ROBUST AUTOMATIC SPEECH RECOGNITION IN LOW-SNR CAR ENVIRONMENTS BY THE APPLICATION OF A CONNECTIONIST SUBSPACE-BASED APPROACH TO THE MEL-BASED CEPSTRAL COEFFICIENTS

Sid-Ahmed Selouani, Hesham Tolba & Douglas O'Shaughnessy

INRS- Télécommunications, Université du Québec
900 de la Gauchetière Ouest,
Québec, H5A 1C6, Canada
{selouani, tolba, dougo}@inrs-telecom.uquebec.ca

ABSTRACT

In this paper, the problem of robust large-vocabulary continuous-speech recognition (CSR) in the presence of highly interfering car noise has been considered. Our approach is based on the noise reduction of the parameters that we use for recognition, that is, the Mel-based cepstral coefficients. This is achieved by the use of a Multilayer Perceptron (MLP) network for noise reduction in the cepstral domain in order to get less-variant parameters. Then, the obtained enhanced features are refined via the Karhunen-Loève Transform (KLT) implemented using the Principal Component Analysis (PCA). Experiments show that the use of the enhanced parameters using such an approach increases the recognition rate of the CSR process in highly interfering car noise environments. The HTK Hidden Markov Model Toolkit was used throughout our experiments. Results show that the proposed hybrid technique when included in the front-end of an HTK-based CSR system, outperforms that of the conventional recognition process based on either a KLT- or an MLP-based preprocessing recognition in severe interfering car noise environments for a wide range of SNRs varying from 16 dB to -4 dB using a noisy version of the TIMIT database.

1. INTRODUCTION

The performance of existing CSR systems, whose designs are predicated on relatively noise-free conditions, degrades rapidly in the presence of a high level of adverse conditions. Several approaches have been studied for achieving noise robustness [1, 2]. In this paper, we focus on optimizing the performance of a CSR system by choosing a suitable distortion measure. The idea of a robust distance measure is to extract relevant features from speech signals which must be insensitive to degradations of the speech signal due to interfering noise or distortions. Many approaches [3] have been used to extract relevant features from a speech signal. Cepstral parameters are well suited to speech signals which must be insensitive to degradations of the transmission channels. Several approaches to obtain a new set of robust parameters were introduced in [5, 6, 7].

In this paper, we propose a novel robust CSR system to be used in car noisy environments. Our approach for noise reduction is applied in the cepstral domain. It is based on the application of a combination of the Karhunen-Loève Transform (KLT) and a Connectionist approach. Each of these two approaches has been successfully used in both speech enhancement and recognition processes. We show in this paper through experiments on highly noisy data that a cepstral noise reduction can be obtained using such an approach and consequently an improvement of the recognition performance.

This paper will be organized into the following sections. In section 2 we describe the basis of the MLP network and the PCA approaches that will be used to describe our proposed hybrid PCA-MLP approach. Then, we proceed in section 3 with the description of the database, the platform used in our experiments and the evaluation of the proposed MLP-PCA-based recognizer in a noisy car environment and the comparison of such a recognizer to both the MLP- and the PCA-based recognizers in order to evaluate its performance. Finally, in section 5 we conclude and discuss our results.

2. PROPOSED ENHANCEMENT APPROACH

2.1. Multilayer Perceptron Network

As mentioned above, the first step that has been proposed to improve the performance of the CSR process in highly noisy car environments in the cepstral domain is the use of a multilayer perceptron (MLP) network. The fact that the noise and the speech signal are combined in a nonlinear way in the cepstral domain motivated us to choose the MLP, since it can approximate the required nonlinear function to some extent [6, 7]. The input of the MLP is the noisy MFCC vector \( \hat{C} \), while the actual response of the network \( \hat{C} \) is computed during a training phase using a convergence algorithm to update the weight vector in a manner to minimize the error between the output \( \hat{C} \) and the desired clean cepstrum value \( C \). The weights of this network are calculated during a training phase with a back-propagation training algorithm using a mean square error criterion [8].
2.2. Signal Subspace Approach

Since the speech signals used throughout this study are corrupted by highly interfering noise, the MLP described in section 2.1 cannot totally suppress the noise effects on the MFCCs. In an attempt to reduce the noise effects on the obtained enhanced features, we proposed to use a subspace approach based on the Karhunen-Loève Transform (KLT) implemented using the Principal Component Analysis (PCA) [10]. This choice was motivated by the success of such an approach in reducing the noise effects in the cepstral domain [9].

