A MINIMUM-MEAN-SQUARE-ERROR NOISE REDUCTION ALGORITHM ON MEL-FREQUENCY CEPSTRA FOR ROBUST SPEECH RECOGNITION

Dong Yu, Li Deng, Jasha Droppo, Jian Wu, Yifan Gong, and Alex Acero

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052
{dongyu; deng; jdroppo; jianwu; ygong; alexac}@microsoft.com

ABSTRACT

We present a non-linear feature-domain noise reduction algorithm based on the minimum mean square error (MMSE) criterion on Mel-frequency cepstra (MFCC) for environment-robust speech recognition. Distinguishing from the MMSE enhancement in log spectral amplitude proposed by Ephraim and Malah (E&M) [7], the new algorithm presented in this paper develops the suppression rule that applies to power spectral magnitude of the filter-banks' outputs and to MFCC directly, making it demonstrably more effective in noise-robust speech recognition. The noise variance in the new algorithm contains a significant term resulting from instantaneous phase asynchrony between clean speech and mixing noise, missing in the E&M algorithm. Speech recognition experiments on the standard Aurora-3 task demonstrate a reduction of word error rate by 48% against the ICSLP02 baseline, by 26% against the cepstral mean normalization baseline, and by 13% against the conventional E&M log-MMSE noise suppressor. The new algorithm is also much more efficient than E&M noise suppressor since the number of the channels in the Mel-frequency filter bank is much smaller (23 in our case) than the number of bins in the FFT domain (256). The results also show that our algorithm performs slightly better than the ETSI AFE on the well-matched and mid-mismatched settings.

Index Terms — MMSE Estimator, MFCC, Noise Reduction, Robust ASR, Speech Feature Enhancement

1. INTRODUCTION

It is generally held that the desirable signal domain to which noise reduction should be applied differs between human listening and automatic speech recognition (ASR). Conventional wisdom posits that the lower the distortion is between the enhanced speech and the clean speech in the domain closest to the "backend" (human perception or machine recognition), the better the enhancement performance will be. For subjective human listening noise reduction is often applied in the spectral-magnitude domain (e.g., spectral subtraction, Wiener filtering, and Ephraim/Malah spectral amplitude MMSE suppressor [5]). Subjective human listening experiments [8] show that speech enhancement becomes more effective when applied to the log-spectral amplitude domain [7].

In this paper, we apply the same line of thinking to speech feature enhancement for ASR applications, where Mel-Frequency Cepstral Coefficients (MFCCs) have been used pervasively as the closest input representation to ASR back-ends. Specifically, we propose a non-linear feature-domain noise reduction algorithm based on the minimum mean square error (MMSE) criterion on MFCCs for environment-robust speech recognition. We explain that the problem of seeking an MMSE estimator on MFCCs can be reduced to seeking a log-MMSE estimator on the Mel-frequency filter bank's output, which can be solved independently for each filter bank channel. We derive the algorithm by assigning uniformly distributed random phases to the real-valued filter bank's outputs and assuming that the artificially generated complex filter bank's outputs follow zero-mean complex normal distributions. We show two key differences between our new suppression rule and the log-MMSE spectral amplitude estimator proposed by Ephraim and Malah (E&M) [7]. First, our suppression rule is applied directly to the MFCC instead of the spectral amplitude. Second, the noise variance used in our algorithm has been derived to contain an additional term resulting from that the clean speech and the noise are not in phase with each other.

Compared with our previous noise robust techniques such as SPLICE (stereo-based piecewise linear compensation for environments) [2][5], the new algorithm has several advantages. First, it does not require a codebook to be constructed using training data and thus is more robust to unseen environment and easier to be deployed. Second, it introduces no additional look-ahead frame delay. Third, it is applied to the MFCC and hence can be easily plugged into the existing feature extraction pipeline. Speech recognition experiments on the Aurora-3 task demonstrate that our proposed algorithm has huge performance advantages over the E&M log-MMSE noise suppressor. Our algorithm is also much more efficient since the number of the channels in the mel-frequency filter bank is usually much smaller than the number of bins in the FFT domain.

