Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition

Tanja Schultz and Alex Waibel
{*tanja@ira.uka.de*}

Interactive Systems Laboratories
University of Karlsruhe (Germany), Carnegie Mellon University (USA)

ABSTRACT

With the distribution of speech technology products all over the world, the portability to new target languages becomes a practical concern. As a consequence our research focuses on the question of how to port LVCSR systems in a fast and efficient way. More specifically we want to estimate acoustic models for a new target language using speech data from varied source languages, but only limited data from the target language. For this purpose we introduce different methods for multilingual acoustic model combination and a polyphone decision tree specialization procedure. Recognition results using language dependent, independent and language adaptive acoustic models are presented and discussed in the framework of our GlobalPhone project which investigates LVCSR systems in 15 languages.

Mit der weltweiten Verbreitung von Sprachtechnologieprodukten wird die schnelle und effiziente Portierung vorhandener Spracherkennungssysteme auf neue Sprachen zu einer Angelegenheit von direkt anwendbarem Nutzen. Aus diesem Grund konzentriert sich unsere Forschung auf die Frage, wie sich ein Spracherkennungssystem, genaugenommen die akustischen Modelle, unter Ausnutzung vorhandener Daten anderer Sprachen in einer neuen Sprache effizient entwickeln lassen. Zu diesem Zweck führen wir unterschiedliche Methoden zur Kombination multilingualer akustischer Modelle ein und definieren die Polyphone Decision Tree Specialization Methode. Es werden zahlreiche Erkennungsexperimente anhand sprachenabhängiger, sprachenunabhängiger und sprachenadaptiver akustischer Modellen vorgestellt und im Rahmen des GlobalPhone Projektes evaluiert. GlobalPhone ist ein Projekt, in dem LVCSR Spracherkennung in 15 verschiedenen Sprachen untersucht wird.

1. Introduction

The state of the art in large vocabulary continuous speech recognition (LVCSR) has advanced substantially for quite a number of languages. Recognition systems developed originally for one language have been successfully ported to several languages, including systems developed by IBM (Cohen et al., 1997), Dragon (Barnett et al., 1996), BBN (Billa et al., 1997), Cambridge (Young et al., 1997), Philips (Dugast et al., 1995), MIT (Glass et al., 1995), and LIMSI (Lamel et al., 1995). The transformation of English systems to such diverse languages like German, Japanese, French, and Mandarin Chinese illustrates that speech technology generalizes across languages and that similar modeling assumptions hold for various languages.

To date, however, extensions have only been performed with well known languages for which large amounts of data are available. To build a recognizer, this data usually includes dozens of hours of recorded and transcribed speech. Unfortunately the assumption that large speech databases can be provided on demand does not hold for several reasons. Firstly, the collection of large databases requires a tremendous amount of time and resources. Secondly, more than 4000 languages exist in the world and about 10% are spoken by at least 100.000 native speakers and therefore might be of potential interest. Which of these languages are of interest for speech recognition applications can change very quickly with the political and economic situation. Finally, in some research areas like non-native speech recognition it is even not possible, for combinatorial reasons, to collect large databases.

As a consequence, our research has focused on the question of how to build a LVCSR system for a new target language using speech data from varied source languages, but only limited data from the target language. For that purpose we first develop monolingual recognition engines on the basis of our recently collected GlobalPhone database in 15 languages. The term *monolingual recognizer* refers to a system which is designed to recognize speech from one language. The goal of creating monolingual recognizers in multiple languages is twofold: We want to investigate differences between languages and highlight resulting challenges for speech recognition in multiple (even less familiar) languages, and we explore systems in diverse languages as a starting point for our main focus, namely the adaptation to new target languages.

To achieve this goal we investigate multilingual LVCSR systems, i.e systems capable of *simultaneously recognizing languages* which have been presented during the training procedure. Particularly we define a global unit set which is suitable to cover 12 languages. Based on this global unit set we evolve and evaluate different techniques to combine the acoustic models of varied languages and call the resulting multilingual acoustic models *language independent*. These language independent acoustic models allow the data and model sharing of various languages to reduce the complexity and number of parameters of a multilingual LVCSR system. Furthermore, these models will be used as seed models for a new target language.

The statistical methods applied to speech and language modeling not only require hours of recorded and transcribed speech, but also pronunciation dictionaries and large text corpora. In our present research we focus mainly on acoustic modeling problems and assume that other resources are given in the target language. This is a reasonable assumption in the read newspaper domain since acquiring the training data for acoustic models is usually the most expensive part of a data collection. Large corpora as well as dictionaries in many languages are distributed by several data consortia. For dictionaries this is actually true for 11 West-European languages, provided by ELRA in 1998 (ELRA, 2000) and in another 6 widespread languages provided by the LDC (LDC, 2000). However, we are aware of the fact that appropriate large text material are, to date, only available in hundreds of languages and pronunciation dictionaries in some tens of the most spread and studied languages. In many languages only little or no written material is available nor in spontaneous spoken domain applications. Therefore, we want to stress here that we address only one aspect of language independent speech recognition, namely the *language independent acoustic modeling* issue.

As mentioned above, the goal of language independent modeling is the acoustic model combination suitable for a simultaneously recognition of all involved source languages. In contrast the goal of *language adaptive* modeling is the adaptation of preexisting models towards an optimal recognition of a new target language, using only limited adaptation data from this target language. Given the data limitation we face two problems: one is to determine suitable seed models for the initialization of acoustic models in the target language and the second problem is the large phonetic mismatch between varied source languages and the target language when extending the phonetic context window for building context dependent acoustic models. Phoneme model of arbitrary context width are called *polyphones*. The use of large phonetic context windows has proven to increase the recognition performance significantly in the monolingual setting. Therefore, it seems natural to extend this idea to the multilingual setting as well. We approach the first problem by using language independent models as seed models. In order to solve the second problem we introduce a procedure of adapting multilingual polyphone decision trees to a target language with very limited adaptation data. In summary we present techniques which enable us to set up a LVCSR recognition engine in a new target language by borrowing speech data from varied source languages but only limited data from the target language itself.

2. The GlobalPhone project

GlobalPhone is a project undertaken at the Interactive Systems Labs which investigates LVCSR in several languages. One goal of this project is the combination of monolingual recognizers into one multilingual engine, which can handle several languages at a time. This concept requires a multilingual database suitable for LVCSR and a combined acoustic model that represents the sounds of all languages involved. In this section we present the multilingual GlobalPhone database and the global unit set which we developed in the framework of this project.

2.1. The GlobalPhone database

This database currently consists of read speech data for the languages Arabic, Chinese (Mandarin and Shanghai dialects), Croatian, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. Along with the English Wall Street Journal (WSJ0, distributed by LDC and French BREF (BREF-Polyglot sub-corpus, distributed by ELRA) databases, this covers 9 of the 12 most widespread languages of the world (a language rank classification can be found for example in Webster's (Webster, 1992)). In each of the languages about 15-20 hours of high quality speech was collected, spoken by 100 native speakers per language. Each speaker read several articles about political and economical topics chosen from national newspapers. All the newspapers are accessible via Internet, so that large text corpora for language modeling can be easily downloaded. Further details about the GlobalPhone project are given in (Schultz et al., 1997).

Table 1: The GlobalPhone database

Language	Abbr	Utts	Spks	Units	Hours
Ch-Mandarin	CH	10181	132	262K	31.2
Ch-Shanghai	WU	2644	41	79K	9.5
Croatian	KR	4499	92	120K	15.9
English (WSJ)	EN	7434	103	129K	15.9
French (Bref)	FR	7516	80	123K	14.7
German	GE	10085	77	132K	18.3
Japanese	JA	13067	144	268K	33.9
Korean	KO	8107	100	417K	21.0
Portuguese	PO	10220	101	208K	26.0
Russian	RU	11111	106	170K	22.2
Spanish	SP	6898	100	171K	22.1
Swedish	SW	11816	98	184K	21.7
Turkish	TU	6950	100	112K	22.2
Total		110528	1364	2083K	269.7

Table 1 gives the numbers of the GlobalPhone database. While the total sum of 270 hours spoken speech is very high, the available data per language is small compared to monolingual databases usually used for the training of a LVCSR system. Throughout the experiments which will be described in the following we investigate ten of the reported languages with 80% of all speakers per language for training, 10% were used as a development set, and the remaining 10% for a test set.

