A SPEECH ENHANCEMENT APPROACH  
E-CMN/CSS FOR SPEECH RECOGNITION IN  
CAR ENVIRONMENTS

M. Shozakai, S. Nakamura and K. Shikano  
Graduate School of Information Science  
Nara Institute of Science and Technology  
Ikoma, Nara, 630-01 Japan

Abstract - This paper proposes the robust speech enhancement approach E-CMN(Exact Cepstrum Mean Normalization)/CSS (Continuous Spectral Subtraction). The E-CMN, which we proposed for compensation of multiplicative distortions[6], calculates two cepstrum mean vectors, one for speech for each speaker and the other for non-speech for each environment. The CSS subtracts continuously average spectra at every frame. A high robustness of the proposed method is clarified by comparative evaluation with alternative methods for speech recognition tasks in car environments.

1 INTRODUCTION

A robust speech interface for facilities in a car cabin and for mobile computing devices is highly needed. A drastic drop in performance that occurs in adverse environments is widely acknowledged to be due to multiplicative distortions and additive noises. This paper investigates comparison of three kinds of speech enhancement approaches to additive noises such as SS(Spectral Subtraction), CSS(Continuous Spectral Subtraction) and MMSE(Minimum Mean Square Error estimation). A higher robustness of CSS to additive noise is clarified. Then, E-CMN(Exact Cepstrum Mean Normalization) for compensation of multiplicative distortions is described, which calculates two cepstrum mean vectors, one for speech for each speaker and the other for non-speech for each environment. The new robust speech enhancement approach E-CMN/CSS is proposed. Furthermore, performances of E-CMN/CSS and the model adaptation approach E-CMN/PMC(Parallel Model Combination) which we have proposed are comparatively evaluated for a speech recognition task in car environments.
An advantage of the E-CMN/CSS over the E-CMN/PMC is finally discussed with respect to SNR (Signal-to-Noise Ratio).

2 MODELING MULTIPLICATIVE DISTORTION AND ADDITIVE NOISE

The long-term average of short-term spectra $S(\omega; t)$ of frequency $\omega$ at time $t$ in a speech frame is called speaker personality and is defined as

$$H_{\text{Person}}(\omega) = \frac{1}{T} \cdot \sum_{i=1}^{T} S(\omega; t)$$

where $T$ is a sufficiently large natural number. The speaker personality may be considered to represent frequency characteristics which depend on the speaker’s vocal tract and vocal cords. The normalized speech spectra is defined as

$$S^*(\omega; t) = S(\omega; t) / H_{\text{Person}}(\omega)$$

The short-time spectra $S(\omega; t)$ is interpreted as generated outputs when the normalized speech spectra $S^*(\omega; t)$ passes through a time-invariant filter of gain $H_{\text{Person}}(\omega)$ which is a multiplicative distortion to $S^*(\omega; t)$. We may find three kinds of multiplicative distortion for $S^*(\omega; t)$ in addition to the $H_{\text{Person}}(\omega)$ in reality[1].

1. **Speaking style** $H_{\text{Style(N)}}(\omega)$: frequency characteristics peculiar to speaking styles (speed, loudness, Lombard effect etc.) which are affected by an additive noise,

2. **Acoustical transmission characteristics** $H_{\text{Trans}}(\omega)$: spatial frequency characteristics from mouth to microphone, and

3. **Microphone characteristics** $H_{\text{Mic}}(\omega)$: frequency characteristics of microphone.

If we assume that speech and noise are additive in the linear spectrum domain, the observed spectra $O(\omega; t)$ is modeled as

$$O(\omega; t) = H^*(\omega) \cdot S^*(\omega; t) + \tilde{N}(\omega; t)$$

$$H^*(\omega) = H_{\text{Mic}}(\omega) \cdot H_{\text{Trans}}(\omega) \cdot H_{\text{Style(N)}}(\omega) \cdot H_{\text{Person}}(\omega)$$

$$\tilde{N}(\omega; t) = H_{\text{Mic}}(\omega) \cdot N(\omega; t)$$

where $N(\omega; t)$ is an environmental additive noise.
3 SPEECH ENHANCEMENT TO AN ADDITIVE NOISE

Three speech enhancement techniques are discussed.

