# Learning Methods in Multilingual Speech Recognition

**Hui Lin**
Department of Electrical Engineering
University of Washington
Seattle, WA 98125
linhui@u.washington.edu

**Li Deng, Jasha Droppo, Dong Yu, and Alex Acero**
Speech Research Group
Microsoft Research
Redmond, WA 98052
{deng,jdroppo,dongyu,alexac}@microsoft.com

## Abstract

One key issue in developing learning methods for multilingual acoustic modeling in large vocabulary automatic speech recognition (ASR) applications is to maximize the benefit of boosting the acoustic training data from multiple source languages while minimizing the negative effects of data impurity arising from language "mismatch". In this paper, we introduce two learning methods, semi-automatic unit selection and global phonetic decision tree, to address this issue via effective utilization of acoustic data from multiple languages. The semi-automatic unit selection is aimed to combine the merits of both data-driven and knowledge-driven approaches to identifying the basic units in multilingual acoustic modeling. The global decision-tree method allows clustering of cross-center phones and cross-center states in the HMMs, offering the potential to discover a better sharing structure beneath the mixed acoustic dynamics and context mismatch caused by the use of multiple languages' acoustic data. Our preliminary experiment results show that both of these learning methods improve the performance of multilingual speech recognition.

## 1 Introduction

Building language-specific acoustic models for automatic speech recognition (ASR) of a particular language is a reasonably mature technology when a large amount of speech data can be collected and transcribed to train the acoustic models. However when multilingual ASR for many languages is desired, data collection and labeling often become too costly so that alternative solutions are desired. One potential solution is to explore shared acoustic phonetic structures among different languages to build a large set of acoustic models (e.g. [1, 2, 3, 4, 5, 6]) that characterize all the phone units needed in order to cover all the spoken languages being considered. This is sometimes called multilingual ASR, or cross-lingual ASR when no language-specific data are available to build the acoustic models for the target language.

A central issue in multilingual speech recognition is the tradeoff between two opposing factors. On the one hand, use of multiple source languages' acoustic data creates the opportunity of greater context coverage (as well as more environmental recording conditions). On the other hand, the differences between the source and target languages create potential impurity in the training data, giving the possibility of polluting the target language's acoustic model. In addition, different languages may cause mixed acoustic dynamics and context mismatch, hurting the context-dependent models trained using diverse speech data from many language sources.

Thus, one key challenge in the learning multilingual acoustic model is to maximize the benefit of boosting the acoustic data from multiple source languages while minimizing the negative effects of data impurity arising from language "mismatch". Many design issues arise in addressing this challenge, including the choice of "language-universal" speech units, the total size of such units, definition of context-dependent units and their size, decision-tree building strategy, optimal weighting of the individual source languages' data in training, model adaptation strategy, feature normalization strategy, etc. In this paper, we focus on two of these design issues.

The first issue we discuss in this paper is the selection of basic units for multilingual ASR. The main goal of multilingual acoustic modeling is to share the acoustic data across multiple languages to cover as much as possible the contextual variation in all languages being considered. One way to achieve such data sharing is to define a common phonetic alphabet across all languages. This common phone set can be either derived in a data-driven way [7, 8], or obtained from phonetic inventories such as Worldbet [9], or International Phonetic Alpha-bet (IPA) [10]. One obstacle of applying the data-driven approach to large-vocabulary multilingual ASR is that building of lexicons using the automatically selected units is not straightforward, while for the pure phonetic approach, the drawback is that the consistency and distinction among the units across languages defined by linguistic knowledge may not be supported by real acoustic data (as we will demonstrate in Section 2). In this paper, we introduce a semi-automatic unit selection strategy which combines the merits of both data-driven and knowledge-driven approaches. The semi-automatic unit identification method starts from the existing phonetic inventory for multiple languages. This is followed by a data-driven refinement procedure to ensure that the final selected units also reflect acoustic similarity. Our preliminary experiment results show that the semi-automatically selected units outperform the units defined solely by linguistic knowledge.

The second issue we address here is the phonetic decision tree building strategy. As we know, context-dependent models are usually utilized for modern large vocabulary ASR system. One commonly used basic unit in context-dependent models is *triphone*, which consists of a center phone along with its left-neighbor and right-neighbor phones. Typically, around 30 to 40 phonemes are required in order to describe a single language. In a monolingual ASR system, a complete triphone-based acoustic model would contain a total of over 60 thousand triphone models with more than 180 thousand hidden Markov model (HMM) states if each triphone is modeled with a 3-state left-to-right HMM. It is generally impossible to train such large acoustic models with supervised learning methods since a huge amount of labeled acoustic data are required, which is not available at present. To address this issue, phonetic decision tree clustering [11] was introduced and is still widely used today. Usually, the decision trees are limited to operate independently on each context independent state of the acoustic model. In other words, no cross center phone sharing is allowed. This design feature is based on the assumption that there is no benefit from clustering different phones together. Such a restriction may be reasonable for a monolingual ASR system, but it may not be suitable for multilingual acoustic modeling since the acoustic properties across multiple languages is less predicable. In this paper, we use a global decision tree, which better describes acoustics of the training data without artificially partitioning the acoustic space. Improvements of using global decision trees are illustrated in our preliminary experimental results.