2.2. Global Unit Set

Our research in language independent and adaptive LVCSR is based on the assumption that the articulatory representations of phonemes are so similar across languages that phonemes can be considered as units which are independent from the underlying language. Based on this assumption the language specific phoneme inventories of N languages can be unified into one global set $\Upsilon = \Upsilon_{L_1} \cup \Upsilon_{L_2} \cup \ldots \cup \Upsilon_{L_N}$. This idea was first proposed by the International Phonetic Association (IPA, 1993) then transfered to automatic speech recognition by Andersen and Dalsgaard (Andersen et al., 1993) and successfully applied to language identification (Andersen and Dalsgaard, 1997; Corredor-Ardoy et al., 1997). According to this idea we differentiate between the group of language independent *polyphonemes*[1] Υ_{LI}, containing phonemes occurring in more than one language, and N remaining groups of language dependent *monophonemes* $\Upsilon_{LD_{L_1}}, \ldots \Upsilon_{LD_{L_N}}$. The set $\Upsilon_{LD_{L_m}}$ contains phonemes only occurring in language L_m, thus $|\Upsilon_{LD_{L_m}}| = 0$, if each phoneme of language L_m has a counterpart in at least one of the remaining $N-1$ languages.

Similarities of sounds are documented in international phonetic inventories like Sampa (Wells, 1989), Worldbet (Hieronymus, 1993), or IPA (IPA, 1993), which classify sounds based on phonetic knowledge. In our research we define a global unit set for 12 languages based on the IPA scheme. Sounds of different languages, which are represented by the same IPA symbol, share one common unit, so-called *IPA-unit*, in this global unit set. Regarding Chinese sounds we abstain from handling tones separately, i.e. the 5 tonal variations of a Mandarin vowel are treated as one vowel. Table 2 summarizes the polyphonemes and monophonemes for all 12 languages. For each polyphoneme the upper half of Table 2 reports the number of languages which share one phoneme. The lower half of Table 2 contains the number and type of monophonemes for each language.

Table 2: Global Unit Set for 12 languages

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

2.3. Unit sharing across languages

We define the *share factor* $\mathrm{sf_N}$ for a set of N languages as the relation between the sum of language specific phonemes and the size of the global unit set, i.e. $\mathrm{sf_N}$ gives the average number of languages sharing the phonemes of the global unit set:

$$\mathrm{sf_N} = \frac{\sum_{i=1}^{N} |\Upsilon_{\mathrm{L_i}}|}{|\Upsilon|}, \quad |\Upsilon| = |\Upsilon_{\mathrm{LI}}| + \sum_{i=1}^{N} |\Upsilon_{\mathrm{LD_{L_i}}}| \quad (1)$$

The share factor is one, if no polyphonemes exist at all and N, if each of the N languages uses the identical phonetic inventory, i.e. $1 \leq \mathrm{sf_N} \leq \mathrm{N}$.

In our case we have 485 language specific phonemes for 12 languages which are applied for the best monolingual systems reported in the next section. According to Table 2 this results in:

[1] polyphonemes should not be confused with polyphones

$$\mathrm{sf}_{12} = \frac{|\Upsilon_{\mathrm{ch}}| + |\Upsilon_{\mathrm{kr}}| + \ldots + |\Upsilon_{\mathrm{tu}}|}{|\Upsilon|} = \frac{485}{162} = 2.99 \tag{2}$$

which implies that, on average, each phoneme of our global unit set is shared by 3 languages, this is a sharing rate of 25% given 12 languages. We also calculate the average share factor over all possible k-tuples ($k = 1, \cdots, 12$) of 12 languages $\binom{12}{k}$ and plot the result in Figure 1. We find two main points: Firstly, the share factor increases with numbers of involved languages, but the increasing rate is much lower than expected. One reason might be the diversity of the languages. Secondly, the range of the share factor strongly depends on the involved languages, implying that the phoneme inventories of some languages are quite similar while others are not.

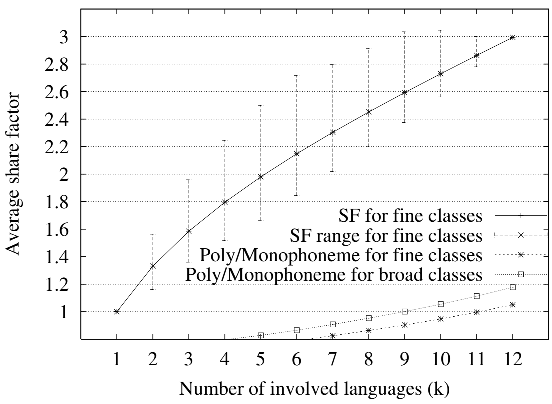


Figure 1: Average and range of the share factor for fine phoneme classes and polyphoneme-to-monophoneme ratio for fine and broad phoneme classes depending on $\binom{12}{k}$ different language groups with $k = 1, \ldots, 12$

Additionally, we calculate the average relation between polyphonemes and monophonemes and plot this in Figure 1 as well. We can see that, on average, the polyphonemes outnumber the monophonemes after 11 languages are included. This might be an indication that we have chosen too fine a partition of the phoneme set. Therefore, we experiment with broader unit classes, which reduce the global unit set up to 40% but as Figure 1 shows, the relation between polyphonemes and monophonemes is not affected significantly, i.e. the ratio of models which will share data of different languages can not be increased by broader partitions of the phoneme set. Since broader classes are contra-productive in terms of monolingual and multilingual speech recognition performance, we prefer fine unit classes instead.

3. Language dependent LVCSR

Based on the GlobalPhone database, we investigate monolingual large vocabulary continuous speech recognition systems in ten languages. For this purpose we use the same speech technology and even the same system architecture, preprocessing, and parameter size across all languages. During development we found a tremendous variation in language specificities.

3.1. Language differences

When comparing the word error rates of the resulting monolingual systems the language specificities have to be taken into account. Therefore, we will first discuss differences between languages and highlight the resulting challenges for speech recognition.

Scripts Many different character types are used in the world's languages. Writing systems fall into two major categories: ideographic and phonologic. In the ideographic scripts, the characters reflect the meaning rather than the pronunciation of a word. Examples for ideographic scripts are the Chinese *Hanzi* and the Japanese *Kanji*. Phonological scripts can be further divided into syllable-based scripts, like Japanese *Kana* or Korean *Hangul*, and alphabetic scripts which are used for the most Indo-European languages, such as Cyrillic script for Russian, or Latin script for English and German.

Letter-to-sound relation Phonologic scripts are easier to handle than ideographic scripts in the speech recognition framework, as in many cases rule-based letter-to-sound mapping tools can be used to generate the pronunciation dictionary needed to guide recognition, while this is usually not possible for ideographic scripts. However, among the languages using alphabetic scripts, the letter-to-sound relation varies considerably. It ranges from nearly one-to-one relation such as for Turkish and for Slavic languages like Russian and Croatian up to languages such as English that requires complex rules and has many exceptions. For the languages using phonologic scripts we implemented letter-to-sound tools as described for Turkish (Çarkı et al., 2000) or Korean (Kiecza et al., 1999); for languages with ideographic script we first built character conversion tools and derived the pronunciation in a second step from the converted strings like for Japanese or Chinese (Reichert et al., 1999).

Sound system Across the world's languages, the sound inventory varies considerably. The size of the phoneme inventory used for speech recognition in GlobalPhone ranges from 29 phonemes (Turkish) to 46 phonemes (Portuguese). The ratio between consonants and vowels in the inventory varies from 4:1 in case of the Croatian language versus 0.8:1 for Portuguese. In spoken speech German is the language with the highest consonants-to-vowel ratio (60%), Portuguese the one with the lowest (50%).

To give a reliable measure of the acoustic difficulties of the languages, we calculated the phoneme-based recognition rate using a phoneme recognizer without any (phoneme) language model constraints. The results indicate significant differences in acoustic confusability between languages, ranging from 33.8% to 46.4% phoneme error rate. The phoneme error rate of a language correlates with the number of phonemes used to model this language as illustrated in Figure 2. Turkish seems to be an exception to this finding. The error analysis showed that this is due to a very high substitution rate between the closed front vowels [e], [i], and [y].

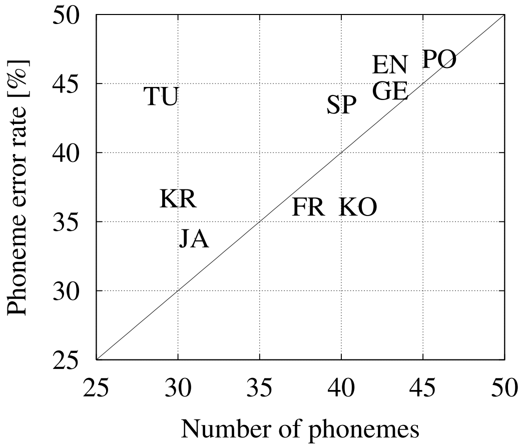


Figure 2: Relation between phoneme error rate and number of modeled phonemes

Many languages belong to the group of tonal languages, in which lexical items are distinguished by contrasts in pitch contour or pitch level on a single syllable, like Mandarin Chinese which differentiate between five tones. Modeling tones separately for Chinese speech recognition increases the phonemic inventory from 48 to 137 phonemes. In pitch languages like Japanese, pitch contrasts are not drawn between syllables but between polysyllabic words. In stress languages individual syllables are stressed. In fixed stress languages like Turkish stress pattern always occur in the same position within a word (Turkish has in general word final stress). Fixed stress languages are easier to model than lexical stress languages like English and German, where the stress position varies across words.

