TOWARDS LANGUAGE INDEPENDENT ACOUSTIC MODELING

P. Beyerlein[1], W. Byrne[2], J. M. Huerta[3], S. Khudanpur[2], B. Marthi[4],
J. Morgan[5], N. Peterek[6], J. Picone[7], W. Wang[8]

Philips Research Laboratories (1)   CLSP, Johns Hopkins University (2)
Dept. ECE, Carnegie Melon University (3)   Depts. CS and Math, University of Toronto (4)
Dept. Foreign Languages, USAMA, West Point (5)   UFAL, Charles University, Prague (6)
ISIP, Mississippi State University (7)   Dept. ECE, Rice University (8)

ABSTRACT

We describe procedures and experimental results using speech from diverse source languages to build an ASR system for a single target language. This work is intended to improve ASR in languages for which large amounts of training data are not available. We have developed both knowledge based and automatic methods to map phonetic units from the source languages to the target language. We employed HMM adaptation techniques and Discriminative Model Combination to combine acoustic models from the individual source languages for recognition of speech in the target language. Experiments are described in which Czech Broadcast News is transcribed using acoustic models trained from small amounts of Czech read speech augmented by English, Spanish, Russian, and Mandarin acoustic models.

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 the 'resource-rich' languages which have such corpora, these statistical 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.

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. This is called 'language independent acoustic modeling' to suggest a similarity in nature to speaker independent modeling. In the current state-of-the-art, speaker independent models are first trained from multiple speakers and then adapted to a specific speaker either before or during recognition. Analogously, language independent modeling is a methodology that combines speech and models from multiple source languages and transforms them for recognition in a specific target language.

As mentioned above, acoustic training data is only one resource needed for statistical ASR. However, we have assumed 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, however, in that dictionaries with pronunciations are available for many languages, as are on-line newspapers and other text. However, we stress that we address here only one aspect of language independent modeling.

We have developed methods to share data and acoustic models between languages. Underlying these methods are 'phone mappings' that describe the similarity of sounds in two different languages. We obtain these phone mappings using both *knowledge-based* and *automatic* methods. The knowledge-based methods rely only on acoustic-phonetic 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. By either approach we can borrow models from several languages simultaneously to cover the phone inventory of the target language. 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 models from several languages simultaneously. 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 by themselves do not perform well in transcribing the target language.

We present below a necessarily brief description of our experiments. Our web site *www.clsp.jhu.edu/ws99/projects/asr* contains complete documentation of our work, some of the language data and models used, and a more extensive bibliography of prior work in language independent and multilingual acoustic modeling.

2. MULTILINGUAL TRAINING AND TEST SETS

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

We chose Czech language Voice of America (VOA) broadcasts as our test domain since news broadcasts contain a variety of different types of speech and are relatively easy to obtain. We chose Czech since we have ongoing projects [2] from which we could borrow resources. We also felt that studying Czech as a rapid-porting task was realistic since, unlike Spanish or Mandarin, there is fairly little knowledge of existing Czech ASR to influence our work. Our final test set consisted of one week of news broadcasts, although due to evolution of our experiments, not all the numbers reported below are directly comparable; see our web site for more detailed reporting.

As our out-of-domain acoustic training data, we used broadcast news recordings in English, Spanish, and Mandarin obtained from the Linguistic Data Consortium. We also used read Russian speech collected at West Point for computer aided foreign language instruction and read Czech speech from the Charles University Corpus of Financial News (CUCFN). All speech was down-sampled to 16KHz as needed. The acoustic models were trained from mel-frequency, cepstral data using HTK [6]. Unless otherwise noted, the source language acoustic models were monophone systems to simplify cross-language mapping; full system descriptions are on our web site.

We built 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 taken from our previous work [2]. After obtaining the performance of this well-trained system, we reduced drastically the size of the acoustic training set and retrained new, impoverished acoustic models. 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 10 hour system. We would then attempt to improve this impoverished system by borrowing from other languages. However, as Table 1 shows, performance on Czech VOA is relatively good despite large variations in training set size and model complexity. This behavior appears to be due to the extremely regular and careful speech used by Czech VOA announcers and not due to a preponderance of speech by individual news anchors or other obvious similarities between training and test sets. We note that we observed similar behavior in experiments with Spanish VOA broadcasts.

From these results we concluded that the Czech VOA speech

Training Data	Model type	WER (%)
12.8 hour	12 mixture, cross-word triphone	27.1
10.0 hour	20 mixture, monophone	27.6
1.0 hour	8 mixture, monophone	30.2
0.5 hour	20 mixture, monophone	31.3

Table 1: Training and Testing on Czech VOA Broadcasts.

