A Combined Phonetic-Phonological Approach to Estimating Cross-Language Phoneme Similarity in an ASR Environment

Lynette Melnar
lynette.melnar@motorola.com

Chen Liu
chen.liu@motorola.com

Abstract

This paper presents a fully automated linguistic approach to measuring distance between phonemes across languages. In this approach, a phoneme is represented by a feature matrix where feature categories are fixed, hierarchically related and binary-valued; feature categorization explicitly addresses allophonic variation and feature values are weighted based on their relative prominence derived from lexical frequency measurements. The relative weight of feature values is factored into phonetic distance calculation. Two phonological distances are statistically derived from lexical frequency measurements. The phonetic distance is combined with the phonological distances to produce a single metric that quantifies cross-language phoneme distance.

The performances of target-language phoneme HMMs constructed solely with source language HMMs, first selected by the combined phonetic and phonological metric and then by a data-driven, acoustics distance-based method, are compared in context-independent automatic speech recognition (ASR) experiments. Results show that this approach consistently performs equivalently to the acoustics-based approach, confirming its effectiveness in estimating cross-language similarity between phonemes in an ASR environment.

1 Introduction

Speech technologists typically use acoustic measurements to determine similarity among acoustic speech models (phone(me) HMMs) and there are a variety of distance metrics available that prove the effectiveness of this method (see Sooful and Botha 2002). Additionally, HMM similarity can be evaluated indirectly through comparison of HMM performances in ASR experiments.

For acoustic measurements, speech data must be accessible for model training. However, speech data unavailability is a practical concern in that most commercially available speech databases are restricted to widely spoken languages in large business markets. The vast majority of languages have not been exposed to intense data collection and resources for these languages are subsequently either limited or completely unavailable. Hence a knowledge-based phoneme distance metric potentially has great value in acoustic modeling for resource-limited languages in that it can predict cross-language HMM similarity in the absence of target-language speech data.

Knowledge-based approaches to HMM similarity generally attempt to identify articulatory similarity between phonemes across languages. The typical strategy is subjective and label-based, where two phonemes are judged to be more or less similar depending on their transcription labels (Köhler 1996; Schultz and Waibel 1997, 2000).

A label-based approach suffers for two obvious reasons. First, phone inventories designed for speech technology applications are predominantly phonemic in orientation. Thus, transcription labels do not transfer with the same phonetic value to other languages, even where international phonetic transcription labels are employed. In a phonemic transcription strategy, transcription labels are gen-

Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology at HLT-NAACL 2006, pages 1–10,
New York City, USA, June 2006.

erally restricted to only the most basic symbols, usually unmodified letters of the Roman alphabet (IPA 1999). Second, phoneme transcription labels fail to capture allophony. The best phonetic definition that a phoneme transcription label can offer is the most typical phonetic realization of that phoneme. Not surprisingly, label-based cross-language transfer experiments have produced poor performance results.

In contrast to the subjective, label-based strategy, researchers in such fields as language reconstruction, dialectometry, and child language development, commonly use automatic feature-based approaches to articulatory similarity between phonemes. In these methods, phonemes are represented by a distinctive feature vector and a phonetic distance or similarity algorithm is used to align phoneme strings between related words (Connolly 1997; Kessler 1995, 2005; Kondrak 2002; Nerbonne and Heeringa 1997; Somers 1998). Significantly, in these approaches, phonological similarity is generally assumed.

In principle, the feature-based approach to phonetic distance admits more precise specification of phonemes because it supports allophonic variance. For example, a standard feature-based approach to allophony representation restricts feature inclusion to only those features relevant to all realizations of the phoneme. Another common approach retains features that are relevant to all allophonic variants, but leaves their values underspecified (Archangeli 1988). However, it is unclear from the literature whether allophony is explicitly addressed in the current feature-based approaches to phoneme similarity.

A strategy for specifying allophony and characterizing phonetic distance between phonemes is only one component in predicting phoneme similarity among diverse languages without acoustic data in an ASR environment. Because HMMs represent phonemes and significant allophones in a language-dependent context, it is necessary to consider the overall constructed target-language HMM system. Thus phonological distance quantities that regulate the priority of source languages for phoneme selection in accordance to their phonological similarity to the target language are also in order.

In this paper, we describe an automated, combined phonetic-phonological (CPP) approach to estimating phoneme similarity across languages in ASR. Elsewhere, we provide the phonetic and phonological distance algorithms (Liu and Melnar 2005, 2006), though offer little linguistic justification of the approach or evaluation of the experiment results due to space limitations. Here, we focus on explaining the linguistic principles behind the algorithms and analyzing the results.

