Borrowing Language Resources for Development of Automatic Speech Recognition for Low- and Middle-Density Languages

Lynette Melnar, Chen Liu

Application and Software Research Center, Motorola Labs Motorola, Schaumburg, IL 60196, USA Lynette.Melnar, Chen.Liu@motorola.com

Abstract

In this paper we describe an approach that both creates crosslingual acoustic monophone model sets for speech recognition tasks and objectively predicts their performance without target-language speech data or acoustic measurement techniques. This strategy is based on a series of linguistic metrics characterizing the articulatory phonetic and phonological distances of target-language phonemes from source-language phonemes. We term these algorithms the Combined Phonetic and Phonological Crosslingual Distance (CPP-CD) metric and the Combined Phonetic and Phonological Crosslingual Prediction (CPP-CP) metric. The particular motivations for this project are the current unavailability and often prohibitively high production cost of speech databases for many strategically important low- and middle-density languages.

First, we describe the CPP-CD approach and compare the performance of CPP-CD-specified models to both native language models and crosslingual models selected by the Bhattacharyya acoustic-model distance metric in automatic speech recognition (ASR) experiments. Results confirm that the CPP-CD approach nearly matches those achieved by the acoustic distance metric. We then test the CPP-CP algorithm on the CPP-CD models by comparing the CPP-CP scores to the recognition phoneme error rates. Based on this comparison, we conclude that the CPP-CP algorithm is a reliable indicator of crosslingual model performance in speech recognition tasks.

1. Introduction

Speech technologists typically use acoustic measurements to determine similarity among acoustic speech models for crosslingual modeling and there are a variety of distance metrics available (e.g., Sooful & 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 principally restricted to high-density languages while the vast majority of languages are low- and middle-density languages. Low- and middle-density languages have not been exposed to intense data collection and resources for these languages are subsequently either limited or completely unavailable.

A high-density language is characteristically a majority language associated with a large, economically advantaged population of speakers, a significant proportion of which regularly uses computers or has computer experience. The language bears official status or non-official predominant use in one or more countries and is also recognized as important by foreign governments. Finally, the language is supported by a writing tradition and has been studied and well documented in various types of language resources. In regards to speech modeling, it is this last factor of language resource availability that proves most significant. In particular, a high-density language is associated with several commercially available speech resources of various types (and quality). Examples of such high-density languages are major dialects of Arabic, English, Mandarin, and Spanish.

A low-density language, in contrast, lacks many of the high-density characteristics. The population of speakers may be quite small, economically disadvantaged, and have little or no computer experience. The language may be considered a minority language in the countries where it is spoken and is not judged to be very significant by foreign governments or language researchers. A writing system may be completely lacking for the low-density language or only non-standardized writing systems may exist in limited use; thus language resources are sparse or non-existent. Many Native American and minor African languages would fall into this category. Finally, a middle-density language shows a balance of those extremes exhibited by high- and low-density languages and would include, for example, many Chinese, Indian, and African languages in the emerging market.

Hence crosslingual modeling and performance prediction that necessitate no target-language speech data have a great value for the resource limited low- and middle-density languages by taking advantage of the well-trained models of, typically, high-density languages. Our approach is specifically suited to target low- and middle-density languages that are (1) associated with a writing system but perhaps lack significant, readily available language resources and (2) related to other, typically high-density languages for which language resources are easily accessible.

While this approach was developed with low- and middle-density languages in mind, it may also be an adequate alternative modeling strategy for any language when database acquisition time or expense is limited.

2. Prior Work

The general objective of crosslingual modeling for ASR is to create a set of acoustic models from one or more source languages for a target language and for that acoustic model set to achieve adequate recognition performance in the target language. To do this, it is generally recognized that the selected source models must be similar in phonetic coverage to the corresponding target-language phonemes. Thus, crosslingual model creation and recognition prediction is largely a problem in phoneme similarity estimation.

Purely knowledge-based approaches to phoneme similarity prediction 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 & Waibel, 1997 & 2000).

