ESTIMATION OF SOURCE LOCATION BASED ON 2-D MUSIC AND ITS APPLICATION TO SPEECH RECOGNITION IN CARS

Takayuki Nagai, Keisuke Kondo, Masahide Kaneko, and Akira Kurematsu

Course in Electronic Engineering, The University of Electro-Communications, Tokyo, Japan.
Email:{tnagai, k-kondoh, kaneko, kure}@apple.ee.uec.ac.jp

ABSTRACT

This paper proposes a speech recognition and an enhancement system for noisy car environments based on a microphone array. In the system, multiple microphones are arranged in 2-dimensional space, surrounding the interior of a car, and the speaker's location is first estimated by our proposed HE (Harmonic Enhanced) 2-D MUSIC (MUltiple SIgnal Classification). Then, 2-D Delay and Sum (DS) is applied to enhance the target speech. Such pre-processing makes robust speech recognition in noisy car environments possible. In the proposed system, not only a driver, but also a fellow passenger can control car electronics by their voices no matter where they are. This is an advantage of the system as well.

The results of the simulation and the preliminary experiment in a real car environment are presented to confirm the validity of our proposed system.

1. INTRODUCTION

Hands free speech recognition and enhancement in cars are important tasks for voice control of car electronics like navigation system, car audio, hands-free mobile phone and so forth. They are also crucial from the traffic safety point of view[1].

In the car environments, there are various noises like ambient background noise that depends on speed of the car, state of the road, position of the windows and so on. Music from the car audio, horns, and directional indicators also degrade speech signals. These noises sometimes make the SNR as low as -15 dB in our experiments. In such noisy environments, error rate of the speech recognition increases enormously. Moreover, since the SNR is quite low, it is more difficult task to remove the noise compared to an indoor environment. A lot of studies have been done to overcome the problem[2]-[6], however, further efforts are required to improve the performance of hands free speech recognition and enhancement.

In this paper, a speech recognition and an enhancement system for car environments is proposed. The system is based on a microphone array. A number of microphones are arranged in 2-dimensional space, surrounding the interior of a car, and the speaker's location is estimated by our proposed HE 2-D MUSIC. Then, 2-D Delay and Sum (DS) is applied to mitigate the interference. Such pre-processing makes robust speech recognition in noisy car environments possible. Another advantage of the method is that it is possible to recognize not only the voice from a driver's seat but also the voice from a passenger seat or a back-seat. The results of the computer simulation are shown to confirm the validity of our proposed system. We also present the result of preliminary experiment in a real car environment.

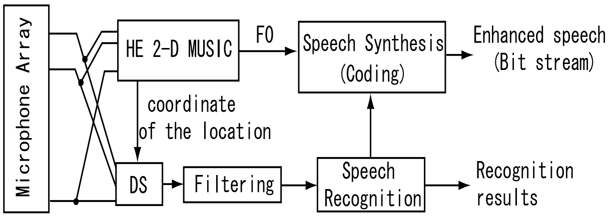
Microphone Array
HE 2-D MUSIC
F0
Speech Synthesis (Coding)
Enhanced speech (Bit stream)
coordinate of the location
DS
Filtering
Speech Recognition
Recognition results

Figure 1: Speech Recognition and Enhancement System

2. PROPOSED SYSTEM

2.1. System Overview

The outline of the proposed speech recognition and the enhancement system is shown in Fig.1. In the system, speaker's position is first estimated from the array input signals by HE 2-D MUSIC described below. The speech signal is emphasized in the DS part based on the estimated speaker's location. The processed speech is recognized at the speech recognition part. When the clean speech is required, like hands-free mobile phone, it is synthesized using the extracted pitch and the recognition results. The synthesized speech is used as an enhanced speech instead of using the output of the DS part directly. This is because the speech signal obtained by the DS part does not have sufficient quality in hearing. Moreover, very low bit-rate speech coding based on HMM[8] is applicable to this scheme.

2.2. Estimation of Speaker's location

2.2.1. Harmonic Enhanced 2-D MUSIC

In general, the direction of arrival (DOA) estimation method like MUSIC considers a planar wave from certain direction θ. In order to apply the MUSIC method to speaker's position estimation in cars, the spherical wave from coordinates (x, y) should be considered, since it is a near field problem.

