TOWARDS MULTILINGUAL SPEECH RECOGNITION USING DATA DRIVEN SOURCE/TARGET ACOUSTICAL UNITS ASSOCIATION

R. Bayeh[1], S. Lin[2], G. Chollet[2], C. Mokbel[1]

[1]University of Balamand, PO Box 100, Tripoli, Lebanon
[2]ENST, CNRS URA820, 46 rue Barrault, 75634 Paris cedex 13, France

ABSTRACT

Multilingual speech recognition pushes to study the acoustic modeling of target language units using one or more source languages' units. This paper presents a study of manual and data driven association of two possible target units with source language's phonemes. The target units studied are words and phonemes. Algorithms for data-driven association are described. While phoneme-to-phoneme association is more practical, words' transcription provides better results. It has been shown that more precise and rich source models are more suitable to determine those association. Experiments are conducted with French as source language and Arabic as target language.

1. INTRODUCTION

Research in multilingual speech recognition has gained increasing interest in the last years [1]. The motivation for such research is twofold, theoretical and practical. At the theoretical level, developing multilingual speech recognition systems pushes towards a factorized and reduced set of acoustic units models which need advanced techniques in acoustic modeling. At the practical level, numerous applications would take advantage of such multilingual systems. As a first step in this direction the present paper describes a study on defining an automatic speech recognition system in a target language based on acoustic models from one or multiple source languages. An important advantage of such approach is to build speech recognition systems for languages or dialects where only small databases exist to train precise systems. The present study also investigates at which level should acoustic units' correspondence between source and target languages be established.

Linguistically speaking, the Arabic language, which is our Target language, does not have a unique phonetic representation despite the presence of several Arabic speaking nations. This is mainly due to the inconsistency of dialects and words used in different regions. All countries share the same written Arabic language however the spoken languages are not common. Therefore, a minimal database based on one dialect would be insufficient for creating and training a full acoustic model. Therefore, we have attempted to design a recognizer for Lebanese dialect relying on a different language's phone models. In the experiments presented in this paper, the target language is the Lebanese dialect of Arabic and the source language is the French language.

As for the acoustic units, different levels may be chosen for every language, phones, triphones, syllables or words. In the present study, phonemes models are considered for the source language. For the target language, we investigated two possible units, the words and the phonemes. In other words, two cases are studied and compared; a direct transcription of Arabic words using French phones and a correspondence between source and target phones.

The correspondence between source and target acoustic units can be determined either manually or automatically. For data driven correspondences, algorithms similar to those used earlier for the determination of variants of pronunciations in lexical modeling [1][4] are used. In this case several training utterances are aligned on a loop of phonemes allowing based on several criteria to determine an optimal data driven transcription of words or a correspondence between phonemes. The theoretical foundation of these inference techniques is given in the next section. Section III presents the databases used in our experiments. Section IV presents the experiments and results obtained when using words as acoustic units for the Target language. Section V presents the experiments and results when phones are used as Target acoustic units. In this case phones correspondence tables are built manually or determined automatically from acoustical data. The relevance of these tables is further investigated in the data-driven approach by using two sets of phonemes models for the source language, the first set used to estimate the correspondence and the second set is used for recognition. Finally, section VI provides conclusions and perspectives.

2. DATA-DRIVEN ACOUSTIC UNITS CORRESPONDENCE

As stated in the introduction, the theoretical foundation of the data-driven correspondence between source and target acoustic units is similar to the one used for automatic determination of variants of pronunciation in lexical

modeling as in [4]. Assume that N utterances $(U_1, \ldots, U_N)$ from the Target language are available for training and that the acoustical models for the Source language are $(\lambda_1,\ldots, \lambda_p)$. We define $(\mu_1, \ldots, \mu_q)$ the q acoustical models for the Target language and we suppose that a target model can be expressed as a function of the source acoustic models:

$$\mu_i = f_i(\lambda_1,\ldots,\lambda_p) \quad i = 1,\ldots q \qquad \text{(Eq. 1)}$$

Given that, in the target language, every utterance is expressed as a succession of target acoustical units, the general optimization problem is to find $\{\hat{f}_i\}$ such as:

$$\{\hat{f}_i\} = \arg\max_{\{f_i\}} p(U_1,\ldots U_N / \mu_1,\ldots,\mu_q) \qquad \text{(Eq. 2)}$$

The optimization to be performed in (Eq. 2) depends on the nature of the source and target acoustical units and on the type of functions $\{f_i\}$. In the present work the source units are phonemes and two cases are considered for the target units, i.e. words and phonemes.