Segmentation Another issue is the segmentation of character strings into natural units. English or Spanish are languages which provides us with a natural segmentation into words which can conveniently be used as dictionary units for speech recognition. The words are long enough to differ from each other in a sufficient number of phonemes, but short enough to be able to cover most material with a reasonable number of different word forms that occur frequently. This is important for the statistical analysis required by the automatic learning processes that modern speech recognition systems rely on. But other languages lack an adequate segmentation. In Japanese and Chinese whole sentences are written in strings of characters without any spacing between adjacent words. In order to determine appropriate dictionary units, the strings of characters have to be segmented manually or by morphological analysis. Details about how we proceed with the segmentation of languages can be found in (Çarkı et al., 2000; Reichert et al., 1999; Kiecza et al., 1999).

Morphology Natural segmentation is one factor which influences the length of a word unit, the other one is the morphology. Languages like German build long word phrases by compounding nouns. Another group of languages, including Korean and Turkish, has a morphologic structure which provides for agglutination and suffixing. The inflection, derivation, and other relationships between words are expressed by constantly concatenating suffixes to the word stem. All these effects result in rapid growth of the number of word forms occurring in a text. As a consequence, poor recognition results are achieved when using a certain set of word forms as dictionary units for speech recognition, and many new word forms are encountered in unseen speech, giving a high Out-Of-Vocabulary (OOV) rate for these languages.

Table 3 gives the size of vocabulary and resulting OOV-rates for ten languages. The OOV-rates differ significantly between these languages. For English we observed the lowest OOV-rates of 0.3% with 64K as well known from the literature. For Korean and Chinese OOV-rates down to 0% are achieved with a 64K vocabulary due to the applied segmentation. Whereas for Turkish we found 13.5% and up to 34% in the case of not segmented Korean word forms.

Table 3: OOV-rates for ten GlobalPhone languages

Language	Vocabulary	OOV-Rate
English	64K	0.3%
Korean	64K	34.0%
Korean (segmented)	64K	0.2%
Turkish	64K	13.5%
German	61K	4.4%
Chinese (segmented)	60K	0%
Portuguese	60K	4.3%
French	30K	4.7%
Croatian	31K	13.6%
Spanish	30K	5.2%
Japanese (segmented)	22K	3.0%

3.2. LVCSR systems in 10 languages

We developed monolingual large vocabulary continuous speech recognition systems in ten languages using our Janus Recognition Toolkit (JRTk) (Finke et al., 1997). Building speech recognition engines for so many languages is associated with considerable effort. Therefore, we tried to optimize the development procedure by automatization. The pronunciation dictionaries were generated by the above mentioned letter-to-sound mapping tools and the language models were calculated based on fully automatically downloaded text resources from the Internet. For the initialization of the acoustic models we applied our fast and efficient bootstrapping algorithm using a language dependent four-lingual phoneme pool (Schultz and Waibel, 1997).

For each language, the acoustic model consists of a fully continuous HMM system with 3000 sub-triphone and sub-quinphone models respectively. The term *sub-polyphone* here refers to a polyphone which is divided into a begin, middle and end state. A mixture of 32 Gaussian components is assigned to each state. The Gaussians are on 13 Mel-scale cepstral coefficients with first and second order derivatives, power, and zero crossing rate. After cepstral mean subtraction a linear discriminant analysis reduces the input vector to 32 dimensions.

The sub-polyphone models are created by applying a decision tree clustering procedure which uses an entropy-based distance measure, defined over the mixture weights of the Gaussians, and a question set which consists of linguistically motivated questions about the phonetic context of a phoneme model (Finke and Rogina, 1997). In each step of clustering the question giving the highest entropy gain is selected when splitting the tree node. The splitting procedure is stopped after reaching the predefined number of 3000 sub-polyphone models.

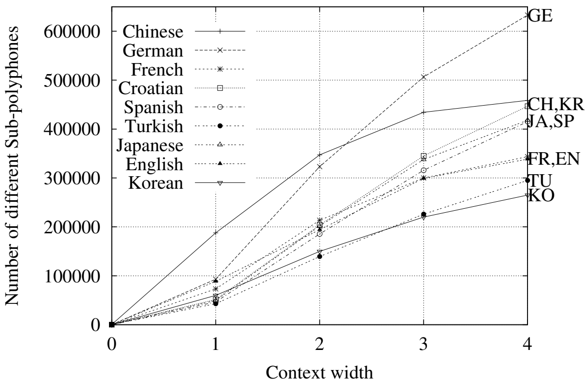


Figure 3: Number of sub-polyphones for different context width

The number of observed polyphones varies between the languages due to differences in the phonotactic structure. Figure 3 shows the number of different polyphones for context width ±1 (triphones), ±2 (quinphones), ±3 (septphones) up to a context window of 4 phonemes to the left and to the right. Due to implementation reasons the number of polyphones is not only affected by the phonotactics but also by the length of the dictionary units, since in JRTk the context window is not extended to more than one phoneme into the neighbor words.

As Figure 3 illustrates German has by far the most polyphones. This can be explained by less restricted phonotactics which also allow consonant clusters. Korean and Turkish have the lowest number of polyphones, the latter might be due to the vowel harmony of the Turkish language. The behavior of Chinese polyphones is a result of the short length of dictionary units after segmentation. For comparison reasons and for further experiments the recognition engines in all languages have the same model size. As a consequence German clustered models represent a greater variety of polyphone types than in other languages.

Figure 4 shows the resulting word error and phoneme error rates for language dependent LVCSR systems in ten languages. As a consequence of segmentation, not in all languages word error rates can be presented. Chinese and Korean are given in character-based error rate, Japanese in hiragana-based error rate. For the remaining languages the error rates are reported based on the natural segmentation. For the reported results we control the OOV-rate by including the test words in the decoders vocabulary list and adding small language model probabilities. Overall the word error rates range between 10% and 20%, the phoneme error rates range between 33.8% and 46.4% as already shown in Figure 2. Comparing these numbers to other LVCSR engines reported elsewhere it should be taken into account that we used only a very limited size of training data per language as can be seen from Table 1 and that the recognizers are built based on automatically generated knowledge sources.

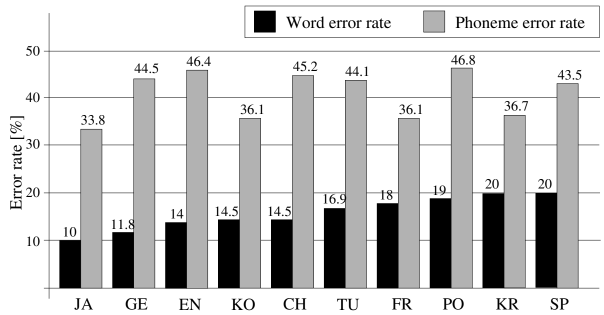


Figure 4: Phoneme and word error rates for LVCSR systems in ten languages

Since the core engines are the same across all languages, performance differences can be explained by the discussed language-specific inherent challenges like letter-to-sound relation, segmentation or word length, phonotactics, and morphology. For the first time it could be shown for a high number of languages that speech technology generalizes across such divers types of languages.

4. Language independent acoustic modeling

Based on the described global unit set and the created monolingual systems we investigate different methods to combine the acoustic models of varied languages to one multilingual acoustic model. The main goals of the model combination are the reduction of the overall amount of acoustic model parameters and the improvement of the model robustness for language adaptation purposes.

4.1. Acoustic model combination

We introduce three different methods for acoustic model combination, the language separate *ML-sep*, the language mixed *ML-mix*, and the language tagged *ML-tag* combination method. Their performance is evaluated in a five-lingual setup for the languages Croatian, Japanese, Korean, Spanish, and Turkish. The evaluated systems applied the same preprocessing and acoustic modeling as the aforementioned monolingual systems, in particular the probability $p(x|s_i)$ to emit x in state s_i is described by a mixture of K_i Gaussian components: $p(x|s_i) = \sum_{k=1}^{K_i} c_{s_i k} N(x|\mu_{s_i k}, \Sigma_{s_i k})$. Figure 5, 6 and 7 illustrate the three different acoustic model combination methods. In these figures the mixture weights c are symbolized as distributions and the Gaussian components $N(x|\mu, \Sigma)$ are symbolized as rounded boxes.

