Towards Language Independent Acoustic Modeling

The 1999 Johns Hopkins University
Language Engineering Workshop

Bill Byrne, Group leader	CLSP Johns Hopkins University
Peter Beyerlein	Philips Research Laboratories
Juan Manuel Huerta	Electrical and Computer Engineering Carnegie Mellon University
Sanjeev Khudanpur	CLSP Johns Hopkins University
Bhaskara Marthi	Computer Science and Mathematics University of Toronto
John Morgan	Department of Foreign Languages USAMA, West Point
Nino Pterek	UFAL Charles University, Prague
Joe Picone	ISIP Mississippi State University
Wei Wang	Electrical and Computer Engineering Rice University

Contents

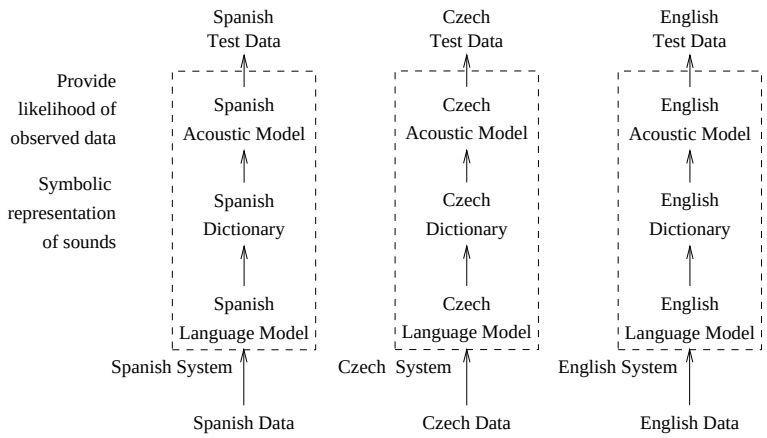

Figure 1: Modeling Multilingual Speech Using Monolingual ASR Systems.

1 Introduction

Language independent acoustic modeling was one of the topics studied at the 1999 Johns Hopkins University Language Engineering Workshop hosted by the Center for Language and Speech Processing. Our work was motivated by the need for speech recognition in languages beyond the well-studied languages of Europe, Asia, and the Americas. The statistical techniques used for speech and language modeling require relatively large amounts of monolingual speech and text as training data. In 'resource-rich' languages which have such corpora, these statistical estimation methods have been shown to work quite well. However, if only small amounts of training data are available in a language, these monolingual techniques are less effective. Our goal was to address this problem by developing techniques that reduce the amount of data needed to model resource-poor languages by borrowing data and models from resource-rich languages.

Multilingual ASR techniques are a significant departure from current practice, in that the best ASR systems are inherently monolingual. Their language models, pronunciation dictionaries, and acoustic models are constructed using data only from the language of interest, and make no use of data or models from other languages. These monolingual techniques are limited in processing multiple languages. Unless the relationships between the individual languages is described and captured, separate systems for each language must be built and made to operate independently as shown in Figure 1. Of course, this should be acceptable if an adequate system can be trained for each language. However, as described above, there are situations in which it may be desirable to borrow data and models across languages.

While in our studies we used multiple languages simultaneously, our goal was not to build a multilingual ASR system capable of recognizing several languages equally well. We intended instead to develop a good monolingual system for a specified target language by borrowing data and models from other languages. In speaker independent ASR, models are first trained using speech from multiple speakers and then adapted to a specific speaker either before or during recognition. Analogously, *language independent acoustic modeling* is a methodology that combines speech and models from multiple source languages and transforms them for recognition in a specific target language.

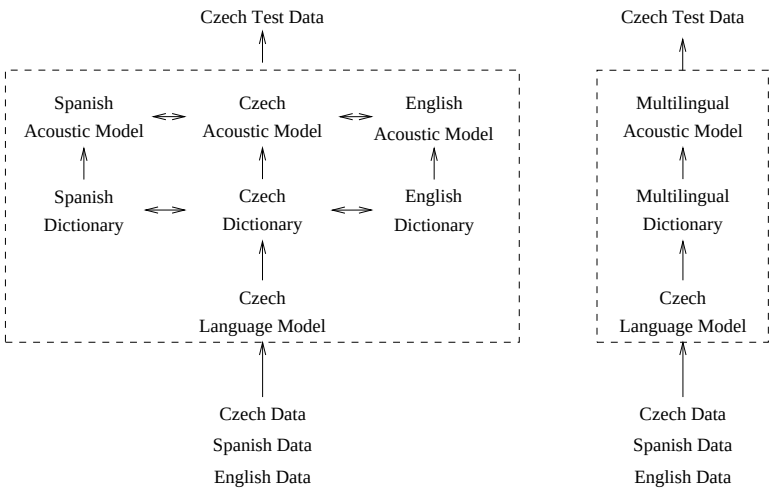

Figure 2: Two Approaches to Language Independent Acoustic Modeling. Monolingual systems can be trained independently and merged prior to use (left) or speech data from several languages can be pooled to train a single set of acoustic models (right).

As mentioned above, acoustic training data is not the only resource needed for statistical ASR. However, we have assumed for our work that language models, pronunciations, and appropriate acoustic processing are available for the target language, and that only transcribed acoustic training data is in short supply. This is not a completely unrealistic scenario in that dictionaries with pronunciations are available for many languages, as are on-line newspapers and other text. However, we stress that in our work we have addressed only one aspect of language independent modeling.

1.1 Previous work

As part of our summer project we conducted a literature review and held a 'paper club' to familiarize ourselves with previous relevant work in cross-lingual and multilingual acoustic modeling. For a recent overview, see the report by Hovey *et al.* [16]. We also found work described in the following references to be relevant to our studies [16, 6, 12, 8, 14, 18, 13, 11, 5]

1.2 Project Overview

We have developed methods for automatic speech recognition that share data and acoustic models across languages. Underlying these methods are *phone mappings* that identify similar speech sounds across languages. We obtain these phone mappings using both *knowledge-based* and *automatic* methods. The knowledge-based methods rely only on acoustic-phonetic categorizations of the individual languages and as such can be used if no data at all is available in the target language. The automatic methods derive phone mappings using small amounts of acoustic data in the target language. Using mappings found by either approach we can borrow models from several languages simultaneously to cover the phone inventory of the target language as depicted in Figure 2. The automatic methods allow additional refinement by borrowing models sub-phonetically at the HMM-state level. This can be especially valuable if the target language contains phones not found

in any of the source languages since these techniques are free to assemble a new phone model from component states of different source language phone models.

While both the automatic and knowledge-based phone mappings can be used without modification to construct recognizers in the target language by borrowing acoustic models from the various source languages, HMM adaptation techniques can also be used to improve the systems using the small amount of target language adaptation data we assume is available.

As a further refinement, we obtained the best recognition performance not from individually adapted source language acoustic models but by using Discriminative Model Combination (DMC) to combine acoustic models from several languages simultaneously. Referring to the left ASR architecture of Figure 2, it is not at all obvious how best to combine hypotheses produced using several sets of acoustic models. DMC provides a principled and effective way to do this. This combination can be done at the sentence or sub-word level, with better performance obtained using phone-level combinations. We note in particular that DMC makes effective use of source language acoustic models that individually do not perform well in transcribing the target language.

We began experiments in *language adaptive training* in an attempt to train a single set of acoustic models using a multilingual training set, as depicted in Figure 2. This work is still on-going, however our work in *Language Adaptive Clustering* provides strong evidence that all the above methods should benefit from acoustic normalization that transforms data and models as part of the phonetic mapping between languages.

1.3 Accomplishments

We summarize here the achievements of our summer project. Novel techniques and results not previously reported in the literature are italicized.

- An experimental framework for language independent acoustic modeling that is cross-domain as well as multilingual
- Creation of Czech language Broadcast News corpus
- Compilation of acoustic phonetic characterizations of English, Czech, Russian, Spanish, and Mandarin
- Development of knowledge-based phone-mappings that allow sharing models across languages when no training data is available
- *Development of automatic methods to derive sub-phonetic mappings to share acoustic HMMs between languages*
- *Development of Language Adaptive Clustering to derive automatic phone-level mappings and cross-language acoustic normalization*
- *Use of phone-level cross-language normalization to improve source language HMM performance*
- Use of HMM adaptation techniques to improve source language HMMs
- *Use of DMC to combine acoustic scores derived from multiple source language acoustic models*

- *Use of ROVER to combine acoustic scores derived from multiple source language acoustic models*
- *Improvement over an monolingual baseline system* using multilingual methods

2 Czech Monolingual Experiments

As part of our research program we have established an experimental framework for language independent acoustic modeling. Since this problem has not been widely studied, we were not able simply to use previously defined training and test sets to evaluate our ideas. We therefore began our work by investigating ASR performance in an attempt to find an appropriate 'operating point' at which to conduct our experiments.

ASR performance is determined by a variety of factors. Performance is generally poor if the speech to be recognized is produced spontaneously, as in conversational speech. Read speech is in general easier to recognize, and, as a special case, 'planned speech' by trained broadcast announcers can be particularly easy to transcribe. Performance also depends on the modeling techniques used, the recording conditions, the amount of data available to train language and acoustic models, and the similarity between the training data and the test material. We considered these last two factors to be especially important in defining our experiments, since performance is generally best given a large training set that closely resembles the test set.

Our initial plan to define our experiments was straightforward. We first decided on Czech language news broadcasts as our test domain. We choose to use news broadcasts because they contain a variety of different types of speech and are relatively easy to obtain. We choose the Czech language because there are ongoing language engineering projects studying Czech from which we would be able to borrow resources. We also felt that studying Czech was a realistic task since, unlike Spanish or Mandarin, there is fairly little knowledge of existing Czech ASR systems to influence our work. To obtain the needed broadcast training and test material, we arranged with the Linguistic Data Consortium to record Czech language Voice of America (VOA) broadcasts.

We decided to build our initial Czech broadcast news system from a ten hour Czech VOA acoustic training set using techniques known to work well in other languages and domains. The language model and pronouncing dictionary were developed in our previous work in transcription of read Czech [3]. Such an effort is a common exercise in training an ASR system using a moderately large amount of homogeneous acoustic training data.

After obtaining the performance of this well-trained system, we planned to reduce drastically the size of the acoustic training set and build a new, impoverished system. Given our past experience and the reported experience of others, we expected that training a system using approximately one hour of acoustic training data would yield an ASR system that performed substantially worse than the initial, well-trained system.

This reduced-size acoustic training set, the impoverished models, along with the dictionary and language model would serve as the baseline for our second set of experiments. We would attempt to improve the performance of the impoverished system by borrowing both acoustic training data and full ASR systems from other languages. In summary, our plan was to begin with a well-trained monolingual system built from homogeneous data and 'back into' a heterogeneous multi-lingual domain by reducing the target language acoustic training data.

As the following account describes, our experiments did not go as we expected. We found that speaking-style effects in the Czech VOA recordings dominated all other factors in ASR performance. The 'planned speech' of the VOA announcers was very easy to recognize, no matter how little data was used to train the system. We therefore were forced to obtain other news broadcasts data for use as our test set. The remainder of this section describes the data and experimental conditions in more detail.

2.1 Czech Speech and Language Data

Read Speech

Our initial experience with Czech ASR is in the transcription of read speech [3]. We developed pilot ASR systems using speech from the Charles University Financial News Corpus (CUCFN). We used the portion of this corpus that consists of recordings of read economic news taken from the Ceskomoravsky Profit Journal. This database consists of speech read by fluent Czech speakers recorded in quiet conditions at 22KHz with 16 bit resolution. The speech was recorded simultaneously with both a Sennheiser head-mounted, close-talking microphone and a desk-mounted microphone. In our work we used the recordings from the desk-mounted microphone channel. Speech from 29 male speakers and 23 female speakers has been collected and verified. Most subjects were native speakers of common Czech, except for some speakers with marked regional accents from North Moravia and South Moravia. There was also one native Russian speaker and one native Macedonian speaker. The first stage of corpus contains a total of 7280 sentences yielding slightly more than 17 hours of speech.