In contrast to the label-based strategy, researchers in such fields as dialectometry, language acquisition, and language reconstruction 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 & Heeringa, 1997; Somers, 1998).

In principle, the feature-based approach 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).

A strategy for specifying allophony and characterizing phonetic distance or similarity between phonemes is only one component in predicting phoneme similarity without acoustic data. Because phonemes necessarily *interact*, it is also important to consider the phonology, or at least, the phonotactics of the overall constructed system. Ideally, the distribution characteristics of the resulting acoustic models will match the distribution characteristics of their corresponding target-language phonemes.

3. Crosslingual Model Creation and Prediction Method

The crosslingual model creation and performance prediction methods described here are based on a combination of metrics characterizing the articulatory phonetic and phonological distribution distances of phonemes from source and target languages. For this reason, we identify the collection of measurements and the approach overall as *Combined Phonetic and Phonological* or CPP (Liu & Melnar, 2005). The only resource assumed for each target language is a high-quality pronunciation dictionary. The first step in the CPP approach is the definition of each source and target language phoneme by a set of 32 distinctive feature categories specifically designed to capture allophonic variation. Next, the phonetic distance between each target-language phoneme and all source-language phonemes is measured. We base our phonetic distance measurement on weighted articulatory features where the value of a weight for a feature is derived from the frequency of the feature in all the source-language lexica. In our experiments, we use the Manhattan distance where the distance between phonemes equals the sum of the absolute values of individual feature distances.

The phonological distance between the source and target languages is then estimated by calculating the lexical monophoneme and biphoneme distribution distances between each source language and the target language. For each language, a pronunciation lexicon is used for this purpose. The pronunciation lexicon and a phonetic description of the pronunciation labels used are the only language resources required for the target language in this approach.

The phonetic distance, monophoneme and biphoneme phonological distance scores are then weighted so that the impact of phonetic distance equals that of phonological distance; these scores are subsequently combined to identify the closest source-language phoneme matches for each target-language phoneme. The pre-existing acoustic models for the top two most similar source-language phonemes for each target-language phoneme are selected for the crosslingual target-language model set. Because phoneme similarity is based on the combined scores of phonetic and phonological distances, we call this metric the *CPP Crosslingual Distance* or CPP-CD. The CPP-CD model is schematized in Figure 1.



Figure 1: Schematization of the CPP-CD approach

The precise formulation of CPP-CD metric is provided in Liu & Melnar (2005) and is not repeated here.

We similarly predict the performance of the crosslingual model set by measuring the weighted sum of relative phoneme distances. Each target-language phoneme is assigned an importance weight based on the phoneme's lexical frequency. Then the contribution and interference effect of all the *donor* phonemes to each target-language phoneme is measured. For each target phoneme, a *matching* and *confusing* phoneme set is

defined. A matching set consists of the most similar (i.e. least distant) donor phonemes; these have a contribution effect in target phoneme identification. A confusing set consists of all the donor phonemes except those in the corresponding matching set; because confusing phonemes might have some phonetic proximity to the target phoneme, they potentially have an interference effect in target phoneme identification.

Since individual phoneme contribution is represented by a distance measure, the total contribution is derived as a harmonic sum of individual contributions. The total interference effect relative to a target phoneme from a set of confusing phonemes is derived from the distance of each individual donor phoneme in the same way. The contribution or interference effect of a donor phoneme to the target phoneme is derived from the same component distance measures used in the derivation of CPP-CD, i.e. it consists of the phonetic, monophoneme, and biphoneme distance measurements. We therefore call the prediction algorithm the *CPP Crosslingual Prediction* or CPP-CP. See Figure 2 for a model of the CPP-CP algorithm.

Because prediction is distance-based, the smaller the CPP-CP score, the higher the predicted performance of the crosslingual models. For a set of languages, we evaluate CPP-CP scores relative to recognition phoneme error rates (recognition results with the crosslingual models on the native speech data). The reliability of the prediction score is verified by its agreement with the phoneme error rate in recognition tests, as shown in the next section.