In the *ML-sep* combination method each language-specific phoneme is trained solely with data from its own language, i.e. no data are shared across languages to train the acoustic models. The multilingual component of *ML-sep* is the feature extraction, since one global LDA-matrix is calculated taking all language-specific phoneme models as LDA classes. Context dependent models are created by applying the described entropy-based decision tree clustering procedure. Provided the above mentioned modeling of emission probabilities, the *ML-sep* combination method can be described as:

$$\text{ML-SEP}: \begin{cases} c_{s_i} \neq c_{s_j} & , \quad \forall i \neq j \\ \mu_{s_i,k} \neq \mu_{s_j,k} , & \forall i \neq j \\ \Sigma_{s_i,k} \neq \Sigma_{s_j,k}, & \forall i \neq j \end{cases}$$

A schematic of the separate acoustic modeling method is shown in Figure 5 for the beginning state of phoneme "M".

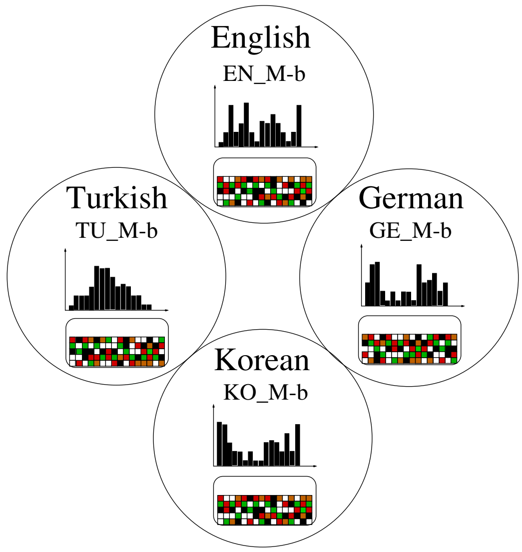


Figure 5: **ML-sep**: Separate acoustic modeling

In the *ML-mix* combination method we share data across different languages to train the acoustic models of poly-phonemes, i.e. phonemes of different languages which belong to the same IPA-unit defined in our global unit set (see Subsection 2.2). During training we do not preserve any information about the language. In other words, for each IPA-unit of the global unit set we initialize one mixture of 16 Gaussian components per state and train the model of this IPA-unit by sharing the data of all languages belonging to the IPA-unit. In the five-lingual case the sharing factor is $sf_5 = 2.1$ which means that, on average, each model is trained with data of two different languages. The context dependent models are created by applying the aforementioned clustering procedure. Since we do not have any language identities, the linguistically motivated questions of the question set are derived from the IPA-reference scheme. The splitting procedure is stopped after reaching a predefined number of 3000 language independent sub-quinphone models, which results in system *ML-mix3000*.

With the function ipa(s_i) which returns the IPA-unit to which s_i belongs, we can describe the *ML-mix* model combination method as:

$$\text{ML-MIX}: \begin{cases} c_{s_i} = c_{s_j} & , \quad \forall i,j : \text{ipa}(s_i) = \text{ipa}(s_j) \\ \mu_{s_i,k} = \mu_{s_j,k} \ , & \quad \forall i,j : \text{ipa}(s_i) = \text{ipa}(s_j) \\ \Sigma_{s_i,k} = \Sigma_{s_j,k}, & \quad \forall i,j : \text{ipa}(s_i) = \text{ipa}(s_j) \end{cases}$$

A schematic of the mixed acoustic modeling method is shown in Figure 6 for the polyphoneme "M" which is shared between all languages.

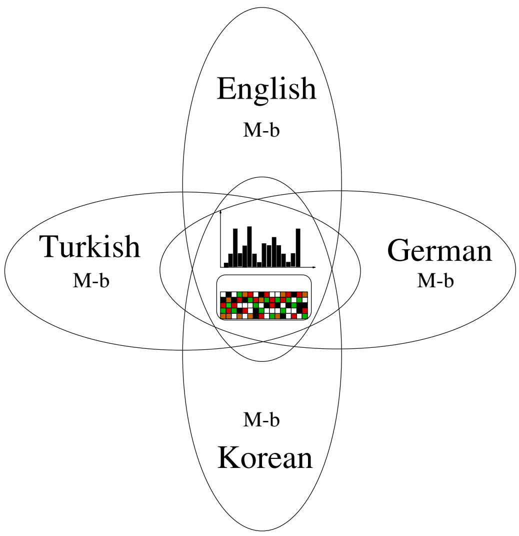


Figure 6: **ML-mix**: Language mixed acoustic modeling

Another way to share phoneme models across languages is performed in the model combination method *ML-tag*. Here each phoneme receives a language tag attached in order to preserve the information about the language the phoneme belongs to. *ML-tag* is similar to *ML-mix* in the sense that they both share all the training data and use the same clustering procedure. But for *ML-mix* the training data are only labelled by phoneme identity, whereas for *ML-tag* the training data is labelled by both phoneme and language identity. The clustering procedure is extended by introducing questions about the language and language groups to which a phoneme belongs. The Gaussian components are shared across languages as in the *ML-mix* method but the mixture weights are kept separately. Therefore, the relative importance of phonetic context and language membership is resolved during the clustering procedure by a data-driven method. The *ML-tag* combination method can be described as:

$$\text{ML-TAG}: \begin{cases} c_{s_i} \neq c_{s_j} & , \quad \forall i \neq j \\ \mu_{s_i,k} = \mu_{s_j,k} \ , & \quad \forall i,j : \text{ipa}(s_i) = \text{ipa}(s_j) \\ \Sigma_{s_i,k} = \Sigma_{s_j,k}, & \quad \forall i,j : \text{ipa}(s_i) = \text{ipa}(s_j) \end{cases}$$

We start with 650,000 different sub-quinphones defined over the five languages and create two fully continuous systems, *ML-tag3000* with 3000 models, and *ML-tag7500* with 7500 models, the latter one being of the same size as five monolingual systems each having 1500 models.

4.2. Simultaneous recognition

We explore the usefulness of our modeling approach by comparing the recognition performance of the monolingual case with the performance which is achieved by the resulting systems from the *ML-sep*, *ML-tag*, and *ML-mix* combination method. The experiments are done for the five languages Croatian, Japanese, Korean, Spanish and Turkish. The comparison focus on the purpose of simultaneously recognizing these languages which are involved for training the multilingual acoustic models. First we compare the monolingual system to the system *ML-sep* which only differs in the multilingual LDA. Compared to the monolingual case the multilingual LDA

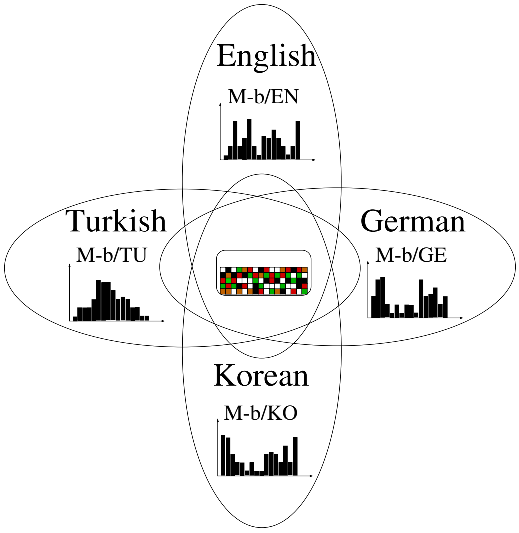


Figure 7: **ML-tag**: Language tagged acoustic modeling

slightly increase the word error rate but not significantly. When we compare the combination methods to each other we found that the system *ML-tag3000* outperforms the mixed system *ML-mix3000* in all languages by an average of 5.3% (3.1% - 8.7%) error rate. Since the collection of the GlobalPhone speech data is uniform in terms of recording and channel conditions we draw the conclusion that preserving the language information achieves better results with respect to simultaneous recognition. The *ML-tag3000* system reduces the model size to 40% compared to the monolingual case (3000 vs 5x1500 models), resulting in a 3.1% performance degradation on average (1.2% - 5.0%). However, not all of the degradation can be explained by the reduction of parameters. This can be derived from the comparison between the monolingual systems and *ML-tag7500*. We still observe an average performance gap of 1.1% (0.3% - 2.4%) when comparing the acoustic modeling with respect to simultaneous recognition of the relevant source languages. The finding coincides with other studies (Bonaventura et al., 1997; Cohen et al., 1997; Köhler, 1998). A detailed description of these experiments can be found in (Schultz and Waibel, 1998a).

