Training Acoustic Models with Speech Data from Different Languages

Chen Liu, Lynette Melnar

Human Interaction Research, Motorola
1295 E. Algonquin Road, Schaumburg, IL 60196, USA
Chen.Liu, Lynette.Melnar@motorola.com

Abstract

We present a technique to train acoustic models for a target language using speech data from distinct source languages. In this approach, no native training data from the target language is required. The acoustic model candidates for each target-language phoneme are automatically selected from a group of existing source languages by means of a combined phonetic-phonological (CPP) metric, developed by incorporating statistically-derived phonetic and phonological distance information (Liu and Melnar, Interspeech 2005). The method assumes availability of sufficient native training data for the source languages and pronunciation lexica for both the target and source languages. Once the model candidates are determined for each target-language phoneme, the target HMMs are trained with the speech data from the source languages by means of a "silkie-hen-on-duck-eggs" strategy – namely the target phoneme model training is embedded in the source phoneme model training. The recognition performance of the resultant models is comparable to that of our previously-reported CPP-derived models built through multi-mixture construction while the size of the current models is only a fraction of the previous models, depending on the number of HMM candidates used for each target phoneme. Utilizing the CPP metric, both versions of the models reach the performance of models generated by a data-driven acoustic-distance mapping approach, far above the general phoneme symbol-based cross-language transfer strategies.

1. Introduction

As ASR-enabled voice products are quickly gaining importance and popularity, building acoustic models efficiently and effectively for a new language becomes a required technology. However, acquiring sufficient native speech data can be costly and time consuming, and may become a bottleneck to model development. Although speech data is commercially available for many world languages in established markets, numerous voice products are increasingly targeting new markets that are often associated with resource-poor languages.

Researchers have long been striving to produce a practical cross-language solution to the problem of data unavailability. Efforts have largely been focused on borrowing resources from resource-rich languages to build acoustic models for a resource-poor language. Representative approaches are roughly divided into two categories, linguistic and acoustic. In the former, phonetic or phonological knowledge is used to select phonemes from the source languages as substitutes for the target-language phonemes. The selection procedure is normally determined either through IPA symbol commonality [1][2] or through a more in-depth consideration of phonological factors by a language expert [3]. On the other hand, the acoustic approach is data-driven, where some native data from the target language is required. The native data is either employed to further improve the performance of the models built with the linguistic approach, or it is used to train raw models, which in turn are used to locate the best candidate models in the source languages through acoustic distance measurement [3]. Generally, data-driven approaches have yielded cross-language models with superior performance relative to approaches using only knowledge-based phoneme mappings [4][5].

However, we previously demonstrated an automated linguistic approach with performance results comparable to acoustic approaches [6][7] that selects the best phoneme candidates efficiently and consistently without any access to target-language speech data. The candidate selection is based on a combined phonetic-phonological (CPP) metric which incorporates phonetic and phonological distance information. Obtaining inter-phoneme distance results solely based on statistically derived linguistic knowledge, the CPP metric is shown to be effective in characterizing similarity between phonemes across languages. The recognition performance of the target models built with the selected candidates reaches the performance level of acoustic approaches.

One practical issue with the previous approach is the resulting large model size. The target models are built directly from the well-trained source models; specifically, each target HMM is a multimixture assembly of the corresponding candidate HMMs. Our experiments show that the performance of the target models increases with the number of source models used for each target phoneme. To obtain a superior performance, normally two or three candidates are used to assemble a target HMM; hence the size of the target HMMs may double or triple. This consequently increases the demand on memory space and also slows down the computation speed by the same magnitude.

In this paper, we present a cross-language training technique by which the acoustic models selected for the target language are *trained* with existing speech data from the source languages. The first step of the method, CPP candidate selection for each target phoneme, is the same as the previous approach. In the model building step, the target HMMs are trained directly on the source data, and the size of the HMMs therefore stays constant. Our experiments show that model performance relative to the multimixture assembly method remains constant.