Broadcast Speech

Satellite transmissions of Voice of America broadcasts were recorded by the Linguistic Data Consortium (LDC) and transcribed at the University of West Bohemia according to protocols developed by LDC for use in Broadcast News LVR evaluation. The recordings span the period February 8 through May 4, 1999. The corpus consists of 46 recordings of 30 minute news broadcasts yielding a total of 23 hours of broadcast material. Portions of the shows containing music, speech in music, or other non-speech material are marked, but these intervals were not transcribed. This yields approximately 19:30 minutes of transcribed material from each 30 minute broadcast, for a total of 20 hours 24 minutes of pure transcribed speech.

Broadcasts from another news source were recorded to complement the VOA collection. Several programs broadcast by Český rozhlas 1 - Radiožurnál (*http://radiozurnal.CRo.cz*) on July 30 and 31, 1999 were recorded at Charles University. The shows contained general news with a mix of discussions, spontaneous and planned speech. The FM broadcasts were recorded directly onto a PC using the CoolEdit (*http://www.syntrillium.com*) program at 22KHZ and 16 bit resolution. The data was then transcribed at CLSP during the workshop. Through this impromptu collection effort we obtained an additional 99 minutes of transcribed speech intended primarily for use as a test set.

Czech Language Models

In our experiments this summer we used language models developed in our previous work on read Czech. The language model vocabulary was 63K words, and we used a bigram language model trained from a 16.5 million word corpus of news text (Lidové Noviny 1991-1994). Table 1 shows the perplexity of representative samples of the three Czech language databases. Although the language model training corpus is from another domain, perplexities and OOV rates are fairly consistent across the different test sets.

Corpus	Perplexity	OOV Rate
CUCFN	737.5	6.7%
VOA	664.0	6.7%
CRo1	763.6	7.8%

Table 1: Test set perplexities and OOV Rates

Training Set Size (hours)	Model Set		WER (%)
	Mixtures	Type	
Czech VOA Test Set 1			
12.8	12	3886 state xword triphone	27.1
10.0	12	monophone	27.6
1.0	8	monophone	30.2
0.5	20	monophone	31.3
Czech VOA 1.0 Hour Acoustic Training Set			
1.0	10	monophone	26.1

Table 2: Training and Testing on Czech VOA Broadcasts. Word Error Rate (WER) changes very little despite large variations in model complexity and training set size.

2.2 Czech Baseline ASR Experiments

We defined a variety of training and test sets in the course of our initial experiments. The initial acoustic training set was drawn from a selection of 40 shows broadcast during the period February 2, 1999 through March 27, with two additional shows from April 30 and May 4. The total amount of transcribed speech in these shows totalled 12.8 hours. In all the experiments we conducted, broadcasts were segmented into individual utterances using boundary information taken from the annotations.

2.2.1 Czech Monolingual ASR Experiments

Our first Czech VOA test set consisted of broadcasts from February 15, March 13, and May 3 1999 totalling 1.0 hours of transcribed speech. The speech feature parameterization employed in training and test are mel-frequency cepstra, including both delta and delta-delta sub-features; cepstral mean subtraction is applied to all features on a per utterance basis. Waveform files were down-sampled to 16KHz. All models were trained using the HTK 'incremental build' procedures beginning from a flat-start.

The first CZ VOA experiments tested a 12 mixture, state clustered, cross-word triphone system. The Word Error Rate (WER) of this system was 27.1% (see Table 2), which we considered to be reasonable given the 6.7% OOV rate. We then investigated the performance of a 20-mixture monophone system. Our hope was that we would be able to use monophone systems in our experiments since this would simplify sharing models across languages, and we did indeed find that this monophone system performed comparably to the triphone system.

Czech VOA Test Set 1				
Speaker Identity	Gender	WER	Gender	WER
Unknown	F	24.05	F	22.47
	F	17.26	F	25.91
	F	31.67	F	33.55
	M	25.14	M	26.61
	M	23.91		
Anchors	M	16.28	M	45.54
	F	25.53	F	24.82
Overall	27.6%			

Table 3: Word Error Rate by Speaker for the 10 Hour Czech VOA 12-Mixture Monophone System. High overall accuracy is not due to a few well-recognized individuals.

Motivated by these results, we decided to study monophone performance as a function of reduced training set size. We expected performance to decrease with large reductions in the training set, however, as detailed in Table 2, we found performance to be largely insensitive to both model complexity and the amount of training data. This experiment is evidence that there is too much self-similarity in this particular training and test set combination for it to be useful for ASR experiments. This is further confirmed by testing the 1.0 hour monophone models on the data used to train them - only a 4% absolute difference is performance is observed between the training and test sets. This is contrary to the expectation that performance on the training data should be much greater than on a fair test set.

2.2.2 Analysis of Czech VOA Performance

We studied the performance by each speaker to see if this self-similarity is due to speech from 'news anchors' dominating both the training and test sets. However, as shown in Table 3, we found that performance varied widely over all the speakers in the test set. In fact, the worst performing speaker was one of the anchors.

We next considered whether the training and test sets were similar because they were collected within a relatively short time, since stories 'in the news' contained frequently occurring words and phrases that might end up being unusually well-trained. We defined another test set to be Czech VOA recordings from the week of May 21, 1999. This ensured a separation of several weeks between the test set and the bulk of the training data. However, as show in Table 4, performance was only slightly worse on these later shows than on the earlier test set. We concluded from this that the similarity between the test and training set was not simply due to their being recorded at about the same time.

2.2.3 Cross-Domain Experiments

As shown in Table 5, we found several surprising results in experiments with our read speech systems and data sets. Most surprisingly, a read speech system trained on 1.0 hour of speech performs significantly better on the Czech VOA data than it does on read speech. Conversely, the

Czech VOA Test Set 2			
Date	WER	Date	WER
05/21/99	36.14	05/22/99	33.43
05/23/99	36.44	05/24/99	37.18
05/25/99	33.34	05/26/99	39.43
05/27/99	37.54	05/28/99	32.89
Overall	35.7%		

Table 4: Daily Word Error Rate of a 1.0 Hour, 20-Mixture Monophone Czech VOA System. This test set was recorded several weeks after the acoustic training set, and performance is only slightly less than found on the earlier test set.

Models	WER (%)		
	CUCFN	VOA Set 1	VOA Set 2
10.0 hr CZ VOA 12-Mixture Monophone	68.0	27.6	
1.0 hr CZ VOA 20-Mixture Monophone	66.1	30.2	28.8
1.0 hr CUCFN 20-Mixture Monophone	47.3		35.7

Table 5: Training and Testing on Czech VOA Broadcasts and CUCFN Read Speech. The read speech CUCFN models perform better across domains than the Czech broadcast VOA models.

Czech VOA systems perform much worse on the read speech. This suggests that the Czech VOA data is more like read speech than much of the speech actually in the read speech corpus.

We were curious whether this self-similarity is a general property of VOA speech, or whether we merely were unlucky with our Czech broadcasts. Juan Huerta performed a quick experiment using the CMU Sphinx III Spanish Hub V broadcast news system with a bigram derived from newspaper stories. Performance of acoustic models trained on 1.0 hour of Spanish VOA speech was measured on 30 minute test sets of Spanish VOA test broadcasts and Spanish language ECO news broadcasts from Mexico. The results given in Table 6 are similar to those we encountered in the Czech VOA data: the system trained on Spanish VOA data performs well on other Spanish VOA data, but generalizes poorly to other Spanish broadcast data.

Our concerns about the general nature of VOA speech prompted us to record the aforementioned news programs broadcast by Český rozhlas 1 - Radiožurnál. For convenience this test set was called CRo1. Unlike the Czech VOA data, performance on this test set varied as expected with reductions

	30 Minute Test Set WER (%)	
Models	Spanish VOA	ECO
1.0 hr Spanish VOA Monophones	22.5	51.7

Table 6: Testing Spanish Broadcast News with 1.0 Hour Spanish VOA Models. Spanish VOA models generalize poorly to ECO newsbroadcast data.

Czech CRo1 Test Set	
Models	WER (%)
13 hour CUCFN 3886 state xword triphone†	42.0
10 hour CUCFN 12-mixture monophone†	55.5
10 hour CUCFN 20-mixture monophone	54.8
10 hour Czech VOA 12-mixture monophone	58.0
1 hour CUCFN 20-mixture monophone	58.6

Table 7: Word Error Rate for the CRo1 News Broadcasts. Performance varies significantly with variations in training set size and model complexity. Experiments marked † were conducted with 22KHz sampled training and test data.

	CUCFN	Czech VOA-1	Czech VOA-2	CRo1
females/males	7/7	8/5	26/6	39/61
females/males utterances	700/699	257/147	836/281	345/466
planned/spontaneous utterances (%)	100/0	100/0	100/0	45/55
studio/outside (%)	100/0	100/0	100/0	85/15
total utterances	1399	404	1117	811
speakers	14	13	32	46
duration (minutes)		60	150	99

Table 8: Characteristics of the Test Set Partitions.

in training data and model complexity. In particular, we observe an absolute reduction of 17% in word accuracy by going from a 13 hour cross-word triphone system to a 1.0 hour monophone system.

2.3 Training and Test Set Definitions

Our initial experiments indicated that our Czech VOA collection is quite well-behaved, in that using only small amounts of acoustic training data yields fairly good word accuracy. Although our experiments do not explain why this is so, this lack of variability makes the Czech VOA unsuitable for use simultaneously in training and testing. A related problem is that since this VOA appears to be similar only to itself and very different from other speech, by studying it we risk obtaining results that are not valid in general. For these reasons we extended the test and training set to include Czech speech outside VOA.

We decided to fix the 1.0 hour CUCFN read speech training set as the in-language acoustic training set. The main test set is the second Czech VOA Test (chosen for its larger size). The CUCFN test set and the CRo1 collection serve as secondary, and more difficult, test sets. In this way we avoid using the Czech VOA data simultaneously in training and testing.

This provides a realistic and interesting training scenario that involves cross-domain as well as multilingual factors. Overall characteristics of the test set partitions are provided in Table 8.

The baseline recognition performance is summarized in Table 9. These are the ‘numbers to

Word Error Rate (%)		
CUCFN	Czech VOA	CRo1
47.3	35.7	58.6

Table 9: Performance of a 1.0 Hour CUCFN 20-Mixture Monophone System.

beat': any experiment that improves over these results by using only the 1.0 hour CUCFN acoustic training set and data borrowed from other languages will be considered a success.

2.4 A Cautionary Note

These experiments with Czech VOA are reported to emphasize that language is just one characteristic of speech and that other conditions, such as speaking style, are significant factors in ASR performance. It is therefore critically important to obtain diverse training and test sets for multilingual experiments. It is also important that results of limited domain experiments, such as training and testing with data from the same news programs, be interpreted cautiously since performance may not carry over to more diverse domains.

3 Cross-Lingual Phone Mappings

3.1 Knowledge-Based Phone Mappings

In some applications, it is highly desirable to be able to develop speech recognition systems without using any acoustic training data. In such situations, borrowing models from other languages for which speech recognition technology is well-developed is extremely attractive. The approaches presented here are referred to as knowledge-based because they exploit linguistic knowledge of the languages and their phoneme inventories, and because they have not been retrained on any target language acoustic data.

The goals of the work presented in this section were two-fold:

(1) to develop baseline performance for target language systems developed from our existing source-language monolingual systems, and

(2) to minimize the amount of target language training data required by developing effective techniques for model combination from the source languages.

In our case, our source languages were English (EN), Spanish (ES), Mandarin Chinese (MD), and Russian (RN). The target language was Czech (CZ). As previously mentioned, these languages were chosen primarily because of the existence of large amounts of data from a similar domain: Broadcast News (BN). Russian was the only exception. Though the Russian data consisted of read speech, Russian is acoustically very close to Czech, and hence provided another important contrastive data point.

