1. KNOWLEDGE-BASED PHONE MAPPINGS

In some applications, it is highly desirable to be able to develop speech recognition systems without a need for any acoustic training data. In such situations, borrowing models from other languages for which speech recognition technology is well-developed is extremely attractive. We refer to these well-developed languages for which data exists as source languages. The language for which we build the new recognizer is referred to as the target language. 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. For reasons that are obvious, we did allow access to language model training data and pronunciation dictionaries in the target language.

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 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. These goals are summarized in Figure 1. 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 requiring no target language training data generally performed in the range of 80% WER; systems allowed some access to target language data to determine phone-level or state-level mappings, but did not do any acoustic retraining, 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%. The second 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.

The recognition technology employed for this study, as previously mentioned, is a standard monophone-based continuous density hidden Markov model LVCSR system. Relevant features of the system are summarized in Table 1. Monophone acoustic models containing approximately 20 mixtures per state were used because these gave performance very competitive with more sophisticated context-dependent phonetic models, and were much easier to manipulate for the experiments described below. The recognition architecture was a synthesis of a finite state machine decoder developed at AT&T, which served as the search engine during decoding, and the acoustic modeling capabilities of Entropic's HTK (v2.2) system, which provided Gaussian statistical modeling calculations. HTK was used for all acoustic training. With this architecture, most recognition experiments required on the order of 350M of memory and less than 4x real-time on a 450 MHz Pentium processor.

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 [1]. 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) [2]. A portion of this analysis is shown in Figure 1. A complete inventory, along with several related resources, can be found in [3]. An example of such a description for a phone is shown in Table 2. 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. Examples of such mappings are given in Figure 3. 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 3. 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 3, and explained in greater detail in Figure 4. The performance on the VOA-1 evaluation for each of these mappings is given in Table 4. 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.

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 5, and the associated performance is given in Table 5. 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 [4] 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 6. Unfortunately, this approach resulted in a slightly degraded performance, as shown in Table 6. 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 6. The corresponding performance is given in Table 6. 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 achieved, as expected. However, overall performance is still below the best monolingual system, and far below the Russian system shown in Table 6. Again, this was a discouraging result.

We proceeded with an analysis of the common error modalities for our best system. This is summarized in Figure 7. We have observed that, though the overall WER is high, performance at the phone-level appears to be quite good. The alignments are plausible, and a majority of the words are only partially misrecognized. Since Czech is an inflected language, this analysis raised some concerns that our language modeling approach was not optimal. For example, a morphologically-based approach might pay dividends if the majority of the errors are occurring on inflections rather than stems (it could be the case that performance at a morphological level is good, and hence the system would be usable for information extraction tasks). This analysis also encouraged us to consider better methods of combining phone models across source languages as a way of making our phone models more language independent (from the previous experiments it is clear that the models are well-tuned to the source languages and corresponding channel characteristics).

1.3. MODEL COMBINATION

We conducted some explorations into ways one could combine models based only on confusion data from the source languages. Our strategy here was simple: generate confusion data by recognizing the source language training data with multiphone systems (and tracking instances where a phone in the reference transcription for one source language phone was modeled by a phone from a different language). The procedure for generating this confusion data involved the standard forced alignment approach used in HMM training, as shown in Figure 8. However, each source language phone was allowed to be represented as any other source language phone. The alignment process would make the best choice using the multiphone pronunciation network previously described.

The confusion data generated by this experiment was quite interesting. The confusion matrix (an extremely large matrix since it has N phones x M languages number of entries) is surprisingly diagonal. An example of some of these confusions is shown in Figure 9. The diagonal nature of this matrix shows how finely tuned the HMMs are to their specific language training data, and most likely the channel conditions of that data as well. Such biases most likely contribute to the lack of success we have had applying source language models to a new domain. Yet, when confusions do occur, they occur in plausible ways (consonants are more consistent than vowels; some vowels generate confusions with their counterparts in other languages). Nevertheless, in an effort to approach what one might consider to be language-independent phones, or IPA phones, we investigated ways to combine phone models using this confusion data.