## 2 Semi-automatic Unit Selection

### 2.1 The Technique

The steps of the semi-automatic unit selection procedure developed in this work are described below:

- We start with a common phonetic inventory, say $\mathcal{I} = \{p_1, p_2, ..., p_n\}$, defined for multiple languages. There are $n$ phonemes defined in this inventory. For convenience, we denote the index set as $\mathcal{N}$, i.e. $\mathcal{N} = \{1, 2, ..., n\}$.
- A separate phonetic inventory $\mathcal{I}_l = \{p_{k,l} \mid k \in \mathcal{S}, \mathcal{S} \subseteq \mathcal{N}\}$ is formed for each langauge $l$. $\mathcal{I}_l$ contains *all* the phones used for language $l$. Language tag is attached to the phone symbol to denote that it belongs to language $l$.

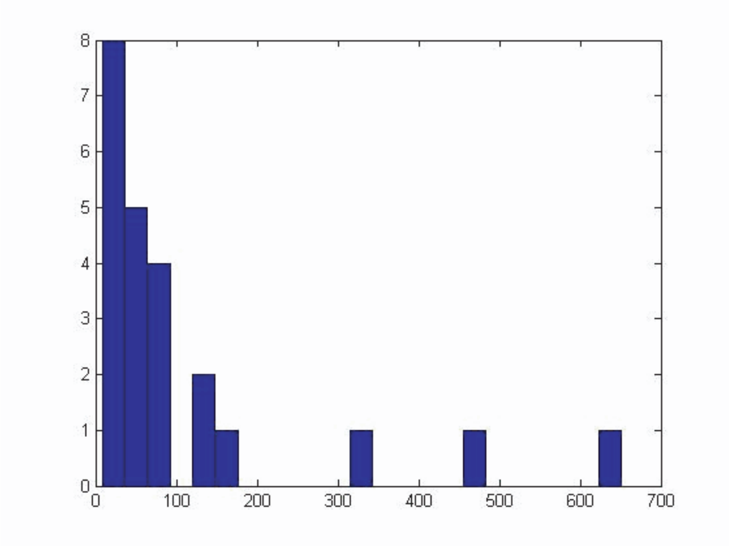

Figure 1: Histogram of KL distances between phones sharing the same symbol UPS (based on IPA)). The numbers on x axis represents the value of KL distances.

- Using the transcribed data, train a HMM $H_{k,l}$ for each monophone $p_{k,l}$ for each language.
- All phones in all languages are clustered and the phones in the same cluster are shared with acoustic data during multilingual training. Specifically, K-mean clustering is performed to all the phones in all languages, where the distance between phones are defined as the Kullback-Leibler (KL) distance between HMMs; I.e. $d(p_{k,l_1}, p_{k,l_2}) = d_{KL}(H_{k,l_1}, H_{k,l_2})$. A new symbol is used to represent all the phones in the same cluster, and these new symbols form our final phonetic inventory $\mathcal{I}_{new}$ across all languages. Mappings from $\mathcal{I}_l$ to $\mathcal{I}_{new}$ are recorded accordingly.
- Obtain a new lexicon for each language $l$ using the mapping from $\mathcal{I}_l$ to $\mathcal{I}_{new}$.

The design in the second step above with the use of language tags is intended to prevent any data sharing across languages since at this intial stage we assume there are no common phones defined among different languages. For example, phoneme $p_{k,l_1}$ and phoneme $p_{k,l_2}$ are treated as two distinct phones, one for language $l_1$ and the other for language $l_2$, but they have both originated from the same phone $p_k$ in the common phonetic inventory $\mathcal{I}$. If we were to fully trust the common phonetic inventory $\mathcal{I}$, $p_{k,l_1}$ and $p_{k,l_2}$ would be identical, and thus the acoustic data for $p_{k,l_1}$ from langauge $l_1$ and data for $p_{k,l_2}$ from language $l_2$ would be shared to represent a common unit $p_k$. Unfortunately, the common phone inventory in our investigation has been found not to accurately reflect the real acoustic similarities. This is illustrated in Figure 1), where a histogram is plotted of the KL distances between the Italian and Spanish phones that have the same symbol in Universal Phone Symbol (UPS) – $d_{KL}(H_{k,italian}, H_{k,spanish}), k \in \mathcal{N}$. The numbers on x axis represents the value of the KL distance. Apparently, at least three symbols result in very different acoustic distributions, indicating that the UPS set could not accurately reflect acoustic similarities across languages. Detailed investigation motivated us to distinguish phones for different languages at this step and to leave the decision of sharing data or otherwise to the clustering algorithm based on the data themselves.

