A NOVEL NOISE ROBUST FRONT-END USING FIRST ORDER VTS IN CONSTRUCTION OF MEL-WARPED WIENER FILTER[*]

Mu Su, Peng Li, Zhuo Wang, Peng Ding, Bo Xu

Hi-Tech Innovation Center, Institute of Automation
Chinese Academy of Sciences, Beijing, China, 100080
Email: {msu, pengli, zwang,pding, xubo} @hitic.ia.ac.cn

ABSTRACT

In this paper, we first review two approaches in the context of robust recognition, e.g. speech enhancement based two-stage mel-warp wiener filtering (MWF) [1] and first-order Vector Taylor Series (VTS) [2] compensation in log power spectrum, which are widely used. A new noise robust front-end is proposed, in which VTS compensation derived statistics are used to construct the mel-warped wiener filter. We will show that this noise robust front end is superior. The experiments results prove that our proposed method does show significant improvement over VTS and MWF.

1. INTRODUCTION

Noise severely degrades the performance of a state-of-the-art recognition system when it is trained in the clean condition. For this reason, robust speech recognition has always been one of the challenging areas of statistical speech research. Noise has two main effects on clean speech. First, it introduces a distortion (usually nonlinear transformation) in the feature space which causes a mismatch between training and recognition conditions. Secondly, for the sake of its random nature, it inevitably causes a loss of information. For analysis convenience, noise is further divided into two categories: additive noise and convolutive noise. In this paper, only additive noise will be considered. There has been considerable research to account for the effect of noise, and hence increase the robustness of the system. The previous work can be broadly classified into four main categories: Using inherently robust feature parameterizations, speech enhancement techniques [3] whose goal is to increase SNR, feature space compensation which aims to recover clean features as accurate as possible given noisy feature observations and adapting the parameters of model to better represent the speech under noise condition.

One of the successful uses of spectral enhancement is two-stage mel-warped wiener filtering as in [1] and shows a comparable performance with other noise reduction techniques. The main reason for this successful use of wiener filtering is that wiener filter is constructed in the perceptually relevant domain, which proves to be advantageous in obtaining lower word error rates [1]. However, its drawbacks also can not be neglected. One of the main drawbacks is the assumption of the precise estimation of wiener filter parameters which is derived from a rather coarse recursive estimation of noise spectrum.

On the other hand, as a successful implementation of feature space domain compensation, VTS [2] has become widely used in robust recognition. As is known, the effect of the additive noise results in a non-linear transformation of the representation space in the log filter-bank-energy (log FBE) domain; VTS is introduced to approximate this non-linearity in log FBE domain by its Taylor series expansion. The resulting clean log FBE estimate is based on the MMSE criteria. Because of the non-linear nature of noisy signal, this linear approximation can lead to a large bias on the estimation of clean speech, especially in extremely low SNR conditions.

In this paper, we present a novel method which uses first-order VTS compensation-derived statistics for the construction of wiener filter in perceptually relevant domain. The paper is organized as follows. In Section 2, our proposed new front-end is presented in detail. The experiments and results are presented and comparisons are made with respect to VTS and MWF and our work is summarized in Section 4.

2. THE PROPOSED NOISE ROBUST FRONT-END

2.1 motivations of the proposed front-end

The motivations of the proposed new front-end are shown below. In [1], Agarwal proposed a novel two-stage mel-warped wiener filter algorithm for robust recognition.

The noisy signal with additive noise can be expressed as $y(t)=x(t)+n(t)$, where $y(t)$, $x(t)$ and $n(t)$ are noisy signal, speech signal and noise. Their auto-correlation and power spectrum counterparts are defined as $R_y=R_x+R_n$ and $P_y(w)=P_x(w)+P_n(w)$ respectively. A wiener filter is then constructed as

[*] The work is supported by National High Technology R&D program 863 of China under contract 2005AA114070