Through the course of our work this summer, we established some important bounds on performance that provide a good deal of perspective on the problem. Systems that use no target language training data generally performed in the range of 80% WER; systems allowed some access to target

IPA Conversion Chart For Consonants

Manner	Language	Place of Articulation																						N
		Bilabial		Labio-dental		Dental		Alveolar		Post-alveolar		Retroflex		Palatal		Velar		Uvular		Pharyn-geal		Glottal		
Plosive	English	p	b					t	d							k	g							6
	Spanish	p	b			t	d									k	g							6
	Mandarin	p	b					t	d							k	g							6
	Russian	p P	b B					t T	d D							k	g G							11
	Czech	p	b					t	d					tj	dj	k	g							8
Nasal	English		m						n								nx							3
	Spanish		m						n						gn									3
	Mandarin		m						n								N							3
	Russian		m M						n N															4
	Czech		m						n						nj		ng							4

Figure 3: An IPA description of the consonant portions of the phone sets used in our experiments.

language data to determine phone-level or state-level mappings performed in the range of a WER of 55%; systems allowed some amount of retraining or systems built from large amounts of target language data achieved performance in the range of 30%.

A third goal of this work was to attempt to close the gap between the knowledge-based systems operating at a WER of 80% and the data-driven systems operating in the range of 55% WER. We attempted to do this only by utilizing a priori information about the proximity of the source languages to the target language, and developing intelligent methods of model combination for the source languages.

3.1.1 Monolingual Cross-Language Baselines

Our first set of baselines involved a simple mapping experiment in which phones from the Czech target language were mapped to their nearest neighbor in a single source language using a similarity measure based on feature-based descriptions of the phones. This is a manual procedure that leverages extensive knowledge of acoustic phonetics [4]. Our approach involved first describing the phones in both the source and target languages in terms of their articulatory positions, a process that leads to a description of the sounds using the International Phonetic Alphabet (IPA) [23]. A portion of this analysis is shown in Figure 3. A complete inventory, along with several related resources, can be found in [21]; an example of such a description for a phone is shown in Table 10. The advantage of this approach is that all languages can, in theory, be represented within the same system. Other advantages include an ability to cluster phones for context-dependent representations using approaches based on acoustic phonetic similarity analogous to what is used in language-dependent recognition.

We next determined the proximity of a sound in the target language to a sound in the source language using this representation, and developed an associated symbol-to-symbol mapping. Exam-

Phone	Description
S	UNVOICED ALVEOLAR FRICATIVE
F	UNVOICED LABIO-DENTAL FRICATIVE
II	HIGH FRONT UNROUND LONG VOWEL

Table 10: An example of a representation of a phone in terms of articulatory positions.

	VOA-1	VOA-2
Source Language	WER(%)	WER(%)
Czech	27.6	23.6
Russian	65.2	60.8
Spanish	79.3	71.7
English	80.9	75.5
Mandarin	91.1	88.7

Table 11: Baseline monolingual system performance.

ples of such mappings are given in Figure 4. While it was possible to achieve reasonable mappings for each language, there are significant variations in the level of detail used in the source language phonetic inventories. Spanish, for example, only used 25 phones, while Russian used 44 phones. Since optimization of the source language systems was beyond the scope of this project, we did not spend a lot of time fine-tuning the phonetic mappings, or designing phone inventories particularly suited to our task. Instead, as a starting point, we used off-the-shelf state-of-the-art existing BN systems. We proceeded to use these mappings to obtain baseline performance of a Czech Broadcast News (CZBN) recognition system using acoustic models from the source languages derived from these mappings. The procedure was quite simple: represent each phone symbol in the Czech lexicon using a corresponding source language phone located from these mappings. The performance of systems constructed in this manner is given in Table 11. Overall, we observe that performance is poor - in the range of 80%WER. It was a great surprise to observe that the Russian acoustic models, though they were trained on read speech, were a close match to the CZBN data, especially considering the differences in microphones, speaking style, and speaking rates. As we subsequently found out, the CZBN data is relatively well-articulated, and fairly easy to recognize at a nominal level of performance. We also observed from these experiments that performance for English and Spanish was comparable, and performance for Mandarin lags the other systems.

Upon observing this degradation of performance for Mandarin, we hypothesized that the phone mapping was a major source of error. Hence, we evaluated four different phone mappings. These mappings are summarized in Figure 4, and explained in greater detail in Figure 5. The performance on the VOA-1 evaluation for each of these mappings is given in Table 12. Though we achieved a very minor improvement in performance (a 0.8% absolute gain), we can conclude that performance is not extremely sensitive to the quality of the manual phone mapping at the level of performance our system was operating at. Hence, we turned our attention to methods for combining multiple languages into a single system.

Czech		English		Spanish		Mandarin				Russian	
v1	Example	v1	v2	v1	v2	v1	v2	v3	v4	v1	v2
a	(ah:2) but	ah	ah	a	a	@	a	@	@	@	@
aa	(aax:2) father	aax	aax	a	a	@	@	@	@	aa	aa
aw	(aw:1) down	aw	aw	a	a u	&	e	&	&	a	a
b	(b:1) blue	b	b	b	b	b	b	b	b	b	b
c	(t s:3) Yeltsin	t	t s	t	t s	Z	c	c	c	c	c
ch	(ch:1) chip	ch	ch	ch	ch	q	C	q	q	chj	chj
d	(d:1) dark	d	d	d	d	d	d	d	d	d	d
dj	(d y:4) due	d	d y	d	d y	d	d	d	d y	dj	dj
e	(eh:1) bet	eh	eh	e	e	E	>	>	>	e	e
ee	(eh:3) long of e	eh	eh	e	e	E	E	E	E	ee	ee
f	(f:1) fix	f	f	f	f	f	f	f	f	f	f
g	(g:1) global	g	g	g	g	g	g	g	g	g	g
h	(hh:2) ahead	hh	hh	j	j	h	x	h	h	x	x
i	(ih:1) hit	ih	ih	i	i	I	i	i	i	ih	ih
ii	(iy:1) he	iy	iy	i	i	i	I	i	i	i	i
j	(y:1) yes	y	y	y	y	r	y	y	y	j	j
k	(k:2) key	k	k	k	k	k	k	k	k	k	k
l	(l:1) loom	l	l	l	l	l	l	l	l	l	l
m	(m:1) meet	m	m	m	m	m	m	m	m	m	m
n	(n:1) noun	n	n	n	n	n	n	n	n	n	n
ng	(nx:1) hang	nx	nx	n	n g	N	N	N	N	nj	nj
nj	(n y:4) new	n	n y	gn	gn	N	N	N	N y	nj	nj
o	(aa:2) hot	aa	aa	o	o	o	o	o	o	o	o
ow	(ow:1) low	ow	ow	o	o	o	o	o	o	o	o
p	(p:2) power	p	p	p	p	p	p	p	p	p	p
r	(r:4) Rome	r	r	r	r	r	r	r	r	r	r
rsh	(r sh:5) n/a	r	r sh	r	r ch	s	r	r	r s	sh	r sh
rzh	(r zh:5) n/a	r	r zh	r	r ll	s	r	r	r S	zh	r zh
s	(s:1) son	s	s	s	s	s	s	s	s	s	s
sh	(sh:1) shape	sh	sh	ch	ch	S	S	S	S	sh	sh
t	(t:2) tomado	t	t	t	t	t	t	t	t	t	t
tj	(t y:4) statue	t	t y	t	t y	t	t	t	t y	tj	tj
u	(uh:2) could	uh	uh	u	u	u	u	u	u	u	u
uu	(uw:1) who	uw	uw	u	u	u	u	u	u	u	u
v	(v:1) victory	v	v	v	v	f	w	w	w	v	v
x	(k hh:3) Loch	k	k hh	j	j	h	h	h	k h	x	x
z	(z:1) zoo	z	z	s	s	s	s	s	s	z	z
zh	(zh:1) pleasure	zh	zh	ll	ll	S	S	S	S	zh	zh

Figure 4: Phone mappings from Czech to our four source languages using an IPA-based feature representation. For some languages, several possible mappings are shown to demonstrate that there is some amount of ambiguity in these mappings.

Czech		Mandarin				
v1	**Example - CZ**	**IPA or Description**	**v1**	**v2**	**v3**	**v4**
a	(ah:2) but	front allophone of /a/	@	a	@	@
aa	(aax:2) father	front allophone of /a/	@	@	@	@
aw	(aw:1) down	schwa, mid central unrounded	&	e	&	&
b	(b:1) blue	p	b	b	b	b
c	(t s:3) Yeltsin	aspirated dental affricate ts	Z	c	c	c
ch	(ch:1) chip	aspirated palatal affricate	q	C	q	q
d	(d:1) dark	t	d	d	d	d
dj	(d y:4) due	t	d	d	d	d y
e	(eh:1) bet	e	E	>	>	>
ee	(eh:3) long of e	lower-mid front unrounded	E	E	E	E
f	(f:1) fix	f	f	f	f	f
g	(g:1) global	k	g	g	g	g
h	(hh:2) ahead	laryngeal or velar fricative	h	x	h	h
i	(ih:1) hit	barred i	I	i	i	i
ii	(iy:1) he	i	i	I	i	i
j	(y:1) yes	retroflex r	r	y	y	y
k	(k:2) key	aspirated k	k	k	k	k
l	(l:1) loom	l	l	l	l	l
m	(m:1) meet	m	m	m	m	m
n	(n:1) noun	n	n	n	n	n
ng	(nx:1) hang	velar nasal	N	N	N	N
nj	(n y:4) new	velar nasal	N	N	N	N y
o	(aa:2) hot	mid back round	o	o	o	o
ow	(ow:1) low	mid back round	o	o	o	o
p	(p:2) power	aspirated p	p	p	p	p
r	(r:4) Rome	retroflex r	r	r	r	r
rsh	(r sh:5) n/a	s	s	r	r	r s
rzh	(r zh:5) n/a	retroflex affricate	zh	r	r	r S
s	(s:1) son	s	s	s	s	s
sh	(sh:1) shape	voiceless retroflex fricative	S	S	S	S
t	(t:2) tornado	t	t	t	t	t
tj	(t y:4) statue	t	t	t	t	t y
u	(uh:2) could	high back rounded	u	u	u	u
uu	(uw:1) who	high back rounded	u	u	u	u
v	(v:1) victory	f	f	w	w	w
x	(k hh:3) Loch	laryngeal or velar fricative	h	h	h	k h
z	(z:1) zoo	dental affricate (ts)	s	s	s	s
zh	(zh:1) pleasure	retroflex affricate	S	S	S	S

Figure 5: Four variations of Czech to Mandarin phone mappings that were explored to diagnose the poor performance of the Mandarin system.

Source Language	VOA-1 WER(%)
Mandarin - v1	91.1
Mandarin - v2	93.7
Mandarin - v3	90.1
Mandarin - v4	89.3

Table 12: Several approaches to Mandarin phone mappings were explored in an effort to improve performance. As we can see, performance was not greatly influenced by the nature of the manual phone mapping.

Source Language	VOA-1 WER(%)
Spanish	79.3
Selective	77.7

Table 13: A comparison of performance using a Spanish-only system, and a system involving a mixture of mappings from three source languages. Though there is a modest improvement in performance, the improvement was not nearly as significant as we had hoped.

3.1.2 Multilingual Phone Mappings

It was evident that a single source language did not provide optimal coverage of Czech. Therefore, it was natural to explore a mapping that involved phones from all source languages based on proximity in the IPA table. Since Russian was clearly acoustically closer to Czech than any of the other source languages, we excluded Russian from the set of source languages for this experiment, so that it would not mask any trends in our knowledge-based systems that might surface. This was somewhat of a cheating experiment in that we began with our best models - the Spanish system. We then replaced phones in cases where other languages appeared to have a closer match. We did include Mandarin even though we had suspicions about the quality of the models.

