A Non-Acoustic Approach to Crosslingual Speech Recognition Performance Prediction

Chen Liu, Lynette Melnar
Motorola Labs, Motorola, Schaumburg, IL 60196, USA
Chen.Liu, Lynette.Melnar@motorola.com

Abstract

Crosslingual acoustic modeling is an effective technique for building acoustic models in the absence of native training data. A small amount of native speech data is still needed for verifying the crosslingual models by running an actual recognition test. In some very stringent yet realistic situations, however, even the test data may not be available. We introduce an algorithm that objectively predicts the recognition performance of crosslingual acoustic models. This approach does not require conducting of actual speech recognition tests with target-language speech data; nor does it depend on any acoustic measurement techniques. The algorithm is based on a series of linguistic metrics characterizing the articulatory phonetic and phonological information of phonemes from both the target and source languages. It is useful both for validating crosslingual models for speech recognition applications, and for making database acquisition decisions that could prove very cost-beneficial.

Index Terms: acoustic modeling, crosslingual modeling, phonetic metric, phonological metric, speech recognition

1. Introduction

Building acoustic models for a new language market is usually slowed down due to a need for a sufficient amount of native speech data since data acquisition is normally very time-consuming and costly. Researchers have found a solution by leveraging crosslingual benefit (e.g., [1][2][3][4]). The models for the target language are built with pre-existing acoustic models from available foreign source languages (called donor languages in this context). In those approaches, a small amount of native data is required during the model development process and performance verification stage. Previously we presented an automated, linguistics-based method which completely eliminated any requirement for the native training data for making the crosslingual acoustic models [5]. Yet like those former approaches, it still needed a small amount of native data for verifying the performance of the crosslingual models by running actual recognition tests. Practically speaking, however, even a small amount of target-language test data is not always available, yet the ability to estimate the performance of the crosslingual models for that language may be very much desired. The most obvious scenario involves speech database acquisition decisions. For example, the savings in cost would be considerable if the decision could be made to not purchase a particular high-priority language database because the performance of the crosslingual models built for that language can reach a practical use level. Further, crosslingual prediction ability allows for the strategic purchasing of databases based on maximum benefit; that is, in making purchasing or collecting decisions, consideration would be given to which high priority language databases among several would be the most useful for donating models for crosslingual application so that subsequent database purchases can be minimized. In this study, we propose a prediction algorithm that estimates the recognition performance of the crosslingual models built in [5] without need for any native data from the new target language.

The prediction algorithm is based on the crosslingual model selection algorithm we previously developed; therefore we provide a brief description of this prior work. Readers can refer to [5] for details.

In our system, the phoneme inventory, containing \( l \) phonemes for a language \( l \), is denoted as \( \{ p_l(i) | i = 1, \ldots, l \} \). A phoneme \( p_l(i) \) is represented by a vector of \( J \) phonetic features,

\[
[v_l(i,1), v_l(i,2), \ldots, v_l(i,J)]^T
\]

where each \( v_l(i,j) \) is a binary feature, i.e., 1 for present and 0 for absent, \( i = 1, \ldots, l \), \( j = 1, \ldots, J \), \( l = 1, \ldots, L \), and the superscript \( T \) denotes vector transposition.

The donor models are selected from all the potentially contributing source languages based on a combined phonetic-phonological (CPP) crosslingual metric, which in turn is derived from more basic linguistic metrics including phonetic distance, monophoneme distribution distance, and biphoneme distribution distance, thus

\[
CPP(i,k) = \alpha_d \cdot |d_k(i,k)|_N + \alpha_p \cdot |D_p|_N + \alpha_g \cdot |D_g^r|_N
\]

where \( CPP(i,k) \) represents the distance between phoneme \( p_l(i) \) from language \( l \) and phoneme \( p_k(j) \) from language \( k \), and both phonemes belong to the same phonological category \( g \) (vowels or consonants). The weights \( \alpha_d \), \( \alpha_p \), and \( \alpha_g \) represent the relative importance of each quantity. We equate the overall importance of phonetics with that of phonology by using the weight values \( \{ \alpha_d, \alpha_p, \alpha_g \} \approx (2,1,1) \). The symbol \( |x|_N \) denotes that the quantity inside is linearly scaled into the range \([0,1]\).