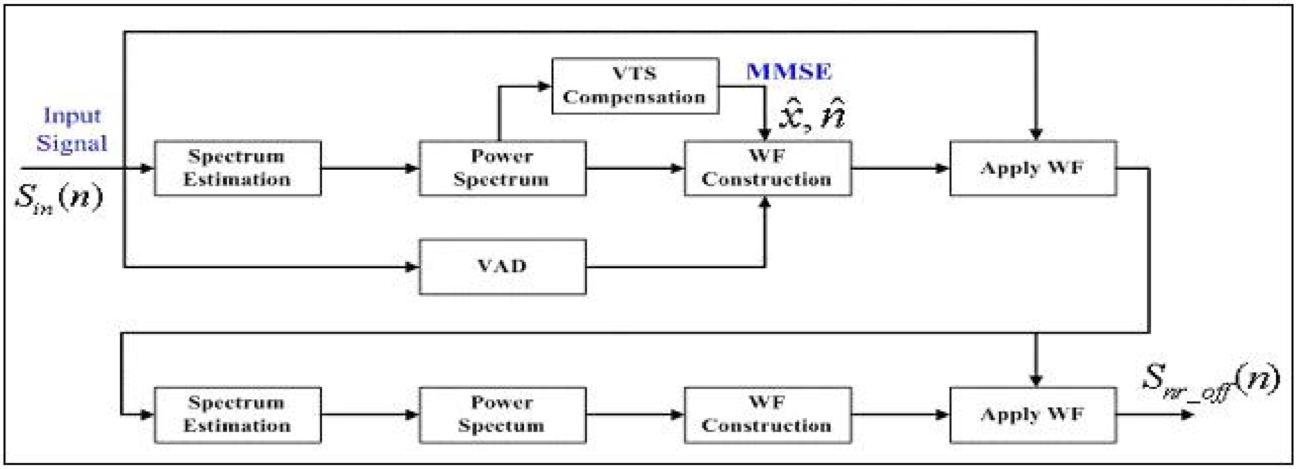

Figure 1 Block Scheme of the Proposed Noise Robust Front-end

$$H(w) = \frac{P_y(w) - \hat{P}_n(w)}{P_y(w)} \qquad (1)$$

The hat on the letter *P* stands for the estimate of power spectrum. While wiener filter minimizes the objective error of the above approximation, many studies have shown that the error minimization in the perceptually relevant domain is more advantageous since the widely used recognition system front-end is in a perceptual domain, mel frequency.

Therefore, wiener filter transfer function can be expressed in mel-warped filter bank domain. Finally, wiener filtering is performed in the time domain, where noisy signal convolves with the impulse response of mel-warped wiener filter.

From the equation 1, it is obvious that the key part of the algorithm is noise estimation. However, the original implementation assumes that the noise statistics for the construction of wiener filter can be precisely estimated using a simple recursive way, which is hardly so in low SNR conditions.

To overcome this, we propose a novel noise robust front-end, in which VTS compensation is introduced to obtain a more systematic noise estimate.

2.2 System Overview

Figure 1 illustrates the block scheme of the proposed noise robust front-end.

The upper part of block diagram shows the first stage mel-warped wiener filtering process, in which MMSE-based first order VTS compensation is performed and the estimated clean log-spectral feature and its noise counterpart are fed into the next block for the construction of mel-warped wiener filter. The lower part shows the second stage mel-warped wiener filtering process, where a simple recursive noise estimation pattern is used for computational efficiency. Since the noise is largely suppressed in the first stage and only residue noise remains, we did not find any significant improvement in our preliminary experiments to implement VTS compensation for the estimation of noise in second stage.

In [8], Wu presented a front-end which is a modified version of mel-warped wiener filtering. The difference is that in our system, VTS compensation is explicitly used to obtain MMSE-based speech and noise statistics. Secondly, speech and noise are estimated in log FBE domain in stead of Cepstrum domain for computational efficiency.

2.3 Noise Estimation Using VTS Compensation

In our proposed front-end, noise is estimated through the use of VTS compensation-based MMSE criteria, which is illustrated below in detail.

2.3.1. Environmental function in log-spectral domain

In log-spectral domain, the effect of additive noise on clean feature satisfies the following equation:

$$y=x+log(I+e^{(n-x)}) \qquad (2)$$

where $y=[y_0,\ldots,y_{K-1}]$, $x=[x_0,\ldots,x_{K-1}]$ and $n=[n_0,\ldots,n_{K-1}]$ denote log-spectral vectors of noisy signal, clean speech and additive noise respectively. It is obvious to see that the effect of additive noise can be regarded as a non-linear transformation of the original clean speech features.

2.3.2. Construction of noisy speech distribution using VTS

In log-spectral domain, the distribution of noise can be modeled as a multivariate Gaussian probability density

$$p(n) = \prod_k N(n_k; \mu_{nk}, \sigma^2_{nk}) \qquad (3)$$

with mean μ_{nk} and variance σ_{nk}^2, while that of speech is modeled as the GMM.

$$p(x) = \sum_m c_m \prod_k N(x_{mk}; \mu_{xmk}, \sigma_{xmk}^2) \quad (4)$$

In [2], the noise model parameters are trained using EM algorithm. In this paper, a VAD-based iterative noise model parameters estimation technique is adopted as in our previous work in [7].

