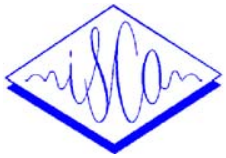

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SPEECH ENHANCEMENT USING NONLINEAR MICROPHONE ARRAY UNDER NONSTATIONARY NOISE CONDITIONS

Hiroshi SARUWATARI †, *Shoji KAJITA*‡, *Kazuya TAKEDA*† *and Fumitada ITAKURA*‡

†Graduate School of Engineering/CIAIR, ‡Center for Information Media Studies/CIAIR, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, JAPAN
E-mail: sawatari@itakura.nuee.nagoya-u.ac.jp

10.21437/Eurospeech.1999-564

ABSTRACT

This paper describes a spatial spectral subtraction method by using the complementary beamforming microphone array to enhance noisy speech signals for speech recognition. The complementary beamforming is based on two types of beamformers designed to obtain complementary directivity patterns with respect to each other. In this paper, it is shown that the nonlinear subtraction processing with complementary beamforming can result in a kind of the spectral subtraction without the need for speech pause detection. To evaluate the effectiveness, speech enhancement experiments and speech recognition experiments are performed based on computer simulations under both stationary and nonstationary noise conditions. In comparison with the optimized conventional delay-and-sum array, it is shown that: (1) the proposed array performs more than 20% better in word recognition rates under the conditions that the white Gaussian noise is used, (2) the proposed array improves the word recognition rate by about 5% when the interfering noise is a single speaker or the overlap of some speakers, (3) the proposed array improves the word recognition rate by more than 10% when the noise is a nonstationary bubble noise.

1. INTRODUCTION

Speech enhancement in noisy environments is a typical and important approach to construct a robust man-machine interface, such as a speech recognition system used in the real world. Among various noise reduction methods, a microphone array is one of the most effective techniques. The Delay-and-Sum (DS) array[1] and the adaptive array[2] are the conventional and popular microphone arrays used for noise reduction. However, they must use a large number of microphones or much computational costs to achieve high performance, especially in low frequency regions. To achieve further improvement, several microphone arrays combined with nonlinear speech processing, such as the Spectral Subtraction (SS) method[3], have been proposed in the recent works[4, 5, 6]. In these methods, however, there exists other problems in terms of degradations of speech quality due to the speech pause detection error or the misestimation of noise directions.

In our recent work[7], a new microphone array system based on nonlinear array signal processing has been proposed. In this system, both complementary beamforming[8] and nonlinear subtraction processing are used to construct the spatial SS without any speech pause detection and the estimation of noise directions. In this paper, to evaluate the effectiveness of the proposed array, speech enhancement experiments and speech recognition experiments are performed based on computer simulations. From the experiments, compared with the conventional DS array, it is shown that the proposed array can perform better in the word recognition rate under both stationary and nonstationary noise conditions.

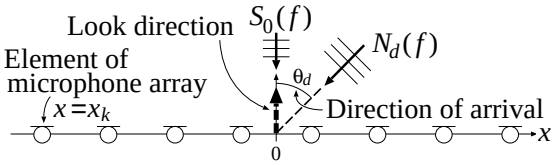

Figure 1: Configuration example of a microphone array and acoustic signals.

This paper is constructed as follows. In the following section, the nonlinear microphone array and its optimization algorithm for directivity patterns are described. In Section 3, some experiments based on computer simulations are performed. After discussions on the results of the experiments, we conclude this paper in Section 4.

2. ALGORITHM

2.1. Nonlinear Microphone Array with Complementary Beamforming

In this study, a straight-line array is assumed. The coordinates of the elements are designated as x_k ($k = 1, \cdots, K$), and the directions of arrival of multiple signals are designated as θ_d ($d = 1, \cdots, D$) (see Fig. 1). Also, the look direction is set to be normal to the array ($\theta = 0$).

