NOISE REDUCTION BASED ON MICROPHONE ARRAY AND POST-FILTERING FOR ROBUST SPEECH RECOGNITION IN CAR ENVIRONMENTS

Junfeng Li, *Xugang Lu* and *Masato Akagi*

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan
{junfeng, xugang, akagi}@jaist.ac.jp

ABSTRACT

Robust speech recognition in car environments has been an important application and attracted great research interests in recent years. Its performance dramatically degrades due to various kinds of noises existing in car environments. To deal with acoustic noises, we have proposed two noise reduction systems which are based on microphone array and post-filtering. In this paper, we first describe the two noise reduction systems previously suggested. Then, we are devoted to investigate the performance improvements of the automatic speech recognition (ASR) system when the two noise reduction systems are used as the front-end processors. The speech recognition experiments were conducted using multi-channel car noise recordings and AURORA-2J speech database, the recognition results are also reported. Some discussions on the proposed noise reduction systems are finally presented based on the experimental results.

1. INTRODUCTION

In the past several decades, hands-free speech processing technology in car environments has been of increased research interests for many applications, such as *automatic speech recognition* (ASR) system. One main problem associated with this technology is that the signals received by the distant microphones are severely corrupted by various kinds of noises. Although a large number of algorithms have been published so far [1]-[5], the problem of suppressing noise signals and improving the performance of the speech recognition systems in car environments is still very interesting and challenging in speech signal processing field. A potential solution is to construct a practically effective and computationally efficient front-end processor with the objective of developing a robust speech recognition system in adverse environments.

A variety of noise reduction algorithms for in-car applications have been reported in the literature [1]-[5]. Matassoni *et al.* [2] adopted the single-channel based schemes, magnitude spectral subtraction and log-MMSE estimator, to suppress the background noise. The noise-suppressed signals were then used to do speech recognition, resulting in the improved recognition rate. In comparison of single-channel technique, multi-channel technique has shown substantial superiority in reducing noise due to its spatial filtering capability. Zhang *et al.* [4] proposed a “constrained switched adaptive beamformer” for ASR system in real car environments. However, its relatively slow convergence rate degrades its performance in dealing with non-stationary noise signals in practical conditions. Moreover, Grenier evaluated the performance of the *generalized sidelobe canceller* (GSC) beamformer in car environments [3]. Further, he pointed out that the GSC beamformer, as a front-end processor for ASR system, is not effective in high-noise conditions.

In this paper, we first show the characteristics of the noise fields in car environments and introduce two noise reduction algorithms based on microphone array and post-filtering [7]-[10]. The suggested noise reduction algorithms are then used as front-end processors for a speech recognition system to improve its robustness and recognition rate in adverse car environments. Speech recognition results are presented to show the effectiveness of two noise reduction algorithms. Finally, we give some discussions on two noise reduction systems.

2. ANALYSIS OF NOISE FIELD IN CAR ENVIRONMENTS

To characterize a noise field, a widely used measure is the *magnitude-squared coherence* (MSC) function, defined as:

$$\Gamma_{x_i x_j}(k,\ell) = \frac{|\phi_{x_i x_j}(k,\ell)|^2}{\phi_{x_i x_i}(k,\ell)\phi_{x_j x_j}(k,\ell)}, \tag{1}$$

This research is conducted as a program for the “Fostering Talent in Emergent Research Fields” in Special Coordination Funds for Promoting Science and Technology by Ministry of Education, Culture, Sports, Science and Technology, Japan.

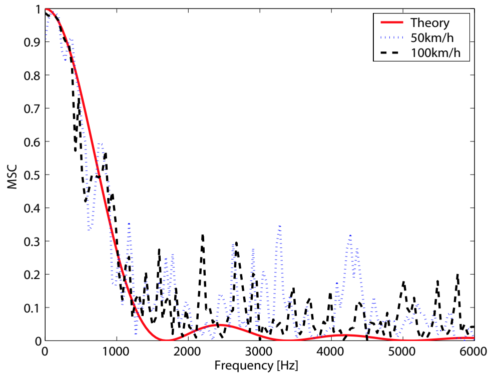

Fig. 1. Magnitude-squared coherence function in car environment (d = 10cm).

where $\phi_{x_ix_j}(k,\ell)$ is the cross-spectral density between two signals $x_i(t)$ and $x_j(t)$; $\phi_{x_ix_i}(k,\ell)$ and $\phi_{x_jx_j}(k,\ell)$ are the auto-spectral densities of $x_i(t)$ and $x_j(t)$, respectively; and k and ℓ are the frequency index and the frame index.