The CPP approach is fundamentally based on articulatory phonetic features and is designed to handle allophonic variation. Feature salience and phonetic distance are automatically calculated and phoneme distance is constrained by statistically-derived phonological similarity biases. Unlike other distinctive feature-based approaches to phoneme similarity, phonological distance is *not* assumed. In testing this approach in cross-language transfer experiments, target-language resources are restricted to lexica and phonology descriptions and do not include speech data.

In the next section, we describe our feature-based phoneme specification method. In section three, we show how our phoneme specification approach is used in calculating phonetic distance between phonemes. Section four describes two other distance metrics that predict phonological similarity between languages. We explain how the three distance metrics combine to quantify cross-language phoneme distance and select target-language phoneme HMM inventories. In section five, we describe the experiments that we conducted to evaluate our approach to phoneme similarity prediction. Here, the CPP method is compared with an acoustic distance method in context-independent speech recognition. We offer our evaluation and conclusions in section 6.

2 Phoneme specification

In the CPP approach to estimating cross-language phoneme similarity, each phoneme in our multilingual ASR dataset is associated with a distinctive feature matrix. Feature categories are fixed for all phonemes, hierarchically related, and binary-valued. Feature-contradiction, associated with allophonic variance, is explicitly addressed through the introduction of a small set of special corollary features.

2.1 The phoneme feature matrix

As noted in the introduction, cross-language phoneme comparison requires accurate feature specification. Because a phoneme comprises one or more allophones which may contrast in particular features, a distinctive feature strategy that allows for feature contradiction is preferred. Omitting contradictory features and underspecifying contradictory values are two well-known methods.

However, cross-language phoneme comparison in a computational environment is greatly facilitated by agreeing on a *fixed* set of *binary-valued* features for all phonemes. A fixed set of distinctive features is favored as this enables cross-class phoneme comparison. A binary-valued system is easy to manipulate and naturally lends itself to mathematical formulation. However, strict binary-valued feature systems only indicate the presence or absence of a feature, and feature contradiction must then be indicated by feature omission - which is not possible in a fixed distinctive feature set.

The phoneme specification method that we employ indicates feature contradiction associated with allophony in a strict binary-valued, fixed set of distinctive features through the introduction of special feature categories. Specifically, we utilize a small set of *corollary* features to mark the occasional, allophonic realizations of some primary features. A corollary feature is defined as a feature that supplements a primary feature in the system. The corollary features mark “occasionality” (associated with context dependency, dialectal variation, speech style variation, etc.) in the primary feature as either present or absent.

2.2 Primary and corollary features

Our feature set includes twenty-six primary articulatory features and six corollary features. The selected primary features conform to a typical set of hierarchically-related distinctive features (e.g. syllabic, sonorant, consonantal, labial, coronal, nasal, continuant, high, low, back, etc.) (Ladefoged 1975). In this hierarchical system, the presence of one feature presupposes the presence of those hierarchically dominant features. For example, the presence of the feature [alveolar] requires the presence of the feature [coronal], and the presence of the feature [nasal] requires the presence of the feature [sonorant]. Significantly, the reverse of these relations is not true. As is explained later in the next section, this feature structure allows for a linguistically-principled determination of feature salience in phonetic distance calculation.

Corollary features are restricted to specifying those primary features that are judged to be most significant to cross-language phoneme comparison *in an ASR environment*. Phoneme inventories designed for ASR comprise both phonemes and significant allophones, where a significant allophone is characteristically both acoustically distinct from the primary allophone and associated with a sufficiently high count of occurrence in the associated speech database. Thus American English ASR inventories regularly include an alveolar tap, a contextually-realized allophonic variant of both /t/ and /d/. Furthermore, pronunciation transcriptions in ASR lexica are typically phonetic - within the context of the phoneme-based inventory. So, word-final voice neutralization in German is overtly indicated throughout the lexicon (e.g. *hund* : h U n t). A typical ASR phoneme then does not represent a true phoneme; rather it encompasses only that phonemic variation that is not explicitly captured by its existing significant allophones in the inventory.

Corollary features specify variance that is not usually overtly indicated in ASR inventories and lexica but that is important to cross-language phoneme comparison in an acoustic, ASR environment. Internal phoneme recognition experiments indicate that generally major class features (syllabic, sonorant, etc.), manner features (nasal, continuant, etc.) and laryngeal features (voice, spread glottis, etc.) are more robustly identified than place features (labial, coronal, etc.); accordingly, the set of corollary features, provided in Table 1, predominantly targets particular major class, manner, and laryngeal features.

