Robust Speech Recognition by Modifying Clean and Telephone Feature Vectors Using Bidirectional Neural Network

Mansoor Vali , Seyyed Ali Seyyed Salehi, Kazem Karimi

Departement of Biomedical Engineering,
Amirkabir University of Technology, Tehran, Iran.
m_vali@bme.aut.ac.ir

Abstract

In this paper we present a new method for nonlinear compensation of distortions, e.g. channel effects and additive noise, in clean and telephone speech recognition. A Bidirectional Neural Network (Bidi-NN) was developed and implemented in order to modify distorted input feature vectors and improve the overall recognition accuracy. Distorted components in feature vectors were estimated in accordance with 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 MLP neural network was trained with these modified feature vectors. Comparing the achieved results with a reference model that was trained with unmodified feature vectors, demonstrate significant improvement in clean and telephone speech recognition accuracy.

Index Terms: robust speech recognition, bidirectional neural network, clean, telephone, feature vectors, variations.

1. Introduction

Noise robustness is a major problem that still remains unresolved in today's speech recognition technology [1]. Many automatic speech recognition systems can achieve high accuracies in well-constrained conditions. However, when there are variations in training and test speech, significant degradation might be observed in performance [2].

Different techniques to overcome these variations fall in two major categories: 1) the model-domain approach, where the speech models in the recognition system are modified or adapted in order to match the statistical characteristics of the distorted noisy test speech; and 2) the feature-domain approach, where the noisy test speech (possibly the "noisy" training speech as well) is modified or improved towards clean speech, without altering the recognition models.

The speech recognition over telephone network is a widespread application. Basically, the distortion sources in telephone network come from two classes: (1) noise contamination including background noise and electrical noise, and (2) channel effect caused by telephone handset and transmission line. Due to these distortion sources, the speech recognition performance will be seriously damaged. In the literature [3], many algorithms have been proposed for compensating the noise effect. However, it is not adequate for overcoming the mismatch problem by only considering the noise effect. Accordingly, the RASTA method [4] was presented for reducing the variability of additive noise and channel effect. Besides, the Signal Bias Removal (SBR) [5], Stochastic Matching (SM) [6] and Channel-Effect-Cancellation [7] methods were also successfully applied for telephone speech recognition.

Missing data approaches have the potential to provide highly robust recognition for speech corrupted by high levels of additive noise and make minimal assumptions about the nature of the noise. They are based on identifying uncorrupted, reliable regions in the frequency domain and adapting recognition algorithms so that classification is based on these regions [8]. The corrupted regions may be either neglected in the process of recognition, or estimated by imputation.

Classification with incomplete patterns generally incorporates missing data techniques during classification whereas learning is accomplished with complete data. This approach is very successful in robust ASR and has been implemented in CDHMM based ASR systems [9] and Neural networks based ASR systems [10].

In telephone speech, the lower (0 to 125Hz) and the higher (3500 to 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 (in comparison with previous work had been done in [10] for noisy speech), is the fact that all frames are identically corrupted and there are no uncorrupted frames available to provide estimation for the missing bands. To overcome this problem, we used clean speech in conjunction with telephone speech and a Bidi-NN was trained with both data simultaneously. Using a feedback branch with a delay in epochs of training and test process of Bidi-NN, input feature vectors were recursively modified to achieve higher recognition accuracy. In addition, the missing components in telephone speech feature vectors were estimated in accordance with the latent knowledge in the hidden layer of the neural network. This knowledge is obtained by simultaneous training with clean and telephone speech and is mostly induced by phonemic content and less influenced by the irrelevant variations in speech signal.

The newly acquired modified feature vectors were then used to train an MLP-based recognition model. Finally, the recognition accuracy of this model would be compared with a reference model that was exclusively trained with original feature vectors to quantify the improvement in phoneme recognition accuracy.