Now, let $S_q(f, \ell)$ be the Fourier transform of a source signal $s_q(t)$ at a certain frame ℓ, where f denotes a frequency. The sound source is located at the point (x_q, y_q). Then, the Fourier transform of the input signals to M number of microphones can be written as

$$[S_q(f,\ell)e^{-j2\pi f\tau_0} \cdots S_q(f,\ell)e^{-j2\pi f\tau_{M-1}}]^T = S_q(f,\ell)\boldsymbol{d}_q(f,x_q,y_q), \qquad (1)$$

where

$$\boldsymbol{d}_q(f,x_q,y_q) = [e^{-j2\pi f\tau_0}\ e^{-j2\pi f\tau_1} \ldots e^{-j2\pi f\tau_{M-1}}]^T \qquad (2)$$

$$\tau_m = \sqrt{(x_q-\hat{x}_m)^2+(y_q-\hat{y}_m)^2}/c \qquad (3)$$

$$(c: the\ velocity\ of\ sound)$$

and $\hat{x}_i, \hat{y}_i$ denote the coordinates of the i-th microphone. Assuming that there exit K localized sound sources, the input vector of the array can be expressed in time-frequency domain as

$$\boldsymbol{S}(f,\ell) = \sum_{q=0}^{K-1} S_q(f,\ell)\boldsymbol{d}_q(f,x_q,y_q) + \boldsymbol{N}(f), \qquad (4)$$

where $\boldsymbol{N}(f)$ denotes an uncorrelated noise. Then $M \times M$ matrix $\boldsymbol{R}_n(f,\ell)$, which corresponds to the noise subspace, is calculated using eigenvectors $\boldsymbol{V}(f) = [\boldsymbol{V}_0(f) \cdots \boldsymbol{V}_{M-1}(f)]$ of the covariance matrix $\boldsymbol{R}(f,\ell)$ of $\boldsymbol{S}(f,\ell)$ as follows;

$$\boldsymbol{R}_n(f,\ell) = \sum_{m=K}^{M-1} \boldsymbol{V}_q(f)\boldsymbol{V}_q(f)^*, \qquad (5)$$

where $*$ stands for Hermitian transpose. Here, the speaker's location at a certain frame ℓ can be estimated by maximizing the following function $P(f,x,y)$ over the whole frequency;

$$P(f,x,y) = \frac{1}{\boldsymbol{d}_q(f,x,y)^*\boldsymbol{R}_n(f,\ell)\boldsymbol{d}_q(f,x,y)}. \qquad (6)$$

When only one sound source(speaker) exists, it is easy to find the speaker's location, since MUSIC function P has only one peak. Although we assume that two or more people don't speak at once, P has multiple peaks in practice because of the noises such as car audio, radio and so forth. In such case, it is hard to distinguish the speaker's position from the location of undesired noise. To solve this problem, we use harmonic structure of the speech. First, the fundamental frequency(F0) is estimated using the method described below. Then, we add up the MUSIC functions as follows:

$$P(x,y) = \sum_{f\in\Omega} \frac{1}{\boldsymbol{d}_q(f,x,y)^*\boldsymbol{R}_n(f,\ell)\boldsymbol{d}_q(f,x,y)},$$

$$\Omega = \{F_0,\ 2F_0,\ \cdots\ kF_0\}, \qquad (7)$$

where F_0 and k denote the fundamental frequency and a positive integer, respectively. Eq.(7) means that $P(x,y)$ is a sum of MUSIC functions $P(f,x,y)$ at which the frequency f is in an integer multiple of estimated F_0. Fig.3 shows the MUSIC function of HE 2-D MUSIC (a) and that of 2-D MUSIC (b). From the figure, one can see that (a) has only one peak whereas (b) has multiple peaks because of the car audio.