Table 1: Corollary features

Corollary Feature	Description
syllabic-occ	positive value marks the occasional realization of the phoneme as a syllabic consonant or glide
voice-occ	positive value marks the occasional voicing of phonemes
labial-occ	positive value marks the occasional rounding of vowels
nasal-occ	positive value marks the occasional nasalization of vowels and glides
rhotic-occ	positive value marks the occasional rhotization of liquids and vowels
spread-occ	positive value marks the occasional aspiration of obstruents

It should be pointed out that allophones that express a place contrast or difference in continuance with the primary realization of a phoneme are typically considered significant allophones in the ASR phoneme system and are therefore overtly represented.

As an illustration of the usefulness of corollary features in cross-language phoneme comparison, consider Table 2 which includes a partial feature matrix for the phoneme /k/ associated with 17 languages and dialects:

Table 2: Partial distinctive feature table

Languages	phoneme	spread glottis	spread -occ
Arabic	k	0	0
Danish	k	1	1
German	k	1	1
British English	k	1	1
U.S. English	k	1	1
Lat. Spanish	k	0	0
Can. French	k	0	0
Parisian French	k	0	0
Italian	k	0	0
Japanese	k	1	1
Dutch	k	0	0
Brz. Portuguese	k	0	0
Eur. Portuguese	k	0	0
Swedish	k	1	1
Korean	k	1	0
Cantonese	k	1	0
Mandarin	k	1	0

Note that the realization of the phoneme /k/ differs across the seventeen languages and dialects in the two features provided: [spread glottis] and [spread-occ]. The presence of the feature [spread glottis], marked by 1, and the non-presence of the corollary feature [spread-occ], marked by 0, indicates that the glottis is always open during the articulation of the phoneme; i.e. this phoneme is consistently associated with aspiration. The precise IPA transcription of this segment is /k^h/. A positive value for the corollary feature [spread-occ] means that the phoneme is only sometimes associated with aspiration. This phoneme has two principle phonetic realizations, marked [k] and [k^h] in IPA notation. A 0 value for the feature [spread glottis] and corollary feature [spread-occ] indicates that the segment is never aspirated. Thus this phoneme is most accurately labeled /k/ in IPA labeling.

Because this methodology incorporates phoneme feature contradiction, overall phonological similarity among languages and dialects is more precisely predicted:

Table 3: Phoneme similarity across languages

phoneme	allophone(s)	language	lang. family
k	k^h, k	Danish	Germanic
		German	Germanic
		Br. Eng.	Germanic
		Amer. Eng.	Germanic
		Japanese	Altaic
		Swedish	Germanic
	k^h	Korean	Altaic
		Mandarin	Sinitic
		Cantonese	Sinitic
	k	Arabic	Afro-Asiatic
		Lat. Span.	Romance
		Parisian Fr.	Romance
		Canadian Fr.	Romance
		Italian	Romance
		Dutch	Germanic
		Brz. Port.	Romance
		Eur. Port.	Romance

Table 3 reveals that Germanic languages tend to only occasionally aspirate /k/, Romance languages avoid aspirating /k/, and Sinitic languages typically aspirate /k/. Of course, closely related languages tend to be phonologically similar.

3 Phonetic distance

Most techniques for measuring phonetic distance between phonemes that do not assume speech data availability are based on articulatory features, though perceptual distance, judged (subjective) distance, and historical distance are also attested (Kessler 2005). We base our phonetic distance measurement on articulatory features because of their cross-linguistic consistency and general availability.

As Kessler notes, standard phonological theory provides no guidance in comparing phonetic distance between phonemes across multiple features (Kessler 2005). In our experiments to date, we use the Manhattan distance where the distance between phonemes equals the sum of the absolute values of individual feature distances. This approach is fairly standard in the literature, though the Euclidean distance has also been reported to attain good results (Kessler 2005).

Because features are known to differ in relative importance (Ladefoged 1969), some researchers apply weights or saliencies to the individual features for distance calculation. Nerbonne and Heeringa (1997), for example, weighted each feature by information gain, or entropy reduction. Kondrak (2002) expressed weights as coefficients that could be changed to any numeric value. He adjusted the coefficients until he achieved optimal performance on aligning cognate words.