In the next section, we review the CPP metric that is presented in detail in [6][7]. In section 3, we introduce the "silkie-hen-on-duck-eggs" cross-language model training method. Then, in section 4, we present the experiments and results of our cross-language transfer utilizing the CPP metric and silkie-hen-on-duck-eggs training strategy. We give our conclusions in section 5.

2. A combined phonetic-phonological metric

2.1. Weighted phonetic distance

2.1.1. Quantitative representation of phonemes

To measure phonetic distance between phonemes, each phoneme is quantitatively represented by a vector of a fixed set of binary-valued phonetic features that characterize articulatory characteristics, such as voicing, place of articulation, and manner of articulation [8][10]. The binary value indicates the presence or absence of an articulatory feature of the phoneme. A phoneme is denoted by $p_l(i)$, where l (=1,…,L) represents the language that includes the phoneme, and i (=1,…,I_l) represents the index of the phoneme in language l. Thus, the phoneme inventory of language l is

$$\{p_l(i) \mid i = 1,\ldots,I_l\}\ . \tag{1}$$

A phoneme $p_l(i)$ is represented by a vector of J features

$$\mathbf{f}[p_l(i)] = [v_l(i,j)]^T = [v_l(i,1),\ldots,v_l(i,j),\cdots,v_l(i,J)]^T \tag{2}$$

where each $v_l(i,j)$ is a binary feature, $i = 1,\cdots,I_l$, $j = 1,\cdots,J$, $l = 1,\cdots,L$, and the superscript T denotes vector transposition.

Two characteristics of the quantitative phonemic representation system are noted here. First, the feature set has internal structure whereby hierarchical relations are expressed. In such a system, the presence of one feature presupposes the presence of those hierarchically dominant features. This feature structure warrants that the feature-based phonetic distance consistently reflects realistic articulation contrast among phones.

Second, corollary features are invented to account for allophonic variation. This innovation allows for feature contradiction between allophones corresponding to the same phoneme. For example, a phoneme /k/ may have two principal phonetic realizations depending on the absence or presence of aspiration (the feature 'spread glottis'): [k] and [k^h], respectively, in IPA notation [10]. The value of the feature 'spread glottis' is equal to 0 for the [k] allophone and 1 for the [k^h] allophone. To resolve this type of feature contradiction, corollary features are introduced to specify the occasional, allophonic realization of phonetic features. We use J=30 features in our system; these include 26 primary phonetic features and four corollary features. Features are chosen so that no two phonemes have identical feature vectors in any given language.

2.1.2. Weighted phonetic distance

A drawback of using binary systems directly is that they neglect the relative importance of individual features. Calculation of phonetic distance with binary vectors might give incorrect results [11][12][13]. To solve this problem, we use weights, or salience, on individual features. The value of a weight for a feature is derived from the frequency of the feature in the lexica of all the source and target languages. Let $c_l[p_l(i)]$ denote the occurrence count of a phoneme $p_l(i)$ in a lexicon of language l, then the frequency of each feature j contributed by the phoneme $p_l(i)$ is $c_l[p_l(i)]v_l(i,j)$, and the frequency of each feature j contributed by all the phonemes in language l is $\sum_{i=1}^{I_l} c_l[p_l(i)]v_l(i,j)$. The global weights derived from all the phonemes in the entire source and target languages are

$$\mathbf{W}(j) = diag\{w(1),\cdots,w(j),\cdots,w(J)\} \tag{3}$$

where

$$w(j) = \frac{1}{L}\sum_{l=1}^{L} w_l(j) = \frac{1}{L}\sum_{l=1}^{L}\frac{\sum_{i=1}^{I_l} c_l[p_l(i)]v_l(i,j)}{\sum_{j=1}^{J}\sum_{i=1}^{I_l} c_l[p_l(i)]v_l(i,j)} \quad j = 1,\cdots,J \tag{4}$$

where *diag*(vector) gives a diagonal matrix with elements of the vector as the diagonal entries. In the definition each language is treated equally. Thus the weights are not subject to the relative size of each language's lexicon.