It was our original intent to combine models using some sort of factor analysis or multidimensional scaling. However, we did not have enough time to fully explore this approach. A much simpler approach that was easier to implement and investigate, was to combine system outputs using an approach known as ROVER [5]. The ROVER process is summarized in Figure 10. ROVER combines hypotheses using a majority voting scheme to produce a consensus hypothesis. It has been shown to provide modest gains in performance in LVCSR experiments. We employed ROVER at the word level and the phone level. The results are shown in Table 7.

ROVER can be easily employed to improve WER by combining the word-level outputs of the monolingual systems (or any group of systems for that matter). As we see in Table 7, this gave a modest improvement in performance for the word-level systems. WER was improved from 65.2% for the Russian system to 62.1% for ROVER system using all source language systems. The same experiment conducted at the phone level showed a similar improvement in performance. Note that the phone-level accuracy of our system, approximately a 35% phone error rate, is surprisingly high given the high WER (ranging from 90% to 65% for the source language systems). Normally, we would expect phone error rates to be higher than the WER. Again, we see this as an encouraging result and believe it points to some language modeling interactions that are not well understood in our experiments.

1.4. CHANNEL NORMALIZATION

Since improving the system without access to target language training data has been hard to do, we became concerned that there were some systematic variations in channel that were preventing

our source language models from performing well. To investigate this hypothesis, we conducted two types of experiments: (1) investigation of the sensitivity of our phone mappings to the silence model (which serves as an estimate of the background channel), and (2) a simple global maximum likelihood linear regression (MLLR) [6] experiment. When mapping models from source languages to target languages, one must make some decision about the best strategy for synthesizing a silence model. In a cheating experiment scenario, we used an actual CZBN silence model, as well as silence models from each of the source languages, but found no great impact on recognition performance (for example, at most, a 1% WER improvement when using the CZBN silence model versus the original source language silence model).

In our MLLR experiment, the source language models for English were adapted to the Czech data by allowing a single global transformation of the model means (a transformation matrix is derived that is applied to each mean vector in the HMM models). The results of this experiment are summarized in Table 8. We observe that global MLLR produced a measurable improvement in performance. However, this improvement was not great enough to offset the language mismatch. The WER for the MLLR-adapted system was not better than that achieved with some basic data-driven phone mapping experiments (which provide WER's in the mid-60% range), nor was it better than that achieved using Russian models trained from a completely different ambient environment and speaking style. Hence, from this simple experiment we conclude that a systematic channel variation is not the primary factor limiting performance. In fact, it seems to suggest that there is a language-dependent shift that must be accounted for in a more complex way.

1.5. SUMMARY

In this study, we attempted to improve speech recognition performance without access to any target language training data. We attempted this using 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 the following URL: *http://www.clsp.jhu.edu/ws99/projects/asr/final_presentation/knowledge_based*.

REFERENCES

[1] D.R. Calvert, *Descriptive Phonetics*, Thieme Inc., New York, New York, USA, 1986.

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

[3] J. Picone, "Knowledge-Based Phone Mappings," *http://www.clsp.jhu.edu/ws99/projects/asr/final_presentation/knowledge_based*, Center For Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland, USA, August 1999.

[4] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Large Vocabulary Speech Recognition," *Proceedings of the International Conference on Spoken Language Processing*, pp. 1819-1822, Sydney, Australia, November 1998.

[5] J. Fiscus, "A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barbara, pp. 347-354, California, USA, December 1997.

[6] C. J. Leggetter, and P. C. Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression," *Proceedings of the ARPA Spoken Language Technology Workshop*, Barton Creek, Texas, USA, February 1995.

