

ACOUSTICAL LETTER

A two-element-microphone-array-based speech recognition system in vehicle environment

Heng Zhang, Qiang Fu* and Yonghong Yan

ThinkIT Speech Lab., Institute of Acoustics, Chinese Academy of Sciences, P.R. China

(*Received 25 September 2007, Accepted for publication 16 January 2008*)

Keywords: speech enhancement, null-forming, auditory subbands, speech recognition
PACS number: 43.72.−p, 43.72.Dv, 43.72.Ne [doi:10.1250/ast.30.51]

1. Introduction

The performance of speech communication and automatic speech recognition (ASR) system is often disturbed by environmental noise. Many techniques featuring microphone arrays have been used to improve the performances mentioned above by enhancing desired speech signal while suppressing noise and interference. Some of these techniques are also of great help to hearing aids.

With the help of microphone arrays, we can choose to focus on signals from a particular direction [1]. Better estimation of signal and noise can also be achieved. The Frost beamformer [2] was one of the first array structures to handle adaptive broadband processing by canceling everything that does not come from the look direction. Later, Griffiths and Jim developed an alternative method [3] called generalized sidelobe canceler (GSC) which effectively reduces the computational complexity as well as provides flexibility to implement beamformers according to different designing principles by using the GSC-structure [4]. Other algorithms include Zelinski's approach of post-filtering [5], which employs auto- and cross-correlation functions of signals to estimate the power spectral density (PSD) of signals and noise.

The scheme of adaptive null-forming based on differential microphone technique was first put forward by Elko and Pong in 1995 [6], and developed by Luo *et al.* [7], which features a simple structure employing two omni-directional microphones in end-fire orientation. Compared to beamforming algorithms using more than 4 microphones, it is hard to achieve sharp receiving pattern with less sensors. Therefore, instead of trying to form a narrow beam aiming at the speech source, null-forming focuses on forming a receiving pattern with a null steered to the noise source adaptively while maintaining the desired signal coming from the front.

In this paper, a modified version of adaptive null-forming with auditory subbands [8] is used to weaken the performance degradation caused by narrow-band effect. The signal is decomposed into several subbands according to auditory masking effect. This increases SNR considerably while preserving the auditory effect. A single channel speech enhancement module is applied afterwards to further enhance the desired speech signal. Experiments shows that with this approach the recognition rate of an ASR system is effectively improved in vehicle environment.

*e-mail: qfu@hccl.ioa.ac.cn

2. Single-band adaptive null-forming scheme [7]

The adaptive null-forming algorithm [7], with two microphone in end-fire orientation, is shown in Fig. 1, in which

Fore	signal received by microphone in the front
Back	signal received by microphone in the back
θ	signal arrival angle
d	the distance between the microphone pair
c	the propagation speed of sound wave
$x(n)$	first-order differential result of upper branch
$y(n)$	first-order differential result of lower branch
$W(n)$	coefficient of adaptive filter
$z(n)$	system output

We take the front microphone as a reference and have $x(n)$, $y(n)$ and $z(n)$ as

$$x(n) = 1 - e^{-j2\pi f\frac{d}{c}(1+\cos\theta)} \tag{1}$$

$$y(n) = e^{-j2\pi f\frac{d}{c}} - e^{-j2\pi f\frac{d}{c}\cos\theta} \tag{2}$$

$$z(n) = 1 - e^{-j2\pi f\frac{d}{c}(1+\cos\theta)} - W(n) \times (e^{-j2\pi f\frac{d}{c}} - e^{-j2\pi f\frac{d}{c}\cos\theta}) \tag{3}$$

where d is the spacing between the two microphones. All the right terms of the equations above should be multiplied by the signal received by the front microphone.

It can be concluded that

$$W(n) = -\frac{\sin\left(\pi f\frac{d}{c}(1+\cos\theta_{\text{null}})\right)}{\sin\left(\pi f\frac{d}{c}(1-\cos\theta_{\text{null}})\right)} \tag{4}$$

when output power is minimized.

