A TWO-STAGE SPECTRAL TRANSFORMATION APPROACH TO FAST SPEAKER ADAPTATION

H.C. Choi and R.W. King

Speech Technology Research Group
Department of Electrical Engineering
University of Sydney

ABSTRACT - A two-stage approach for using spectral transformation to perform supervised speaker adaptation for continuous speech recognition is proposed. In the first stage, all the speech data of the training speakers are first transformed to the spectral space of a reference speaker. Then a set of adaptive Hidden Markov Models (HMMs) is trained for this reference speaker using his/her speech data and the transformed speech data of the other training speakers. Once the models have been trained, they are used for all new speakers. In the second stage, speaker adaptation is performed by transforming the feature vectors of the speech of a new speaker to the spectral space of this reference speaker. Recognition is then performed using the pre-trained models. The effectiveness of this approach is investigated using the DARPA Resource Management (RM1) continuous speech corpus.

INTRODUCTION

It is commonly agreed that a well trained speaker-dependent recognition system always performs better than its speaker-independent counterpart. However, since speaker-dependent models are speaker-specific, a new set of models has to be trained for each new speaker if high recognition performance is required. Thus, as far as a new speaker is concerned, the training procedure of a speaker-dependent recognition system has two main problems. The first one is the considerable effort and time required of the new speaker to provide sufficient training data. The second one is that a significant lead time is then required for training the models before the new speaker can actually use the system.

As remedies to address the above problems, various speaker-adaptive methods have been proposed. These methods have been used to either adapt a new speaker to a recognition system or alternatively, adapt a recognition system to a new speaker by utilising only a small amount of training speech data from the new speaker. Basically, these methods can be classified into three categories, namely, spectral transformation (Class et al., 1990; Huang, 1992; Knohl & Rinscheid, 1993; Choi & King, 1994a,b), speaker classification (Imamura, 1991; Nakamura & Akabane, 1991; Ljolje, 1993; Kosaka & Sagayama, 1994) and re-estimation of model parameters (Lee et al, 1991; Necioglu, 1992; Takami et al., 1992; Lee & Gauvain, 1993; Kosaka et al, 1993; Huo et al., 1994).

All these methods can tackle the first problem quite well by largely reducing the amount of training data required. However, most adaptive methods still have the second problem because of their time-consuming adaptation processes. One example is the use of re-estimation of model parameters for adaptation. Since this approach requires re-training of the recognition models during adaptation, it is not so suitable for fast adaptation if there are many models or model parameters in the recognition system. Hence for applications which require fast adaptation, the first two approaches listed above are more appropriate. But since the implementation of speaker classification adaptation requires a different set of models for each class, more resources may be needed if this approach is used.

When spectral transformation is used for speaker adaptation, one can perform the transformation either from the spectral space of a new speaker to that of a reference speaker or in reverse from the spectral space of a reference speaker to that of a new speaker. Generally speaking, one will have a much faster adaptation process but a lower recognition accuracy after adaptation for the same amount of training data if the transformation in the first direction is chosen to transform the feature vectors. Since fast adaptation is required for practical applications, the approach, which uses transformation in the first direction is more appropriate. Many methods have been proposed to improve the adaptation performance of this approach

and all of them concentrate on improving the estimation of the transformation. Here we propose a new approach which also improves the training of the basic recognition system so as to further improve the adaptation performance. In this approach, the adaptation process is divided into two stages. In the off-line stage, spectral transformations are used to improve the training of a basic recognition system. Since this stage is done off-line, the processing time in this stage will not affect the on-line adaptation processing time. In the on-line stage, spectral transformations are used to transform the feature vectors of those new speakers to the spectral space of a reference speaker.

In this work, the use of the proposed two-stage spectral transformation approach to supervised speaker adaptation for continuous speech recognition is investigated. Context independent phones are modelled by continuous density Hidden Markov Models (HMMs) in the recognition systems. The spectral transformation required for each speaker is estimated by using canonical correlation analysis (CCA) in order to achieve fast adaptation.