The principle of this approach is to decompose the vector space of the noisy signal into a signal-plus-noise subspace and a noise subspace. Enhancement is performed by removing the noise subspace and estimating the clean signal from the remaining signal space. Such decomposition is performed by applying the KLT to the noisy zero-mean normalized MFCC vector \( \mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_N]^T \). Assuming that \( \mathbf{C} \) has a symmetric non-negative autocorrelation matrix \( \mathbf{R} = \mathbf{C} \mathbf{C}^T \) with a rank \( r \leq N \), \( \mathbf{C} \) can be represented as a linear combination of the eigenvectors \( \beta_1, \beta_2, \ldots, \beta_r \), which corresponds to the eigenvalues \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0 \), respectively. That is, \( \mathbf{C} \) can be calculated using the following orthogonal transformation:

\[
\mathbf{C} = \sum_{k=1}^{r} a_k \beta_k, \quad k = 1, \ldots, r, \tag{1}
\]

where the coefficients \( a_k \), which are called the principal components, are given by the projection of the vector \( \mathbf{C} \) in the space generated by the eigenvector basis, as follows:

\[
a_k = \mathbf{C}^T \beta_k, \quad k = 1, \ldots, r. \tag{2}
\]

In [10], the linear estimation of the clean vector \( \mathbf{C} \) is performed using two perceptually meaningful estimation criteria as follows:

\[
\hat{\mathbf{C}} = \sum_{k=1}^{r} W_k a_k \beta_k, \quad k = 1, \ldots, r, \tag{3}
\]

where \( W_k \) is a weighting function given by:

\[
W_k = \left[ \frac{\lambda_k}{\lambda_k + \sigma_n^2} \right]^{\gamma}, \quad k = 1, \ldots, r, \tag{4}
\]

where \( \sigma_n^2 \) is the noise variance and \( \gamma \geq 1 \).

An alternative choice for \( W_k \) which results in a more aggressive noise suppression is given by:

\[
W_k = \exp \left\{ -\frac{\nu \sigma_n^2}{\lambda_k} \right\}, \quad k = 1, \ldots, r, \tag{5}
\]

where the value of the parameter \( \nu \) is to be fixed experimentally.

Instead of dealing with only one of the above mentioned weighting functions, we chose to use a combination of them. That is, when the signal is less corrupted by noise, the weights \( W_k \) defined by Equation 4 are used. On the other hand, the weights defined by Equation 5 are used for highly corrupted signals. In fact, this was not chosen in an arbitrary way, but our choice was guided by what is called the reconstruction’s quality function, denoted by \( Q \). \( Q \) is defined as the ratio of the sum of the eigenvalues used to reconstruct \( \hat{\mathbf{C}} \) to the sum of all the eigenvalues, as follows:

\[
Q = \frac{\sum_{k=1}^{r} \lambda_k}{\sum_{k=1}^{N} \lambda_k}. \tag{6}
\]

The first- and second-order derivatives of \( Q \) are given by:

\[
\Delta Q = \frac{\lambda_{r+1}}{\sum_{k=1}^{N} \lambda_k}, \tag{7}
\]

and

\[
\Delta \Delta Q = \frac{\lambda_{r+1} - \lambda_r}{\sum_{k=1}^{N} \lambda_k} \quad \text{for} \quad \lambda_{r+1} > \lambda_r, \tag{8}
\]

where \( \lambda_k \), \( k = 1, \ldots, N \) are the eigenvalues of the clean signal. Given the fact that the magnitudes of the low-order eigenvalues are higher than the magnitudes of the high-order ones, then the effect of the noise on the low-order eigenvalues is less than that for high-order ones. Thus, the variations of \( \Delta \Delta Q \) for a certain noise variance \( \sigma_n^2 \) and a certain value of \( r \) tend to zero for higher order eigenvalues. Thus, the \( Q \)-acceleration function \( (\Delta \Delta Q) \) permits determining the optimal component order, denoted \( r_t \), at which we switch between the use of the two weighting functions defined by Equations 4 and 5.

2.3. MLP-PCA-Based CSR system

The front-end of the proposed MLP-PCA-based CSR system shown in Figure 1 consists of two levels of enhancement. In the first level, the noisy 13-dimensional vector (12 MFCCs + energy) are fed to an MLP network in order to reduce the noise effects on such a vector. This first pre-processing does not require any a-priori knowledge about the nature of the corrupting noisy signal, which permits dealing with any kind of noise. Moreover, using this enhancement technique we avoid the noise estimation process that necessitates a speech/non-speech pre-classification, which could be not accurate for low SNRs. It is also interesting to note that such a technique is less complex than many other enhancement techniques that need to either model or compensate for the noise. However, this enhancement technique requires a large amount of data in order to train the network well [11].

Once we obtained the enhanced vector, it is fed to the PCA module, which represents the second enhancement level. This module refines the enhanced vector by projecting them in the subspace generated by a weighted version of the eigenvectors of the clean signal as shown in section 2.2. The motivation behind the use of a second level of enhancement after using the MLP network is to compensate for the limited power of the MLP network for enhancement outside the training space [11]. Then, the first derivatives are computed from the refined enhanced obtained vector in order to produce a 26-dimensional vector upon which the HMMs, that model the speech subword units, were trained.
3. EXPERIMENTS

3.1. Database

In the following experiments the TIMIT database, described in [12], was used. The TIMIT corpus contains broadband recordings of a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States, each reading 10 phonetically rich sentences. To simulate a noisy environment, car noise was added artificially to the clean speech. To study the effect of such noise on the recognition accuracy of the CSR system that we evaluated, the reference templates for all tests were taken from clean speech. On the other hand, the dr1 subset of the TIMIT database was chosen from the available database to evaluate the recognition system.