With the approximation $\sin\theta \approx \theta$ within the frequency range of interest, (4) can be approximated as

$$W(n) = -\frac{1+\cos\theta_{\text{null}}}{1-\cos\theta_{\text{null}}} \tag{5}$$

Output power can be formulated as:

$$E[z^2(n)] = E[(x(n) - W(n)y(n))^2] = R_{xx} - 2W(n)R_{xy} + W^2(n)R_{yy} \tag{6}$$

Minimizing (6) leads to

$$W_{\text{opt}} = \frac{R_{xy}}{R_{yy}} \tag{7}$$

which can be calculated iteratively.

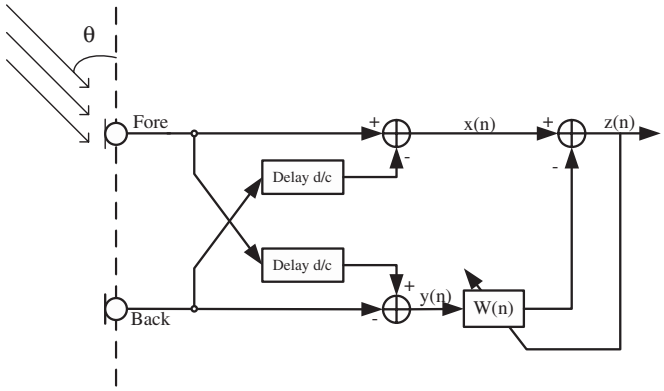

Fig. 1 Adaptive null-forming.

3. Auditory subband null-forming [8]

From [7], it can be learned that the relation between θ_{null} and $W(n)$ is monotonic. That is, one $W(n)$ decides only one θ_{null}. But considering the frequency cue we lose while simplifying (4) to get (5), this is not exactly the case. We can presume an example, when $d = 0.0425\,\text{m}$, $c = 340\,\text{m/s}$, $W(n) = -0.5$, and see θ_{null} vary at different frequency.

Freq.(Hz)	1,000	2,000	3,000	4,000
$\theta_{\text{null}}(^\circ)$	110.75	114.39	124.88	167.63

That is to say, when the noise source comes from an angle of θ and we get a $W(n)$ adaptively, there is only one frequency point f_{opt} at which the receiving pattern has a null at θ. At other frequencies far away from f_{opt}, the null will deviate so that the system can not cancel these components as effectively as those near f_{opt}.

Furthermore, when the number of noise sources is more than one, the algorithm has a difficulty to steer the null to the right direction. θ_{null} will either converge between directions of the noise sources as an effect of average or vibrate between them. The noise reduction effect will thus be weakened.

To solve these problems, subbanding scheme, which is a common practice in array applications [9], is empolyed. And a system is developed as shown in Fig. 2. Differential results of upper and lower branch are decomposed into a series of subbands respectively. Adaptive null-forming is implemented in each subband, after which the results are combined to make the final output. This enables the system to form a null separately in each band so the effect of deviation with frequency mentioned above is reduced. And angle of null in each band can be steered to different directions when under the circumstance of multiple interferences.

As speech is concerned, the energy of desired signal mainly centralizes in low frequencies, so the signal in this area appears to be more *colorful*, while in higher frequencies, signal energy appears to be much weaker. So it is reasonable that nonuniform filterbanks, instead of the uniform ones, should be used to make the low frequency bands narrower to proceed explicit analysis while in the high frequency bands, the bandwidth should be broader to contain more signal energy in order that the adaptive filters may converge more smoothly.

The subbands in this approach are made according to the Bark frequency group [10]. The signal components within each group are judged integrally by brain. Thus the enhanced speech will sound more natural if division in frequency domain is made according to this biological basis. Our experiments show that better result can be achieved by employing auditory subbands compared with some other subbanding schemes. The filterbank for subbanding features FIR filters with an order of 100 to provide frequency response with stop band adequately narrow.

