Using AR HMM State-Dependent Filtering for Speech Enhancement

Driss Matrouf, Jean-Luc Gauvain

Spoken Language Processing Group
LIMSI-CNRS, BP 133, 91403 Orsay cedex, FRANCE
{driss, gauvain}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

In this paper we address the problem of enhancing speech which has been degraded by additive noise. As proposed by Ephraim et al., autoregressive hidden Markov models (AR-HMM) for the clean speech and an autoregressive Gaussian for the noise are used. The filter applied to a given frame of noisy speech is estimated using the noise model and the autoregressive Gaussian having the highest *a posteriori* probability given the decoded state sequence. The success of this technique is highly dependent on accurate estimation of the best state sequence. A new strategy combining the use of cepstral-based HMMs, autoregressive HMMs, and a model combination technique, is proposed. The intelligibility of the enhanced speech is indirectly assessed via speech recognition, by comparing performance on noisy speech with compensated models to performance on the enhanced speech with clean-speech models. The results on enhanced speech are as good as our best results obtained with noise compensated models.

INTRODUCTION

Speech enhancement has been investigated by many researchers. However, most of the approaches use limited prior information about speech. For example, spectral subtraction techniques [7] can be applied to noisy speech in the same manner as to any other noisy signal. Another example is the CDCN algorithm [3] which can be seen as an enhancement technique in the cepstral domain. It uses prior information about the speech cepstra, contained in a codebook, to estimate additive and convolutive noises and compensates for them using the MMSE criterion. This kind of *a priori* information may not be sufficient to enhance speech without reducing intelligibility.[1]

Using a Maximum *A Posteriori* (MAP) approach, Lim and Oppenheim [9] proposed a time-varying autoregressive Gaussian model to enhance the speech signal, where both the model and the signal are directly estimated from the noisy signal. The estimation is done iteratively, once over the time-varying AR models assuming that the clean signal is available (the first estimation of clean speech signal is the noisy speech signal) and once over the clean speech using the estimated models and the AR noise model. This estimation cannot really converge properly as the number of unknown variables (AR models and clean speech) is large with respect to the number of known variables (noisy speech). To solve this problem, Ephraim et al. [6] proposed to use an AR-HMM framework and to estimate the models from clean speech training data rather than from the given noisy signal.

To find the mode of the *a posteriori* probability density function (pdf) of the clean speech, Ephraim et al. used an iterative procedure based on the EM algorithm [6]. When the initialization is inappropriate, this iterative procedure converges towards a local maximum which can be far from the optimal solution. This is often the case with very noisy speech signals. Logan and Robinson [8] proposed a model combination technique in the autoregressive HMM framework to better initialize the iterative procedure. The initialization is done by decoding the noisy speech frames with a speech recognizer based on noise compensated AR-HMMs. However, it is well known that cepstral-based HMM recognizers are more efficient at decoding speech than AR-HMMs especially for large vocabulary applications. Here we extend this latter approach using a cepstral-based HMM recognizer for initialization. We use two sets of acoustic models instead of one: the first one is a cepstral-based HMM (with $\Delta cepstrum$ and $\Delta^2 cepstrum$) which is used to find a better initialization for the iterative process; the second model is an autoregressive one and is used to estimate the optimal time-varying filters. Estimation of the clean speech is obtained by applying this time-varying filter to successive frames of the noisy speech signal. For a given noisy frame, we first find the cepstral Gaussian with the highest *a posteriori* probability by decoding the speech with the compensated cepstral-based models. This decoding gives a frame/cepstral-Gaussian alignment, where each Gaussian in the cepstral-based HMM corresponds to an autoregressive Gaussian in the AR-HMM. The optimal filter is then estimated using this autoregressive Gaussian and the noise autoregressive Gaussian. The cepstral-based HMMs and the AR-HMMs are trained in such a way that there is a one-to-

[1] In the CDCN case, the intelligibility is from the view point of speech recognizer.

one mapping between the two sets of models at the Gaussian level. To do so, we first estimate the cepstral-based HMM, and then we use the statistics of the last iteration to estimate the AR-HMM parameters.

The following sections describe the implementation of this approach for speech enhancement. Experimental results with different noises are given to demonstrate the improvement of speech quality and speech recognition performance with enhanced speech is measured.