2. Database

Two sentences from FARSDAT database [11] uttered by 200 speakers were chosen for our clean speech database. In addition, two sentences uttered by 64 different speakers of TELEPHONIC FARSDAT database were used for our telephone speech. Telephone handsets and transmission lines were different for different utterances. Thus, the variations of environmental conditions were taken into consideration. 75% of clean and telephone utterances were used for training and the rest were kept for test. These two long sentences consist of all Persian (Farsi) phonemes, and therefore provide a reliable deduction for our context-dependent study. Our phonemes set consist of 33 context independent phonemes, as well as silence. Sampling rates of clean and telephone speech signals were 16 kHz and 8 kHz respectively.

3. Feature vectors

Probably the most popular feature extraction method for clean speech is using Mel-Frequency Cepstral Coefficient (MFCC) parameters [12]. In presence of environmental variations, MFCC may not be an

appropriate choice in that it is excessively sensitive to the effect of communication channels and noise. Logarithms of Filter Bank Energies (LFBE), what was perceived in our previous study [13] are logarithms of energies of a filter bank in mel-scale, while MFCC parameters are obtained by performing DCT on the same LFBE parameters. Thus, if communication channel or noise, distorts some parts of the spectrum, in case of LFBE, only the components that correspond to those parts are distorted, while in case of MFCC, performing DCT will spread this distortion to the whole parameters. Therefore, it seems to be more relevant to choose LFBE feature vectors for models which are to be simultaneously trained with both clean and telephone speech.

3.1. Feature extraction method

In order to obtain LFBE parameters, a filter bank of 18 was assigned for the 0-8 kHz bandwidth of clean speech. Telephone speech signal is within 125 to 3700 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 MLP neural networks were trained simultaneously with both clean and telephone speech features, all feature vectors had to be of the same dimension. Therefore, the missing parameters in telephone feature vectors were assigned zero values.

The normalization method used in this study was similar to Cepstral Mean Substraction (CMS) method [14], i.e. for each utterance, the overall mean value was subtracted from feature vectors and the result was divided by the standard deviation.

4. Reference speech recognition model

An MLP neural network was designed as a reference model for clean and telephone speech recognition. This model, shown in Fig. 1, was solely intended for phoneme classification of feature vectors. The neural network consisted of one hidden layer with 100 units. The number of hidden units was chosen in a way that the number of training patterns were 4 to 10 times the number of neural network weights [15]. Going below 4 times would degrade the generalization capability of the network and going over 10 times would result in enormous networks that were cumbersome to train.

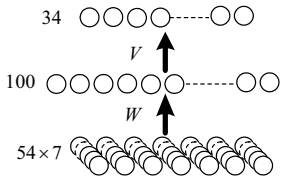

Figure 1. *Structure of reference model*

The input to the MLP consisted of 7 time frames (the frame that had to be trained or classified, along with 3 preceding and 3 succeeding frames), i.e. 7×54=378 input units. The outputs were 34 units, corresponding to the number of Persian phones that we had defined. The activation function used for all neurons was the tangent hyperbolic function and for efficient training the target values were set to -0.9 or 0.9.

For 10 times, the network was trained with the whole training set, including both clean and telephone speech; each time with different initial random weights. The average recognition accuracies for clean and telephone speech on the test set were 84.1% ± 0.2 and 69.7% ± 0.1, respectively.

The chart in Fig. 2 shows the obtained recognition accuracies for 6 categories of phonemes. Obviously, 1) In both clean and telephone speech, vowels had the highest recognition accuracy. 2) Telephone channels mostly damaged recognition accuracies for plosive and affricative phonemes.