4. System evaluation

4.1. Simulation result

To test the performance of the proposed system, a computer based experiment is carried out in which a small room of $5\,\text{m} \times 4\,\text{m} \times 3\,\text{m}$ with a reverberant time of approximate 300 ms is simulated using image method [11]. Two

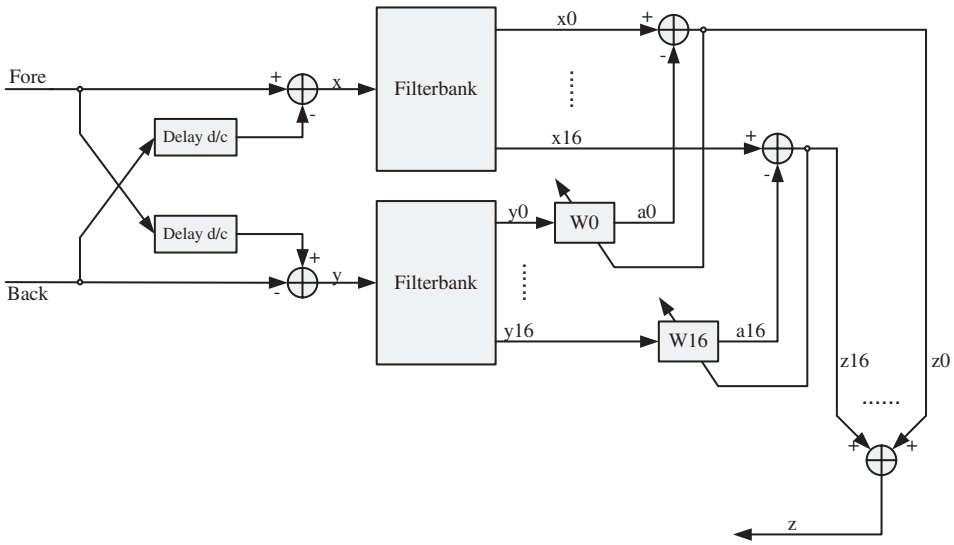

Fig. 2 Auditory subband null-forming.

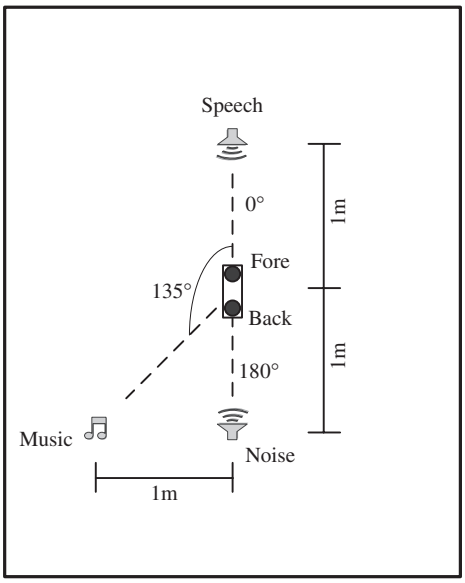

Fig. 3 Simulation experiment.

microphones are placed in the center of the room when speech and noise sources are assigned as Fig. 3.

Speech source is placed in front of the array, at the direction of 0° and is 1 m away from the microphones. White Gaussian noise is placed in the back, at 180° with the same distance away as speech. A non-stationary interference (music) is at 135°, 1 m in the left of Gaussian noise. The spacing between the two microphones is set to be 4.25 cm.

In one group of the experiments, the music source is mute. There is only one source of noise under this circumstance. And in the other group, there are two interference sources. In both groups, signals are recorded at a sampling frequency of 8,000 Hz and with interferences in different intensities. All signals are simulated to be recorded ideally so calibration is not necessary here. The system performance is recorded in Table 1.

SNR_{in} indicates the input SNR measured at one of the two microphones. And $SNR_{\text{gain}}^{\text{ori}}$, $SNR_{\text{gain}}^{\text{uni}}$ and $SNR_{\text{gain}}^{\text{audi}}$ denote the improvement on SNR by original single-band null-forming algorithm, subband null-forming using uniformly divided subbands, and the proposed method, respectively. Here the number of bands and filter length of the uniform filterbank are the same as the auditory filterbank used in the proposed algorithm. Compared to the original single band null-forming and null-forming within uniformly divided subands, the auditory null-forming can constantly bring about an SNR improvement of 2–3 dB.

4.2. Recognition result

Recognition experiment is carried out in vehicle environment. An array formed by two omni-directional microphones is placed at dashboard. About 350 Chinese phrases are uttered by a male voice coming from half a meter away from the array. The recorded signal is processed by the algorithm mentioned above and is further enhanced by a single channel speech enhancement module employing the minimum statistics noise estimation method [12] and Wiener filtering. The recognition result by an in house HMM-based embedded ASR system [13] is listed in Table 2.