We define the phonetic distance between phonemes $p_l(i)$ and $p_t(k)$ in the form of a Manhattan distance, which is expressed as

$$d_{lt}(i,k) = \left\|\mathbf{W}(j)(\mathbf{f}[p_l(i)] - \mathbf{f}[p_t(k)])\right\|_1 = \sum_{j=1}^{J} w(j)\left|v_l(i,j) - v_t(k,j)\right| \tag{5}$$

where $i = 1,\cdots,I_l$, $k = 1,\cdots,I_t$, and the weights, given in a diagonal matrix $\mathbf{W}(j)$, are dependent upon the feature identity j.

2.2. Phonological distances

Although phonetic features are very important in phoneme specification, a phoneme's realization is ultimately dependent on the overall phonology of the language to which it belongs. For example, a phoneme identified as /i/ in a language that has five vowel phonemes, /i e a o u/, is more likely to correspond to an /i/ in another language that has the same vowel contrasts than, say, to some other language that has only a three-way contrast. Phoneme inventories, then, naturally provide constraints on subphonemic variance. Using this logic, inventory similarity between languages can assist in indicating allophonic similarity between phonemes already corresponding in feature specification. In this section, we define two distance metrics to characterize cross-language phonological similarity.

2.2.1. Monophoneme distribution distance

Monophoneme distribution distance characterizes the difference in lexical phoneme distribution between two languages. Specifically, the distribution, or normalized histogram, of the phonemes is obtained from a large lexicon of a language, with the probability in the distribution corresponding to the frequency of a phoneme in the lexicon. The monophoneme metric is a typological comparison that is based on two principal classes of information: (1) types of sounds and (2) frequencies of these sounds in the lexicon. The former class, types of sounds, is directly associated with phoneme inventory correspondence while the latter, phoneme frequency, concerns relative phoneme importance. Note that in order for the distribution to be an unbiased representative of a language, we derive it from a typical lexicon instead of a database.

Because the phoneme inventories of the two languages to be compared may not be identical, we first need to define a combined inventory for them

$$\{p_{lt}(m) \mid m = 1,\ldots,I_{lt}\} = \{p_l(i) \mid i = 1,\ldots,I_l\} \cup \{p_t(k) \mid k = 1,\ldots,I_t\} \tag{6}$$

where $p_{lt}(m)$ is a phoneme in the combined inventory where there are total I_{lt} phonemes.

The frequency of the phoneme $p_{lt}(m)$ in language l can be expressed as

$$\rho_l[p_{lt}(m)] = \frac{c_l[p_{lt}(m)]}{\sum_{i=1}^{I_l} c_l[p_l(i)]}, \qquad m = 1, \cdots, I_{lt} \tag{7}$$

where $c_l[p_{lt}(m)]$ is the occurrence count of phoneme $p_{lt}(m)$ in a lexicon of language *l*. If a phoneme $p_{lt}(m)$ does not exist in the language, its frequency would be zero. The difference of phoneme frequencies between the two languages can be calculated as

$$d\rho_{lt}[p_{lt}(m)] = \left|\rho_l[p_{lt}(m)] - \rho_t[p_{lt}(m)]\right| \qquad m = 1, \cdots, I_{lt} \tag{8}$$

Then the monophoneme distribution distance between the target language *t* and source language *l* is

$$D\rho_{lt} = \sum_{m=1}^{I_{lt}} d\rho_{lt}[p_{lt}(m)]\,. \tag{9}$$

The distance is calculated between the target language and every one of the source languages.

In view of the known differences in phonological characteristics between vowels and consonants, we make separate calculations for the vowel and consonant categories. Thus Eq. (9) becomes

$$D\rho^g_{lt} = \sum_{p_{lt}(m)\in g} d\rho_{lt}[p_{lt}(m)] \tag{10}$$

where *g*=vowels or consonants.

2.2.2. *Biphoneme distribution distance*

The biphoneme distribution distance metric characterizes the difference in lexical distribution of phoneme pairs, or biphonemes, between two languages. It explicitly provides a biphoneme inventory, permissible phonotactic sequences, and phonotactic sequence importance. It also implicitly incorporates phoneme inventory and phonological complexity information.

