Two-Stage Feature Compensation Of Clean And Telephone Speech Signals Employing Bidirectional Neural Network

Iman Esmaili[1], Mansour Vali[1], Jahanshah Kabudian[2]

[1]Shahed University
[2] Research Center for Intelligent Signal Processing (RCISP)
{ iesmaili, vali}@shahed.ac.ir, kabudian@rcisp.ac.ir

ABSTRACT

In this paper, we continue our previous work on nonlinear feature compensation of distortions in clean and telephone speech recognition systems. We have shown that Bidirectional Neural Network (Bidi-NN) can compensate nonlinearly-distorted components of feature vectors. In this study, we present a new effort to improve recognition accuracy on clean and telephone speech data by employing a two-stage feature compensation technique for recovering optimal (from a classification point of view) Log-Filter Bank Energies (LFBE). These new features are achieved by training a new Bidi-NN with compensated features and considering compensated feature as the input data to Bidi-NN. We also achieved MFCC features by applying discrete cosine transform (DCT) to compensated Log-Filter Bank Energies (LFBE) features. HMM phone models are trained on these modified features. By using the two-stage compensated features, we obtained an absolute improvement of 4.73% and 9.29% in phone recognition accuracy compared to baseline system in clean and telephone conditions respectively. We also obtained an absolute improvement of 25.67% in phone recognition accuracy for the system which was trained on clean data but tested on telephone data. These results show excellency of NN-based nonlinear compensation of speech feature vectors in HMM-based speech recognition systems.

Index Terms—Bidirectional neural network (Bidi-NN), hidden markov model, robust speech recognition, telephone speech recognition.

1. INTRODUCTION

The recent progress of automatic speech recognition (ASR) systems shows high speech recognition accuracy under controlled conditions. However, the performance degrades when the recognizer deployed under the environments mismatched to the training environment. The degradation of the performance of the ASR system is due mainly to the ambient noise and transmission channel.

In principle, robust ASR can be achieved in four different ways (and combinations of them): (1) By appropriate extraction of robust features in the front-end [1], (2) By transformation of noisy speech features to clean speech features [2], (3) By adaptation of the references towards the current environment [3], or (4) By including noisy and/or distorted references in the training database [4]. Making use of robust ASR approaches, the recognition performance for the addressed type of degradation (namely the one which was taken into account in the development of the ASR system) can be improved. Unfortunately, it cannot be guaranteed that a recognizer will perform similarly well for new types of degradations, or for combinations of them [5].

Missing-feature technique has been known to be effective in improving speech recognition performance in additive background noise conditions [6]. This method depends mostly on the characteristics of speech that are resistant to noise, rather than on the characteristics of the noise itself. The missing feature method consists of two steps. The first step is estimation of a "mask" which determines which spectral parts of the noisy input speech are unreliable [7]. The second step is to reconstruct the unreliable regions or bypass them for other processing.

In telephone speech, the lower (0-125Hz) and the higher (3400-8000Hz) bands are completely lost and thus, might be considered as wide-band speech with missing data in these regions. What makes missing data techniques not to be applicable to telephone speech recognition is the fact that all frames are identically corrupted and there are no uncorrupted frames available to provide estimation for the missing bands.

In our previous work [8], we proposed a new algorithm for training the Bidirectional neural network (Bidi-NN). In this method, input feature vectors were recursively enhanced to achieve higher recognition accuracy. In addition, the missing components in telephone speech feature vectors were estimated according to the latent knowledge in the hidden layer of the neural network. This knowledge is obtained by training with clean and telephone speech simultaneously and is mostly induced by phonemic content and less influenced by the irrelevant variations in speech signal. An approach for reconstructing missing features in band-limited speech is suggested in [9] by considering the correlation between reliable components and missed components, however in our work missed components are reconstructed not only with the features of band-limited data but also with the latent knowledge which is obtained from training of Bidi-NN.

We showed in [8] that the mentioned techniques are very successful for neural network based ASR system. However, Hidden Markov Model (HMM) is the most popular and the most successful tool for speech recognition [10], thus we continue this work by training phone based HMMs on the enhanced features. We trained a new Bidi-NN with compensated feature vectors and achieved two-stage compensated feature vectors by considering previous compensated feature vectors as the input data to the Bidi-NN. We also achieved MFCCs by applying DCT to the LFBE features. Results show that employing each of these new steps leaded to an ASR system superior to that of our previous work.

