A NOISE-ROBUST ASR FRONT-END USING WIENER FILTER CONSTRUCTED FROM MMSE ESTIMATION OF CLEAN SPEECH AND NOISE

Jian Wu *

The University of Hong Kong,
Department of Comp. Sci. and Info. Sys.,
Pokfulam Road, Hong Kong, China
jwu@csis.hku.hk

Jasha Droppo, Li Deng, Alex Acero

Microsoft Research,
One Microsoft Way,
Redmond WA 98052, USA
{jdroppo,deng,alexac}@microsoft.com

ABSTRACT

In this paper, we present a novel two-stage framework of designing a noise-robust front-end for automatic speech recognition. In the first stage, a parametric model of acoustic distortion is used to estimate the clean speech and noise spectra in a principled way so that no heuristic parameters need to set manually. To reduce possible flaws caused by the simplifying assumptions in the parametric model, a second-stage Wiener filtering is applied to further reduce the noise while preserving speech spectra unharmed. This front-end is evaluated on the Aurora2 task. For the multi-condition training scenario, a relative error reduction of 28.4% is achieved.

1. INTRODUCTION

It is well known that designing a noise-robust front-end is one of the essential issues in practical deployment of speech recognition technology. Among all kinds of front-ends for speech recognition proposed over past few decades, Mel-frequency cepstral coefficients (MFCC) are widely adopted because of their superiority in clean speech recognition. However, current automatic speech recognition systems are often used in environments with unexpected background noise. Their recognition rates suffer considerably because the MFCC is not immune to the distortion of speech spectra caused by the background noise. Therefore, several approaches based on *spectral enhancement* or *feature-domain compensation* have been developed upon MFCC to clean the noisy features so that they become more noise-insensitive.

On *spectral enhancement*, much work has been done by using traditional signal processing techniques, such as spectral subtraction and Wiener filtering. Such strategies have many successful applications in signal processing (e.g. [6]) because the reconstruction of clean spectra is close to optimal if the true noise spectra is given and the autocorrelation functions of the signal and the noise are known. One drawback of those techniques, however, is that they often need to implicitly assume a single Gaussian distribution on speech signals or even more heuristic assumptions to estimate the noise spectra. These assumptions are often unable to hold in many situations. Moreover, some of critical parameters in such approaches need to be finely tuned according to the specific task.

Recently, a second kind of method that uses probabilistic modelling and statistical learning techniques has attracted more attention (e.g. [5, 7, 9]). These parametric *feature-domain compensation* methods ignore the constraints in *spectral enhancement*. They use a parametric generative model of noise and speech, with a deterministic or stochastic mapping from these hidden features to the noisy observation. The generative model for speech or noise can be either a Gaussian mixture model (*GMM*) or a hidden Markov model(*HMM*). Given this structure, these methods can take advantage of the optimality properties of Maximum Likelihood (*ML*) estimation derived in a principled way [2]. However, they may also suffer from the ill-defined composed model. In most cases, the generating mechanism of noisy cepstra from clean speech and noise cepstra is highly nonlinear or even unknown. In order to have a computationally tractable model, a procedure of linearization is unavoidable. This makes the composed model quite different from the true distribution of noisy speech and thus causes a dramatic bias on the estimation of clean speech even when the true distribution of clean speech and noise is known.

In this paper, we propose a novel approach, to derive a noise-robust front-end, which takes advantage of signal-processing-based *spectral enhancement* and statistical *feature-domain compensation*. It assumes the noise cepstra of each utterance is generated from a single stationary source and estimates the distribution parameters of the source according to the sufficient statistics accumulated from current noisy speech. It tracks the clean speech and noise cepstra dynamically for each frame, which provides a first-stage estimation of clean speech power spectra and noise power spectra. These estimates are used to construct a Wiener filter for the original noisy speech power spectra, which removes noise while preserving speech spectra. As a byproduct of this work, we also show that the spectral enhancement in a perceptually relevant domain is more advantageous in achieving lower word error rates (e.g. [1]) by examining the use of MFCC and high resolution cepstral coefficient (*HRCC*) which is produced directly from the power spectra without Mel-bank smoothing, under the identical framework.