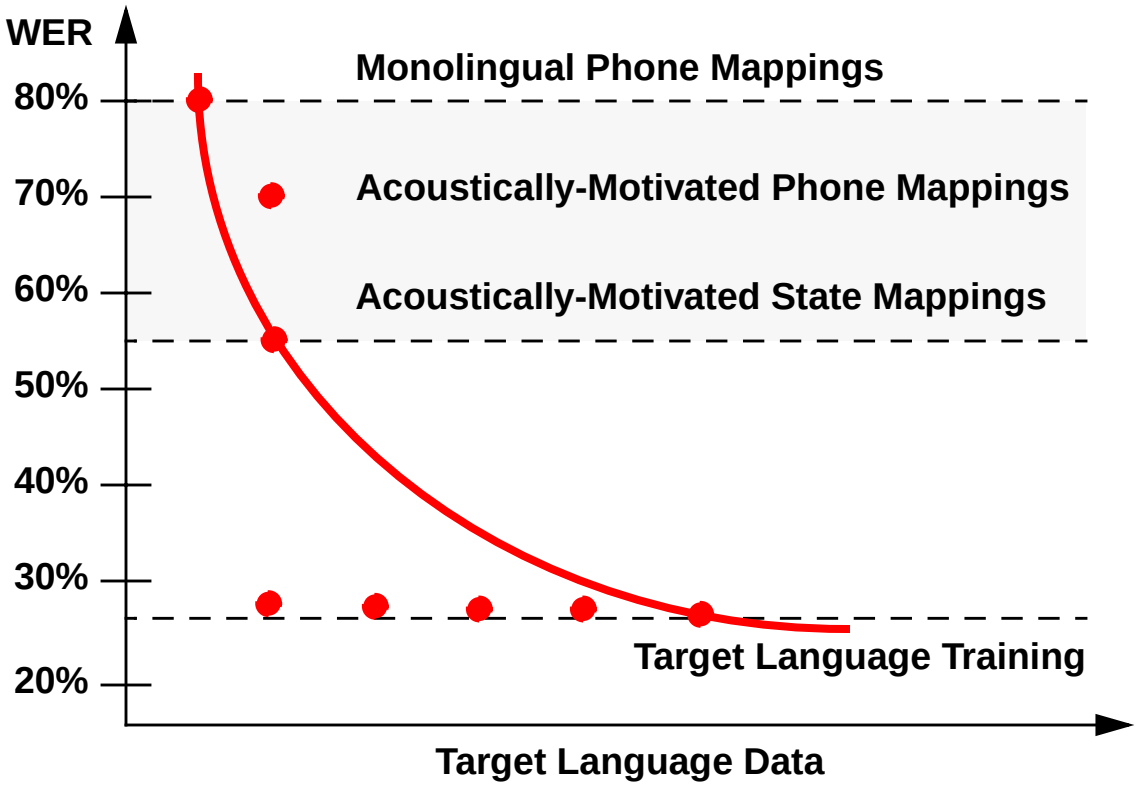


Figure 1. An overview of word error rate (WER) as a function of the amount of target language data. Our goal in exploiting linguistic knowledge to develop target language phone mappings is to approach the performance of techniques that require use of target language data. In this case the difference in performance is approximately 80% WER for knowledge-based approaches, and 55% WER for acoustically-motivated approaches.

IPA Conversion Chart For Consonants

Manner	Language	Place of Articulation																						N
		Biliabial		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 2. An IPA description of the consonant portions of the phone sets used in our experiments.

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) tornado	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 3. 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.