(1) SS (Spectral Subtraction) [2]

An estimated spectra of additive noise $\hat{N}(\omega; t)$ is calculated in frames which are judged as non-speech by VAD (Voice Activity Detection). An estimated spectra $\hat{S}(\omega; t)$ of $S(\omega; t)$ is calculated as follows:

$$
\hat{N}(\omega; t) = \begin{cases} 
\hat{N}(\omega; t - 1) & \text{in speech frame} \\
\gamma \cdot \hat{N}(\omega; t - 1) + (1 - \gamma) \cdot O(\omega; t) & \text{otherwise} 
\end{cases}
$$

(6)

$$
\hat{S}(\omega; t) = \begin{cases} 
O(\omega; t) - \alpha \cdot \hat{N}(\omega; t) & \text{if } O(\omega; t) - \alpha \cdot \hat{N}(\omega; t) > \beta \cdot O(\omega; t) \\
\beta \cdot O(\omega; t) & \text{otherwise} 
\end{cases}
$$

(7)

where $\alpha$, $\beta$, $\gamma$ are called over-estimation factor, flooring factor and smoothing factor respectively. An estimation of $\hat{N}(\omega; t)$ depends on an accuracy of VAD. An existance of so-called “musical noise” is inevitable due to inaccuracy of $\hat{N}(\omega; t)$ estimation.

(2) CSS (Continuous Spectral Subtraction) [3]

An estimated value of additive noise $\hat{N}(\omega; t)$ is continuously updated in every frames without VAD as follows:

$$
\hat{N}(\omega; t) = \gamma \cdot \hat{N}(\omega; t - 1) + (1 - \gamma) \cdot O(\omega; t).
$$

(8)

An estimated spectra $\hat{S}(\omega; t)$ is calculated in the same way as SS using equ.(7). An estimation of $\hat{N}(\omega; t)$ does not depend on an accuracy of VAD. But, because speech spectra affect $\hat{N}(\omega; t)$ estimation, there is a problem that week spectral components following strong spectral components are masked out. It leads to speech spectra distortion. The paper[3] proposes HMM composition method with considering residual spectra after CSS as additive noises. However, there is no description of CSS’s robustness to additive noise in the paper.

(3) MMSE (Minimum Mean Square Error estimation) [4]

Under the assumption that occurances of speech and noise follow mutually independent Gaussian distribution, the estimated spectra $\hat{S}(\omega; t)$ are obtained by Bayesian estimator

$$
\hat{S}(\omega; t) = G_{\text{MMSE}}(\bar{\xi}_\omega, \bar{r}_\omega, q_\omega) \cdot O(\omega; t).
$$

(9)
The gain $G_{\text{MMSE}}(\xi_\omega, \zeta_\omega, q_\omega)$ is estimated with minimum mean square error principle of (log-)amplitude of short-term spectra where $\xi_\omega$, $\zeta_\omega$, $q_\omega$ are estimated a priori SNR, a posteriori SNR and assumed probability of non-existence of speech at frequency $\omega$. The MMSE, which has been recently applied to canceling noises for a hands-free cellular phone[5], seems promising for speech recognition in adverse environments.

Spectrograms for word “ai” uttered by a Japanese female are shown in Fig.1, where clean speech and noisy speech(10dB SNR with car noise) are processed with no processing(NO), SS, CSS and MMSE.

We define a variability measure of MFCC(Mel-Frequency Cepstrum Coefficient)s between SNR1 and SNR2 as

$$D_{\text{Variability}}^{\text{SNR1,SNR2}} = \frac{N}{N-1} \sum_{n=1}^{N} \sum_{i=1}^{l} \left( c_i^{\text{SNR1}}(n) - c_i^{\text{SNR2}}(n) \right)^2.$$  \hspace{1cm} (10)

where $c_i^{\text{SNR}}(n)$ denotes i-th MFCC in frame n at SNR. $N$ and $l$ are a number of frames and an order of MFCC vector respectively. Fig.2 shows variability measures $D_{\text{Variability}}^{\text{clean}}$, $D_{\text{Variability}}^{\text{10dB}}$, $D_{\text{Variability}}^{\text{20dB}}$, for SS, CSS and MMSE, calculated from 10 order MFCC vector in speech frames of 65 words of 4 speakers. The symbol $\infty$ means clean data. 20dB and 10dB mean that car noise is added to the clean data.
at SNR 20dB and 10dB respectively. Fig.2 shows that the variability measure for CSS is lower than those for SS and MMSE. It suggests that CSS gives more similar MFCC vectors in a wide range of SNR. We can observe this robust property of CSS by comparing spectrograms of Fig.1 (c.1) and (c.2). On the other hand, speech spectral distortion by CSS is noticeable by comparing Fig.1 (a.1) with (c.1).

