ACOUSTIC MODEL INTERPOLATION FOR NON-NATIVE SPEECH RECOGNITION

Tien-Ping Tan, Laurent Besacier

CLIPS-IMAG Laboratory, UMR CNRS 5524
BP 53, 38041 Grenoble Cedex 9, France
tien-ping.tan@imag.fr, laurent.besacier@imag.fr

ABSTRACT

This paper proposes three interpolation techniques which use the target language and the speaker's native language to improve non-native speech recognition system. These interpolation techniques are manual interpolation, weighted least square and eigenvoices. Each of them can be used under different situation and constraints. In contrast to weighted least square and eigenvoices methods, manual interpolation can be achieved offline without any adaptation data. These methods can also be combined with MLLR to improve the recognition rate. Experiments presented in this paper show that the best non native adaptation method, combined with MLLR can give 10% WER absolute reduction on a French automatic speech recognition system for both Chinese and Vietnamese native speakers.

Index Terms— non-native ASR, interpolation, adaptation

1. INTRODUCTION

Automatic speech recognition applications are becoming increasing popular. However, as automatic speech recognition matured, speech recognition performance on non-native speakers is still low.

Recently, there were increased interests in the research of non-native speech recognition. Works in non-native speech recognition take into account variations of speech from non-native speakers compare to native. Most of them focus on acoustic model adaptation [1] and pronunciation lexicon improvement [2]. Getting non-native speech for acoustic modeling is often difficult and in some cases unfeasible. Therefore, research in acoustic model adaptation attempt to use limited non-native speech [3] or the speaker's mother tongue (source language) to improve the target language acoustic models. We were encouraged to use only the speaker's mother tongue to improve the target model for non-native speech recognition, because studies have shown that non-native speakers often substitute target language phonemes with their native language phonemes [4]. Furthermore, it is relatively easier to access the corpus of a particular language compare to non-native speech. In this paper, we will explore different interpolation methods to improve the baseline acoustic model.

This paper is organized as following. In *section 2*, we describe our approach to find non-native speaker's phoneme substitution. In *section 3*, we will use the knowledge of phoneme substitution for adapting the acoustic models. *Section 4* gives the experimental results. Finally, conclusions are drawn in *section 5*.

2. PHONEME SUBSTITUTION BY NON-NATIVES

The approaches to determine the target language phoneme substitution by non-natives can be divided into two groups: knowledge based and data driven. In knowledge based approaches, the most probable phoneme confusion can be obtained through perception test, acoustic phonetic analysis or simply through the analysis of the IPA table of both the target and the source languages. On the other hand, data driven methods can be achieved by using phoneme distances [5], statistical phoneme confusion matrix [6] and others.

We used a combination of knowledge from the IPA table and followed by a confusion matrix to find the possible non-native's phoneme substitution. Using the IPA information, we can know similar phonemes which exist in different languages. For new target language phonemes, not existing in the source language, their sound class, manner of articulation, place of articulation and other phonetic knowledge can help to decide the native phoneme to be used. On the other hand, diphthong can be considered as two corresponding phonemes. As for the phoneme confusion matrix, we performed a forced alignment on the non-native speech using target language acoustic model. Subsequently, automatic phoneme recognition is executed using the acoustic model of the source language on the same data. The probability that the source language phoneme matches the target language phoneme is measured for all target language phonemes through time alignment. The one with the highest probability is then selected. Details on the phoneme substitution process we used are given in *section 4*.

3. NON-NATIVE SPEAKER ADAPTATION

We describe here three different interpolation techniques to improve an existing target language acoustic model using the source language. Note that the source language speech used for adaptation involves different speakers than the non-native speakers in our corpus. In general, these approaches consist of first determining the possible target language phoneme substitution (as described in the preceding section). Next, the phoneme mapping information is used to create a new source language acoustic model, which has the same configuration as the target language acoustic model, so that both can be interpolated.

3.1. Manual Interpolation

Manual interpolation can be performed by estimating a-priori weights to be multiplied to the acoustic model of the target and source language.

With the knowledge of the target language phoneme substitution, it is still not ready to directly perform interpolation because the Gaussians in each corresponding phoneme in the source and target language acoustic model are still in a state of mismatch. One way is by selecting the nearest Gaussian to a particular Gaussian in the target language acoustic model from the source language acoustic model using certain distance measure for example Euclidean distance and divergence measure [1]. In cases where the original corpus is not available, this can give a reasonably good result. In situation where we have access to the original speech corpus, it is better to recreate the source language Gaussians, based on the target language acoustic model.

