A STUDY OF MULTILINGUAL SPEECH RECOGNITION

Fuliang Weng, Harry Bratt, Leonardo Neumeyer, and Andreas Stolcke

Speech Technology And Research Laboratory
SRI International
Menlo Park, California
http://www.speech.sri.com

ABSTRACT

This paper describes our work in developing multilingual (Swedish and English) speech recognition systems in the ATIS domain. The acoustic component of the multilingual systems is realized through sharing Gaussian codebooks across Swedish and English allophones. The language model (LM) components are constructed by training a statistical bigram model, with a common backoff node, on bilingual texts, and by combining two monolingual LMs into a probabilistic finite state grammar. This system uses a single decoder for Swedish and English sentences, and is capable of recognizing sentences with words from both languages. Preliminary experiments show that sharing acoustic models across the two languages has not resulted in improved performance, while sharing a backoff node at the LM component provides flexibility and ease in recognizing bilingual sentences at the expense of a slight increase in word error rate in some cases. As a by-product, the bilingual decoder also achieves good performance on language identification (LID).

1. INTRODUCTION

The aim of this work is to develop a multilingual speech recognizer capable of decoding a word string in any of a given set of languages. LID is achieved simultaneously. Our approach is to treat all words as equal tokens regardless of the languages they belong to. The statistics of the acoustic and language models are estimated using a multilingual speech database with orthographic transcriptions. Language-specific knowledge is incorporated into the system through the dictionary of pronunciations used by the Hidden Markov Models (HMMs) and by specifying phone classes that may contain phones from different languages. In this initial system, the phone sets do not overlap across languages. The proposed approach has some interesting characteristics:

- HMMs of allophones (of any language) that belong to the same classes and share similar contexts could potentially share the same Gaussian codebooks. This work will investigate the effect of Gaussian sharing on recognition performance.
- Multilingual LMs can be used for improving LID performance, thus allowing us to incorporate a high-level knowledge source for LID at the lexical level.
- The system is capable of recognizing sentences spoken in more than one language. Mixing words from different languages, or *code-switching*, is common within linguistic communities where there is general familiarity with more than one language.
- In real-time multilingual applications, a single decoder can be used. Alternative approaches usually do LID followed by a language-specific recognizer or require multiple recognizers to run in parallel.

Section 2 describes training issues in the multilingual system. Section 3 discusses multilingual recognition experiments. Section 4 presents LID related results. Section 5 gives a brief summary of our work and future directions.

2. MULTILINGUAL TRAINING ISSUES

2.1. Acoustic Training Issues

In this work, we experimented with bilingual (English/Swedish) recognition systems for the Air Travel Information System (ATIS) domain [6]. To build the Swedish version of the ATIS database, the English transcriptions were translated to Swedish. The Swedish prompts were then read by 100 subjects (50 male, 50 female). For rapid experimentation we used 4000 male utterances per language as training data [1].

Our main motivation for sharing acoustic parameters across the two languages is to make better use of available data in training Gaussian codebooks. That is to say, when features of the training data from the two languages are located closely in the acoustic space, they are used in training a common codebook.

One important issue in multilingual training is the *sharing granularity*: that is, at what level sharing should occur. Since many phones from the two languages are similar, we started by sharing phone classes. The phone classes were organized based on place of articulation for vowels and manner of articulation for consonants. The English and Swedish phones were grouped according to the following 11 classes: front rounded vowels, front unrounded vowels, central vowels, back vowels, diphthongs, semivowels and glides, nasals, sibilants, other fricatives, voiced plosives, and unvoiced plosives, as listed

Phone Classes	English	Swedish
Front unrounded vowels	iy y ih eh ae ey	i: i e: e ae: ae a
Front rounded vowels		y y: u: oe: oe oe2 oe2:
Central vowels	uh ax ah	u ae2 ae2:
Back vowels	aa ao ow uw w	aa aa: o o: a: ow w
Diphthongs	aw oy ay	ay
Glides	er r l	r l rl j
Nasals	m n ng	m n rn ng
Sibilants	ch jh z zh s sh	s tj rs sj sh
Fricatives	f th v dh hh	f v h th
Voiced plosives	b d g	b d rd g
Unvoiced plosives	p t k	p t rt k

Table 1. Swedish and English phone classes.

in Table 1. Notice that this table also includes some Swedish phones that are borrowed from English. These phones come predominantly from the many U.S. city names found in the ATIS corpus.