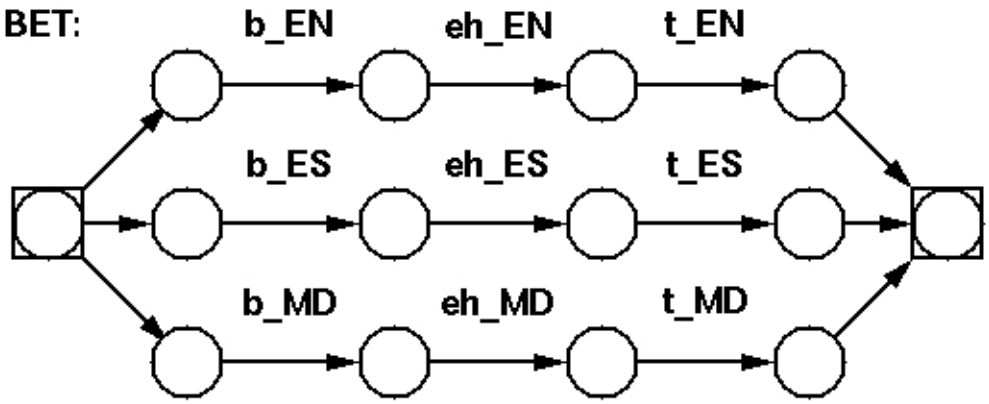
Czech		Mandarin				
v1	**Example – CZ**	**IPA or Description**	**v1**	**v2**	**v3**	**v4**
a	(ah:2) but	front allophone of /a/	@	a	@	@
aa	(aax:2) father	front allophone of /a/	@	@	@	@
aw	(aw:1) down	schwa, mid central unrounded	&	e	&	&
b	(b:1) blue	p	b	b	b	b
c	(t s:3) Yeltsin	aspirated dental affricate 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 4. Four variations of Czech to Mandarin phone mappings that were explored to diagnose the poor performance of the Mandarin system.

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 5. A selective phone mapping that uses phones from three source languages to model Czech.

Parallel Pronunciations:



All-Phone Approach:

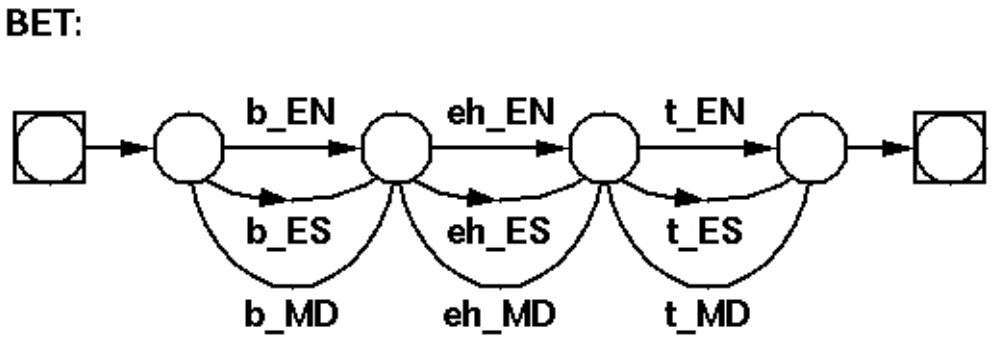
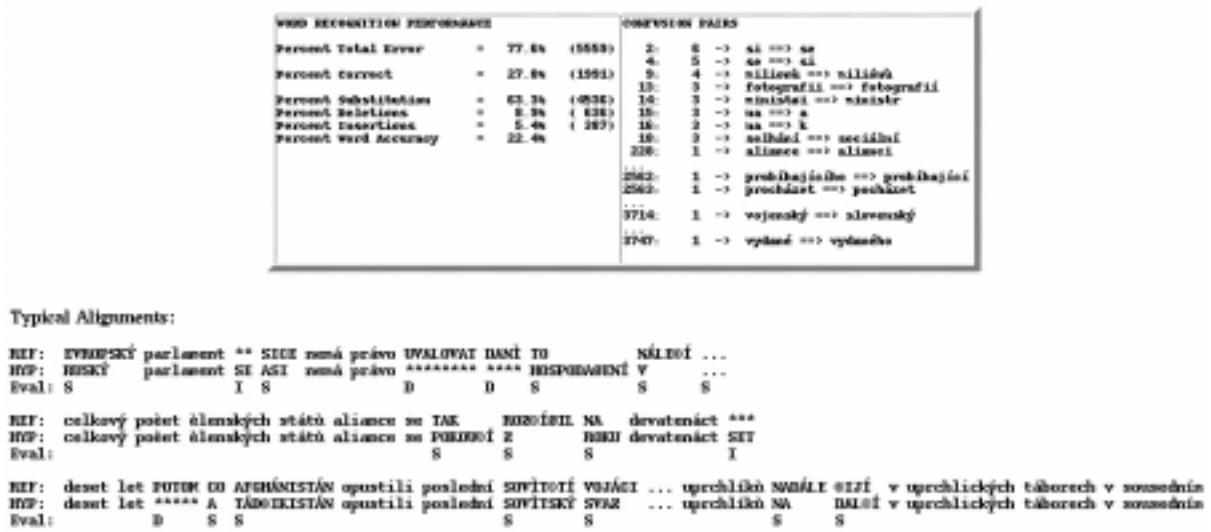