The paper is organized as follows. In Section 2, an overview of our approach is briefly illustrated and introduced. In Section 3, the basic assumption and operation of each sub-component in the proposed front-end are explained in more detail. The experiments and results on Aurora2 database are presented in Section 4 and our work is summarized in Section 5.

2. SYSTEM OVERVIEW

Fig 1. illustrates the block scheme of the proposed front-end. The upper part shows the process of estimating HRCC or MFCC y from the input noisy signal s. The only difference between these

*The work was carried on when the author was an intern at Microsoft Research

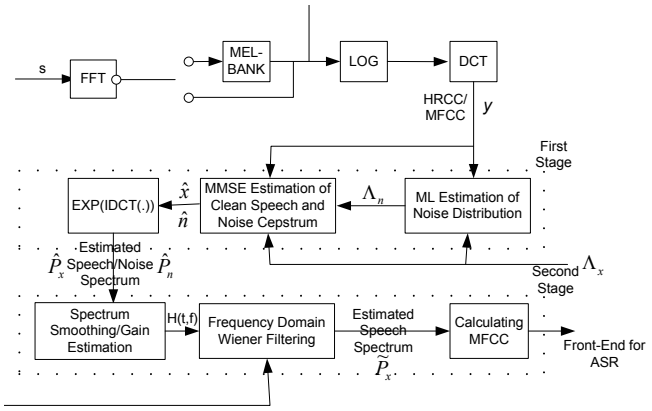

Fig. 1. Block scheme of the proposed front-end

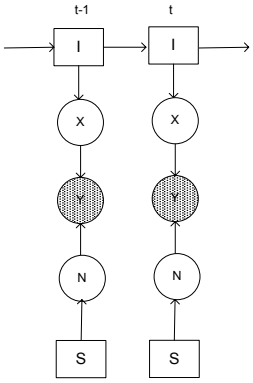

Fig. 2. Directed acyclic graphs specifying conditional independence relations for the statistical model based feature compensation

two features is that the Mel-bank filtering is not performed on the FFT-derived spectra in calculating HRCC to keep the high resolution of the features on spectra and all cepstral coefficients are remained to avoid any lose of information. Then the estimated cepstra of noisy speech are fed to the middle part of the scheme to estimate (Mel-warped) clean speech spectra and (Mel-warped) noise spectra. The middle layer is a typical application of statistical model based speech enhancement. It loads a pre-trained GMM or HMM, Λ_x, of clean speech cepstra and estimates the noise distribution parameters $\hat{\Lambda}_n = \arg\max_{\Lambda_n} p(Y|\Lambda_x, \Lambda_n)$ for current utterance $Y = \{y_1^T\}$ using EM algorithm. Then based on the up-to-date noise distribution, an MMSE estimation of clean speech and noise cepstra are computed for each frame of input speech,

$$\begin{aligned} \hat{x}_t &= \int x p(x|y_t, \Lambda_x, \Lambda_n) dx \ , \\ \hat{n}_t &= \int n p(n|y_t, \Lambda_x, \Lambda_n) dn \ . \end{aligned} \tag{1}$$

As mentioned above, the *feature-domain compensation* based on statistical approach has to take the risk of biased estimation due to the necessary approximation. In our approach, we perform a further step of frequency domain Wiener filtering based on the (Mel-warped) power spectra of $|\hat{P}_x|^2$ and $|\hat{P}_n|^2$ to reduce the influence of the biased estimation on clean speech spectra. Compared with the conventional *spectral enhancement* approach based on Wiener filtering, the noise spectra used to design the filter parameter are estimated in a more systematic fashion and thus the process of estimation becomes more portable.

The output of the Wiener filtering, $|\tilde{P}_x|^2$, is treated as the final enhanced speech spectra and the conventional MFCC are computed as the front-end used for the speech recognition. In the next section, some details of above process will be further explained.