The paper is organized as follows: Section two describes structure of neural network for feature compensation. Section three shows block diagram of our speech recognition system. In section four, the details of experiments are presented, and finally section five concludes this paper.

2. BIDIRECTIONAL NEURAL NETWORK

As shown in Fig. 1, Bidi-NN comprises of two parts: an MLP neural network, and a feedback path from the hidden layer to the input layer.

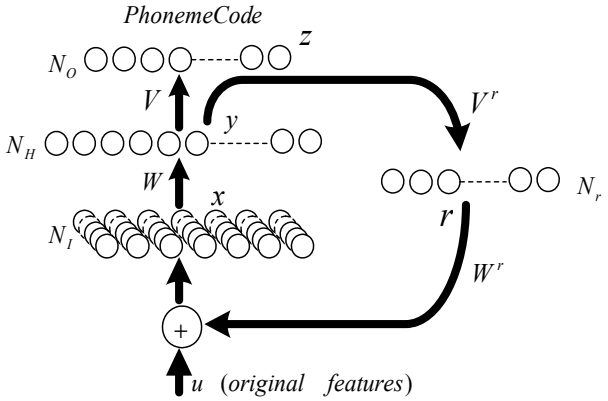

Fig. 1. Bidi-NN strucrure for compensation of input speech feature

N_I, N_H, N_r, N_O are the number of input units, feed-forward hidden layer units, feedback hidden layer units, and output units respectively. u is the representation vector (original speech feature vector) , r is the feedback hidden layer output vector, W and V are the weight matrices, x is the input to the feed-forward hidden layer, z is the output vector and y is the output of the feed-forward hidden layer.

We used Log-Filter Bank Energies (LFBE) spectral parameters as representation vector of speech signal. The LFBE spectral parameter is logarithm of energies of a Mel-scaled filter bank. In case of telephone speech signal, some filter bank values (0-125Hz and 3400-8000Hz) are missed. Since Bidi-NN is trained with both clean and telephone speech features simultaneously, all feature vectors had to be of the same dimension. Therefore, the missing parameters in telephone feature vectors were set to zero. A complete description of training algorithm was described in [8], thus we will briefly review how the Bidi-NN works.

In each epoch (e.g. n^{th} epoch), the Bidi-NN is trained with all frames (clean and telephone data) and the weights are modified according to the training algorithm. The input speech feature vector is modified like as follows:

$$x_i[n] = \lambda u_i + \sum_{l=0}^{N_r} r_l[n] w_{li}^r \qquad i = 1,2,...,N_I \qquad (1)$$

It is shown in Eq. (1) that a fraction of original speech feature vector (representation vector), i.e. λu (where $0 < \lambda < 1$) is summed up with a linear function of feedback hidden layer output values r, to form the modified input feature vector in n^{th} epoch. In the first epoch ($n = 1$), the input is considered to be $x_i[1] = u$.

The feedback path in Bidi-NN was intended to modify the corrupted or missing components of input feature vectors according to the latent knowledge in the feed-forward hidden layer. The latent knowledge in the hidden layer is mostly induced by the phonemic content of the speech signal, and most variations of the input features that are not beneficial for phoneme recognition have been discarded. In the feedback path, modified components corresponding to each of the input feature vector elements are estimated by means of a nonlinear function and are then summed up with corresponding components in feature vectors. The feature vectors are logarithms of the energies of Mel-scaled filter bank. Thus, the aforementioned summation is equivalent to multiplying each of the filter bank energies by a compensation term which is adaptively estimated by the feedback path.

3. THE ASR SYSTEM

The main blocks of our ASR system are shown in Fig. 2. The Bidi-NN is trained with both telephone and clean feature vectors (LFBE spectral parameters). After the convergence of the neural network, original features were iteratively exposed to the network for $n = 1, 2, \ldots, N$ where n is the number of epochs. After the first epoch, the values of all units of the feed-forward hidden layer were obtained for the whole test set. These values were used to modify the input feature vectors in the next epoch, n = 2. Results showed that we have no real gain in performance for $n > 3$ [8]. In this step, the feature vectors will be improved, and in case of telephone speech signal, missing features will be reconstructed too. These compensated features will be used for training the HMM phone models.