First, using two types of complementary weight vectors[8] of element $\boldsymbol{g} = [g_1, \cdots, g_K]$ and $\boldsymbol{h} = [h_1, \cdots, h_K]$, we construct the signal spectra $S^{(g)}(f)$ and $S^{(h)}(f)$. Here, "*complementary*" implies one of the following conditions: "directivity pattern gain $|\boldsymbol{g}\boldsymbol{a}_d(f)| \gg$ directivity pattern gain $|\boldsymbol{h}\boldsymbol{a}_d(f)|$" or "directivity pattern gain $|\boldsymbol{g}\boldsymbol{a}_d(f)| \ll$ directivity pattern gain $|\boldsymbol{h}\boldsymbol{a}_d(f)|$" for an arbitrary direction d (see Fig. 2). The exception is that the gain of both directivity patterns is unity with respect to the look direction. Here, $\boldsymbol{a}_d(f)$ is a steering vector and is defined as follows

$$\boldsymbol{a}_d(f) \equiv [a_{1,d}(f), \cdots, a_{k,d}(f), \cdots, a_{K,d}(f)]^{\mathrm{T}}, \quad (1)$$

$$a_{k,d}(f) \equiv \exp[j2\pi f \cdot x_k \cdot \sin(\theta_d)/c], \quad (2)$$

where c is the velocity of sound. As for the signals $S^{(g)}(f)$ and $S^{(h)}(f)$, the following equations are applicable for the target speech signal arriving from the look direction, $S_0(f)$, and noise signals

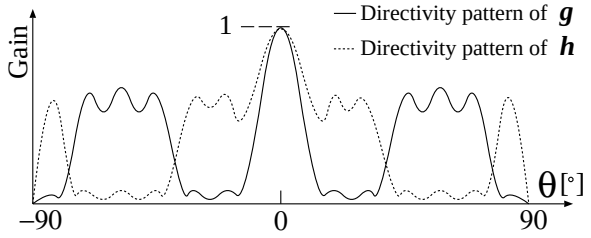

Figure 2: Example of directivity patterns using the complementary beamforming.

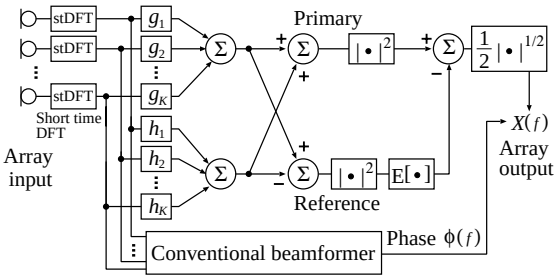

Figure 3: Block diagram of nonlinear microphone array with complementary beamforming shown in Eq. (8).

arriving from other directions, $N_d(f)$.

$$S^{(g)}(f) \;=\; S_0(f) + \sum_{d=1}^{D} \boldsymbol{g}\boldsymbol{a}_d(f) \cdot N_d(f) \tag{3}$$

$$S^{(h)}(f) \;=\; S_0(f) + \sum_{d=1}^{D} \boldsymbol{h}\boldsymbol{a}_d(f) \cdot N_d(f) \tag{4}$$

The sum of Eqs. (3) and (4) is designated as the primary signal, $S^{(\mathrm{p})}(f)$, and the difference is designated as the reference signal, $S^{(\mathrm{r})}(f)$. These can be given as

$$S^{(\mathrm{p})}(f) = 2S_0(f) + \sum_{d=1}^{D} \{\boldsymbol{g}\boldsymbol{a}_d(f) + \boldsymbol{h}\boldsymbol{a}_d(f)\} \cdot N_d(f), \tag{5}$$

$$S^{(\mathrm{r})}(f) = \sum_{d=1}^{D} \{\boldsymbol{g}\boldsymbol{a}_d(f) - \boldsymbol{h}\boldsymbol{a}_d(f)\} \cdot N_d(f). \tag{6}$$

If the directivity patterns $|\boldsymbol{g}\boldsymbol{a}_d(f)|$ and $|\boldsymbol{h}\boldsymbol{a}_d(f)|$ are designed to be complementary, and if there is no correlation among arriving signals, the following approximation holds:

$$\begin{aligned}
&\mathrm{E}\left[\left|\sum_{d=1}^{D}\{\boldsymbol{g}\boldsymbol{a}_d(f) + \boldsymbol{h}\boldsymbol{a}_d(f)\}\cdot N_d(f)\right|^2\right] \\
&\quad\approx \sum_{d\in\Omega_g} |\boldsymbol{g}\boldsymbol{a}_d(f)|^2 \cdot \mathrm{E}\left[|N_d(f)|^2\right] \\
&\qquad + \sum_{d\in\Omega_h} |\boldsymbol{h}\boldsymbol{a}_d(f)|^2 \cdot \mathrm{E}\left[|N_d(f)|^2\right] \\
&\quad\approx \mathrm{E}\left[\left|\sum_{d=1}^{D}\{\boldsymbol{g}\boldsymbol{a}_d(f) - \boldsymbol{h}\boldsymbol{a}_d(f)\}\cdot N_d(f)\right|^2\right] \\
&\quad= \mathrm{E}\left[\left|S^{(\mathrm{r})}(f)\right|^2\right] \qquad (7) \\
&\Omega_g \;\equiv \left\{d \,\middle|\, d \text{ for } |\boldsymbol{g}\boldsymbol{a}_d(f)| \gg |\boldsymbol{h}\boldsymbol{a}_d(f)|\right\} \\
&\Omega_h \;\equiv \left\{d \,\middle|\, d \text{ for } |\boldsymbol{g}\boldsymbol{a}_d(f)| \ll |\boldsymbol{h}\boldsymbol{a}_d(f)|\right\}.
\end{aligned}$$

Therefore, the expectation value of the power spectrum of the noise component in the primary signal (the second term on the right hand side of Eq. (5)) can be approximated by that of the reference signal.

Using the primary and reference signals, without any speech pause detection, we can construct the spatial SS processing by

$$X(f) \equiv \frac{1}{2}\cdot\left||S^{(\mathrm{p})}(f)|^2 - \mathrm{E}\left[|S^{(\mathrm{r})}(f)|^2\right]\right|^{1/2}\cdot e^{j\phi(f)} \tag{8}$$

where $X(f)$ represents the complex spectrum of the speech signal recovered by the proposed method. Also, $\phi(f)$ is an appropriate phase function; for example, the phase function can be obtained by a conventional DS beamformer. Figure 3 shows a block diagram of this array system.

In this algorithm, noise reduction processing is conducted frame by frame, and the expectation value of $|S^{(\mathrm{r})}(f)|^2$ in Eq. (8) is approximately calculated by averaging the power spectra of reference signals over some frames. This interframe-averaged power spectrum is designated as $\left\langle|S^{(\mathrm{r})}(f)|^2\right\rangle$ hereafter.

2.2. Restriction Algorithm for Over-subtraction

In the proposed array processing, when confronted with the non-stationary noises, the instantaneous power spectrum of the reference signal at a frame, $|S^{(\mathrm{r})}(f)|^2$, can be frequently smaller than the interframe-averaged power spectrum of the reference signal, $\left\langle|S^{(\mathrm{r})}(f)|^2\right\rangle$. In such a case, the over-estimation about $\mathrm{E}[|S^{(\mathrm{r})}(f)|^2]$ and over-subtraction from the primary signal in Eq. (8) arise. This over-subtraction causes a degradation of the recovered speech quality.

To avoid this degradation, the following algorithm is introduced in Eq. (8) and conducted frame by frame.

$$|X(f)| \;\equiv\; \begin{cases} \frac{1}{2}\cdot\left||S^{(\mathrm{p})}(f)|^2 - |S^{(\mathrm{r})}(f)|^2\right|^{1/2}, \\ \quad \left(\text{if } \beta\cdot\left\langle|S^{(\mathrm{r})}(f)|^2\right\rangle > |S^{(\mathrm{r})}(f)|^2\right) \\ \frac{1}{2}\cdot\left||S^{(\mathrm{p})}(f)|^2 - \left\langle|S^{(\mathrm{r})}(f)|^2\right\rangle\right|^{1/2}, \\ \quad (\text{otherwise}), \end{cases} \tag{9}$$

where β is a threshold parameter to decide if $\left\langle|S^{(\mathrm{r})}(f)|^2\right\rangle$ on the current frame is appropriate to use in the subtraction processing. In the experiments, β is set to be 0.1. In this algorithm, the instantaneous power spectrum of the reference signal itself is used in the subtraction processing when the power spectrum of the reference signal is decided to be sufficiently smaller than its averaged value. Thus, using this procedure, the restriction for the over-subtraction is achieved.