The phonetic distance \( d_l(i,k) \) in Eq. (2) characterizes the difference in weighted distinctive feature distribution between two phonemes, and it is calculated in the following way

\[
d_l(i,k) = \sum_{j=1}^{J} w(j) |v_l(i,j) - v_k(j,k)| \quad i = 1, \ldots, l \quad k = 1, \ldots, L
\]

The feature-importance weights are derived from the feature occurrence count of a phoneme \( p_l(i) \) in the lexica of the source and target languages \( l = 1, \ldots, L \), specifically,

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

where \( c_l(p_l(i)) \) refers to the occurrence count of a phoneme \( p_l(i) \) in the lexica of the source and target languages \( l = 1, \ldots, L \), specifically,
where the term \( c_j[p_t(i)]v_t(i,j) \) is the frequency of each feature \( j \) contributed by the phoneme \( p_t(i) \), and 
\[
\sum_{j} c_j[p_t(i)]v_t(i,j)
\]
represents the frequency of each feature \( j \) contributed by all the phonemes in language \( l \).

The monophoneme distribution distance \( D\rho^g \) and biphoneme distribution distance \( D\rho^g \) in Eq. (2) characterize the difference in lexical phoneme distribution and biphoneme distribution, respectively, between two languages \( l \) and \( t \). The distribution, or normalized histogram, of the phonemes or biphonemes of a language is obtained from a large lexicon, with the probability in the distribution derived from the frequency of a phoneme or biphoneme in the lexicon. Inferred from the definition of phoneme inventory for a single language, a combined inventory for both languages as well as the result of algorithm verification tests. We also give some practical examples of applying the technique.

2. Method

The prediction algorithm is based on a weighted sum of some specially defined relative inter-phoneme distances. The prediction score for the crosslingual models of language \( t \) is calculated as follows:
\[
P_t = \sum_{k} D[p_t(k)]\psi_t[p_t(k)]
\]
(10)
where \( \psi_t[p_t(k)] \) is the importance weight of phoneme \( p_t(k) \) in the target language \( t \). In our approach, the importance weight is represented by the frequency of the phoneme in a typical lexicon of the language:
\[
\psi_t[p_t(k)] = \frac{c_t[p_t(k)]}{\sum_{i} c_t[p_t(i)]}
\]
(11)
where \( c_t[p_t(k)] \) is the occurrence count of phoneme \( p_t(k) \) in the lexicon. The quantity \( D[p_t(k)] \) in Eq. (10) is derived as
\[
D[p_t(k)] = \frac{D_\gamma[p_t(k)]}{D_\gamma[p_t(k)]}
\]
(12)
and includes the contribution effect \( D_\gamma[p_t(k)] \) of all the donor phonemes to the target phoneme \( p_t(k) \) and simultaneously the interference impact \( D_\gamma[p_t(k)] \) of all the confusing phonemes to the target phoneme. The set of donor phonemes, denoted by \( S[p_t(k)] \) or \( S \), consists of all the phonemes selected from the source languages to represent the target phoneme \( p_t(k) \). The confusing phoneme set, \( \bar{S}[p_t(k)] \) or \( \bar{S} \), is composed of the donor phonemes selected to represent the other phonemes except \( p_t(k) \) in the target language; they might have some linguistic proximity to the target phoneme \( p_t(k) \) and cause confusion or substitution in speech recognition.

The whole contribution to the target phoneme \( p_t(k) \) from a set of donor phonemes \( S[p_t(k)] \) is obtained from the contribution \( D[p_t(i)][p_t(k)] \) of each individual donor \( p_t(i) \) in the following manner:
\[
D_\gamma[p_t(k)] = \sum_{\beta[p_t(i)] \in S[p_t(k)]} D[p_t(i)][p_t(k)]^{-1}
\]
(13)
Since individual contribution is represented by a distance measure (see Eq. 15), the total contribution is derived from individual contributions as a harmonic sum, which is normally used to compute a summation of individual effects that decrease with distance. Likewise, the total interference effect to the target phoneme \( p_t(k) \) from a set of confusing phonemes \( \bar{S}[p_t(k)] \) is derived from the distance \( D[p_t(i)][p_t(k)] \) of each individual \( p_t(i) \) in the same way:
\[
D_\gamma[p_t(k)] = \sum_{\beta[p_t(i)] \in \bar{S}[p_t(k)]} D[p_t(i)][p_t(k)]^{-1}
\]
(14)
In Eqs. (13) and (14), the contribution or interference effect of a phoneme \( p_t(i) \) from language \( l \) to the target phoneme \( p_t(k) \) in language \( t \) is embodied by a combinatorial distance measure, similar to CPP in Eq. (2), namely,
\[
D[p_t(i)][p_t(k)] = \alpha_p[D_\rho^t][p_t(k)] + \alpha_p[D_\gamma^t][p_t(k)] + \alpha_p[D_\gamma^t][p_t(k)]
\]
(15)
where the symbol $\{\cdot\}$ denotes a scaling operation, while the other quantities and symbols including the phonetic distance $d_{\theta}(i,k)$, monophoneme distribution distance $D\rho_{\theta}^g$, and biphoneme distribution distance $D\gamma_{\theta}^g$, the superscript $g$, and the weights $\alpha_d, \alpha_{\rho}$, and $\alpha_{\gamma}$ are the same as in Eq. (2). In contrast to CPP crosslingual model selection algorithm, two new weights $\eta_{\rho}$ and $\eta_{\gamma}$ are introduced to represent the effect of the numbers of monophonemes and biphonemes (or the sizes of the phoneme and biphoneme inventories) in the target language $t$ as they play a role in the recognition performance of models (understandably, recognition error tends to be higher for languages with greater numbers of phonemes and biphonemes). Thus:

\[ \eta_{\rho} = \frac{I_t}{\kappa_{\rho}} \]

\[ \eta_{\gamma} = \frac{I'_t}{\kappa_{\gamma}} \]

where $I_t$ and $I'_t$ are the numbers of monophonemes and biphonemes in the target language $t$, respectively; whereas $\kappa_{\rho}$ and $\kappa_{\gamma}$ are normalization factors. We use the average monophoneme number and biphoneme number across all the available source languages as the normalization factors, thus:

\[ \kappa_{\rho} = \text{avg}(I_t) \]

\[ \kappa_{\gamma} = \text{avg}(I'_t) \]

Because the prediction score is based on distance measures, it follows that a relatively small prediction score $P_t$ forecasts a relatively high performance of the crosslingual models. Derived solely in the linguistic domain, a prediction score $P_t$ does not numerically associate with any recognition score such as phoneme or word error rate (PER or WER). However, the verification experiment in the next section shows that the prediction score is certainly an indication of performance of crosslingual models. In the experiment, we will use PER as reference since PER obtained with an open grammar is a more direct indication of the quality of individual phonetic models.

3. Experiment

Twenty languages are used for which we have test data available. These are American English (en-US), Brazilian Portuguese (pt-BR), British English (en-GB), Canadian French (fr-CA), Cantonese (zh-CN-Cant), Czech (cs-CZ), Danish (da-DK), Dutch (nl-NL), Egyptian Arabic (ar-EG), European Portuguese (pt-PT), German (de-DE), Italian (it-IT), Japanese (ja-JP), Korean (ko-KR), Latin-American Spanish (es-LatAm), Mandarin (zh-CN-Mand), Parisian French (fr-FR), Russian (ru-RU), Shanghaiinese (zh-CN-Shang), and Swedish (sv-SE). For each language, a native monolingual model set had been built by training with native speech data; it is only used as source material for building the crosslingual models.

The acoustic features are 39 MFCC features including cepstral, delta, and delta-delta. The databases are diverse in design and include CallHome, EUROM, SpeechDat, etc. The utterances are phrases and sentences. In each of the tests, we select one language as the target language, and construct its crosslingual models with source material selected from the other nineteen languages. A CPP distance score is calculated for each target phoneme and the top two candidate phonemes are chosen. Their associated native acoustic models are used for building the crosslingual model set. The PERs of the crosslingual models are obtained with the native target test data. Because maximum language coverage is desirable for crosslingual model evaluation, even language databases of relatively poor quality, including data insufficiency, are included in this experiment. Thus, some PERs in the recognition tests are very high.

![Figure 1: Agreement between prediction and PER.](image)

The PERs of such crosslingual models on the native target test data is shown in Fig. 1. The same top two candidate phonemes for each target phoneme are also used in computation of the prediction score (Eq. 10) for each target language. The prediction scores for all the twenty languages are also displayed in Fig. 1. Derived with different scaling and in different domains, the prediction score and actual PER are not comparable numerically. However, the reliability of the prediction score can be judged from its consistency and correspondence with PER. For this reason, all the languages in Fig. 1 are sorted so that the prediction score increases monotonically. The figure shows that the PER trend is similar to that of prediction; in particular, it shows that a language with a high prediction score generally has a high PER. To objectively and quantitatively evaluate the correspondence between the prediction and PER, we employ the well-known Kendall’s $\tau$ distance [7] in statistics, i.e.,

\[ \tau = \frac{4P}{n(n-1)} - 1 \]

where $P$ is the number of concordant pairs (A concordant pair is a pair of languages for which the two measures agree), and $n$ is the number of total languages, i.e., 20. The normalized Kendall’s $\tau$ distance is 0.75, indicating a very strong agreement between the two measures, prediction and PER.