A diffuse noise field has been shown to be a reasonable model for many practical noise environments [6]. The theoretical MSCs of a perfect diffuse noise field against frequency are plotted in Fig. 1, along with the measured MSCs using real-world car noises. From Fig. 1, some characteristics of car noise environments can be easily observed: (i) car noise environment can be modelled as a diffuse noise field; (ii) MSC in car conditions is a frequency-dependent measure; (iii) noises on different microphones are high-correlated in the low frequencies and low-correlated in the high frequencies.

3. NOISE REDUCTION ALGORITHMS BASED ON MICROPHONE ARRAY AND POST-FILTERING

Considering an array with 3 microphones in a noisy environment, shown in Fig. 2, the observed signal on each microphone is composed of desired speech signal, localized noise and non-localized noise. Here, localized noise includes noise component coming from some determinable directions (point noise sources), eg., passenger's interfering speech and other passing car noise. While non-localized noise includes noise components coming from all directions, such as reverberated noise signals in car environments. The objective of this research is to reduce both localized and non-localized noises simultaneously while keeping the desired speech distortionless with the goal of improving the recognition rate and robustness of ASR systems. To implement this idea, we constructed the noise reduction systems, shown in Fig. 3, which consists of localized noise suppression and non-localized noise suppression, detailed in the following.

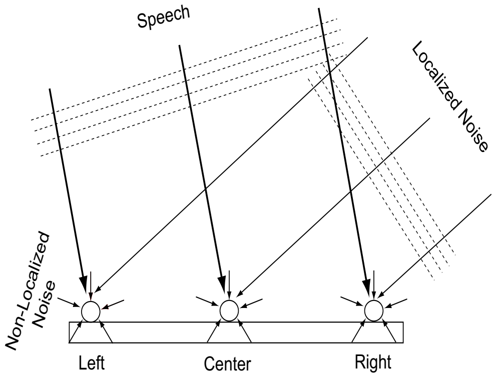

Fig. 2. Microphone array and signal model.

3.1. Localized noise suppression [7][10]

The basic idea of suppressing localized noise is first to estimate the spectrum of localized noise and then to subtract it from that of noisy observation.

To estimate localized noises, the authors have proposed a hybrid noise estimation technique, which combines a subtractive beamformer based multi-channel estimation technique and a soft-decision based single-channel estimation technique, yielding more accurate spectral estimates for localized noises [7]. The spectrum of localized noise, $\hat{N}^c(k)$, calculated by the hybrid technique, is given by:

$$\hat{N}^c(k,\ell) = \begin{cases} \hat{N}^c_m(k,\ell), & \text{not array nulls} \\ \hat{N}^c_s(k,\ell), & \text{array nulls} \end{cases} \qquad (2)$$

where $\hat{N}^c_m(\lambda,\omega)$ and $\hat{N}^c_s(\lambda,\omega)$ are the estimated spectrum for localized noise by the multi-channel technique [7] and the single-channel technique [12], respectively. Furthermore, we presented a *robust and accurate speech absence probability* (RA-SAP) estimator which makes full use of the characteristic of the high estimation accuracy of the multi-channel technique and considers the strong correlation of speech presence uncertainty between adjacent frequencies and consecutive frames [10]. This RA-SAP estimator further enhances the estimation accuracy of this hybrid estimation technique. The estimated spectrum of localized noise is then reduced from that of the noisy observation by non-linear spectral subtraction.

3.2. Non-localized noise suppression

To further suppress non-localized noise, we have presented two post-filters, described in the following.

3.2.1. Post-filter 1 [8][10]

This post-filter is based on the *optimally-modified log-spectral amplitude* (OM-LSA) estimator which is characterized by

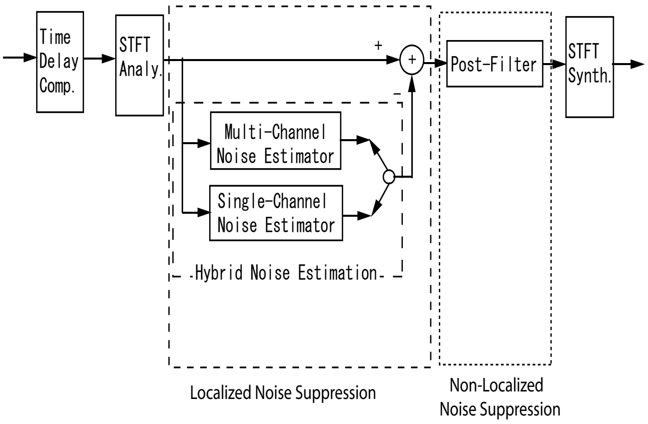
Time Delay Comp.
STFT Analy.
+
+
−
Post-Filter
STFT Synth.
Multi-Channel Noise Estimator
Single-Channel Noise Estimator
-Hybrid Noise Estimation-
Localized Noise Suppression
Non-Localized Noise Suppression