If the word is considered to be the basis for target acoustic unit, the nature of $\{f_i\}$ will determine the optimization algorithm. If f_i is the transcription of i^{th} target word using the source phonemes, we should find the phoneme sequence leading to the highest likelihood for the training sequences of the i^{th} word. As in[4], an N-best alignment of the i^{th} word training utterances on a loop of source phonemes models provides several solutions and the solution leading to the maximum likelihood is kept. If k possible transcriptions are possible, then k solutions will be selected among the different alignments.

When phonemes are considered as basic target units, the $\{f_i\}$ represent source-to-target phonemes association. This association has to be inferred from training utterances. For this purpose, training utterances are segmented into phonemes. These utterances are then aligned on the source phonemes loop. For every target phoneme the corresponding acoustic segments are selected. The target phoneme is associated with the top m source phonemes having maximum overlap with its acoustic segments.

Once the correspondence between source units and target units determined, the parameters of the source models may be adapted to better describe the distribution of the training target utterances. Several techniques may be used for adaptation purposes [2]. The experiments conducted in this work are limited to the MLLR adaptation.

3. DATABASES AND PHONETIC REPRESENTATIONS

Two databases are used in our experiments, Swiss French Polyphone [5] database and an Arabic (Lebanese dialect) database. The Swiss French Polyphone consists of approximately 10 sentence samples and 28 isolated words for 5,000 speakers. Two sets of phone models are trained on this database. The first set of phone models, referred to as PL16 in this paper, consists of 42 phone models including 2 silence models and 6 closure models. The second set, PL32, is a total of 36 phone models with the plosive phone models trained as a single phone. All models are trained on 9938 phonetically rich sentences spoken by 1000 speakers (500 men, 500 women). In addition, the two model sets are three-state left-to-right Hidden Markov Models (HMMs), where the PL16 set has 16 component Gaussian mixture distributions associated with each state whereas the PL32 set has each state emission density comprising 32 Gaussian mixture components.

The Arabic database, on the other hand, is a SpeechDat like telephone database collected at the University of Balamand (UOB). This database contains 923 isolated word samples collected from approximately 50 male and female speakers between the ages 18 and 25. 550 samples, approximately 14 word utterance per speaker, of this database are used as the target language database and are divided into two sets, one for training (208 samples) and the other for testing (342 samples). The training utterances were segmented into phonemes and used for alignment and adaptation.

HTK [2] was used to train these models. As feature vectors, 13-components MFCC vectors are extracted from 25.6ms windows every 10ms. The first- and second-order derivatives are associated to the static MFCC vectors leading to a feature vector with 39 components.

For phonetic representation, the IPA (International Phonetic Alphabet), which is a well-known and useful tool in exploring phonetic similarity across languages, is used along with SAMPA, the keyboard representation of its sample at this stage for all Arabic transcriptions [3]. The French phone models, however, relied on S2 standard and were changed to SAMPA in the second approach.

4. TARGET LANGUAGE WORDS ASSOCIATED WITH SOURCE LANGUAGE PHONEMES

In the first case, phonetic sequences using the source model set were created either manually or automatically for each target word in the corpus. These sequences were then used along with start and end silence nodes for the creation of word networks. Finally, adaptation using the training utterances was conducted and performance was evaluated using the test utterances. The following sections describe the determination of the sequences both manually and automatically and their corresponding results.

4.1. Manual transcription

The acoustic evidence for each word in the target corpus was used to determine its best phonetic representation in this approach. Although the corpus consists of Lebanese

dialect samples only, the accents of different geographical regions led to more than one transcription for each word in some cases. Some example sequences are shown in *Table 1*. Models for the Arabic vocabulary were built based on these transcriptions. The training utterances were used to adapt those models.

