ROBUST SPEECH RECOGNITION IN CAR ENVIRONMENTS

M. Shozakai, S. Nakamura, K. Shikano

Nara Institute of Science and Technology
Takayama 8916-5
Ikoma, Nara 630-01, JAPAN

ABSTRACT

A user-friendly speech interface in a car cabin is highly needed for safety reasons. This paper will describe a robust speech recognition method that can cope with additive noises and multiplicative distortions. A known additive noise, a source signal of which is available, might be canceled by NLMS-VAD(Normalized Least Mean Squares with frame-wise Voice Activity Detection). On the other hand, an unknown additive noise, a source signal of which is not available, is suppressed with CSS(Continuous Spectral Subtraction). Furthermore, various multiplicative distortions are simultaneously compensated with E-CMN(Exact Cepstrum Mean Normalization) which is speaker-dependent/environment-dependent CMN for speech/non-speech. Evaluation results of the proposed method for car cabin environments are finally described.

1. INTRODUCTION

The concept of ITS(Intelligent Transport Systems) was proposed and is promoted in many countries. Drivers, cars, roads and information network systems are connected with wireless communication technologies. There must be user-friendly human machine interface for drivers to have an easy access to various information of traffic, road construction, dynamic route guidance for navigation and so forth. The VODIS(Voice Operated Driver's Information Systems) project[2] was launched in Europe to realize a robust speech interface for command and control applications for car facilities such as a car navigation, a car stereo and a cellular phone.

There exist both additive noises and multiplicative distortions in car cabin environments. Table 1 lists the additive noises of four types. Here, the known additive noise with its source signal available might be cancelled with an adaptive filter approach like NLMS(Normalized Least Mean Squares)[3]. On the other hand, the stationary unknown additive noises can be effectively suppressed with noise cancellation methods such as SS(Spectral Subtraction)[4]. Furthermore, the CMN(Cepstrum Mean Normalization)[5] approach was proposed to compensate multiplicative distortion of microphone characteristics.

This paper is organized as follows. In Section 2 we separately discuss three algorithms for robust speech recognition, 1)NLMS-VAD(NLMS with frame-wise Voice Activity Detection) for canceling a known additive noise, 2)CSS(Continuous Spectral Subtraction)[6] for suppressing a stationary unknown additive noise, and 3)E-CMN(Exact CMN)[1] for compensating a multiplicative distortion. In Section 3 the combined approach

Table 1. Additive noises in a car cabin.

	known	unknown
stationary	engine etc.	road, wind, air-conditioner etc.
non-stationary	car stereo speaker out, navigation guide, traffic information guide etc.	bump, wiper, winker, conversation, noise when passing a car running to opposite direction etc.

NLMS-VAD/CSS/E-CMN is proposed to realize a robust speech recognizer in car cabin environments and is evaluated with a speaker independent large vocabulary word recognition task. Finally we summarize our proposal and outline our future work.

2. ROBUST SPEECH RECOGNITION "NLMS-VAD/CSS/E-CMN"

2.1 Modeling Additive Noise and Multiplicative Distortion in a Car Cabin

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

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

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

$$S^*(\omega;t) = S(\omega;t) / H_{Person}(\omega). \quad (2)$$

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 $H_{Person}(\omega)$ 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 $H_{Person}(\omega)$ in reality[9] as follows;

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

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

(3)*microphone characteristics* $H_{Mic}(\omega)$: 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

$$O(\omega;t) = H^*(\omega)\cdot S^*(\omega) + \tilde{N}(\omega;t) + \tilde{E}(\omega;t), \quad (3)$$

$$H^*(\omega) = H_{Mic}(\omega)\cdot H_{Trans}(\omega)\cdot H_{Style(N)}(\omega)\cdot H_{Person}(\omega), \quad (4)$$

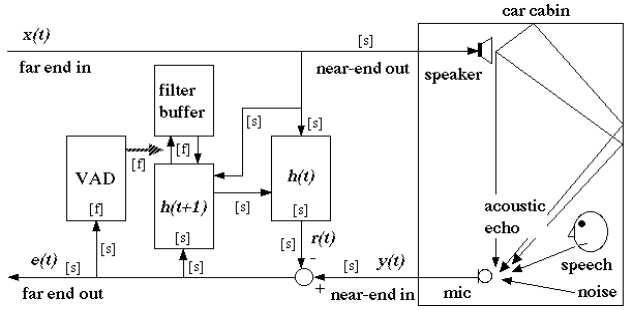
$$\tilde{N}(\omega;t) = H_{Mic}(\omega)\cdot N(\omega;t), \quad (5)$$

$$\tilde{E}(\omega;t) = H_{Mic}(\omega)\cdot E(\omega;t), \quad (6)$$

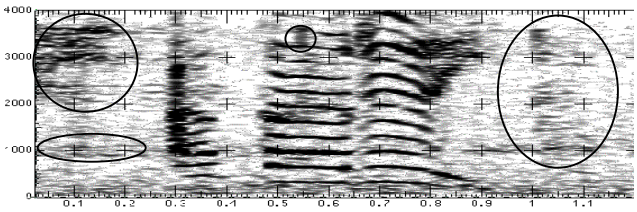
where $N(\omega;t)$, $E(\omega;t)$ are an unknown additive noise and a known one respectively.