We propose to do this by first mapping each phoneme in the pronunciation dictionary of the source language to the phoneme of the target language using the phoneme substitution information that we found previously. Note that several source phonemes can be mapped to the same target phoneme. Instead of using Baum Welch algorithm to recreate the Gaussians, we adapt the initial target language acoustic model (MLLR then MAP) using the source language corpus and pronunciation dictionary. This will model a source language acoustic model even with small amount of speech. The resulted acoustic model has the same number of Gaussians as the target language acoustic model, and at the same time all the Gaussians are clustered accordingly. Weights for each model are then predicted, and a new model is created with the following formula:

$$p_{\text{Adapt}} = w \cdot p_{\text{Target}} + (1\text{-}w) \cdot p_{\text{Source}} \qquad (1)$$

where p_{Adapt}=adapted acoustic model, p_{Target}=target language acoustic model, and p_{Source}=source language acoustic model. w= weight, $0<=w<=1.0$

3.2. Weighted Least Square

Manual interpolation relies on fixing a-priori weights for the acoustic models. In certain situation when we are able to obtain some target speech from the non-native speaker, we would like to predict the weights to be applied on the acoustic models. Here we attempted to use weighted least square (WLS) to predict the weights. The equation (1) can be rewritten as a matrix formulation:

$$A\,x = b \qquad (2)$$

where,

$$A = \begin{pmatrix} p_{\text{Target}} & p_{\text{Source}} \end{pmatrix} \quad x = \begin{pmatrix} w1 \\ w2 \end{pmatrix} \quad b = \begin{pmatrix} p_{\text{Adapt}} \end{pmatrix}$$

Using some non-native speaker's adaptation utterances, we can derive the speaker's means by forced-aligning them using the initial target language acoustic model. Since only a few utterances are used, some of the Gaussians in the vector *b* will be zeros. We ignore all the Gaussians in *A* and *b* where the Gaussian in b is zero.

We can solve the above equation and find *x*, given the least errors using the least square formula. Since not every mean has the same weight, we applied weights to the least square formula [7]. Variances ($\boldsymbol{\Sigma}$) are used as the weights for the weighted least square formula, so

$$A^T A\,\varkappa = A^T\, b \qquad (3)$$

$$A^T \boldsymbol{\Sigma}^{-1} A\,\varkappa = A^T \boldsymbol{\Sigma}^{-1}\, b \qquad (4)$$

3.3. Eigenvoices

Eigenvoices method has been successfully used in speaker adaptation [8]. In the previous two approaches, interpolation is performed on the acoustic models. Instead for eigenvoices, we can see it as an interpolation of eigenvectors. The standard eigenvoices technique is applied here. However, we attempt to expand the eigenspace by creating a non-native space using the speaker's native language.

Speaker adaptation in eigenvoices is achieved by creating a speaker space and subsequently finding the speaker we want to adapt on that space. The first step to create a speaker space is to create a speaker dependent acoustic model for each speaker. For each target language speaker dependent acoustic model, the process is the usual one, where we first create a speaker independent acoustic model. We subsequently derive speaker dependent model for each speaker, by adapting the speaker independent model that we have created using the speech from each speaker with a combination of MLLR and then MAP adaptation.

Next, we created the components for non-native space by going through similar steps we used to create the source language acoustic model for the previous interpolation methods. The only difference is that in eigenvoices, we have to use MLLR and MAP to adapt the target language speaker independent acoustic model to speaker dependent acoustic models using the source language speech from each speaker.

After the speaker dependent acoustic models for target and source languages are created, the means of the acoustic models are written out, each as a sequential vector which is known as supervector. Next, principal components analysis (PCA) or singular value decomposition (SVD) can be used to find the eigenvectors which define the eigenspace.

Not all eigenvectors will be used. A subset of eigenvectors which has among the highest eigenvalues (principal components) is selected for interpolation. The projection methods in PCA, MLED [9] or others [10] are among the approaches which can be used to find the interpolation weights.