3.2. Noise Reduction

In this study, the MLP network was trained using noisy speech at different values of SNR varying from 16 dB to -4 dB using the above-mentioned algorithm. The architecture of the network that has been used throughout all our experiments consists of three layers. The input layer consists of 13 neurons, while the hidden layer and the output layer consists of 26 and 13 neurons, respectively. The input to the network is the noisy 12-dimensional MFCC vector in addition to the energy. The weights of this network are calculated during a training phase with a back-propagation algorithm with a learning rate equal to 0.25 and a momentum coefficient equal to 0.09. The obtained weight values are then used during the recognition process in order to reduce the noise in the enhanced obtained vector that is inputted to the PCA module as shown in Figure 1. The eigenvectors used in the PCA reconstruction module are weighted by the gain function according to the optimal choice of $W_k$ given by either Equation 4 or 5. In our experiments, such optimal choice was based on the variations of $\Delta Q$. These variations are shown in Figure 3 for different SNR values. We found through experiments as shown in Figure 3 that $r = 7$ is a convenient value for switching between the use of either Equation 4 or 5 for the computation of the reconstructed vector according to Equation 3. It was found that the use of such a combination leads to an optimization of such weights. Figure 2 shows the first four MFCCs for a signal that has been chosen from the test set. It is clear from the comparison illustrated in this figure that the processed MFCCs, using the proposed hybrid approach, are less variant than the noisy MFCCs and closer to the original ones.

3.3. Recognition Platform

In order to recognize the continuous speech data that has been enhanced as mentioned above, the HTK-based speech recognition system described in [13] has been used throughout all experiments. HTK is an HMM-based speech recognition system. The toolkit can be used for isolated or continuous whole-word-based recognition systems. The toolkit was designed to support continuous-density HMMs with any numbers of state and mixture components. It also implements a general parameter-tying mechanism which allows the creation of complex model topologies to suit a variety of speech recognition applications.

In all our experiments, 12 MFCCs were calculated on a 30-msec Hamming window advanced by 10 msec each frame. Then, an FFT is performed to calculate a magnitude spectrum for the frame, which is averaged into 20 triangular bins arranged at equal Mel-frequency intervals. Finally, a cosine transform is applied to such data to calculate the 12 MFCCs. Moreover, the normalized log energy is also found, which is added to the 12 MFCCs to form a 13-dimensional (static) vector. This static vector is then expanded to produce a 26-dimensional (static+dynamic) vector upon which the hidden Markov models (HMMs), that model the speech subword units, were trained. The baseline system used for the recognition task uses either a mono-, bi- or tri-phone Gaussian mixture HMM system.

4. RESULTS

Three different sets of experiments has been carried out on the noisy version of the TIMIT database at different values of SNR which varies from 16 dB to -4 dB. In the first set of these experiments, we tested an MLP-based CSR system. We found through experiments that using only the MLP as a pre-processing approach to enhance the MFCCs that were used for recognition
using single mixture Gaussian HMMs leads to an improvement in the accuracy of the word recognition rate as shown in Table 1.

These tests have been repeated using a CSR system based on a PCA preprocessing. The obtained results showed also an improvement in the accuracy of the word recognition rate. However, the comparison of the MLP- and the PCA-based CSR systems show that the MLP-based system outperforms the PCA-based one for all the different values of SNR which varies from 16 dB to -4 dB as shown in Table 1.

Finally, we repeated these tests using the proposed approach which combines the PCA with the MLP in the pre-processing stage as mentioned in section 2. The obtained results showed an improvement in the accuracy of the word recognition rate compared to both the MLP-based and the PCA-based CSR system for all the different values of SNR which varies between 16 dB to -4 dB as shown in Table 1.

5. CONCLUSION

In this paper, a new robust CSR system based on the use of an MLP-PCA hybrid enhancement noise reduction approach in the cepstral domain in order to get less-variant parameters. First, the MFCCs were enhanced using an MLP network. Then, the obtained enhanced features are refined via the KLT implemented using the PCA. Experiments show that the use of the enhanced parameters using such a hybrid approach increases the recognition rate of the CSR process in highly interfering car noise environments for a wide range of SNRs varying from 16 dB to -4 dB using a noisy version of the TIMIT database.

We are currently continuing the effort towards the improvement of the performance of the designed system by modifying the PCA parameters by the use of evolutionary algorithms in order to optimize the weight function \( W_k \).

6. REFERENCES


