MULTI-MICROPHONE NOISE REDUCTION TECHNIQUES FOR HANDS–FREE SPEECH RECOGNITION –A COMPARATIVE STUDY–

Joerg Bitzer[†], *Klaus Uwe Simmer*[‡] *and Karl-Dirk Kammeyer*[†]

[†] University of Bremen, FB–1, Dept. of Telecommunications
P.O. Box 330 440, D–28334 Bremen, Germany, email: bitzer@comm.uni-bremen.de
[‡] Aureca GmbH,
Mozartstrasse 26, D–28203 Bremen, Germany, email: uwe.simmer@aureca.com

ABSTRACT

In this paper we will describe different multi-microphone noise reduction techniques as for example front-end systems for a command speaker-independent word recognizer in an office environment. Our focus lies on examining the recognition rate if the noise source is not gaussian and stationary, but a second speaker in the same room. In this case standard noise reduction techniques like spectral subtraction will fail, and only multi-microphone techniques can raise the recognition rate by using the spatial information of the speakers. We will compare the delay-and-sum-beamformer, different superdirective solutions and the post-filter approach. Our results show that these techniques are the right choice for hands–free speech recognition systems. Furthermore, we will give a deeper insight into the superdirective design for microphone arrays.

1. INTRODUCTION

More and more users of personal computers work with speech recognition devices to control their systems. The input microphone is almost always headset mounted. This restriction is very uncomfortable. To avoid this restriction, hands–free devices are the right choice, but the recognition rate decreases dramatically, as the signal-to-noise ratio (SNR) decreases. Many publications address this problem with the focus on broad-band, slowly varying noise. Single- and multi-microphone approaches are known [1, 2, 3, 4]. This contribution deals with the problem of a second speaker in the same room. Therefore, the interference signal is coloured and non-stationary. For a two-channel system a possible solution is published in [5]. This algorithm is a derivation of a two-channel generalized sidelobe canceller [6]. But it can be shown that this kind of algorithms fails in reverberant environments [7]. In this contribution we will focus on non-adaptive solutions. In section two we describe the different approaches for multi-microphone noise reduction techniques. Especially the superdirective design is explained and a design procedure based on the coherence function is given. Section three shows the results of the noise reduction experiments.

2. NOISE REDUCTION ALGORITHMS

Noise reduction in a reverberant environment is a difficult problem. We will focus on multi-microphone algorithms which have good capabilities to suppress diffuse noise.

2.1. Beamformer

A general description of a non-adaptive beamformer is given by

$$y(n) = \sum_{i=0}^{N-1} x_i(n) * a_i \tag{1}$$

where N is the number of receivers, x_i the input signal at the sensor i, and $(*)$ denotes the convolution with an arbitrary designed filter a_i, which includes the steering delays. Obviously, this system can also be realized in the frequency domain in order to avoid the convolution operation. Additionally, the steering is much easier in the frequency domain, as the fractional delay is only a multiplication with a linear phase term. We are using a standard overlap-add block-processing with zero-padding, a hamming window and a 50% overlap. For a single frequency bin the beamformer is given by

$$Y(\omega) = \sum_{i=0}^{N-1} X_i(\omega) A_i(\omega) \quad . \tag{2}$$

The only problem is the design of a good filter $A_i(\omega)$. The optimal solution for the design problem under robustness constraints is well-known and given in [8]. Our focus in this work lies on a unified description of the superdirectivity design for different noise fields. In contrast to other publications [9, 8, 10, 11, 12], we will show that the complex coherence function plays a keyrole in the design of good beam patterns. The complex coherence is defined as the normalized cross-power spectral density of two sensors.

$$\Gamma_{X_0X_1}(\omega) = \frac{P_{X_0X_1}(\omega)}{\sqrt{P_{X_0X_0}(\omega)P_{X_1X_1}(\omega)}} \quad (3)$$

We define a coherence matrix for all sensor pairs at one frequency ω.

$$\Gamma = \begin{pmatrix} 1 & \Gamma_{X_0X_1} & \Gamma_{X_0X_2} & \cdots & \Gamma_{X_0X_{N-1}} \\ \Gamma_{X_1X_0} & 1 & \Gamma_{X_1X_2} & \cdots & \Gamma_{X_1X_{N-2}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \Gamma_{X_{N-1}X_0} & \Gamma_{X_{N-1}X_1} & \Gamma_{X_{N-1}X_2} & \cdots & 1 \end{pmatrix} \quad (4)$$