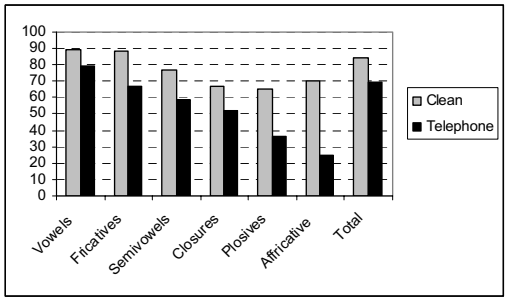

Figure 2. *Categorized recognition accuracies for reference model*

5. Bidirectional neural network

5.1. Structure

As shown in Fig. 3, Bidi-NN consists of a network identical to the reference model MLP, and a feedback branch from the hidden layer to the input layer. This feedback branch consists of a hidden layer of 40 units with tangent hyperbolic activation function, and full-connected W^r and V^r weights. The output of the feedback branch is a vector of N_I elements, which is combined with corresponding components in the input layer as in Eq. (1). N_I is the number of input units of Bidi-NN.

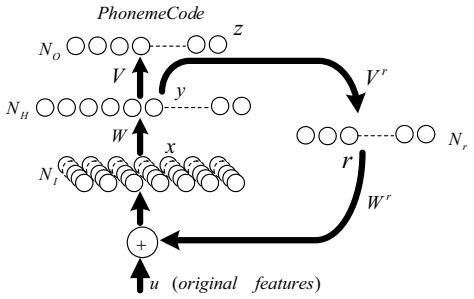

Figure 3. Structure of Bidi-NN network for modifying input feature vectors

The feedback branch in Bidi-NN was intended to modify the corrupted or missing components in input feature vectors, according to the knowledge, latent in the feed-forward hidden layer. The reason is that, the feed-forward network is more capable in learning the well correlated and outnumbered clean speech features compared to the more corrupted and dispersed telephone speech features. In addition, the knowledge latent in the hidden layer is mostly induced by the phonemic content of speech signal, and most variations in input features that are not beneficial for phoneme recognition have been discarded. Therefore, in order to improve or modify the input feature vectors, it is best to use the information in this layer.

In the feedback branch, modified components for each of the input feature vectors are estimated by means of a nonlinear function and are then summed up with correspondent components in feature vectors.

5.2. Training algorithm

The Bidi-NN and the reference MLP were trained similarly, with utilizing back propagation method in order to reduce the gradient of error. Thus, feed-forward and feedback weights were modified

in a manner to learn the phonemic content of feature vectors. The training algorithm was different from a common MLP, in that the inputs to the Bidi-NN were combined from the original feature vectors and the modified vectors. Modified vectors were obtained from the hidden layer values in the previous epoch. The complete training algorithm is expressed in Eq.s (1) to (13).

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

$$r_l[n] = f(\sum_{j=0}^{N_H} y_j[n-1] v_{jl}^r) \qquad l = 1,2,...,N_r \tag{2}$$

$$y_j[n] = f(\sum_{i=0}^{N_I} x_i[n] w_{ij}) \qquad j = 1,2,...,N_H \tag{3}$$

$$z_k[n] = f(\sum_{j=0}^{N_H} y_j[n] v_{jk}) \qquad k = 1,2,...,N_O \tag{4}$$

$$E = \frac{1}{N_O} \sqrt{\sum_{k=1}^{N_O} (z_k[n] - d_k)^2} \tag{5}$$

$$\delta_k^z[n] = (z_k[n] - d_k) \dot{z}_k[n] \tag{6}$$

$$\delta_j^y[n] = (\sum_{k=1}^{N_O} \delta_k^z[n] v_{jk} + \sum_{k=1}^{N_r} \delta_k^r[n] v_{jk}^r) \dot{y}_j[n] \tag{7}$$

$$\delta_i^x[n] = \sum_{j=1}^{N_H} \delta_j^y[n] w_{ij} \tag{8}$$

$$\delta_l^r[n] = (\sum_{i=1}^{N_I} \delta_i^x[n] w_{li}^r) \dot{r}_l[n] \tag{9}$$