The rest of the paper is organized as follows. In Section 2, we formulate the MMSE estimation problem on MFCC and show how the problem can be reduced to the log-MMSE estimation on the Mel-frequency filter bank's outputs. In Section 3, we first describe the development of our non-linear noise reduction algorithm in detail. Then, we illustrate how the parameters used in the algorithm are estimated with a focus on the computation of a novel variance term induced by the phase difference between clean speech and mixing noise. We describe the evaluation procedure on the Aurora-3 task and report the experimental results in Section 4. In Section 5 we conclude the paper.

2. PROBLEM FORMULATION

Without lack of generality, we denote x as clean speech. We assume that x is corrupted with independent additive noise n to become noisy speech y:

$$y(t) = x(t) + n(t), \quad (1)$$

where t is the time-sample index. Our goal is to find the MMSE estimate $\hat{c}_x(k)$ against each separate dimension k of the clean speech MFCC vector c_x given noisy MFCC vector c_y.

There are three reasons for choosing the dimension-wise instead of the full-vector MMSE criterion. First, each dimension of

the MFCC vector is known to be relatively independent of each other and hence diagonal covariance matrices are usually used in modeling the MFCC space in ASR. Second, the dynamic range of MFCC is different between different dimensions. If the MMSE criterion were applied to the full MFCC vector, each dimension would need to be weighted differently so that the error would not be dominated by one or two dimensions. Choosing the appropriate weights not only is difficult but also introduces unnecessary computational overhead. Third, the dimension-wise MMSE criterion decouples different dimensions, making the algorithm easier to develop and to implement.

The solution to the MMSE problem for each element of the dimension-wise MFCC vectors is the conditional expectation of

$$\begin{aligned}\hat{c}_x(k) &= E\{c_x(k)|m_y\} = E\left\{\sum_b a_{k,b}\log m_x(b)\,\middle|\,m_y\right\}\\ &= \sum_b a_{k,b}E\{\log m_x(b)\,|m_y\},\end{aligned} \tag{2}$$

where $a_{k,b}$ are the discrete cosine transform coefficients, m_y and m_x are the Mel-frequency filter bank's output in power for the noisy and clean speech respectively, b is the filter bank channel id.

The additive assumption (1) for speech-noise mixing in time domain gives the same relationship in frequency domain:

$$Y(f) = X(f) + N(f). \tag{3}$$

where $Y(f)$, $X(f)$, and $N(f)$ are discrete Fourier transformation (DFT) of noisy speech waveform y, clean speech waveform x, and noise waveform n. We further assume that $m_x(b)$ be independent of $m_y(b')\ \forall b' \neq b$ given $m_y(b)$ and thus it can be reconstructed solely from $m_y(b)$. Then, (2) can be further simplified to

$$\hat{c}_x(k) \cong \sum_b a_{k,b}E\{\log m_x(b)\,|m_y(b)\}. \tag{4}$$

The problem is thus reduced to finding the log-MMSE estimator of the Mel-frequency filter bank's output

$$\widehat{m}_x(b) = exp\big(E\{\log m_x(b)\,|m_y(b)\}\big). \tag{5}$$

There can be many different solutions to (5) based on different assumptions on the noise and noisy speech models. In the following section, we derive one of the solutions

3. THE MFCC-MMSE ESTIMATOR

At the first glance of (5), it appears that the E&M log-MMSE magnitude spectral suppressor could be directly applied to the filter bank output by converting the power spectral to the magnitude spectral first and then converting it back once the suppression is done. Our experiments on Aurora-3 showed that such a naive approach gave poor results (see detail in Section 4). This motivates a more principled approach to be described in this section.