A summary of the resulting mapping is shown in Figure 6, and the associated performance is given in Table 13 Though we achieved modest improvements in performance (1.6% absolute WER), we did not achieve performance comparable to data-driven mapping methods discussed later.

Our next attempt to understand the deficiencies of the knowledge-based system was to explore a series of experiments in which the recognition system was allowed to choose the best combination of phones at runtime (rather than fixing these via a mapping prior to recognition). First, we explored a parallel pronunciation approach [22] in which each item in the lexicon was allowed to be represented as a sequence of phones from a single language. This was implemented using pronunciation networks, and is summarized in Figure 7.

Unfortunately, this approach resulted in a slightly degraded performance, as shown in Table 14. This result was somewhat discouraging, since we had hoped that the additional degrees of freedom would offset any systematic acoustic bias between the two domains. The next obvious thing to try was to allow the recognition system to mix and match phones from all source languages. This approach, referred to as a multiphone approach, is also summarized in Figure 7. The corresponding performance is given in Table 14. The multiphone approach was an attempt to let the recognizer find the best realization of a phone, rather than fixing this based on a priori linguistic knowledge. We can see that a minor improvement in performance over the parallel pronunciation system was

Spanish	Selective
a	a_ES
aa	@_MD
aw	aw_EN
b	b_ES
c	t_ES
ch	ch_ES
d	d_ES
dj	d_ES
e	e_ES
ee	E_MD
f	f_ES
g	g_ES
h	j_ES
i	ih_EN
ii	i_ES
j	y_ES
k	k_ES
l	l_ES
m	m_ES
n	n_ES
ng	nx_EN
nj	gn_ES
o	o_MD
ow	o_ES
p	p_ES
r	r_ES
rsh	r_ES
rzh	r_ES
s	s_ES
sh	sh_EN
t	t_ES
tj	t_ES
u	u_ES
uu	u_ES
v	v_ES
x	j_ES
z	z_EN
zh	zh_EN

Figure 6: A selective phone mapping that uses phones from three source languages to model Czech.

achieved, as expected. However, overall performance is still below the best monolingual system, and far below the Russian system shown in Table 14. Again, this was a discouraging result.

Source Language	VOA-1 WER (%)
Czech (CZ)	27.6
Russian (RN)	65.2
Spanish (ES)	79.3
English (EN)	80.9
Mandarin (MD)	91.1
Parallel Prons.	83.0
Multi-Phone Prons.	80.1

Table 14: Performance for two approaches as mixing phones from multiple languages. The parallel pronunciation approach constrains words to use phones from the same language. The multi-phone approach allows the system to mix and match phones from any language. As we can see, the latter system resulted in a minor improvement in performance, but did not exceed the performance of the baseline system.

3.2 Automatic Generation of Phone and Sub-Phonetic Mappings

The purpose of this work was to generate automatically a set of phonetic mappings from a pool of well trained languages (the source languages) to a single language (the target language) where there is little data to train a large set of acoustic models. To address this problem, we developed a methodology to derive automatically these mappings both at the phonetic and at the subphonetic levels.

Several uses can be made of these mappings once they are obtained: for example, they can be used to assemble acoustic models in the target language using components obtained from the source languages acoustic model inventories, they can be used to derive initialization models for further adaptation or retraining methods, or they can be used to borrow data for acoustic modeling.

In the following sections we will describe the *Confusion Matrix* approach to finding cross-lingual mappings as well as the criteria we explored in our investigations, we will describe how we mixed models coming from several source models, and we finally present our experimental results. In the final subsection we present directions for further work.

3.2.1 The Confusion Matrix Approach to Cross-Lingual Phonetic Similarities

Figure 8 below presents a segment of speech in the target language. Let X denote the phonetic segmentation and labels assigned to the utterance segment. These segments can be obtained through human intervention or automatically, by force-aligning the segment transcriptions. Let Y denote the output of a phonetic recognition of the same speech segment in a given source language. The phonemes that will appear in this string Y are not phones that belong in the phonetic inventory of the language in which the sentence was uttered, however, for a sufficiently long segment of speech the co-occurrences between phone in the string X and phones in the string Y will reflect the similarity, at least from the recognizer's point of view, between phones in both languages.

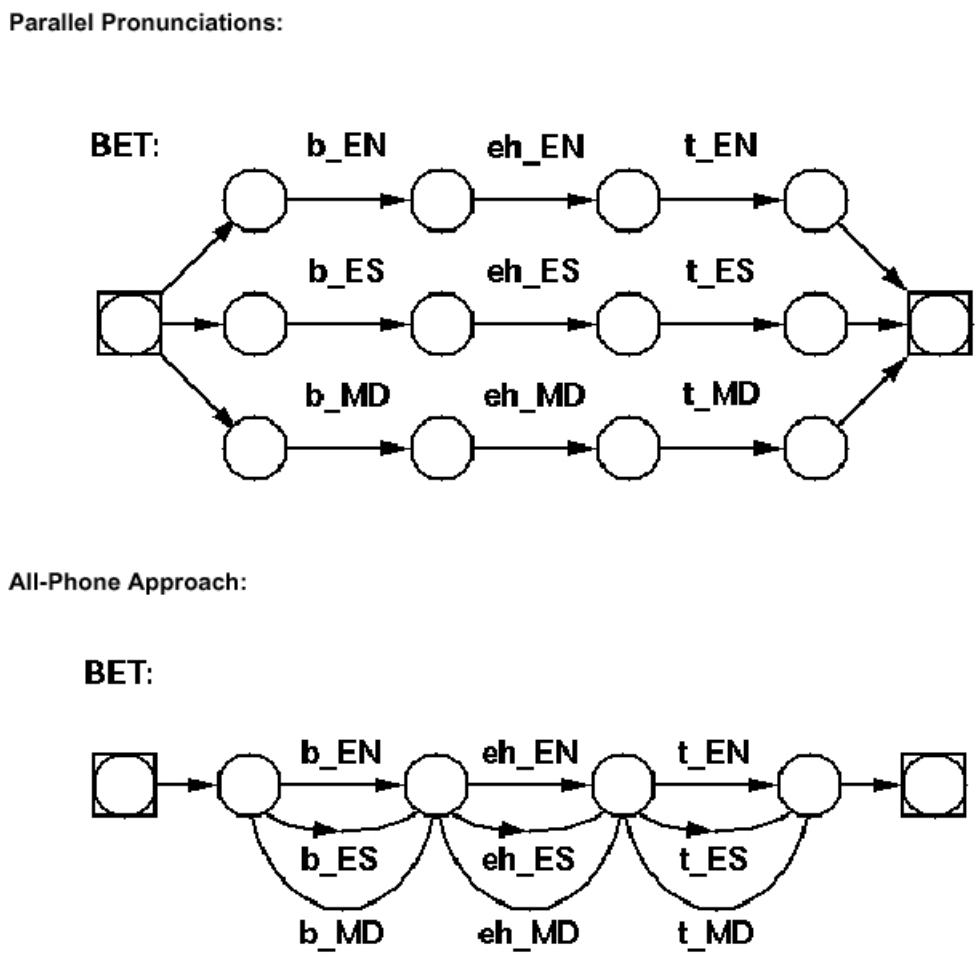

Figure 7: Two approaches to mixing multiple source language acoustic models without the use of acoustic training data. In the first approach, the recognizer is constrained at the lexical level to phones from a single source language to represent a word. In the second approach, the recognizer can mix and match phones from any source language.

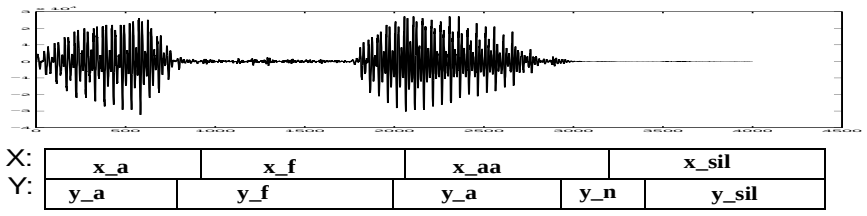

Figure 8: Speech Segment with Automatic Labelling Compared to the Reference Transcription.

Once a criterion for co-occurrence between two phonetic labelings of the acoustic segments is defined (e.g., a minimum number of overlapping frames, etc.), we can arrange the phones of the source language and target language into a matrix that contains the counts of co-occurrences between the n^{th} and k^{th} phones of the source and target languages, respectively, in the (n, k) entry of the matrix. This matrix of co-occurrences is the confusion matrix.

After the confusion matrix between the phones of two languages is obtained, we derive mappings from this matrix. Given a source phone (in the n^{th} row), we would like to select the phone in the target language that best matches it (i.e., choose the best matching k^{th} column). To do this we can simply choose the column with the highest count. A better method takes into account the number of times the k^{th} source language phone was hypothesized by dividing the counts of the bin (n, k) by the accumulated counts of the column k.

We extended this technique to the state level, motivated by our intuition that some phones in the target language seemed hard to match. To obtain the subphonetic mapping, we broke each HMM in the source and target language into its conforming states and derived an HMM from each of these states. Using these new, sub-phone HMMs we constructed a new confusion matrix. As expected, we found that some of these hard-to-match target language phones were modeled by assembling new models from phonetic subunits from other languages.

We observed that when many states and phones from various languages were competing to represent any given target model, several models seemed to give high counts and thus be close candidates for a reasonable match. We explored the possibility of including several of these best matching candidates by combining the Gaussian models in their mixtures after weighting them accordingly. We established the weights used in this state combination to be proportional to the normalized number of counts corresponding to the map. Table 15 shows an example of the best 3 matches between ENglish, MAndarin and SPanish, to represent the 3 states of the Czech vowel HMM *aa* and their corresponding weights.

Experimental Results

Table 16 below shows recognition experiments we conducted using mappings derived from confusion matrices. Column one refers to the languages employed to derive the source inventory of

State	Source Language HMM State Matches with Scores		
aa_1	en_ay_1 / 0.12	en_aa_1 / 0.11	ma_a_1 / 0.09
aa_2	en_aw_2 / 0.19	en_aa_2 / 0.17	en_aax_2 / 0.12
aa_3	en_ay_2 / 0.28	en_aw_3 / 0.23	en_ao_3 / 0.06

Table 15: Example Cross-Lingual State-Level Mappings Ordered by Confusion Score; *e.g.* Czech data aligned as state 1 of the Czech HMM *aa* is best modeled by state 1 of the English HMM *ay*.

Source(s)	Mapping	Method	n-best	WER(%)
Czech	Phone	baseline		38.01
EN	Phone	manual	1	>80
SP	Phone	manual	1	>80
EN	Phone	matrix	1	68.31
SP	Phone	matrix	1	68.67
EN	State	matrix	1	64.75
SP	State	matrix	1	70.03
MA	State	matrix	1	79.69
EN+SP+MA	State	matrix	1	62.28
EN+SP+MA	State	matrix	3	55.77
EN+SP+MA	State	matrix-2	3	54.38
EN+SP+MA	State	matrix-2 + LAC	3	48.80

Table 16: Recognition Performance Using Automatic Methods. The Czech baseline system and the knowledge-based system are included for comparison.

phonetic units; column two refers to the type of phonetic unit employed (i.e., phone or state); column three refers to the method employed to determine the mapping (i.e., manual, matrix based, or matrix based with Language Adaptive Clustering); column four refers to the number of best matching source language phonetic units from the confusion matrix (ranked by their normalized counts) that were used to assemble the target language phonetic unit; column five shows the corresponding Word Error Rate. The first line presents our baseline, in which monophone Czech models trained on approximately one hour of Broadcast News data are used to recognize a separate episode of Broadcast News data. The following two lines show the recognition results obtained using a typical human based mapping from the source languages English (E) and Spanish (S) respectively. When mappings are obtained using the matrix approach the word error rate drops below 70% (lines 4 and 5). State level mappings help reduce the error rate of the English mappings. The best results are obtained when three source languages are included (English, Spanish and Mandarin Chinese) and state mappings are obtained for both one state to one state mapping (line 9) and best three states to one Czech state (lines 10 and 11). The best number is below 55% WER. The difference between line 10 and line 11 is due to the presence (line 11) or absence (line 10) of count normalization of the columns in the confusion matrix.