Figure 6. Two approaches to mixing multiple source language acoustic models without the use of acoustic training data. In the first approach, the recognizer is constrained at the lexical level to phones from a single source language to represent a word. In the second approach, the recognizer can mix and match phones from any source language.



WORD RECOGNITION PERFORMANCE

Percent Total Error	=	77.6%	(5553)
Percent Correct	=	27.8%	(1991)
Percent Substitution	=	63.3%	(4536)
Percent Deletions	=	8.3%	(626)
Percent Insertions	=	5.4%	(387)
Percent Word Accuracy	=	22.4%	

CONFUSION PAIRS

2:	6	->	sí ==> se
4:	5	->	se ==> sí
9:	4	->	milicek ==> milióvů
13:	3	->	fotografii ==> fotografií
14:	3	->	ministeri ==> ministr
15:	3	->	aa ==> a
16:	2	->	aa ==> k
18:	2	->	nelhůní ==> sociální
220:	1	->	aliance ==> alianci
...			
2562:	1	->	prohlhajícího ==> prohlhající
2562:	1	->	prochlzet ==> pochlzet
...			
3714:	1	->	vojenský ==> slovenský
...			
3747:	1	->	vydané ==> vydaného

Typical Alignments:

```
REF:  EVROPSKÝ parlament ** SIDE nemá právo UVALOVAT DANÌ TO          NÁLEDÍ ...
HYP:  RUSKÝ    parlament SE ASI  nemá právo ******** **** HISPODAŘENÍ V      ...
Eval: S                  I  S               D        D    S           S      S

REF:  celkový počet èlenských států aliance se TAK     BRZDÍŠEL NA   devatenáct ***
HYP:  celkový počet èlenských států aliance se POLEDNÍ Z        RŮKU devatenáct SET
Eval:                                          S      S        S               I

REF:  deset let POTOM CO AFGHÁNISTÁN opustili poslední SOVĚTŠTÍ VOJÁCI ... uprchlíků NAVÁLE ŽIJÍ  v uprchlických táborech v sousedním
HYP:  deset let ***** A  TÁDŽIKISTÁN opustili poslední SOVĚTSKÝ SVAZ   ... uprchlíků NA     DALŠÍ v uprchlických táborech v sousedním
Eval:           D     S  S                            S       S               S      S
```

Figure 7. An analysis of the performance of our best knowledge-based system. The phonetic accuracy appears to be quite good, though the word-level performance is lacking, perhaps due to language modeling issues (Czech is an inflected language).