TWO-STAGE ADAPTATION APPROACH

The spectral differences between two speakers can be compensated if a vector in the spectral space of a speaker can be transformed appropriately to the corresponding vector in the spectral space of another speaker. This transformation can be computed using the corresponding pairs of vectors which are obtained by performing dynamic time warping (DTW) on the same context of speech uttered by these two speakers. The main concept of spectral transformation is shown in Figure 1.

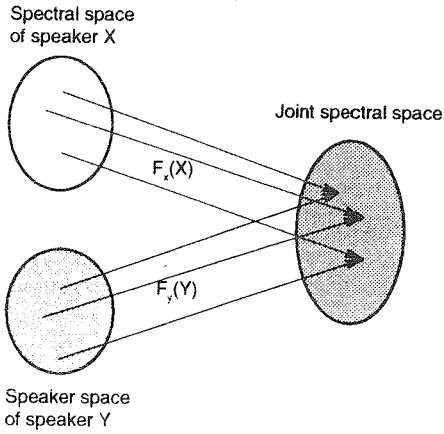

Figure 1 : Concept of Spectral Transformation

In this work, CCA is used to compute the required transformations. Basically, CCA determines two transformations which each of them transforms the spectral space of each speaker into a joint spectral space. The transformations are made so that the correlations between corresponding components of the transformed vectors are maximised. Once the transformations are found, one of them is inverted and then combined with the other one to form a single transformation. More details on the use of CCA to estimate a spectral transformation can be found in our previous work (Choi & King, 1994a,b).

The proposed adaptation process is divided into two stages, namely, an off-line stage and an on-line stage. In the first stage or the off-line stage, all the speech data of the training speakers are first transformed to the spectral space of a reference speaker. The required transformations are estimated by using CCA and ten adaptation sentences from each of the speakers are used. Then a set of HMMs is trained for this reference speaker using the training speech data of this reference speaker and the transformed speech data of the other training speakers. Once the models have been trained up, they are used as a basic recognition system for all new speakers. The use of spectral transformation in this stage is to reduce the mismatch among the speech data of those different training speakers and hence improve the recognition accuracy of the basic system. Since this stage is done off-line, the actual on-line processing time required for adaptation will not be affected.

In the second stage or the on-line stage, speaker adaptation is performed by transforming the feature vectors of the speech of a new speaker to the spectral space of this reference speaker. Recognition is then performed using the pre-trained basic recognition system. The same recognition system is used for all new speakers but the individual transformation for each new speaker is estimated accordingly by using CCA. Since this approach does not require any modifications to the basic recognition system during the on-line adaptation process, a fast adaptation can be achieved. A block diagram of the whole approach is shown in Figure 2.

