ROBUST AUTOMATIC SPEECH RECOGNITION BY THE APPLICATION OF A TEMPORAL-CORRELATION-BASED RECURRENT MULTILAYER NEURAL NETWORK TO THE MEL-BASED CEPSTRAL COEFFICIENTS

Michel Héon, Hesham Tolba and Douglas O'Shaughnessy

INRS-Télécommunications, Université du Québec
16 Place du Commerce, Verdun (Île-des-Soeurs),
Québec, H3E 1H6, Canada
{heon,tolba, dougo}@inrs-telecom.uquebec.ca

ABSTRACT

In this paper, the problem of robust speech recognition 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. A Temporal-Correlation-Based Recurrent Multilayer Neural Network (TCRMNN) for noise reduction in the cepstral domain is used in order to get less-variant parameters to be useful for robust recognition in noisy environments. Experiments show that the use of the enhanced parameters using such an approach increases the recognition rate of the continuous speech recognition (CSR) process. The HTK Hidden Markov Model Toolkit was used throughout. Experiments were done on a noisy version of the TIMIT database. With such a pre-processing noise reduction technique in the front-end of the HTK-based continuous speech recognition system (CSR) system, improvements in the recognition accuracy of about 17.77% and 18.58% using single mixture monophones and triphones, respectively, have been obtained at a moderate SNR of 20 dB.

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 [7]. In this paper, we focus on optimizing the performance of an 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 [1, 11, 5] have been used to extract relevant features from a speech signal. Cepstral parameters are well suited to speech recognition due to their compact orthogonality [1]. Unfortunately, cepstral features are highly sensitive to noise. It was shown in [10] that cepstral distributions for clean data are well behaved and approximately normal, but in the presence of noise, their profiles are changed significantly and this consequently degrades the performance of an CSR system. It was found that the cepstrum coefficients have the additional advantage that one can derive from them a set of parameters which are invariant to any fixed frequency-response distortion introduced by either the adverse environments or the transmission channels [4]. Several approaches to obtain a new set of robust parameters were introduced in [6, 3, 8, 12].

In this paper, we propose a novel robust CSR system to be used in additive noisy environments. Our approach for noise reduction is applied in the cepstral domain. It is based on the application of a Recurrent Multilayer Neural Network (RMNN) to the mel-based cepstral coefficients (MFCCs) on a frame-by-frame basis, while taking into account the correlation effects of the neighbor MFCCs. This approach is tested using the TIMIT database contaminated by additive Gaussian Noise (AGN). We proved via experiments that our approach outperforms the Cepstral Noise Reduction (CNR) approach [12] and the obtained MFCCs are very close to the MFCCs representing the clean speech over a wide range of signal-to-noise ratio (SNR) levels.

In order to reduce the noise effect on the MFCCs that will be used in recognizing noisy speech, we designed a 3-layer RMNN. The entry layer of such a network consists of 84 neurons, and a 12-neuron output layer without hidden layers. The output neurons represent the 12 processed cepstral coefficients to be predicted to represent the present frame, t. The input layer is divided into two groups. The first group consists of the noisy MFCCs belonging to the present frame, t, and the preceding 4 frames. These latter are used in order to cope with the temporal correlation between the MFCC coefficients for successive frames. The second group consists of the output of the network at the time instants $t-1$ and $t-2$, which represents the recurrent part of the network. In order to calculate the weighting coefficients of such a network, a minimum mean square error criterion is used during the training phase. The rule applied to train the proposed network is the *Norm-Cum-Delta Rules* [9], where each neuron uses a *tanh* transfer function.

2. CEPSTRAL COEFFICIENTS

The cepstral coefficients are used to describe the short-term spectral envelope of a speech signal. The cepstrum is the inverse Fourier transform of the logarithm of the short-term power spectrum of the signal. By the logarithmic operation, the vocal tract transfer function and the voice source are separated. Consequently, the pulse sequence originating from the periodic voice source reappears in the cepstrum as a strong peak at the quefrency lag T_o. The advantage of using such coefficients is that they reduce the dimension of a speech spectral vector while maintaining its identity. There are two ways to obtain the cepstral coefficients: FFT cepstral and LPC cepstral coefficients.

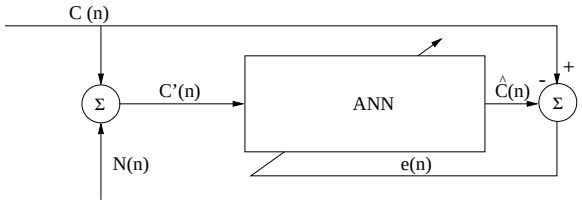


Figure 1: A Simplified Architecture of a Generic Identification System.

