

USING NOISE ROBUST FEATURES FOR SPEECH ENHANCEMENT

*Mohamed F. Mansour and **Mohsen A. Rashwan

* Suez Canal University, Faculty of Computers and Information Systems, Ismailia, Egypt. IEEE member

** Cairo University, Faculty of Engineering, Department of Electronics and Communication, Giza , Egypt

ABSTRACT

A New system for speech enhancement, that uses hidden markov models (HMM) with noise-robust features, is developed. This system alleviates the error in decoding the noisy speech, that was observed in the conventional model-based enhancement system [2] . This decoding error results in a degradation in the performance of the conventional system from its theoretical upper limits. A comparative test, using additive white noise, has shown that the proposed system is superior to conventional model-based speech enhancement system at all input SNR.

1. INTRODUCTION

Model Based speech enhancement [2,4,5] is one of the important approaches for improving the quality of the speech signal corrupted by additive noise. In this approach, autoregressive hidden markov models (AR-HMM) are used to represent both the clean speech and the noise process. During enhancement, using Maximum A Posteriori (MAP) estimation of the enhanced speech [4], the input noisy signal is aligned to a set of composite states which represent the clean speech and the noise process. After aligning, a set of filter templates is applied to the input noisy signal to recover the clean one. The correct decoding of the input noisy signal, to produce the optimal composite states sequence representing it, is crucial to the success of this approach. It was noted that, there are some degradation of model-based enhancement systems from their theoretical upper limits due to the effect of the additive noise on the clean speech. The upper limit performance is obtained by using the clean speech (which is assumed to be available) for decoding, rather than the noisy one. In this paper, We suggest the use of the noise-robust features for aligning the states so as to minimize the effect of the noise on the decoded state sequence. We develop an enhancement system, that uses some cepstrum-based features, namely the Adaptive Component Weighted (ACW) cepstrum, for decoding the noisy speech. The spectral characteristics of the clean speech and the noise process are described independent of the type of features used for decoding. The resulting algorithm is simpler in calculations than the conventional system as it doesn't use composite states for the clean speech and the noise. The overall performance is improved compared with conventional model-based speech enhancement by nearly 10%.

2. USING NOISE ROBUST FEATURES

2.1. The Conventional System

In the conventional model-based speech enhancement system, AR-HMMs are used to represent both the clean speech signal and the noise-process. Composite states are used to represent the noisy speech. Each composite state consists of a state from the clean speech model, and a state from the noise-only model.

The output probability of the composite state is a gaussian distribution (not autoregressive), whose mean vector and covariance matrix are calculated from the mean and covariance of the building states[4]. Each composite state is associated with a filter template (Wiener Filter), whose spectral characteristic is defined in terms of the covariance matrices of the building states as described in [4,5].

we adopt the Maximum A Posteriori (MAP) estimation approach[4], for enhancement. During the enhancement session, the noisy speech is decoded using the Viterbi algorithm to find the optimal composite state sequence that represents the noisy speech. The filters associated with each composite state on the optimal state sequence are applied in order to the noisy speech to recover the enhanced signal.

2.2. Rationale

The theoretical upper limit performance of the conventional model-based enhancement system is obtained by using the clean speech for decoding other than the noisy one as shown in Figure (1). It was found that, the performance of the conventional model-based speech enhancement degrades from its theoretical upper limits due to the effect of the additive noise.

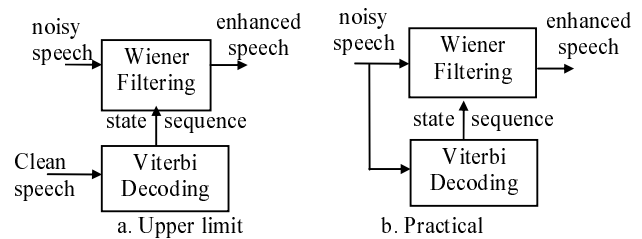


Figure 1. Configurations for MAP estimation

To overcome this problem, we suggest the use of features, that are insensitive to the additive noise, for decoding, other than the Linear Prediction Coefficients (LPC), that are used in the conventional model-based enhancement.

We choose to use Adaptive Component Weighted (ACW) cepstrum [1] as an example for the noise-robust features, because it shows high noise robustness in speaker identification applications.

3. THE PROPOSED SYSTEM

3.1. Basic Idea

The basic idea of our system is to use the noise robust features for decoding the input noisy utterance. Toward this goal, we build ergodic HMM in the proposed features space, rather than AR-HMM for the case of the conventional system.. This HMM that models the clean speech, is used directly to model the noisy one in the enhancement session. Each state in the HMM

is associated with a filter template in the form of Wiener filter as that used in the conventional system. The filter templates are estimated from the clean speech spectral templates at the training session and the noise template at the enhancement session as will be discussed later.