3.1 The Suppressor Rule

To develop a rigorous approach, we construct three complex variables $M_x(b)$, $M_n(b)$ and $M_y(b)$ such that

$$\begin{aligned}|M_x(b)| = m_x(b) &= \sum_f w_b(f)|X(f)|^2,\\ |M_n(b)| = m_n(b) &= \sum_f w_b(f)|N(f)|^2,\\ |M_y(b)| = m_y(b) &= \sum_f w_b(f)|Y(f)|^2.\end{aligned} \tag{6}$$

where $w_b(f)$is the fixed b-th Mel-frequency filter's weight for the frequency bin f. Many $M_x(b)$, $M_n(b)$ and $M_y(b)$ would satisfy (6), among which we choose the ones with uniformly distributed random phases $\theta_x(b)$, $\theta_n(b)$, and $\theta_y(b)$ (which can be considered as the weighted sum of the phases over all the DFT bins). Selecting such phases enables us to make the assumption that complex variables $M_x(b)$ and $M_y(b) - M_x(b)$ both follow the zero-mean complex normal distributions.

Since $M_y(b)$ contains all information there is in $m_y(b)$, (5) can be rewritten as

$$\widehat{m}_x(b) = exp\big(E\{\log m_x(b)\,|M_y(b)\}\big). \tag{7}$$

Following a similar approach developed in [7], we can find the solution to (7) as

$$\begin{aligned}\widehat{m}_x(b) &= exp\big(E\{\log m_x(b)\,|m_y(b)\}\big)\\ &= G\big(\xi(b), v(b)\big)m_y(b),\end{aligned} \tag{8}$$

where the gain is

$$G\big(\xi(b), v(b)\big) = \frac{\xi(b)}{1+\xi(b)}exp\left\{\frac{1}{2}\int_{v(b)}^{\infty}\frac{e^{-t}}{t}dt\right\} \tag{9}$$

In (9), the quantity

$$v(b) = \frac{\xi(b)}{1+\xi(b)}\gamma(b) \tag{10}$$

is defined by the adjusted *a- priori* SNR for each filter bank:

$$\xi(b) \stackrel{\text{def}}{=} \frac{\sigma_x^2(b)}{\sigma_d^2(b)}, \tag{11}$$

and by the adjusted *a-posteriori* SNR:

$$\gamma(b) \stackrel{\text{def}}{=} \frac{m_y^2(b)}{\sigma_d^2(b)}. \tag{12}$$

Then, the MMSE estimator for the MFCC becomes

$$\begin{aligned}\hat{c}_x(k) &\cong \sum_b a_{k,b}E\{\log m_x(b)\,|m_y(b)\}\\ &= \sum_b a_{k,b}\log\Big(G\big(\xi(b), v(b)\big)m_y(b)\Big).\end{aligned} \tag{13}$$

We would like to point out two essential differences between noise suppression rule (9) and that proposed in [7]. First, our suppression rule is applied to MFCC (after applying to the *power* spectral domain of the filter bank's output) instead of to the magnitude spectral domain as in [7]. Second, the *a priori* and *a posteriori* SNRs defined in (11) and (12) are different from those defined in [7]. Because of the use of the filter bank, they need to be adjusted to include not only the noise (in the power spectral and not the spectral magnitude domain) variance $\sigma_n^2(b) = E\{m_n^2(b)\}$, but also the additional variance $\sigma_\varphi^2(b)$ resulting from instantaneous phase differences between the clean speech and the mixing noise. That is,

$$\sigma_d^2(b) \cong \sigma_n^2(b) + \sigma_\varphi^2(b), \tag{14}$$

3.2 Estimation of $\sigma_n^2(b)$ and $\sigma_x^2(b)$

In our current implementation, the noise variance $\sigma_n^2(b)$ is estimated using a minimum controlled recursive moving-average noise tracker similar to the one described in [1]. $\sigma_x^2(b)$ is estimated using the same decision-directed approach as that described in [6].