Similar to the monophoneme distribution distance, the distribution of biphonemes in a language is obtained based on the frequency of a biphoneme in a large lexicon. The biphoneme inventory for the target language *t* is expressed as

$$\{q_t(k) \mid k = 1, \ldots, I'_t\} \tag{11}$$

while the biphoneme inventory for a source language *l* is

$$\{q_l(i) \mid i = 1, \ldots, I'_l\} \tag{12}$$

Then the combined biphoneme inventory for the two languages to be compared is

$$\{q_{lt}(n) \mid n = 1, \ldots, I'_{lt}\} = \{q_l(i) \mid i = 1, \ldots, I'_l\} \cup \{q_t(k) \mid k = 1, \ldots, I'_t\} \tag{13}$$

where $q_{lt}(n)$ is a biphoneme in the combined inventory where there are total I'_{lt} biphonemes. For a phoneme at the beginning or end of a word, $q_{lt}(n)$ takes the format of “void+phoneme” or “phoneme+void”, respectively.

The frequency of a biphoneme $q_{lt}(n)$ in language l can be expressed as

$$\gamma_l[q_{lt}(n)] = \frac{c_l[q_{lt}(n)]}{\sum_{i=1}^{I'_l} c_l[q_l(i)]}, \qquad n = 1, \cdots, I'_{lt} \tag{14}$$

where $c_l[q_{lt}(n)]$ is the occurrence count of biphoneme $q_{lt}(n)$ in a lexicon of language *l*. The difference of biphoneme frequencies between the two languages is

$$d\gamma_{lt}[q_{lt}(n)] = \left|\gamma_l[q_{lt}(n)] - \gamma_t[q_{lt}(n)]\right| \quad n = 1, \cdots, I'_{lt} \tag{15}$$

Then the biphoneme distribution distance between the target language *t* and source language *l* is

$$D\gamma_{lt} = \sum_{n=1}^{I'_{lt}} d\gamma_{lt}[q_{lt}(n)]\,. \tag{16}$$

Similarly, the distance is better characterized within the categories of vowels and consonants separately. In our algorithm we count each biphoneme twice, the first time as a left-contact biphoneme and second time as a right-contact biphoneme. Thus

$$D\gamma^g_{lt} = \sum_{\text{right of } q_{lt}(n)\in g} d\gamma_{lt}[q_{lt}(n)] + \sum_{\text{left of } q_{lt}(n)\in g} d\gamma_{lt}[q_{lt}(n)] \tag{17}$$

where *g*=vowels or consonants.

Use of phonological metrics ensures that the overall model pool will have a bias toward a reduced set of phonologically similar languages. Because our final model pool is meant to represent a single language system, i.e., the target language, it is reasonable to expect that similarity in languages of the model pool provides consistency in the target HMM system [14].

2.3. A combined phonetic-phonological (CPP) metric

Finally, a metric is developed by combining the above-mentioned phonetic and phonological distances. Since the three distances are from different domains, normalization is necessary before combination. The normalization, aimed at extracting the relative ranking between candidate phonemes and languages, is a linear processing that scales the score range from each domain into the range [0 1].

The combined phonetic-phonological metric (CPP) is defined as

$$CPP_{lt}(i,k) = \alpha_d \cdot [d_{lt}(i,k)]_N + \alpha_\rho \cdot [D\rho^g_{lt}]_N + \alpha_\gamma \cdot [D\gamma^g_{lt}]_N \tag{18}$$

where $CPP_{lt}(i,k)$ represents the distance between phoneme $p_l(i)$ from language *l* and phoneme $p_t(k)$ from language *t*, and both phonemes belong to the same phonological category g (vowels or consonants). The weights α_d, α_ρ, and α_γ represent the relative importance of each quantity. We equate the overall importance of phonetics with that of phonology by using the weight values ($\alpha_d, \alpha_\rho, \alpha_\gamma$)=(2,1,1). The symbol $[\cdot]_N$ denotes that the quantity inside is linearly scaled into the range [0 1]. For $D\rho^g_{lt}$ and $D\gamma^g_{lt}$, the original range is determined by scores of all the source languages. Their scaling is done once for a target language *t*. While for $d_{lt}(i,k)$, we found that it is better to do scaling once for each target phoneme $p_t(k)$, and the original range is determined by scores of a group of candidate phonemes that includes at least one phoneme from each source language.

