ON THE FRAME-BASED AND SEGMENT-BASED NONLINEAR SPECTRAL TRANSFORMATION FOR SPEAKER ADAPTATION

Fikret S. Gurgen[1] and H. C. Choi

Speech Technology Research Group
Department of Electrical Engineering
University of Sydney

ABSTRACT - This study presents speaker adaptation using nonlinear spectral transformation of frame-based and segment-based feature vectors for a continuous density hidden Markov model (CDHMM) speech recognizer. A multilayer perceptron (MLP) structure is used for the nonlinear mapping of speech frames and segments and the adaptation results are compared with those of canonical correlation analysis (CCA) technique (frame-based) which is a linear method. Experiments are performed using isolated words from the TI-46 database. It is concluded that the linear transformation gives better recognition rate after the frame-based adaptation but in the segment-based case, the nonlinear transformation improves the recognition rate after the adaptation in comparing to frame-based linear transformation due to the addition of dynamic information of speech.

INTRODUCTION

Between-speaker variations are one of the main error sources of speech recognition systems. It is generally agreed that a speaker-dependent (SD) system with sufficient training data can easily perform the recognition with an error rate that is two to three times better than that of a speaker independent (SI) system. When the amount of speaker specific training data is limited, the performance improvement is not realized. One way to obtain good performance is to employ the speaker-independent models and to adapt the new speaker with a minimum amount of data. This is often called as speaker adaptation (Lee et al., 1991). Among the various approaches to speaker adaptation such as speaker clustered models (Rabiner et al, 1989), spectral transformation (Choukri, 1986; Choi & King, 1994)or speaker conversion, the adaptation of the model vectors (Brown et al., 1983), we investigate advantages and drawbacks associated with spectral transformation which uses feature parameters.

Speaker variations include various factors such as vocal tract, pitch, speaking speed, cultural and background differences. Given two different speakers, there is no simple mapping function that can cover all these variations. In the context of spectral transformation, we focus on finding out two mapping functions to transform the pieces from the speech of those two speakers. Mapping functions can be classified as linear or nonlinear. The speech of a speaker can also be divided into in frames or larger parts, segments, depending on the specific application. We study the effect of nonlinear mapping functions on the spectral transformations using frames and segments of speech signal.

In the last decade, neural networks (NN) (Lippmann, 1991; Gurgen, 1992; Hush, 1993) have attracted considerable attention as learning systems that optimally adjust their parameters from the training data to function as a nonlinear system between input and output spaces. Since the main functions in daily life are nonlinear, the nonlinear mapping becomes extremely important. In the speech processing, nonlinear mapping is found to be very useful in the recognition and the enhancement. In this paper, we use a variety of static multilayer perceptron (MLP) structures with back propagation (BP) training algorithm as a nonlinear mapping tool for the study of spectral transformation between two speakers. As a comparison, a canonical correlation analysis (CCA) is used as a linear mapping tool. In this case, frames of spectral data are employed since linear mapping can be defined at each point which corresponds to multidimensional a frame in the spectral sample space. Each frame of the new speaker in the spectral space is transformed to the corresponding frame of reference speaker.

1. is on leave from Bogazici University, Computer Engineering Dept. Bebek 80815 Turkey

Isolated words from the TI-46 database are used for the evaluation of nonlinear mapping of frames and segments of speech. Since the vocabulary size is small, we don't need any special design such as codeword-dependent neural network (CDNN) (Huang, 1992) for the improvement of the adaptation.

NONLINEAR SPECTRAL TRANSFORMATION BY MLPS

A MLP is a feedforward NN which is trained by supervised training algorithm such as BP algorithm. In order to be used in spectral transformation, a MLP should have a number of features: First, we should use sufficient number of connections between units in each layer of MLP to learn the complex nonlinear mapping function between input and output space. Second, MLP structure performs better in small/medium size tasks, thus, the feature vectors have to be grouped into smaller units such as frames/segments and then, various MLPs can learn the nonlinear mapping within the corresponding region. The smallest unit of the feature vector can be frames of speech signal. Then, the combination of frames which is called segment can be used as a whole in the mapping for the speaker adaptation. In this study, we first employ a framed-based MLP structure (Figure 1) which transforms each frame x_f of a new speaker x to the corresponding frame y_f of a reference speaker y by minimizing an objective function:

$$\sum O(y_f - T(x_f))$$

where $O(x_f,y_f)$ can be a mean square error (MSE) function

$$MSE = (y_f - T(x_f))^2$$

or a cross entropy (CE) function

$$CE = y_f \ln(T(x_f)) - (1-T(x_f)) \ln(1-y_f)$$

Each basic unit of MLP has a nonlinear sigmoid function with a threshold Q. The weights between basic units are computed through the BP algorithm. In the training phase, the unknown weights are decided using the input-output (new speaker-reference speaker) pairs of spectral data. In the testing or adaptation phase, the input frames/segments of new speaker are applied to MLP and the adapted data is obtained for the usage of speaker-dependent recognition system. The nonlinear mapping between the input and the output frames is easily learned by the MLP since the complexity of the task is not high.

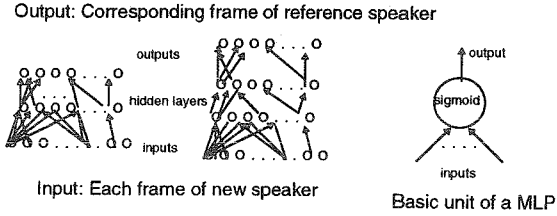

Figure 1. 1 and 2 layer MLP for nonlinear spectral transformation

To show full potential of a nonlinear mapping tool in spectral transformation, we extend the usage of MLPs in frame-based mapping to segment-based mapping (Fukuzawa et al., 1992). It is observed that a MLP structure covering all the frames of the utterances is not powerful enough to learn the nonlinear mapping between speakers. As a result, new MLP structures are constructed (Figure 2). Spectral information of each frame of a segment is transformed by independent 1-hidden layer MLPs and the dynamic information of frames is conveyed to the output by three ways: The first way is to use independent frame-based MLPs for each segment by minimising a common objective function. The

second way is to additionally include the cross connections from the neighboring frames of each side. The last way is to employ separate hidden layer nodes for dynamic information of neighboring frames. The common property of segment-based nonlinear mapping methods that are used here is to minimise a common objective function for the segments in the training phase of MLP.

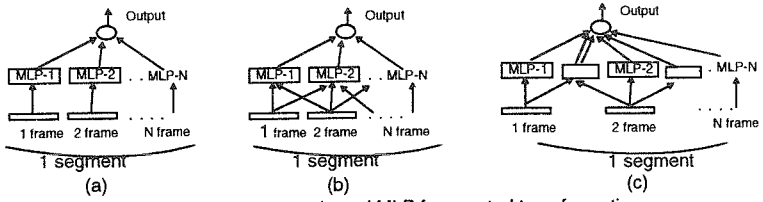

Figure 2. Segment-based MLP for spectral transformation
a. Independent MLPs with common objective function
b. Additional cross-connections from neighbor frames
c. Additional separate hidden nodes for neighbor frames

LINEAR SPECTRAL TRANSFORMATION BY CCA

Since each frame of the speech of a speaker can be described as a point in the n-dimensional feature space, the spectral transformation between two speakers can be linearly performed. In the n-dimensional space, it is possible to find a matrix (with nonzero determinant) that transforms the points of new speaker subspace into the points of reference speaker subspace or into a joint subspace. As a first choice of linear transformation, a matrix A can be found that minimises the mean square error (MSE) between the frames of speaker y and transformed frames of speaker x (Choi & KIng, 1994). As a more general linear transformation, canonical correlation analysis (CCA) method can be used (Figure 3).

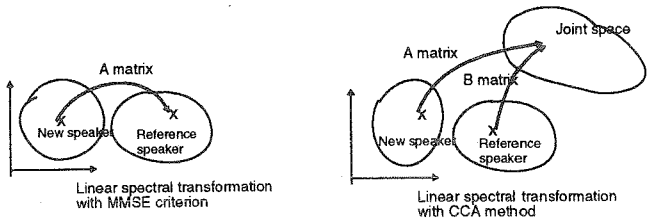

Figure 3. Linear spectral transformation by MMSE criterion and CCA.

The CCA is a linear transformation that transforms the spectral subspace of each speaker into a joint subspace. Moreover, more constraints are imposed on the each transformation. Let

$$u = T_x(x_f) = Ax_f \quad ; \quad v = T_y(y_f) = By_f$$

where A and B are transformation matrices to be determined. In order to estimate A and B, we minimise the MSE

$$MSE = E\{(u_i - v_i)^2\} \quad \text{with constraints } E\{u_i\} = E\{v_i\} = 1$$

where u_i and v_i show the i^{th} component of the vector u and v.The matrices A and B can be found using CCA (Anderson, 1984). One remark is that CCA method can only be applied to spectral transformation as a linear method in the case of frame-based mapping since point by point transformation is possible by the matrices in the n-dimensional space.

