Ste06] Volker Steinbiss. Human language technologies for europe. Work comissioned by ITC-irst, Trento, Italy to Accipio Consulting, Aachen, Germany, April 2006.

[STNE+93] E. G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. Automatic speech recognition without phonemes. In *Proceedings of the 3rd European Conference on Speech Communication and Technology EUROSPEECH'93*, pages 129–132, Berlin, Germany, September 1993. ISCA.

[Sto02] A. Stolcke. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of the 7th International Conference on Spoken*

Language Processing (ICSLP 2002), pages 901–904, Denver, CO, USA, 2002. ISCA.

[Stü04] Sebastian Stüker. Multilingual articulatory features. Diplomarbeit, Universität Karlsruhe *TH*, 2004.

[Stü08a] Sebastian Stüker. Integrating thai grapheme based acoustic models into the ml-mix framework - for language independent and cross-language asr. In *Proceedings of the First International Workshop on Spoken Languages Technologies for Under-resourced languages (SLTU)*, Hanoi, Vietnam, May 2008.

[Stü08b] Sebastian Stüker. Modified polyphone decision tree specialization for porting multilingual grapheme based asr systems to new languages. In *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4249–4252, Las Vegas, NV, USA, April 2008. IEEE.

[Sut03] William J. Sutherland. Parallel extinction risk and global distribution of languages and species. *Nature*, 423(15):276–279, May 2003.

[SW98a] Tanja Schultz and Alex Waibel. Development of multilingual acoustic models in the globalphone project. In *Proceedings of the First Workshop on Text, Speech, Dialogue — TSD'98*, pages 311–316. Masaryk University Press, September 1998.

[SW98b] Tanja Schultz and Alex Waibel. Language independent and language adaptive large vocabulary speech recognition. In *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP'98)*. ISCA, December 1998.

[SW98c] Tanja Schultz and Alex Waibel. Multilingual and crosslingual speech recognition. In *Proceedings of the DARPA Workshop on Broadcast News Transcription and Understanding*. DARPA, 1998.

[SW00] T. Schultz and A. Waibel. Polyphone decision tree specialization for language adaptation. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, volume 3, pages 1707–1710, Istanbul, Turkey, June 2000. IEEE.

[SW01] T. Schultz and A. Waibel. Language independent and language adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(1-2):31–51, August 2001.

[SZR$^+$06] Sebastian Stüker, Chengqing Zong, Jürgen Reichert, Wenjie Cao, Muntsin Kolss, Guodong Xie, Kay Peterson, Peng Ding, Victoria Arranz, Jian Yu, and Alex Waibel. Speech-to-speech translation services for the olympic games 2008. In *Proceedings of Machine Learning for Multimodal Interaction - Third International Workshop, MLMI 2006, in Lecture Notes in Computer Science, Vol. 4299, 2006, XII*, pages 17–20, Bethesda, MD, USA, May 2006. Springer.

[USN98] Ulla Uebler, Michael Schssler, and Heinrich Niemann. Bilingual and dialectal adaptation and retraining. In *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP'98)*. ISCA, December 1998.

[van79] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, second edition, 1979.

[VNT96] Stephan Vogel, Hermann Ney, and Christopher Tillmann. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, volume 2, pages 836–841, Copenhagen, Denmark, August 1996. Association for Computational Linguistics Morristown, NJ, USA.

[WBNS97] Fuliang Weng, Harry Bratt, Leonardo Neumeyer, and Andreas Stolcke. A study of multilingual speech recognition. In *Proceedings the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, pages 359–362, Rhodes, Greece, September 1997. ISCA.

[WF08] Alex Waibel and Christian Fügen. Spoken language translation. *IEEE Signal Processing Magazine*, 70, May 2008.

[WGC01] M. Wester, S. Greenberg, and S. 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 EUROSPEECH 2001 Scandinavia*, pages 1729–1732, Aalborg, Denmark, September 2001.