2.3. Optimization of Directivity Patterns

To make Eq. (8) proper as an estimation of the recovered signal, we design directivity patterns so that the noise component in the expectation value of the power spectrum in the primary signal is

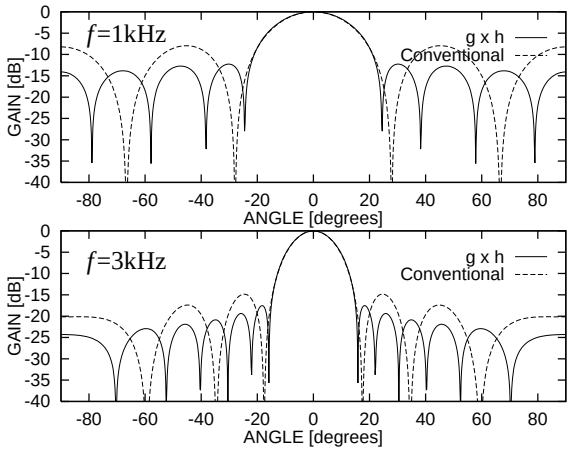

Figure 4: Directivity patterns at 1 kHz (top), and at 3 kHz (bottom). In each figure, the solid lines show the directivity patterns $|\boldsymbol{g}\boldsymbol{a}_d(f)\cdot\boldsymbol{h}\boldsymbol{a}_d(f)|$ in the proposed array and the broken lines show the directivity patterns of the optimized conventional DS array.

decreased. Here, using the $\mathrm{E}[|S^{(\mathrm{p})}(f)|^2]$ instead of the $|S^{(\mathrm{p})}(f)|^2$ in Eq. (8), the estimated power spectrum of Eq. (8) is given as

$$\begin{aligned}|\hat{X}(f)|^2 &= (1/4)\cdot\left|\mathrm{E}\left[|S^{(\mathrm{p})}(f)|^2\right]-\mathrm{E}\left[|S^{(\mathrm{r})}(f)|^2\right]\right| \\ &= \Bigg|\mathrm{E}\left[|S_0(f)|^2\right] \\ &\quad +\sum_{d=1}^{D}\mathrm{Re}\left[\boldsymbol{g}\boldsymbol{a}_d(f)\cdot(\boldsymbol{h}\boldsymbol{a}_d(f))^*\right]\cdot\mathrm{E}\left[|N_d(f)|^2\right]\Bigg| \\ &\le \mathrm{E}\left[|S_0(f)|^2\right]+\sum_{d=1}^{D}|\boldsymbol{g}\boldsymbol{a}_d(f)\cdot\boldsymbol{h}\boldsymbol{a}_d(f)|\cdot\mathrm{E}\left[|N_d(f)|^2\right].\end{aligned} \tag{10}$$

Eq. (10) indicates that the gain for the target signal is one, and the gain for the noise is $|\boldsymbol{g}\boldsymbol{a}_d(f)\cdot\boldsymbol{h}\boldsymbol{a}_d(f)|$. Accordingly, to reduce the noise component in Eq. (10), it is not necessary to produce sidelobes which have small $|\boldsymbol{g}\boldsymbol{a}_d(f)|$ and $|\boldsymbol{h}\boldsymbol{a}_d(f)|$ individually, but to design them so as to obtain a small $|\boldsymbol{g}\boldsymbol{a}_d(f)\cdot\boldsymbol{h}\boldsymbol{a}_d(f)|$ in the directivity patterns.

An eight-element array with the interelement spacing of 5 cm is assumed in the design and the weight vectors $\boldsymbol{g}$ and $\boldsymbol{h}$ are calculated based on the above-mentioned criterion for each frequency independently. The solid lines in Fig. 4 show the directivity patterns $|\boldsymbol{g}\boldsymbol{a}_d(f)\cdot\boldsymbol{h}\boldsymbol{a}_d(f)|$ for 1 and 3 kHz as typical frequencies. For comparison, we also plot the optimized directivity patterns for a conventional DS array based on a single weight vector. It is evident from Fig. 4 that the ability of the sidelobe reduction in the proposed array is improved by about 5 dB for each frequency region.