In [1] the use of the Mel-scale (Equation (1)) in the derivation of cepstral coefficients was introduced. It was shown in this study that such a scale improves the performance of speech recognition systems over the traditional linear scale. The Mel scale is a mapping from a linear to a nonlinear frequency scale based on human auditory perception. An approximation to the Mel-scale is:

$$mel(f) = 2595\ log_{10}\ (1 + \frac{f}{700}), \qquad (1)$$

where f corresponds to the linear frequency scale. For the MFCC computations, N critical bandpass filters that roughly approximate the frequency response of the basilar membrane in the cochlea of the inner ear are selected. These filters span $156 - 6844$ Hz and are spaced on the Mel-frequency scale defined in equation (1), which is roughly linear below 1 kHz and logarithmic above this frequency. The filters are triangular and multiplicatively scaled by the area. These filters are applied to the log of the magnitude spectrum of the signal, which is estimated on a short-time basis. To obtain the MFCCs, C_n, a discrete cosine transform, is applied to the output of the N filters, X_k, as follows:

$$C_n = \sum_{k=1}^{N} X_k\ cos\Big(\frac{\pi\ n}{N}(k - 0.5)\Big), n = 1, 2, ..., M, \qquad (2)$$

where M is the number of the cepstral coefficients, N is the analysis order and $X_k, k = 1, 2, ..., N$, represents the log-energy output of the k^{th} filter. For the MFCC computations, 20 triangular bandpass filters were used.

3. RMNN NOISE REDUCTION NETWORK

3.1. Training

Fig. 1 shows the general identification algorithm. In accordance with this general algorithm, we proposed an RNN for the noise reduction in the cepstral domain as shown in Fig. 2. This RNN is a three-layer network with all the outputs of the output layer fed back to the input layer. The shown RNN is a dynamic system with an output depending on all previous inputs. Hence, it incorporates the dynamic information of the input speech cepstral signal for adapting noisy cepstral coefficients to clean ones. The RNN is first trained by the output delayed back-propagation algorithm described in section 3.2. Such a training phase permits the adjustment of the weight values in order to obtain an estimated output $\hat{C}(n)$ similar to the desired clean cepstrum value $C(n)$.

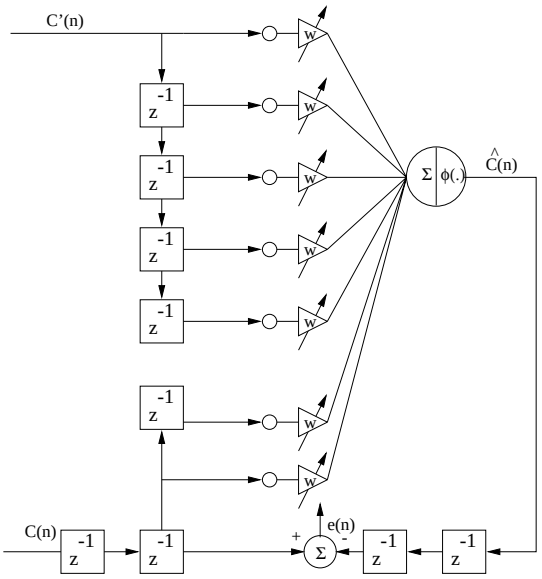


Figure 2: Training RNN Topology.

Minimization of the error between the estimated and the desired values is performed using a quadrature mean error criterion (QEM) given by:

$$QEM = \overline{e^2(n)} = \frac{1}{P}\sum_{p=1}^{P}\big(C_p(n) - \hat{C}_p(n)\big)^2, \qquad (3)$$

where P is the dimension of the MFCC vector.

3.2. Weights Modification Algorithm

Given the input vector $\mathbf{C}^{'}(n) = [C_1^{'}(n), ..., C_p^{'}(n)]^T$ and the weight vector $\mathbf{W}(n) = [\theta(n), W_1(n), ..., W_p(n)]^T$, where p is the number of the neurons at the input of the network and $\theta(n)$ is a threshold value, the actual response of the network $\hat{C}(n)$ 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 $\hat{C}(n)$ and the desired response $C(n)$ as follows [14]:

1. **Initialization**
   W(0) = random(-0.1,0.1).
2. **Activation**
   Assign values to both the input and the desired output of the network: $C^{'}(n)$ and $C(n)$.
3. **Output Computation**
   $\hat{C}(n) = \tanh(\mathbf{W^T}(\mathbf{n})\mathbf{C}^{'}(\mathbf{n}))$
4. **Weight Vector Adaptation**
   for each iteration:
   $m(n+1) = m(n) + \eta_1 e(n) C^{'}(n)$
   for each iteration modulo epoch=0:

   - $W(n+1) = W(n) + m(n) + \eta_2 a(n)$,
   - $a(n) = m(n)$,
   - m(n+1) = 0,

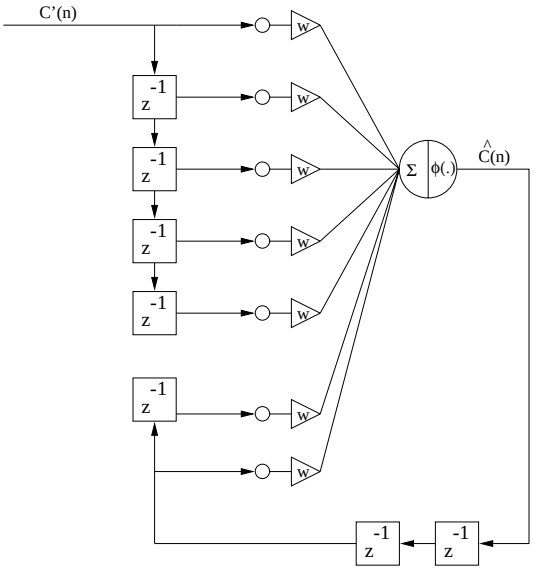


Figure 3: Noise Reduction RNN Topology.

where $0 \leq \eta_1, \eta_2 \leq 1.0$
and the error $e(n)$ is given by $e(n) = C(n) - \hat{C}(n)$.

5. $n = n + 1$; go back to step 2.

3.3. Noise Reduction

The noise reduction using the proposed approach is shown in Fig. 3. The particularity of such a topology is the use of two output-delayed values as the input to the network. This in turn forces the network to take into consideration the preceding estimated values in order to invoke the estimation of the succeeding values. The two delayed $\hat{C}(n)$ allows centering, in the time-domain, the input noisy signal with respect to the output signal. That is, at a time instant t, the network restores $\hat{C}(t)$ with the values: $C^{'}(t-2), C^{'}(t-1), C^{'}(t), C^{'}(t+1), C^{'}(t+2)$, using the optimum weights obtained during the training phase.

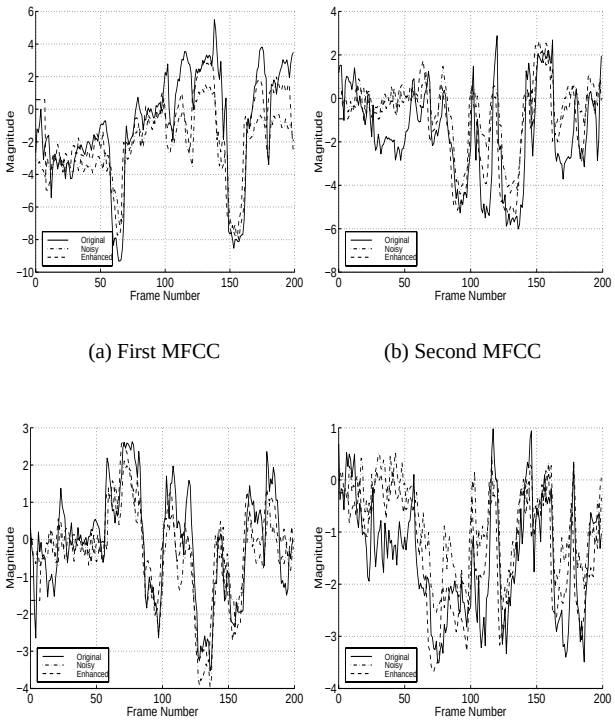
3.4. HTK

The speech recognition system used in our experiments, HTK, is completely described in [13]. 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.

4. EXPERIMENTS

4.1. Database

In the following experiments the TIMIT database, described in [2], 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, white Gaussian noise was added artificially to the clean



(a) First MFCC (b) Second MFCC

(c) Third MFCC (d) Fourth MFCC

Figure 4: Comparison between original, noisy and enhanced MFCCs.

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.

4.2. Noise Reduction

In this study, the RMNN was trained using noisy speech with a 20 dB SNR using the above-mentioned algorithm. The obtained weight values are then used in order to reduce the noise in the MFCCs input to the CSR system as shown in Fig. 3. It is clear from the comparison illustrated in Fig. 4 that the processed MFCCs, using the proposed recurrent neural network, are less variant than the noisy MFCCs and closer to the original MFCCs. These results reflect the improvement of the recognition accuracy when such coefficients were used for the recognition of continuous speech, as shown in Tables 1 and 2.