To avoid calculating eigenvectors k times, we use the model expansion[7]. In this method, the eigenvector $\boldsymbol{V}_m(f)$ is expanded over the basis functions $g_n(f)$ as

$$\boldsymbol{V}_m(f) = \sum_{n=0}^{N} \boldsymbol{V}_{mn} g_n(f), \quad m \in \{K,\ \cdots,\ M-1\} \qquad (8)$$

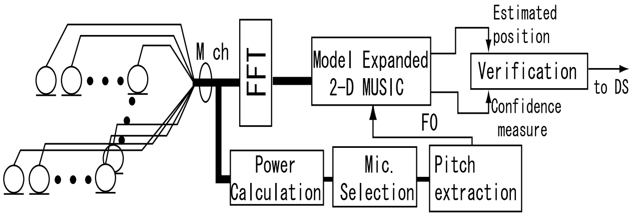

Figure 2: Details of the proposed HE2-DMUSIC

where $\boldsymbol{V}_{mn}$ denotes n-th expanded eigenvector. In order to obtain the frequency dependent eigenvectors $\boldsymbol{V}_m(f)$, M input signals at a certain frequency f are expanded over $g_n(f)$. Then, the covariance matrix of an expanded input signals is diagonalized to obtain frequency independent $\boldsymbol{V}_{mn}$. Finally, set of vectors $\boldsymbol{V}_m(f)$ can be computed by Eq.(8). In later experiments, we use $g_n(f) = (f - 1/4)^n$ with $N = 1$, where f represents normalized frequency ($f \in \{0\ \cdots\ 1/2\}$). Fig.2 illustrates the details of HE 2-D MUSIC.

2.2.2. Microphone Selection and Pitch Estimation

To use the harmonic structure, a fundamental frequency F_0 must be estimated. Prior to the F_0 estimation, we must decide which microphone to use for the pitch detection part. Four microphones are chosen according to their powers. Then F_0 is estimated using the inputs signals of selected four microphones separately and F_0 with highest confidence measure C_{pitch} [9] is selected. Spectrum comb analysis is used to detect F_0, since the method requires low computational cost and has robustness to noise. The confidence measure C_{pitch} as in [9] is used for voiced/unvoiced classification. When the current frame is classified as unvoiced, we interpolate the results of preceding and following frames, since HE 2-D MUSIC is not applicable. The speech period is also detected by using both confidence measure C_{pitch} and power of the frame.

2.2.3. Verification

The result of HE 2-D MUSIC (estimated speaker's location) is verified by the following confidence measure C_{music};

$$C_{music} = \frac{max[P(x,y)] - mean[P(x,y)]}{max[P(x,y)]}, \qquad (9)$$

where $max[P(x,y)]$ and $mean[P(x,y)]$ denotes a maximum value and an average of P(x,y), respectively. When $C_{music} < 0.5$, the estimated coordinate is discarded and is interpolated using the results of foregoing and following frames.

2.3. Delay and Sum (DS) Beamformer

The DS emphasizes the target speech by adding an appropriate amount of delays to the input signals followed by summing them up. As a result of delay compensation, main lobe is formed to the coordinate of the speaker's location that is estimated by the HE 2-D MUSIC method at each frame. This results in an enhanced speech. In our spherical

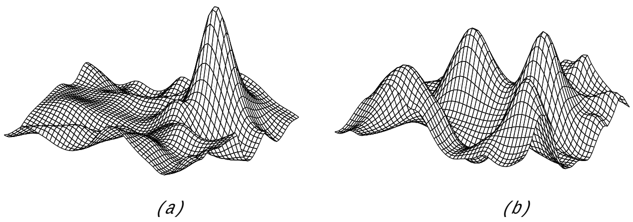

Figure 3: MUSIC Function P. (a)Using a harmonic structure. (b)Without using a harmonic structure.

model, frequency response of DS at (x, y) can be written as follows:

$$H_{xy}(f) = \sum_{m=0}^{M-1} \left[w_m A_m(x,y) \exp\{-j2\pi f(D_m(x,y)/c - \tau_m)\} \right], \quad (10)$$

where

$$D_m(x,y) = \sqrt{(x-\hat{x}_m)^2 + (y-\hat{y}_m)^2},$$

and w_m denotes a weight to m-th microphone. $A_m(x, y)$ represents a function, which corresponds to an attenuation of sound (it is inversely proportional to $D_m(x, y)$ in general). Appropriate choice of w_m can improve the magnitude response of DS, however, we use an unit weight in the later experiments, since it is difficult to obtain a suitable w_m. As is well known, the DS does not work sufficiently at low frequencies. Therefore the components below 100[Hz] are filtered out, since it has little importance for speech recognition.

3. SIMULATED DATA EXPERIMENT

3.1. Conditions for the simulation

The computer simulation of our proposed system was carried out to evaluate the performance. Sixteen microphones were assumed to be arranged as in Fig.4. The sampling rate is 16 kHz and the speech data was synthesized as if the speaker had been located at (0.2, 0.2), which corresponded to the left rear seat. The background noise, which was recorded in the real car, was added to each input speech. We also added the sound of music as the directive noise.