3. A “silkie-hen-on-duck-eggs” training method

We employ the regular 3-state, left-right, multimixture, continuous-Gaussian HMMs for the acoustic models and assume that the models from all the source and target languages have the same topology. Once the top candidates are determined from the CPP metric for each target phoneme, the next step is to build the target HMMs using the candidate information. In our previous method [6], the candidate models are well trained with their native speech data. Each target HMM is constructed in the form of a multimixture model with the mixture components for a certain state adopted from the pdfs of the candidates corresponding to the same state. The original mean and variance values are maintained while the mixture weights are scaled down so that the new weights add up to one for each state.

Normally the performance of the assembled target models increases with the number of candidates used. For example, the performance of models built for Spanish (the details of the experiment are given in the next section) is 70.09% WER with one candidate, 93.06% WER with two candidates, and 93.50% WER with three candidates, per target phoneme, respectively. However, the size of the target models increases drastically with the number of the candidates, which consequently results in greater computation load and memory requirement. Therefore, this model building method is principally suited for research stage development. The obvious advantage of building target models with ready-to-use source models is that the procedure is very fast in that it entirely circumvents model retraining. However, for implementation, the issue of size must necessarily be addressed.

One immediate remedy to the size challenge is to merge the mixtures in close proximity. However, given the lack of native target-language data, it is impossible to tune and update the models through training, which is a necessary step after merging.

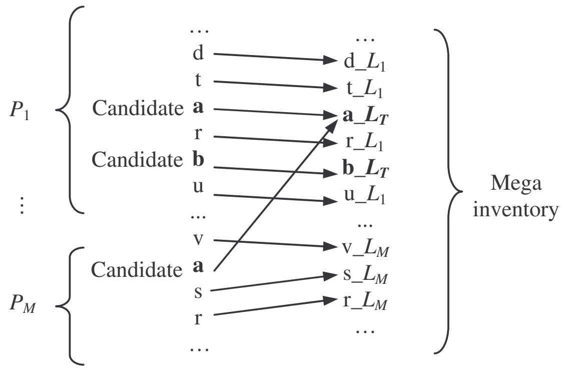
Instead, we have developed a method of training the target models with data from just those source languages from which the models are selected, hereinafter "donor languages". All the speech data from the donor languages is collectively referred to as *mega* data, and is used in training. The key to this method is the composition of a mega phoneme inventory. All the phoneme symbols of the donor languages are attached with an ID tag indicating the languages they belong to, except for those phonemes that have been selected as candidates for the target phonemes, which are tagged with the target-language ID. The ID-tagged mega phoneme inventory (as illustrated in Fig. 1) includes all the phonemes from the donor languages and is used to retranscribe all the pronunciation lexica and phoneme transcriptions. Thus, for the mega data, we build a mega inventory, a mega lexicon, and a mega transcription set.

The mega inventory is expressed as

$$\bigcup_{l'=1}^{M}\{p_{l'}(i)\mid i=1,\dots,I_{l'} \text{ except when } p_{l'}(i)\Rightarrow p_t(k)\} \cup \{p_t(k)\mid k=1,\dots,I_t\} \tag{19}$$

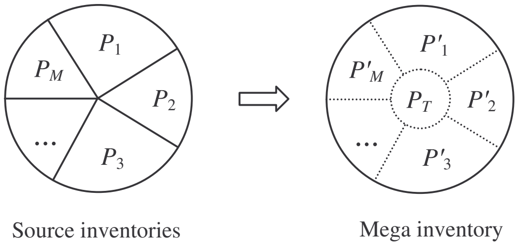
where $p_{l'}(i)\Rightarrow p_t(k)$ means $p_{l'}(i)$ is a selected candidate for the target phoneme $p_t(k)$. Multiple candidates are permissible for a single target phoneme. As shown in Fig. 2, the target inventory is embedded as a subset in the mega inventory.