3. PROPOSED FRONT-END

3.1. Statistical Estimation of Clean Speech Spectra and Noise Spectra

3.1.1. Graphical Representation of the Parametric Model

Fig. 2 illustrates a graphical representation of the parametric model which is also widely used in previous work such as [3, 4, 5, 9]. It assumes that the clean speech cepstra is generated from a random process with Markov chain, which is usually described as an HMM, and a GMM for noise cepstra. Y,X,N,I and S stand for the random variables of noisy, clean, noise cepstra, index of state generating the clean speech and index of Gaussian components producing noise, respectively. For the sake of simplification, in the following we will only discuss the case of GMM for clean speech and single Gaussian for noise with no loss of generality. For example, it is supposed that $x \sim \sum_i c_i \mathcal{N}(m_x(i), \Sigma_x(i))$ and $n \sim \mathcal{N}(m_n, \Sigma_n)$. $\Sigma_x(i)$ and Σ_n represent diagonal covariance matrixes.

Suppose that we ignore other effects, such as linear channel, and only consider the additive noise. The parametric model of noisy speech cepstra given x and n can be approximated by

$$y_t = C \log(\exp(C^{-1} x_t) + \exp(C^{-1} n_t)) + \epsilon_t \ , \tag{2}$$

where y_t, x_t, n_t are noisy speech, clean speech and noise cepstra at t-th frame, respectively. ϵ_t is the residue error which is assumed to be a zero-mean Gaussian random variable. C is the corresponding matrix of DCT.

3.1.2. ML Estimation of Background Noise Distribution

The basic assumption of the statistical estimation in this paper is that the distribution parameter for noise frames, $\Lambda_n = \{m_n, \Sigma_n\}$, are fixed but unknown while the parameter of clean frames, $\Lambda_x = \{c_i, m_x(i), \Sigma_x(i)\}$, are known in advance by training over available clean speech. The problem we want to solve in this section is, given y_1^T and Λ_x, to estimate the distribution of noise for current utterance, Λ_n, to maximize the joint probability of $P(Y, X, N, I|\Lambda_x, \Lambda_n)$.

Similar to the derivation in [10], the general equation to update m_n and Σ_n by an EM process is

$$\hat{m}_n = \frac{\sum_t \sum_i p(i|y_t) E\{n_t|y_t, i\}}{\sum_t \sum_i p(i|y_t)} \tag{3}$$

$$\hat{\Sigma}_n = diag\Big[\frac{\sum_t \sum_i p(i|y_t) E\{n_t n_t'|y_t, i\}}{\sum_t \sum_i p(i|y_t)} - \hat{m}_n \hat{m}_n'\Big] \tag{4}$$

where

$$p(i|y) = \frac{c_i p(y|i)}{\sum_k c_k p(y|k)} , \quad (5)$$

$$p(y|i) = \int\int p(y|n,x)p(n)p(x|i)dndx . \quad (6)$$

In order to have a closed-form solution of the integration in Eq. (6), the nonlinear function (2) is often approximated by its Taylor series expansion on the operating points x_0 and n_0,

$$y_t \approx A_0 + G(x_0,n_0)(x_t - x_0) + (I - G(x_0,n_0))(n_t - n_0) + \epsilon_t$$

where

$$A_0 = C\log[\exp(C^{-1}x_0) + \exp(C^{-1}n_0)]$$

and

$$G(a,b) = C\frac{\exp(C^{-1}a)}{\exp(C^{-1}a) + \exp(C^{-1}b)}C^{-1} .$$

Under this Taylor series approximation, the distribution of y given x and n becomes a normal distribution,

$$p(y|x,n) = \mathcal{N}(y; A_0 + G(x_0,n_0)(x - x_0) + (I - G(x_0,n_0))(n - n_0), \Sigma_\epsilon) , \quad (7)$$

where Σ_ϵ is the covariance matrix of ϵ which is often assumed to be diagonal. And,

$$p(i|y) = \mathcal{N}(y; \mu_y(i), \Sigma_y(i)) ,$$

where

$$\mu_y(i) = A_0 + G_0(m_x(i) - x_0) + (I - G_0)(m_n - n_0)$$
$$\Sigma_y(i) = (I - G_0)\Sigma_n(I - G_0)' + G_0\Sigma_x(i)G_0' + \Sigma_\epsilon .$$

For simplicity of computation, G_0 is assumed as a matrix with the same diagonal elements as $G(x_0,n_0)$ but zero elements in off-diagonal positions, which results in a diagonal matrix of $\Sigma_y(i)$.