The propagation vector of the desired speech signal for a linear sensor array is

$$\mathbf{d} = \left[e^{-j\frac{\omega}{c}d_{0c_s}\cos\theta}, \cdots, 1, \cdots, e^{-j\frac{\omega}{c}d_{N-1c_s}\cos\theta}\right]^T \quad (5)$$

where d_{0c_s} denotes the distance of the sensor 0 to the center sensor c_s, c the speed of sound, and θ the direction of arrival. Thus, the optimal filter coefficients can be computed according to [8]:

$$A = \frac{\Gamma^{-1}\mathbf{d}}{\mathbf{d}^*\Gamma^{-1}\mathbf{d}} \quad (6)$$

The main advantage of this representation over the others is that all knowledge about the coherence function of noise fields can be used for the design of the array coefficients $A_i(\omega)$.

Let us start with some well-known theoretically defined noise fields. For the uncorrelated noise field the complex coherence is zero at all frequencies. The solution of equation 6 leads to the delay-and-sum-beamformer $|A_i(\omega)| = 1/N$ plus the linear phase term given by the steering direction. This is obvious, as the delay-and-sum-beamformer is the optimal solution for uncorrelated noise. If we assume an isotropic noise field in three dimensions (diffuse noise field), the coherence is given by [13]

$$\Gamma_{X_iX_j}(\omega) = \frac{\sin(\frac{\omega d_{ij}}{c})}{\frac{\omega d_{ij}}{c}} \quad (7)$$

The result of equation 6 leads to the standard superdirective beamformer.

For the isotropic noise field in two dimensions the coherence is[14]

$$\Gamma_{X_iX_j}(\omega) = J_0\left(\frac{\omega d_{ij}}{c}\right) \quad (8)$$

where J_0 is the zero-th-order bessel function of the first kind. This leads to the solution of [12] as an improved superdirective design for speech enhancement.

Neither solutions can be used directly for array design, because the results required infinite precision of the sensors. In other words, all sensors must have exactly the same gain and phase response and no sensor noise is allowed. To avoid this problem Gilbert and Morgan [9] recommended to add a small scalar at the main diagonal of the cross-correlation matrix. The drawback of this approach is that only a constant with no physical interpretation is used. Our solution is slightly different. We want to retain the interpretation as a coherence matrix. The noise variance of the sensors can be included in the coherence function. For example in a diffuse noise field, plus an uncorrelated noise with variance σ_n^2, the coherence is

$$\Gamma_{X_iX_j}(\omega) = \frac{\sin(\frac{\omega d_{ij}}{c})}{\frac{\omega d_{ij}}{c}(1 + \frac{\sigma_n^2}{P_{NN}(\omega)})} \quad (9)$$

where $P_{NN}(\omega)$ is the assumed noise power spectral density of the diffuse noise field. Now it is possible to compute coefficients with an optimized constraint for every desired sensor-noise-to-room-noise-ratio. Therefore, a physical interpretation of the additive constant is given. Typical values are ratios about $-20dB \cdots -40dB$.

In order to get an optimal design for a specific noise field it is possible to include a-priori information in the coherence function. Either as a result of the theoretical analysis of the noise field (for instance if the direction of a single interferer is known), or by measuring the complex coherence of the noise field.

2.2. Beamformer and Post-Filter

Delay-and-sum-beamformers can be extended by so-called post-filter structures. These filters estimate the transfer function with the help of the auto- and cross-power spectral densities of the different input channels. The transfer function of an optimal post–filter can be estimated as [15]

$$\widehat{W}(\omega) = \frac{\frac{2}{N^2 - N}\sum_{i=0}^{N-2}\sum_{j=i+1}^{N-1} Re\left\{P_{X_{ij}}\right\}}{P_{Y_bY_b}} \quad (10)$$