3.2. Proposed Models

Ergodic HMM in the ACW cepstrum space is used. The output probability of each state is a continuous distribution described as a mixture of gaussians, each has a diagonal covariance matrix. The number of states used is typically 5, and the number of gaussian distributions per state ranges from 3 to 16. It was found that the performance of the system is nearly the same for a wide range of the number of states and number of distributions within a single state. The spectral template of the clean speech associated with each state has the form of Autoregressive model of order 10-12.

3.3. Gain Adaptation

One of the key factors for the successful use of the HMM, is the matching of the gain conditions between the training and testing sessions. In our system, this problem is solved by calculating the maximum likelihood estimate of the gain contour at the two sessions. we adopt the same approach described in [3] at both the training and enhancement sessions.

3.4. Training

The training of the HMM is performed using the Segmental K-means algorithm[10]. The training procedure is concluded in the following steps:

1. Initializing the HMM parameters , using K-means vector quantization of the training data.
2. Initializing the gain contour of each of the input training utterances. The value of the gain contour at each time frame is set to LPC residual power of this frame.
3. Segmenting the input training data between states by applying the Viterbi algorithm [6] on each input utterance of the training data.
4. Reestimating the HMM parameters from the data assigned to each state using the relations described in [10].
5. Estimating the spectral templates associated with each state from the frames assigned to it. This is done by first, calculating the average autocorrelation vector of the frames assigned to each state according to the relation :

$$r_i(j) = \frac{1}{N_i} \sum_{n=1}^{N_i} \sum_{m=1}^{K-j} x_n(m)x_n(m+j) \quad (1)$$

where : $r_i(j)$ is the average autocorrelation vector of the frames assigned to state i , N_i is the number of frames assigned to state i during the training session, $x_n(m)$ is the m th sample of the frame number n assigned to state i , and K is the input frame length. From this autocorrelation vector the autoregressive model of this state is estimated using the Levinson-Durbin algorithm [9]. This AR model is used to estimate the spectral template of the clean speech associated with each state.

6. Reestimating the gain contour of each of the input training utterances. The new gain value at each time frame is equal to the residual power that results from filtering that frame by the

inverse AR filter, associated with the state on the optimal state sequence at this time.

7. If the final likelihood of the training data at two subsequent iterations is less than a certain threshold stop, otherwise repeat from step 3.

3.5. Enhancement

The enhancement is an iterative algorithm. First, the noise spectral template is estimated from the first period of the input noisy utterance (typically 200 ms), which is assumed to be silence. The gain contour is initialized as described in the training section. At each iteration, the following steps are done:

1. Estimating the filter templates associated with each of the HMM states from the clean speech spectral template of this state, the noise spectral template, and the value of the gain contour at each time frame. The filter has the form

$$H(\omega) = \frac{g_t^2 \cdot |S_x(\omega)|^2}{g_t^2 \cdot |S_x(\omega)|^2 + |S_n(\omega)|^2} \quad (2)$$

where: $S_x(\omega)$ is the normalized spectrum of the speech signal,

$S_n(\omega)$ is the spectrum of the noise process , g_t^2 is the value of the gain contour at time t .

2. Decoding the input speech, using the Viterbi algorithm to find the optimal state sequence that best represents the input. At the first iteration, the noisy speech is used for decoding but at subsequent iterations the enhanced speech from the previous iteration is used. It should be mentioned that, the features used for expansion are the proposed noise robust features.
3. Applying a sequence of filters, associated with states on the optimal state sequence, to the noisy signal so as to produce the enhanced speech at this iteration.
4. Reestimating the gain contour, in the same way described in training section.

The iteration stops, if the difference in the final likelihood of the enhanced utterance at two successive iterations is less than a certain threshold. The overall enhancement system is shown in figure (2).

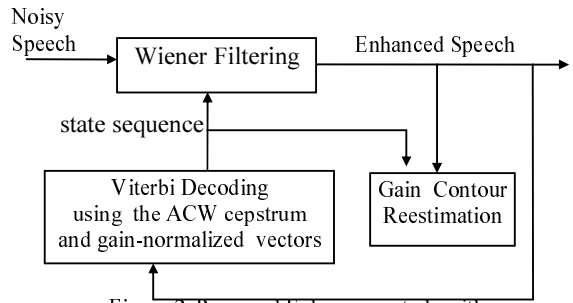


Figure 2. Proposed Enhancement algorithm

4. RESULTS

The proposed system is compared with the conventional model-based speech enhancement with gain-adapted MAP estimation criteria [4]. The training data for both is five minutes of a single speaker utterance (the same as test speaker).