Fig. 3. Block diagram of proposed algorithms.

the following gain function [12]:

$$G(k,\ell) = G_{H_1}(k,\ell)^{1-q(k,\ell)} G_{min}^{q(k,\ell)}, \tag{3}$$

where G_{min}, $q(k,\ell)$, $G_{H_1}(k,\ell)$ are a constraint constant, the *speech absence probability* (SAP) at spectral subtraction output and the gain function of the traditional MMSE-LSA estimator when speech is surely present defined in [11].

According to Eq. (3) and the Bayes' rule, the performance of this post-filter is greatly dependent on the *a priori* SAP. To further improve its performance, we proposed a new estimator for the *a priori* SAP based on the unchanged coherence characteristic of the noise field at spectral subtraction output. Under the assumption of a diffuse noise field, MSCs are first computed at spectral subtraction output. The MSC spectra are then divided into two parts: high frequency region with low MSCs and low frequency region with high MSCs. In the high frequency region, the MSC spectra are further divided into E sub-bands and averaged across the frequencies in each sub-band, obtaining the average MSC $\bar{\Gamma}_e(k,\ell)$ in e-th sub-band. The *a priori* SAP is calculated as: if a high averaged coherence (higher than a threshold $Tmax_e$) is detected, a speech present state is detected presumably; if a low averaged coherence (low than a threshold $Tmin_e$) is detected, a speech absent state is detected presumably. For the MSCs in [$Tmax_e, Tmin_e$], the *a priori* SAPs are determined by the linear interpolation. While, in the low frequency region, we calculate an average MSC $\bar{\Gamma}(k,\ell)$, averaged across the frequencies over the transient frequency f_t which is calculated by $f_t = c/(2d)$, where c and d are the speed of sound and distance of microphones. Using this average MSC $\bar{\Gamma}(k,\ell)$, the *a priori* SAPs are determined following the same ideas in the high frequency region. The estimated *a priori* SAPs are incorporated into the post-filter with the purpose of improving the noise reduction performance of this post-filter [10].

3.2.2. *Post-filter 2* [9]

This post-filter is developed with a hybrid structure, which applies a modified Zelinski post-filter in the high frequencies and a Wiener filter in the low frequencies, to deal with correlated and uncorrelated noise components with the assumption of a diffuse noise field [9]. Under this assumption, uncorrelated noises are found in the frequencies over the transient frequency f_t. Since transient frequencies are determined by the microphone spacings, we can determine the different transient frequencies according to the distances between different microphone pairs. Furthermore, the different transient frequencies divide the full frequency band into some sub-bands. In each sub-band (except the lowest sub-band), noise signals are mutually weakly correlated for the individual frequency of interest on the microphones of the corresponding pair sets. Thus, the spectral densities of desired speech and noisy signal can be estimated from the cross- and auto- spectral densities of multi-channel inputs. Thus, the gain function of the modified Zelinski post-filter is given by:

$$G_{mz}(k,\ell) = \frac{\frac{1}{|\Omega_m|}\sum\limits_{\{i,j\}\in\Omega_m} \Re\{\phi_{x_i'x_j'}(k,\ell)\}}{\frac{1}{|\Omega_m|}\sum\limits_{\{i,j\}\in\Omega_m}\left[\frac{1}{2}\left(\phi_{x_i'x_i'}(k,\ell)+\phi_{x_j'x_j'}(k,\ell)\right)\right]}, \tag{4}$$

where Ω_m is the microphone pair set for m-th sub-band, $x_i^{'}$ is the spectral subtraction output in i-th channel.

In the low sub-band, we adopt a single-channel technique to estimate a Wiener filter. The gain function of this Wiener filter is:

$$G_s(k,\ell) = \frac{SNR_{priori}(k,\ell)}{1 + SNR_{priori}(k,\ell)}, \tag{5}$$

where $SNR_{priori}(k,\ell)$ is the *a priori* SNR, which is updated in a decision-directed scheme which significantly reduces the residual "musical noise" as detailed in [11]. A soft-decision based approach is used to estimate the noise spectrum under speech presence uncertainty [12], which can update the noise estimate even in speech active periods, improving its performance in dealing with non-stationary noise.