4. EXPERIMENTS

The experiments were performed on our non native French corpus [11] using CMU Sphinx ASR. There are two groups of non-native speakers: Chinese and Vietnamese. Each speaker read about a hundred sentences related to tourism domain. The baseline 16 Gaussians target language context independent acoustic model was created using BREF120 corpus [12], while for the source language, we had a 15 hours of Vietnamese corpus [13] and a 5 hours of Mandarin Chinese [14]. The general domain trigram language model was created using *Le Monde* newspapers text, and subsequently interpolated with a tourism domain language model (from NESPOLE project). The average WER for native Vietnamese and Chinese speakers is high and stands at 60.6% and 58.5% respectively. The high difficulty of the database is confirmed by human perception test[1] which showed average WER of 12.1% and 11.3% respectively.

4.1. Determining the Phoneme Substitution

We performed a time-alignment scoring to obtain the phoneme confusion matrix. By comparing only the similar phonemes in French-Vietnamese and French-Chinese (accounts about half of the total phonemes in both languages), we found there were about 24% and 28% of wrong classification in Vietnamese and Chinese respectively from the phoneme confusion results. This has prompted us to use IPA table to find the common source-target phoneme substitutions when possible. For new target (French) phonemes, the new target language voiced plosives were matched to unvoiced version in the source language. The same applied for fricatives. For Vietnamese speakers, the French phonemes /ʃ/ and /ʒ/ were substituted by Vietnamese /s/ and /z/ respectively. For the French glide /ɥ/, it was mapped by the source glide /w/. As for vowels, replacement is based on the vowel chart and other studies [4]. For Chinese speakers for example, we replace /ɔ/, /ɑ/ and /ɛ/ with /o/, /a/ and /e/ respectively. For others which we were not sure, phoneme confusion matrix results were being used.

4.2. Manual Interpolation

We evaluated the model created using the method proposed in *section 3.1* against the model created by selecting the nearest Gaussian for each corresponding target state from the source acoustic model using Euclidean distance. The weights in the range of zero until one were gradually assigned to the models.

The results showed that the proposed interpolated model has average WER which is always lower than the baseline. Native Vietnamese has the lowest average WER when the weights for French and Vietnamese are at 0.3 and 0.7. On the other hand, the lowest average WER for native Chinese is when the weights for French and Chinese are at 0.2 and 0.8 respectively. This may indicate that native Chinese speech is slightly more accented.

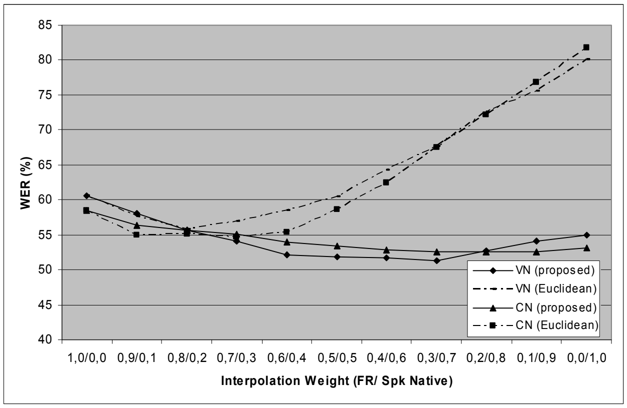

Figure 1. Graph shows manual interpolation at different weights for native Vietnamese (VN) and native Chinese (CN) speakers of French (FR). FR weight at 0.0 denotes baseline results.

4.3. Weighted Least Square

Three utterances were used to predict the interpolation weights for each speaker. The results show that if some speech from the speaker is available it is advantageous to use it to predict the weights.

Native Speaker	Baseline	FR=0.5, VN/CN=0.5	Manual Int. (best result)	WLS
Vietnamese	60.6	51.9	51.3	51.3
Chinese	58.5	53.4	52.5	53.2

Table 1. Comparing WER of manual interpolation and WLS

4.4. Eigenvoices

From BREF120 corpus, a French speaker independent acoustic model was initially created. From the speaker independent acoustic model, subsequently 120 speaker dependent models were derived. Using the French speaker independent model as the initial model, we have also created 29 speaker dependent models from Vietnamese corpus and 20 from Chinese corpus. The means of the speaker dependent models were written out as supervectors. To evaluate the performance, we performed speaker adaptation

[1] Human listeners were asked to transcribe non-native utterances, where an unlimited number of replays for each utterance were permitted.

with eigenvectors created using 120 French supervectors and also when the additional supervectors from the speaker's native language were added. MLED adaptation was applied using 20 principal components. Table 2 below shows that adding speaker's native speech (fourth column) improves the performance compared to using French only speech (third column). Since the recording condition and speaker are different from the test, we conclude that the improvement is resulted from the use of the speaker's native language.