Because of the non-linear nature of environmental function, the noisy speech is not Gaussian. However, using VTS expansion of the environmental function, noisy signal statistics can also be modeled as a GMM in equation (5). As in [7], it can be concluded that first-order Taylor series expansion is accurate enough to give desirable performance, in the following deduction, first-order Taylor series will be adopted, and then the parameters of *p(y)* can be formulated as:

$$p(y) = \sum_m c_m \prod_k N(y_{mk}; \mu_{ymk}, \sigma_{ymk}^2) \quad (5)$$

in which

$$\mu_{ymk} = x_{xmk} + \log(1 + e^{(\mu_{nk} - \mu_{xmk})})$$

$$\sigma_{ymk}^2 = \sigma_{xmk}^2 + \Delta_{mk} \quad (6)$$

$$\Delta_{mk} = (\sigma_{xmk}^2 + \sigma_{nk}^2)\beta_{mk}^2 - 2\beta_{mk}\sigma_{xmk}^2$$

$$\beta_{mk} = (1 - \frac{1}{1 + e^{(\mu_{nk} - \mu_{xmk})}})$$

2.3.3. Log-spectral Compensation based on MMSE criteria

Since speech log-spectra are modeled as GMM and each component of them is independent of each other, the compensation algorithm can be expressed as a weighted sum of the MMSE estimators over all components:

$$\hat{n} = y - \hat{x} \quad (7)$$

$$\hat{x} = \sum_m E[x|y,m] p(m|y) \quad (8)$$

where

$$p(m|y) = \frac{c_m p(y|m)}{\sum_i c_i p(y|i)}$$

$$E[x|y,m] = y - E[f(\Delta)|y,m] \quad (9)$$

in which $f(\Delta) = f(n-x) = \log(I + e^{(n-x)})$ stands for the non-linear component of the environmental function. Expanding $f(\Delta)$ at point $\mu_\Delta = \mu_n - \mu_x$ using first-order VTS, equation (8) can be rewritten as

$$E[x|y,m] \approx y - E[f(\mu_\Delta)|y,m] = y - f(\mu_\Delta) \quad (10)$$

2.4 Detailed Specification of the Proposed Front-end

In this subsection, detailed procedures of the proposed front-end are listed below.

Firstly, the noisy speech signal (frame length 25ms, frame shift 10ms) $S_{in}(n)$ is fed into spectrum estimation block, where it is windowed by hanning window and 512 point FFT is performed.

Secondly, in power spectrum block, power spectrum is obtained and smoothed, and then mel filter bank operation is performed.

Thirdly, VTS Compensation is performed with MMSE-based noise estimate $\hat{n}$ in log-spectral domain using equation (5)(6)(7)(8) and (10), from which mel-warped wiener filter is constructed using equation (1).

Finally, the impulse response of the constructed wiener filter is convolved with the input noisy signal to get the enhanced signal.

The same process repeats itself in second stage, except that VAD block is omitted and a simple recursive style is used to estimate the noise in log-spectral space. Normal MFCC process is performed to the denoised speech signal $S_{nr\_off}(n)$ to form final feature vectors.

3. EXPERIMENT RESULTS

Our baseline recognition system is a speaker-independent large vocabulary continuous speech recognizer. The acoustic model is a class-triphone model, which is trained using decision tree to generate state-tied triphone. The decoding approach is based on the time synchronous one-pass decoder with trigram language model [6].

In our system, each frame is represented by a 42-dimensional feature vector that consists of normalized log energy, 12 mel-frequency cepstral coefficients (MFCCs) and pitch along with their first and second differentials. The acoustic analysis uses 25ms windows and 10ms frame shift.

The training set and test set are all from the standard Mandarin continuous speech corpuses under the project 863.The training set contains 48000 sentences from 83 male speakers.

In the baseline system, the clean acoustic model is trained using these 48000 utterances. While for the rest of the noise robust front-end, these sentences are equally split up into 20 subsets with each subset adding one noise at one SNR. The four noise styles from NOISEX92 [5] database are Babble, White, Factory and Leopard. The 5 SNRs are Clean, 20dB, 15dB, 10dB and 5dB.The test set, namely corpus98, contains 600 sentences from 10 male speakers. 4 different kinds of noises (Babble, White, F16, Destroy) at 5 SNRs (20dB, 15dB, 10dB, 5dB, 0dB) are added to the test set to form 12000 sentences, in which Babble, White are highly matched with training set, while F16 and Destroy represent the mismatched conditions from training set. All the utterances are recorded at a sampling rate of 16KHz with 16bit resolution. Our proposed front-end is compared against that of VTS and MWF [1].