2.2 Cancellation of Known Additive Noise

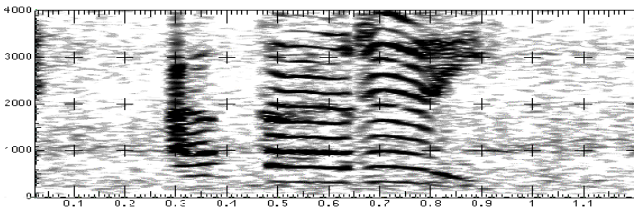
The known additive noises in Table 1 might be cancelled by the adaptive filter technique. The most typical algorithm NLMS is used in this paper. The important issue is how to control start and stop of FIR coefficients update by NLMS. When an speech signal exists in the adaptive filter input, the update of FIR coefficient must be stopped. On the other hand, the update should be continued unless there is the speech signal. The detection of the speech signal must be accurate. However, there are various unknown additive noises which make the detection difficult. A simple end-point detection using signal power does not work well because it is quite difficult to distinguish the speech signal and the stationary additive noise. A more robust speech detection method is highly required. In this paper, we use the frame-wise VAD algorithm[7] which is standardized for GSM cellular phones. This algorithm is capable of adapting to varying background additive noise level. However, it was observed that the detection performance at low SNR below 10dB is poor. So, we modified the algorithm so that it is capable of detecting speech interval at lower SNR. The adaptive filter enhanced with the frame-wise VAD is called NLMS-VAD(NLMS with frame-wise VAD). The block diagram of NLMS-VAD is shown in Fig.1(a), where [s], [f] indicate sample-wise, frame-wise operations respectively. The cancelled signal is fed into the VAD in which speech/non-speech detection is carried in frame-wise fashion. If non-speech is detected, the estimated FIR coefficients $h(t+1)$ is saved in filter buffer. Otherwise, $h(t+1)$ is replaced by the FIR coefficients saved in the filter buffer in order to prevent them from being deteriorated due to a delay of frame-wise VAD operation. Fig.1(b) shows a spectrogram of microphone input of Japanese word “akarui”(0.28-0.9 sec.) uttered in a running car cabin with a car stereo playing a music. Fig.1(c) shows spectrogram of the output of NLMS-VAD, which is obtained from a microphone input. It is observed that the acoustic reverberations which are indicated with black circles in Fig.1(b) are effectively suppressed by NLMS-VAD. The maximum and average of ERLE(Echo Return Loss Enhancement) for this test data are 9.3dB and 4.5dB respectively. The result of VAD is shown in Fig.1(d) which suggests that the VAD works reliably.



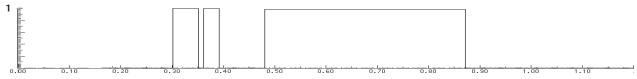
(a)block diagram of NLMS-VAD



(b)Input to NLMS-VAD



(c)Output from NLMS-VAD



(d)VAD decision

Fig.1 NLMS-VAD

2.3 Suppression of Unknown Additive Noise

The most typical approach of canceling stationary unknown additive noise is SS. On the other hand, MMSE has been studied as a promising noise cancellation technique for a hands-free cellular phone[10]. Recently, a advantage of CSS over SS and MMSE was reported[11]. The CSS is formulated as follows;

$$\hat{N}(\omega;t) = \gamma\cdot\hat{N}(\omega;t-1) + (1-\gamma)\cdot O(\omega;t). \quad (7)$$

$$\hat{S}(\omega;t) = \begin{cases} O(\omega;t) - \alpha\cdot\hat{N}(\omega;t) \\ \quad if\ O(\omega;t) - \alpha\cdot\hat{N}(\omega;t) > \beta\cdot O(\omega;t) \\ \beta\cdot O(\omega;t) \quad otherwise \end{cases} \quad (8)$$

An estimated value of additive noise $\tilde{N}(\omega;t)$ is continuously updated in every frames regardless of VAD result in CSS on the contrary to SS. Then, an estimated spectra $\hat{S}(\omega;t)$ is calculated. Because speech spectra affect $\tilde{N}(\omega;t)$ estimation, there is a problem that week spectral components following strong spectral components are masked out. It leads to distortion of speech spectra.