4.3. Analysis of language questions

In this section we describe the pertinence of language information coded in the acoustic models. For this purpose we take the polyphone decision tree of *ML-tag*.

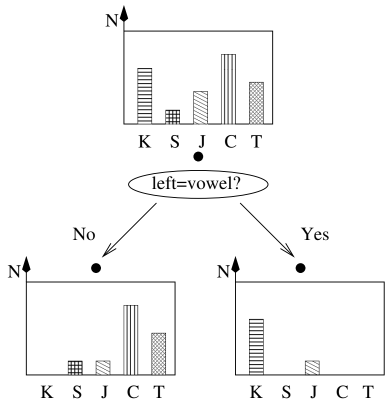


Figure 8: Language distribution

We compute the language distribution for each tree node as pictured in Figure 8 and calculate the language entropy gain by traversing the given tree. This gain D_H is calculated as $D_H = p(n_{yes})H_{yes} + p(n_{no})H_{no} - (p(n_{yes}) + p(n_{no}))H_{org}$ where H_n is the entropy of the distributions in node $n \in \{org, yes, no\}$ defined as $H_n = \sum_{i=1}^{l} p_n(i) \log_2 p_n(i)$ und $l = 5$. The resulting sum of the entropy gain D_H is plotted over the number of clustered sub-polyphones in Figure 9.

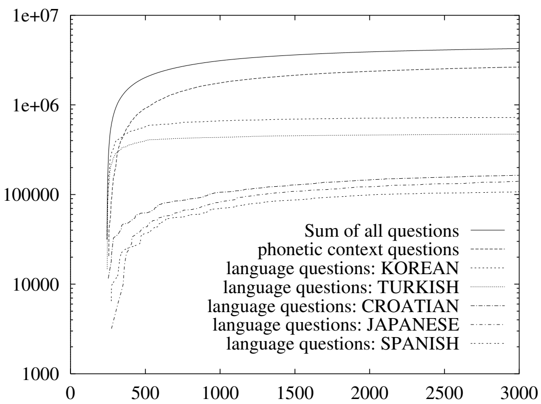


Figure 9: Entropy gain plotted over number of clustered sub-polyphones

This procedure enables us to analyze the ratio of language questions compared to phonetic questions. The curve *sum of all questions* gives the overall language entropy gain of all questions, whereas the curve *phonetic context questions* shows the entropy gain belonging to non-language questions. The big gap between both curves indicates that major parts of the entropy gain result from language questions. The remaining five curves give the contribution of questions belonging to only one language. It is shown that questions about Korean and Turkish are more important than questions about other languages, especially in the beginning of clustering. This indicates that sounds in those two languages seem to be different from the rest. Both results demonstrate that language questions are frequently asked and are especially more important in the beginning of the splitting process than questions about the phonetic context of a phoneme. After about 3000 splits the main part of language information are clustered out, which implies that in our case multilingual systems with more than 3000 polyphone clusters are composed largely of language dependent acoustic models.

5. Language adaptive acoustic modeling

Currently an important cost factor for developing LVCSR systems for new languages is the need for large amounts of transcribed audio data for training accurate acoustic models. To accommodate potential variations in the amount of training data available for the target language, we address three issues:

- No Data: Cross-language transfer
- Limited Data: Language adaptation
- Large amount of Data: Bootstrapping approach

The term cross-language transfer refers to the technique of using a recognition system on a new language without having ever seen any training data of the language in question. Research in this area investigates whether cross-language transfer between two languages of the same family performs better than across family borders (Constantinescu and Chollet, 1997), and whether the number of languages used for training the original acoustic transfer models influences the performance on the target language (Gokcen and Gokcen, 1997; Schultz and Waibel, 1998b). Some results indicate a relation between language similarity and cross-language performance (Bub et al., 1997; Constantinescu and Chollet, 1997). Furthermore, others (Bub et al., 1997) and our experiments have clearly shown that multilingual transfer models outperform monolingual ones (Schultz and Waibel, 1998a).

In the language adaptation technique, an existing recognizer is adapted to the new target language using very limited training data. Ongoing research (Wheatley et al., 1994; Köhler, 1998; Schultz and Waibel, 1998c) concentrates on two issues: The amount of adaptation data needed to get reasonable results and finding suitable acoustic models to start from. As expected, the language adaptation performance is strongly related to the amount of data used for adaptation. Wheatley et al. demonstrate that the number of training speakers is more critical than the number of training utterances (Wheatley et al., 1994). We investigate the issue of finding suitable initial models, comparing the effectiveness of multilingual acoustic models to monolingual models (Schultz and Waibel, 1998c). Once more our conclusion match those of other studies (Köhler, 1998); i.e. multilingual models outperform monolingual ones (Schultz and Waibel, 1998c).

The key idea in the bootstrapping approach is to initialize the acoustic models of the target language recognizer using seed models developed for other languages. After this initialization step, the resulting system is completely rebuilt using large amounts of training data from the target language. We had already applied this approach in earlier studies (Osterholtz et al., 1992) to bootstrap a German recognizer from English. Wheatley et al. (Wheatley et al., 1994) proved that cross-language seed models achieve lower word error rates than flat starts or random models. Recently, we demonstrated the usefulness of a global unit set and multilingual acoustic models as seed models (Schultz and Waibel, 1997).

Previous approaches for language adaptation have been limited to context independent acoustic models. Since for the language dependent case wider contexts increase recognition performance significantly, we investigate whether such improvements extend to the multilingual setting. The use of wider context windows raises the problem of phonetic context mismatch between source and target languages. To measure this mismatch we define the coverage coefficient. In order to approach the mismatch problem we introduce a method for polyphone decision tree adaptation.

5.1. Phonetic context mismatch

We define the coverage coefficient $\mathrm{cc_N(L_T)}$ of the target language L_T to be:

$$\mathrm{cc_N(L_T)} = \frac{|\Upsilon_{\mathrm{L_T}} \cap \Upsilon|}{|\Upsilon_{\mathrm{L_T}}|} = 1 - \frac{|\Upsilon_{\mathrm{LD_{L_T}}}|}{|\Upsilon_{\mathrm{L_T}}|} \qquad (3)$$

While the share factor sf defined in Section 2 measures the average sharing of all phonemes in the global unit set over all languages, the coverage coefficient cc gives us the portion of phonemes in the target language L_T which are covered by phonemes of the global unit set. The coverage coefficient is zero, if no phoneme of the target language L_T has a counterpart in the global unit set, and one if each phoneme is covered, i.e. $0 \leq \mathrm{cc(L_T)} \leq 1$.

The idea of phoneme coverage can be extended naturally to models of various context width. Based on the above definition we now introduce monophone coverage, triphone coverage and in general polyphone coverage. We further distinguish between the coverage of polyphone types and polyphone occurrences. For the latter the frequency of a polyphone is taken into account to reflect that coverage of frequent polyphones is more important than coverage of less frequent ones with respect to recognition performance.

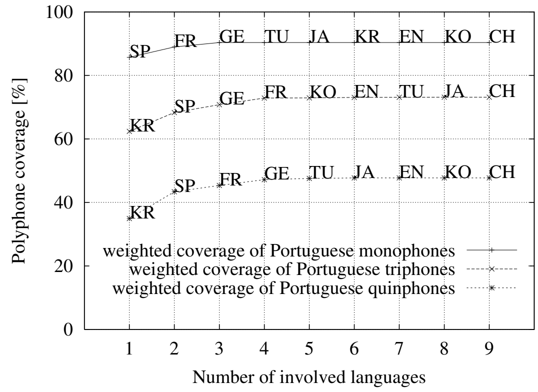


Figure 10: Portuguese polyphone coverage by nine languages

In the following we will apply the polyphone decision tree specialization procedure to adapt the multilingual recognition engine to the target language Portuguese. To examine how well the 46 Portuguese phonemes and resulting polyphones are covered by a given language pool, we calculated the coverage with respect to the global unit set (without Portuguese). The coverage indicates how well a generic polyphone decision tree fits to the target language Portuguese. The percentage coverage $\mathrm{cc(Po)} \times 100$ is plotted in Figure 10 for context width zero (monophones), one (triphones) and two (quinphones). The calculation of plotted coverage proceeds as follows: We select the language among all pool languages which achieves the highest coverage for Portuguese. Then we remove this language from the pool and calculate the coverage between Portuguese and each language pair resulting from the combination of removed language plus remaining pool language. The procedure is repeated for triples and so forth. Thus in each step we determine the language which maximally complements the polyphone set.