To better understand the effect of sharing model parameters across languages at the acoustic level, two contrasting sets of phonetically tied mixture (PTM) acoustic models were trained. One set of models allows English and Swedish phone classes to share Gaussian codebooks (called *shared acoustic models*); the other does *not* allow phone classes of the two languages to share common codebooks (called *non-shared acoustic models*). For the non-shared acoustic models, there is a total of 23 classes: 10 English phone classes and 11 Swedish phone classes, plus one pause phone and one reject phone. For the shared acoustic models, there are 13 phone classes, where each pair of corresponding phone classes from the two languages gets merged to form a single bilingual phone class.

In addition to the two sets of PTM acoustic models just described, two sets of genonic acoustic models were trained [2]. Notice that in a genonic system, HMM allophones of a given class share the same Gaussian codebook, and the sets of HMM states that share the same mixture components are determined automatically using agglomerative clustering techniques. These two genonic acoustic models were booted from their corresponding PTM acoustic models. Therefore, the shared genonic acoustic models allow codebooks to be shared among the phone states across the two languages, and the non-shared acoustic models forbid this.

The shared phone classes were motivated by linguistic evidence, and may not be optimal in terms of acoustic features. Ideally, if we have enough data, we should start with all the phones of the two languages and let the clustering algorithm decide which phones should be in one class. For simplicity, we took a shortcut in this process, and trained a third set of acoustic models using a new, larger, set of phone classes. These phone classes were motivated from the clustering map obtained from the agglomerative clustering in the process of training the shared genonic models. The new phone clusters used in the system are given in Table 2. We were unable, however, to obtain good results with this set of phone classes, and further investigation on this issue will be a future goal.

Phone Classes	English	Swedish
Front unrounded high vowels	iy y ih ey	i: i
Front unrounded non-high vowels	eh ae	e: e ae: ae a
Front rounded vowels		y y: u: oe: oe oe2 oe2:
Central vowels	uh ax ah	u ae2 ae2:
Back high vowels	ow uw w	o o: ow w
Back non-high vowels	aa ao	aa aa: a:
Diphthongs	aw oy ay	ay
Eng. r	axr er r	
Eng. l	l	
Swe. l		l rl
Other swe. glides		r j
Nasals	m n ng	m n rn ng
Fricatives	f th v dh hh	f v h th
Sibilants	ch jh z zh s sh	s tj rs sj sh
Labial plosives	b p	b p
Coronal plosives	d t	d t rd rt
Unvoiced plosives	g k	g k

Table 2. Swedish and English phone clusters.

2.2. Language Model Training Issues

In the construction of the LM components of the recognition systems, statistical grammars were trained in the form of bigram backoff models [4]. For the purpose of comparison, monolingual and bilingual LMs were created separately. The monolingual LMs were trained using text from a single language. The bilingual LMs were trained using the pooled English and Swedish data. The latter resulted in a bilingual LM with a single backoff node. Using a single backoff node permits hypotheses with words in both languages, and makes the system able to deal conveniently with code-switching phenomena. A negative side effect of this shared backoff node is the increased possibility of confusion among words from the two languages, as will be shown in the experimental results described below. For training, we used the available in-domain English text with 220,000 words and in-domain Swedish text with 100,000 words. Although there is twice as much English than Swedish training data, the results in the next section show that the English test set still has a higher perplexity than the Swedish test set, and word error on the English data is somewhat higher. An alternative would have been to use equal amounts of training data for both languages, but this would have resulted in less balanced recognition performance.

Another bilingual constrained LM (LM_C) was built by combining the two monolingual LMs into a probabilistic finite state grammar (PFSG) as shown in Figure 1. The two monolingual LMs share the same initial and final nodes, but there is no transition going from the English subgrammar to the Swedish subgrammar or vice-versa. Therefore, the resulting bilingual LM should behave

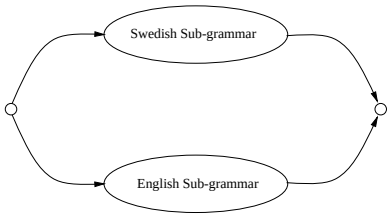

Figure 1. The constrained LM (LM_C).