4. EXPERIMENTS AND RESULTS

The studied noise reduction systems were used as front-end processors for a speech recognizer. Their performance was evaluated in terms of speech recognition rate in various car noise environments.

Two noise reduction systems were constructed. The first, referred to as Algorithm 1, was composed of the microphone array-based localized noise suppression followed by the post-filter 1. The second, referred to as Algorithm 2, was composed of the microphone array-based localized noise suppression followed by the post-filter 2. The noise reduction systems were first applied on the multi-channel noisy input signals and outputted enhanced speech signals, which entered the speech recognition system. Thus, the performance improvement caused by the noise reduction systems is evaluated based on the recognition rate.

To assess the performance of the proposed noise reduction algorithms, an equally-spaced linear array consisting of three microphones with inter-element spacing of 10cm was mounted above the windshield in a car. The array was about 50 cm apart from and directly in front of the driver (target speech source). The noise recordings were performed across all channels simultaneously, which were mainly composed of engine noise, high air-condition noise and the noise coming from frication between tyres and road. The multichannel noise signals are first re-sampled to 8kHz before doing experiments.

The speech data were selected from AURORA-2J database for training and testing. The acoustic model was trained using 8440 sentences, uttered by 55 persons. For testing, we generated two sets of noise-corrupted data. The first data set (Set A) involved the addition of the randomly selected segments of the multi-channel car noise across 1001 test sentences in AURORA-2J at different SNR levels from 0dB to 20dB in 5dB steps. The second data set (set B) involved the addition of the multi-channel car noise and a secondary speakers speech (passengers interference), which was Japanese digit /ichi/, with DOA of 60 degree to the right, across 1101 test sentences in AURORA-2J at different SNR levels same as above. Data set B corresponds to a realistic context for a typical car environment where a passenger is speaking.

The signals were pre-emphasized with a coefficient 0.97. A hamming window of 32ms length with 16ms frame rate was used. The first 12 dimensions of de-correlated log compressed Mel energy spectrum was chosen (the zero-th order coefficient was discarded). Combining with the log power energy, we got 13 dimensional static feature vector. Together with their first and second order dynamic values, 39 dimensional feature vectors were formed. The acoustic models consist of ten digits, one silence and short pause models. Each distribution of digit has 18 states with 16 output distributions. Silence model has 5 states with 3 distributions, and short pause model has 3 states with one distribution. Each distribution of digit has 20 Gaussians while that of silence and short pause has 36 Gaussians. Each model was trained as a left-to-right topology with three states (without skip among states) by using Baum-Welch algorithm with a flat-starting embedded training. Standard Viterbi decoding technique was used for recognition.

The recognition results for testing data sets A and B are shown in Fig. 4 and Fig. 5, respectively. As Fig. 4 shows, the proposed two algorithms offer high recognition rate than the traditional algorithm (delay-and-sum beamformer followed by the Wiener post-filter) and the noisy inputs at all SNR levels, especially in low SNRs. In comparison of the algorithm 1, the algorithm 2 provides much higher speech recognition rate in all conditions. This superiority is caused by the low speech distortion introduced by the algorithm 2 with regard to the algorithm 1, although the algorithm 1 was proven to be able to improve the intelligibility in subjective evaluations [10].

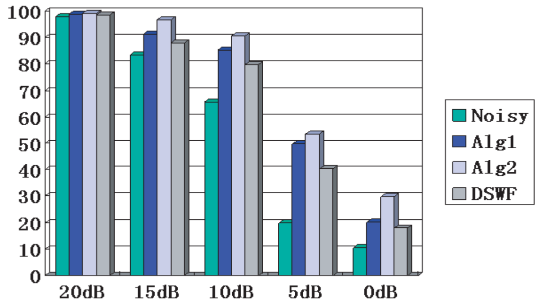

Fig. 4. Speech recognition results for testing data set A.

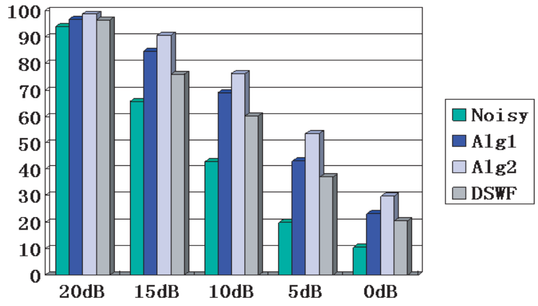

Fig. 5. Speech recognition results for testing data set B.

