ACOUSTIC-PHONETIC UNIT SIMILARITIES FOR CONTEXT DEPENDENT ACOUSTIC MODEL PORTABILITY

Viet Bac Le, Laurent Besacier*, Tanja Schultz***

* CLIPS-IMAG Laboratory, UMR CNRS 5524
BP 53, 38041 Grenoble Cedex 9, FRANCE

** Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh, PA, USA

email: viet-bac.le@imag.fr

ABSTRACT

This paper addresses particularly the use of acoustic-phonetic unit similarities for portability of context dependent acoustic models to new languages. Since the IPA-based method is limited to a source/target phoneme mapping table construction, an estimation method of the similarity between two phonemes is proposed in this paper. Based on these phoneme similarities, some estimation methods for polyphone similarity and clustered polyphonic model similarity are investigated. For a new language, first a polyphonic decision tree is built with a small amount of speech data. Then, clustered models in the target language are duplicated from the nearest clustered models in the source language and adapted with limited data to the target language. Results obtained from the experiments demonstrate the feasibility of these methods.

1. INTRODUCTION

Nowadays, computers are heavily used to communicate via text and speech. Text processing tools, electronic dictionaries, and even more advanced systems like text-to-speech or dictation are readily available for several languages. However, the implementation of Human Language Technologies (HLT) requires significant resources, which have only been accumulated for a very small number of the 6900 languages in the world. Among HLT, we are particularly interested in Automatic Speech Recognition (ASR). Therefore, we are interested in new techniques and tools for rapid portability of speech recognition systems when only limited resources are available. Resource sparse languages are typically spoken in developing countries, but can nevertheless have many speakers. In this paper, we investigate Vietnamese, which is spoken by about 70 million people, but for which only very few usable electronic resources are available.

In crosslingual acoustic modeling, previous approaches have been limited to context independent models [1, 2, 3]. Monophonic acoustic models in target language were initialized using seed models from source language. Then, these initial models could be rebuilt or adapted using training data from the target language.

Since the recognition performance is increased significantly in wider contexts, the crosslingual context dependent acoustic modeling portability and adaptation can be investigated. J. Köhler [4] used HMM distances to calculate the similarity between two monophonic models. This method can be extended to context dependent models. A triphone similarity estimation method based on phoneme distances was first proposed by B. Imperl [5] and used an agglomerative clustering process to define a multilingual set of triphones. One problem in portability of context dependent acoustic models is that the context mismatch across languages increases dramatically for wider contexts. T. Schultz [6] proposed PDTS (Polyphone Decision Tree Specialization) to overcome this problem. In PDTS, the clustered multilingual polyphone decision tree is adapted to the target language by restarting the decision tree growing process according to the limited adaptation data in the target language. While PDTS is purely data-driven method, the intention of this paper is to explore a knowledge-based approach.

In this work, we investigate a new method for this crosslingual transfer process. We do not use the existing decision tree in source language but build a new decision tree just with a small amount of data from the target language. Then, based on the acoustic-phonetic unit similarities, some crosslingual transfer and adaptation processes are applied.

In this paper, we start in section 2 by proposing different acoustic-phonetic unit similarities estimation methods. In section 3 these similarities are applied to port context independent and dependent acoustic models across languages. The experimental framework and results are presented in section 4. Section 5 concludes the work and gives some future perspectives.

2. ACOUSTIC-PHONETIC UNIT SIMILARITIES

The research in crosslingual acoustic modeling is based on the assumption that the articulatory representations of phonemes are so similar across languages that phonemes can be considered as units which are independent from the underlying language [6]. Based on this assumption, we proposed in this section some methods for estimating the similarities of some phonetic units (phoneme, polyphone, clustered polyphone) which will be further used in crosslingual context dependent acoustic modeling.

2.1. Phoneme Similarity

In our work, both *data-driven* and *knowledge-based* methods are applied and proposed to automatically or manually obtain the phoneme similarities across languages.

2.1.1 Data-driven methods

The acoustic similarity between two phonemes can be obtained automatically by calculating the distance between two acoustic models (HMM distance [4], Kullback-Leibler distance, Bhattacharyya distance, Euclidean distance [7] or by calculating a confusion matrix [1, 2]). A confusion matrix is calculated by applying a source language phoneme recognizer to a small amount of target language acoustic data, which was already phonetized with the target language acoustic units. Note that in the basic phoneme recognizer we use, all phonemes have the same probability to appear. Then, each entry of the confusion matrix is

normalized by dividing it through the number of occurrences of all corresponding phonemes in the source language [3].