As expected, the coverage decreases dramatically for wider contexts. With a nine language pool (Russian and Swedish are not involved), the coverage of Portuguese monophones achieves 91%, drops to 73% for triphones and to 47% for quinphones. After incorporating the three main contribution languages the coverage for monophones cannot be increased any further. When enlarging the context width to one, coverage saturates after four languages. For a context width of two we observed that at least five languages contribute to the quinphone coverage rate. Therefore, we expect that increasing the context width requires more languages.

We experiment with removing the main contribution languages from the pool, i.e. we remove one of the languages Spanish, Croatian and French. Removing Spanish could nearly be compensate by German plus Croatian, and vice versa. This indicates that these three languages cover similar portions of the Portuguese polyphone set. It is not possible to compensate for the removal of French by including other languages as French provides unique polyphones not found elsewhere. In this case the missing phonemes are nasal vowels which are frequent in Portuguese. We conclude from this observation that, when designing a language pool for adaptation purposes, it is more critical to find a complementary set of languages than to cover a large number of languages. Calculating the polyphone coverage across languages helps to determine a complementary language set.

Table 4 summarizes the triphone coverage for 10 languages. The coverage of triphone types is given in the upper row, of triphone occurrences in the lower row. For example 33.6% of Japanese triphone occurrences are covered by German triphones, whereby 22.3% of the triphone types are responsible for this coverage rate. On the other hand only 19.5% of all German triphone occurrences are covered by Japanese triphones. This effect is due to the Japanese phonotactics which only allow consonant vowel combinations but no consonant clusters.

From analyzing the coverage in Figure 10 and Table 4 we draw the conclusion that a polyphone decision tree, even build on several languages, can not be applied successfully to a new language without adaptation.

Table 4: Triphone coverage matrix for 10 languages; 2 numbers are given for each matrix entry (i, j) meaning that language i is covered by language j with triphone types (upper number) and triphone occurrences (lower number)

	CH	EN	FR	GE	JA	KO	KR	PO	SP	TU
CH	100	0.3 6.8	0.1 5.8	0.1 5.3	0.1 4.2	0.0 5.3	0.1 4.2	0.1 5.4	0.1 5.3	0.2 4.9
EN	0.6 5.2	100	6.5 18.6	5.4 18.1	1.8 8.9	3.4 11.6	1.5 7.7	0.9 6.6	1.3 6.6	3.8 9.2
FR	0.1 3.9	9.7 16.4	100	29.0 53.3	10.2 22.7	11.2 28.7	25.8 45.5	18.4 36.4	17.4 41.3	23.1 35.6
GE	0.1 3.9	5.5 19.6	19.8 41.6	100	9.3 19.5	7.2 18.2	18.6 34.9	13.6 28.0	12.9 28.3	12.9 26.1
JA	0.2 2.5	4.5 9.9	16.8 37.4	22.3 33.6	100	9.8 25.6	16.0 29.2	11.0 27.6	13.6 31.2	25.9 52.5
KO	0.1 4.1	4.9 16.1	10.9 35.0	10.3 36.3	5.8 24.9	100	10.2 38.6	8.0 30.8	9.3 38.4	9.1 26.1
KR	0.2 1.8	3.2 5.0	37.0 64.7	39.0 68.8	14.0 28.2	15.0 34.5	100	31.0 63.0	34.3 61.8	31.5 50.4
PO	0.4 2.3	2.0 4.6	28.0 49.5	30.2 57.9	10.2 26.7	12.5 37.5	32.9 62.5	100	33.5 57.5	19.8 39.9
SP	0.2 2.5	2.7 5.6	23.5 60.1	25.4 60.2	11.2 34.0	12.9 40.1	32.2 64.2	29.7 58.2	100	17.5 41.0
TU	0.8 5.4	8.9 18.3	36.3 52.0	29.6 46.0	24.8 46.1	14.6 33.0	34.4 50.1	20.4 38.6	20.3 39.6	100

5.2. Polyphone decision tree specialization

In order to overcome the problem of the observed mismatch between represented context in the multilingual polyphone decision tree and the observed polyphones in the new target language, we propose the Polyphone Decision Tree Specialization (PDTS) procedure. In PDTS the clustered multilingual polyphone decision tree is adapted to the target language by restarting the decision tree growing process according to the limited adaptation data available in the target language (Schultz, 2000).

Figure 11 illustrates the polyphone cluster tree for the middle state of the phoneme d^j before adaptation. Dur-

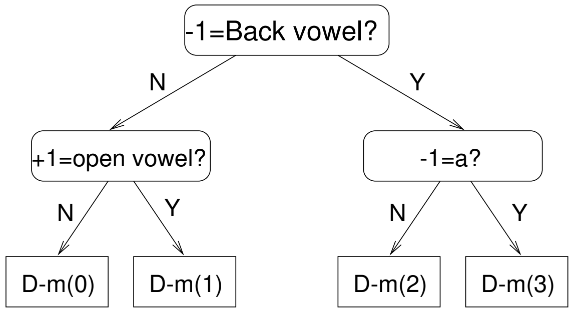


Figure 11: Tree **before** Polyphone Decision Tree Specialization

ing the clustering procedure only three splits resulting in four leaf nodes were used to capture the phonetic context of d^j in the multilingual data. However, in the Portuguese language this phoneme is very frequent and occurs in very different contexts. Traversing this non-adapted tree during decoding Portuguese speech would lead to very poorly estimated residual class models, since the context questions do not reflect the Portuguese contexts.

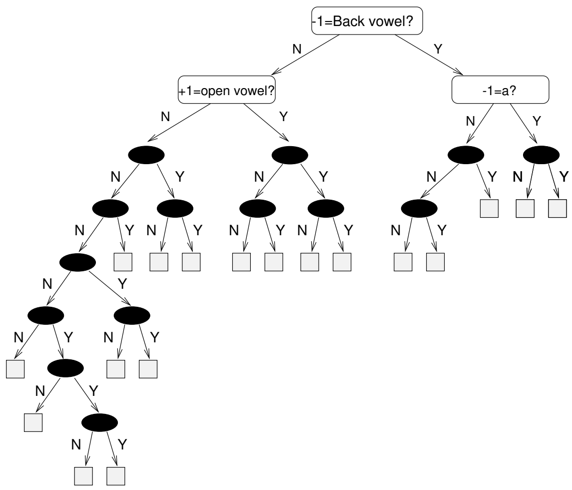


Figure 12: Tree **after** Polyphone Decision Tree Specialization

Figure 12 shows the decision tree for the middle state of the same phoneme d^j after applying PDTS. The former tree was further clustered according to 14 additional questions, resulting in 18 leaf nodes. The re-growing process is completed after reaching a predefined number of new leaf nodes depending on the amount of training data. The adapted decision tree now represents valid contexts of the Portuguese d^j and is expected to improve the recognition results for Portuguese input. This will be evaluated in the experiments described in Section 6.

6. Comparative experiments

In the following experiments we investigate the benefit of the acoustic model combination and the polyphone decision tree specialization (PDTS) for the purpose of adaptation to the Portuguese language. The above-described five-lingual recognition systems are ported to Portuguese using different amounts of data. We assume that a Portuguese dictionary as well as the recordings and transcriptions of some spoken utterances are given. The dictionary mapping is done according to an heuristic IPA-based mapping approach (Schultz and Waibel, 1998c). A subset of 300 utterances from 10 test speakers is used to carry out the experiments. The test dictionary has about 7300 entries, the OOV-rate is set to 0.5% by including the most common words of the test set into the dictionary. A trigram language model with Kneser/Ney backoff scheme is calculated on a 10 million word corpus from Agency France Press (LDC95T11, distributed by LDC) interpolated with the GlobalPhone training data leading to a trigram perplexity of 297. For adapting the acoustic models we use 15 minutes, 25 minutes, and 45 minutes of speech spoken by 8 speakers. We also experiment with 45 minutes spoken by 16 speakers, and 90 minutes spoken by 16, 32, and all 78 training speakers.

Figure 13 summarizes the experiments which have been performed to improve the Portuguese LVCSR system. The row labelled **SystemId** gives the name which is used to identify the developed systems. The row **Data** refers to the amount of adaptation data (0-90 minutes of spoken speech). **Quality** explains whether the phonetic alignments are *initially* created based on the multilingual recognition engine or assumed to be available in *good* quality. The term **Method** is related to the porting approach which is applied: Cross-language transfer (CL), adaptation (Viterbi or MLLR), and bootstrapping technique (Boot). *Viterbi* refers to one iteration of Viterbi training along the given alignments. *MLLR* is the Maximum Likelihood Linear Regression (Leggetter and Woodland, 1995), and *Boot* refers to the iterative procedure: creating alignments, Viterbi training, model clustering, training, and writing improved alignments. The item **Tree** describes the origin of the polyphone decision tree: *'–'* refers to context independent modeling, *LI* is the generic language independent polyphone decision tree of system *ML-mix3000*, *LD* is the language dependent tree which is built exclusively on Portuguese data, and *PDTS* refers to the adapted LI polyphone tree after applying PDTS.