For the speech recognition part, we used five state left-to-right HMMs. Feature parameters contain 16 dimensional MFCCs, 16 dimensional Δ MFCCs, log energy and Δ log energy. Each model is re-estimated by the high-pass filtered speech (male speaker's 201 sentences).

3.2. Speech Recognition Experiments

The word recognition was performed under various SNRs. We used 119 words (voice commands such as, turn on a radio, turn the volume up, etc.) spoken by three male speakers. The recognition results are given in Fig.5. The results for an un-processed speech (original input) are also presented for a comparison purpose. From the figure, the improvement in the word recognition rate can be seen in all SNRs. We have compared the proposed method with Non-linear Spectral Subtraction (NSS) as well. NSS has shown to be effective in the car speech recognition [2] and widely used. The results suggest the effectiveness of the proposed method.

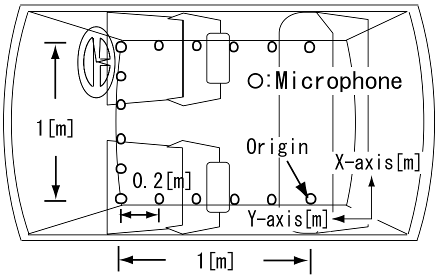

Figure 4: Configuration of microphone array in the car.

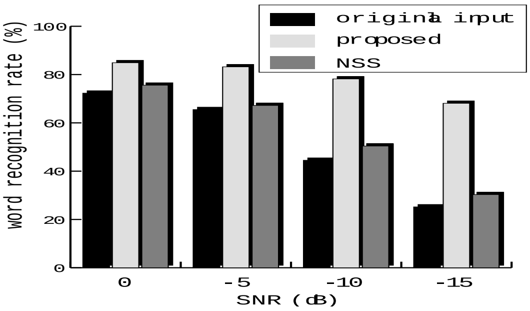

Figure 5: Word recognition rates in various SNRs

4. PRELIMINARY TEST IN THE CAR

4.1. Settings

We have tested the system under a real car environment as a preliminary test. The car is equipped with 16 microphones, 16 channel amplifier, 16 channel A/D converter, laptop PC, and a power supply. Microphones are mounted equally spaced (20 cm) on the front and both sides of the car as in Fig.4. Fig.6 shows the inside of the car, in which we collected the speech data. Three people (a driver, a passenger on the side seat, and a passenger on the back seat) rode in the car. The car was driven about 70 km/h during the speech period. Speech data was collected under various conditions like radio on/off, window open/close, air conditioner on/off, and their combinations.

4.2. Estimation of Speaker's Location and DS

The collected speech data were processed in an off-line manner. Fig.7(a) and (b) shows x and y coordinates of estimated speaker's location at each frame. In this example, the speaker is a male on the side seat and he was asked to fix his mouth position nearly at the point (0.3,0.7). The radio was on and windows were opened. The figure shows that the speaker's position can be estimated almost correctly even under such a cruel condition.

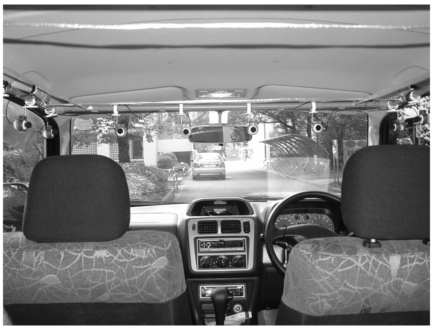

Figure 6: The car used in our experiment.

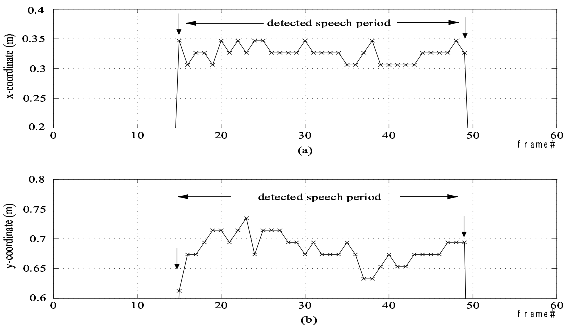

Figure 7: Estimated speaker's position

Fig.8 (a), (b), and (c) illustrates the speech waveform recorded with one microphone (closest to the speaker), the waveform after high-pass filtering (components below 100 Hz are filtered out), and the speech waveform after DS processing, respectively. From the figure, it turns out that the proposed system worked well under the real car environments.