3. EXPERIMENTS AND RESULTS

In this section, computer simulations are performed to examine the applicability of the proposed method. The performance of proposed array shown in Fig. 4 is compared with the optimized conventional DS array shown in Fig. 4 with respect to the word recognition test.

Table 1: Analysis Conditions for CSR Experiments

Frame Length	25 msec
Frame Shift	10 msec
Feature Vector	12 MFCC + ΔMFCC + ΔΔMFCC ΔPOWER + ΔΔPOWER
Vocabulary	68
Grammar	no grammar

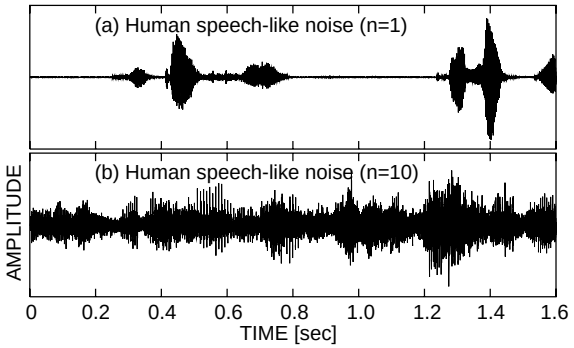

Figure 5: Waveform examples of HSLN with different numbers of superpositions n, (a) nonstationary signal sounds like a single speaker ($n = 1$), (b) nonstationary noise like a bubble noise ($n = 10$).

3.1. Conditions for Experiments

All sound data prepared in this experiments were sampled at 12 kHz with 16 bit resolution. To remove the noise components in lower frequency regions, which cannot be reduced by the conventional DS or proposed arrays, all sound data received by microphones are filtered by a highpass filter. The filter has the gradual transition spectral envelope as follows: the cut off frequency is set to be 500 Hz and the transient characteristic is 14 dB/oct.

Noise reduction processing is conducted frame by frame under the following conditions: the frame length is 21.3 msec, the frame shift is half of the frame length, and the window function is rectangular. The interframe-averaged power spectrum, $\left\langle|S^{(\mathrm{r})}(f)|^2\right\rangle$, is calculated by averaging the power spectra of reference signal over 10 frames.

3.2. Experiment Using Stationary Noise

We generate noisy signals by artificially adding white Gaussian noises to clean speech signals with different signal-to-noise ratios (SNRs) from -10 to 10 dB. The noises are assumed to arrive from a single direction, 50°.

The HMM continuous speech recognition (CSR) experiment is performed in a speaker dependent manner. For the CSR experiment, 10 sentences of one female speaker are used as test data, and the monophone HMM model is trained by 140 phonetically balanced sentences. Both test and training set are selected from the ASJ continuous speech corpus for research. The rest of conditions are summarized in Table 1.

Figure 6 shows the results of the word recognition rates for different input SNRs. As shown in this figure, the recognition rate using a single microphone only is quite low, and the conventional DS array and proposed array are effective to improve the word recognition rate. As compared with the results of the conventional DS array, the proposed method improves the recognition rate by

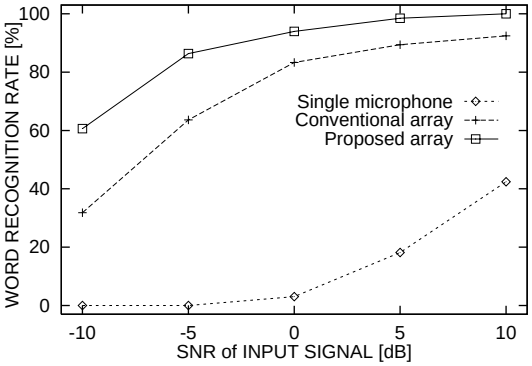

Figure 6: Word recognition rate for different input SNRs under the white Gaussian noise conditions.

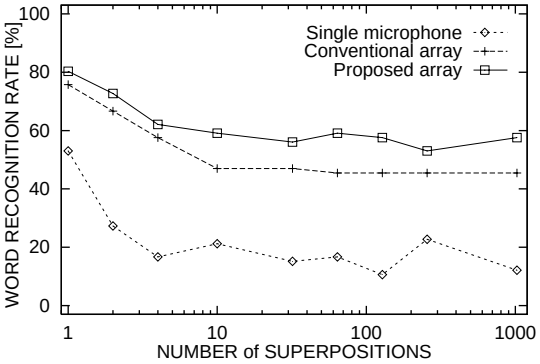

Figure 7: Word recognition rate for different numbers of superpositions in HSLN.

more than 20% under -5 and -10 dB conditions. This indicates that the proposed array is applicable to the speech recognition system under noisy conditions, especially in low speech quality conditions.