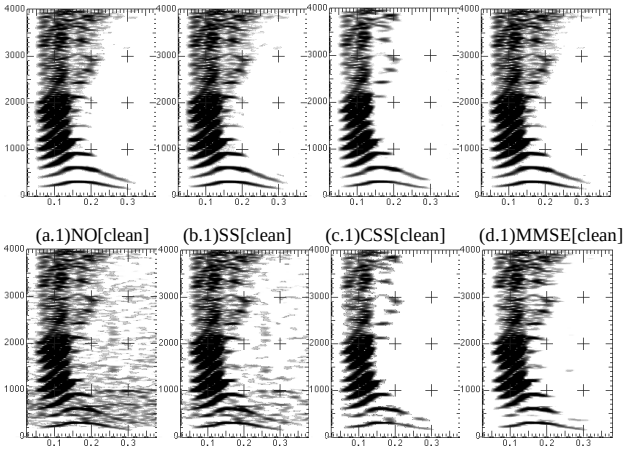


Fig.2: Effect of speech enhancement techniques(horizontal axis : sec., vertical axis: frequency).

Spectrograms for word "ai" uttered by a Japanese female are shown in Fig.2, 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

$$D_{Variability}^{SNR1,SNR2} = \frac{1}{N} \cdot \sum_{n=1}^{N} \sum_{i=1}^{I} \left(c_i^{SNR1}(n) - c_i^{SNR2}(n) \right)^2 , \quad (9)$$

where $c_i^{SNR}(n)$ denotes i-th MFCC in frame n at SNR. N and I are a number of frames and an order of MFCC vector respectively. Fig.3 shows variability measures $D_{Variability}^{\infty,20dB}$, $D_{Variability}^{\infty,10dB}$ for SS, CSS and MMSE, calculated from 10 order MFCC vector in speech frames of 65 words of 4 speakers. The symbol ∞ means clean data. 20dB and 10dB mean that car noise is added to the clean data at SNR 20dB and 10dB respectively. Fig.3 shows the variability measure for CSS is lower than those for SS and MMSE. It suggests 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.2 (c.1) and (c.2). On the other hand, speech spectral distortion by CSS is noticeable by comparing Fig.2 (a.1) with (c.1).

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 and test data of 4 speakers 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 2. 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.

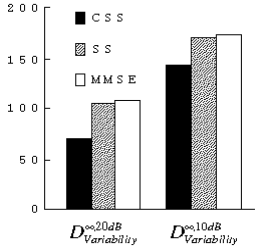


Fig.3: Variability measure of MFCC.

Table 2: Average word recognition rates.

SNR	20dB	10dB
NO	91.9%	84.6%
SS	96.5%	89.6%
CSS	95.4%	93.8%
MMSE	96.9%	91.9%

(property-2)CSS has slightly worse performance than SS and MMSE at higher SNR due to inevitable spectral distortion.

2.4 Suppression of Unknown Additive Noise

We proposed the E-CMN algorithm which is capable of compensating various kinds of multiplicative distortion collectively by normalizing input spectra[1]. The algorithm is described as follows;

(*Estimation Step*) : Two cepstrum mean vectors are calculated. One, obtained from speech frames of sufficiently-long utterance, 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.

This E-CMN is compared with a conventional CMN[5]. The recognition task is speaker-independent 520 Japanese words using 54 context-independent tied-mixture HMMs which are trained from clean speech. The acoustic analysis uses 8kHz sampling, 32ms frame length and 20ms frame shift. The parameters are 10 MFCC(Mel-Frequency Cepstrum Coefficient)s, 10 Delta MFCCs and Delta energy. The number of shared Gaussian distributions are 256, 256 and 64 respectively. One measured impulse response from a mouth of dummy head equipped in a driver's seat to omni-directional microphone mounted on driver's sun-visor, $H_{Trans}(\omega)$, is convoluted with evaluation data. No additive noise is added to the evaluation data. The speech/non-speech decision is done by the same VAD mentioned above. 10 words are used to calculate speaker/environment-dependent cepstrum means by (*Estimation Step*). Word accuracy for no processing, the conventional CMN and E-CMN applied to both HMM training data and evaluation data are 80.1%, 90.8% and 93.3% respectively. It is suggested that E-CMN enables a simultaneous compensation of four kinds of multiplicative distortions described

in subsection 2.1 by operating as an equalizer in frequency domain[1].

2.5 NLMS-VAD/CSS/E-CMN

We propose the combination of NLMS-VAD, CSS and E-CMN. A microphone input is processed with NLMS-VAD to cancel a known additive noise. A mixed signal of left and right channel sources of car stereo system is given as a reference(far-end in) of NLMS-VAD. Secondly, the output of NLMS-VAD is noise-suppressed with CSS. Finally, spectra obtained after CSS is converted to cepstrum domain parameter which is equalized with E-CMN. The VAD result in NLMS-VAD is reused for E-CMN.