3. EXPERIMENTS AND RESULTS

To demonstrate the necessity of arrays for noise reduction when a second speaker is the disturbance we examined the following situation. In figure 1 the simulated situation is shown. A main speaker is straight in line with the endfire array of four omnidirectional microphones. The interference speaker is relatively close to the microphone array, but 1.2m (60°) left of the main beam. This situation could occur in an office with more than one person using speech as the input device to control the computer. The reverberation is simulated by the image method in the frequency domain

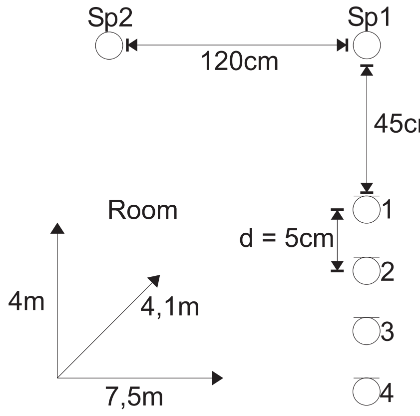

Figure 1: Configuration for the speech recognition experiment

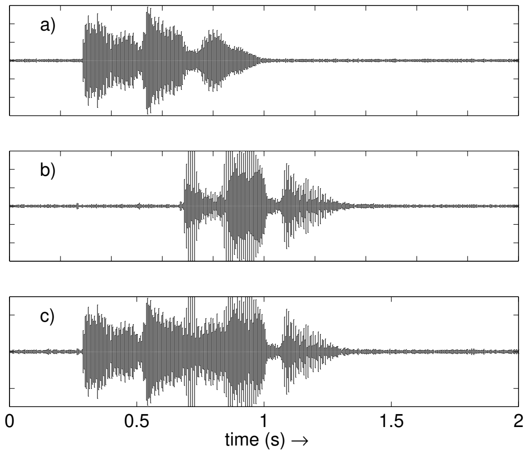

Figure 2: Example for a speech signal. a) = Original, b.) Interference, c.) Sum of desired and interfering speaker

in order to get fractional delays. The reverberation time τ_{60} is set to $300ms$ and $100ms$. Our speech-recognition system is a speaker-independent isolated word-recognizer. The feature vector consists of short-time modified coherence coefficients (SMC) [16] of order twelve. The mapping routine is dynamic time warping. The recognizer was trained with clean speech (16 speakers, 4 utterances, 40 words including numbers and short commands).
The test material was speech from 4 different speakers (not included in the training phase (4 utterances, 40 words) convolved with the room impulse response for the endfire direction and additionally mixed with speech convolved with the impulse response of the interference direction. The voice activity detection (VAD) is assumed to be perfect (hand-labeled), but VAD for this situation is another unsolved problem. Figure 2 shows an example of a mixed signal.
We tested four different algorithms as noise reduction front-ends.

1. Delay and sum beamformer (D&S, optimal for uncorrelated noise).
2. Standard superdirective (SSD, optimal for diffuse noise)
3. Superdirective with a-priori information (SDOPT, optimal for this situation).
4. D&S + Post-filter

Recognition Rate (%)		
	τ_{60}=300ms	τ_{60}=100ms
No noise, no reverb.	93.3	
No noise, reverb. (1.Mic)	92.7	
Mixed signal (1.Mic)	42.2	48.6
Mixed signal (4.Mic)	34.5	42.2
D&S	41.3	47.5
SSD	62.0	62.3
SDOPT	62.3	68.6
Post-filter	42.2	46.6

The results show, that the recognition rate is fair in the undisturbed case, even in highly reverberant environments. But if a second speaker is in the room the recognition rate decreases dramatically, and the influence of the reverberation is higher. The result for the fourth microphone in contrast to the first microphone is a loss of 6-8%. As a first result we can see that for a single microphone system the microphone should be as close to the speaker as possible. Surprisingly, the rate is neither increased by the D&S nor by the post-filter structure, even though we are using four microphones now. This result can be explained if we look at the beam-pattern in figure 3. The directivity for the delay and sum beamformer is very small for arrays with a distance of only $5cm$. Since the noise reduction performance of the post-filter structure is directly connected to the capabilities of the D&S to suppress noise [15], this structure in not efficient for this problem. The noise reduction performance can be optimized by increasing the distance.
Only the superdirective beamformer algorithms increase the recognition rate significantly. The optimized version includes information about the direction of the noise source and therefore it reduces the interference signal much better. In figure 3 the beam-pattern at 750Hz is shown. The jammer direction (60°) can be seen clearly. However, due to reverberation additional jammer directions appear which cannot be suppressed by superdirective beamformers.