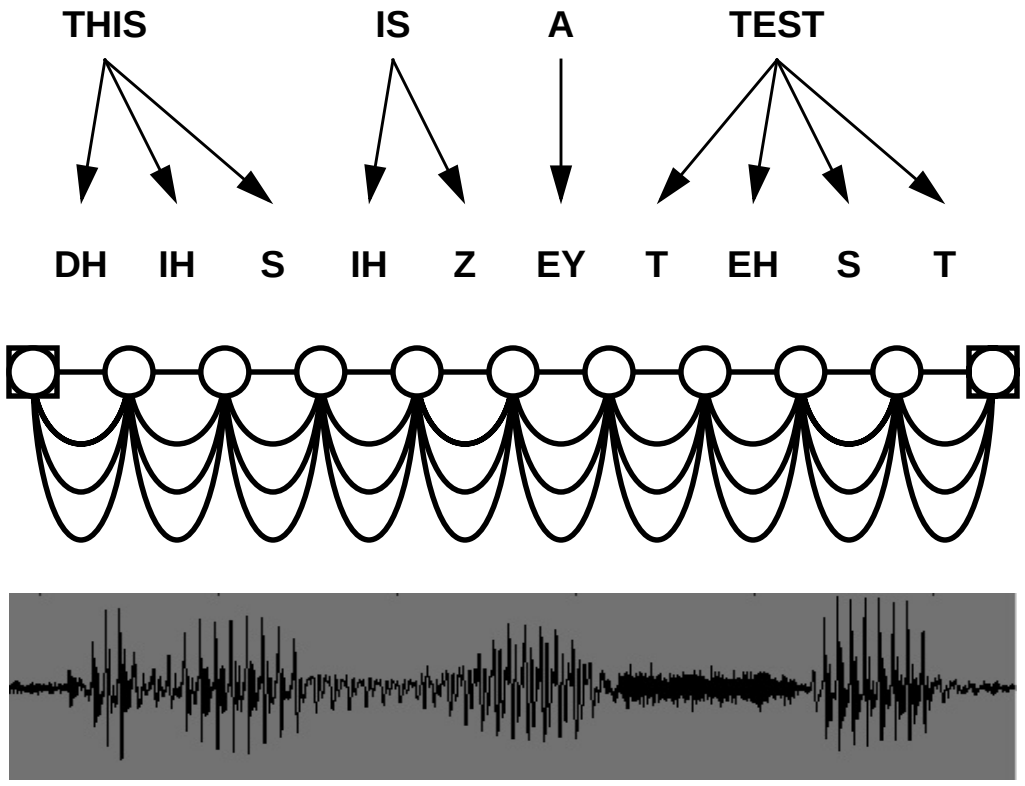


Figure 8. A forced alignment approach was combined with the multi-phone pronunciation model to produce phone confusion data for the source language training databases.

	@_RN	E_MD	I_MD	R_MD	U_MD	a_ES	a_MD	a_RN	aa_EN	aa_RN	aax_EN
@_RN	3201	0	0	0	0	0	0	0	0	0	1
E_MD	0	5639	0	0	0	7	0	0	0	0	0
I_MD	0	0	1343	0	0	0	0	0	0	0	0
R_MD	16	0	0	767	0	0	0	0	0	0	0
U_MD	0	0	0	0	1638	0	0	0	0	0	0
a_ES	0	3	0	0	0	47762	782	75	16	146	200
a_MD	1	0	0	0	0	2169	12120	37	16	40	309
a_RN	0	0	0	0	0	1660	315	3011	5	0	0
aa_EN	0	0	0	0	0	471	223	12	316	5	0
aa_RN	0	0	0	0	0	0	0	0	0	3346	100
aax_EN	0	0	0	0	0	0	0	0	0	0	5446

Figure 9. An example of phone confusion data generated on the source language training databases using the multiphone approach. The vowel “a” appears to be a good example of a vowel that varies significantly across languages, and could benefit from some sort of model combination.