Normally, the confusion matrix represents the likelihood of the confusion between two phonemes. Thus, we can use these phoneme confusions to evaluate phoneme similarities. Let M, N be numbers of phonemes in source and target language. Let A(M,N) be the confusion matrix. The similarity $d(s_i, t_j)$ between phoneme t_j in the target language and phoneme s_i in the source language is calculated as:

$$d(s_i, t_j) = A_{i,j} \quad (1)$$

where $A_{i,j} \in [0,1]$, i=1..M, j=1..N.

2.1.2. Proposed knowledge-based method

Traditionally, knowledge-based methods had been applied to find the phoneme of the source language that best matches a phoneme in the target language [1, 6]. However, no knowledge-based method is known that allows to calculate the similarity between two phonemes. Thus, in this section, we propose a new knowledge-based method to calculate the phoneme similarity. As we know, similarities of sounds are documented in international phonetic inventories like the International Phonetic Alphabet (IPA)[1] which classifies sounds based on phonetic knowledge.

Based on the IPA phoneme classification we propose a *bottom-up algorithm* to determine a distance-based similarity between two phonemes. This algorithm consists of two steps: *top-down classification using a hierarchical graph* and *bottom-up phoneme distance estimation.*

a) Step 1: Top-down classification

Figure 1 shows a hierarchical graph where each node is classified into different layers. To each node we manually assigned a group of phonemes following the IPA phoneme classification scheme. Each group of phoneme has a user-defined similarity value assigned that represents the similarity of the elements within this group. All nodes corresponding to the same layer obtain the same similarity value. Let k be the number of layers and G_i be the user-defined similarity value for layer i (i = 0...k-1). In our work, we investigated several settings of k and G_i and set G = {0.9; 0.45; 0.25; 0.1; 0.0} with k = 5 based on a cross-evaluation in crosslingual acoustic modeling experiments.

To grow this graph, we start with the group PHONEME, which contains all the phonemes, at layer 0 and divide it into a CONSONANT group and VOWEL group at layer 1. This top-down classification is applied with increasingly specified grouping criteria until each group contains only one phoneme.

b) Step 2: Bottom-up estimation

To estimate the distance between two phoneme s and t, we locate them in the leaves of the graph and then trace back from their respective leaves until the nearest common parent node is reached. The similarity between s and t is thus given by the similarity value of layer i, which contains this parent node, we have:

$$d(s, t) = G_i \quad (2)$$

For example, the parent node of vowel [i] and [u] is CLOSE, we have:

$$d([i], [u]) = G_2 \quad (= 0.25 \text{ in our experiment }).$$

[1] http://www2.arts.gla.ac.uk/IPA/ipa.html

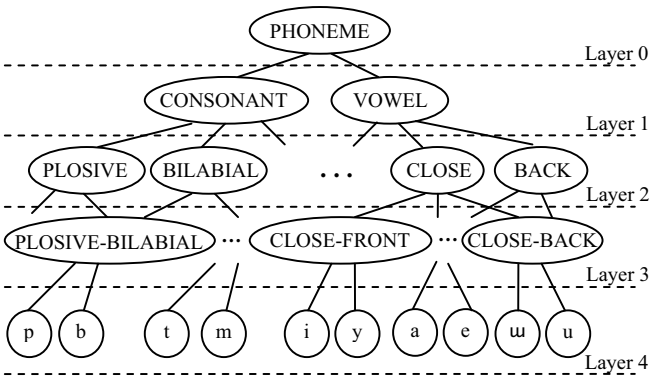

Figure 1 : Hierarchical graph for phoneme similarity

2.2. Polyphone Similarity

Let L be the left and the right context length of a polyphone. We assume that the context length of polyphones in source and target language are the same. If not, a context normalization procedure is needed. Let S be the phoneme set in source language, T be the phoneme set in target language.

Let $P_S = (s_{-L}, s_{-L+1},\ldots, s_{-1}, \mathbf{s_0}, s_1,\ldots, s_L)$ and $P_T = (t_{-L}, t_{-L+1},\ldots, t_{-1}, \mathbf{t_0}, t_1,\ldots, t_L)$ be polyphones in source and target language, where $s_{-L},\ldots, s_{-1}, s_0, s_1,\ldots, s_L \in S$ and $t_{-L},\ldots, t_{-1}, t_0, t_1,\ldots, t_L \in T$ denote the central phoneme, left phonemes or right phonemes of P_S and P_T.