3.3. Experiment Using Nonstationary Noise

To evaluate the noise reduction ability of the proposed array for a nonstationary noise, speech recognition experiments are performed using the human speech-like noise (HSLN)[9] as an interfering noise. HSLN is a kind of bubble noise generated by superimposing independent speech signals. By changing the number of superpositions, we can simulate the various noise conditions. For example, the HSLN of one or several superpositions can be considered as a nonstationary signal which sounds like a single speaker or the overlap of some speakers. When the number of superpositions is set to be some dozens, the HSLN becomes a nonstationary signal which sounds like the bubble noise (see Fig. 5). The CSR experiment is performed using the HSLN in the same manner introduced in Sect. 3.2. In this experiment, the input SNR is fixed at 0 dB.

Figure 7 shows the results of the word recognition rates with different numbers of superpositions in HSLN. As compared with the results of the optimized conventional DS array, by applying the proposed method, it is shown that: (1) the improvement in recognition rate of about 5% is obtained when the HSLN of one or several superpositions is used as an interfering noise, (2) the improvement of more than 10% is obtained when the HSLN of several dozens of superpositions or more is used. As shown in these results, the proposed array is applicable to the speech recognition system under the nonstationary noise conditions.

4. CONCLUSION

This paper describes a spatial spectral subtraction method by using the complementary beamforming microphone array to enhance noisy speech signals for speech recognition. From the experiments using the white Gaussian noise, compared with an optimized conventional delay-and-sum array, it is shown that the proposed array can improve a word recognition rate by more than 20% where the input SNR condition is -5 or -10 dB. From the experiments using the human speech-like noise as a nonstationary noise, (1) the proposed array can improve the word recognition rate by about 5% when the interfering noise is a single speaker or the overlap of some speakers, (2) the proposed array can improve the word recognition rate by more than 10% when the noise is the nonstationary bubble noise.

5. ACKNOWLEDGMENT

The authors are grateful to Dr. Mitsuo Komura in SECOM. CO., LTD., who is a co-proposer of the complementary beamforming technique, for his suggestions and discussions on this work. This work was supported by Grant-in-Aid for Science research from the Ministry of Education (JSPS Research Fellowships for Young Scientists).

6. REFERENCES

[1] J. L. Flanagan, J. D. Johnston, R. Zahn and G. W. Elko: "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol.78, no.5, pp.1508–1518 (1985).

[2] L. J. Griffiths and C. W. Jim: "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. AP*, vol.30, no.1, pp.27–34 (1982).

[3] S. F. Boll: "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP*, vol.27, no.2, pp.113-120 (1979).

[4] H. Y. Kim, F. Asano, Y. Suzuki and T. Sone: "Speech enhancement based on short-time spectral amplitude estimation with two-channel beamformer," *IEICE Trans. Fundamentals*, vol.E79-A, no.12, pp.2151–2158 (1996).

[5] J. Meyer and U. Simmer: "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," *Proc. ICASSP 97*, vol.2, pp.1167–1170 (1997).

[6] M. Mizumachi and M. Akagi: "Noise reduction by paired-microphones using spectral subtraction," *Proc. ICASSP 98*, vol.2, pp.1001–1004 (1998).

[7] H. Saruwatari, S. Kajita, K. Takeda and F. Itakura: "Speech enhancement using nonlinear microphone array with complementary beamforming," *Proc. ICASSP 99*, vol.1, pp.69–72 (1999).

[8] H. Saruwatari and M. Komura: "Synthetic aperture sonar in air medium using a nonlinear sidelobe canceller," *IEICE Trans. A*, vol.J81-A, no.5, pp.815–826 (1998) (in Japanese).

[9] D. Kobayashi, S. Kajita, K. Takeda, and F. Itakura: "Extracting speech features from human speech like noise," *Proc. ICSLP 96*, vol.1, pp.418–421 (1996).