3. EVALUATION

3.1 Recognition Task

The recognition task is speaker independent 520 Japanese words with 54 context-independent tied-mixture HMMs. A clean speech is added with a car noise with SNR 10dB and used as a training data for the HMMs after enhanced by CSS and E-CMN. The acoustic analysis condition, acoustic parameters and number of shared Gaussians are the same as those mentioned in subsection 2.4. The known acoustic reverberation and the unknown car noise were recorded in a car cabin in idling, running at 60kmph on city road and running at 100kmph on express way. During the recording, 5 music sources(pops, jazz, rock, classic and rakugo(narration of Japanese comical tale)) were played by a car stereo. These mixed additive noises are added to a clean speech convoluted with the same impulse response in subsection 2.4.

3.2 Performance Analysis

Averages of word recognition rates for 5 music sources at idling, 60kmph and 100kmph are shown in Fig.4. Three cases, case 1:w/o Speaker Out, case 2:w/ Speaker Out & w/o NLMS-VAD, case 3:w/ Speaker Out & w/ NLMS-VAD, are compared. A recognition rate for the case without any additive noises is 80.0%. RRE(Recovery Rate of Error) is defined as

$$[\{r(case\,3)-r(case\,2)\}/\{r(case\,1)-r(case\,2)\}]\times 100 \quad (10)$$

where $r(case\,x)$ represents average word recognition rate for *case x*. We get the RREs over 80% for all driving conditions. Table 3 shows word recognition rates in case 3 for each music source. Although there exist some fluctuations of recognition rates, rather stable performance for any music is realized. We observe that some residual music reverberation after NLMS-VAD might be masked out by (property-1) of CSS.

4. SUMMARY

This paper proposed the NLMS-VAD/CSS/E-CMN which is robust to existence of a known additive noise, a stationary unknown additive noise and a multiplicative distortion in adverse car environments. One of future research goals is to study how to cope with non-stationary additive noises such as conversational speech in a car cabin, a noise of bump, a noise generated in passing a car running to opposite direction and so forth.

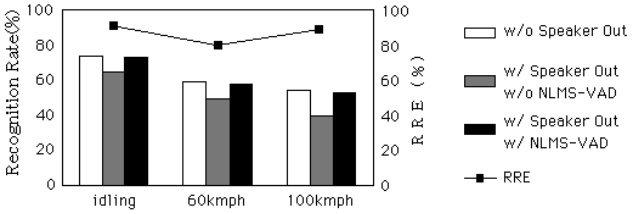


Fig.4: Average ord recognition rate.

Table 3: Word recognition rate for each music source.

	idling	60kmph	100kmph
pops	72.4%	57.1%	53.2%
rock	73.2%	59.2%	49.3%
jazz	73.9%	55.8%	50.8%
classic	72.9%	57.1%	54.2%
rakugo	73.7%	58.5%	54.1%

5. REFERENCES

[1] Shozakai, M., Nakamura, S. and Shikano, K., "A Non-iterative Model-Adaptive E-CMN/PMC Approach for Speech Recognition in Car Environments", *Proc. EUROSPEECH, pp.287-290,* Rhodes, Greece, 1997.

[2] Pouteau, X., Krahmer, E. and Landsbergen, J., "Robust Spoken Dialogue Management for Driver Information Systems", *Proc. EUROSPEECH, pp.2207-2210,* Rhodes, Greece, 1997.

[3] Haykin, S., *Adaptive Filter Theory, 2nd ed.* Englewood Cliffs, NJ, Prentice-Hall, Boston, 1991.

[4] Boll, S. "Supression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. ASSP-27, no.2, pp.113-120,* 1979.

[5] Furui, S., "*Cepstral Analysis Technique for Automatic Speaker Verification*", *IEEE Trans. ASSP-29, no.2, pp.254-272,* 1981.

[6] Nolazco Flores, A. and Young, S. J., "Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation", *Proc. ICASSP, pp.I-409-412,* 1994.

[7] Recommendation GSM 06.32.

[8] Ephraim, Y. and Malah, D., "Speech Enhancement Using a Minimum Mean Square Error Short Time Spectral Amplitude Estimator", *IEEE Trans. ASSP-32, no.6, pp.1109-1121,* 1984.

[9] Acero, A., *Acoustical and Environmental Robustness in Automatic Speech Recognition,* Kluwer Academic Publishers, Boston, 1992.

[10] Scalart, P. and Benamar, A., "A System for Speech Enhancement in the Context of Hands-Free Radiotelephony with Combined Noise Reduction and Acoustic Echo Cancellation", *Speech Communication, 20, pp.203-214,* 1996.

[11] Shozakai, M., Nakamura, S. and Shikano, K., "A Speech Enhancement Approach E-CMN/CSS for Speech Recognition in Car Environments", *Proc. IEEE Workshop of ASRU,* Santa Barbara, 1997.