Robust Speech Recognition by Improvement Missing Features using Bidirectional Neural Network

Hojat Mohammadnejad
Engineering Faculty
Shahed University
Tehran, Iran
Mohammadnejad@shahed.ac.ir

Mansoor Vali
Engineering Faculty
Shahed University
Tehran, Iran
vali@shahed.ac.ir

Abstract— **In this paper we present a new method for nonlinear compensation of mismatches, e.g. additive noise, on clean and noisy speech recognition. We were inspired by the human recognition system in development and implementation of a new Bidirectional Neural Network (BNN). This procedure, results in improvement of input features and consequently increasing the overall recognition accuracy. The feedforward weights of this network are trained using both clean and noisy speech features. The results demonstrate significant improvements in clean and especially noisy speech recognition accuracy compared to reference model trained on unimproved features.**

Keywords- Bidirectioal Neural Netwrok; Multi Layer Perceptron; Speech recognition;

I. 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 mismatches between training and test speech, significant degradation might be observed in performance [2].

Two major classes of techniques to overcome these mismatches include: 1) the model-domain approach, where the speech models in the recognizer are modified or adapted to match the statistical properties of the unmodified noisy test speech; and 2) the feature-domain approach, where the noisy test speech (possibly the "noisy" training speech as well) is modified or enhanced to move toward clean speech as closely as possible.

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.

However, improvements obtained from these techniques are not significant compared to human performance. Considering the human auditory system in voice recognition introduces the new and interesting missing data method [4]. In this method primary recognition is accomplished by reliable areas and domains. Unreliable areas are corrected afterwards. This action is iterated until the recognition is completed and the primary pattern is corrected.

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 [5] and Neural networks based ASR systems [6].

For missing-feature methods to be practicable, the unreliable components must be identified without a priori knowledge of the true SNR of spectrographic elements. Conventionally this is done by maintaining a running estimate of the noise spectrum and using this to estimate which elements of the spectrogram are unreliable. This method has the disadvantage of requiring that the spectrum of the corrupting noise be estimated, a problem that is almost intractable when the noise is non-stationary.

Missing feature approaches are so sensitive to rate of true identification of reliable and unreliable components and because of that the most difficult aspect of missing feature methods is the estimation of the spectrographic masks that identify unreliable spectral components.

To overcome this problem, we will use a new method to remove this identification step from all compensation process. We used clean speech in conjunction with noisy speech that is simultaneously trained to a Bidirectional Neural Network. This network was inspired by computational concepts of human mind and having feed-back connections in addition to feed-forward connections is an effort to approach the performance, flexibility, robustness and the reliability of human speech recognition system.

With the proposed algorithm for training the BNN, the input feature vectors are iteratively improved to achieve higher recognition accuracy. In addition, the missing components in noisy speech feature vectors are estimated in accordance with the latent knowledge in the hidden layer of the Neural Network. This knowledge has the most information of phonemes and least information of other speech variations. This is the main ability of proposed method.

The newly acquired modified feature vectors are then used to train an MLP-based recognition model. Finally, the recognition accuracy of this model will be compared with a reference model which was exclusively trained with initial

978-1-4244-7484-4/10/$26.00 ©2010 IEEE

feature vectors to quantify the improvement in phonemes recognition accuracy.

II. Missing feature methods

Missing feature methods use Logarithmic Filter-Bank Energies (LFBE) features to identify unreliable components. It is assumed that all components of a spectrogram with SNR above some threshold T are reliable estimates of the corresponding components of clean speech and the others are unreliable and must be estimate.

In this methods the unreliable components of the spectrogram, are estimated based on the reliable components and the known statistical properties extracted from LFBE of clean speech corpus. Recognition can then be performed either with the complete LFBE so derived, or with cepstral coefficients derived from them.

III. Feature Vectors

As explained above, LFBE method has been used for feature extraction to achieve improved noisy features. In the other words, we used LFBE features to compensation system to get improved features. Briefly, LFBE features represent power spectral parameters of spectrogram, in other word LFBE features are time-frequency display of speech information. Noisy speeches have some corrupted component in time-frequency display of speech information so we can identify these components as missing LFBE features and try to modify them. If the LFBE parameters are transformed to MFCC features before modification process the noise inside noisy LFBE features are propagated to all parameters of MFCC in recognition vectors. In this study we used Bidirectional neural network to compensate missing LFBE features.