In order to calculate the expectation of n_t and $n_t n_t'$ for each frame, we should also know $p(n|y,i)$, which is calculated by

$$p(n|y,i) = \frac{p(n)\int p(y|n,x)p(x|i)dx}{p(y|i)} = \mathcal{N}(n; \mu_{n|y}(i), \Sigma_{n|y}(i)) , \quad (8)$$

where

$$\mu_{n|y}(i) = m_n + A(y,i) ,$$
$$\Sigma_{n|y}(i) = [(I - G_0)\Sigma_n(I - G_0)' + G_0\Sigma_x(i)G_0' + \Sigma_\epsilon]^{-1} (G_0\Sigma_x(i)G_0' + \Sigma_\epsilon)\Sigma_n ,$$

and

$$A(y,i) = [(I - G_0)\Sigma_n(I - G_0)' + G_0\Sigma_x(i)G_0' + \Sigma_\epsilon]^{-1} (I - G_0)\Sigma_n(y - \mu_y(i)) .$$

Then it is easy to derive that

$$E\{n_t|y_t,i\} = \mu_{n|y_t}(i) , \quad (9)$$

$$E\{n_t n_t'|y_t,i\} = \Sigma_{n|y}(i) + \mu_{n|y_t}(i)\mu_{n|y_t}'(i) , \quad (10)$$

and thus

$$\hat{m}_n = m_n + \frac{\sum_t\sum_i p(i|y_t)A(y_t,i)}{\sum_t\sum_i p(i|y_t)} \quad (11)$$

$$\hat{\Sigma}_n = diag\Big[\frac{\sum_t\sum_i p(i|y_t)[A(y_t,i)A(y_t,i)' + \Sigma_{n|y_t}(i)]}{\sum_t\sum_i p(i|y_t)} - (\hat{m}_n - m_n)(\hat{m}_n - m_n)'\Big] . \quad (12)$$

Accordingly, the noise distribution parameters, Λ_n can be updated iteratively.

Note that, the convergence of Eq. (11) and (12) would become very slow if Σ_n is small because $A(y,i)$ will become very small. Therefore, similar to [10], the convergence of iteration can be sped up by updating m_n to maximize $P(Y, I|\Lambda_x, \Lambda_n)$ instead of $P(Y, X, N, I|\Lambda_x, \Lambda_n)$. By setting the derivative of corresponding auxiliary function, in M-step of EM process, with respect to m_n as zero, we can have

$$\hat{m}_n = m_n + \frac{\sum_t\sum_i p(i|y_t)(I - G_0)\Sigma_y^{-1}(i)\big[y_t - \mu_y(i)\big]}{\sum_t\sum_i p(i|y_t)(I - G_0)\Sigma_y^{-1}(i)} \quad (13)$$

Eq. (12) is still used to update $\hat{\Sigma}_n$.

As mentioned above, there were several algorithms based on the parametric model shown in Eq. (2) and different statistic learning approaches, proposed in recent years (e.g. [3, 4, 9]). In [3] and [4], the noise was assumed to be generated from a nonstationary source, and an sequential EM algorithm or a prior evolution process are used to track the changing noise. In both approaches, a forgetting mechanism is introduced to remove the influence of the past observations. But such algorithms may expectedly [4] be undesirable if the noise source is stationary or changed very slowly. It is because the forgetting factor is prefixed at a non-zero value [3] or at a zero value [4], rather than made adaptive. Therefore, given the application scenario where the noise source is stationary and waiting for a batch of observations is affordable, it would be more desirable to estimate the noise distribution by a batch-mode EM algorithm, which is used in our approach in this paper. However, there is no fundamental difficulties in replacing the sequential algorithm of noise tracking in [3, 4] into the first-stage processing shown in Fig. 1 to deal with the nonstationary noise tracking. [9] uses the same parametric model as described in this section. But in our approach, we update $\hat{m}_n$ using Eq. (13) so that the convergence is much faster.