In our approach, weights are derived from the lexica of all the considered languages. Specifically, the value of a weight for a feature is derived from the frequency of the feature in the lexica. Each language is treated equally in this approach; thus, the weights are not subject to the relative size of a language's lexicon.

Because our phoneme specification method incorporates hierarchical relations between features, feature weights are necessarily interdependent. Hierarchically dominant features are more frequently attested than their subordinate features and thus receive more weight. Further, hierarchically superior features tend to correspond to major phonetic categories (sonorant, consonantal, syllabic, etc.), which are expected to be more contrastive or distant to each other than sister subordinate categories. Thus, in a hierarchical feature system, lexical frequency of features is a reasonable indication of feature importance in phonetic contrast or distance.

In the following two subsections the phonetic distance algorithm is described.

Quantitative representation of phonemes

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

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

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

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

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.

Weighted phonetic distance

As mentioned, the value of a weight for a feature in the present phonetic distance approach is derived from the frequency of the feature in the lexica of all the considered 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 all the languages are

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

where

(4)

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

where *diag*(vector) gives a diagonal matrix with elements of the vector as the diagonal entries. 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

(5)

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

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*.

4 Phonological distance metrics

Although our phoneme specification approach is designed to account for allophonic variance, not all variation is captured. Because of this, the effectiveness of measuring phonetic distance as a standalone strategy to predicting cross-language phoneme similarity is compromised. Furthermore, phonetic distance does not determine relative phoneme similarity in the not atypical scenario where two or more phonemes share the same phonetic distance to some target phoneme. In order to address these problems, phonological distance metrics are used to bias cross-language phoneme similarity predictions toward languages that have similar phoneme inventories and phoneme frequency distributions. The general idea is that the more similar the phoneme inventory and relative importance of each corresponding phoneme between languages, the more likely it is that the corresponding phonemes will be more similar.

Phonological distance consideration is especially desirable in an ASR environment because ultimately HMMs corresponding to those source-language phonemes predicted to be most similar to target-language phonemes must interact in a system that is intended to reflect a single target language. Use of phonological metrics then ensures that the overall model pool will have a bias toward a reduced set of phonologically similar languages, and it is reasonable to expect that similarity in languages of the model pool provides consistency in the target HMM system (see Schultz and Waibel 2000).

In this section, we define two distance metrics to characterize cross-language phonological similarity. One is based on monophoneme inventories while the other is based on biphoneme inventories.

4.1 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. We derive the distribution from a lexicon as we consider it more representative of a language's phonology than a particular database.

The monophoneme distribution 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 is directly associated with phoneme inventory correspondence while the latter concerns relative phoneme importance.

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

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

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

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

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

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

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

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

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

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

where *g*=Vowels or Consonants.

4.2 Biphoneme distribution distance

The biphoneme distribution distance metric characterizes the difference in lexical distribution of phoneme pairs, or biphonemes, between two languages. Similar to the monophoneme distribution distance, the distribution of biphonemes in a language is obtained based on the frequency of biphonemes in a large lexicon.

The biphoneme metric indicates how phonemes can combine in a language and how important these combinations are. Though the phonotactics provided in this approach is limited to only a sequence of two, the overall biphoneme inventory and distribution provides important phonological information. For example, it indicates if and to what extent consonants can cluster. Some languages tend to disfavor consonant clustering, like the Romance languages, while others allow for broad clustering, like the Germanic languages. It also indicates if and to what extent vowels may cooccur. Many languages require an onset consonant so vowels will never co-occur; other languages have no such restriction.

The biphoneme metric then yields types of information that are distinct from the monophoneme metric. 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

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

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

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

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

(13)

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

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

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

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

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

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

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

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

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

where *g*=Vowels or Consonants.

4.3 CPP phoneme distance

For phoneme similarity prediction, we unite the phonetic and phonological distance metrics to arrive at the CPP phoneme distance measurement. Since the three distances are from different domains and provide distinct types of information, normalization is necessary before combination. The normalization, aimed at extracting the relative ranking between source phonemes and languages, is a linear processing that scales the score range from each domain into the range [0 1].

We equate the overall importance of phonetics with that of phonology by providing a weight of 2 to the phonetic score and 1 to each of the phonological scores. By doing this, a source-language phoneme can have a greater phonetic distance to some target-language phoneme than other source-language phonemes but a lower phonological distance and receive a lower overall phoneme distance score. It is because phonological distance is considered as important as phonetic distance that the overall constructed target-language model pool will tend to be restricted to a subset of phonologically similar languages.