However, some discrepancies are observed between the prediction scores and the phoneme errors. For example, the cs-CZ and de-DE model sets are predicted to perform worse than they actually do, while the en-GB and zh-CN-Shang models are predicted to perform better than the verification.

---

1 The scaling scheme is $[A]_t = A/(A_{\text{max}} - A_{\text{min}})$, where $A$ can be $d_{\theta}(i,k)$, $D\rho_{\theta}^g$, or $D\gamma_{\theta}^g$. That is, the quantity is scaled as if it is processed in the normalization but without the zero line adjustment $\frac{A_{\text{min}}}{A_{\text{max}} - A_{\text{min}}}$. 

2721
tests indicate. These discrepancies are believed to be caused by the following four factors: (i) inconsistency in data quality and task complexity across languages (due to database availability); (ii) sub-optimal native model quality for some languages (due to training data insufficiency); (iii) non-unary mapping between the acoustic and non-acoustic phonological domains (thus inherently limiting the capability to predict performance in one domain based solely on the data in another); and (iv) omitting other potentially useful linguistic parameters in prediction. It is noted that the first two factors could be alleviated by conducting this same experiment using only language databases developed under the same conditions and guidelines (for example, all SPEECON databases).

Based on the comparison between prediction and PER, we suggest that languages with prediction score in the range between 0 and 1 (close to the left end of the language axis) are most likely to achieve acceptable recognition scores.

4. Applications

A number of applications can be designed with the prediction. We give a simple illustrative example here where we supplement our twenty language set with eleven additional target languages for which no native test data is assumed available: European Spanish (es-ES), Finish (fi-FI), Greek (el-GR), Hindi (hi-IN), Hungarian (hu-HU), Indian English (en-IN), Norwegian (no-NO), Persian (fa-IR), Polish (pl-PL), Romanian (ro-RO), and Turkish (tr-TR). Prediction scores are computed only with the original twenty languages as source languages. The new languages are sorted in an ascending order along with the original twenty languages according to their prediction score in Fig. 2. Therefore, the crosslingual performance for the new languages can be predicted by the range to which the prediction score belongs or by their adjacent neighbors whose PERs are known.

It is observed that of the newly added languages, the crosslingual models for es-ES and ro-RO are expected to have acceptable quality (their prediction scores are in the 0-1 range). The crosslingual models for the new languages closest to the right end of the language axis, on the other hand, such as fa-IR, hi-IN, and en-IN, would be expected to perform poorly in speech recognition tasks.

From a linguistic point of view, the predicted performances of these languages are reasonable. We would expect that the best prediction scores would be associated with target languages which are phonologically similar to several source languages, such as might be the case with related languages or contact languages. es-ES and ro-RO are both Italic languages and are thus related to six of the source languages, i.e. es-LatAm, it-IT, fr-CA, fr-FR, pt-BR, and pt-PT. It is worth noting that there is very little speech data commercially available for ro-RO; ELDA (Evaluations and Language Resources Distribution Agency) currently prices the BABEL Romanian database at 6,000 EUR (non-member, commercial rate). From the same distributor, Castilian Spanish speech databases range from 20,000 EUR (SpeechDat(M)) to 80,000 EUR (SpeechDat-Car). Based on these costs, potential savings in not purchasing ro-RO or es-ES speech databases is minimally 26,000 EUR.

en-IN, hi-IN and fa-IR, on the other hand, are not phonologically similar to any of the twenty source languages. hi-IN is the only Indic language and fa-IR the only Iranian language in the total source language set. While en-IN is classified a Germanic language (like en-US, en-GB, de-DE, etc.), phonologically it is heavily influenced by the speech environment in which it evolved, i.e. the Indic and Dravidian languages of India.

![Figure 2: Crosslingual prediction on 31 languages.](image)

5. Conclusion

We described a novel algorithm for predicting crosslingual speech recognition performance. This prediction tool does not require any speech data from the target language nor does it employ acoustic measurements; rather it is based only on articulatory phonetic and phonological features of the phonemes in the source and target languages. We also presented an example to showcase the practical value of the prediction algorithm.

6. References