$$v_{jk} = \alpha v_{jk} + \eta \delta_k^z[n] y_j[n] \tag{10}$$

$$w_{ij} = \alpha w_{ij} + \eta \delta_j^y[n] x_i[n] \tag{11}$$

$$w_{ij}^r = \alpha w_{ij}^r + \eta \delta_i^x[n] r_l[n] \tag{12}$$

$$v_{jl}^r = \alpha v_{jl}^r + \eta \delta_l^r[n] y_j[n-1] \tag{13}$$

In each epoch (e.g. n^{th} epoch), the Bidi-NN is trained with all frames and the weights are modified according to Eq. (1) to (13). 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. It is shown in Eq. (1) that a fraction of representation vector u (where $0 < \lambda < 1$) is summed up with a linear function of feedback hidden layer values r, to form the input feature vector in n^{th} epoch. In the first epoch ($n = 1$) the input is considered to be $x_i[1] = u$. According to Eq. (2), the values for feedback hidden layer are obtained as a nonlinear function of feed-forward hidden layer values (y) for epoch $n–1$.

W^r and V^r in Eq. (1) and Eq. (2) are respectively the weight matrices for feedback branch in Bidi-NN. Obviously, the input to the network for each epoch is the summation of a fraction of original feature vector (λu) and a nonlinear function of y values obtained from the previous epoch. Thus, it is necessary for the whole training set, that the y values from the previous epoch are recorded to be used in the next epoch.

Feed-forward hidden layer values (y) and network's outputs (z) are also obtained from Eq. (3) and Eq. (4), were W and V are the weight matrices for the first and the second layer of the feed-forward network, respectively. Output error for each input feature vector, is obtained from Eq. (5), where d_k is the desired value of output units for that input feature vector. Hence, the gradient of error in different layers is calculated according to Eq.s (6) to (9). Finally, for each input feature vector, network weights are iteratively modified according to Eq.s (10) to (13). The training coefficient η is 0.001 and momentum coefficient α, is 0.2.

6. Experiments

In case of distorted signals affected by environment noise or channel effects (telephone speech), the derivative parameters will be definitely more distorted, which is resulted from high pass differential filter. Hence, we applied our algorithm to improve static features by means of the Bidi-NN. Afterwards, the correspondent delta and acceleration parameters were added to them to form 54-element feature vectors, which were later used for the speech recognition system.

λ parameter has a significant role in the training process. It varies between 0 and 1 and determines the ratio of original feature vectors that should be incorporated with predicted vectors (which are achieved from hidden layer values during the previous epoch) to result the improved feature vectors. Hence, when λ approaches zero, it means original feature vectors are less emphasized, and when it approaches 1, it means that the original feature vectors are more emphasized. Since in the telephone speech feature vectors, the out of frequency range elements were assigned zero values, these elements are predicted only from hidden layer values. This can be seen in Eq.(1) where the first term is zero and the second term is used to form these elements. Therefore, although changing the value of λ does not directly affect the improvement of zero elements, but it has drastic effect on reconstruction of the other non-zero elements of the feature vectors.

The best value for λ was determined empirically, i.e. by finding the highest accuracy for the test and training sets. Therefore, five distinct Bidi-NN were trained for five different values of λ: 0.3, 0.5, 0.6, 0.7 and 1. The Bidi-NNs were trained with both clean and telephone speech features, according to Eq.s (1) to (13). The number of units in different layers were: $N_I = 7 \times 18$, $N_H = 100$, $N_r = 40$, N_O=34.

After the convergence of all the networks, for each network test set was iteratively exposed to the network for n = 1, 2, …, N, according to Eq.s (1) to (4). Each time, the recognition accuracy was recorded. n = 1 means that only the feed-forward part of the Bidi-NN was used for recognition. After the first epoch, the values of all units in 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.