EXPERIMENTS AND RESULTS

TI-46 database of isolated-words is used for the experiments. The database is in English and its vocabulary includes 10 digits, 26 letters and 10 computer-related words. There are 8 male (m_i, i=1,..,8) and 8 female speakers (f_i, i=1,..,8). In the experiments, reference speakers with the same sex of a new speaker are chosen for adaptation purpose and speakers f1 and f3 are taken as new speakers and speakers f2, f4, f6, f8 are used as reference speakers. For each speaker, 10 utterances of each word are designated as training tokens while the other 16 utterances are designated as testing tokens. In addition, 3 tokens of each word from the training set of a new speaker are used to form 3 sets of adaptation tokens.

In the preprocessing stage, Mel-Frequency Cepstrum Coefficients (MFCC) are computed as feature parameters. A Hamming window of 25 ms. is shifted every 10 ms. to obtain a frame vector. For each frame of the utterances, a 12 dimensional MFCC vector is computed. In the adaptation stage, only the 12 MFCCs and normalised energy are used. In the recognition stage, 26-dimensional vector that consists of 12 MFCCs, 12 delta MFCCs, normalised log-energy and delta log-energy are used.

A continuous density hidden Markov model (CDHMM) is employed for the recognition. It is a left-to-right model with 6 states. Each state includes 3 mixtures/states. The training set of a speaker is used to train the speaker-dependent models for that speaker.

The MLPs used in the experiments consists of 1 input layer, 1 (1 layer and 2 layers for frame-based adaptation) hidden layer and 1 output layer. For each frame, there are 13 input units and 13 output units. In the frame-based adaptation, 13 hidden units in the 1 hidden layer case and 19 X 19 hidden units for 2 hidden layer case are used. In the segment-based adaptation, 1 hidden layer structure is extended for the usage with common objective function, cross-connections from neighbor frames and separate hidden nodes for neighbor frames cases. A symmetric sigmoid function transfers the weighted inputs with a threshold from previous layer to the next layer. A random initiation with the value between the range of -0.5 and 0.5 is used for all the weights. The BP algorithm which is used for the training of the networks has some features: We started the experiments with a small amount of training data and the sample size is increased depending on the error computed at the output. The momentum and the learning rate of the BP algorithm are adjusted dynamically through the training. We also employ two kinds of objective function: MSE and CE.

In the first part of the experiments, the frame-based adaptation is considered with nonlinear and linear mapping. Each frame of the new speakers' feature parameters is independently adapted to corresponding frame of the reference speaker's feature parameters. The nonlinear mapping of the MLP is obtained by two objective criterions which are MSE and CE and the recognition experiments with CDHMM are performed. The average results of all experiments (f1, f3 are new speakers and f2, f4, f6,f8 are reference speakers) are shown in the Table 1.

Table 1: The average performance of frame-based nonlinear mapping (%)

Frame-based nonlinear mapping	MSE criterion	CE criterion
MLP	75.2	75.8

The performance of CE criterion slightly better than that of MSE. For the speaker adaptation purpose, CE becomes favorable since it optimises cross entropy information between the frames of reference and new speakers instead of considering only MSE. Table 2 shows the performances of linear and nonlinear spectral mapping for frame-based adaptation.

Table 2: The average performance of frame-based linear and nonlinear mapping (%)

Frame-based nonlinear mapping	Linear (CCA) method	Nonlinear (MLP with CE criterion) method
Performance	78.2	75.8

The linear mapping with transformation matrices of the CCA method outperforms the nonlinear mapping of the MLP using in frame-based adaptation (average 2.5% performance improvement). In a very small partition of vector space such as in frames case, there is no extra need to employ a powerful nonlinear method. The point to point mapping of a linear method becomes successful. Also the computation time of a linear method is less than that of a nonlinear method.