The distance-based similarity of P_S and P_T is calculated as a weighted sum of distance between corresponding source/target phonemes along their context:

$$d(P_S, P_T) = \propto_0 .d(s_0, t_0) + \propto_1 .[d(s_{-1}, t_{-1}) + d(s_1, t_1)] + \ldots + \propto_L .[d(s_{-L}, t_{-L}) + d(s_L, t_L)] \quad (3)$$

where $\propto_0, \propto_1,\ldots \propto_L$ are contextual weight coefficients which represent the influence of contextual phoneme to the central phoneme; $d(s_k, t_k)$ is the phoneme distance (k = -L,…L). In the same way, the triphone similarities are calculated in [5].

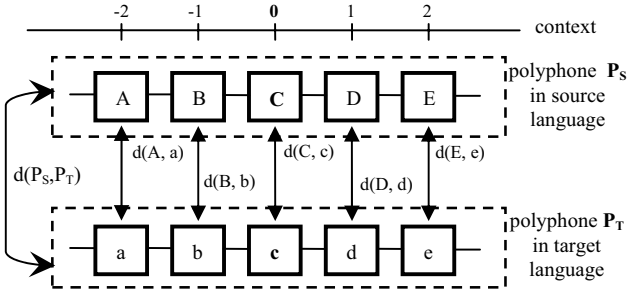

Figure 2 : Distance-based polyphone similarity

Figure 2 shows an example of the similarity between polyphone P_S = (A B **C** D E) and P_T = (a b **c** d e) in the source and target language.

For each polyphone of the target language, the nearest polyphone P_S* in source language is obtained that satisfies the following relation:

$$\forall P_S \in S,\ d(P_S^*, P_T) = \min[\ d(P_S, P_T)\ ] \quad (4)$$

2.3. Clustered Polyphonic Model Similarity

Since the number of polyphones in a language is very large (e.g.,

over 100,000 triphones for English), a limited training corpus usually does not cover enough occurrences of every polyphones. As a consequence many polyphones in the test set have never been seen in training. Thus, we need to find models that are accurate and trainable in acoustic modeling. A decision tree-based clustering (figure 3) or an agglomerative clustering [5] procedure is needed to cluster and model similar polyphones in a clustered polyphonic model.

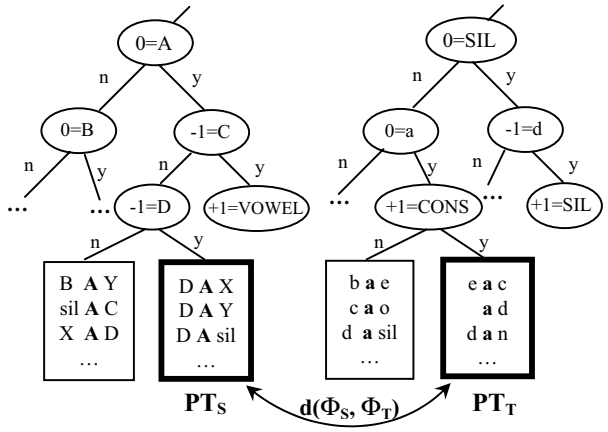

Figure 3 : Clustered polyphone similarity across languages

Therefore, for crosslingual context dependent modeling, a clustered polyphonic model similarity evaluation method must be proposed to find two nearest clustered polyphonic models across languages (figure 3).

Let $\Phi_S = (P_{S1}, P_{S2}\ldots P_{Sm})$ be a clustered polyphonic model of m polyphones in the source language and $\Phi_T = (P_{T1}, P_{T2},\ldots, P_{Tn})$ be a clustered polyphonic model of n polyphones in the target language, the similarity between Φ_S and Φ_T is the average of all distances between any two polyphones in Φ_S and Φ_T. We have:

$$d(\Phi_S, \Phi_T) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d(P_{Si}, P_{Tj})}{m.n} \quad (5)$$

For each clustered polyphone set in the target language, the nearest clustered polyphone set P_S^* in source language is obtained if it satisfies the following relation:

$$\forall\Phi_S,\ d(\Phi_S^*, \Phi_T) = \min\ [d(\Phi_S, \Phi_T)] \quad (6)$$

3. CROSSLINGUAL ACOUSTIC MODELING

3.1. Context Independent Acoustic Model Portability

For context independent acoustic modeling, the phonetic unit is the monophone and a distance between monophone models is calculated. Φ_S and Φ_T are calculated using the distance between two phonemes. Equation (5) leads to:

$$d(\Phi_S, \Phi_T)\ ] = d(P_S, P_T) = d(s, t) \quad (7)$$

where d(s, t) is calculated by equation (1) or (2).