Table 1 – Manual Word Transcriptions

Arabic Pronunciation	French Transcription
3amil	an mm ei ll ai mm ei ll
akfil	aa kk ff ei ll
…	…

4.2. Automatic transcription

To automatically associate each of the target words with a sequence of phonemes, recognition was conducted on the training utterances using an elementary loop of phonemes to produce N-best results. The three recognized sequences with the highest likelihood are selected for each word. These transcriptions (e.g. Table 2) referred to as automatic phonetic transcriptions were used to build a new model which was adapted and evaluated.

Table 2 –Automatic Word Transcriptions

Arabic Pronunciation	French Transcription
3amil	ff an in mm ai ll ss ff an in mm ee ll ss ff an in mm ee ii ll un
akfil	aa pp ff ii ll ss ai pp ff ei ll aa kk ff ei nn ll
…	…

Figures 3 and 4 show the recognition results after different iterations of MLLR adaptation. Automatic transcription produces better performance with simpler source model PL16, while we notice the opposite tendency with a more precise model PL32. Our interpretation is that more data is necessary to adapt PL32 models.

5. PHONEMES CORRESPONDENCE TABLES

Word is the largest target language unit which would offer the best modeling precision using other language smaller units. However, this requests to regenerate a new lexical dictionary for the target language. It would be practical if an association of smaller units could be found. Therefore, phonemes association was studied. The mapping between the target and source units was done both manually and automatically in a tabular way.

These tables, together with lexical description of target words in terms of target language phones, were then used to create the new model by copying the source unit model and renaming it according to the target phone it represents. Finally, as was the case in the previous method, recognition and MLLR adaptation were conducted.

5.1. Manual correspondence

Similar to the manual transcription method, correspondence was done relying on human expertise. An association table, sample of which is shown in Table 3, was manually created relating each Arabic phone to one or more French phones.

5.2. Automatic correspondence

Table 3 – Phone Correspondence

Arabic Symbol	Arabic Word	French Symbol	French Word
b	ba:b (door)	bb	bateau
T	Tala:T (three)	ss	salut
…	…	…	…

To automatically create a correspondence table, the first step was simple recognition of the target utterances in the training set using a simple source units' loop. Upon comparing the resulting transcriptions with the original ones, the ratio of correspondence of each target unit to each source unit was calculated. Finally, the French unit with the highest ratios for each target phone was joined to create an association file similar to Table 3. A sample of the automatic correspondence results obtained for each set of phone models is shown in Figure 1.

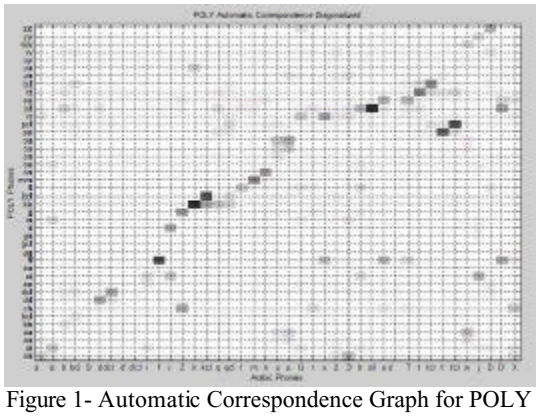

Figure 1- Automatic Correspondence Graph for POLY

Both automatic approaches include the use of the PL16 automatic mapping tables for the PL32 model sets experiments and vice versa. These are referred to as PL16_32 and PL32_16 respectively.

Figures 3 and 4 produce the results for phonemes correspondence and compare them to those obtained with

automatic word transcriptions. Here, for both models, PL16 and PL32, automatic correspondence provides better results. Our interpretation is that correspondence method introduces more constraint since one phoneme-to-phoneme correspondence is selected. In comparison, transcription approaches, whether manual for PL32 or automatic for PL16, overcome correspondence approaches which is expected intuitively. In order to study and compare the capabilities of each method and every model to determine association and to use those associations in acoustical modeling several experiments are conducted and reported in the Figures 5 and 6.