Training Set	CUCFN	VOA
1.0 hr VOA	66.1%	28.8%
1.0 hr CUCFN	47.3%	35.7%

Table 2: WER in Training and Testing on Czech VOA Broadcasts and CUCFN Read Speech Using 20 Mixture Monophone Models.

was too self-similar to be used as both training and test data. We therefore investigated a cross-domain training scenario in which a small amount of read speech from the CUCFN corpus would serve as the Czech language training data. After comparing performance across the mono-lingual Czech read and broadcast domains (Table 2), we decided to fix the 1.0 hour CUCFN read speech training set as the Czech language acoustic training set and to attempt to improve performance on the Czech VOA test data by borrowing from English, Mandarin, Spanish and Russian. This provides a realistic and interesting training scenario that involves cross-domain as well as multilingual factors.

These experiments with Czech VOA are reported as a cautionary note 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. KNOWLEDGE-BASED PHONE MAPPINGS

In some applications, it is highly desirable to develop speech recognition systems without any acoustic training data. In such situations, borrowing models from other languages for which speech recognition technology is well-developed is an attractive idea. 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.

Our initial experiments involved simple mappings 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 [3]. 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) [4].

The advantage of this approach is that all languages can, in theory, be represented within the same system. We 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. 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. We used these mappings to obtain baseline performance 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 map-

Source Language : Czech VOA WER (%)	
Russian : 60.8	Spanish : 71.7
English : 75.5	Mandarin : 88.7

Table 3: Performance Using Knowledge Based Phone Mappings.

pings. 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 VOA data, especially considering the differences in microphones, speaking style, and speaking rates. We also observed from these experiments that performance for English and Spanish was comparable, and performance for Mandarin lags the other systems.

It was evident from the construction of the mappings 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. Though we achieved modest improvements in performance (1.6% absolute WER), we did not achieve performance comparable to data-driven mapping methods discussed next.

Our next attempt to understand deficiencies in the knowledge-based system was to explore a series of experiments in which the recognition system was allowed to chose the best combination of phones at runtime. First, we explored a parallel pronunciation approach [5] in which each item in the lexicon was represented as a sequence of phones from a single language implemented using pronunciation networks. Unfortunately, this approach resulted in slightly degraded performance even though we had hoped that the additional degrees of freedom would offset any systematic acoustic bias between the two domains. We next tried a *multiphone* approach that allowed the recognition system to mix and match phones from all source languages as an attempt to let the recognizer find the best realization of a phone, rather than fixing this based on a priori linguistic knowledge. We found minor improvement in performance over the parallel pronunciation system, as expected. However, overall performance is still below the best monolingual system, and far below the Russian monolingual system. In these experiments 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 be better if the majority of the errors occur on endings 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.

4. AUTOMATIC GENERATION OF PHONE AND STATE LEVEL ACOUSTIC MAPPINGS ACROSS LANGUAGES

We developed a general methodology to derive cross-language mappings automatically both at phonetic and sub-phonetic levels. We call our approach the *Confusion Matrix* approach to finding cross-lingual mappings. These confusion matrices are tables of acoustic similarity between phones across languages. They are obtained by first performing a mono-lingual phonetic labeling of the target language acoustic data using the target language phone set - this can be done manually or via forced-alignment using HMMs; we use the latter approach. Phonetic recognition of this data is then performed using acoustic models from each of the source languages; for this we used simple, unweighted, phone-loop recognizers. This yields multiple phonetic segmentations of the target language acoustic data in the source language phone inventories.

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 nth 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 nth 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 kth 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 seemed hard to match from one language to another. 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 described above how we established the best mapping for each phone/state of the target language. We found out 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 in proportion to the normalized number of counts corresponding to the map.

Table 4 shows recognition experiments we conducted using mappings derived from confusion matrices. For comparison in this experiment, monophone Czech models trained on 1 hour of Czech give 38% WER. When mappings are obtained using the phone-level confusion matrix approach, the word error rate drops below 70%. State-level mappings further reduce the error rate of the English mappings. Better results are obtained when multiple source languages are included (English, Spanish and Mandarin), and state mappings are obtained for both state-to-state mapping and best three states to a single Czech state (the 3-state method). The best result is below 55% WER. The 3-state methods reported differ in the presence (54.4%) or absence (55.8%) of count normalization of the columns in the confusion matrix.