Figure 2: Schematization of the CPP-CP approach

Because the prediction score is based on distance, the smaller the CPP-CP score, the higher the predicted performance of the crosslingual models. For a set of languages, we evaluate CPP-CP scores relative to recognition phoneme error rates (recognition results with the crosslingual models on the native speech data). Where the prediction scores match the general trend of the recognition phoneme error rates, we consider the relative prediction scores reliable.¹

In general, recognition error is expected to be higher for those languages having larger biphoneme inventories. A higher number of biphonemes in a language correlates to more phonotactic possibilities and thus greater potential confusability among monophoneme models. Furthermore, since the CPP-CD algorithm is based on phonetic and phonological distances, it follows that overall target-source language *proximity* is also significant. Two languages may be said to be proximate if they are closely related linguistically or if the populations of speakers intermingle or otherwise habituate coterminous areas (i.e. are contact languages (Trask, 1996)). A higher degree of phonological similarity is expected among related and contact languages than among languages that lack such proximate relations.

In addition to biphoneme inventory size and language proximity, the quality and consistency of the language resources must be considered important performance In particular, negative performance factors factors. include: (1) inconsistency in data quality and task complexity across languages due to database availability and (2) sub-optimal native model quality for some languages due to training data insufficiency. In conducting these experiments, we have made no effort to improve or harmonize the existing language resources: rather, our intent is to test the CPP strategy with our current databases. This reflects a very practical business scenario where time is not available for extended database processing.

4. Experiments, Results, and Analysis

4.1 CPP-CD

To test our CPP-CD approach to modeling, we first compare it to both an acoustic distance approach and to native monolingual modeling in ASR experiments targeting five high-density languages for which we have speech data for testing: Latin American Spanish, Italian, Japanese, Danish, and European Portuguese. For ease of presentation, we refer to each major target or source dialect as a language in the remainder of this paper.

For the acoustic model distance measurement, we adopt the Bhattacharyya metric (Mak & Barnard, 1996). The reference models are built with the top two native models chosen from source languages based on their acoustic distance from the corresponding native target model.

For native language modeling, a native monolingual model set had been built by training with native speech data for each of the target languages. The acoustic features are 39 regular MFCC features including cepstral, delta, and delta-delta. The databases included CallHome, VAHA, EUROM, SpeechDat, and GlobalPhone, among others. Because the models trained with native language speech data are used in measuring model distance in the Bhattacharyya metric, it is expected to work better than our CPP-CD metric, which only estimates acoustic similarity indirectly through articulatory phonetic similarity and overall phonological

¹ At the time of LREC submission, the details of the CPP-CP approach were undergoing internal clearance review; for this reason, formulae are not presented.

similarity.

For both the CPP-CD and Bhattacharyya approaches, we use twenty languages from six major language groups defined by genetic relation: (i) Afro-Asiatic: Egyptian Arabic; (ii) Altaic: Japanese, Korean; (iii) Germanic: American English, British English, Danish, Dutch, German, Swedish; (iv) Italic: Brazilian Portuguese, Canadian French, European Portuguese, Italian. Latin-American Spanish, Parisian French; (v) Sinitic: Cantonese, Mandarin, Shanghainese; and (vi) Slavic: Czech, Russian. Among these twenty languages, three may be considered middle-density languages in the sense that they are relatively underrepresented in terms of available speech resources.² These are Brazilian Portuguese, Cantonese, and Shanghainese. For each crosslingual experiment, the target language is left out of the source language pool for model selection.

The word recognition results of these experiments are provided in Table 1, along with biphoneme inventory sizes.

Observe that the performance of CPP-CD constructed models nearly matches the performance of the native models for Spanish and surpasses those for Italian. The CCP-CD approach performs better than the Bhattacharyya acoustic distance approach for Italian, Spanish, and Japanese and not as well for Portuguese and Danish. Overall the CPP-CD and Bhattacharyya acoustic approaches perform similarly. The average recognition result using the Bhattacharyya-derived models is 82.89% while the average CPP-CD result is 82.65%; thus, the difference is trivial.