Concerning the recognition results shown in Fig. 5, we can know that the proposed algorithms also demonstrate high recognition rate at all SNRs. In this noise condition when passenger is speaking, the recognition accuracy went down greatly for unprocessed noisy testing data and the enhanced data by the traditional algorithm. While, the proposed algorithms can deal with both passenger's interference and diffuse background noise, improving the recognition rate in this noise condition. For the comparison of the proposed algorithms, the algorithm 2 still show better performance improvements than the algorithm 1, due to the low distortion it introduces.

5. DISCUSSIONS

The traditional noise reduction algorithm, a delay-and-sum beamformer followed by a Wiener post-filter, does not perform well in car environments. This is because the delay-and-sum beamformer (3ch) only provides very limited noise reduction performance, and the Wiener post-filter does also fail in the low frequencies which are high correlated, where high noise energies are congregated. Furthermore, the very limited noise reduction performance results in its ineffectiveness in improving speech recognition rate.

The proposed algorithm 1 shows the high performance improvements in all conditions, compared with the traditional algorithm. However, its performance is sensitive to the implementation parameters, e.g., $Tmax_e$ and $Tmin_e$. The sensitive parameters dramatically degrade the noise reduction performance, introducing large speech distortion and further greatly degrade the speech recognition performance.

The proposed algorithm 2 offers the highest speech recognition rate among the tested algorithms. This improvement is attributed to the fact that the algorithm 2 can deal with all kinds of noise signals with very low speech distortion. Moreover, its performance is also immune to the implementation parameters in practical adverse environments. The reliable high noise reduction performance of the algorithm 2 further results in the high recognition rate when it is used as the front-end processor of the speech recognition system.

6. CONCLUSIONS

In this paper, we first introduced two noise reduction algorithms we proposed earlier based on microphone array and post-filtering. The main concentration was then put on improving the performance of speech recognition systems when the suggested algorithms were used as front-end processors. The speech recognition results using real-world car noise recordings show that: the proposed algorithms give higher speech recognition rate than the traditional algorithm at all SNRs in all noise conditions; and the algorithm 2 outperforms the algorithm 1 in improving recognition rate of ASR system in all tested environments. This performance improvement can be attributed to the fact that algorithm 2 is able to deal with various kinds of noise and preserve the speech components (low speech distortion) simultaneously.

7. REFERENCES

[1] A. Mrutti, P. Coteetti, *et al.*, “On the development on an in-car speech interaction system at IRST”, In *Proc. of Special Workshop in Maui (SWIM)*, Hawaii, Jan, 2004.

[2] M. Matassoni, M. Omologo and C. Zieger, “Experiments of in-car audio compensation for hands-free speech recognition”, In *Proc. ICASSP2000*.

[3] Y. Grenier, “A microphone array for car environments”, *Speech Communication*, vol. 12, no. 1, pp. 25-39, 1993.

[4] X.X. Zhang and John H.L. Hansen, “CSA-BF: A constrained switched adaptive beamformer for speech enhancement and recognition in real car environments”, *IEEE trans. on speech and audio processing*, vol. 11, no 6, pp.733-744, 2003.

[5] M. Nakayama, *et al.* “An evaluation of in-car speech enhancement techniques with microphone array steering”, In *Proc. ICA2004*, Kyoto, Japan, 2004. M. S.

[6] M. Brandstein and D. Ward (eds.), “Microphone Arrays: Signal Processing Techniques and Applications”, Springer-Verlag, Berlin, 2001.

[7] J. Li and M. Akagi, “Noise reduction using hybrid noise estimation techniques and post-filtering”, In *Proc. ICSLP2004*, pp. 2705-2708, Korea, 2004.

[8] J. Li, X. Lu and M. Akagi, “A noise reduction system in arbitrary noise environments and its applications to speech enhancement and speech recognition”, In *ICASSP2005*, USA.

[9] J. Li and M. Akagi, “A hybrid microphone array post-filter in diffuse noise field”, To appear in *Eurospeech2005*, Portugal.

[10] J. Li and M. Akagi, “A noise reduction system based on hybrid noise estimation technique and post-filtering in arbitrary noise environments”, To appear in *Speech Communication*, 2005.

[11] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimation”, *IEEE trans. on acoustic. speech and signal processing*, vol. 33, no. 2, pp.443-445, 1985.

[12] I. Cohen and B. Berduo, “Speech enhancement for non-stationary noise environments”, *Signal processing*, vol. 81, no. 11, pp. 2403-2418, 2001.