5. CONCLUSION

In this paper, we proposed HE 2-D MUSIC and its application to a speech recognition and an enhancement system for noisy car environments. We showed that the combination of HE 2-D MUSIC and 2-D DS improves the word recognition rate through the simulation. Results of the preliminary experiment in a real car environment was also presented.

Speech recognition experiments using real data (now in progress), consideration of an array configuration, and investigation of the speech synthesis part are left for future research. Moreover, the experiment should be continued using a larger speech data collected in various car environments.

6. REFERENCES

[1] H.R.Pfitzinger, "The Collection of Spoken Language Resources in Car Environments", Proc. of Int. Conf. on Language Resources and Evaluation, vol.2, pp.1097-1100, Granada (May 1998)

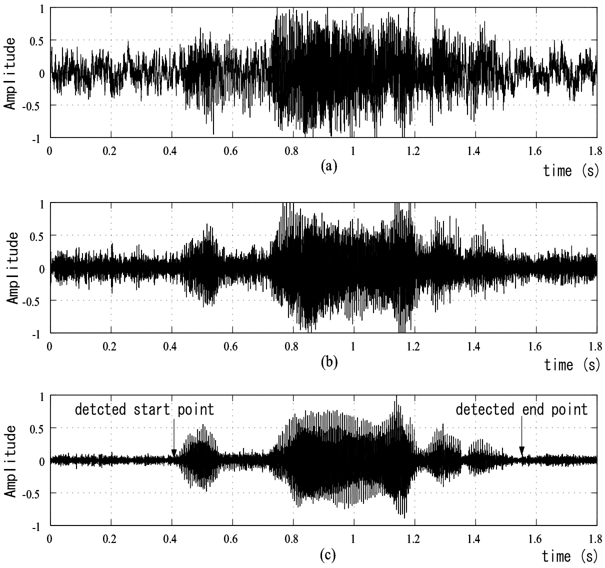

Figure 8: Speech waveforms. (a) Noisy speech. (b) Filtered speech. (c) After DS processing.

[2] C.E.Mokbel and G.F.A.Chollet, "Automatic Word Recognition in Cars", IEEE Trans. on Speech & Audio Process., vol.3, no.5, pp.346-356 (Sep.1995)

[3] S.Nordlholm, I.Claesson, and I.Bengtsson, "Adaptive Array Noise Suppression of Handsfree Speaker Input in cars", IEEE Trans. on Vehicular Technology, vol.42, no.4, pp.514-518 (Nov. 1993)

[4] S.Nordlholm, I.Claesson, and M.Dahl, "Adaptive Microphone Array Employing Calibration Signals: An Analytical Evaluation", IEEE Trans. on Speech & Audio Process., vol.7, no.3, pp.241-252 (May 1999)

[5] J.Hernando, C.Nadeu, and J.B.Marino, "Speech recognition in a noisy car environment based on LP of the one-sided autocorrelation sequence and robust similarity measuring techniques", Speech Communication, vol.21, pp.17-31 (1997)

[6] J.Meyer and K.U.Simmer, "Muti-Channel Speech Enhancement in a Car Environment using Wiener Filtering and Spectral Subtraction", proc. of Int. Conf. on Acoustics, Speech, and Signal Processing, pp.1167-1170 (Apr. 1997)

[7] Y.Grenier, "Wideband Source Location Through Frequency-Dependent Modeling", IEEE Trans. on Signal Processing, vol.42, no.5, pp.1087-1096 (May 1994)

[8] K.Tokuda, T.Masuko, J.Hiroi, T.Kobayashi, and T.Kitamura, "A very Low Bit Rate Speech Coder Using HMM-Based Speech Recognition/Synthesis Techniques", Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing, vol.2, pp.609-612 (May 1998)

[9] D.Wu, M.Tanaka, R.Chen, L.Olorenshaw, M.Amador, and X.Menendez-Pidal, "A Robust Speech Detection Algorithm for Speech Activated Hands-Free Applications", Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing, vol.4 (May 1999)