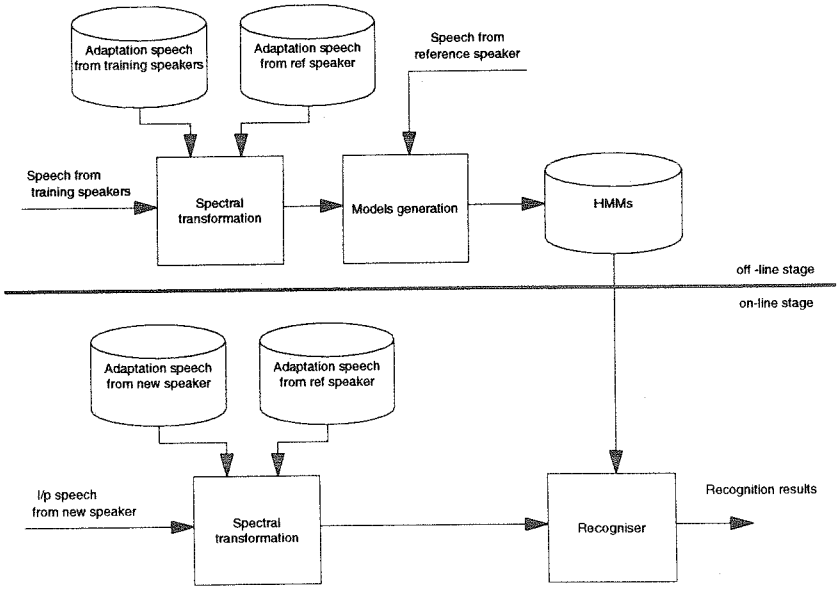

Figure 2 :- Two-stage adaptation approach

EXPERIMENTS

Speech database

The DARPA Resource Management (RM1) continuous speech corpus is used in the adaptation experiments. All 12 speakers from the speaker-dependent portion are used with 7 of them being assigned as new speakers (4 males and 3 females) and another 5 of them being assigned as training speakers (3 males and 2 females). Out of those 5 training speakers, one male and one female are selected as reference speakers. The speaker assignment is made so that all new speakers have different dialects, thus making the adaptation task a severe test of this adaptation approach. Test data for each new speaker consist of 100 sentences from the development-test session while 10 sentences from the rapid adaptation session of each speaker are used as adaptation speech data. On average, the total length of these 10 adaptation sentences is around 40 seconds.

Front-end processing

In both the training of models and speech recognition stages, the speech of a speaker is represented by its 12 mel-frequency cepstrum coefficients (MFCCs), 12 delta MFCCs, normalised log-energy and delta log-energy at a rate of 10 ms with a Hamming window of 25 ms. However, only those 12 MFCCs and normalised log-energy are used in the adaptation process in order to reduce the size of the transformation matrix. The delta values of those transformed feature vectors are computed after adaptation.

Recognition models

Recognition is based on 48 context-independent phone models with a bigram grammar. Each phone model is represented as a left-to-right continuous density Hidden Markov Model (CDHMM) with 3 states and 3 mixtures per state. Speaker-dependent (SD) models for each reference speaker are trained using 600 sentences from that particular speaker. Whilst the speaker-independent (SI) models are trained using all the training sentences (total 3000) from those 5 training speakers. Moreover, reference speaker-adaptive (RSA) models are trained using the appropriately transformed speech data of these 3000 sentences.

RESULTS

Each result reported here is the average performance over the different speaker combinations and it is reported in terms of percentage (%) of word error which includes substitution, deletion and insertion errors. Since there are 7 new speakers and 2 reference speakers, 14 cross-speaker conditions can be tested. Without adaptation, the average cross-speaker and speaker-independent test results are as shown in Table 1.

Table 1 :- Test results before adaptation

Test	Word error (%)
Cross-speaker	38.4
SI	9.8

Since different sets of recognition models can be used for the same transformation, a few sets of adaptation experiments can be done to demonstrate the effects of choosing different sets of models as the basic recognition system. In this work, three sets of models have been used one by one as the basic recognition system and they are speaker-dependent (SD), speaker-independent (SI) and reference speaker-adaptive (RSA) models respectively. Average test results after adaptation are shown in Table 2.

Table 2 :- Test results (word error in %) after adaptation

	Number of adaptation sentences		
Basic recognition models	1	5	10
SD	15.6	11.6	11.1
SI	13.1	9.1	9.3
RSA	11.3	8.6	8.3

DISCUSSION

As observed from the above results, it is found that the adaptation performance of spectral transformation is quite dependent on the basic recognition models being used. For a given transformation, the better the set of basic recognition models is trained, the better the recognition result after adaptation is. Comparing the values of those results in each column of Table 2, it is found that the two-stage approach, which uses the RSA models as basic recognition system, has the best adaptation performance.

It is interesting to note that although both sets of SI and RSA models are trained using the same source of speech data, the use of spectral transformations to pre-process the speech data in the training of the RSA models has generated a better set of models. Provided that each transformation is estimated accurately in the off-line stage, the transformed speech data of those training speakers can be treated as if they are speech data from the same reference speaker. Therefore, the training of the RSA models can be viewed as if it is a speaker-dependent one and hence better recognition performance is expected.