![Variability measure of MFCC.](image)

We compare the performances of SS, CSS and MMSE with speaker-independent Japanese 65 words recognition task. A whole word HMM, which has two states per phoneme, is trained from training data for each word. Each state has one Gaussian distribution with diagonal covariance. Acoustic analysis is done with 8kHz sampling, 32ms frame length, 20ms frame shift. 10 MFCCs are used as acoustic parameters. Car noise is added to both training data of 36 speakers(18 males/18 females) and test data of 4 speakers(2 males/2 females) with the same SNR. The training data and the test data are noise-cancelled by the same speech enhancement technique. The average word recognition rates are shown in Table 1.

<table>
<thead>
<tr>
<th>SNR</th>
<th>20dB</th>
<th>10dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>91.9%</td>
<td>84.6%</td>
</tr>
<tr>
<td>SS</td>
<td>96.5%</td>
<td>89.6%</td>
</tr>
<tr>
<td>CSS</td>
<td>95.4%</td>
<td>93.8%</td>
</tr>
<tr>
<td>MMSE</td>
<td>96.9%</td>
<td>91.9%</td>
</tr>
</tbody>
</table>

Table 1: Average word recognition rates.

We can summarize two properties of the CSS here.

(property-1) CSS gives higher performance than SS and MMSE at lower SNR due to low value of variability measure.

(property-2) CSS has slightly worse performance than SS and MMSE at higher SNR due to inevitable spectral distortion.
We proposed the E-CMN algorithm for normalizing speech by compensating various multiplicative distortions as follows\cite{6}:

**Estimation Step**: Two cepstrum mean vectors are calculated. One, obtained from speech frames of sufficiently-long utterances, is speaker-dependent. The other, obtained from non-speech frames, is environment-dependent.

**Normalization Step**: The speaker-dependent cepstrum mean vector for speech is subtracted from the input cepstrum vector in speech frames. The environment-dependent cepstrum mean vector for non-speech is subtracted from the input cepstrum vector in non-speech frames.

We propose two variations of E-CMN/CSS method by combining E-CMN for multiplicative distortions with CSS for additive noises as shown in Fig. 3.
E-CMN/CSS(clean) : Speech models like HMM are trained from clean speech processed by CSS and E-CMN.

E-CMN/CSS(noisy) : Speech models like HMM are trained from noisy speech (clean speech with noise added) processed by CSS and E-CMN.

We compare performance of the E-CMN/CSS method with that of the E-CMN/PMC method we have proposed[6]. A brief description of the E-CMN/PMC is as follows;

(1) The clean HMMs are trained from clean speech which are normalized by the E-CMN(Estimation & Normalization Steps).
(2) The clean HMMs are adapted, by HMM composition techniques[7, 8], with noise HMM and the multiplicative distortion $H^*(\omega)$ estimated by the E-CMN(Estimation Step).

The recognition task is speaker-independent 520 Japanese words with 54 context-independent tied-mixture HMMs which are trained from speech database for 40 speakers. The acoustic analysis uses 8kHz sampling, 32ms frame length and 20ms frame shift. The parameters are 10 MFCCs, 10 delta MFCCs and a delta energy. The number of shared Gaussian distributions are 256, 256 and 64 respectively. The impulse response measured from the mouth of dummy head located at driver's seat to the omni-directional microphone mounted on driver's sun-visor is convoluted with evaluation data (2 males and 2 females) as $H_{\text{Trans}}(\omega)$. Fig.4 shows recognition performances for E-CMN/CSS(noisy), E-CMN/CSS(clean), E-CMN/PMC and no adaptation. The SNR of training data for E-CMN/CSS(noisy) is 10dB in this simulation.

![Recognition Rate (%) vs SNR (dB)](image)

Fig. 4: Comparison of E-CMN/CSS and E-CMN/PMC.
The E-CMN/CSS (noisy) outperforms the E-CMN/PMC in a wide range of SNR except 29dB. It seems that the (property-1) of CSS prevents a significant degradation of recognition performance at lower SNRs.

(2) The performance of E-CMN/CSS (noisy) at 29dB SNR is slightly worse than that of E-CMN/PMC because of the (property-2) of CSS.

5 CONCLUSION

The new speech enhancement approach E-CMN/CSS was proposed and compared with the model adaptation approach E-CMN/PMC for a speech recognition task in car environments. The advantage of E-CMN/CSS over E-CMN/PMC at lower SNR was shown.

6 REFERENCES