ENHANCEMENT PROCESS

Let us first consider the case of an observation generated by a single Gaussian. Let $\mathbf{y}$ be a noisy frame ($\mathbf{y} \in \mathbf{R}^K$) and $f_{\lambda_\mathbf{X}}$ be the pdf of the clean frame $\mathbf{x}$ corresponding to the noisy frame $\mathbf{y}$. Assuming that $\mathbf{x}$ is generated by an autoregressive process, its pdf $f_{\lambda_\mathbf{X}}(\mathbf{x})$ is defined as

$$f_{\lambda_\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{K}{2}} |S_\mathbf{X}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}^{'} S_\mathbf{X}^{-1} \mathbf{x})\} \qquad (1)$$

where $S_\mathbf{X}$ is the autocorrelation matrix: $S_\mathbf{X} = \sigma^2_\mathbf{X}(A^{'}_\mathbf{X} A_\mathbf{X})^{-1}$, for which $\sigma^2_\mathbf{X}$ is the variance of the innovation process of the AR source, and $A_\mathbf{X}$ is a $K \times K$ lower triangular Toeplitz matrix in which the first $p+1$ elements of the first column constitute the coefficient of the AR process: $a_{i,0 \leq i \leq p}$ where $a_0 = 1$.

Similarly, let $f_{\lambda_\mathbf{n}}$ be the pdf of the additive noise which is also assumed to be an autoregressive Gaussian. The enhancement problem consists of estimating the clean frame $\mathbf{x}$ using $f_{\lambda_\mathbf{X}}$, $f_{\lambda_\mathbf{n}}$ and $\mathbf{y}$. The MAP estimation $\hat{\mathbf{x}}$ is defined as follows:

$$\hat{\mathbf{x}} = \underset{\mathbf{X}}{\text{argmax}} \log\{h(\mathbf{x}, \mathbf{y})\} \qquad (2)$$

where $h(\mathbf{x}, \mathbf{y})$ is the joint pdf of $\mathbf{x}$ and $\mathbf{y}$. Since the noise is additive and independent of the signal, we have:

$$h(\mathbf{x}, \mathbf{y}) = f_{\lambda_\mathbf{X}}(\mathbf{x}) f_{\lambda_\mathbf{n}}(\mathbf{y} - \mathbf{x}). \qquad (3)$$

This maximization leads to the Wiener filter. The Fourier transform of the estimated clean frame is calculated as:

$$\hat{X}(\theta) = \frac{\Gamma_\mathbf{X}(\theta)}{\Gamma_\mathbf{X}(\theta) + \Gamma_\mathbf{n}(\theta)} Y(\theta) \qquad (4)$$

where $Y(\theta)$ is the Fourier transform of the noisy speech frame, and $\Gamma_\mathbf{X}(\theta)$ and $\Gamma_\mathbf{n}(\theta)$ are the power spectral densities associated with the two AR processes. The spectral densities are obtained as follows:

$$\Gamma_\mathbf{X}(\theta) = \frac{\sigma^2_\mathbf{X}}{|\Psi_\mathbf{X}(\theta)|^2}, \qquad (5)$$

$$\Gamma_\mathbf{n}(\theta) = \frac{\sigma^2_\mathbf{n}}{|\Psi_\mathbf{n}(\theta)|^2} \qquad (6)$$

where $\Psi_\mathbf{X}(\theta)$ and $\Psi_\mathbf{n}(\theta)$ are the Fourier transforms of the prediction coefficients for the clean speech and for the noise respectively.

Considering the more general case of an AR-HMM process, let $\mathbf{y} = \mathbf{y}_{t, t=1,\ldots,T} / \mathbf{y}_t \in \mathbf{R}^k$ be a sequence of noisy frames corresponding to a noisy sentence (T is the number of frames in the sentence). The MAP estimate of the clean speech frames is obtained iteratively using the EM algorithm. At each iteration k, the Fourier transform $\hat{X}^k_t$, $\{t = 1, \ldots, T\}$ of the estimated clean frame $\hat{\mathbf{x}}_t(k)$, is obtained from the noisy frame as follows:

$$\hat{X}^{k+1}_t(\theta) = \left[\sum_{\beta,\gamma} p_t(\beta, \gamma | \hat{\mathbf{x}}(k)) H^{-1}_{\gamma|\beta} \right]^{-1} Y_t(\theta), \qquad (7)$$

where, $p_t(\beta, \gamma | \hat{\mathbf{x}}(k))$ is the probability of being in Gaussian γ of state β at time t, given that $\hat{\mathbf{x}}(k)$ is generated by the AR-HMM, and $H_{\gamma|\beta}$ is the Wiener filter associated with the autoregressive Gaussian γ of state β and the autoregressive Gaussian of the noise (Equation 4).