### 2.2 Experiments

In our experiments, we use the universal phone set (UPS), which is a machine-readable phone set based on the IPA, to represent the language universal speech units. In most cases, there is a one-to-one mapping between UPS and IPA symbols, while in a few other cases UPS is a superset of IPA. For example, UPS includes some unique phone labels for commonly used sounds such as diphthongs, and nasalized vowels, while IPA treats them as compounds. Generally, UPS covers sounds in various genres, including consonants, vowels, suprasegmentals, diacritics, and tones. Table 1 illustrates the number of different types of UPS units for the two languages (Italian and Spanish) used in this experiment.

Table 1: Number of vowel , consonant, suprasegmentals, and diacritics units for the 2 languages used in this experiment

| | **vowel** | **consonant** | **suprasegmentals** | **diacritics** |
|---|---|---|---|---|
| **Italian** | 11 | 24 | 1 | 0 |
| **Spanish** | 6 | 21 | 0 | 0 |

To cover these two languages, we only need 44 units (including four other symbols used for silence and noise). That is, $|\mathcal{I}| = 44$ in our case. Monophone HMMs with single Gaussian per state were trained separately for these two languages. The KL distance between phones which share the same UPS symbol, $d_{KL}(H_{k,italian}, H_{k,spanish}), k \in \mathcal{N}$, were calculated and the histogram is plotted in Figure 1. To gain insight into what value of the distance actually indicates "dis-similarity", the distances between different phones within the same language are also estimated. For Spanish, the estimated average distance is 213; for Italian, it is 335. These values are smaller than some values shown in Figure 1 for the same symbol across the two languages, which indicates that the use of UPS as is would necessarily introduce "language mismatch".

After adding language tag as introduced in Section 2.1, we have $|\mathcal{I}_{italian}| = 40$ and $|\mathcal{I}_{spanish}| = 31$. This gives a total of 71 monophone units for the two languages. These 71 units are further clustered resulting a final phone set with 47 units ($|\mathcal{I}_{new}| = 47$).

Table 2: Training set descriptions

| **Language** | **Corpus** | **#. Speaker** | **Hours** |
|---|---|---|---|
| **Italian** | ELRA-S0052 | 989 | 23.5 |
| **Spanish** | ELRA-S0065 | 992 | 33.9 |

Some statistics of the data used for training are shown in Table 2. The training procedure used in this experiment is described below. 13 MFCCs were extracted along with their first and second time derivatives, giving a feature vector of 39 dimensions. Cepstral mean normalization was used for feature normalization. All the models mentioned in this paper are cross-word triphone models. Phonetic decision tree tying was utilized to cluster triphones. A set of linguistically motivated questions were derived from the phonetic features defined in the UPS set. The number of tied states, namely *senones*, can be specified at the decision tree building stage to control the size of the model. The top-down tree building procedure is repeated until the increase in the log-likelihood falls below a preset threshold. The number of mixtures per senone is increased to four along with several EM iterations. This leads to an initialized cross-word triphone model. The transcriptions are then re-labeled using the initialized cross-word triphone models, which were used to run the training procedure once again - to reduce number of mixture components to one, untie states, re-cluster states and increase the number of mixture Gaussian components. The final cross-word triphone is modeled with 12 Gaussian components per senone.

Table 3: Test set descriptions

| **ID** | **Corpus** | **#. Utterances** | **#. Speaker** | **Envronments** |
|---|---|---|---|---|
| **Test I** | ELRA-S0052 | 2140 | 99 | Office/home |
| **Test II** | ELRA-S0116 | 2199 | 400 | Office/home/street/public place/vehicle |
| **Test III** | PHIL18 | 4827 | 197 | Quite environment |
| **Test IV** | PHIL42 | 3916 | 129 | Office/home/Quite environment |

In testing, we are interested in telephony ASR under various environments, including home, office, and public places. We chose Italian as our target language, which is observed during the language-universal training. Several test sets were used as shown in Table 3. In all of our experiments, the standard Microsoft speech recognition engine was used for acoustic modeling and decoding.

Table 4 shows the word error rates (WER) results of different methods on the four test sets. The row with “Monolingual training” refers to the procedure where we only used the data for Italian to train the acoustic models. For multilingual training, data for both Italian and Spanish were used. We had 3000 senones based on the amount of data (about 20 hours) for monolingual training, while for the multilingual training, we had 5000 senones since more training data (about 50 hours) were used in the training. For fair comparisons, we also increased the number of senones for the monolingual model, which was stopped at around 4600 when the problem of data insufficiency was detected. It can be seen that multilingual acoustic modeling outperforms monolingual training on Test sets II, III and IV. Semi-automatic unit selection described in this paper is shown to be effective with significant improvements on Test II and III compared with using UPS.