Equation (6) leads to:

$$\forall\Phi_S,\ d(\Phi_S^*, \Phi_T) = \min\ [d(\Phi_S, \Phi_T)\ ] = \min[\ d(s, t)\ ] \quad (8)$$

By applying equation (8), a *phoneme mapping table* between source and language can be obtained. Based on this mapping table, the acoustic models in the target language can be borrowed from the source language and adapted by a small amount of target language speech data (see [3] for more details).

3.2. Context Dependent Acoustic Model Portability

In this section, a context dependent acoustic model portability method is proposed based on the phonetic similarities described in the previous section.

Firstly, by using a small amount of speech data in the target language, a decision tree for polyphone clustering (PT_T) can be built. We suppose that such a decision tree (PS_S) is also available in the source language (figure 3).

Secondly, by applying the equation (5), we can evaluate the distance between any two source/target clustered polyphonic models. That allows us, by applying the equation (6), to determine for each model in target language, the most similar model in the source language. This model is then copied into the acoustic model in the target language.

Finally, while acoustic models borrowed directly from the source language did not perform very well, an adaptation procedure (Viterbi training, MLLR, MAP) can successfully be applied with a small amount of speech data in the target language (see also [6]).

4. EXPERIMENTS AND RESULTS

This section presents our experiments in portability of context dependent acoustic models to new language using acoustic-phonetic unit similarities. Experiments and results in crosslingual context independent modeling were already presented in [3].

4.1. Experimental framework

4.1.1. ASR system

All recognition experiments use the JANUS toolkit [8] developed at the ISL Laboratories. The model topology is a 3- state left-to-right HMM with 48 Gaussian mixtures per state. The pre-processing of the system consists of extracting a 43 dimensional feature vector every 16 ms. The features consist of 13 MFCCs, energy, the first and second derivatives, and zero-crossing rate. An LDA transformation is used to reduce the feature vector dimensionality to 32.

Since Vietnamese language is a monosyllabic and tonal language with 6 tones (figure 4), we used syllables rather than words as recognition units (*syllable-based ASR system*). Furthermore, in the described experiments, the Vietnamese phones are modeled without tone indication. Since tone is a discriminative feature in Vietnamese, decisions between two different words with the same phone sequence but two different tones, are made by the language model.

4.1.2. Vietnamese Text and Speech Resources

<table>
<tr><td colspan="5">Tonal syllables (6,492)</td></tr>
<tr><td colspan="4">Base syllables (2,376)</td><td rowspan="3">Tone (6)</td></tr>
<tr><td rowspan="2">INITIAL(22)</td><td colspan="3">FINAL (155)</td></tr>
<tr><td>Medial(1)</td><td>Nucleus(16)</td><td>Ending(8)</td></tr>
</table>

Figure 4 : The phonological hierarchy of Vietnamese syllables

Firstly, since there are 6,492 syllables in the Vietnamese language (figure 4), a vocabulary of 6,492 syllables was extracted from a 40,000 word vocabulary. Then a pronunciation dictionary

for Vietnamese was built by applying our *VNPhoneAnalyzer* [9] on this syllable vocabulary.

Secondly, documents were gathered from Internet and filtered for building a text corpus. After data preparation, the text corpus has a size of 868 MB. A syllable-based statistical trigram language model was trained from this text corpus by using the SRILM toolkit [10] with a Good-Turing discounting and Katz backoff for smoothing. It is very important to note that with this toolkit, the unknown words are removed in our case, since we are in the framework of closed-vocabulary models. The perplexity value evaluated on our test corpus is *108.5*.

Finally, speech data was extracted from the *VNSpeechCorpus* [9], which was built at CLIPS-IMAG and MICA laboratories. In order to build a polyphonic decision tree and to adapt the crosslingual acoustic models, 2.25 hours of data spoken by 8 speakers were used. The test set contains 400 utterances spoken by 3 speakers different from the training speakers.

4.2. Experimental Results

4.2.1. Baseline System

By using 2.25 hours of Vietnamese speech data, decision trees for 500, 1000 and 2000 sub-triphone models were built respectively by a clustering procedure. These models are trained using LDA calculation, codebooks initialization (kmeans) and 6 iterations of Viterbi training.

4.2.2. Comparative Experiments

For crosslingual experiments, we use multilingual context dependent models (MM6-Mix with 12,000 sub-quinphone models) developed by ISL Laboratories [6]. Speech data from six languages were used to build these models: Arabic, Chinese, English, German, Japanese and Spanish. After the crosslingual transfer procedure, initial sub-models were adapted with 2.25 hours of Vietnamese speech data by 6 iterations of Viterbi training.