Native Speaker	Baseline	FR supervectors	FR + VN/CN supervectors
Vietamese	60.6	54.7	51.9
Chinese	58.5	52.7	51.5

Table 2. Average WER of eigenvoices using 20 components

To verify that this improvement is actually due to the adding of the source language data, and not to the bigger number of supervectors, we evaluated the performance of our system by varying the number of FR supervectors used and setting the number of principal components used at constant. The results show that when the number of target language (FR) supervectors reaches 40, the adaptation has already reached an optimum state for native Vietnamese and native Chinese, where subsequent results do not vary much (less than 1%) after that. This shows that the improvement observed in Table 2 is due to the use of the source language to create the eigenvectors for non-native speaker adaptation.

4.5. Comparing approaches

To evaluate the performance further, we compare our approach with MLLR. We also perform combined adaptation using our methods with MLLR. For both Chinese and Vietnamese speakers, the best non-native adaptation method, combined with MLLR, gives 10% WER absolute reduction.

Native Speaker	Baseline	MLLR	WLS	Egv.	WLS + MLLR	Egv + MLLR
Vietnamese	60.6	53.4	51.3	51.9	50.1	50.1
Chinese	58.5	51.5	53.2	51.5	50.5	48.9

Table 3. Comparing WER of different approaches for non-native speaker adaptation

5. CONCLUSION

We have presented three types of interpolation techniques to improve target language acoustic model using source language for non-native speakers. The methods require the use of source language corpus. As for future work, we will investigate whether there is a better way to determine the Gaussian for interpolation when we only have the source language acoustic model.

6. REFERENCE

[1] S. Witt, "Use of Speech Recognition in Computer Assisted Language Learning," University of Cambridge, Ph.D Thesis, 1999.

[2] S. Goronzy, *Robust Adaptation to Non-Native Accents in Automatic Speech Recognition*, Springer Verlag, German, 2002.

[3] Z. Wang and T. Schultz, "Non-Native Spontaneous Speech Recognition through Polyphone Decision Tree Specialization," *Proc. Eurospeech-03*, Geneva, pp. 1449-1452, 2003.

[4] J. Flege, "The Production of 'New' and 'Similar' Phones in a Foreign Language: Evidence for the Effect of Equivalence Classification," *Journal of Phonetics* vol. 15, pp. 47-65, 1987.

[5] B.H. Juang and L.R. Rabiner, "A Probabilistic Distance Measure for Hidden Markov Models," *AT&T Technical Journal* vol. 64, no. 2, pp. 391-408, 1985.

[6] J.J. Humpries and P. Woodland, "The Use of Accent-Specific Pronunciation Dictionaries in Acoustic Model Training," *ICASSP-98*, Seattle, vol. 1, pp. 317-320, 1998.

[7] D.C. Montgomery, E.A. Peck and G. Geoffrey Vining, *Introduction to Linear Regression Analysis, 3rd Edition*, Wiley, 2001.

[8] R. Kuhn, P. Nguyen, L. Goldwasser, N. Niedzielski, S. Fincke and M. Contolini, "Eigenvoices for Speaker Adaptation," *ICSLP 98*, Sydney, pp. 1774-1777, 1998.

[9] P. Nguyen, "Fast Speaker Adaptation," Technical report, Eurecom, 1998.

[10] R.J. Westwood, "Speaker Adaptation Using Eigenvoices," University of Cambridge, MPhil Thesis, 1999.

[11] T.P. Tan and L. Besacier, "A French Non-Native Corpus for Automatic Speech Recognition," *LREC 2006*, Genoa, pp. 1610-1613, 2006.

[12] L.F. Lamel, J.L. Gauvain and E. M., "BREF, a Large Vocabulary Spoken Corpus for French," *Eurospeech-91*, Genoa, pp. 505-508, 1991.

[13] V.B. Le, T. Do-Dat, E. Casteli, L. Besacier and J.F. Serignat, "Spoken and written language resources for Vietnamese," *LREC 2004*, Lisbon, pp. 599-602, 2004.

[15] CCC Corpora, Chinese Corpus Consortium, http://www.d-ear.com/CCC/corpora.htm