Source(s)/Method	WER	Source(s)/Method	WER
EN/Phone	68.3	SP/Phone	68.7
EN/State	64.8	SP/State	70.0
MA/State	79.7	EN+SP+MA/State	62.3
EN+SP+MA/3-State	55.8	EN+SP+MA/3-State	54.4

Table 4: WER(%) Using Automatic Phone Mappings.

Acoustic Scores and Phonetic Classes	WER(%)
N-Best oracle	19.8
first best (baseline)	34.0
$V_{ru} + C_{ru} + S_{ru} + V_{sp} + C_{sp} + S_{sp}$	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 6: DMC Rescoring of 1000-best Lists. The combination uses knowledge based mappings, the Czech language model, and the Czech, Spanish, Russian, and English vowel, consonant and silence models

5. ACOUSTIC 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 hour of Czech read speech largely compensates for these differences, as shown in Table 5. Furthermore, while performance improves significantly, the adapted systems do not individually improve over the monolingual Czech systems.

Source	Mixtures / Type	Unadapted	MLLR+MAP
MA 10 hr.	20 /monophone	88.7	63.0
SP 10 hr.	20 / monophone	71.6	50.9
RU 3 hr.	20 / monophone	60.8	45.3
EN 10 hr.	20 / monophone	75.7	47.2
EN 10 hr.	8 / triphone		35.1
EN 72 hr.	12 / triphone		32.7
CZ 1 hr.	20 / monophone	33.4	
CZ 1 hr.	6 / triphone	30.7	

Table 5: Adaptation WER(%) of Systems with Varying Complexities and Amounts of Source Language Training Data

6. DISCRIMINATIVE MODEL COMBINATION OF MULTIPLE SOURCE LANGUAGE ACOUSTIC MODELS

Discriminative model combination [1] aims at an optimal integration of all available acoustic and language models into one log-linear posterior probability distribution. The coefficients of the log-linear combination are estimated on training samples using discriminative methods to obtain an optimal classifier. For example, a multilingual combination at the sentence level of scores from Czech, Spanish, and Mandarin acoustic models has the following form for a sentence hypothesis w given the acoustic data x

$$\lambda_{lm} L_{cz}(w) + \lambda_{cz} A_{cz}(x|w) + \lambda_{sp} A_{sp}(x|w) + \lambda_{ma} A_{ma}(x|w)$$

where $L_{cz}(w)$ is the Czech language model likelihood, $A_{cz}(x|w)$, $A_{sp}(x|w)$, $A_{ma}(x|w)$ are the Czech, Spanish, and Mandarin acoustic model likelihoods. The parameters λ are optimized to minimize WER on a held-out set of Czech data.

Although the results are not reported in detail here, we find that DMC rescoring at the sentence level does not improve over the monolingual Czech performance. However, performance can be improved by applying DMC at the phoneme-class level. For example, the acoustic likelihood $A_{cz}(x|k)$ can be separated by the

---

[1] References and procedures are in the HTK documentation [6].

contribution of vowels, consonants, and silence models. Parameters can then be introduced to define a posterior distribution based on these language-specific phonetic classes:

$$\lambda_{lm} L_{cz}(k) + \lambda_{cz,V} V_{cz}(x|k) + \lambda_{cz,C} C_{cz}(x|k) + \lambda_{cz,S} S_{cz}(x|k).$$

From the results in Table 6 we conclude that the structuring into phoneme classes improves performance over combination at the sentence level. Furthermore, combination of multilingual phoneme-class models performs better than the monolingual Czech systems, even when the monolingual systems are optimized using DMC.

7. CONCLUSION

We have presented a methodology for language independent acoustic modeling. We found that both knowledge-based and automatic methods can be used to derive cross-lingual phonetic mappings. Model adaptation and discriminative model combination can then be used to further improve and merge systems from diverse languages. Additional experiments, particularly in language adaptive training, can be found on our web site.

REFERENCES

[1] P. Beyerlein, "Discriminative Model Combination", ICASSP, Seattle, 1998.

[2] W. Byrne *et al.* "Large Vocabulary Speech Recognition for Read and Broadcast Czech", 1999 Workshop on Text Speech and Dialog, Marianske Lazne, Czech Republic.

[3] D. Calvert, Descriptive Phonetics, Thieme, New York, 1986.

[4] Handbook of the International Phonetic Alphabet, Cambridge University Press, Cambridge, UK, 1999.

[5] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Large Vocabulary Speech Recognition," IC-SLP, Sydney, Australia, 1998.

[6] S. Young *et al.* The HTK Book, Entropic, Inc. 1999.