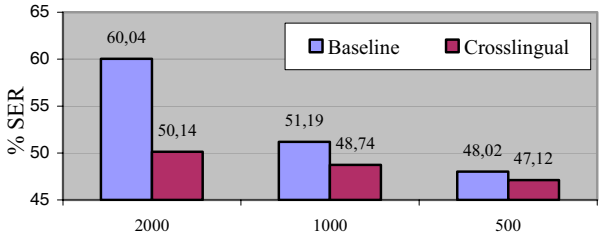

Figure 5 : Performance (syllable error rate) of baseline system and crosslingual method with different numbers of sub-triphone models

Figure 5 shows the Syllable Error Rate (SER) of the baseline system and the proposed crosslingual system. The crosslingual system improves 1.87%, 4.79% and 16.49% of absolute SER for 500, 1000, and 2000 sub-triphone models respectively. As the number of clustered sub-models increases, SER of the baseline system increases proportionally since the amount of data per model decreases due to the limited training data. However, the crosslingual system is able to overcome this problem by indirectly using data in other languages.

Figure 6 presents the influence of adaptation data size and number of speakers on the baseline system and two methods of phoneme similarity estimation: *proposed knowledge-based* and *data-driven using confusion matrix*. We find that the knowledge-based method outperforms the data-driven method.

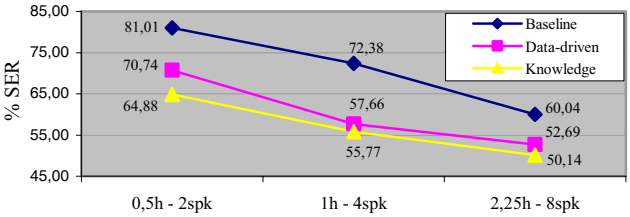

Figure 6 : Performance of phoneme similarity estimation methods with different amount of adaptation data (2000 sub-triphone models used): a) baseline system b) data-driven c) proposed knowledge-based

5. CONCLUSIONS AND PERSPECTIVES

This paper presents different methods of estimating the similarities between two acoustic-phonetic units. Based on these similarities, some crosslingual context independent and dependent acoustic modeling methods are proposed in our work. By using 2.25 hours of Vietnamese adaptation data, results from the obtained baseline system are outperformed by the proposed system (up to 16.49% of absolute SER). We note that, by using the vocabulary of 6,492 syllables, our syllable-based system almost covers all of the possible words in Vietnamese language (LVCSR). The potential of our method is demonstrated even though the use of trigrams the in syllable-based language modeling might be insufficient to obtain acceptable error rates (best SER is 47.12% obtained with 2.25h Vietnamese data only).

In the future, we will investigate word-based ASR systems to obtain the most likely recognition unit in Vietnamese language. We also plan to try different size of polyphone context and different contextual weight coefficients in order to obtain the suitable crosslingual acoustic models.

6. REFERENCES

[1] P. Beyerlein et al., "Towards language independent acoustic modeling", *ASRU'99*, Keystone, CO, USA, December 1999.

[2] R. Bayeh et al., "Towards multilingual speech recognition using data driven source/target acoustical units association", *ICASSP'04*, vol. I, pp. 521-524, Montreal, Canada, May 2004.

[3] V. B. Le, L. Besacier, "First steps in fast acoustic modeling for a new target language: application to Vietnamese", *ICASSP'05*, vol. 1, pp. 821-824, Philadelphia, PA, USA, March 2005.

[4] J. Köhler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds", *ICSLP'96*, pp. 2195-2198, Philadelphia, PA, USA, October 1996.

[5] B. Imperl et al., "Agglomerative vs. Tree-based clustering for the definition of multilingual set of triphones", *ICASSP'00*, vol. 3, pp. 1273-1276, Istanbul, Turkey, June 2000.

[6] T. Schultz, A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition", *Speech Communication*, vol. 35, no. 1-2, pp. 31-51, August 2001.

[7] J. J. Sooful, E. C. Botha, "An acoustic distance measure for automatic cross-language phoneme mapping", *PRASA'01*, pp. 99-102, South Africa, November 2001.

[8] M. Finke et al., "The Karlsruhe-Verbmobil Speech Recognition Engine", *ICASSP'97*, vol. 1, pp. 83-86, Munich, Germany, 1997.

[9] V. B. Le et al, "Spoken and written language resources for Vietnamese", *LREC'04*, pp. 509-602, Lisbon, Portugal, May 2004.

[10] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit", *ICSLP'02*, vol. 2, pp. 901-904, Denver, CO, USA, September 2002.