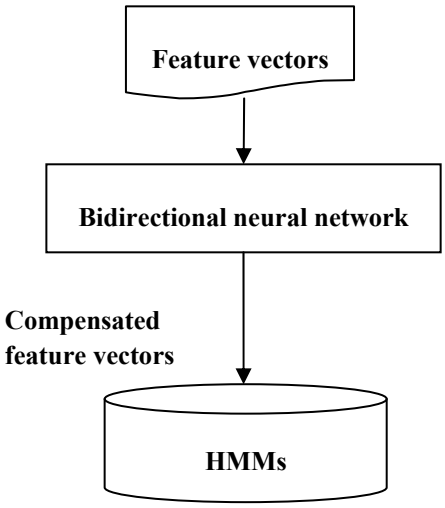

Fig. 2. Block diagram of our ASR system. The reconstructed feature vectors are obtained from the Bidi-NN. These modified features will be used for training the HMM phone models.

4. EXPERIMENTS

4. 1. Database

We have employed out techniques in speaker-independent phone recognition for Persian (Farsi) spoken language. Four hundred sentences of the standard Farsi phonetically-balanced continuous speech database *FarsDat* [12] which uttered by 200 speakers were chosen as our clean speech database. In addition, 128 sentences uttered by 64 different speakers of Telephone FarsDat (*TFarsDat*)

[11] database were used as our telephone speech database. Telephone handsets and transmission lines were different for different utterances, thus the variations of transmission conditions were taken into consideration. The training and the test sets were 3/4 and 1/4 of the whole clean and telephone speech databases. Our phonemes set consist of 34 context independent phones (including silence). Sampling rates of clean and telephone speech signals were 16 kHz and 8 kHz respectively.

4.2. Feature Extraction

In order to obtain LFBE parameters, 18 Mel-scaled filters were used for the 0-8 kHz bandwidth of clean speech. Telephone speech signal is within 125-3400 Hz and only 13 filters (2^{nd} to 14^{th}) were covered. Therefore, 18 parameters for clean speech and 13 parameters for telephone speech were obtained. By incorporating delta and acceleration coefficients, each feature vector of clean speech included 54 parameters, while for telephone speech it consisted of 39 parameters. Since all feature vectors had to be of the same dimension, the missing parameters in telephone feature vector (1^{st}, 14^{th}, 15^{th}, 16^{th}, 17^{th}, 18^{th} and their delta and acceleration) were set to zero. These missed parameters will be recovered by the Bidi-NN.

4.3. Bidirectional Neural Network (Bidi-NN)

A Bidi-NN as shown in Fig. 1 was designed as feature compensator. The input to the Bidi-NN consisted of 7 consecutive speech vectors (the current frame, along with 3 preceding and 3 succeeding frames), i.e. 7×54=378 input units. The neural network consisted of one hidden layer with 100 hidden units. The outputs were 34 units, corresponding to the number of Persian phones that we have defined.

4.4. Hidden Markov Models

HTK tools are used for training and testing the HMM phone models [13]. Thirty-four phone models were trained. The HMM phone models have 3 states and 16 diagonal-covariance Gaussian PDFs per state.

4. 5. Evaluation

To evaluate the ASR system, we trained the HMM phone models with LFBE feature vectors and considered it as baseline of clean speech experiments. According to Fig. 2, the Bidi-NN was simultaneously trained with both clean and telephone speech signals. We achieved compensated feature vectors by considering the original feature vectors as the input data to the Bidi-NN.
To test the maximum strength of Bidi-NN in feature compensation, we trained another Bidi-NN with compensated feature vectors and achieved two-stage compensated feature vectors by considering the compensated feature vectors as the input data to the Bidi-NN. Two different HMM phone models sets are trained with one-stage compensated and two-stage compensated feature vectors. We also achieved MFCC feature vectors by applying DCT to LFBE features. We performed all above-mentioned experiments for MFCC feature vectors too. As shown in Fig. 3, in LFBE-based systems, the recognition accuracy was improved 3.97% for one-stage compensated and 5.26% for two-stage compensated feature vectors compared to the baseline. In MFCC-based systems, recognition accuracy was improved 3.36% for one-stage compensated and 4.73% for two-stage compensated feature vectors compared to the baseline. This improvements show the ability of Bidi-NN in reduction of irrelevant variations (e.g. speaker variation, channel variation, etc) from speech signals.