Acoustic Training Using Automatic Mappings

We used the best mapping described above (last row in the table), to derive Czech models from the source languages. Using these models as starting points for further iterations of Baum Welch training did not seem to give any noticeable advantage over training from a flat-start. However, the performance of the best mapping could be improved substantially by performing phonetically motivated cepstral mean normalization as described in the next section: the WER was brought down in this way to 48.8%.

3.3 Language Adaptive Clustering

Here we examine a novel method to find cross-lingual phone mappings using a modified version of vector quantization [15]. The key feature here is that we allow the source language data to be acted upon by language-specific transformations, in the hope that these transformations will model differences in recording conditions as well as differences in the pronunciation patterns of the languages. We stress that these cross-lingual transformations are useful not only in improving clustering; they can be applied directly to models and data to remove cross-language variability.

Vector-quantization, also called K-Means clustering, is well-known, as is the LBG algorithm used to obtain VQ codebooks. Given a set of data vectors, the goal is to find a finite set of centroids, or codewords, that will be used to represent the data so that the total distortion is minimized. For a collection of data vectors $\{x_i\}$, the minimum distortion vector quantizer attempts to find a set of codewords or centroids, $C = \{C_k\}$, to minimize the design objective

$$\sum_i \min_{C_k \in C} |C_k - x_i|^2.$$

The preimages of codewords, i.e the vectors that share a common codeword, are known as clusters. Intuitively, the idea is that each cluster contains vectors that are "close together". The LBG algorithm is an iterative procedure that alternates between recomputing centroids and recomputing clusters. This is combined with "splitting": to increase the number of clusters, each centroid is perturbed to create a new codeword.

The outcome of this procedure can be represented as a binary tree where each node at the k^{th} stage represents a cluster, and its children are the clusters into which it has been split. In the case of multilingual acoustic data, if vectors of a Czech phone and vectors of, say, a Spanish phone end up in the same cluster, then that Spanish phone should be mapped to the Czech phone.

For our application, we introduce two modifications. The VQ training technique does not ensure that all the samples of a given phone belong to the same cluster, so we modify the step of the algorithm in which clusters are recomputed. Rather than assign a codeword to each vector individually, we find the codeword that best describes all instances of a phone in each language.

The VQ procedure could be run with this modification to cluster phones across languages. However, we observed (see Table 18) that the differences between languages dominate differences between phones: by the second level of splitting, the clusters become extremely language-specific, i.e. each cluster contains mostly phones from only one language. This defeats the purpose of the procedure, which was to find phone clusters across languages.

Our second modification is to introduce cross-language transformations to eliminate these broad cross-language differences. Specifically, for each cluster and source language, we allow a member of a family T of transformations to act on that language's data in the cluster.

We have defined a modified VQ objective function that incorporates these transformations. Let $x_i^{p,l}$ denote the i^{th} sample of phone p from language l. The modified objective function that describes a clustering Czech, Spanish, Russian, Mandarin, and English phone is

$$\sum_p \min_{C_k \in C} \sum_i |C_k - x_i^{p,cz}|^2 + \sum_{l \in \{ma,sp,ru,en\}} \min_{C_k \in C} \min_{T^{p,l} \in T} \sum_i |C_k - T^{p,l}(x_i^{p,l})|^2.$$

Note that no transformation is applied to the target language data. In this way we hope to find the best target language codewords along with mappings from the source language data to the target language codewords.

We considered two possible families of transformations: rotations

$$T^{p,l}(x_i^{p,l}) = W^{p,l} x_i^{p,l}$$

and additive shifts

$$T^{p,l}(x_i^{p,l}) = x_i^{p,l} + b^{p,l}.$$

In either case, the LBG algorithm was modified as follows.

Language Adaptive LBG Clustering

Given a set of codewords $\{C_k\}_{k=1}^K$ and a set of transformations $\{T^{p,l}\}$, one iteration of the modified LBG procedure is summarized as follows

Find codewords for all phones in each language :	$r^{p,l}$	$=$	$\text{argmin}_k \sum_i \lvert C_k - T^{p,l}(x_i^{p,l})\rvert^2$
Reestimate all centroids $C_k,\ k = 1, \ldots, K$:	X_k	$=$	$\{x_i^{p,l} : r^{p,l} = k\}, \quad c_k = \overline{X_k}$
	C_k	$\leftarrow$	c_k
Reestimate transforms for source language phones :	k	$=$	$r^{p,l}$
	$T^{p,l}$	$\leftarrow$	$\text{argmin}_{t \in T} \sum_i \lvert c_k - t(x_i^{p,l})\rvert^2$

As the procedure shows, after recomputing the centroids and clusters, we add another step, in which the transformation is recomputed to be the best possible member of T given the new centroids and clusters.

In the case of rotations, we reestimate these transformations by arranging all the samples of a phone p from language l into a matrix $X = [x_i^{p,l}]$, and with the new centroid c_k fixed, we use a procedure analogous to least-squares regression to find the optimum transformation

$$W^{p,l} = c_k \ 1' \ X' \ (XX')^{-1}$$

where $'$ denotes transposition. In the case of shifts, the transformation is simpler:

$$b^{p,l} = c_k - \overline{x^{p,l}}$$

where the last term is the mean vector of the samples from phone p of language l.

	WER(%)
mapping alone	86.4
mapping + transformation	71.6%

Table 17: WER of Cross-Language Phone Clustering using Additive Shift.

Results

For these experiments, the target language was Czech and the source languages were Russian, Spanish, and Mandarin. For each language, we used about 1.0 hours of acoustic data in the form of 39-dimensional mel-frequency cepstral coefficients with their first and second differences. The data was labelled at the phone level via automatic alignment. The Czech data was CUCFN speech aligned by the 1.0 hour monophone models.

To establish some initial baselines, we ran the VQ algorithm without any language-specific transformations. Clusters were split successively until they contained only one Czech phone. The results are shown in Table 18. It can be seen that the clusters are predominantly multi-lingual, and even worse, many Czech phones can be found in clusters alone.

We next tried to force phones to map across languages, by looking for the four closest Spanish and Mandarin phones for each Czech phone. Results are shown in Table 20. As can be seen, the mappings from Table 19 do not look reasonable, and it is therefore necessary to apply some kind of transformation to the non-Czech data.

Next we moved on to the actual experiments involving transformations. In the first case, using rotations, there was a disappointing lack of cross-language clustering; phones still tended to stay with others of the same language. This is shown in Table 20. In the second case, using an additive shift, there was much more cross-language clustering. The next step was to use the results of this latter cross-language clustering to generate a mapping. The results of this are shown in Table 21. We tested this mapping in two ways. First, we simply ran a recognition experiment on the Czech data by replacing each Czech phone with its source language phone, and the results are given in Table 17. This gave relatively poor results - the word error rate was 86.4%. However, this approach does not take into account the additive shifts that were used while clustering. So the next experiment was to apply the appropriate additive shifts to the means of the Gaussian mixtures for each source language phone's HMM. This significantly reduced word error rate to 71.6%, which is comparable to the other automatic phone-level methods.

Applications to Other Methods

Another point worth mentioning in connection with cross-language transformations is that they can also be used in conjunction with other methods of generating mappings. Specifically, given a mapping, we could try to improve its performance by applying additive shifts to the source language HMMs as above. We tried doing this with the mappings generated by using a state-level confusion matrix. This resulted in a significant drop in error rate from 54.5% to 48.8%.

Cluster	Phone : Source Language
1	g:CZ e:SP l:SP d:SP m:SP r:SP n:SP i:SP g:SP ll:SP gn:SP n:MA i:MA
2	ee:CZ rr:SP at:MA d:MA
3	sil:SP k:SP t:SP a:SP p:SP
4	sp:SP t:MA g:MA p:MA
5	s:SP ch:SP x:MA j:MA q:MA
6	x CZ j:SP s:MA c:MA
7	l:CZ o:SP u:SP
8	f:CZ f:SP h:MA z:MA f:MA k:MA
9	rzh:CZ
10	rshCZ
11	aa:CZ a:CZ
12	a:MA
13	ow:CZ aw:CZ o:MA
14	o:CZ amp:MA gt:MA
15	uu:CZ w:MA u:MA
16	u:CZ
17	ng:CZ b:SP
18	h:CZ
19	v:CZ b:CZ v:SP
20	b:MA
21	r:CZ
22	d:CZ y:SP
23	e:CZ j:CZ i:CZ e:MA
24	y:MA %:MA r:MA l:MA
25	ii:CZ
26	m:CZ dj:CZ m:MA
27	n:CZ
28	nj:CZ
29	z:CZ x:SP
30	s:CZ c:CZ
31	p:CZ t:CZ
32	k:CZ sp:MA
33	sil:MA
34	sil:CZ sp:CZ
35	zh:CZ
36	tj:CZ
37	sh:CZ
38	ch:CZ

Table 18: VQ Phone Clusters Found Without Cross-Language Transformations.

Czech	Best	Second	Third	Fourth
sil	sp:MN	f:SP	s:MN	r:MN
p	c:MN	p:MN	s:SP	gn:SP
rsh	t:MN	p:SP	c:MN	e:MN
e	f:MN	f:MN	x:SP	l:MN
j	p:MN	g:MN	x:MN	y:MN
i	f:SP	b:MN	sil:SP	d:MN
v	x:SP	sil:MN	q:MN	@:MN
aa	s:MN	sp:SP	z:MN	y:SP
m	z:MN	t:MN	j:MN	rr:SP
ii	sil:SP	z:MN	t:MN	%:MN
n	j:SP	x:SP	f:SP	g:SP
o	sp:SP	t:SP	sp:MN	n:MN
s	sil:MN	j:SP	j:SP	v:SP
l	p:SP	sp:MN	p:MN	ll:SP
x	j:MN	k:SP	ch:SP	k:SP
h	q:MN	j:MN	f:MN	b:SP
a	x:MN	d:MN	t:SP	sp:SP
u	k:SP	c:MN	k:SP	w:MN
r	t:SP	k:MN	sp:SP	at:MN
k	k:MN	ch:SP	sil:MN	l:SP
nj	b:MN	q:MN	p:SP	sil:MN
f	ch:SP	h:MN	k:MN	e:SP
ch	g:MN	x:MN	d:MN	r:SP
sp	h:MN	s:MN	g:MN	u:SP
z	d:MN	b:SP	h:MN	m:MN
t	s:SP	g:SP	b:MN	m:SP
c	at:MN	d:SP	e:SP	b:MN
b	g:SP	m:MN	y:SP	ch:SP
uu	@:MN	w:MN	g:SP	i:MN
d	r:MN	sil:SP	@:MN	t:SP
ee	b:SP	rr:SP	%:MN	p:SP
rzh	rr:SP	@:MN	y:MN	n:SP
zh	y:SP	r:SP	d:SP	d:SP
dj	w:MN	n:MN	at:MN	g:MN
ow	r:SP	v:SP	r:MN	p:MN
sh	m:MN	l:SP	ll:SP	o:SP
tj	a:MN	l:MN	i:SP	j:SP
ng	d:SP	r:MN	o:SP	f:SP
g	n:MN	y:MN	b:SP	u:MN
aw	o:SP	y:SP	r:SP	i:SP

Table 19: The Four Closest Spanish and Mandarin Phones for Each Czech Phone Measured Without Transformation.