3.3 Estimation of $\sigma_\varphi^2(b)$

Inclusion and estimation of $\sigma_\varphi^2(b)$ are one major novelty of this

work, which considerably contributes to the performance gain shown in Section 4. Rigorous estimation is difficult and we developed the following approximate method (details omitted):

$$\sigma_{\varphi}^2(b) = E\left\{\left(\sum_f 2|X(f)||N(f)|\cos\varphi(f)\,w_b(f)\right)^2\right\}$$
$$= 2\sum_f w_b^2(f)E\{|X(f)|\}^2\,E\{|N(f)|\}^2 \qquad (15)$$
$$\cong 2\frac{\sum_f w_b^2(f)}{\left(\sum_f w_b(f)\right)^2}\sqrt{\frac{\sigma_x^2(b)}{\sigma_n^2(b)}}\sigma_n^2(b)$$

in our current implementation. Note that (15) depends on $\sigma_n^2(b)$ and $\sigma_x^2(b)$. Therefore, $\sigma_{\varphi}^2(b)$ needs to be estimated after estimating $\sigma_n^2(b)$ and $\sigma_x^2(b)$.

.

4. PERFORMANCE EVALUATION

We have conducted extensive speech recognition experiments on the standard Aurora-3 task [5] to evaluate the performance of the non-linear MMSE noise reduction algorithm on MFCC described so far in this paper.

4.1 Experimental Setup

The Aurora-3 task consists of noisy digit recognition sub-tasks under realistic automobile environments. In the Aurora-3 corpus, each utterance is labeled as coming from either a high, low, or quiet noise environment, and as being recoded using a close-talk microphone or a hands-free, far-field microphone.

Based on the languages, the task can be classified into four separate digit recognition sub-tasks. For each language, three experimental settings are defined for the evaluation:

Well-matched – Both the training and the testing set contain all combinations of noise environments and microphones.

Mid-mismatch – The training set contains quiet and low noise data recorded using the far-field microphone, and the testing set contains the high noisy data recorded using the far-field microphone. The mismatch is mainly caused by the noise.

High-mismatch – The training set contains close-talk data from all noise classes, and the testing set contains high noise and low noise far-field data for testing. The mismatch is caused mainly by channel distortion.

All speech recognition results reported in this section use the HMMs trained in the manner prescribed by the scripts included with the Aurora-3 task. The HMMs used consist of 16-state whole-word models for each digit in addition to the "sil" and "sp" models. The 39-dimenion features used in our experiments contain the 13-dimention static MFCC features and their delta and delta-delta features. The parameters (such as smoothing factors and the size of the minimum tracking windows) used for noise tracking are similar to those described in [1].

4.2 Experimental Results

The purpose of our experiments is to examine to what extent our new algorithm is effective for its designed purpose: noise robustness under the additive noise environment. With this goal in mind, we have conducted a series of experiments to compare our algorithm with other noise robust algorithms such as the conventional E&M log-MMSE magnitude spectral suppressor (which operates on the much more expensive DFT domain) and the ETSI's advanced front end (AFE).

In all the results reported in this section, the ICSLP02 baseline refers to the baseline system using the standard WI007 front-end (Figure 1). The AGN/CMN baseline is the system with the WI007 frontend and a standard active gain normalization and cepstral mean normalization algorithm (Figure 2). In both the MFCC-MMSE and the E&M log-MMSE systems we applied the noise suppression algorithms on top of the AGN/CMN baseline system (Figure 3 & 4).

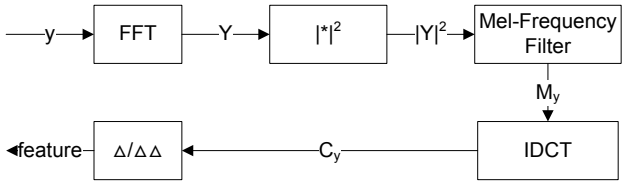

Fig. 1: Feature extraction pipeline for ICSLP02 baseline system.