Once the mega inventory, mega lexicon, and mega phoneme transcription set are created, any generic training strategy, such as Baum-Welch training and Viterbi training, can be employed naturally on the mega data in the same way as training a set of acoustic models for a single language. If we liken the model training to an egg-hatching procedure, the training of the embedded target models is analogous to a silkie hen hutching some duck eggs besides her own eggs. Hence we call the method a "silkie-hen-on-duck-eggs" strategy.



P_1, …, P_M : individual source inventories
L_1, …, L_M , L_T : language ID

Figure 1: Tagging phoneme symbols with language ID.



P_1, P_2, …, P_M : individual source inventories
P'_1, P'_2, …, P'_M : source inventories less the candidate phonemes
P_T : target inventory

Figure 2: Inventories before and after merging.

4. Experiments

We use 17 languages in the experiments testing the CPP approach to cross-language transfer: Italian, Canadian French, US English, Swedish, European Portuguese, Mandarin, Latin-American Spanish, Japanese, Korean, Danish, German, Cantonese, British English, Parisian French, Brazilian Portuguese, Dutch, and Egyptian Arabic. For each language, a native monolingual model set is generated by training with its native speech data. The acoustic features are 39 regular MFCC features including cepstral, delta, and delta-delta. The databases include CallHome, EUROM, and SpeechDat, among others. The benchmark performance of the native models approximates 95% for most languages. In each of the following experiments, we first pick one language as the target language, leaving the remaining languages as candidate source languages. The CPP scores are then calculated for each target-language phoneme and the top two candidate source-language phonemes are chosen. Their associated acoustic models are used for acoustic model construction. Experiments are performed on the following four target languages: Italian, Latin-American Spanish, European Portuguese, and Danish.

Each recognition task includes about 3000 utterances of digit strings, command words, and sentences.

4.1. Baselines

Two benchmark results are used as baselines. Baseline #1 is the performance of the native monolingual, context-independent models from each target language.

Table 1: Performance of baseline models.

Target Language	Baseline #1 (Native models)	Baseline #2 (Acoustic mapping)
Spanish	94.49	88.61
Italian	98.42	98.27
Danish	94.36	72.95
Portuguese	96.31	77.91

Baseline #2 is the performance of a constructed model set for each target language. These models are built with the top two candidate models chosen from source languages based on their acoustic distance from the corresponding natively trained target model (i.e., Baseline #1 version). The Baseline #2 models are built as multimixture assemblies. Hence, Baseline #2 provides a benchmark about the capacity of the source language models in replacement of the target language models. However, the Baseline #2 benchmark is not a theoretically strict upper bound for cross-language transfer because the distribution measurement in the acoustic space is probabilistic. We adopt the widely used Bhattacharyya metric for the distance measurement [15].

4.2. Experiments on cross-language models

The first experiment is conducted on target models built in the form of multimixture assemblies. Based on the CPP scores, the top two candidate HMMs are used for each target phoneme. The results are given in the first column of Table 2.

Table 2: Performance of the cross-language models.

Target Language	CPP-based (Assembled)	CPP-based (Trained)	Model size ratio
Spanish	93.06	94.21	0.51
Italian	98.52	98.01	0.51
Danish	70.15	68.54	0.52
Portuguese	72.74	73.43	0.50

In the second set of experiments, target models are trained with a set of mega data using the silkie-hen-on-duck-eggs approach, where a distinct mega phoneme inventory and mega data set is constructed for each of the four target languages. Recall that for each target language, the mega data consists of the native speech data from the donor source languages, which are in turn determined by the mega inventory. That is, only source languages that donate HMMs for corresponding target-language phonemes contribute to the mega data. As in the previous experiment set, the two top source-language candidates are selected for each target-language phoneme. The models are trained with the normal Baum-Welch method. The results are given in the second column of Table 2. In spite of using two candidates per target phoneme, the number of mixtures in each target HMM can be restored to the same level as the number of mixtures in a source HMM. Therefore, the size of the trained target models is half of the size of the assembled target models (see the third column of Table 2). As shown in Table 2, however, the performance of the models is not affected by the size reduction.