Test Lang.	Shared Acoustic Model	Shared Language Model	PTM	Genones	Genones with LM_C
English	No	No	7.40	7.04	
	Yes	No	7.79	6.89	
	No	Yes	7.73	7.25	7.21
	Yes	Yes	8.28	7.12	6.93
Swedish	No	No	6.72	6.03	
	Yes	No	7.15	6.03	
	No	Yes	7.96	7.02	6.29
	Yes	Yes	8.00	7.19	6.08

Table 3. English/Swedish word error rates for various speech recognition systems.

similarly to the two monolingual LMs in most cases. That is, it will not allow a hypothesis with words from both languages, and the constrained LM score for a single language hypothesis will be the same as the corresponding monolingual LM score for that hypothesis. Effectively, this approach allows us to run the two language-specific recognizers in parallel, choosing the language whose best hypothesis gives the higher score[5, 3]. We will discuss it further in Section 4.

3. MULTILINGUAL RECOGNITION

In the recognition experiments, six PTM and eight genonic speech systems were constructed. All the English experiments were tested on a set of 443 sentences with 4660 words, while all the Swedish experiments were tested on a set of 267 sentences with 2337 words.

Of the six PTM systems, four systems consist of bilingual acoustic components and monolingual LM components. The remaining two of the six PTM systems have both bilingual acoustic and language model components: one of them has shared acoustic parameters across the two languages, and the other one does not. The bilingual PTM system with shared acoustic parameters uses 13 phone classes. In this case, phones in the same class share the same Gaussian codebook. The non-shared PTM system is trained using 23 classes. Each language-specific set of phones has a separate codebook.

Test	No. Train Sent.	Vocab Size	OOV (%)	Perplexity	
				Non-Shared LM	Shared LM
Eng	20K	1662	0.2	22.4	23.8
Swe	11K	1264	0.3	14.9	17.7

Table 4. Comparison of English and Swedish language models.

Six of the eight genonic systems were booted from their corresponding PTM systems. The shared system has the same amount of Gaussian components as the non-shared system, to maintain a constant ratio of Gaussian components to training vectors. The other two genonic systems have shared acoustic components and use the constrained LM (LM_C), described in Section 2.

In addition to these systems, we have trained English-only and Swedish-only systems. The recognition results are the same as the systems with bilingual non-shared acoustic components and monolingual LM components, as expected. Therefore, no details of the results are given here.

The initial results for the 16 systems are summarized in Table 3. The *Shared Acoustic Model* column indicates whether Gaussian codebooks are shared across languages in the acoustic components of the 16 systems. A "Yes" in the *Shared Language Model* column means that the LM has a single backoff node shared by both languages and the system uses a single decoder for the two languages, while "No" in this column indicates that the system uses a monolingual LM. The *PTM*, *Genones* and *Genones with LM_C* columns indicate that the systems are PTM systems, genonic systems, and genonic systems with LM_C as their LM components, respectively.

From the table, it is clear that the genonic system significantly outperforms the PTM system in most cases, as expected. However, we must point out that these differences in the word recognition accuracy between the PTM and genonic systems are limited by the total amount of training data available (4000 utterances per language).

We also observe that sharing acoustic parameters does not seem to affect the word error rate in general. On the other hand, using the bilingual LM, with a common backoff node, built from bilingual text results in a significant degradation in performance. This degradation is more significant for the Swedish test set: 6.03% to 7.02% for the genonic system with no acoustic parameter sharing. The same case in English results in an increase from 7.04% to 7.25%, which is statistically insignificant. This result could be associated with the unbalanced amounts of English and Swedish LM training data and the greater increase in perplexity in the Swedish test set compared to the English test set (see Table 4).

The LM_C language model described in Section 2 was constructed to test this hypothesis by balancing the two monolingual LMs. The recognition results in the LM_C column in Table 3 show that there is no difference compared with the corresponding systems with the monolingual components.