For all the aforementioned networks, the modification of test feature vectors has increased recognition accuracies. For instance, for the network with λ=0.6, recognition accuracies for N = 6 epochs are illustrated in Fig. 4, where improvements can be seen for both clean and telephone speech. After n = 3 epochs, the recognition accuracy did not change. Similar results were obtained for other Bidi-NNs. Recognition accuracies for clean and telephone speech after 3 epochs are shown in Table (1). The first row of the table shows the results for an MLP NN (similar to feedforward section of Bidi-NNs), which was trained with these 18-element feature vectors. The results for all 5 Bidi-NNs are better than the result of MLP NN.

We can conclude that the appropriate value of λ is somewhere between 0.5 and 1. In order to find the accurate value of this parameter, it is necessary to improve feature vectors for the whole clean and telephone speech in training and test sets according to Eq. (1). This procedure was performed for Bidi-NNs with λ = 0.5, 0.6, 0.7 and 1. We perceived that missing components in telephone feature vectors (with initial zero values) were estimated.

In order to evaluate the quality of these modified feature vectors in phoneme recognition, each feature vector was accompanied by its

correspondent delta and acceleration coefficients, and for every data set belonging to each λ, an MLP neural network identical to the reference model was trained with these newly acquired feature vectors. The recognition accuracies of these models for clean and telephone speech are illustrated in Table (2).

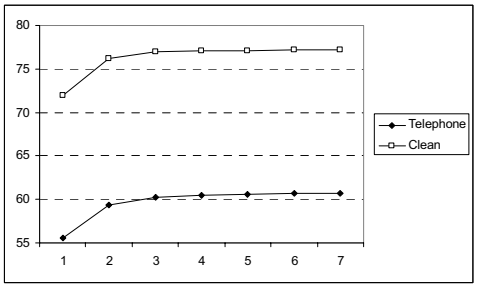

Figure 4. *Recognition accuracies for 6 epochs of improvement in test feature vectors* (λ=0.6)

Table 1. Recognition accuracies of phonemes for clean and telephone speech in Bidi_NNs

Neural Network	Clean speech	Telephone speech
MLP	76.9	59.4
Bidi-NN: λ =1	77.4	60.4
Bidi-NN: λ =0.7	77.4	60.5
Bidi-NN: λ =0.6	77.2	60.7
Bidi-NN: λ =0.5	76.9	60.3
Bidi-NN: λ =0.3	76.7	59.7

Table 2. Recognition accuracies of phonemes for improved Feature vectors

Neural Network	Clean speech	Telephone speech
MLP: Reference	84.1	69.7
MLP: λ =1	84.7	72.1
MLP: λ =0.7	85.4	72.7
MLP: λ =0.6	85.5	72.9
MLP: λ =0.5	84.9	72.1

Comparison of the results shows that the best accuracy rate was obtained for λ=0.6. The recognition accuracies of this model for clean and telephone speech were 85.5% and 72.9% respectively (Table 2, row 4). Thus, applying the Bidi-NN for modification of feature vectors resulted in 1.4% and 3.2% increase in the recognition accuracies for clean and telephone speech, respectively.

For a deeper understanding of the results, the increase in recognition accuracies for different categories of phonemes is shown in Fig. 5. Obviously, for telephone speech, the best results were achieved for Affricative and Plosive phonemes that were mostly affected by telephone channel. An advantage of Bidi-NN is that it increases the recognition accuracy for clean speech as well. This improvement is due to the reduction of a series of irrelevant variations, e.g. in speakers, microphones, etc.

7. Conclusions

In this study, we extracted logarithmic spectral feature vectors for clean and telephone speech, with identical dimensions and correspondent components. Mean Subtraction Normalization was used to reduce the linear channel effects on the feature vectors. Both clean and telephone speech feature vectors were simultaneously trained to an MLP-based recognition model as a reference model. Afterwards, 5 Bidirectional Neural Networks (Bidi-NN) - for 5 different values of λ - were trained with static features. By iteratively exposing feature vectors to Bidi-NNs, they were modified based on a nonlinear method to reduce the irrelevant variations and increase the phoneme recognition accuracy. By adding delta and acceleration coefficients to the modified feature vectors, several MLP neural networks identical to the reference model were trained with these 54-element feature vectors. The best result was achieved for λ=0.6 where the phoneme recognition accuracies of modified clean and telephone feature vectors were increased by 1.5% and 3.2%, respectively in comparison with the reference model.