4.3. Discussion

Comparison of the results in Table 2 with the baseline results in Table 1 shows that the performance of cross-language models based on the CPP metric is equivalent overall to the performance of models constructed by an acoustic distance strategy, regardless of the way the cross-language models are generated. The performance of CPP-based cross-language models reaches the performance of the native models for Spanish and Italian, while it is far lower for the other three languages. We explain this difference by noting that neither Spanish nor Italian have phonemes that are unattested in the other languages in our dataset while Japanese, Danish, and Portuguese all contain phonemes that are absent in the source languages (see [7] for further discussion).

5. Conclusions

In this paper, we presented a cross-language training method, the "silkie-hen-on-duck-eggs" strategy, where target-language acoustic model training is embedded in source-language acoustic model training. The candidate phonemes are initially selected from source languages using the CPP metric, an automated, purely linguistic-based approach. The silkie-hen-on-duck-eggs training technique replaces the previous cross-language strategy where two or more source models are used directly (without retraining) for each target-language phoneme and achieve recognition results comparable to data-driven methods. The silkie-hen-on-duck-eggs strategy proves to be effective at building target-language models as small as natively trained models and achieving recognition performances as high as the much larger combination of source models.

6. References

[1] Köhler J., "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," *ICASSP'98*, 417-420, 1998.

[2] Schultz, T. and Waibel, A., "Multilingual and crosslingual speech recognition," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Va., 1998.

[3] Schultz, T. and Waibel, A., "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Comm.*, 35, 31-51, 2001.

[4] Sooful, J. J. and Botha, E. C., "Comparison of acoustic distance measures for automatic cross-language phoneme mapping," *ICSLP'02*, 521-524, 2002.

[5] Kienappel, A. K., Geller, D., and Bippus, R., "Cross-language transfer of multilingual phoneme models," *ISCA Tutorial and Research Workshop ASR 2000, Automatic Speech Recognition: Challenges for the New Millennium*, Paris, 2000.

[6] Byrne, W. *et al.*, "Towards language independent acoustic modeling," *ICASSP'00*, 1029-1032, 2000.

[7] Liu, C. and Melnar, L., "An automated linguistic knowledge-based cross-language transfer method for building acoustic models for a language without native training data," *Interspeech'05*, 1365-1368, Lisbon, 2005.

[8] Melnar, L. and Liu, C., “HMM similarity prediction: An enhanced articulatory phonetic measurement approach,” *3rd Conf. Experimental Phonetics*, Santiago de Compostela, Spain, 2005.

[9] Chomsky, N. and Halle, M., *The Sound Pattern of English*, Harper & Row, New York, 1968.

[10] IPA, *Handbook of the International Phonetic Association*, Oxford University Press, 1999.

[11] Connolly, J. H., “Quantifying target-realization differences,” *Clinical Linguistics & Phonetics*, 11:267–298, 1997.

[12] Kessler, B., “Computational dialectology in Irish Gaelic,” *Proc. 6th Conf. European Chapter of ACL*, 60–67, 1995.

[13] Somers, H. L., “Similarity metrics for aligning children’s articulation data,” *Proc. 36th Annual Meeting ACL and 17th Int. Conf. Comp. Ling.*, 1227–1231, 1998.

[14] Schultz, T. and Waibel, A., “Polyphone Decision Tree Specialization for Language Adaptation”, *ICASSP’00*, Istanbul, 2000.

[15] Mak, B. and Barnard, E., “Phone clustering using the Bhattacharyya distance,” *ICSLP’96*, 2005-2008, 1996.