Cluster	Phones
1	g:CZ l:SP o:SP s:SP rr:SP b:SP m:SP a:SP r:SP n:SP i:SP y:SP sp:SP gn:SP sil:MA h:MA w:MA at:MA n:MA y:MA i:MA s:MA o:MA t:MA x:MA a:MA %:MA d:MA m:MA e:MA gt:MA c:MA u:MA r:MA j:MA q:MA l:MA u:SP v:SP g:SP z:MA @:MA
2	h:CZ sil:SP sp:MA
3	z:CZ j:SP k:MA
4	l:CZ
5	r:CZ
6	o:CZ
7	aw:CZ
8	ii:CZ
9	nj:CZ
10	ng:CZ
11	rzh:CZ
12	zh:CZ
13	f:SP g:MA b:MA p:MA
14	rsh:CZ
15	c:CZ
16	s:CZ
17	aa:CZ
18	a:CZ
19	ee:CZ
20	e:CZ
21	j:CZ
22	i:CZ
23	u:CZ uu:CZ
24	ow:CZ p:SP
25	m:CZ
26	n:CZ
27	v:CZ
28	d:CZ
29	dj:CZ k:SP e:SP d:SP t:SP ll:SP ch:SP x:SP f:MA
30	b:CZ
31	x:CZ
32	tj:CZ
33	sh:CZ
34	ch:CZ
35	f:CZ
36	t:CZ
37	p:CZ k:CZ
38	sil:CZ sp:CZ

Table 20: Phone Clusters Found Using Rotations.

Czech Phone	Source Language Phone	Czech Phone	Source Language Phone
sil	sp:RU	k	k:RU
p	p:SP	t	t:RU
z	c:RU	rsh	fj:RU
ch	chj:RU	g	gj:RU
dj	m:SP	h	v:RU
d	d:SP	ng	g:RU
n	n:RU	uu	u:RU
ii	y:SP	j	j:RU
e	e:RU	ee	e:SP
aa	a:RU	x	x:RU
f	z:RU	b	g:SP
v	v:SP	m	m:RU
ow	u:SP	u	l:RU
s	s:RU	c	ch:SP
tj	sj:RU	l	ee:RU
sh	shj:RU	a	aa:RU
o	o:RU	aw	u:RU
r	y:RU	i	i:RU
nj	gn:SP	rzh	ch:SP
zh	zj:RU	rr	r:SP

Table 21: Phone Mapping Found Using Clustering with Additive Shifts.

3.4 Discussion

Knowledge-Based Methods

In developing knowledge-based methods for cross-lingual phone mapping, we attempted to improve speech recognition performance without access to any target language training data using only linguistic knowledge about the acoustic phonetic structure of each language. We learned that proximity of the source language models to the target language is presently a stronger correlate than anything we can do based on linguistic knowledge and phonetic mappings. We also showed that accounting for some language-dependent bias between the source languages and the target language is not a trivial matter. It seems characterization of the proximity of the target language in an acoustic sense might be a worthwhile topic for further research, as well as a more controlled study of channel-independent acoustic representations. Data and resources related to the information presented in this section can be found on the web at
http://www.clsp.jhu.edu/ws99/projects/asr/final_presentation/knowledge_based .

We proceeded with an analysis of the common error modalities for our best system. We have observed that, though the overall WER is high, performance at the phone-level appears to be fairly good; for example, the Russian system phone error rate was 36.3%. The alignments are plausible, and a majority of the words are only partially misrecognized.

Automatic Methods

The phone confusion method described in this section helped us obtain automatically derived mappings at the phonetic and subphonetic level between a pool of well trained languages and our target language. It is reasonable to expect that the noise and acoustic conditions in which these source languages were recorded will influence and to a large extent determine the phonetic mappings obtained in this way. In other words, the approach described above does nothing to remove any sort of acoustic bias that will influence the phonetic mapping outcome. We showed that this approach helps well to develop a basic set of mappings which will be of help in approaches described later in this report (e.g., DMC). We demonstrated that by combining states from source languages HMMs we can get better mappings than by using phones. It is worthwhile to devote some future efforts into the problem of acoustic bias removal before deriving the acoustic mappings

Language Adaptive Clustering

The above algorithm with additive shifts gave mappings whose performance was comparable to other automatically generated phone mappings. Furthermore, we found that additive shifts can also improve other methods, such as the confusion matrix based approach. Possible extensions of this approach include obtaining mappings at the state level, and using a broader class of transformations, such as affine transformations.

4 Cross-Language Acoustic Model Adaptation

Despite the substantial differences between the quality of phone mappings obtained by knowledge-based and automatic state-level phone mappings, adaptation using MLLR and MAP [1] on the 1.0

[1] References and procedures are in the HTK documentation [25].

Source	Mixtures / Type	Unadapted	GMLLR	MLLR+MAP
MA 10 hours	20 /monophone	88.7		63.0
SP 10 hours	20 / monophone	71.6		50.9
RU 3 hours	20 / monophone	60.8	70.7	45.3
EN 10 hours	20 / monophone	75.7		47.2
CZ 1 hour	20 / monophone	33.4		
CZ 1 hour	6 / triphone	30.7		

Table 22: Effect of Adaptation on Source Language System. Adaptation is via 1 global MLLR transformation, followed by a 4 class MLLR transformation, followed by MAP adaptation. Test set is VOA-2.

Training Data	System Mixtures/Type	Adaptation Steps	WER(%)	
			VOA-2	CUCFN
EN 10.3 hours	12/triphone	4xMLLR+1xMAP	35.1	47.6
EN 10.3 hours	12/triphone	4xMLLR+4xMAP	32.6	44.1
EN 72.0 hours	12/triphone	4xMLLR+4xMAP	32.7	42.1
CZ 1 hour	20/monophone		33.4	47.3
CZ 1 hour	6/triphone		30.7	37.1

Table 23: Effect of Adaptation on English Broadcast News Systems. Number of training iterations for each adaptation procedure are included.

hour of Czech read speech largely compensates for these differences, as shown in Tables 22 and 23. Furthermore, while performance improves significantly, the adapted systems do not individually improve over the monolingual Czech triphone system.

5 Multilingual Discriminative Model Combination

Discriminative model combination [1, 2] aims at an optimal integration of all given acoustic and language models into one log-linear posterior probability distribution. As opposed to the maximum entropy approach, the coefficients of the log-linear combination are estimated on training samples using discriminative methods to obtain an optimal classifier.

Given the posterior distribution $\pi(k|x)$ that observation x belongs to class k, the decision rule that results in a minimum expected number of classification errors is the so-called Bayes' decision rule. For a given observation x of unknown class membership, find the class $k(x)$ such that:

$$\forall k' = 1, ..., K; k' \neq k: \quad \log \pi(k|x) - \log \pi(k'|x) \geq 0. \tag{1}$$

The function $g(x, k, k') = \log(\pi(k|x)/\pi(k'|x))$ in (1) describes the class boundaries and is referred to as discriminant function [7, 10].

In our problem of recognizing continuously spoken sentences, the observation is a sequence of feature vectors $x_1^T = (x^1, \ldots, x^T)$, which has to be classified into a word sequence

$w_1^S = (w^1, ..., w^S)$. However, the true distribution $\pi(w_1^S|x_1^T)$ of natural human speech is unknown. Therefore $\pi(w_1^S|x_1^T)$ has to be approximated by a model distribution $p(w_1^S|x_1^T)$.

A widely used training criterion for the distribution p is the maximum likelihood criterion. The assumption is that we know the functional form of the probability distribution p, but not the parameters. Using the maximum likelihood criterion the parameters are estimated on training samples.

The resulting distribution p is then "plugged in" to the Bayes' decision rule: For a given observation x_1^T of unknown class membership, find the class $w_1^S(x_1^T)$ such that:

$$\forall {w'}_1^{S'} \neq w_1^S : \quad \log p(w_1^S|x_1^T) - \log p({w'}_1^{S'}|x_1^T) \geq 0. \tag{2}$$

Rewriting the discriminant function

$$\begin{aligned} g(x_1^T, w_1^S, {w'}_1^{S'}) &= \log p(w_1^S|x_1^T) - \log p({w'}_1^{S'}|x_1^T) \\ &= \log[p(w_1^S)p(x_1^T|w_1^S)] - \log[p({w'}_1^{S'})p(x_1^T|{w'}_1^{S'})], \end{aligned} \tag{3}$$

we obtain the well-known decomposition of p into a language model probability $p(w_1^S)$ and an acoustic-phonetic likelihood $p(x_1^T|w_1^S)$. Since p typically deviates from the true distribution π, the decision rule (3) will deviate from Bayes' decision rule, thus leading to a suboptimal classifier.

To overcome this limitation, discriminative methods can be applied [17, 20]. The goal of discriminative parameter optimization is to be able to correctly discriminate the observations rather than to fit the distributions to the observed data.

A simple example for the discriminative approach is the optimization of the so-called language model factor λ of the discriminant function:

$$g(x_1^T, w_1^S, {w'}_1^{S'}) = \log[p(w_1^S)^\lambda p(x_1^T|w_1^S)] - \log[p({w'}_1^{S'})^\lambda p(x_1^T|{w'}_1^{S'})]. \tag{4}$$

Experiments [19] show that a value λ with $\lambda \neq 1$ gives a minimum word error rate. The deviation from value $\lambda = 1$ is caused by the deviation of the language model probability $p(w_1^S)$ and the deviation of the likelihood $p(x_1^T|w_1^S)$ from their "true" values.

Let us assume that we are given M different acoustic and language models, which are identified by numbers $j = 1, \ldots, M$. From model j we can compute the posterior probability $p_j(k|x)$ of a hypothesized class k given an observation x. These models are now log-linearly combined into a distribution of the exponential family:

$$p^{\Pi}_{\{\Lambda\}}(k|x) = e^{-\log Z_\Lambda(x) + \sum_{j=1}^M \lambda_j \log p_j(k|x)} \tag{5}$$

The coefficients $\Lambda = (\lambda_1, ..., \lambda_M)^T$ can be interpreted as weights of the models j within the model combination (5). The value $Z_\Lambda(x)$ is a normalization constant. As opposed to the maximum entropy approach, which leads to a distribution of the same functional form, the coefficients Λ are optimized with respect to the decision error rate of the discriminant function (6):

$$\log \frac{p_\Lambda(k|x)}{p_\Lambda(k'|x)} = \sum_{j=1}^M \lambda_j \log \frac{p_j(k|x)}{p_j(k'|x)} \tag{6}$$

This approach is called *Discriminative Model Combination*. If only one acoustic and one language model are combined, DMC will optimize the so called language weight,or language model factor,. DMC allows for the integration of any model into an optimal decoder, since the weight λ_j of the model j within the combination depends on its ability to provide information for correct classification.

5.1 DMC Training

Thus far, DMC has been used to optimize large vocabulary continuous speech recognition (LVCSR) systems at the sentence level, although it can also be applied to other problems in pattern recognition due to its general formulation. In LVCSR systems, the spoken utterance is used as observation x and any hypothesized sentence can be regarded as class k. For DMC training we are given a set of sentences $n = 1, \ldots, N$. For each of the training sentences we are given the acoustic observation x_n and the correct class assignment k_n, i.e. k_n is the correct transcription of x_n. Using a preliminary decoding we can define the set of rival classes $k \neq k_n$ and we can compute the number of word errors of the rival class k with the help of the Levenshtein distance $\mathcal{L}(k_n, k)$. The model combination should then minimize the word error count $E(\Lambda)$:

$$E(\Lambda) = \sum_{n=1}^{N} \mathcal{L}\left(k_n, \arg\max_{k \neq k_n}\left(\log \frac{p_\Lambda(k|x_n)}{p_\Lambda(k_n|x_n)}\right)\right)$$

on representative training data to assure optimality on an independent test set. Since this optimization criterion is not differentiable we approximate it in analogy to the well-known MCE training by a smoothed word error count:

$$E_{MWE}(\Lambda) = \sum_{n=1}^{N} \sum_{k \neq k_n} \mathcal{L}(k, k_n) S(k, n, \Lambda), \tag{7}$$

where $S(k, n, \Lambda)$ is a smoothed indicator function. $S(k, n, \Lambda)$ should be close to one if the classifier (6) will select hypothesis k and it should be close to zero if the classifier (6) will reject hypothesis k. One possible indicator function with these properties is

$$S(k, n, \Lambda) = \frac{\left(p^{\Pi}_{\{\Lambda\}}(k|x_n)\right)^{\eta}}{\sum_{k'}\left(p^{\Pi}_{\{\Lambda\}}(k'|x_n)\right)^{\eta}}, \tag{8}$$

where η is a suitable constant. Optimization of $E_{MWE}(\Lambda)$ with respect to Λ leads to an iterative gradient descent scheme. Another possible indicator function with similar properties is the following second order function:

$$S(k, n, \Lambda) = \begin{cases} 1 & g > A \\ \left(\frac{g+B}{A+B}\right)^2 & -B < g < A \\ 0 & g < -B \end{cases} \tag{9}$$

where

$$g = \log \frac{{p^{\Pi}}_{\{\Lambda\}}(k|x_n)}{{p^{\Pi}}_{\{\Lambda\}}(k_n|x_n)}$$

which gives a closed form matrix solution for Λ. The values A, B determine the form of the 2-nd degree function and the set of hypotheses used for the training. Both indicator functions lead to similar and reasonable DMC coefficients λ_j.