When the segment-based adaptation is considered, the necessity of implementing nonlinear mapping becomes unavoidable. Thus, a MLP is again considered as a nonlinear mapping tool. On the primary experiments, it is observed that to employ a single, large MLP for whole utterance is not successful. Even though it would have been successful, the issues such as the training time, performance, modularity,.. would prevent us using a single MLP. As a result, we have decided to implement independent 1-hidden layer MLPs and the dynamic information of frames is conveyed to the output by using independent frame-based MLPs for each segment by minimising a common objective function, by additionally including the cross connections of neighboring frames to same independent modules or by additionally employing separate hidden layer nodes for neighboring frames in a whole piece of word utterance. CE and MSE objective functions are also used. It is found that the best performance is obtained by employing separate hidden layer nodes for neighboring frames (Figure 3c) using CE criterion. As a result, for the rest of the experiments, we employ the configuration in Figure 3c with CE criterion as the nonlinear spectral mapping tool between reference and new speakers.Table 3 denotes the performance of a nonlinear mapping tool (semi-independent MLPs) using the segment-based transformation and compares it with linear mapping tool (on the frame-based transformation case).

Table 3: The average performance of linear (frame-based) and nonlinear (segment-based) mapping (%)

	Linear (CCA) method	Nonlinear method (MLPs with extra hidden units with CE criterion)
Performance	78.2	80.6

The segment-based nonlinear mapping increases the recognition rate an average of 2.5% over the linear method because the valuable dynamic information from the neighboring frames are included to estimate the nonlinear mapping for spectral transformation. The proposed segment-based nonlinear mapping is implemented by the partial nonlinear mappings. Each independent MLP maps a part of the new speaker's vector space into corresponding subspace of reference speaker and an independent MLP module also utilises the neighboring information. The computation time for the segment-based transformation is increased due to the larger size of the network. For those applications that need faster adaptation time, linear mapping becomes advantageous since it basically depends on matrix operations. But for the applications that need better adaptation at the cost of more computation, it is a good choice to employ segment-based nonlinear mapping.

CONCLUSION AND DISCUSSION

Linear and nonlinear methods are used for spectral transformation using frames and segments. Linear method (CCA here) that uses independent feature frames of speech turns out to be easy to use, fast approach of the adaptation. It does not consider the dynamic information of speech signal. On the frame-based spectral transformation, linear methods become efficient in terms of both adaptation performance and simplicity.

Nonlinear methods becomes efficient for the segment-based transformation since it gives the opportunity to extract the dynamic information. Among the proposed semi-independent MLP models to deal with the segments, the model that employs separate hidden nodes for neighboring frames extracts the dynamic information better than the other models. Also, CE is better than MSE in the context of spectral transformation. Most of the dynamic information is assumed to be included in the neighboring frames. The derivation of further dynamic information causes the network to become more complex and difficult to train.

REFERENCES

Anderson T W (1984), " An Introduction to Multivariate Statiscal Analysis," J Wiley & Sons, 2 nd ed., Chapter 12.

Brown P F, Lee C H and Spohr J C (1983), "Bayesian Adaptation in Speech Recognition," ICASSP'83, pp. 761-764.

Choi H C, King R W (1994), " Speaker Adaptation through Spectral Transformation for HMM based Speech Recognition," Intl. Symp. on Speech, Image Processing and Neural Networks (ISSIPNN'94), pp. 686-689.

Choukri K, Chollet G and Grenier Y (1986), "Spectral transformations through cannonical correlation analysis for speaker adaptation inASR," ICASSP'86, pp. 2659-2552.

Fukuzawa K, Komori Y, Sawai H and Sugiyama M (1992), "A segment-based speaker adaptation neural network applied to continuous speech recognition," ICASSP'92, pp. 433-437.

Gurgen F S (1992), "Phoneme recognition neural networks," Intl Symp. on Computer and Information Sciences VII (ISCIS VII)," pp 569-572.

Huang X (1992), "Speaker Normalization for Speech Recognition," IEEE Intl Conf. on Acoustics, Speech and Signal Processing (ICASSP'92), pp. 465-468.

Hush D R, Horne B G (1993), "Progress in Supervised Neural Networks," IEEE Signal Processing Magazine, pp. 8-39, Jan.

Lee C H, Lin C H, Juang B H (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.

Lippmann R. P. (1991), " Review of Neural Networks for Speech Recognition," Reading in Speech Recognition Edted by Alex Waibel& Kai Fu Lee, Morgan Kaufmann Publishers, Inc. pp. 374-392.

Rabiner L, Lee C and Wilpon J (1989), "HMM Clustering for Connected Word Recognition," ICASSP'89, pp. 405-408.