Table 1 Performance comparison — Simulations (Unit: dB).

Group	Single interference			Double interferences		
SNR_{in}	−8.9	0.9	11.8	−10.8	−0.7	7.3
$SNR_{\text{gain}}^{\text{ori}}$	9.5	9.2	9.3	11.2	11.0	11.1
$SNR_{\text{gain}}^{\text{uni}}$	9.6	9.7	9.2	11.9	11.5	10.9
$SNR_{\text{gain}}^{\text{audi}}$	12.2	12.3	10.9	13.9	13.5	12.1

Table 2 Recognition performance (correct rate%).

Cases	*Noisy*	*Audi.*	*Audi.* + *SC*
Case 1	64.3	70.7	77.3
Case 2	86.7	88.6	88.6
Case 3	15.2	35.9	29.0

In Case 1, the vehicle is running on a city high road at a speed of about 75 km/h with windows closed. In Case 2 and 3, the vehicle is parked by roadside with windows open. Specially in Case 3, the radio system is turned on and the nearest loudspeaker from the array is about 35 cm away. *Noisy*, *Audi.* and *Audi.*+*SC* in Table 2 represent the unprocessed signal (recorded by one of the two microphones), the result of auditory subband null-forming and the final result (processed by both auditory subband null-forming and single channel module), respectively. It's obvious that in normal applications (Case 1) the subband null-forming and single channel module can both provide an improvement of 6–7%. In Case 2 when it's more quiet, the entire system can also contribute about 2%. And in most challenging occasion of Case 3, the subband null-forming improves the performance greatly (though it's still poor). And the result after single channel becomes worse because interfering speech and other transient noise aggravate the signal distortion brought by single channel process.

5. Summary

We propose a two-element-microphone-array based speech recognition system featuring auditory subband null-forming and a single channel speech enhancement module. Experiments show that the proposed methods effectively improve the speech quality and recognition rate. Further work may concern the combination of null-forming and post-filtering to deal with transient interferences as in Case 3 mentioned in previous chapter.

Acknowledgement

This work is partially supported by MOST (973 program2004CB318106), National Natural Science Foundation of China (10574140, 60535030), The National High Technology Research and Development Program of China (863 program, 2006AA010102, 2006AA01Z195). The authors would like to thank the anonymous reviewers who help to improve the work.

References

[1] D. G. Manolakis, V. K. Ingle and S. M. Kogon, *Statistical and Adaptive Signal Processing* (McGraw-Hill Education and Tsinghua University Press, Beijing, 2000).

[2] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, **60**, 926–935 (1972).

[3] L. J. Griffths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, **30**, 27–34 (1981).

[4] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications* (Springer-Verlag, Berlin, 2001).

[5] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," *ICASSP-88*, Vol. 5, pp. 2578–2581 (1988).

[6] G. W. Elko and A.-T. N. Pong, "A simple adaptive first-order differential microphone," *ASSP Workshop*, Vol. 15, pp. 169–172 (1995).

[7] F.-L. Luo, J. Yang, C. Pavlovic and A. Nehorai, "Adaptive null-forming scheme in digital hearing aids," *IEEE Trans. Signal Process.*, **50**, 1583–1590 (2002).

[8] H. Zhang, Q. Fu and Y. Yan, "Adaptive null-forming algorithms with auditory sub-bands," *ISCSLP 2006*, LNAI 4274, pp. 248–257 (2006).

[9] W. H. Neo and B. Farhang-Boroujeny, "Robust microphone arrays using subband adaptive filters," *Vision, Image and Signal Processing, IEE Proc.*, Vol. 149-1, pp. 17–25 (2002).

[10] K. Yi, B. Tian and Q. Fu, *Speech Signal Processing* (*Chinese*) (NDIP, Beijing, 2000).

[11] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, **65**, 943–950 (1979).

[12] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. Eur. Signal Processing Conf. '94*, pp. 1182–1185 (1994).

[13] J. Shao, J. Han and Y. Yan, "An efficient approach of constructing search space for embedded speech recognition," *NCMMSC '2005, Technical Acoustics*, Vol. 24, pp. 157–160 (2005).