To provide the functionality of code-switching and to keep the word error rate low, one N-Best rescoring experiment was conducted using an additional knowledge source, language identity score (LIS). This knowledge source is used to penalize (but not to forbid) the recognition

Test Lang.	Non-Shared Acoustic Models		Shared Acoustic Models	
	Word Miss (%)	Sent Miss (%)	Word Miss (%)	Sent Miss (%)
English	0.4	2.7	0.5	2.9
Swedish	0.8	3.7	0.8	5.2

Table 5. Recognition errors with words from both languages.

Test Lang.	Non-Shared Acoustic Models	Shared Acoustic Models
	Sent Miss (%)	Sent Miss (%)
English	0	0.2
Swedish	0.4	0

Table 6. Language identification errors after taking simple majority of words in hypothesis.

hypotheses that have words from both languages. This language identity score is defined as follows:

$$\mathrm{LIS}(X) \quad = \quad \frac{\max_{i \in \{E,S\}} \text{words in } X \text{ from language } i}{\text{total number of words in } X}$$

where *E* refers to English and *S* to Swedish, and *X* is an utterance.

Results show that there is only a small improvement in word error rate for N-Best rescoring using LIS, even though fewer 1-best hypotheses contain words from both languages after rescoring.

To conclude, the bilingual system provides an efficient way of pruning unlikely hypotheses from the two languages by using a single Viterbi decoder. The shared acoustic models lead to a compact system, although it has not improved recognition accuracy so far. Sharing LMs across languages offers the flexibility for code-switching at the expense of a slight increase in word error rate in some cases.

4. LANGUAGE IDENTIFICATION

We analyzed the LID performance of the bilingual systems. Table 5 shows the percentage of words and sentences that contain a word in the other language for the LM trained on bilingual texts. We observe that less than 1% of the recognized words have the wrong language identity. Even better LID performance can be obtained by taking a simple majority of the words in a hypothesis (Table 6). The constrained LM also achieves very good LID performance (Table 7), although our task is not directly comparable to those commonly used in LID research [7]. Compared with LID systems using multiple large vocabulary continuous speech recognizers [5, 3], our system uses a single Viterbi algorithm to prune hypotheses in a multilingual space, which enables us to eliminate unlikely language candidates at an early stage. Furthermore, because of the sharing of acoustic models, our system is more compact and offers real time performance.

Test Lang.	Non-Shared Acoustic Models	Shared Acoustic Models
	Sent Miss (%)	Sent Miss (%)
English	0	0
Swedish	0.7	0.4

Table 7. Language identification errors when using the constrained LM.

5. SUMMARY

We investigated the effect of sharing acoustic and language models for multilingual speech recognition. Results show that sharing parameters across two languages maintains good performance. As a by-product, the bilingual systems also show good results on LID. In future experiments, we will research various ways of optimizing the unconstrained bilingual language model and new approaches for sharing acoustic models across languages, and we will also include more languages in the system.

ACKNOWLEDGMENT

The authors would like to express their thanks to Jaan Kaja for discussions in creating Swedish phone classes. This work was part of the SLT project sponsored by Telia Research AB in Sweden.

REFERENCES

[1] D. Carter, J. Kaja, L. Neumeyer, M. Rayner, F. Weng, and M. Wiren. Handling compound nouns in a Swedish speech-understanding system. In *Proceedings of ICSLP-96*, 1996.

[2] V. Digalakis, P. Monaco, and H. Murveit. Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognizers. *IEEE Transactions on Speech and Audio Processing*, pp. 281–289, 1996.

[3] J. Hieronymus and S. Kadambe. Robust spoken language identification using large vocabulary speech recognition. In *Proceedings of ICASSP-97*, 1997.

[4] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3), 1987.

[5] S. Lowe, A. Demedts, L. Gillick, M. Mandel, and B. Peskin. Language identification via large vocabulary speaker independent continuous speech recognition. In *Proceedings of ARPA Human Language Technology Workshop*, 1994.

[6] P. J. Price. Evaluation of spoken language systems: the ATIS domain. In *Proceedings of 1990 DARPA Workshop on Speech and Natural Language*, 1990.

[7] M. Zissman and A. Martin. Language identification overview. In *Proceedings of the Fifteenth Annual Speech Research Symposium*, 1995.