There exist other non-linear estimation methods (Huang, 1992; Knohl & Rinscheid, 1993) which have been used successfully to estimate the required transformation. However, they are not so suitable for on-line adaptation. Since all these methods require the training of a certain type of neural network to learn the transformation, a long processing time is required before the transformation can be made available for adaptation. Therefore, CCA is chosen to estimate the transformation because it is linear, simple to implement, capable to handle most speaker variations and most important, its computation is fast.

Further improvement to the adaptation performance can be made possible if class-dependent (e.g. sex, accent, phonetic classes) transformations are used. The challenges will then be how to choose the right classes and class sizes that are effective and meanwhile will not increase the amount of computation beyond the tolerable limit. The issues involved will be studied in our future work.

CONCLUSION

The use of this two-stage spectral transformation approach to speaker adaptation for continuous speech recognition is found to be promising. Using the average SI test result before adaptation as reference, it is found that this approach has achieved, on average, an error reduction of 15.3% when only ten adaptation sentences are used. Since this approach does not require any modifications to the basic recognition system during the on-line stage and its adaptation is fast, it is particularly attractive for the future development of on-line speaker adaptive recognition system.

ACKNOWLEDGEMENT

H.C. Choi is supported by the Australian Postgraduate Award (APA).

REFERENCES

Choi H.C. & King R.W. (1994a), "Speaker Adaptation through Spectral Transformation for HMM Based Speech Recognition", Proc. Int. Sym. on Speech, Image Processing and Neural Networks, ISSIPNN94, pp. 686-689.

Choi H.C. & King R.W. (1994b), "Fast Speaker Adaptation through Spectral Transformation for Continuous Speech Recognition", Proc. Int. Conf. on Spoken Language Processing, ICSLP94, pp. 2219-2222.

Class F. et al. (1990), "Fast Speaker Adaptation for Speech Recognition Systems", Proc. ICASSP90, pp. 133-136.

Huang X. (1992), " Speaker Normalization for Speech Recognition", Proc. ICASSP92, Vol. I, pp. 465-468.

Huo Q. et al. (1994), "Bayesian Learning of the SCHMM Parameters for Speech Recognition", Proc. ICASSP94, Vol. I, pp. 221-224.

Imamura A. (1991), "Speaker-Adaptive HMM-Based Speech Recognition with A Stochastic Speaker Classifier", Proc. ICASSP91, pp. 841-844.

Knohl L. & Rinscheid A. (1993), "Speaker Normalization and Adaptation Based on Feature-Map Projection", Proc. Eurospeech93, pp. 367-370.

Kosaka T. et al. (1993), "Dynamic Approach to Speaker Adaptation of Hidden Markov Networks for Speech Recognition", Proc. Eurospeech93, pp. 363-366.

Kosaka T. & Sagayama S. (1994), "Tree-Structured Speaker Clustering for Fast Speaker Adaptation", Proc. ICASSP94, Vol. I, pp. 245-248.

Lee C.H. et al. (1991), "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models", IEEE Trans. on Signal Processing, Vol. 39, No. 4, pp. 806-814.

Lee C.H. & Gauvain J. (1993), "Speaker Adaptation Based on MAP Estimation of HMM Parameters", Proc. ICASSP93, Vol. II, pp. 558-561.

Ljolje A. (1993), "Speaker Clustering for Improved Speech Recognition", Proc. Eurospeech 93, pp. 631-634.

Nakamura S. & Akabane T. (1991), " A Neural Speaker Model for Speaker Clustering', Proc. ICASSP91, pp. 853-856.

Necioglu B.F. (1992), "A Bayesian Approach to Speaker Adaptation for the Stochastic Segment Model", Proc. ICASSP92, Vol. I, pp. 437-440.

Takami J.I. et al. (1992), "Speaker Adaptation of the SSS (Successive State Splitting)-Based Hidden Markov Network for Continuous Speech Recognition", Proc. Int. Conf. on Speech Science and Technology, SST92, pp. 437-442.