3.1.3. Update of Σ_ϵ

In the above equations, it is not mentioned how to estimate the covariance matrix, Σ_ϵ, of the residue error ϵ. Although it can also be derived analytically with an iterative EM process by

$$\hat{\Sigma}_\epsilon = diag\Big[\frac{\sum_t\sum_i p(i|y_t)E\{\epsilon_t\epsilon_t'|y_t,i\}}{\sum_t\sum_i p(i|y_t)}\Big] ,$$

in our approach, the exact estimation is not adopted because it will involve a lot of computations. Instead, we either set it as zero or approximate it using the following formula,

$$\hat{\Sigma}_\epsilon = max\Big(0, \Sigma_\epsilon + diag\Big[\frac{\sum_t\sum_i p(i|y_t)\big[(y_t - \mu_y(i))(y_t - \mu_y(i))' - \Sigma_y(i)\big]}{\sum_t\sum_i p(i|y_t)}\Big]\Big) \quad (14)$$

The *max* operation is to ensure the value of the diagonal convariance matrix Σ_ϵ are nonnegative.

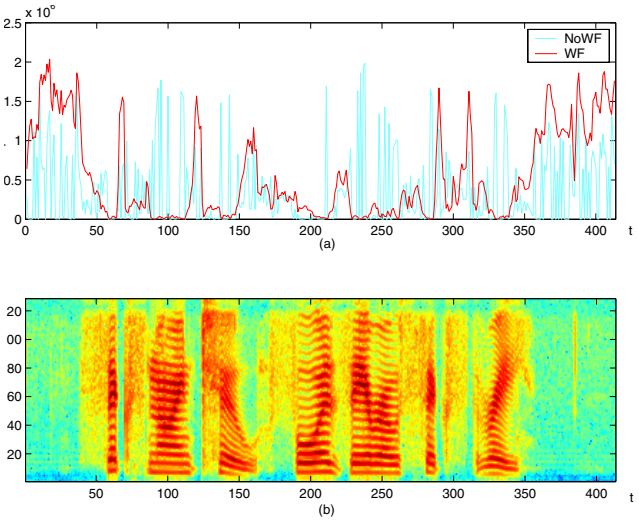

Fig. 3. (a) Square error of estimated C0 compared with the clean cepstra ($SquareError = |C0_{est} - C0_{clean}|^2$); (b) Corresponding linear spectra of clean speech.

3.1.4. MMSE Estimation of Clean Speech and Noise Cepstra

According to (1), given the distribution parameters of noise cepstra, Λ_n, the noise cepstra is easily estimated by

$$\begin{aligned} \hat{n}_t &= \int np(n|y_t)dn = \sum_i p(i|y_t) \int np(n|y_t, i)dn \\ &= \sum_i p(i|y_t)\big[m_n + A(y_t, i)\big] \ . \end{aligned} \tag{15}$$

Similarly,

$$\hat{x}_t = \sum_i p(i|y_t)\big[m_x(i) + B(y_t, i)\big] \ , \tag{16}$$

where

$$\begin{aligned} B(y,i) &= [(I - G_0)\Sigma_n(I - G_0)' + G_0\Sigma_x(i)G_0' + \Sigma_\epsilon]^{-1} \\ & \quad G_0\Sigma_x(i)(y - \mu_y(i)) \ . \end{aligned}$$

3.2. Frequency Domain Wiener Filtering

Given the estimate of the cepstral coefficients for each frame, $\hat{x}_t$ and $\hat{n}_t$, their power spectra $|\hat{P}_x|^2$ and $|\hat{P}_n|^2$ can be reconstructed. It is no doubt that the enhancement process described as above can be applied on either HRCC or MFCC to enhance the spectra or a Mel-warped spectra. The (Mel-warped) power spectra is smoothed in both time and frequency domain to reduce variances due to erroneous estimation in the first stage. Then, modified Wiener filtering, given by the following equation, is estimated for each frame,