5.2 Log-linear Structuring of Distributions

DMC allows for the optimization of any log-linear distribution. Now, the idea is to find suitable factorizations of the distribution. Each of the factors may then be weighted independently, leading to a log-linear functional form of the distribution. The following 2 sections show examples of structuring the overall distribution into a log-linear form.

Assume that we are given acoustic models $A_i, i = 1, \ldots, I$ and language models $L_j, j = 1, \ldots, J$. These models can be interpolated into one log-linear distribution:

$${p^{\Pi}}_{\{\Lambda\}}(k|x) = \frac{\prod_i A_i(x|k)^{\lambda_i} \prod_j L_j(k)^{\lambda_j}}{\sum_{k'} \prod_i A_i(x|k')^{\lambda_i} \prod_j L_j(k')^{\lambda_j}}. \quad (10)$$

Define:

$$p_{A,i}(k|x) = \frac{A_i(x|k)}{\sum_{k'} A_i(x|k')} \quad (11)$$

$$p_{L,j}(k|x) = L_j(k) \quad (12)$$

Then we can write:

$${p^{\Pi}}_{\{\Lambda\}}(k|x) = \frac{\prod_i p_{A,i}(k|x)^{\lambda_i} \prod_j p_{L,j}(k|x)^{\lambda_j}}{\sum_{k'} \prod_i p_{A,i}(k'|x)^{\lambda_i} \prod_j p_{L,j}(k'|x)^{\lambda_j}} \quad (13)$$

Thus we are able to handle the optimization of the interpolation of acoustic as well as of language models in a uniform way. Both model types may be interpolated optimally at the same time using DMC.

5.3 Structuring the Distribution into Phonetic Classes

The following structuring of a distribution was first applied to decompose the distribution into word classes [24]. The idea here is to segment the overall score $\log p(x|k)$ of the sentence k into the phones h found in the sentence k:

$$\log p(x|k) = \sum_{h \in k} \log p(x^h|h) \quad (14)$$

To reduce the number of classes, here we clustered all phonemes into the three classes: vowels, consonants, silence.

$$\log p_V(x|k) = \sum_{h \in k} \delta(h, V) \log p(x^h|h) \quad (15)$$

$$\log p_C(x|k) = \sum_{h \in k} \delta(h, C) \log p(x^h|h) \quad (16)$$

$$\log p_S(x|k) = \sum_{h \in k} \delta(h, S) \log p(x^h|h) \quad (17)$$

After segmenting the sentence probability into these 3 phonetic classes we can again build a log-linear distribution:

$$p^{\Pi}{}_{\{\Lambda\}}(k|x) = \frac{p(k)^{\lambda_L} p_V(x|k)^{\lambda_V} p_C(x|k)^{\lambda_C} p_S(x|k)^{\lambda_S}}{\sum_{k'} p(k')^{\lambda_L} p_V(x|k')^{\lambda_V} p_C(x|k')^{\lambda_C} p_S(x|k')^{\lambda_S}} \quad (18)$$

The idea behind this functional form of the distribution is that the vowels, consonants and silence models may have a different importance for classification and should be weighted differently.

5.4 Combination of Multiple Source Language Acoustic Models

5.4.1 Sentence-Level Model Combination

The most direct use of DMC to merge models from multiple languages is to combine them at the sentence-level. For example, a combination of Spanish and Mandarin models has the following functional form:

$$\log p^{\Pi}{}_{\{\Lambda\}}(k|x) = C + \lambda_{LM} L_{cz}(k) + \lambda_{cz} A_{cz}(x|k) + \lambda_{sp} A_{sp}(x|k) + \lambda_{ma} A_{ma}(x|k) \quad (19)$$

where $L_{cz}(k)$ is the Czech language model, $A_{cz}(x|k)$ is the Czech acoustic model, $A_{sp}(x|k)$ is the Spanish acoustic model and $A_{ma}(x|k)$ is the mandarin acoustic model.

5.4.2 Phonetic Class Combination

A potentially more powerful approach is to use DMC with language-specific phonetic classes by applying the ideas of Section 5.3. In the example given here, DMC is applied at the phoneme-class model level, i.e. for each of the three languages a vowel-model, a consonant-model and a silence-model were created, summing up to 9 different acoustic models:

$$\begin{aligned} \log p^{\Pi}{}_{\{\Lambda\}}(k|x) = {} & C + \lambda_{LM} L_{cz}(k) \\ & + \lambda_{cz,V} V_{cz}(x|k) + \lambda_{cz,C} C_{cz}(x|k) + \lambda_{cz,S} S_{cz}(x|k) \\ & + \lambda_{sp,V} V_{sp}(x|k) + \lambda_{sp,C} C_{sp}(x|k) + \lambda_{sp,S} S_{sp}(x|k) \\ & + \lambda_{ma,V} V_{ma}(x|k) + \lambda_{ma,C} C_{ma}(x|k) + \lambda_{ma,S} S_{ma}(x|k). \end{aligned} \quad (20)$$

where $V_{cz}(x|k)$ is the Czech vowel-class model, $C_{cz}(x|k)$ is the Czech consonant-class model, $S_{cz}(x|k)$ is the Czech silence model, and the remaining class models are defined similarly for Spanish and Mandarin.

5.5 DMC Experiments on the VOA Test Sets

The software used at the workshop was restricted to the use of n-best lists for DMC-training and DMC-decoding. Note that DMC can be applied directly on dense lattices, as is done at Philips

DMC models	WER(%)
oracle	19.54
anti-oracle	45.04
first best (baseline)	24.76
$A_{cz} + L_{cz}$	24.36
$A_{cz} + L_{cz} + WordPenalty$	23.78
$A_{cz} + A_{sp} + A_{ma} + L_{cz} + WordPenalty$	24.01

Table 24: Sentence-Level DMC Results on the VOA-1 Test Set. The experiment used 100-best lists, knowledge based Spanish and Mandarin phoneme mappings, combination of the Czech language model, the Czech acoustic model, the Spanish acoustic model, and the Mandarin acoustic model

Research Laboratories in the HUB4 system. This simplified our experiments; when working with N-Best lists, the handling of the hypotheses for DMC-training and decoding becomes trivial. On the other hand, the choice of the size of the N-Best lists turned out to be crucial for the obtained results, which will be shown in the next few sections.

Experiments on the VOA-1 Test Set

In the first set of experiments, knowledge based mappings of Spanish (SP) and Mandarin (MA) context independent phonemes to Czech (CZ) phonemes were applied. These mappings were created prior to the workshop. Both the Spanish and the Mandarin systems were trained on about 10 hours of acoustic Broadcast News data. Using the mappings, Czech phoneme models were generated by plugging in the corresponding Spanish phoneme model. The same procedure was applied for the Mandarin models. Thus we arrived at a Spanish-Czech and a Mandarin-Czech system.

The baseline Czech monophone system (CZ) was trained on 1 hour Czech VOA data. Using this system, 100-best lists were decoded on the Czech VOA-1 test corpus. The task was now to beat the 1 hour Czech VOA system, with help of the Spanish-Czech and the Mandarin-Czech system. For this reason, the Spanish and Mandarin models were scored on the 100-best lists of the Czech VOA-1 test corpus. Next DMC was applied, using the held-out method: train DMC coefficients on the first/second half of the corpus and test on the second/first half, and add up the error counts obtained on both tests.

Since the N-Best lists were quite short and the Spanish and Mandarin knowledge-based models performed badly, - A free decoding of Spanish and Mandarin models on Czech data gives a word error rate of about 80-90% - no real gain was expected. The following two sections summarize the experiments on this test set at the sentence-level and on the phoneme-class level.

Using a sentence-level model combination as described in Equation 19, produced results summarized in Table 24. From these we can conclude that Spanish and Mandarin models do not help, which may be due to the small size of the n-best lists or due to the weakness of the models. Another important observation is that DMC automatically found out that the given Spanish and Mandarin models are weak, so these models received a small weight within the model combination and the performance of the overall system did not degrade! In addition we see that DMC optimized automatically the language weight; this is a well-known effect.

The next experiment used for DMC phonetic-class combination. To have a fair comparison,

DMC Model	WER(%)
oracle	19.5
anti-oracle	45.0
first best (baseline)	24.8
$A_{cz} + L_{cz} + WP$	23.8
$V_{cz} + C_{cz} + S_{cz} + L_{cz} + WP$	23.2
$V_{cz} + C_{cz} + S_{cz} + V_{sp} + C_{sp} + S_{sp} + + V_{ma} + C_{ma} + S_{ma} + L_{cz} + WP$	23.5

Table 25: Phonetic-Class DMC Results on the VOA-1 Test Set. The experiment used 100-best lists, knowledge based Spanish and Mandarin phoneme mappings, combination of the Czech language model, the Czech, Spanish, and Mandarin vowel, consonant, and silence models.

the monolingual baseline experiment was defined by applying DMC to the Czech phoneme-class models only. From Table 25 we can see that the structuring into phoneme classes improves the Czech system from 23.8% to 23.2% but that, again, the Spanish and Mandarin Models do not help. Note that the slight improvement from 23.8% to 23.2% is gained by adding only 3 free parameters to the system only.

Experiments on the VOA-2 Test Set

A new experiment definition was created to overcome the limitations described above. The size of the N-Best lists were increased significantly and more advanced models were applied. During the workshop period, knowledge based mappings of Spanish (SP), Russian (RU), and English (EN) models to Czech models were created. The Czech baseline monophone system was trained on 1 hour of Czech read speech data which may be a useful startup when building a system in a very new language. Using this system 1000-best lists were decoded on the Czech VOA-2 test corpus. The Spanish, Russian, and English models were scored on these 1000-best lists. DMC was again applied using the held-out method. This setup is reasonable and should give more realistic results than the VOA-1 setup in the previous section.

With this new test set, the experiments on the sentence level as well as on the phoneme-class level were repeated. The overall combination of the given source models leads to the following functional form:

$$\begin{aligned} \log p^{\Pi}{}_{\{\Lambda\}}(k|x) &= C + \lambda_{LM} L_{cz}(k) + \lambda_{cz} A_{cz}(x|k) + \lambda_{sp} A_{sp}(x|k) \\ &\quad + \lambda_{ru} A_{ru}(x|k) + \lambda_{en} A_{en}(x|k). \end{aligned} \tag{21}$$

The performance of the various single acoustic models is summarized in Table 26. The error rates were obtained by a free decoding using the source-language acoustic models and the Czech language model on the Czech VOA-2 test set.

The systems in Table 26, which are marked with a '*', were combined into one decoder using DMC. Results are presented in Table 27, from which we can see that the Spanish and Russian models help to improve the system and that the English triphone models help even more. The overall word error rate can be reduced significantly from 33.4% (free decoding of Czech 1 hour system) down to 29.2% (N-Best decoding of multilingual system combination). This result was

Source Language Acoustic Model	VOA-2 WER(%)
29 hour Spanish BN Monophone	71.1 *
4 hour Russian Monophone	60.6 *
10 hour English BN Triphone adapted to Czech	35.1 *
1 hour Czech CUCFN Monophone	33.4 *
1 hour Czech CUCFN Triphone	30.7
10 hour Czech CUCFN Triphone	27.1

Table 26: Recognition Performance of Source-Language and Target-Language Systems.