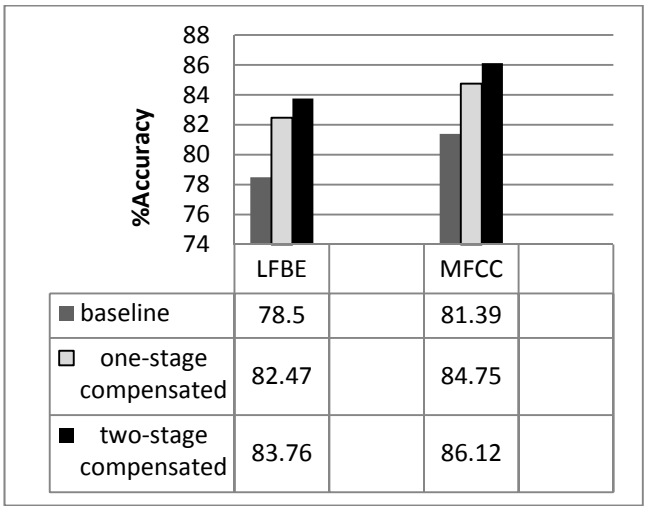

Fig. 3. Phone recognition accuracy for baseline, one-stage compensation and two-stage compensation of feature vectors for LFBE and MFCC feature vectors in clean speech database.

We performed the same experiment for telephone speech signals and achieved one-stage compensated and two-satge compensated feature vectors for LFBE and MFCC features. As shown in Fig. 4, in LFBE-based systems, the recognition accuracy was improved 7.05% for one-stage compensated and 8.17% for two-stage compensated feature vectors compared to the baseline. In MFCC-based systems, recognition accuracy was improved 6.07% for one-stage compensated and 9.29% for two-stage compensated feature vectors compared to the baseline. As the results show, the MFCC-based systems outperform the LFBE-based ones. This is true for both telephone and clean speech experiments.
We have also performed another experiment to show the ability of Bidi-NN as feature compensation system. We trained four different types of models:
A: models trained with non-compensated clean speech features and tested with non-compensated telephone speech features.
B: models trained with one-stage compensated clean speech features and tested with one-stage compensated telephone speech features.
C: models trained with two-stage compensated clean speech features and tested with two-stage compensated telephone speech features.
D: models trained with non-compensated telephone speech features and tested with non-compensated telephone speech features.
Since the MFCC has better results compared to LFBE, this experiment performed on MFCC-based systems. Table 1 shows the recognition accuracy of these four models. The difference in performance between model A and model D shows the impact of the mismatch and the need for robust methods. One-stage

compensated feature based system (B) shows significant improvement compared to non-compensated system (A). The results for two-stage compensated system (C) show 25.67% absolute improvement in recognition accuracy compared to non-compensated system and it is close to the system that trained with telephone speech data. The results obviously show that the Bidi-NN successfully reconstructs the missing components of telephone feature vectors.

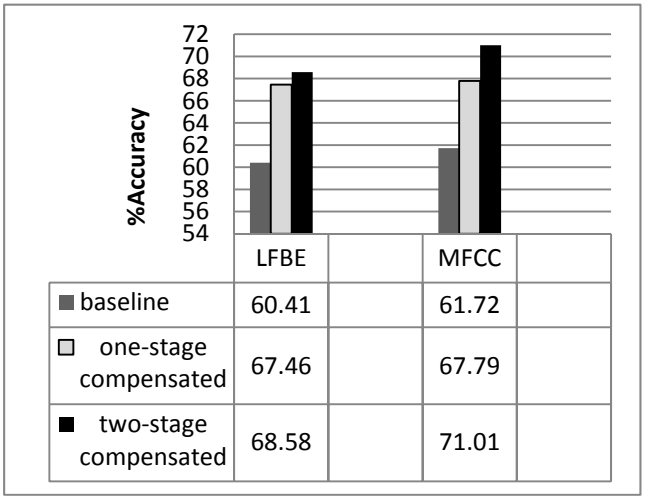

Fig. 4. Phone recognition accuracy for baseline, one-stage compensation and two-stage compensation of feature vectors for LFBE and MFCC feature vectors in telephone speech database.

Table 1. Phone recognition accuracies of models A, B, C and D.