$$|H(t,f)| = \frac{|\hat{P}_y(t,f)|^2 - \alpha(t)|\hat{P}_n(t,f)|^2}{|\hat{P}_y(t,f)|^2} \ , \tag{17}$$

where

$$|\hat{P}_y(t,f)|^2 = |\hat{P}_x(t,f)|^2 + |\hat{P}_n(t,f)|^2 \ , \tag{18}$$

and $\alpha(t)$ is a factor that avoids over estimation of the noise spectra. Its value varies from 0.6 to 0.95 according to the local SNR computed from $|\hat{P}_x|^2$ and $|\hat{P}_n|^2$. t and f are the time and frequency indices, respectively (If MFCC is used, then f is the indices of the Mel filter bank).

The finally estimated speech power spectra is given by applying the Wiener filtering on the original power spectra of noisy speech,

$$|\tilde{P}_x(t,f)|^2 = |P_y(t,f)|^2 \cdot |H(t,f)| \ . \tag{19}$$

In the following, the MFCCs derived from the power spectra $|\hat{P}_x|^2$ are used in the experiments shown as "No WF", while "WF" stands for the experiments using MFCCs derived from $|\tilde{P}_x(t,f)|^2$.

Table 1. Square error of other cepstra over the whole utterance ($*10^5$)

	C1	C2	C3	C4	C5	C6
No WF	2.72	0.76	0.29	0.19	0.16	0.13
WF	2.08	0.54	0.27	0.14	0.16	0.10
	C7	C8	C9	C10	C11	C12
No WF	0.11	0.07	0.06	0.05	0.03	0.03
WF	0.09	0.07	0.05	0.04	0.03	0.03

The square error of estimated first order MFCC (C0) with/without Wiener filtering are plotted in Fig. 3(a) while Fig. 3(b) is the corresponding clean speech log spectra. It is found from the figure that, in general, the proposed Wiener filtering can reduce the square error to some extent, especially for the voicing portion of speech, although it may introduce higher errors in the non-speech part. The overall square errors of other order cepstra are listed in Table 1. Wiener filtering reduces the squared error for almost every cepstral component. It also provides a strong evidence supporting our assumption that combining the modified Wiener filtering with *feature-domain compensation* can further reduce unwanted noise beyond the MMSE estimate while keeping the speech spectra unharmed.

4. EXPERIMENTS AND RESULTS

The task used to verify our front-end is the speaker independent recognition of connected digit strings. The baseline recognition results presented in this section are produced on the Aurora2 database using the modified reference of WI007 for Aurora front-end evaluation [8]. In this modified front-end, for each frame, a 39-dimensional feature vector is generated, which consists of 12 MFCCs and MFCC of order 0, plus their first and second order derivatives. Another modification on WI007 is that the cepstra are computed based on the power spectral density instead of the magnitude spectra. In all of our tests the complex back-end based on whole-word model with 20 Gaussian components with diagonal covariance matrices in each state of CDHMM is used. Each left-to-right CDHMM consists of 18 states. Besides, a pause model, "sil", is created to model the silence before/after the digit string and the short pause between any two digits.

In the baseline system of clean-condition training, all of the CDHMMs are trained from the clean speech while in the baseline system of multi-condition training, they are trained from the collection of 8440 utterances that come from 20 subsets representing 4 different noise scenarios (i.e., *suburban train, babble, car* and

exhibition hall) at 4 different SNRs (i.e., 20dB, 15dB, 10dB and 5dB) and the *clean condition*.

Besides the baseline system, we also train a GMM with 256 Gaussian components over the static HRCC or MFCC of the clean training set as the reference model of clean speech for the parametric model based compensation module.

The test set of Aurora2 task consists of three different parts. For the Test Set A, the same four types of noises as those in training set are added to its subsets, but with 7 different SNRs. For the Test Set B, another 4 types of noises (i.e., *restaurant, street, airport* and *train station*) are added to its subset with also 7 SNRs. For the Test Set C, *suburban train* and *street* noises are used as the additive noise sources but the speech and noise are filtered with a MIRS characteristic while the G.712 characteristic is used in training set as well as the first two test sets.