The success of this procedure for enhancing the noisy speech signal is highly dependent on the estimation of the *a posteriori* probabilities $p_t(\beta, \gamma | \hat{\mathbf{x}}(k))$. These probabilities are estimated using the "backward-forward" procedure. If these probabilities are estimated using acoustic models trained on clean speech data, the iterative process is likely to converge towards suboptimal solution, especially when the signal-to-noise ratio (SNR) of the noisy signal is low. To have a better estimate of these probabilities, we decode the noisy speech signal $\mathbf{y}$ using acoustic models which were obtained by adapting clean cepstral HMMs with the test noise, i.e. using the best available noisy speech model.

Since no prior knowledge of the background noise characteristics is available, model compensation has to be performed using only the test data. The compensated models are obtained by adapting the models trained on clean speech. Various techniques have been proposed to combine a clean speech model with a noise model, including log-normal approximation, numerical integration, and data driven approaches[2]. In this work we use a data-driven model combination (DDMC) approach, where the usual approximations are avoided by directly using the original training speech samples instead of generating speech samples from the models[1, 5]. This approach is computationally inexpensive in comparison to other proposed approaches, even though it requires reading all of the training data form disk. We assume that the Gaussian *posteriori* probabilities for a given training frame remain unchanged after adding the test noise. The basic steps of the enhancement process are shown in Figure 1. The two main components are the initialization using cepstral-based HMM with DDMC and AR-HMM state-dependent filtering.

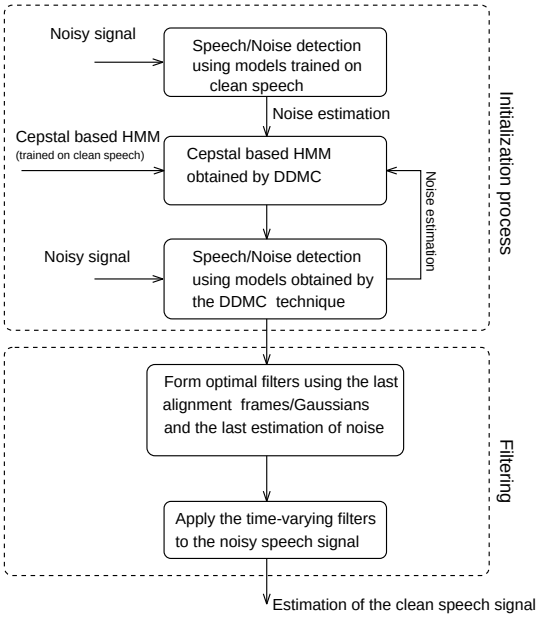

Figure 1: *Speech enhancement using the AR-HMM state-dependent filtering. Cepstral-based HMM and model combination are used for initialization.*

EXPERIMENTAL RESULTS

To evaluate this technique, we used 22,148 utterances from the MASK[2] corpus[4] from a total of 460 speakers. Data from 450 speakers were used for training and data from 10 speakers were kept aside for test. This data was collected using a close-talking microphone with an average SNR of 35dB. The speech signal is bandlimited to 8kHz and sampled at 16 kHz. The recognizer is speaker-independent and capable of recognizing continuously spoken spontaneous speech in real-time with a recognition vocabulary of 1500 words and a bigram language model.

For both cepstral and AR-HMMs we used a 30ms frame window (480 samples) and an 10ms frame rate. For the cepstral HMM, the feature vector is composed of 13 MFCC and their first and second derivatives. Cepstral mean removal is performed for each sentence. The order of the AR Gaussians has been fixed to 16 (for the clean speech AR-HMMs and for the noise model). We use 608 context-dependent phone models. Each phone model is a left-to-right CDHMM with Gaussian mixture observation densities typically having 20 components.