6.1. The Golden line

In the best case we have an entire database for the target language containing dozens of hours of recorded and transcribed speech together with a dictionary and large text corpora. The performance which can be achieved based on such knowledge sources represents our golden line. To determine this golden line we train a Portuguese systems with 16.5 hours of spoken speech from the GlobalPhone database and test the final system based on the aforementioned dictionary and language model. The resulting Portuguese system (SystemId S14) achieves a word error rate of 19.0%. In the following experiments we explore how close we can get to this number by applying the above-defined methods.

6.2. Transfer procedure

According to our finding that language independent models outperform language dependent ones when using them as seed models for a new target language and the fact that the ML-mix combination method performs better than ML-tag for cross-language transfer, we use *ML-mix3000* as the basis system for the adaptation to Portuguese. We start with exploiting the effect of different transfer procedures as summarized in Table 5.

Table 5: Transfer procedure

SystemId	Method	Word Error [%]		Improvement	
		CI	CD		
S2 / S1	Cross-language	69.1	72.0	17.4%	30.7%
S4 / S6	Adaptation	57.1	49.9	-	6.8%
S3	Bootstrapping	-	46.5		

The systems S1 and S2 represent the cross-language transfer approach for context-dependent (CD) and context-independent (CI) modeling respectively. For these systems only the data of the five source languages has been applied for training the acoustic models, no adaptation is performed before decoding the Portuguese speech. Overall, this leads to poor results; the context independent system (S2) slightly outperforms the context dependent system (S1), therefore, the initial alignments are written with system S2. These initial alignments of 15 minutes Portuguese speech are used for adaptation, which leads to 17.4% word error rate reduction in the context independent (S2 $\rightarrow$ S4), and to 30.7% word error rate reduction in the context dependent case (S1 $\rightarrow$ S6). The improvement through context dependent modeling (S4 $\rightarrow$ S6) indicate that the language independent polyphone tree covers some parts of Portuguese phonotactics. However, system S3 which results from the iterative bootstrapping procedure on the same adaptation data, outperforms system S6, i.e. a system with a polyphone decision tree build solely on Portuguese data achieves better results than a system with a non-adapted generic polyphone decision tree trained from various languages, provided that 15 minutes of adaptation data are available.

6.3. Acoustic model training

We compare the training methods which have been applied to the acoustic models. Viterbi refers to one iteration of Viterbi training along the given alignments, MLLR is the Maximum Likelihood Linear Regression. Although, MLLR was originally designed for speaker adaptation, the results in Table 6 show that it can be successfully applied to language adaptation. Provided that 15 minutes of Portuguese speech are given for adaptation, MLLR outperforms the Viterbi training by 4.4%.

Table 6: Acoustic model training

SystemId	Method	Word Error [%]	Improvement
S5	Viterbi	52.2	4.4%
S6	MLLR	49.9	

6.4. PDTS

Next we investigate the effect of specializing the polyphone decision tree according to the proposed PDTS procedure. We compare the PDTS specialized polyphone tree (S10) to non-adapted language independent trees (S6, S8) and to language dependent trees which are trained solely on Portuguese adaptation material (S3, S9). The results are summarized in Table 7 for 15 minutes and 25 minutes adaptation data respectively. The language independent polyphone trees are outperformed by the language dependent ones if no tree specialization is applied. The performance difference increases from 6.8% to 19.2% after the amount of adaptation data is extended to 25 minutes. However, the PDTS adapted tree (S10) significantly outperforms even the language dependent tree in system S9 by 11.9% which means that the knowledge and phonotactics of several languages stored in the polyphone decision tree can be transfered successfully to a new target language.

Table 7: The PDTS method [WE in %]

SystemId	Method	Alignments		Improvement	
		15 min initial	25 min good		
S6/S8	ML-Tree	49.9	40.6	6.8%	19.2%
S3/S9	Boot	46.5	32.8	-	11.9%
S10	PDTS	-	28.9		

6.5. Adaptation data

The phonetic alignments of the Portuguese adaptation utterances are initially created by the multilingual recognition system S2 (initial alignments). In order to accelerate our adaptation process we create improved phonetic alignments which we assume to be available (good alignments). Furthermore, we evaluate the effect of extending the adaptation data, from 15 to 25, then to 45, and finally to 90 minutes of spoken speech.

Table 8: Quality of adaptation data

SystemId	Data	Quality	WE [%]	Improvement
S6	15 min	initial alignments	49.9	13.2%
S7	15 min	good alignments	43.3	

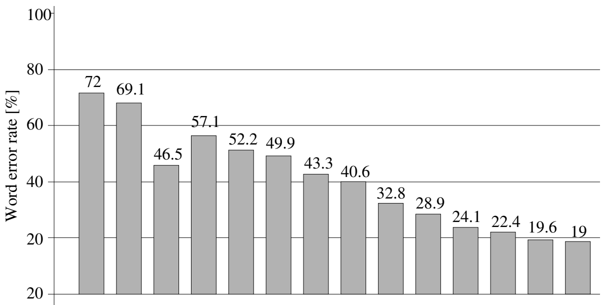
Improving the alignment quality decreases the word error rate by 13.2% as can be seen from Table 8. Nearly doubling the amount of adaptation data gives 16.6% and 12.5% improvement, whereas we achieved 7.1% by doubling the number of adaptation speakers, reported in Table 9. Further extension of the number of speakers did not lead to any improvements.

Table 9: Amount of adaptation data

SystemId	Data	Speakers	WE [%]	Improvement
S10	25 min	8	28.9	
S11	45 min	8	24.1	16.6%
S12	45 min	16	22.4	7.1%
S13	90 min	16	19.6	12.5%

6.6. Résumé

Figure 13 summarizes the word error rates on the Portuguese language for all above-described systems. As expected the recognition of Portuguese speech on the five-lingual recognizer *ML-mix* is poor when no adaptation is performed (S1, S2). System S2 is used to write initial phonetic alignments for adapting the context independent multilingual system (S4) and the context dependent system by Viterbi training (S5) and MLLR (S6). Adaptation by MLLR achieves the highest improvements. In S3 the initial alignments are used to completely rebuild a Portuguese system after bootstrapping from multilingual seed models. The comparison of S6 and S3 indicates that the bootstrap technique outperforms the adaptation when no polyphone decision tree specialization and only 15 minutes of adaptation data has been applied. Nevertheless, the word error rate of 46.5% achieved by the best system S3 is still unsatisfying.



SystemId	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
Data	0		15 minutes					25 minutes			45 8	16	90	16.5h
Quality	-		initial alignments				good alignments							
Method	CL		Boot	MLLR	Vit	MLLR			Boot	MLLR				Boot
Tree	LI	-	LD	-	LI				LD	PDTS				LD

Figure 13: Language adaptation to Portuguese; systems identified by SystemId according to the used amount of adaptation data (0 to 90 minutes of spoken speech), quality of given alignments, applied porting method (Cross-language=CL; Adaptation=MLLR or Viterbi training; Bootstrapping) and type of polyphone tree (context independent='–'; language dependent=LD; language independent=LI; specialized by PDTS=PDTS)

We obtain a significant performance boost from improving alignments (S6 $\rightarrow$ S7) and doubling the amount of adaptation data (S7 $\rightarrow$ S8). While the bootstrapping approach leads to 32.8% (S9), applying the PDTS method leads to a significant improvement of 12% (S9 $\rightarrow$ S10) achieving 28.9% word error rate. This result shows that knowledge from other languages can successfully be adapted to the target language. By extending the amount of adaptation data we achieve another improvement to 24.1% word error (S10 $\rightarrow$ S11). Doubling the number of speakers results in 22.4% error rate (S11 $\rightarrow$ S12). Finally we reach 19.6% word error rate applying the PDTS method based on 90 minutes adaptation data (S13). This result compares to 19.0% word error rate of our golden line (S14) given a large Portuguese database of 16.5 hours training data. The complete adaptation procedures runs on a 300MHz SUN Ultra and takes only 3-5 hours real-time.

7. Summary and Conclusion

In this article we addressed language dependent, language independent, and language adaptive acoustic modeling for read speech recognition using a high number of different languages. Based on the multilingual GlobalPhone database we built monolingual LVCSR systems for ten languages and highlighted language differences and the resulting challenges for speech recognition. Several methods were introduced to combine the language dependent acoustic models to language independent ones. The latter allow data and model sharing across languages and were applied for simultaneous recognition in a compact language independent LVCSR system.