The feature-based phoneme distance metric is defined as

(18)

$$CPP(i,k) = \alpha_d \cdot [d_{lt}(i,k)]_N + \alpha_\rho \cdot [D\rho_{lt}^g]_N + \alpha_\gamma \cdot [D\gamma_{lt}^g]_N$$

where $CPP(i,k)$ 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. As mentioned, ($\alpha_d, \alpha_\rho, \alpha_\gamma$)=(2,1,1). The symbol $[\cdot]_N$ denotes that the quantity inside is linearly scaled into the range [0 1]. For $D\rho_{lt}^g$ and $D\gamma_{lt}^g$, 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 any source language.

5 Experiments

To test our CPP approach to phoneme similarity prediction, we compared it to an acoustic distance approach in ASR experiments. Because native language speech data is used in measuring model distance in the acoustic approach, it is expected to work better than the knowledge-based approach, which only estimates acoustic similarity indirectly through articulatory phonetic distance and overall phonological distance.

5.1 Model construction

We employ the regular 3-state, left-right, multimixture, continuous-Gaussian HMMs as the acoustic models and assume that the models from all the source and target languages have the same topology except that the number of mixtures in a state may vary. Once the top source phonemes are determined from our feature-based phoneme distance metric for each target phoneme, the target HMM is constructed by gathering all the mixtures for a corresponding state from the source candidates. The original mean and variance values are maintained while the mixture weights are uniformly scaled down so that the new weights add up to one for each state. It is possible to weigh mixtures according to the relative importance of the candidates if the importance as reflected by the phoneme distance metric has a significantly large difference. The transition probabilities are adopted from the top one candidate model.

5.2 CPP phoneme model construction

We used the 17 languages and dialects provided in Table 2 in the experiments testing our CPP phoneme distance approach to phoneme HMM similarity. For each language, a native monolingual model set had been built by training with native speech data. The acoustic features are 39 regular MFCC features including cepstral, delta, and delta-delta. The individual ASR databases derive from a variety of projects and protocols, including, but not limited to, CallHome, EUROM, SpeechDat, Polyphone, and GlobalPhone. In each of the following experiments, we select one language as the target language, and construct its acoustic models by using all the other languages as source languages. A phoneme distance score is calculated for each target phoneme and the top two candidate source-language phonemes are chosen for HMM model construction. We conducted experiments with Italian, Latin American Spanish, European Portuguese, Japanese, and Danish as target languages.

5.3 Acoustic model construction

In the acoustics distance approach, models are built with the top two models chosen from source languages based on their acoustic distance from the corresponding native target model. For these experiments, we adopt the widely used Bhattacharyya metric for the distance measurement (Mak and Barnard 1996). It should be noted that the recognition performance of the acoustics-constructed models is not a theoretically strict upper bound for HMM similarity because the measurement in the acoustic space is probabilistic.

5.4 Results

Each recognition task includes about 3000 utterances of digit strings, command words, and sentences. The word accuracy results in Table 4 include the native baseline performance, i.e. the performance of the native monolingual, context-independent models from each target language, as well as the acoustics-based and feature-based performances. These results show that the performance of models selected by the CPP phoneme distance approach is equivalent overall to that of models selected by acoustic distance.

Table 4: Model performance

Target Language	Native Baseline	Acoustic Distance	CPP Distance
Lat. Spanish	94.49	88.61	93.06
Italian	98.42	98.27	98.52
Japanese	95.36	76.72	78.76
Danish	94.36	72.95	70.15
Eur. Portuguese	96.31	77.91	72.74

The performance of models selected by the CPP approach nearly matches the performance of the *native* models for Latin American Spanish and surpasses those for Italian. This approach performs better than the acoustic distance approach for Latin American Spanish, Italian, and Japanese and not as well for Danish and European Portuguese.

6 Evaluation and conclusion

We suggest four principal performance factors to explain the results provided in Table 4: (1) rare phonemes in the target-language inventory; (2) target-language inventory complexity; (3) degree of source-language phonological distance to the target language; (4) reliability of source-language models. Because the CPP approach has only been tested on five languages, we consider this analysis preliminary.

Regarding the first factor, rare phonemes in the target-language inventory, it is worth noting that neither Latin American Spanish nor Italian has phonemes whose exact feature specifications are unattested in phonemes from other languages in our dataset. For these languages, all phonemes have exact source-language matches. In contrast, Japanese, Danish, and European Portuguese each contain phonemes with feature specifications unique to their language. Based on this analysis, we propose that, *all other factors being equal*, the greater the overall phoneme correspondence between the target language and the source languages, the better the target-language HMM performance.