The speech enhancement algorithm has been applied to speech signals degraded with additive noise. In this paper we present results using three types of noise taken from the NOISEX-92 database [10]: white noise, Lynx noise and F16 jet noise. Synthesis of the enhanced signal from the individually processed frames was done using the standard overlap-add technique with a Hanning window.

[2]The MASK spoken language system provides access to train travel information.

Improvement of speech quality

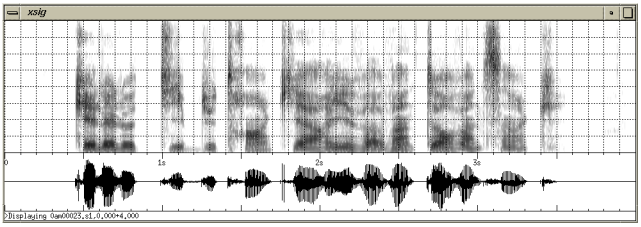

Figure 2: *Spectrogram of clean speech: "quel est le type du train qui arrive à 20 heures 25."*

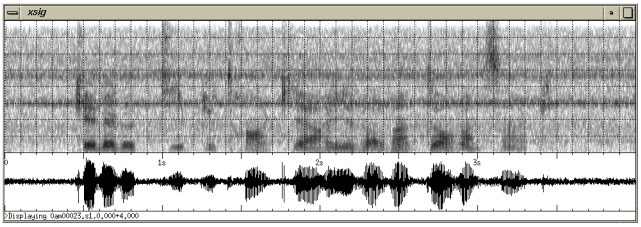

Figure 3: *Noisy speech (SNR=5.7dB) generated by adding F16 Jet noise to the signal in Figure 2.*

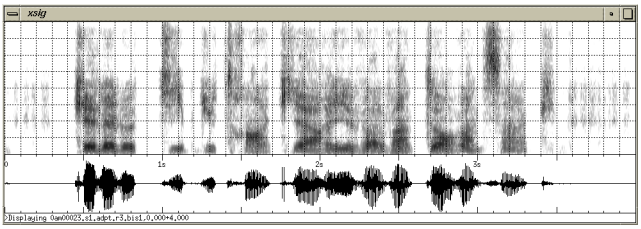

Figure 4: *Enhanced version of noisy speech in Figure 3 using AR-HMM state-dependent Wiener filters.*

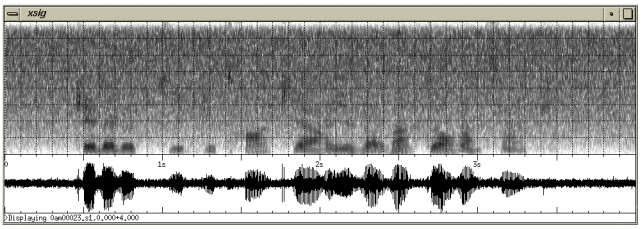

Figure 5: *Noisy speech (SNR=1dB) generated by adding white noise to the signal in Figure 2.*

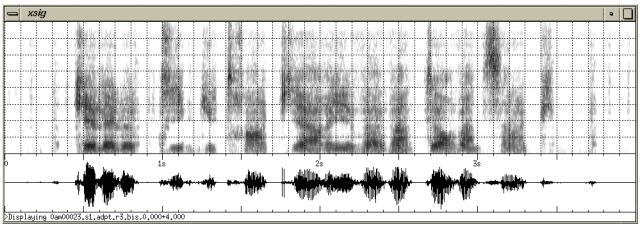

Figure 6: *Enhanced version of noisy speech in Figure 5 using AR-HMM state-dependent Wiener filters.*

A spectrogram of an original clean speech signal is shown in Figure 2. Noisy versions obtained by adding F16 jet noise and white noise are shown in Figures 3 and 5 respectively, with the enhanced signals obtained by Wiener filtering shown in Figures 4 and 6 respectively. Visually a significant decrease in noise can be seen. Listening to the enhanced signal we hear only a slight distortion and no musical noise. Although not illustrated here, enhancement of noisy speech generated by adding Lynx noise was less effective, because the energy of Lynx noise is concentrated in low frequency band, causing more loss of speech information.