4.1. Effects of using Wiener Filtering

Table 2 shows the accuracy on Test Set A under the clean-condition training scenario. *BASELINE* stands for the baseline system. The second raw, shown as *No WF*, are the results of that using the enhanced speech spectra output from the first stage, $|\hat{P}_x|^2$, to calculate MFCC. The third raw, shown as *WF*, list the results of that using the $|\tilde{P}_x|^2$, the output of Wiener filtering, to calculate the MFCC. In this experiment, it is found that the system using our front-end can reduce the word error rates by over 45%.

Table 3 shows the results of multi-condition training produced by the same systems used in Table 2. It is observed that the purely parametric model based approach can not guarantee that the word error rate can be reduced in all cases of multicondition training, especially for the low-SNR utterances. The overall accuracy is even worse than that of baseline system. But with the help of modified Wiener filtering, the word error rates of every SNR level are reduced dramatically by 16% over baseline system.

Table 2. Aurora2 Accuracy (%) (Cleancondition Training: Set A, HRCC, GMM for Λ_x)

	20db	15db	10db	5db	0db	-5db
BASELINE	97.61	93.10	78.46	49.44	24.27	13.34
No WF	98.58	96.85	91.85	78.23	47.94	18.63
WF	98.97	97.03	89.72	76.21	42.62	11.25

	BASELINE	No WF	WF
Average from 20db to 0db	68.58	82.69	80.91

Table 3. Aurora2 Accuracy (%) (Multicondition Training: Set A, HRCC, GMM for Λ_x)

	20db	15db	10db	5db	0db	-5db
BASELINE	99.00	98.30	96.81	91.45	69.88	29.55
No WF	99.05	98.37	96.15	89.44	68.66	30.86
WF	99.13	98.66	97.32	92.85	74.62	34.30

	BASELINE	No WF	WF
Average from 20db to 0db	91.09	90.33	92.52

4.2. GMM vs HMM for Λ_x

In this experiment, the use of more detail model for clean speech model Λ_x is under examination. From the original GMM, we create a new HMM prototype with 256 states, each having a single Gaussian with same mean and covariance matrix as one of Gaussian components in GMM. By designing an initial transition matrix, we allow the transition can occur between any two states. And then several EM iterations are performed to update the parameters of this new HMM. It is easy to extend the equations in Section 3.1 to the case of HMM for Λ_x by calculating $p(i|y_t)$ using the HMM. The results shown in Table 4 imply that, with a more precise description of the clean speech, we can have further improvement on the performance with the proposed front-end, although it is not too much.

Table 4. Aurora2 Accuracy (%) (Multicondition Training: Set A, HRCC with Wiener filtering)

	20db	15db	10db	5db	0db	-5db
BASELINE	99.00	98.30	96.81	91.45	69.88	29.55
GMM	99.13	98.66	97.32	92.85	74.62	34.30
HMM	99.19	98.61	97.23	92.89	76.17	36.51

	BASELINE	GMM	HMM
Average from 20db to 0db	91.09	92.52	92.82

4.3. HRCC vs MFCC

Above experiments are all carried on the high resolution cepstra coefficients, which contain complete information of power spectra. However, it has been indicated by many studies in the past that the error minimization in a perceptually relevant domain is more advantageous (e.g.[1]). In our approach, we can use HRCC in the first stage to enhance the high resolution power spectra, $\hat{P}_x$, or MFCC to recover the Mel-warped power spectra. The following experiments, shown in Table 5, are designed to compare the recognition results of using these two alternatives. It can be shown that using MFCC is much better than that of HRCC, although they only contain a rough information of the power spectrum.