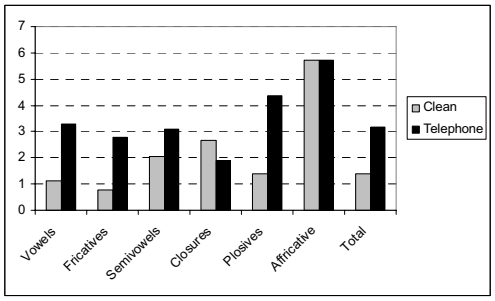

Figure 5. the increase in recognition accuracies for different categories of phonemes in a new MLP model trained with improved features

8. References

[1] Pearce, D. and Aalborg, Ed., "ESE2 Special Sessions on Noise Robust Recognition," in Proc. Eur. Conf. *Speech Communication*, Denmark, Sept. 2001.

[2] Furui, S., "Robust Methods in Automatic Speech Recognition and Understanding", *Proc. Eurospeech*, pp. 1993-1997, GENEVA, Switzerland, 2003.

[3] Gong, Y., "Speech Recognition in Noisy Environments: A Survey," *Speech Communication*, vol. 16, pp. 261-291, 1995.

[4] Hermansky, H. Morgan, N. Bayya, A. and Kohn, P., " Compensation for the Effect of the Communication Channel in Auditory-like Analysis of Speech (RASTA-PLP)," in *Proc. EUROSPEECH*, vol. 3, pp. 1367-1370, 1991.

[5] Rahim, M. G. and Juang, B. H., "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition," *IEEE Trans. Speech Audio Process*., vol. 4, pp. 19-30, 1996.

[6] Sankar, A. and Lee, C. H. "Robust Speech Recognition Based on Stochastic Matching," in *Proc. ICASSP*, pp. 121-124, 1995.

[7] Chien, J. T., Lee, L. M., and Wang, H. C. "A Channel-Effect Cancellation Method for Speech Recognition over Telephone System," *IEE Proc. Visual Image and Signal Processing*., vol. 142, no. 6, pp. 395-399, 1995.

[8] Parveen, S. and Green, P., " Speech Recognition with Missing Data Techniques using Recurrent Neural Networks," Neural Information Processing Systems 14, *MIT Press*, 2001.

[9] Barker, J. Josifovski, L. Cooke, M. P. and Green, P. D. "Soft Decisions in Missing Data Techniques for Robust Automatic Speech Recognition ", *ICSLP,* Beijing, China, 2000.

[10] Morris, A. et al. "A Neural Network for Classification with Incomplete Data: Application to Robust ASR, " *ICSLP,* Beijing, China, 2000.

[11] Bijankhan, M. et al, "FARSDAT-The Speech Database of Persian Spoken Language," SST-94, Perth, pp.826-831, 1994.

[12] Davis, S. B. and Mermelstein, P. "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. ASSP,* Vol. 28, pp. 357–366, 1980.

[13] Vali, M. and Seyyed Salehi, S. A., " Evaluation of MFCC and LFBE Features for Robust Clean and Telephone Speech Recognition," *10th Annual Computer Society of Iran Computer Conference,* February 2005.

[14] Jain, P. and Hermansky, H., "Improved Mean and Variance Normalization for Robust Speech Recognition, " *Proc. of ICASSP* , Salt Lake City, 2001.

[15] Blumer, A. Ehrenfeucht, A. Haussler, D. and Warmuth, M., "Learnability and the Vapnik-Chervo-Nenkis Dimension," J. Ass. Comput. Match., vol.36, no.4, pp. 929-965, 1989.