Recognition with enhanced speech

In this section we describe experiments which indirectly assess the intelligibility of the enhanced signal via speech recognition. The performance of the recognizer on noisy speech with matched models is compared its performance on the enhanced speech with clean-speech models. Experiments have been carried out with test signals degraded by 3 types of additive noise: white noise, F16 jet noise and Lynx noise. The word error rates for the different configurations are shown in Table 1. Besides the Wiener filter given by equation 4 (filter 1), we have also used the square root of the Wiener filter (filter 2).

	Word error rates (%)		
Test configuration	F16	Lynx	White
Clean test data	5.9	5.9	5.9
Noisy test data	55.4	60.7	79.9
Compensation (DDMC)	13.6	21.4	15.2
Enhancement (filter 1)	13.9	21.7	15.2
Enhancement (filter 2)	12.2	20.3	14.5

Table 1 *Average word error rates for different test configurations. Column 1 corresponds to F16 jet noise (SNR=6.4dB), column 2 to Lynx noise with (SNR=5.5dB) and column 3 to white noise (1dB).*

Table 1 shows that the speech recognizer performance deteriorates significantly when the system is trained on clean data and tested on noisy data. For example, training on clean speech and testing on noisy speech (SNR=1dB) generated by adding white noise, the word error rate increases from 5.9% (testing on clean speech) to 79.9%. Compensation for the test noise using DDMC significantly decreased the word error rates (compare rows 2 and 3). Using acoustic models trained on clean speech and the enhanced speech signal gives word error rates similar to those obtained with model combination. Testing with the enhanced speech based on filter 2, gives relative improvements of 4.6% for white noise, 10.3% for F16 jet noise and 5.1% for Lynx noise, compared to the results obtained using model combination. These results indicate that for the speech recognizer there is no decrease in intelligibility of the enhanced speech relative to the noisy speech. As can be predicted theoretically, the performance obtained by training and testing under matched training/testing conditions or enhancing the speech signal should be comparable when the same underlying speech models are used.

CONCLUSION

In this paper we have experimented with a MAP approach using autoregressive HMMs to enhance speech signal degraded by additive noise. This technique introduced by Ephraim et al. [6] requires a reasonable initialization of the EM estimation procedure. Here we have described a new initialization technique relying on the use of cepstral-based HMMs and a data-driven model combination technique [5]. This initialization process makes this speech enhancement technique effective in enhancing noisy speech even with very low SNR. Experiments were carried out using three types of noise taken from the NOISEX-92 database. We found the enhanced speech to be perceptively quite good, with only a slight distortion and no musical noise, although no formal perceptive tests were carried out. Speech recognition experiments with the enhanced speech show that there is no decrease of intelligibility from the viewpoint of the recognizer.

REFERENCES

[1] J.L. Gauvain, L. Lamel, G.Adda, D.Matrouf, "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task," *ICASSP-96*.

[2] M. Gales, S. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *Computer Speech & Language*, **9**(4), Oct. 1995.

[3] A. Acero, R.M. Stern, "Environmental Robustness in Automatic Speech Recognition," *IEEE Acoustics, Speech & Signal Processing*, pp. 849-852. April 1990.

[4] L. Lamel, S. Rosset, S. Bennacef, H. Bonneau-Maynard, L. Devillers, J.L. Gauvain, "Development of Spoken Language Corpora for Travel Information," *Eurospeech'95*, Madrid.

[5] D.Matrouf, J.L. Gauvain, "Model compensation for noises in test and training data," *ICASSP-97*.

[6] Y. Ephraim, D. Malah, B.H Juang, "On the Application of Hidden Markov Models for Enhancing Speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37. NO. 12. December 1989.

[7] S.F Boll, " Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE*, pp. 113-120, 1979.

[8] B. T. Logan, A. J. Robinson, "Enhancement and Recognition of Noisy Speech within an AutoRegressive HMM Framework Using Noise Estimates from The Noisy Signal," *ICASSP-97*.

[9] J.S. Lim, A.V. Oppeenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Siganl Process.*, Vol. 26, pp. 197-210, June 1978.

[10] A.P. Varga, H.J.M Steeneken, M. Tomlinson, D. Jones,"The NOISEX-92 study on the effect of additive noise on automatic speech recognition," *In Technical Report, DRA Speech Research Unit*, 1992.