Currently, most of the automatic speech recognition systems make use of representation parameters based on Mel Frequency Cepstral Coefficients (MFCC) for train and test of recognition model [7]. It is well known, however, that Logarithmic Filter-Bank Energies features are a suboptimal feature domain for recognition and that cepstral coefficients derived from LFBE typically provide significantly greater recognition accuracy. In fact, in some cases using noisy cepstral coefficients results in higher recognition accuracy than the use of LFBE derived from clean speech.

IV. Feature Extraction Methods

In this article, we used two kinds of features:

1. The LFBE parameters, that are Logarithmic Filter-Bank Energies uniformly distributed in mel-frequency scale within the bandwidth of speech signal. For clean and noisy speech that the bandwidth of signals is within 0-16KHz, there exist 20 filters. Therefore, we have extracted 20 parameters for clean and noisy speech. By adding their delta and acceleration coefficients, feature vectors include 60 parameters. These features are used for Bidirectional neural network training to get improved LFBE features for both clean and noisy speech.

2. MFCC features, since we want to train recognition model to obtained overall recognition accuracy, LFBE features are transformed to MFCC feature vectors to increase recognition accuracy. Most speech recognizers preprocess incoming feature vectors in various ways in order to improve recognition accuracy. For example, recognizers usually use cepstral features derived from LFBE parameters, rather than the LFBE themselves, because cepstral vectors are known to result in significantly greater accuracy [8]. Finally, by using DCT transform, we have extracted 13 MFCC parameters for clean and noisy speech. By adding their delta and acceleration coefficients, feature vectors include 39 parameters that are used to train and test of MLP neural network.

Other common type of preprocessing of incoming feature vectors for improve recognition accuracy include mean normalization of feature vectors. The normalization method should be appropriate for the neural network structure used in this study (i.e. equalizing the dynamic range of all components of feature vectors). Thus, for each component of feature vectors, mean value is subtracted and the result is divided by the standard deviation [8].

V. Reference Speech recognition Model

We design an MLP neural network, as our reference model for clean and noisy speech recognition. The task of this model is limited to phoneme classification of feature vectors. This network, as is shown in Fig. 1, consists of one hidden layer with 100 hidden units. This number of hidden units is obtained from computation cost and relationship between amount of training data and number of weights [11]. The input to the MLP consists of 7 frames (current frame, 3 frames in the past and 3 frames in the future) i.e. 273 (7×39) inputs. The outputs are 34 neurons equal to number of Persian phones according to our definition. The activation function used for all neurons is a tangent hyperbolic function and for efficient training the target values are set to 0.9 or -0.9.

Training is based on gradient descent method and backpropagation algorithm and feature vectors of just clean speech that are trained for several times to the network. When the defined epochs in training procedure is finished, training of the network is stopped and test procedure with both clean and noisy speech is started.

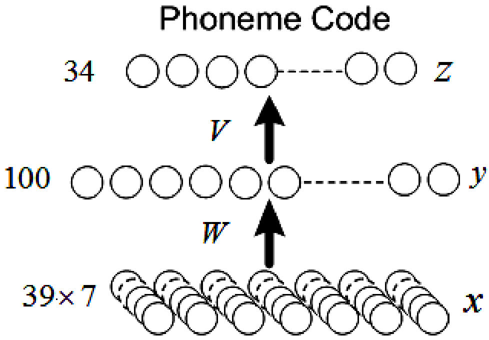

Figure 1. Structure of MLP reference model

VI. BIDIRECTIONAL NEURAL NETWORK

A. *Structure*

The structure of BNN network is as shown in Fig. 3. This network is consisted of a MLP neural network like the reference model and a new branch, connected between the hidden layers to the input layer. This feedback connection includes one hidden layer with 50 neurons with nonlinear hyperbolic tangent functions and full connected weights V_r and W_r .Connection of W_r weights to input of network are linear. The output of BNN like the MLP reference model is the binary presentation of the phonemes, corresponding to the input feature vector of middle frame.