4. CONCLUSION

In this contribution we have shown that superdirective beamformers are a good choice to suppress interfering signals which have the same statistics as the desired signal by using spatial information. The results clearly show that reverberation decreases recognition rates significantly and therefore

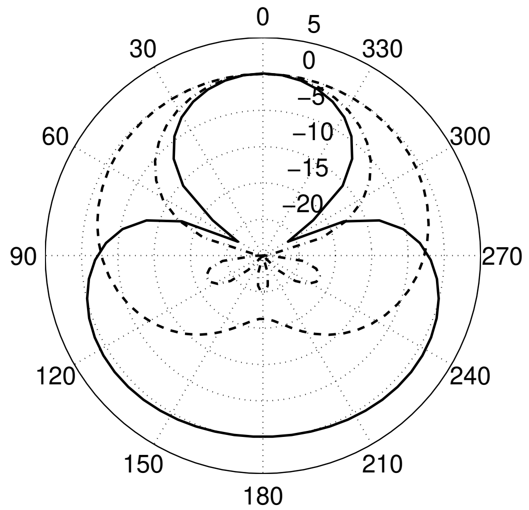

Figure 3: Beam pattern at 750Hz: – – delay and sum, · – superdirective, — superdirective with a-priori information

special care has to be taken for the design of office speech-recognition systems. Additionally, a new interpretation of the superdirective design of beamformers has been given.

5. REFERENCES

[1] J. A. Rex and E. S. J., "An optimal microphone array for speech reception an a car," in *Proc. EURASIP European Signal Proc. Conference (EUSIPCO)*, (Edingburgh,UK), pp. 1752–1755, 1994.

[2] M. C. E. and G. Chollet, "Automatic word recognition in cars," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 346–356, Sep 1995.

[3] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Hands free continuous speech recognition in noisy environment using a four microphone array," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 860–863, 1995.

[4] Q. Lin, C. Che, D. S. Yuk, L. Jin, B. Vries, J. Pearson, and J. Flanagan, "Robust distant-talking speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, vol. 1, pp. 21–24, 1996.

[5] A. Shamsoddini and P. Denbigh, "Enhancement of speech by suppression of interference," in *Int. Conf. on Signal Processing*, (Beijing, China), pp. 753–756, 1996.

[6] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas Propagation*, vol. 30, pp. 27–34, 1982.

[7] J. Bitzer, K. U. Simmer, and K. D. Kammeyer, "Multichannel noise reduction –algorithms and theoretical limits–," in *Proc. EURASIP European Signal Proc. Conference (EUSIPCO)*, vol. 1, (Rhodes, Greece), pp. 105–108, Sep 1998.

[8] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 35, pp. 1365–1375, Oct 1987.

[9] E. Gilbert and S. Morgan, "Optimum design of directive antenna arrays subject to random variations," *Bell System Technical Journal*, pp. 637–663, May 1955.

[10] R. Zelinski, "Mikrofon-arrays mit superdirektiven Eigenschaften zur Sprachsignalverarbeitung," *Frequenz*, vol. 50, pp. 198–204, Sep/Oct 1996. in German.

[11] P. L. Chu, "Superdirective microphone array for a set-top videoconferencing system," in *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.

[12] M. Doerbecker, "Speech enhancement using small microphone arrays with optimized directivity," in *Proc. Int. Workshop on Acoustic Echo and Noise Control*, (London, Great Britain), pp. 100–103, Sep 1997.

[13] A. G. Piersol, "Use of coherence and phase data between two receivers in evaluation of noise environments," *Journal of Sound and Vibration*, vol. 56, no. 2, pp. 215–228, 1978.

[14] B. F. Cron and C. Sherman, "Spatial-correlation functions for various noise models," *Journal of the Acoustical Society of America (JASA)*, vol. 34, pp. 1732–1736, Nov 1962.

[15] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 240–259, May 1998.

[16] D. Mansour and B. H. Juang, "The short-time modified coherence representation and noisy speech recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 6, pp. 795–804, 1989.