DMC Model	VOA-2 WER(%)
oracle	19.8
anti-oracle	56.6
first best (baseline)	34.0
$L_{cz} + A_{cz}$	32.7
$L_{cz} + A_{cz} + A_{ru}$	32.5
$L_{cz} + A_{cz} + A_{ru} + A_{sp}$	32.3
$L_{cz} + A_{cz} + A_{ru} + A_{sp} + A_{en}$	29.2

Table 27: Results of Sentence-Level DMC on the VOA-2 Test Set. The experiment used 1000-best lists, knowledge based Spanish, Russian, and English phoneme mappings, combination of the Czech language model, the Czech acoustic model, the Spanish acoustic model, the Russian acoustic model, and the English acoustic model

DMC model	WER(%)
oracle	19.8
anti-oracle	56.6
first best	34.0
$L_{cz} + A_{cz}$ (baseline)	32.7
$L_{cz} + V_{cz} + C_{cz} + S_{cz}$	32.1
$L_{cz} + A_{cz} + A_{ru} + A_{sp}$	32.3
$L_{cz} + V_{cz} + C_{cz} + S_{cz} + V_{ru} + C_{ru} + S_{ru} + V_{sp} + C_{sp} + S_{sp}$	31.8
$L_{cz} + A_{cz} + A_{ru} + A_{sp} + A_{en}$	29.2
$L_{cz} + V_{cz} + C_{cz} + S_{cz} + V_{ru} + C_{ru} + S_{ru} + V_{sp} + C_{sp} + S_{sp} + V_{en} + C_{en} + S_{en}$	28.9

Table 28: Results of Phonetic-Class DMC on the VOA-2 Test Set. The experiment used 1000-best lists, knowledge based Spanish, Russian, and English phoneme mappings, combination of the Czech language model, the Czech, Spanish, Russian, and English vowel, consonant and silence models

not expected since the combination was done at the sentence level and only 5 free parameters were optimized.

The next question addressed was whether the results at the sentence level can be further improved when applying DMC to phonetic-classes. To compare the results with a corresponding Czech baseline system, the phonetic class model combination was optimized using the Czech models only, as in Equation 19. The following overall system was created and

$$\begin{aligned}
\log p^{\Pi}_{\{\Lambda\}}(k|x) &= C + \lambda_{LM} L_{cz}(k) \\
&\quad + \lambda_{cz,V} V_{cz}(x|k) + \lambda_{cz,C} C_{cz}(x|k) + \lambda_{cz,S} S_{cz}(x|k) \\
&\quad + \lambda_{sp,V} V_{sp}(x|k) + \lambda_{sp,C} C_{sp}(x|k) + \lambda_{sp,S} S_{sp}(x|k) \\
&\quad + \lambda_{ru,V} V_{ru}(x|k) + \lambda_{ru,C} C_{ru}(x|k) + \lambda_{ru,S} S_{ru}(x|k) \\
&\quad + \lambda_{en,V} V_{en}(x|k) + \lambda_{en,C} C_{en}(x|k) + \lambda_{en,S} S_{en}(x|k). \qquad (22)
\end{aligned}$$

Table 28 summarizes the results obtained after optimizing the above model combinations.

5.6 Conclusions and Summary

For the VOA-1 setup we can draw the following conclusions:

- The knowledge based Mandarin and Spanish models did not help. Since DMC found this automatically, the system performance did not degrade.
- The n-best lists were too short.
- The language weight was optimized automatically.
- The results matched our expectations.

From the VOA-2 results of Table 28 we can conclude that the structuring into phoneme classes improves the system and that the interpolation of multilingual phoneme-class models performs better than the interpolation of multilingual systems.

The conclusions from this work using DMC for language independent acoustic modeling can be condensed into the following points:

- DMC works in the language independent environment.
- The size of the n-best lists is crucial. The size should be chosen so that the word error rate of the oracle and of the first-best decoding differ significantly.
- A multilingual interpolation of systems performs better than the monolingual 1 hour system.
- Poor models from remote languages can still contribute to the performance of the overall model combination if they are optimally weighted by DMC-training.
- A log-linear structuring into phonetic classes seems to improve the classification.
- The best performance is achieved by a multilingual interpolation of phonetic class models.

6 Multilingual Model Combination Using ROVER

An approach to multilingual model combination that is straightforward to implement and investigate is to combine system hypotheses using ROVER [9]. ROVER combines hypotheses using a majority voting scheme to produce a consensus hypothesis and has been shown to provide modest gains in performance in LVCSR experiments. We will use ROVER as a means of combining alternative hypotheses derived from different source language acoustic models.

ROVER uses the NIST alignment algorithm to create a word transition network, taking the first input system as a reference and aligning with it the remaining hypotheses. ROVER then uses a voting algorithm to determine the best path of word transitions between all nodes of the network. Word-level confidence scores in the range [0-1] are needed by the voting algorithm were derived in two ways. One method used to derive these scores was to linearly map the acoustic scores of the hypotheses in the 1000 element N-Best list into the confidence range [0-1]. The second method was to use the performance accuracy of the system as a constant confidence score for all hypotheses of that system; we found this latter method to yield the best results.

ROVER was used in two experiments: to combine the four knowledge-based systems that used each of the source language acoustic HMM models and a knowledge based mapping, and to produce a baseline for comparison with the DMC system. The results of the knowledge-based experiments are described in Table 29. We found that the best ROVER output using the accuracy scores of each system as confidence scores produced a WER of 62%. The same experiment performed at the phone-level gives similar improvement.

As a simple experiment to compare ROVER and DMC, the 1000-Best hypotheses from the Czech system are rescored with optimized language model weights and combined using ROVER with confidence scores from system accuracies [2]. The results are summarized in Table 30, but we find DMC and ROVER to give similar performance.

[2] As a result of being restricted to the Czech 1000-Best list, performance improves over that reported in Table 29.

Source	WER(%)	Source	WER(%)
Russian	65.2	Spanish	79.3
English	80.9	Mandarin	90.1
Rover	62.0		

Table 29: ROVER Combination of Source Language Hypotheses Using Knowledge-Based Phone Mappings

Method	Source	WER(%)
-	Czech	33.3
-	Russian	60.8
-	Spanish	71.6
ROVER		33.1
DMC		33.0

Table 30: Comparison of DMC and ROVER in Czech Monolingual 1000-Best List Rescoring.

6.1 Discussion

Using ROVER in the knowledge based phone mappings, we obtain only a small improvement over the best monolingual system, but it provides additional evidence that some benefit can be obtained by combining scores from multiple language systems and we stress again that in this experiment no target language data at all is used in these experiments.

The advantage of ROVER is that it can combine different hypotheses from the different systems, whereas DMC needs the intersection of hypotheses of different systems. In other words DMC needs different scores from different systems of the same hypothesis. Thus ROVER is simpler to implement, although this difference is minor in our N-Best rescoring experiments, since the same hypotheses are rescored by all systems.

7 Conclusion

We have presented the results of our experiments in language independent acoustic modeling. We studied both knowledge-based and automatic methods to derive cross-lingual phonetic and sub-phonetic mappings, and found that the automatic methods performed significantly better than the knowledge-based methods.

Acoustic HMM adaptation further improved the source language models, although not to the point that they performed better than monolingual Czech systems. However, multilingual interpolation with adapted source-language acoustic models was effective in improving the performance of monolingual systems. Surprisingly, even source-language models that perform poorly when used individually can contribute to the overall combination when their contribution is determined by DMC-training. In summary, we have developed a methodology in which cross-language phonetic mappings, acoustic adaptation, and discriminative model combination can be used to improve monolingual systems trained from small amounts of speech.

8 Acknowledgments

We thank M. Riley and F. Pereira of ATT for use of their large vocabulary decoder. Satellite news broadcast recordings were done under contract by the Linguistic Data Consortium, Philadelphia, PA, USA. Language modeling data has been provided by the Lidove Noviny Publishers, Prague, Czech Republic. We gratefully acknowledge the invaluable assistance provided by our colleagues in the Czech Republic at the Charles University and the University of West Bohemia.

References

[1] P. Beyerlein. Discriminative model combination. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

[2] P. Beyerlein. Discriminative model combination. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 481–484, 1998.

[3] W. Byrne, J. Hajic, P. Ircing, F. J. S. Khudanpur, J. McDonough, N. Peterek, and J. Psutka. Large vocabulary speech recognition for read and broadcast Czech. In *Workshop on Text Speech and Dialog, Marianske Lazne, Czech Republic*, 1999.

[4] D. Calvert. *Descriptive Phonetics*. Thieme Inc., New York, 1986.

[5] A. Constantinescu and G. Chollet. On cross-language experiments and data-driven units for ALISP (automatic language independent speech processing). In *ASRU*, pages 606–613, 1997.

[6] L. Deng. Integrated multilingual speech recognition - impact on chinese spoken language processing. In *Proceedings of the International Symposium on Chinese Spoken Language Processing, Singapore*, 1998.

[7] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley, New York, 1973.

[8] V. Fischer, Y. Gao, and E. Janke. Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognizer. In *Proceedings of the International Conference on Spoken Language Processing*, 1998.

[9] J. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–354, 1997.

[10] K. Fukunaga. *Introduction to Statistical Pattern Recognition, Second Edition*. Academic Press, 1990.

[11] P. Fung, C. Y. Ma, and W. K. Liu. MAP-based cross-language adaptation augmented by linguistic knowledge: from English to Chinese. In *EUROSPEECH*, pages 871–874, 1999.

[12] P. Geutner, M. Finke, and P. Scheytt. Adaptive vocabularies for transcribing multilingual broadcast news. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998.

[13] P. Geutner, M. Finke, P. Scheytt, A. Waibel, and H. Wactlar. Transcribing multilingual broadcast news using hypothesis driven lexical adaptation. In *DARPA Hub4 Workshop*, 1998.

[14] P. Geutner, M. Finke, and A. Waibel. Phonetic-distance-based hypothesis driven lexical adaptation for transcribing multilingual broadcast news. In *Proceedings of the International Conference on Spoken Language Processing*, 1998.

[15] R. M. Gray. Vector quantization. *IEEE ASSP Magazine*, pages 4–29, April 1984.

[16] E. Hovy, N. Ide, R. Frederking, J. Mariani, and A. Zampolli. Multilingual information management: Current levels and future abilities, *http://www.cs.cmu.edu/ ref/mlim/* . Technical report, US NSF, April 1998. .

[17] B. H. Juang, W. Chou, and C.-H. Lee. Speech recognition and coding - new advances and trends. In A. J. R. Ayuso and J. M. L. Soler, editors, *Statistical and Discriminative Methods for Speech Recognition*. Springer-Verlag, Berlin-Heidelberg, 1995.

[18] J. Kohler. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2195–2198, 1996.

[19] K.-F. Lee. *The Development of the SPHINX System*. Kluwere Academic Publishers, Boston, 1989.

[20] H. Ney. On the probabilistic interpretation of neural network classifiers and discriminative training criteria. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, pages 107–119, 1995.

[21] J. Picone. Knowledge-based phone mappings. Technical report, Center For Language and Speech Processing, Johns Hopkins University, August 1999. *http://www.clsp.jhu.edu/ws99/projects/asr/final_presentation/knowledge_based* .

[22] T. Schultz and A. Waibel. Language independent and language adaptive large vocabulary speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1819–1822, 1998.

[23] The International Phonetic Association. *Handbook of the International Phonetic Alphabet*. Cambridge University Press, Cambridge UK, 1999. also *http://www.arts.gla.ac.uk/IPA/fullchart.html* .

[24] D. Vergyri. personal communication.

[25] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book*. Entropic, Inc., 1999.