Provided that speech databases are limited in general, we approached the problem of porting acoustic models to a new target language by borrowing models and data from various languages but using only a limited amount of adaptation data from the target language. We explored the relative effectiveness of language independent acoustic models with a wider context in combination with a polyphone decision tree specialization (PDTS) method.

The PDTS method gave 12% relative improvement compared to a recalculation of a language specific polyphone tree and 28% compared to a non specialized multilingual polyphone tree. In summary, we achieved 19.6% word error rate when adapting language independent acoustic models to the Portuguese language using only 90 minutes of spoken Portuguese speech. This compares to 19.0% of a full trained system on 16.5 hours of spoken Portuguese speech. The adaptation procedures runs on a 300MHz SUN Ultra and takes only 3-5 hours real-time.

As a consequence the introduced techniques allow to set up LVCSR systems in a new target language without the need of large speech databases in that language. In combination with the letter-to-sound mapping tools and a full automatically downloading of text resources from the Internet, LVCSR systems in read speech could be developed very efficiently.

8. Acknowledgment

We thank all members of the Interactive Systems Laboratories and the GlobalPhone team. This research would not have been possible without the great enthusiasm of all team members during the collection and validation of the database. We would also wish to acknowledge the anonymous reviewers for their thoughtful comments on an earlier version of this article.

9. References

Andersen, O., Dalsgaard, P., and Barry, W., (1993). Data-Driven Identification of Poly- and Mono-phonemes for four European Languages. In: Proc. Eurospeech, Berlin 1993, pp. 759-762.

Andersen, O., and Dalsgaard, P., (1997). Language Identification based on Cross-language Acoustic Models and Optimised Information Combination. In: Proc. Eurospeech, Rhodes 1997, pp. 67-70.

Barnett, J., Corrada, A., Gao, G., Gillik, L. Ito, Y., Lowe, S., Manganaro, L., and Peskin, B., (1996). Multilingual Speech Recognition at Dragon Systems. In: Proc. ICSLP, Philadelphia 1996, pp. 2191-2194.

Bonaventura, P., Gallocchio, F., and Micca, G., (1997). Multilingual Speech Recognition for Flexible Vocabularies. In: Proc. Eurospeech, Rhodes 1997, pp. 355-358.

Bub, U., Köhler, J., and Imperl, B. (1997). In-Service Adaptation of Multilingual Hidden-Markov-Models. In: Proc. ICASSP, Munich 1997, pp. 1451-1454.

Billa, J., Ma, K., McDonough, J., Zavaliagkos, G., Miller, D. R., Ross, K. N., and El-Jaroudi, A., (1997). Multilingual Speech Recognition: The 1996 Byblos Callhome System. In: Proc. Eurospeech, Rhodes 1997, pp. 363-366.

Çarkı, K., Geutner, P. and Schultz, T., (2000). Turkish LVCSR: Towards better Speech Recognition for Agglutinative Languages. In: Proc. ICASSP, Istanbul 2000, pp. 1563–1566.

Cohen, P., Dharanipragada, S., Gros, J., Monkowski, M., Neti, C., Roukos, S., and Ward, T., (1997). Towards a Universal Speech Recognizer for Multiple Languages. In: Proc. Automatic Speech Recognition and Understanding (ASRU), St. Barbara CA 1997, pp. 591-598.

Constantinescu, A., and Chollet, G., (1997). On Cross-Language Experiments and Data-Driven Units for ALISP. In: Proc. Automatic Speech Recognition and Understanding (ASRU), St. Barbara CA 1997, pp. 606-613.

Corredor-Ardoy, C., Gauvain, J.L., Adda-Decker, M., and Lamel, L., (1997). Language Identification with Language-independent Acoustic Models. In: Proc. Eurospeech, Rhodes 1997, pp. 355-358.

Dugast, C., Aubert, X., and Kneser, R., (1995). The Philips Large-Vocabulary Recognition System for American English, French, and German. In: Proc. Eurospeech, Madrid 1995, pp. 197-200.

ELRA, The European Language Resources Association. Available as: http://www.icp.grenet.fr/ELRA/home.html.

Finke, M. Geutner, P., Hild, H., Kemp, T., Ries, K., and Westphal, M., (1997). The Karlsruhe-Verbmobil Speech Recognition Engine. In: Proc. ICASSP, Munich 1997, pp. 83–86.

Finke, M., and Rogina, I., (1997). Wide Context Acoustic Modeling in Read vs. Spontaneous Speech. In: Proc. ICASSP, Munich 1997, pp. 1743–1746.

Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., Seneff, S., and Zue, V., (1995). Multi-lingual Spoken Language Understanding in the MIT Voyager System. Speech Communication 17, 1-18.

Gokcen, S., and Gokcen, J., (1997). A Multilingual Phoneme and Model Set: Towards a universal base for Automatic Speech Recognition. In: Proc. Automatic Speech Recognition and Understanding (ASRU), St. Barbara CA 1997, pp. 599-603.

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

IPA, (1993). The International Phonetic Association (revised to 1993) - IPA Chart. Journal of the International Phonetic Association 23.

Kiecza, D., Schultz, T., and Waibel, A., (1999). Data-Driven Determination of Appropriate Dictionary Units for Korean LVCSR. In: Proc. International Conference on Speech Processing (ICSP), Seoul 1999, pp. 323–327.

Köhler, J., (1998). Language Adaptation of Multilingual Phone Models For Vocabulary Independent Speech Recognition Tasks. In: Proc. ICASSP, Seattle 1998, pp. 417-420.

Lamel, L., Adda-Decker, M., and Gauvain, J.L., (1995). Issues in Large Vocabulary Multilingual Speech Recognition. In: Proc. Eurospeech, Madrid 1995, pp. 185-189.

Leggetter, C., and Woodland, P., (1995). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. Computer Speech and Language 9, 171-185.

LDC, The Linguistic Data Consortium. Available as: http://www.ldc.upenn.edu.

Osterholtz, L., Augustine, C., McNair, A., Rogina, I., Saito, H., Sloboda, T., Tebelskis, J., Waibel, A., and Woszczyna, M., (1992). Testing Generality in JANUS: A Multi-lingual Speech Translation System. In: Proc. ICASSP, San Francisco 1992.

Reichert, J., Schultz, T., and Waibel, A., (1999). Mandarin Large Vocabulary Speech Recognition using the GlobalPhone Database. In: Proc. Eurospeech, Budapest 1999, pp. 815–818.

Schultz, T., (2000). Multilinguale Spracherkennung: Kombination akustischer Modelle zur Portierung auf neue Sprachen. Dissertation, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, 2000.

Schultz, T., and Waibel, A., (1997). Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets. In: Proc. Eurospeech, Rhodes 1997, pp. 371-374.

Schultz, T., and Waibel, A., (1998a). Multilingual and Crosslingual Speech Recognition. In: Proc. DARPA Workshop on Broadcast News Transcription and Understanding, Lansdowne VA 1998, pp. 259-262.

Schultz, T., and Waibel, A., (1998b). Adaptation of Pronunciation Dictionaries for Recognition of Unseen Languages. In: Proc. SPIIRAS International Workshop on Speech and Computer, St. Petersburg 1998, pp. 207-210.

Schultz, T., and Waibel, A., (1998c). Language Independent and Language Adaptive LVCSR. In: Proc. ICSLP, Sydney 1998, pp. 1819-1822.

Schultz, T., Westphal, M., and Waibel, A., (1997). The GlobalPhone Project: Multilingual LVCSR with Janus-3. In: Proc. SQEL, 2nd Workshop on Multi-lingual Information Retrieval Dialogs, Plzeň 1997, pp. 20-27.

Wells, C. J., (1989). Computer-coded Phonemic Notation of Individual Languages of the European Community. Journal of the International Phonetic Association 19, 32-54.

Wheatley, B., Kondo, K., Anderson, W., and Muthusamy, Y., (1994). An Evaluation of Cross-language Adaptation For Rapid HMM Development in a new language. In: Proc. ICASSP, Adelaide 1994, pp. 237-240.

Webster's, (1992). New Encyclopedic Dictionary. Black, Dog & Leventhal.

Young, S. J., Adda-Decker, M., Aubert, X., Dugast, C., Gauvain, J.L., Kershaw, D. J., Lamel, L., Leeuwen, D. A., Pye, D., Robinson, A. J., Steeneken, H. J. M., and Woodland, P. C., (1997). Multilingual Large Vocabulary Speech Recognition: The European SQALE Project. Computer, Speech, and Language 11, 73-89.