TOWARDS LANGUAGE INDEPENDENT ACOUSTIC MODELING

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

CLSP, Johns Hopkins University (1)
Philips Research Laboratories (2)
Dept. ECE, Carnegie Mellon 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 other than the well-studied European and Asian languages as spoken by the majority populations of Europe, Asia, and America. 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 hoped instead to develop a good monolingual system for a specific target language by borrowing data and models from other languages. Calling this 'language independent acoustic modeling' is meant to suggest a similarity with 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.

Our work focused on the development of 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 states of different source language phone models.

While both the automatic and knowledge-based phone mappings can be used directly to construct recognizers in the target language by borrowing acoustic models from the various source languages, we found it beneficial to use HMM adaptation techniques to improve the source language 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) [1] 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.

This work was supported by the National Science Foundation under Grant No. #IIS-9820687, and carried out at the 1999 Workshop on Language Engineering, Center for Language and Speech Processing, Johns Hopkins University.

We present here a brief description of our experiments. Our web site *www.clsp.jhu.edu/ws99/projects/asr* contains a final report 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 (*e.g.* [2, 3, 4, 5, 6, 7, 8]). This paper expands upon an earlier report of our work [9] and presents additional results in cross-language adaptation and discriminative model combination for multilingual acoustic modeling.

2. TRAINING AND TEST SETS FOR MULTILINGUAL ACOUSTIC MODELING

As part of our research program we established an experimental framework for language independent acoustic modeling. We selected 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 [10] from which we could borrow resources. We also felt that studying Czech as a rapid-porting task was realistic since, unlike English, Spanish or Mandarin, there is relatively 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 results reported here are directly comparable; see our web site for more detailed reporting.

As our source-language 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 downsampled to 16KHz as needed. The acoustic models were trained from mel-frequency, cepstral data using HTK [11]. Unless otherwise noted, the source language acoustic models were multiple mixture monophone systems to simplify cross-language mapping; full system descriptions are on our web site.

After comparing performance across the monolingual Czech read and broadcast domains, we decided to use 1.0 hour of CUCFN read speech as our Czech acoustic training set. After first training monolingual Czech ASR systems using this one hour of speech, our goal was to improve their performance on Czech VOA 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. The experimental results that led to this decision can be found in our final report (see also [9]), however we note that language is just one characteristic of speech and that other conditions, such as speaking style, are also significant factors in ASR performance. It is therefore critically important to obtain diverse training and test sets for multilingual experiments to ensure that cross-language effects are not dominated by other factors. As a related point, it is also important that results of limited domain experiments, such as training and testing with data from the same news program, 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 for new languages 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 we studied to address this problem are referred to as knowledge-based because they exploit linguistic knowledge of the languages and their phoneme inventories, and because they do not involve retraining using 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 when available. Our approach involved first describing the phones in both the source and target languages in terms of their articulatory positions; a complete inventory of the features and mappings used is available online.

We then 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 mappings. Overall, we observed 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, performed relatively well on the Czech 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.

4. AUTOMATIC GENERATION OF PHONE AND STATE LEVEL ACOUSTIC MAPPINGS

We next investigated a general methodology that makes use of small amounts of target language speech 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 monolingual 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 used 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 parallel 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 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

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

Table 1: WER(%) Using Automatic Phone Mappings.

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 seemed hard to match across languages. To obtain the sub-phonetic mapping, we broke each HMM in the source and target language into its states and derived single state HMMs 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 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 might also be considered as matching candidates. 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 1 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.

4.1. Language Adaptive Clustering

We examined a novel method to find cross-lingual phone mappings using a modified version of vector quantization [12]. The key feature here is that we allow the source language data to be acted upon by language-specific transformations. The goal is to learn transformations that normalize acoustic variability across languages while performing phone clustering.

We used a modified VQ objective function to incorporate these transformations. Let $x_i^{p,l}$ denote the i^{th} sample of phone p from language l. The quality of a set of codewords $C = \{C_k\}$ and transforms $\{T^{p,l}\}$ is measured on on a set of Czech, Spanish, Russian, Mandarin, and English data as $\sum_p \min_{C_k \in C} \sum_i |C_k - x_i^{p,cz}|^2 + \sum_{l \in \{ma,sp,ru,en\}} \min_{C_k \in C} \min_{T^{p,l}} \sum_i |C_k - T^{p,l}(x_i^{p,l})|^2$. 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 so that, after recomputing the centroids and clusters, the transformations were recomputed given the new centroids and clusters.

Given one hour of aligned data in each language, we found that this technique was comparable to the phone-level automatic methods described above. We found that it is crucial to apply the transforms learned during clustering. For example, one mapping initially yielded 86.4% WER, which was reduced to 71.6% WER after transformation of the source language HMM mean vectors by the additive shifts found during clustering. Encouraged by this, we applied the per-phone additive transformations to the best 3-state automatic alignment described above and found that the reported WER fell from 54.4% to 48.8%. While this approach falls far short of incorporating cross-language modeling into acoustic training, it suggests that simple normalization techniques can capture significant cross-language variability.