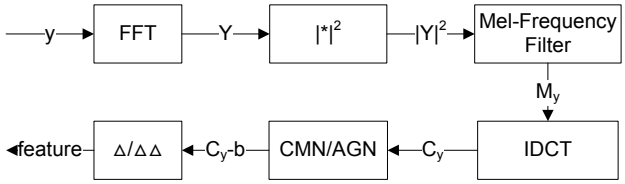

Fig. 2: Feature extraction pipeline for AGN/CMN baseline system.

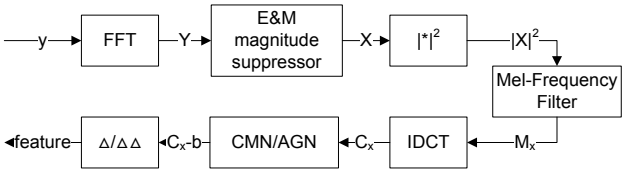

Fig. 3: Feature extraction pipeline for the E&M log-MMSE system, where the suppressor is applied to the DFT bins.

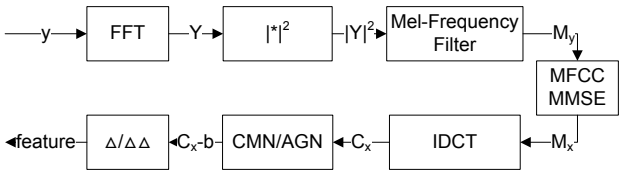

Fig. 4: Feature extraction pipeline for the MFCC-MMSE system.

Tables 1 and 2 summarize the average absolute recognition word error rate (WER) results and the relative improvements for the above four systems, respectively, plus the naïve implementation of the E&M log-MMSE algorithm for the Filter-Bank output magnitude (labeled as "FB Output Magnitude" in the final column of Table 1). We observe that the new MFCC-MMSE approach has achieved over 48% WER reduction relative to the ICSLP02 baseline system, over 25% WER reduction to the AGN/CMN baseline system, and over 13% WER reduction to the conventional E&M log-MMSE algorithm while saving considerable computational cost (23 vs. 256 frequency channels for estimation). We also observe that directly applying the E&M log-MMSE noise suppressor to the magnitude spectrum of the Mel-frequency filter bank output gives only slight gain over the AGN/CMN baseline. Detailed results on each sub-tasks of our MFCC-MMSE noise suppressor are reported in Table 3.

Table 1: Summary of absolute WER in the Aurora-3 task under five different experimental settings

Summary of Aurora 3 Absolute Word Error Rate				
	Well	**Mid**	**High**	**Average**
ICSLP02 Baseline	8.96%	21.96%	48.85%	**23.48%**
AGN/CMN	6.87%	16.52%	31.11%	**16.31%**
E&M log-MMSE	5.57%	12.79%	29.23%	**14.01%**
MFCC-MMSE	5.08%	12.26%	23.26%	**12.13%**
FB Output Magnitude	6.87%	15.21%	31.29%	**15.89%**

Table 2: Summary of relative WER reduction in the Aurora-3 task

Summary of Aurora-3: Relative Performance Improvement			
Relative to →	**ICSLP02 Baseline**	**AGN/CMN**	**E&M log-MMSE**
AGN/CMN	30.55%	--	--
E&M log-MMSE	40.33%	14.08%	--
MFCC-MMSE	48.33%	25.59%	13.41%

Table 3: Detailed Aurora-3 absolute WER results under the MFCC-MMSE experimental setting.

Aurora-3 Word Error Rate Using MFCC-MMSE					
	Finnish	**Spanish**	**German**	**Danish**	**Average**
Well (x40%)	3.54%	5.90%	5.20%	5.66%	**5.08%**
Mid (x35%)	15.12%	5.39%	10.67%	17.84%	**12.26%**
High (x25%)	17.99%	34.77%	10.78%	29.49%	**23.26%**
Overall	**11.21%**	**12.94%**	**8.51%**	**15.88%**	**12.13%**

To further understand the effectiveness of our new algorithm, we have evaluated its performance against ETSI AFE with the results reported in Table 4. Analysis of these results shows that our new MFCC-MMSE approach has comparable performance on the well matched and mid-mismatched settings where noise distortion is the dominant cause of the mismatch. In fact, when counting errors under these two conditions only, the MFCC-MMSE system achieves an average of 8.43% WER, slightly low than that with the ETSI AFE (8.67% WER on average). Our approach, however, performs considerably worse than the ETSI AFE system under the high-mismatched setting. This is attributed mainly to the fact that the distortion in the high-mismatched setting is largely caused by channel distortion, which has not been handled in the design of our system but was carefully handled by the ETSI AFE.