Since the vector space of hidden layer, is constituted from training clean speech in conjunction with just one of assumed SNRs simultaneously, the knowledge of phonemes is latent in this layer due to training with clean speech. So this model is also capable of reconstruct the missing parts of noisy feature vectors regarding to this knowledge.

Training and test process of this network at the next section could be helpful to understand these capabilities of BNN network [9].

B. *Training algorithm*

Training of BNN network consists of two steps. At the first step, feedforward MLP neural network like reference model in this paper is trained and the weights of feedback are not trained.

At the second step, the feedback branch is added from hidden layer to input of feedforward network and training continues for feedback weights. The weights of feedback are trained based on gradient descent method and backpropagation algorithm, while the weights of feedforwaed network are fixed in second step. So errors of output are backpropagated through feedforward layers to improve only weights V_r and W_r .

The input of each iteration for all the training data set is computed and is dependent to hidden layer values, from previous iteration, So for each iteration we need to keep all previous values of hidden layer corresponding to training data set [9].

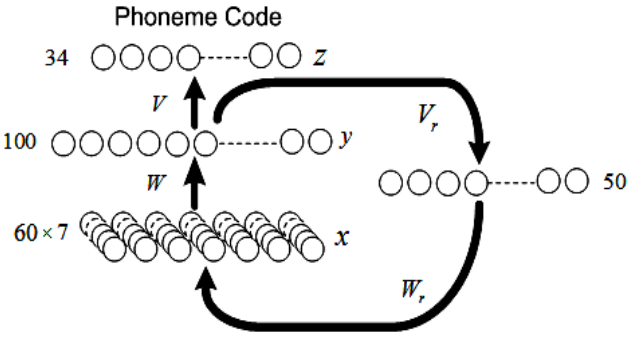

Figure 2. Structure of BNN network

VII. DATABASE

Two sentences of FARSDAT database uttered by 200 speakers are allocated for our clean speech database. Also corrupted mentioned clean speech database with additive white Gaussian noise by difference SNRs (20dB to 0dB), are used for our noisy speech database. %75 of clean and noisy speakers is considered for training and the rest are considered for the test. These two long sentences consist of all Persian (Farsi) phonemes, so the considered databases are adequate to achieve valid results. Our phonemic set consists of 34 context independent phonemes plus silence. Sampling rates of clean and noisy speech signals are 16 kHz.

VIII. EXPERIMENTS

In this experiment after training the feedforward weights of BNN model with both clean and noisy features, training continues for feedback weights with the same clean and noisy features. Note that to train a Bidirectional Neural Network we used clean speech in conjunction with just one of assumed SNRs in noisy speech database, simultaneously. Then we made improved feature vectors with trained BNN for all test speech databases such as clean speech or noisy speeches that have been corrupted in any condition of SNRs.

$BNN(i)$ refers to a Bidirectional Neural Network that is trained using clean speech in conjunction with noisy speech corrupted by $SNR = i \quad dB$ simultaneously, then all clean speech and noisy speech (corrupted by every SNR) are improved by trained $BNN(i)$.

TABLE 1 RECOGNITION ACCURACY WITH UNIMPROVED AND IMPROVED MFCC FEATURES

Recognition Accuracy	***Without Improve***	***BNN (0)***	***BNN (5)***	***BNN (10)***	***BNN (15)***	***BNN (20)***
Clean Speech	83.79%	85.24%	85.80%	85.71%	85.84%	85.74%
SNR-20	78.70%	81.53%	81.77%	82.30%	82.28%	81.73%
SNR-15	73.22%	77.67%	77.86%	78.57%	78.34%	77.08%
SNR-10	64.84%	71.27%	71.56%	72.31%	71.26%	69.77%
SNR-5	53.54%	61.43%	61.36%	61.70%	59.83%	58.66%
SNR-0	39.78%	49.65%	46.99%	48.79%	46.10%	45.62%

Afterwards an MLP neural network is trained based recognition model using the newly acquired modified feature vectors of clean speech. Finally, the recognition accuracy of this model will be compared with a reference model which was exclusively trained with initial feature vectors to quantify the improvement in phonemes recognition accuracy.