Target	Biphon.	Native	Acoustic	CPP-CD
Language	Inv.	Baseline	Distance	CFF-CD
Italian	613	98.42	98.27	98.52
Spanish	520	94.49	88.61	93.06
Japanese	643	95.36	76.72	78.76
Portuguese	776	96.31	77.91	72.74
Danish	980	94.36	72.95	70.15

Table 1: Model performance comparison (word accuracy %)

As noted, we expect crosslingual monophoneme models to perform better for languages with fewer biphonemes. This expectation is validated in Table 1. The two languages with the lowest number of biphonemes are Italian and Spanish and their crosslingual models have the best word recognition performances. The largest biphoneme inventory belongs to Danish, and both Danish crosslingual model sets have the worst performance results. For Japanese and Portuguese, the relative performance differs using the two distance metrics. Portuguese has the third best performance and Japanese, the fourth, using the Bhattacharyya method, while these languages are swapped in performance rank using the CPP-CD approach.

Three of the five test languages belong to the Italic branch of the Indo-European language family (Spanish, Italian, and Portuguese), and there are a total of six Italic source languages. This leaves five Italic source languages for each of the three Italic test languages. The test languages include one Germanic language (Danish); this language is also associated with five closely related languages. Besides these close inter-group relationships, Germanic and Italic languages are distantly related to each other and have a long history of close contact. In contrast with this, Japanese is only related to one source language, Korean, though does have a long history of contact with the Sinitic languages.

Based solely on language proximity, we would predict the Italic and Germanic languages to perform the best. Spanish and Italian do conform to this expectation, while Danish and Portuguese do not. Above, Danish's crosslingual model performance is correlated with Danish's large biphoneme inventory size. We observe here that Portuguese has the second highest number of biphonemes and suggest that this might likewise account for the relatively low crosslingual model performances.

In conclusion, the CPP-CD crosslingual performance results are comparable to those derived from the acoustic model distance measurement and conform to expectations based on the known performance factors of biphoneme inventory size and language proximity.

4.1 CPP-CP

We now use the CPP-CP algorithm to generate prediction scores for each of the five test languages. Recall that phoneme error percentage is the phoneme recognition result using the CPP-CD-derived models on the native speech data. If the trend of the prediction scores matches the trend of the phoneme error rates, we consider the relative prediction scores reliable.



Figure 3: Comparison of phoneme error rates and CPP-CP scores (1)

² We consulted the speech resource catalogues of the Linguistic Data Consortium (LDC), European Language Resources Association (ELRA), and Appen Ltd. to informally evaluate speech database availability. For both Shanghainese and Cantonese only one speech database is available, while Brazilian Portuguese is associated with three speech databases.

Figure 3 confirms that the trend of the CPP-CP scores does match that of the phoneme error rates, and, based on these results, we may tentatively conclude that these CPP-CP scores are indicative of actual recognition performance. The results presented in Figure 3 furthermore suggest that target-language crosslingual model sets that have a CPP-CP score of less than 1 are likely to achieve acceptable recognition performance levels (Spanish, Italian, and Japanese CPP-CP models average a word accuracy of 90.11%).

Testing with additional, lower-density target languages substantiates this judgment. In a second experiment, we create CPP-CD-derived models for the three identified middle-density languages, Brazilian Portuguese, Cantonese, and Shanghainese. We use the same twenty source languages selected for the first experiment and likewise compare their crosslingual phoneme error rates and CPP-CP scores.

Consider now in Table 2 the biphoneme inventory size and number of closely related languages corresponding to each of the eight test languages.³ Brazilian Portuguese is another Italic language and thus is closely related to five of the remaining nineteen source languages. Like European Portuguese, the Brazilian Portuguese biphoneme inventory is large; in fact, with 1046 biphonemes, Brazilian Portuguese has the largest test language biphoneme inventory.

Target	Biphoneme	Related
Language	Inventory	Languages
Average	665.75	4.125
Cantonese	320	2
Shanghainese	428	2
Spanish	520	5
Italian	613	5
Japanese	643	1
Eur. Portuguese	776	5
Danish	980	5
Br. Portuguese	1046	5

 Table 2: Comparison of performance factors among the eight test languages

Cantonese and Shanghainese are both Sinitic languages; besides being closely related to each other, they are also genetically related to Mandarin. Their biphoneme inventories are the smallest with 320 and 428 biphonemes respectively.