Table 4: Comparison between the MFCC-MMSE system and the AFE on Aurora-3.

Aurora-3 Wrord Error Rate AFE on Aurora 3			
	Well	**Mid**	**High**
ETSI AFE	4.70%	13.21%	12.75%
MFCC-MMSE	5.08%	12.26%	23.26%

5. SUMMARY AND CONCLUSIONS

In this paper, we present a new, highly efficient non-linear noise reduction algorithm using the MMSE criterion in the MFCC domain for noise-robust ASR. We describe the algorithm and the parameter estimation methods, show the differences between our algorithm and the conventional E&M log-MMSE noise suppressor, and demonstrate its effectiveness in the standard Aurora-3 task.

This new, model-free approach to MFCC feature enhancement and for noise-robust ASR has several key features. First, it does not require a codebook (unlike SPLICE) be constructed using training data, hence it is highly robust to general unseen acoustic environments and it is easy to deploy in our practical ASR system. Second, it is computationally efficient compared with the conventional E&M log-MMSE noise suppressor since the number of the frequency channels in the Mel-frequency filter bank is much smaller than the number of bins in the DFT domain. Third, it introduces no look-ahead frame delay. Fourth, it is designed to apply to filter bank's outputs and hence can be easily plugged into the feature extraction pipeline of many commonly used ASR systems including our own. The proposed approach as developed so far, however, only deals with additive noises and has not been developed to handle channel distortions. Our current work involves expanding on this capability. We are also investigating the combination of the current algorithm, which does not rely on any data, with the data-driven approach (as exploited in SPLICE) to take advantage of the mutual strengths.

7. ACKNOWLEDGEMENTS

The authors of this paper would like to thank Dr. Jay Stokes at Microsoft Research for valuable discussions.

REFERENCES

[1] I. Cohen and B. Berdugo. "Noise estimation by minima controlled recursive averaging for robust speech enhancement," IEEE Signal Proc. Letters, Vol. 9, 2002, pp. 12-15.

[2] L. Deng, J. Droppo, and A. Acero. "Recursive estimation of non-stationary noise using iterative stochastic approximation for robust speech recognition," IEEE Trans. Speech & Audio Processing, Vol.11, 2003, pp. 568-580.

[3] L. Deng, J. Droppo, and A. Acero. "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," IEEE Trans. Speech & Audio Processing, Vol. 12, 2004, pp. 218-233.

[4] L. Deng, J. Droppo, and A. Acero. "Enhancement of log-spectra of speech using a phase-sensitive model of the acoustic environment," IEEE Trans. Speech & Audio Processing, Vol. 12, 2004, pp. 133-143.

[5] J. Droppo, L. Deng, and A. Acero. Evaluation of SPLICE on the Aurora 2 and 3 Tasks, in Proc. Int. Conf. on Spoken Language Processing. Denver, Colorado, Sep, 2002.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoustics, Speech and Signal Proc, Vol. ASSP-32, pp. 1109-1121, 1984.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoustics, Speech and Signal Proc, vol. ASSP-33, pp. 443–445, 1985.

[8] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," Proc. ICASSP, Vol.I, 2006, pp. 153-156.

[9] D. Macho, L. Mauuary, B. Noé, Y-M Cheng, D. Ealey, D. Jouve, H. Kelleher, D. Pearce, F. Saadoun, "Evaluation of a noise-robust DSR front-end on aurora databases," Proc. Interspeech 2002, pp. 17-20.