Source / Mixtures / Type	Unadapted	Adapted
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	78.8	32.6 †
EN 72 hr. / 12 / triphone	72.1	32.7 ‡
CZ 1 hr. / 20 / monophone	33.4 †	-
CZ 1 hr. / 6 / triphone	30.7 ‡	-

Table 2: WER(%) After MLLR+MAP Adaptation of Source Language Systems Using 1 Hour of Read Czech.

5. ACOUSTIC ADAPTATION

We explored acoustic model adaptation as a way to transform well-trained source language HMM systems using only a small amount of the target language training data. In these experiments we found that despite the substantial variation in the quality of the phone mappings obtained by knowledge-based and automatic state-level phone mappings, adaptation using MLLR and MAP (see the HTK Book [11] for a description of the methods used) on the 1.0 hour of Czech read speech largely compensated for these differences, as shown in Table 2. Furthermore, while performance improves significantly, the adapted systems do not individually improve over the best monolingual Czech systems.

6. DISCRIMINATIVE MODEL COMBINATION OF MULTIPLE SOURCE LANGUAGE ACOUSTIC MODELS

We explored the use of DMC [1] to improve upon the adaptation of single language source ASR systems. DMC aims at an

DMC Monophone Baseline: $L_{cz}, A_{cz\dagger}$	30.9
$L_{cz}, VCS_{cz\dagger}$	30.8
$L_{cz}, VCS_{ru\dagger}, VCS_{sp\dagger}, VCS_{cz\dagger}$	30.8
$L_{cz}, A_{cz\dagger}, A_{ru\dagger}, A_{sp\dagger}, A_{en\dagger}$	28.9
$L_{cz}, VCS_{cz\dagger}, VCS_{ru\dagger}, VCS_{sp\dagger}, VCS_{en\dagger}$	28.5
DMC Triphone Baseline: $L_{cz}, A_{cz\ddagger}$	28.1
$L_{cz}, A_{en\ddagger}, A_{cz\ddagger}$	27.4
$L_{cz}, VCS_{en\ddagger}, VCS_{cz\ddagger}$	27.1
N-Best oracle	19.8

Table 3: WER(%) in DMC Rescoring of 1000-Best Lists.

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

We have found that DMC rescoring at the sentence level improves over the monolingual Czech performance and that it is also possible to apply DMC at the phoneme-class level for further improvement. For example, the acoustic likelihood $A_{cz}(x|k)$ can be separated by the contribution of vowels, consonants, and silence models. Parameters can then be introduced to define a posterior distribution based on these language-specific phonetic classes: $\lambda_{cz,V}V_{cz}(x|k) + \lambda_{cz,C}C_{cz}(x|k) + \lambda_{cz,S}S_{cz}(x|k)$. For brevity, this is denoted as $\lambda_{cz} \cdot VCS_{cz}$.

The source language systems described in Table 2 were combined at the sentence level and at the phone class level. In these experiments we combine acoustic scores obtained by forced alignment using N-Best hypotheses generated by the Czech monophone system. The results reported in Table 3 are based on direct optimization of the DMC WER using the simplex downhill method, known as amoeba search [13]. We found this produced better results than methods that approximate WER by a smooth cost function (see our earlier results [9]); we hypothesize that for a small number of parameters, direct minimization is effective and avoids approximation errors. We found 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, *i.e.* optimized over L_{cz} and A_{cz}.

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 subphonetic 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.

Acknowledgement We thank M. Riley and F. Pereira of ATT for use of their large vocabulary decoder.

8. REFERENCES

[1] P. Beyerlein, "Discriminative model combination," *Proc. ICASSP*, pp. 481–484, 1998.

[2] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," *Proc. EUROSPEECH*, pp. 371–374, 1997.

[3] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy, "An evaluation of cross-language adaptation for rapid HMM development in a new language," *Proc. ICASSP*, pp. 237–240, 1994.

[4] P. Cohen, *et al.*, "Towards a universal speech recognizer for multiple languages," *Proc. ASRU*, pp. 591–598, 1997.

[5] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *Proc. ICASSP*, pp. 1819–1822, 1998.

[6] J. Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *Proc. ICASSP*, pp. 2195–2198, 1996.

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

[8] A. Constantinescu and G. Chollet, "On cross-language experiments and data-driven units for ALISP," *Proc. ASRU*, pp. 606–613, 1997.

[9] P. Beyerlein, *et al.*, "Towards language independent acoustic modeling," *Proc. ASRU*, 1999.

[10] W. Byrne, *et al.* "Large vocabulary speech recognition for read and broadcast Czech," *Proc. Wkshp. on Text Speech and Dialog, Marianske Lazne, Czech Republic*, 1999.

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

[12] R. M. Gray, "Vector quantization," *IEEE ASSP Magazine*, pp. 4–29, April 1984.

[13] J. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, vol. 7, pp. 308–313, 1965.