If we hold average values as a relative indication of performance, we see that among the eight languages, only two, Italian and Spanish are above average for both factors.

Figure 4 provides the phone error rates and prediction scores for the complete set of test languages. As in Figure 3, we see that the trends of the CPP-CP scores and phone error percentages correspond. Among the middle-density languages, only Cantonese has a prediction score below 1, indicating that its crosslingual model set is expected to achieve a practical use recognition level. We suggest that the very small biphoneme inventory size of Cantonese contributes to this relatively low prediction score.



Figure 4: Comparison of phoneme error rates and CPP-CP scores (2)

We note that the Shanghainese crosslingual models perform less well than predicted relative to the other languages. This is surprising given that the Shanghainese biphoneme inventory size is very small and the number and genetic type of related languages are the same as those of Cantonese. However, the Shanghainese database has not been subject to validation that would confirm its overall quality and consistency. We therefore admit the possibility that the poor Shanghainese crosslingual model performance is attributable to substandard native-language resources.

5. Conclusion

The experiments presented herein demonstrate that the CPP strategy is an effective method for both selecting monophone models for crosslingual recognition and predicting the recognition performance of the derived model sets. Because the CPP approach requires no target-language acoustic data, it is especially useful for creating and validating crosslingual model sets for target languages lacking speech data resources, such as low- and middle-density languages. As many emerging markets are populated by low- and middle-density language speakers, this tool can be of great assistance in entering a market quickly and cost-effectively.

6. References

- Archangeli, D. (1988). Aspects of underspecification theory. Phonology, v. 5, pp. 183-207.
- Connolly, J.H. (1997). Quantifying target-realization differences. Clinical Linguistics & Phonetics, v. 11, pp. 267–298.
- Hughes, B. & Maxwell, M. (2006). Frontiers in linguistic annotation for lower-density languages. Proc. Frontiers in Linguistically Annotated Corpora, COLING-ACL,

³ Only *related* language information is provided here. Establishing relevant contact language status is beyond the scope of this study.

pp. 29-37.

- Kessler, B. (1995). Computational dialectology in Irish Gaelic. Proc. 6th Conf. European Chapter of ACL, pp. 60–67.
- Kessler, B. (2005). Phonetic comparison algorithms. Transactions of the Philological Society, v. 103, pp. 243-260.
- Köhler J. (1996). Multilingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. ICSLP'96, pp. 2195-2198.
- Kondrak, G. (2002). Algorithms for Language Reconstruction, Ph.D. thesis, University of Toronto.
- Liu, C. & Melnar, L. (2005). An automated linguistic knowledge-based cross-language transfer method for building acoustic models for a language without native training data. Interspeech'2005-Eurospeech, pp. 1365-1368.
- Mak, B. & Barnard, E. (1996). Phone clustering using the Bhattacharyya distance. ICSLP'96, 2005-2008.
- Nerbonne, J. & Heeringa, W. (1997). Measuring dialect distance phonetically. Proc. 3rd Meeting ACL Special Interest Group in Comp. Phonology, pp. 11-18.
- Schultz, T. & Waibel, A. (1997). Fast bootstrapping of LVCSR systems with multilingual phoneme sets. Proc. Eurospeech 97, pp. 371-374.
- Schultz, T. & Waibel, A.. (2000). Polyphone decision tree specialization for language adaptation. Proc. of ICASSP. Istanbul.
- Somers, H.L. (1998). Similarity metrics for aligning children's articulation data. Proc. 36th Annual Meeting ACL and 17th Int. Conf. Comp. Ling., pp. 1227–1231.
- Sooful, J.J. & Botha, E.C. (2002). Comparison of acoustic distance measures for automatic cross-language phoneme mapping. ICSLP'02, pp. 521-524.
- Trask, R. (1996). A Dictionary of Phonetics and Phonology. London: Routledge, p. 90.