Table 4: WER (%) results on the four test sets for Italian

| Method | #. Phones | #. Senones | Test I | Test II | Test III | Test IV |
|---|---|---|---|---|---|---|
| **Monolingual training** | 40 | 3000 | 3.62 | 5.57 | 6.34 | 17.90 |
| **Monolingual training** | 40 | 4600 | 3.82 | 5.68 | 7.67 | 19.61 |
| **Multilingual training (UPS)** | 44 | 5000 | 4.05 | 5.21 | 5.13 | 17.78 |
| **Multilingual training (semi-auto)** | 47 | 5000 | 4.04 | 4.99 | 5.01 | 17.82 |

## 3 Global Phonetic Decision Tree

### 3.1 The Technique

The standard way of clustering triphone HMM states is to use a set of phonetic decision trees. One tree is built for every state of every center phone. The trees are built using a top-down sequential optimization process. Initially, each of the trees starts with all possible phonetic contexts represented in a root node. Then a binary question is chosen which gives the best splits the states represented by the node. Whichever question creates two new senones that maximally increase the log likelihood of the training data is chosen. This process is applied recursively until the log likelihood increase is less than a threshold.

Instead of using a different decision tree for every context independent phone state, we use a single global phonetic decision tree that starts with all states in the root node. The question sets explored during the clustering includes questions about the current state, about the current center phone, and about the current left and right context phone classes. In contrast, the conventional decision tree building would only use the context questions. Other than that, the global decision tree building procedure is the same as the standard procedure. Using a global decision allows cross-center phone and cross-center state clustering. We believe that such joint clustering could discover a better sharing structure beneath the mixed acoustic dynamics and context mismatch caused by multiple languages.

### 3.2 Experiments

The experimental setup was the same as introduced in Section 2.2. Instead of using traditional decision tree building procedure, we built a single global decision tree during the multilingual training. The new model was compared to that produced without global decision tree optimization. As shown in Table 5, using global decision tree has positive effects consistently on all of the four test sets, supporting our claim that global phonetic decision explores state tying structure that better describes the training data, and thus is a better option for multilingual ASR. Note the global decision tree method experimeted here was not on the semi-automatic selected units. We will explore combining the two learning methods in our future work and further performance improvements are expected.

Table 5: WER (%) results on the four test sets for Italian

| Method | #. Phones | #. Senones | Test I | Test II | Test III | Test IV |
|---|---|---|---|---|---|---|
| **Monolingual training** | 40 | 3000 | 3.62 | 5.57 | 6.34 | 17.90 |
| **Monolingual training** | 40 | 4600 | 3.82 | 5.68 | 7.67 | 19.61 |
| **Multilingual training** | 44 | 5000 | 4.05 | 5.21 | 5.13 | 17.78 |
| **Multilingual training (global DT)** | 44 | 5000 | 3.99 | 5.01 | 4.91 | 17.24 |

## 4 Sumamry and Conclusions

In this paper, we reported our development and experimental results for two learning methods in multilingual speech recognition. The key issue that the learning methods are addressing is how to balance between boosting acoustic training from multiple languages and reduing acoustic data impurity arising from language mismatch. Both learning methods, one on the use of new cross-lingual speech units and another on the use of a global decision tree, are shown to produce superior speech recognition performance over the respective baseline systems. There is vast opportunity to develop new learning methods in the space of multilingual speech recognition.

## References

[1] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, 2001.

[2] T. Schultz and A. Waibel, "Language Independent and language adaptive large vocabulary speech recognition," *Proc. ICSLP*, 1998.

[3] W. Byrne et al., "Towards language independent acoustic modeling," *Proc. ICASSP*, 2000.

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

[5] Li Deng, "Integrated-multilingual speech recognition using universal phonological features in a functional speech production model," *Proc. ICASSP*, 1997.

[6] E. Garcia, E. Mengusoglu, and E. Janke, "Multilingual acoustic models for speech recognition in low-resource devices," *Proc. ICASSP*, 2007.

[7] O. Anderson, P. Dalsgaard, and W. Barry, "On the use of data-driven clustering technique for identification of poly- and mono-phonemes for four european languages," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1, pp. 121–124, 1994.

[8] J. Köhler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," *Proc. ICSLP*, 1996.

[9] James L. Hieronymus, "Ascii phonetic symbols for the world's languages: Worldbet," *AT&T Bell Laboratories, Technical Memo*, vol. 23, 1994.

[10] International Phonetic Association, "Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet," pp. 1–204, 1999.

[11] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," *Proceedings of the workshop on Human Language Technology*, 1994.