[Wika] Wikimedia commons — the writing systems of the world. http://commons.wikimedia.org/wiki/File:WritingSystemsoftheWorld4.png.

[Wikb] Wikipedia — the free encyclopedia. http://www.wikipedia.org.

[Wikc] Wikipedia — the free encyclopedia, ethnologue list of most-spoken languages. http://www.wikipedia.org/Ethnologue_list_of_most-spoken_languages.

[Wikd] Wikipedia — the free encyclopedia, globalization. http://en.wikipedia.org/wiki/Globalization.

[WKAM94] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy. An evaluation of cross-language adaptation for rapid hmmdevelopment in a new language. In *Proceedings the 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 237–240, Adelaide, SA, Australia, April 1994. IEEE.

[Wol99] Roald Wolff. Adaption von kontextenscheidungsbäumen auf neue sprachen. Studienarbeit, Universität Karlsruhe TH, 1999.

[WPG96] P. C. Woodland, D. Pye, and M. J. F. Gales. Iterative unsupervised adaptation using maximum likelihood linear regression. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, pages 1133–1136, Philadelphia, PA, USA, October 1996. ISCA.

[WRN$^+$98] Todd Ward, Salim Roukos, Chalapathy Neti, Jerome Gros, Mark Epstein, and Satya Dharanipragada. Towards speech understanding across multiple languages. In *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP'98)*. ISCA, December 1998.

[WS03] Zhirong Wang and Tanja Schultz. Non-native spontaneous speech recognition through polyphone decision tree specialization. In *Proceedings of the 8th European Conference on Speech Communication and Technology EUROSPEECH'03*, pages 1449–1452, Geneva, Switzerland, September 2003. ISCA.

[WSK07] Matthias Wölfel, Sebastian Stüker, and Florian Kraft. The isl rt-07 speech-to-text system. In *Proceedings of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop (RT-07)*, Baltimore, MD, USA, May 2007.

[Wur98] Stephen A. Wurm. Methods of language maintenance and revival with selected cases of language endagerment in the world. In Kazuto Matsumura, editor, *Studies in Endangered Languages*, pages 191–211. Hituzi Syobo, 1998.

[YS03] H. Yu and T. Schultz. Enhanced tree clustering with single pronunciation dictionary for conversational speech recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology EUROSPEECH'03*, pages 1869–1872, Geneva, Switzerland, September 2003. ISCA.

[Yu02] Peter K. Yu. Bridging the digital divide: Equality in the information age. *Cardozo Arts & Entertainment Law Journal*, 20(1):1–52, 2002.

[YW93] S.J. Young and P.C. Woodland. The use of state tying in continous speech recognition. In *Proceedings of the 3rd European Conference on Speech Communication and Technology EUROSPEECH'93*, pages 2203–2206, Berlin, Germany, September 1993. ISCA.

[ŽK06] Andrej Žgank and Zdravko Kačič. Conversion from phoneme based to grapheme based acoustic models for speech recognition. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006, ICSLP)*, pages 1587–1590, Pittsburgh, PA, USA, September 2006. ISCA.

[ŽKD$^+$05] Andrej Žgank, Zdravko Kačič, Frank Diehl, Jozef Juhar, Slavomir Lihan, Klara Vicsi, and Gyorgy Szaszak. Graphemes as basic units for crosslingual speech recognition. In *Proceedings of the COST278 Final Workshop and ITRW on Applied Spoken Language Interaction in Distributed Environments (ASIDE 2005)*, Aalborg, Denmark, November 2005. ISCA.

[ZSCB98] George Zavaliagkos, Manhung Situ, Thomas Colthurst, and Jayadev Billa. Using untranscribed training data to improve performance. In *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP'98)*, pages 891–894. ISCA, December 1998.

[ZW97] P. Zhan and M. Westphal. Speaker normalization based on frequency warping. In *Proceedings the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1039–1042, Munich, Germany, April 1997. IEEE.