Models	Phone recognition accuracy (%)
A	30.98
B	53.31
C	56.65
D	61.72

5. CONCLUSIONS

Robust ASR systems usually perform well for the degradation which was taken into account in the development of the system but it cannot be guaranteed that a recognizer will perform similarly well for new types of degradations, or for combinations of them.

In this paper, we introduced a flexible model that is able to cope with a variety of distortions. We achieved one-stage compensated feature vectors by considering the original feature vectors as the input data to the Bidi-NN. We also achieved two-stage compensated feature vectors by training a new Bidi-NN with one-stage compensated feature vectors and considering the one-stage compensated feature vectors as the input data to the Bidi-NN. We obtained an absolute improvement of 4.73% and 9.29% in phone recognition accuracy compared to baseline system in clean and telephone data respectively. The results show that the MFCC-based systems outperform the LFBE-based ones.

To show the ability of our system we also performed another experiment. In this experiment, systems which were trained with clean speech data were tested with telephone speech data. The recognition accuracy was improved 25.67% for two-stage compensated feature vectors compared to the system with no compensation. Results corroborate our hypothesis about the ability of Bidi-NN in nonlinear feature compensation of clean and telephone speech signals.

6. ACKNOWLEDGEMENT

This research was supported by Iran Telecommunication Research Center (ITRC) under contract T-500-3133.

7. REFRENCES

[1] B. Kotnik, D. Vlaj B. Horvat, "Efficient Noise Robust Feature Extraction Algorithms for Distributed Speech Recognition (DSR) Systems," *International Journal of Speech Technology*, 6, pp 205–219, 2003.

[2] N. Morales, D.T. Toledano, J.H.L. Hansen, J. Colás, J. Garrido, "Statistical Class-based MFCC Enhancement of Filtered and Band-Limited Speech for Robust ASR," *Proc. InterSpeech*, pp. 2629–2632, Portugal, Sep. 2005.

[3] H.G. Hirsch, "HMM Adaptation for Applications in Telecommunication," *Speech Communication*, vol. 34, pp 127-139, 2001.

[4] M.L. Seltzer, A. Acero, "Training Wideband Acoustic Models Using Mixed-Bandwidth Training Data for Speech Recognition," *IEEE Trans. Speech and Audio Processing*, vol. 15, no. 1, Jan. 2007.

[5] S. Moller, *Quality of Telephone-Based Spoken Dialogue Ssystems*, Springer Science plus Business Media Inc., USA, 2005.

[6] B. Raj, M. L. Seltzer, R. M. Stern, "Reconstruction of Missing Features for Robust Speech Recognition," *Speech Communication.*, vol. 43, no. 4, pp. 275–296, 2004.

[7] W. Kim, R.M. Stern, "Band-Independent Mask Estimation for Missing-Feature Reconstruction in the Presence of Unknown Background Noise," *Proc. IEEE ICASSP,* pp. 305–308, France, May 2006.

[8] M. Vali. , S.A. Seyyed Salehi. K. Karimi, "Robust Speech Recognition by Modifying Clean and Telephone Feature Vectors Using Bidirectional Neural Network," *Proc. InterSpeech*, pp. 2072-2075, USA, Sep 2006.

[9] W. Kim and J. H. L Hansen, "Time–Frequency Correlation-Based Missing-Feature Reconstruction for Robust Speech Recognition in Band-Restricted Conditions," *IEEE Trans. Speech and Audio Process*, vol. 17, no. 7, September 2009.

[10] L.R. Rabiner, B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[11] M. Bijankhan, J. Sheykhzadegan, M.R. Roohani, R. Zarrintare, S.Z. Ghasemi, M.E. Ghasedi, "TFarsDat – The Telephone Farsi Speech Database," *Proc. EuroSpeech*, Geneva, Switzerland, 2003.

[12] M. Bijankhan, J. Sheikhzadegan, M.R. Roohani, Y. Samareh, C. Lucas, M. Tebyani, "FarsDat – The Speech Database of Farsi Spoken Language," *Proc. 5th Australian Int. Conf. Speech Science and Technology (SST)*, pp. 826-831, Perth, Australia, 1994.

[13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, *The HTK Book (version 3.4)*,.Cambridge Univ. Eng. Dept., Cambridge, U.K. 2009.