In our experiment, Noise Adaptive Training (NAT) style is used [4] to dramatically reduce the overall acoustic variation across the range of noise types and levels.

Table 1 shows the average recognition accuracy over different noises of the proposed front-end, compared with that of VTS and MWF. In high SNR conditions, except for the baseline, the superiority of the proposed method over the others is small. While at SNR 10dB and 5dB, the proposed front-end outperforms MWF by 1.92% and 2.46%, VTS by 6.14% and 13.61%. Unfortunately, the advantage of the proposed one over MWF disappears at SNR 0dB. At 0dB SNR, the magnitude of noise and that of speech is almost the same, which makes it hard for the energy-based VAD to tell which is noise or speech. In our front-end framework, VAD is used as a flag for iterative update of noise statistics in equation (3). Also in Table 1, the performance of VTS drops drastically below SNR 15dB. The reason for this dramatic drop is that VTS feature compensation suffers a risk of biased estimation of clean speech and noise due to the first-order expansion approximation even when the true distribution of clean speech and noise is known, especially in low SNR conditions.

	20dB	15dB	10dB	5dB	0dB
Baseline	85.29	78.76	63.70	34.00	9.68
VTS	86.58	83.98	76.99	60.08	28.08
MWF	86.60	85.43	81.21	71.23	46.07
Proposed	87.29	86.72	83.13	73.69	46.09

Table 1 Averaged Recognition Accuracy (%) over different noises of the proposed front end compared with that of VTS, MWF and Baseline System

Table 2 gives the average recognition performance of different noise robust front-end over different SNR conditions. Our proposed algorithm shows a moderate increase over MWF both in matched conditions and mismatched conditions.

	Babble	White	Destroy	F16
Baseline	66.75	42.42	50.69	57.28
VTS	75.05	63.71	59.97	69.84
MWF	75.54	72.14	73.38	76.96
Proposed	75.83	73.42	74.40	77.87

Table 2 Averaged Recognition Accuracy (%) over different SNRs of the proposed front end compared with that of VTS, MWF and Baseline System

To test the performance of the proposed front-end in real environment, the evaluation set (Eva04) of National Evaluation sponsored under the project 863 is introduced. Eva04 consists of 100 sentences recorded male speaker in real noise environments, such as in street, in coffee bar and etc. The average SNR is about 10dB. Table 3 shows the performance of the proposed front-end against VTS, MWF and baseline system.

	Baseline	VTS	MWF	Proposed
Accuracy (%)	33.3	64.1	65.2	66.7

Table 3 Recognition Accuracy of proposed front-end against that of VTS, MWF and Baseline

4. CONCLUSIONS

In this paper, a novel noise robust front-end is explored and superior performance can be obtained. The contribution of this paper lies in the idea of using VTS compensation to derive the estimated speech and noise for the construction of mel-warped wiener filter. The experiments show that significant reductions in WER by at least 1~2% absolute are observed compared to VTS and MWF.

5. REFERENCES

[1] A.Agarwal, Y.M.Cheng, “Two-stage Mel-warped Wiener Filter for Robust Speech Recognition” *Proc. IEEE ASRU workshop 1999.*

[2] P. J. Moreno, B. Raj, R. M. Stern, “A Vector Taylor Series Approach for environment-independent Speech Recognition” *In Proc. ICASSP*, 1996.

[3] “Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms” *ETSI ES 202 050 v.1.1.1* (2002-10), October 2002.

[4] L. Deng, A. Acero, M. Plumpe, X. D. Huang, “Large Vocabulary Speech Recognition under Adverse Acoustic Environments” *In Proc. ICSLP*,2000.

[5] A. Varga, H. J. M. Steeneken, “Assessment for Automatic speech Recognition: Ⅱ. NOISEX-92: A database and experiment to study the Effect of Additive Noise on Speech Recognition Systems” *Speech Communication*, 12:247-251,1993.

[6] S. Gao, B. Xu, H. Zhang, B. Zhao, C. Li, T. Huang, “Update of progress of Sinohear: Advanced Mandarin LVCSR System At NLPR” *In Proc. ICSLP*, 2000.

[7] G. H. Ding, B. Xu, “Exploring High-performance Speech Recognition in Noisy Environments Using High-order Taylor Series Expansion” *In Proc. ICSLP,* 2004.

[8] J. Wu, J. Droppo, L. Deng, A. Acero, “A Noise-Robust ASR Front-End Using Wiener Filtering Constructed from MMSE Estimation of Clean Speech and Noise”, *Proc. IEEE ASRU workshop*, 2003.