4.3. 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. 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,

	$\epsilon_{Sub}(\%)$	$\epsilon_{Del}(\%)$	$\epsilon_{Ins}(\%)$	$C_{Wrd}(\%)$
Clean Cep	26.59	13.97	1.88	59.44
Noisy Cep	46.92	22.11	1.98	30.97
Enhanced Cep	35.35	14.18	1.67	50.47

Table 1: Comparisons of the recognition performance of the RNN-Based HTK CSR system to the baseline HTK using single mixture monophones and the dr1 subset of the TIMIT database when contaminated by AGN, SNR=20dB.

	$\epsilon_{Sub}(\%)$	$\epsilon_{Del}(\%)$	$\epsilon_{Ins}(\%)$	$C_{Wrd}(\%)$
Clean Cep	18.67	6.15	1.67	75.18
Noisy Cep	37.96	12.20	3.55	49.84
Enhanced Cep	28.28	7.72	2.92	63.40

Table 2: Comparisons of the recognition performance of the RNN-Based HTK CSR system to the baseline HTK using single mixture triphones and the dr1 subset of the TIMIT database when contaminated by AGN, SNR=20dB.

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 as described in [4] to form a 12-dimensional (static) vector. This static vector is then expanded to produce a 24-dimensional (static+dynamic) vector upon which the HMMs, that model the speech subword units, were trained. The static vector is extended by appending the first-order difference of the static coefficients. The baseline system used for the recognition task uses either a mono- or tri-phone Gaussian-mixture HMM system.

The speech data is segmented into 30-msec frames with 10-msec overlapping. Each frame is weighted by a 512-point Hamming window, and then the DFT using 512-point FFT of that frame is computed. Then the feature vector is calculated for each frame. Each vector is composed of 12 static MFCCs, plus the dynamic coefficients. This leads to a 24-element vector per frame.

5. RESULTS

Applying the overall proposed recognizer to the noisy version of the TIMIT database with a SNR of 20 dB, and carrying on some experiments proved that the recognition accuracy has increased significantly when the RMNN is used before performing the recognition. In order to evaluate the performance of our proposed system, we compared the performance of the RNN-based HTK recognizer to the baseline HTK recognition system. The relative changes in the word correctness rate, C_{Wrd}, when using our proposed system for testing on a subset of the TIMIT database using single mixture Gaussian models over the baseline HTK are shown in Tables 1 and 2.

6. CONCLUSION

In this paper, a new robust CSR system based on RNN has been described. This was realized by the inclusion of such a network in the pre-processing enhancement algorithm used in the recognition process. We proved via experiments that the proposed RNN-based recognition system system is robust and outperforms the baseline recognition system in an AGN environment.

We are currently continuing the effort towards the inclusion of our proposed RMNN in the front-end of an automatic speech recognition system in order to test the new enhanced parameters.

7. REFERENCES

[1] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-28(4), pp. 357–36, 1980.

[2] William M. Fisher, George R. Dodington, and Kathleen M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specification and Status", Proc. DARPA Workshop on Speech Recognition, pp. 93–99, 1986.

[3] Sadaoki Furui, "Cepstral Analysis Techniques for Automatic Speech Verification", IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP–29(2), pp. 254–272, April 1981.

[4] Sadaoki Furui, "Speaker–Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP–34(1), pp. 52–59, February 1986.

[5] J. Hernando and C. Nadeu, "A Comparative Study of Parameters and Distances for Noisy Speech Recognition", Proc. Eurospeech–91, pp. 91–94, 1991.

[6] B. Juang, L. Rabiner, and J. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition", IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-35(7), pp. 947–954, July 1987.

[7] Jean-Claude Junqua and Jean-Paul Haton, "Robustness in Automatic Speech Recognition", Kluwer Academic Publishers, 1996.

[8] F. Liu, R. Stern, A. Acero, and P. Moreno, "Environment Normalization for Robust Speech Recognition Using Direct Cepstral Comparison", In Proc. ICASSP–94, pp. 61–64, 1994.

[9] Neural Ware Software Group, "Advanced Reference Guide, Software Reference for Professional II/Plus and NeuralWorks Explorer", Neural Ware Inc., 1993.

[10] J. P. Openshaw and J. S. Mason, "On the Limitations of Cepstral Features in Noise", Proc. ICASSP–94, pages 49–52, April 1994.

[11] Manfred R. Schroeder, " Direct (Nonrecursive) Relations Between Cepstrum and Predictor Coefficients", IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP–29(2):pp. 297–301, April 1981.

[12] Helge B.D. Sorensen, "A Cepstral Noise Reduction Multi–layer Neural Network", Proc. ICASSP–91, pages 933–936, 1991.

[13] Cambridge University Speech Group, "HTK Hidden Markov Model Toolkit", Entropic Research Laboratories Inc., Cambridge, December 1993,

[14] Simon Haykin, "Neural Networks – A Comprehensive Foundation", Macmillan College Publication Company Inc., 1994.