In general, it appears that target languages associated with inventories that are greater in size than their least phonologically distant source languages perform worse than target languages associated with smaller inventories relative to their closest source languages. For example, the vowel systems of Danish, European Portuguese, and Japanese are the most complex of the five target languages, with Danish having 26 vowels, European Portuguese having 14 vowels, and Japanese having ten vowels. In sharp contrast, Latin American Spanish has only five vowels and Italian has seven. Both Latin American Spanish and Italian are phonologically similar to other Romance languages in the dataset that have greater vowel contrasts: Brazilian Portuguese (13 vowels), European Portuguese (14 vowels), Parisian French (17 vowels) and Canadian French (19 vowels). Here, we suggest that target languages that have a similar or lesser number of phoneme contrasts compared to the source languages are more likely to achieve higher recognition performances, *all other factors being equal*.

Relative phonological distance of the source languages to the target language and reliability of source language models additionally impact target-language ASR performance. Consider Table 5 where the difference in these factors for Italian and European Portuguese are given. First, Italian and European Portuguese are both Romance languages and our dataset includes a total of six, presumably phonologically similar, Romance languages and dialects. However, the recognition results of the models selected by both the feature-based and acoustics-based phoneme distance method are very different for the two languages.

Table 5: Phonological distance and native baseline performance factors in target-language recognition

Target Language	Italian	Eur. Portuguese
Top 3 least distant langs.	(1) Lat. Spanish (2) Parisian Fr. (3) Brz. Port.	(1) Brz. Port. (2) Lat. Spanish (3) Canadian Fr.
Avg. phonolog. distance of top 3 langs.	0.7399	0.8945
Avg. phonolog. distance of top 1 lang.	0.5757	0.8248
Avg. native baseline of top 3 langs.	89	91.94
Native baseline of top 1 lang.	94.49	84.25

If we compare the phonological distances between the least distant source languages to Italian and European Portuguese, we observe that Italian's closest languages are less distant overall than European Portuguese's closest languages.

Because the phonologically least distant source languages contribute the majority of target-language HMMs, it is reasonable to expect that lesser phonological distance to the target language by a greater number of source languages is likely to result in a better target-language HMM performance, *all other factors being equal*.

Finally, note the substantial discrepancy in native baseline performance between the phonologically least distant source languages for Italian and European Portuguese. The majority of selected models for Italian derive from Latin American Spanish which is associated with a high native recognition baseline. European Portuguese models, on the other hand, largely come from Brazilian Portuguese which has a much lower native baseline. This suggests that the most reliable source-language HMMs, as judged from their native recognition performance, contribute to better target-language recognition performance, *all other factors being equal*.

In future work, we intend to test our CPP phoneme similarity approach on new target languages and expand the preliminary evaluation provided here. In particular, we are interested to what extent this method can predict recognition performance for new target languages.

References

Archangeli, D., “Aspects of Underspecification Theory”. *Phonology* 5:183-207, 1988.

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

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

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

Kessler, B., “Phonetic comparison algorithms,” *Transactions of the Philological Society*, 2005

Köhler J., “Multilingual phoneme recognition exploiting acoustic-phonetic similarities of sounds,” *ICSLP’96*, 2195-2198, Philadelphia, 1996.

Kondrak, G., *Algorithms for Language Reconstruction*, Ph.D. thesis, University of Toronto, 2002.

Ladefoged P., “The measurement of phonetic similarity,” *Int Conf on Comp Linguistics*, Stockholm, Sweden, 1969.

Ladefoged P. *A Course in Phonetics*. Harcourt Brace Jovanovich, New York, 1975.

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.

Liu, C. and Melnar, L., “Training acoustic models with speech data from different languages,” *MULTILING’06*, Stellenbosch, 2006.

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

Nerbonne, J. and Heeringa, W., “Measuring dialect distance phonetically,” *Proc. 3rd Meeting ACL Special Interest Group in Comp. Phonology*, 1997.

Schultz, T. and Waibel, A., “Fast bootstrapping of LVCSR systems with multilingual phoneme sets,” *Eurospeech 97*, 1:371-373, 1997.

Schultz, T. and Waibel, A.., “Polyphone Decision Tree Specialization for Language Adaptation”, In *Proc. of ICASSP 2000*. Istanbul, 2000.

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.

Sooful, J. J. and Botha, E. C., “Comparison of acoustic distance measures for automatic cross-language phoneme mapping,” *ICSLP’02*, 521-524, 2002.