Nonoverlapped frames are used with rectangular window. The frame length is 20 ms. The filtering is done using short time FFT and the Overlap-add convolution technique [8]. The models used for both is ergodic HMM with five states. The output probability of the states is in the form of mixture of gaussians. The noisy signal is an additive white noisy signal at different input SNR.

First, we compare our proposed system, with the ACW cepstrum as the used features, to the conventional system in both the practical configuration and the theoretical upper limit configuration (Figure 1). The results are shown in tables 1,2. The degradation is calculated as follows:

$$\text{Degradation} = \frac{\text{Upper Limit Improvement} - \text{Practical Improvement}}{\text{Upper Limit Improvement}}$$

where the improvement is the difference between input and output SNR's.

<i>Input SNR (dB)</i>	<i>Output SNR (dB) practical / upper limit</i>	<i>Degradation (%)</i>
5	13.62 / 14.94	9
10	17.55 / 18.24	8
15	20.91 / 21.73	12
20	24.5 / 25.0	10

Table 1. Performance of Conventional Gain-Adapted MAP algorithm

<i>Input SNR (dB)</i>	<i>Output SNR (dB) practical / upper limit</i>	<i>Degradation (%)</i>
5	13.97 / 14.7	7
10	17.82 / 18.73	6
15	21.49 / 21.8	4
20	24.9 / 25.05	3

Table 2. Performance of the proposed system with ACW cepstrum

As seen from the two tables, the degradation from the upper limit is reduced using our proposed system, and this improvement is very obvious at high input SNR.

Next, the proposed system is tested using the postfilter cepstrum [7], instead of the ACW cepstrum. It was found that, the performance is approximately the same as the ACW cepstrum case. This shows the flexibility of the proposed system to accommodate any other type of noise-robust features without any change of its structure.

As seen from the above results, the proposed system reduces the percentage degradation from the upper limits and improves the overall performance by 10-15 %.

5. CONCLUSION AND FURTHER WORK

In the work developed in this paper, a new system that uses noise robust features for speech enhancement is implemented. The rationale behind this approach is to alleviate the degradation from the theoretical upper limits that was noted in the conventional model based enhancement systems, due to the effect of the additive noise. The proposed system reduces the degradation from the upper limits and improves the overall performance by 10-15%. In addition, the proposed system is simpler in calculations as no composite models are needed. The models built for the clean speech are used for the noisy one

because the change in the characteristics of the noise-robust features is minor.

Several extensions can be added to our work, e.g. using better noise-robust features to completely remove the degradation, using better models for the noise process according to its nature, and trying models adaptation during enhancement to compensate for the mismatch between training and enhancement sessions.

REFERENCES

- [1] K.T.Assaleh, and R.J.Mammone, "New LP-derived features for speaker identification", IEEE trans. Speech, and Audio Processing, vol. 2, No. 4, pp. 630-638, October 1994.
- [2] Y.Ephraim, "Statistical Model-Based Speech Enhancement Systems", Proc. IEEE, vol. 80, no. 10, pp. 1524-1555, October 1992
- [3] Y.Ephraim, "Gain-Adapted Hidden Markov Models for Recognition of Clean and Noisy Speech", IEEE trans. on Acoustics, Speech, and Signal Processing, vol. 40, No. 6, pp. 1303-1316, June 1992.
- [4] Y.Ephraim, "Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models", IEEE trans. on Acoustics, Speech, and Signal Processing, vol. 40, No. 4, pp. 725-735, April 1992.
- [5] Y.Ephraim, D. Malah and B.H.Juang, "On the Application of Hidden Markov Models for Enhancing Noisy Speech", IEEE trans. on Acoustics, Speech, and Signal Processing, vol. 37, No. 12, pp. 1846-1856, December 1989.
- [6] G.D.Forney, "The Viterbi Algorithm", Proc. IEEE, vol. 61, pp. 268-278, March 1973.
- [7] R.J.Mammone, X. Zhang and R.P.Ramachandran, "Robust Speaker Recognition, A Feature-based Approach", IEEE Signal Processing magazine, vol. 13 No. 5, pp. 58-71, September 1996.
- [8] A.V.Oppenheim and R.W.Schafer, "Digital Signal Processing", Prentice-Hall (1975), pp. 110-115.
- [9] T.W.Parsons, "Voice and Speech Processing", McGraw-Hill 1987. pp. 141-145
- [10] L.R.Rabiner, J.G. Wilpon and B.H.Juang, "A segmental k-means training procedure for connected word recognition", AT&T Tech. Journal, pp. 21-40, May-June 1986.