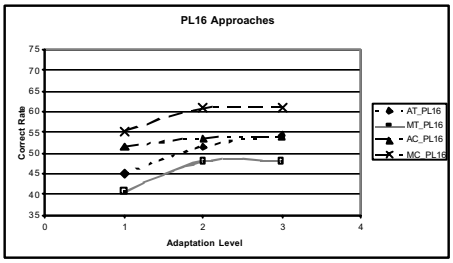

Figure 2- PL16 word recognition for all methods.

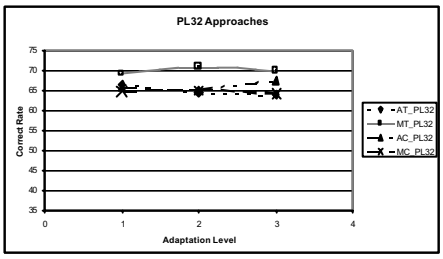

Figure 3- P32 word recognition for all methods.

Figure 5 shows the recognition results for different automatic transcriptions. It is clear that a more precise model PL32 is preferable to determine the transcriptions that can be used successfully with a less precise model PL16 that is more suitable for adaptation with few data. Regarding the capability of the more precise model to determine better associations the Figure 6 confirms this tendency. However, these models need to be maintained since phoneme correspondence is more constrained.

6. CONCLUSIONS AND PERSPECTIVES

In the general framework of multilingual speech recognition, this paper presents a work on corresponding acoustic units between a source language and a target language. The phonemes are considered as the source language acoustic units. Two target units are studied; the word and the phonemes. In each case, the association can be set manually or inferred automatically in a data-driven approach. For the latter direction algorithms are developed and presented. The source language is the French language while Arabic (Lebanese dialect) was considered as the target language. The effect of the precision of the source language models is investigated. The experiments' results permitted to conclude that the data-driven approach is generally more suitable. It is generally better to use more precise models to infer the association. This inferred association can be used with other less precise models that may be more suitable for acoustic adaptation. As a final conclusion, we may say that although phoneme correspondence is more appropriate to build general purpose recognition models, word transcription provides better results.

Finally, several perspectives exist for this work. The optimal choice for both source and target acoustic units must be determined. Building a multilingual set of acoustical units is another perspective.

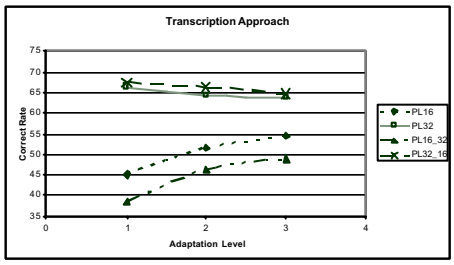

Figure 4- Automatic Transcription recognition results.

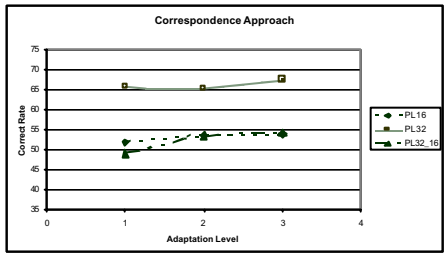

Figure 5- Automatic correspondence recognition results.

Acknowledgment: This work is partly supported by the CEDRE project n. (2001 T F 49 /L 42).

7. REFERENCES

[1] E. Wong et al., "Multilingual Phone Clustering for Recognition of Spontaneous Indonesian Speech Utilizing Pronunciation Modeling Techniques", *Proc. EuroSpeech'03*, Vol., pp 3133-3136, 2003

[2] J. Odell, D. Ollason, P. Woodland, S. Young, J. Jansen, "The HTK Book for HTK V3.2", *Cambridge University Press*, Cambridge, UK, 2002.

[3] IPA, Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet, *Cambridge University Press*, 1999.

[4] H. Mokbel, D. Jouvet, "Automatic derivation of multiple variants of phonetic transcriptions from acoustic signals," *Proc. EuroSpeech'97*, Vol. 3, 1997.

[5] G. Chollet et al, "Swiss French Polyphone and PolyVar: Telephone Speech Databases to model inter- and intra-speaker variability", *IDIAP-RR 96-01*, 1996.