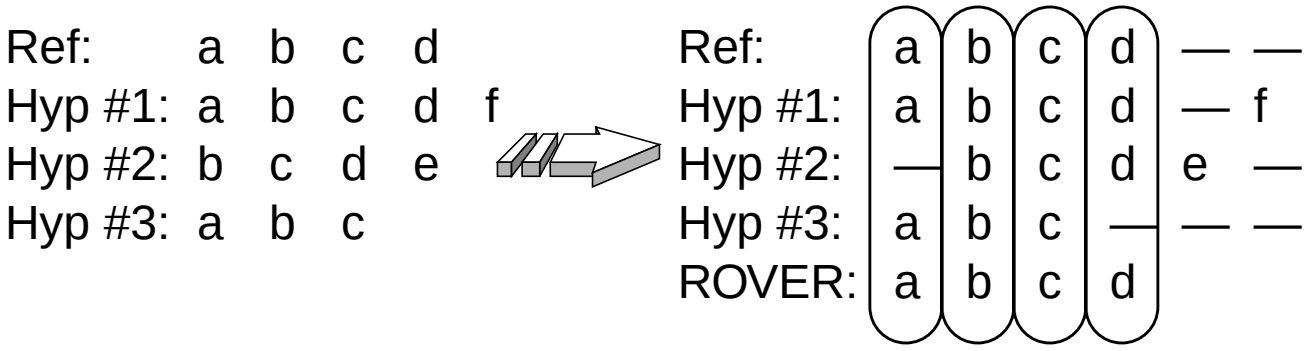


Figure 10. An overview of a method for combining hypotheses to improve system performance known as ROVER. ROVER essentially using a majority voting scheme.

System	Data
Front-End: 16 kHz Sample Frequency 10 msec frames / 25 msec window 12 Mel-Scaled Cepstral Coefficients (MFCCs) Energy	Training: English: VOA Broadcast News (10 hrs) Spanish: VOA Broadcast News (30 hrs) Mandarin: VOA Broadcast News (10 hrs) Russian: Read Speech (4 hrs)
Acoustic Models: Continuous Density HMMs Three-State, Left To Right Topology Monophones (~40) Gaussian Mixtures (~20 per state)	Held Out: Czech Voice of America Broadcasts Broadcast News: CZBN-1 (1.5 hrs) Financial News: CUCFN-1 (1 hr)
Language Model: 63K Vocabulary Bigram LM Trained on 16M Words Approx. 6M Bigrams Test Set Perplexity: 650	Eval: Czech Voice of America Broadcasts Broadcast News: VOA-1 (1 hr) Broadcast News: VOA-2 (2.5 hrs) Broadcast News: VOA-3 (1.7 hrs)

Table 1. An overview of the speech recognition system and databases used in this study.

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

Table 2. An example of a representation of a phone in terms of articulatory positions.

Czech Broadcast News		VOA-1	VOA-2
Source Language	Target Language	WER	WER
Czech (CZ)	Czech (CZ)	27.6	23.6
Russian (RN)	Czech (CZ)	65.2	60.8
Spanish (ES)	Czech (CZ)	79.3	71.7
English (EN)	Czech (CZ)	80.9	75.5
Mandarin (MD)	Czech (CZ)	91.1	88.7

Table 3. Baseline monolingual system performance.

Czech Broadcast News	VOA-1			
Source Language	WER	Dels	Subs	Ins
Mandarin - v1	91.1	15.0	72.5	3.4
Mandarin - v2	93.7	16.2	74.3	3.2
Mandarin - v3	90.1	29.8	59.4	0.9
Mandarin - v4	89.3	28.7	59.7	0.9

Table 4. 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.

Czech Broadcast News	VOA-1			
Source Language	WER	Dels	Subs	Ins
Spanish	79.3	10.6	63.5	5.3
Selective	77.7	9.1	63.1	5.6

Table 5. 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.

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

Table 6. 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.

Word-Level Performance:

Czech Broadcast News		VOA-1
Source Language	Target Language	WER
Czech (CZ)	Czech (CZ)	27.6
ROVER	Czech (CZ)	62.1
Russian (RN)	Czech (CZ)	65.2
Spanish (ES)	Czech (CZ)	79.3
English (EN)	Czech (CZ)	80.9
Mandarin (MD)	Czech (CZ)	91.1

Phone-Level Performance:

Czech Broadcast News		VOA-1
Source Language	Target Language	WER
ROVER	Czech (CZ)	38.4
Russian	Czech (CZ)	36.3

Table 7. A summary of word-level and phone-level error rates for a ROVER experiment.

Czech Broadcast News		VOA-2
Source Language	Target Language	WER
Czech (CZ)	Czech (CZ)	27.6
English (EN)	Czech (CZ)	75.5
English with Global MLLR	Czech (CZ)	70.7

Table 8. An experiment designed to investigate the impact of any channel variations shows that a simple global MLLR adaptation is not able to improve performance to a level where knowledge-based mappings are competitive with data-driven techniques.