In test process for each noisy or clean speech in the input of MLP, recognition accuracy is computed from total accuracy for all phonemes without silence. Table 1 shows the recognition accuracy for reference model and modified features obtained by $BNN(i)$, where is $i = 0,5,10,15 \,\&\, 20$. This train and test processes applied for any condition of SNRs alternatively to get the goal that is possible to improve overall recognition accuracy with a BNN that is trained via clean speech in conjunction with just one of assumed SNRs in noisy speech database.

Also the obtained improvement rate using $BNN(i)$ where is $i = 0,5,10,15 \,\&\, 20$ for any corrupted speech by every SNRs has been represented in Fig. 3. Also the average of these improvement rates display in Table 2.

The results as are shown in Table 2 and Fig. 3, demonstrate the high improvement rates in recognition accuracy compared to baseline accuracy. As the result according to Table 1, the best improvement in recognition accuracy compared to baseline accuracy is obtained by $BNN(10)$ that is refers to a Bidirectional Neural Network that is trained using clean speech in conjunction with noisy speech corrupted by $SNR = 10 \quad dB$ simultaneously, then all clean speech and noisy speech (corrupted by every SNR) are improved by trained $BNN(10)$.

IX. Discussion

The results of experiments demonstrate capability of BNN to correct the distorted parts of logarithmic spectral features of noisy and clean speech data due to additive noise effects without thinking about missing parts as reliable or unreliable. This is caused by nonlinear processing in neural network layers and creating attraction domains.

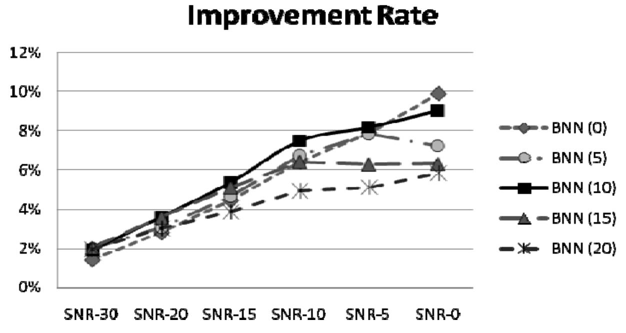

Figure 3 Improvement Rate obtained with BNN(i). where is $i = 0,5,...,20$

TABLE 2 AVARAGE OF IMPROVEMENT IN RECOGNITION ACCURACY

Average of Improvement	***BNN (0)***	***BNN (5)***	***BNN (10)***	***BNN (15)***	***BNN (20)***
	5.49%	5.24%	5.92%	4.96%	4.12%

We have shown in Fig. 3 improvement rates of recognition accuracy for each condition of SNRs. The best improvements are 9.01% when the noisy speech features that are corrupted by SNR=0 dB is improved by $BNN(10)$ that is demonstrated by square symbols in Fig. 3.

X. Conclusion

In this research, we extracted logarithmic spectral feature vectors with the same dimension for both clean and noisy speech and used them to train one model based on bidirectional neural networks. Then by iterative processing, all feature vectors were improved and missing parts of noisy features were predicted by a nonlinear function from hidden layer values of network.

Then one MLP neural network, as a reference model, was trained with the MFCC feature vectors extracted with improved clean speech LFBE features. In the best results in experiment we attained 9.01% and 1.92% increasing in noisy and clean speech recognition accuracy respectively

REFERENCES

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

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

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

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

[5] J. Barker, L. Josifovski, M. Cooke, P. Green, "Soft Decisions in Missing Data Techniques for Robust Automatic Speech Recognition ", ICSLP-2000, Beijing.

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

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

[8] B. Raj and R. M. Stern, “Improving recognition accuracy in noise by using partial spectrographic information”, IEEE SIGNAL PROCESSING MAGAZINE, SEPTEMBER, 2005.

[9] M. Vali , S. A. Seyyed Salehi, “Improvement of Feature Vectors for Cleanand Telephone Speech Recognition Using Bidirectional Neural Network”, Interspeech, 2006.