Table 5. Aurora2 Accuracy (%) (Multicondition Training: Set A, GMM for Λ_x, and Wiener filtering is applied)

	20db	15db	10db	5db	0db	-5db
BASELINE	99.00	98.30	96.81	91.45	69.88	29.55
HRCC	99.13	98.66	97.32	92.85	74.62	34.30
MFCC	99.26	98.79	97.39	93.44	78.02	40.79

	BASELINE	HRCC	MFCC
Average from 20db to 0db	91.09	92.52	93.38

4.4. Full Evaluation on Aurora2

Table 6 lists the full evaluation results of proposed front-end on Aurora2 under multi-condition training, including Test Set B, which

is distorted by the noise different from the training set, and Test Set C with a different channel distortion. In this experiment, cepstral mean normalization is performed after speech enhancement. Comparing the results of Set A, B and C, it can be observed that the accuracy in Test Set B and C are almost same as that of Set A for our proposed front-end, which results in an relative error reduction of 28.4% over baseline systems.

Table 6. Aurora2 Accuracy (%) (Multicondition Training: Full Evaluation, GMM for Λ_x, MFCC with Wiener filtering)

	Noise1	Noise2	Noise3	Noise4	Overall
SET A	94.95	91.87	93.99	93.32	93.53
SET B	93.03	93.13	94.05	93.76	93.50
SET C	94.40	92.85	-	-	93.62
Ave					93.53
Baseline					90.96

5. SUMMARY

In this paper, we present a two-stage framework where a parametric-model-based feature compensation is followed by frequency-domain Wiener filtering, to derive a noise-robust front-end for automatic speech recognition. In this framework, the clean speech and noise spectra are first estimated in a principled fashion under the MMSE criterion. It is followed by the second-stage processing module, which is based on the modified Wiener filtering, to remedy the possible flaw in enhanced speech spectra while removing the noise as much as possible. One of the advantages of this two-stage framework is that, there is almost no parameters to be set manually and thus it works for various noisy environments without any adjustment.

It is obvious that there remains residue distortion even after the enhancement as described above. In [11], an advanced statistical model, namely switching linear Gaussian hidden Markov model, was proposed to compensate the non-stationary distortion in noisy speech, and a significant improvement has been observed [12]. It thus would be interesting to build this advanced model over the enhanced speech presented in this paper and verify whether the word error rates can be further reduced. Furthermore, since the Aurora2 database used in the experiments of this paper is artificially created, it would also be interesting to apply the proposed front-end on more real noisy database to verify its effectiveness in our future work.

6. REFERENCES

[1] A. Agarwal and Y.-M. Cheng, "Two-stage Mel-warped Wiener filter for robust speech recognition", *Proc. IEEE-ASRU workshop 1999*.

[2] H. Attias and L. Deng, "A new approach to speech enhancement by a microphone array using EM and mixture models", *Proc. ICSLP 2002*, Denver, US.

[3] L. Deng, J. Droppo and A. Acero, "Recursive noise estimation using iterative stochastic approximation for steroe-based robust speech recognition", *Proc. IEEE ASRU workshop 2001*, ITALY.

[4] L. Deng, J. Droppo and A. Acero, "Incremental Bayes learning with prior evolution for tracking nonstationary noise statistics from noisy speech data", *Proc. IEEE ICASSP 2003*, Hong Kong, CHINA.

[5] J. Droppo, L. Deng and A. Acero, "A comparison of three non-linear observation models for noisy speech recognition", *Proc. Eurospeech 2003*, Geneva, Switzerland, September 2003.

[6] "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms", ETSI ES 202 050 v1.1.1(2002-10), October 2002.

[7] B. J. Frey, L. Deng, A. Acero, T. Kristjansson, "ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition", *Proc. Eurospeech 2001*.

[8] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", *ISCA ITRW ASR2000*, Paris, FRANCE, September 2000.

[9] T. Kristjansson, B. Frey, L. Deng and A. Acero, "Joint estimation of noise and channel distortion in a generalized EM framework", *Proc. IEEE ASRU workshop 2001*, ITALY.

[10] A. Sanka and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 3, pp. 190-202, 1996.

[11] J. Wu and Q. Huo, "A switching linear Gaussian hidden Markov model and its application to nonstationary noise compensation for robust speech recognition", *Proc. Eurospeech 2003*, Geneva, Switzerland, September 2003.

[12] J. Wu and Q. Huo, "Several HKU approaches for robust speech recognition and their evaluation on Aurora connected digit recognition tasks", *Proc. Eurospeech 2003*, Geneva, Switzerland, September 2003.