ELSEVIER

Speech Communication 34 (2001) 3–12

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Multi-microphone noise reduction techniques as front-end devices for speech recognition ☆

Joerg Bitzer [a,*], Klaus Uwe Simmer [b], Karl-Dirk Kammeyer [c]

[a] *Houpert Digital Audio, Fahrenheitstrasse 1, D-28359 Bremen, Germany*
[b] *Aureca GmbH, Mozartstrasse 26, D-28203 Bremen, Germany*
[c] *Department of Telecommunications, University of Bremen, FB-1, P.O. Box 330 440, D-28334 Bremen, Germany*

Abstract

In this paper, we describe different multi-microphone noise reduction techniques as front-ends for a speaker-independent isolated 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 fail, whereas multi-microphone techniques can raise the recognition rate by using spatial information. We compare the delay-and-sum beamformer, superdirective beamformers, and two post-filter systems. A new adaptive post-filter for superdirective beamformers (APES) is introduced. Our results show that multi-microphone techniques can increase the recognition rate significantly and that the new APES system outperforms related techniques.

Zusammenfassung

In dieser Arbeit werden verschiedene Verfahren zur mehrkanaligen Geräuschreduktion als Eingabegeräte bei einem sprecherunabhängigen Einzelworterkenner vorgestellt. Der Schwerpunkt der Arbeit liegt darin, die Veränderung der Erkennungsleistung zu untersuchen, wenn die Störung durch einen zweiten Sprecher und somit durch eine nichtstationäre und nicht-gaußverteilte Quelle verursacht wird. Für diesen speziellen Fall versagen einkanalige Geräuschreduktionsverfahren, während die Ausnutzung räumlicher Information die Erkennungsrate erhöhen kann. Untersucht wurden dabei nicht-adaptive Verfahren wie der Delay-and-Sum Beamformer, superdirektive Beamformer und adaptive Post-Filter Ansätze. Ein neues Verfahren, das auf einem Post-Filter unter Ausnutzung der besonderen Eigenschaften der superdirektiven Beamformer basiert, wird vorgestellt. Die Ergebnisse zeigen, dass die Ausnutzung räumlicher Information zu einer signifikanten Steigerung der Erkennungsleistung führt und dass der neu entwickelte Algorithmus bessere Ergebnisse liefert als alle anderen untersuchten Verfahren.

Résumé

Dans cette contribution, nous décrivons différentes techniques multi-capteurs de réduction de bruit à la prise de son pour la reconnaissance de mots indépendante du locuteur appliquée à un environnement de bureau. Nous examinons le taux de reconnaissance si la source perturbatrice n'est pas un bruit gaussien et stationnaire, mais un second locuteur

☆ This paper is an extended version, the original paper was entitled "Multi-microphone noise reduction techniques for hands-free speech recognition – a Comparative Study", and it was presented at Robust99.

* Corresponding author.
E-mail addresses: j.bitzer@hda.de (J. Bitzer), uwe.simmer@aureca.com (K.U. Simmer), kammeyer@comm.uni-bremen.de (K.-D. Kammeyer).

0167-6393/01/$ - see front matter
PII: S0167-6393(00)00042-X

présent dans le même local. Dans ce cas, les techniques de réduction de bruit classiques comme la soustraction spectrale sont inefficaces, alors que les méthodes multi-microphones peuvent améliorer le taux de reconnaissance en utilisant l'information spatiale. Nous comparons l'antenne retard-somme, l'antenne super-directive, et deux techniques de post-traitement. Un nouveau post-filtre adaptatif pour les antennes super-directives (APES) est proposé. Nos résultats montrent que les méthodes multi-capteurs peuvent améliorer significativement le taux de reconnaissance, parmi lesquelles la méthode APES se détache en performances.

Keywords: Superdirective beamformer; Multi-microphone noise reduction; Coherence; Speech recognition; Microphone arrays

1. Introduction

An increasing number of PC users work with speech recognition devices in order to control their systems. The input microphone is almost always headset mounted. Those devices are very uncomfortable and restricting. 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 (Rex and Elliot, 1994; Mokbel and Chollet, 1995; Giuliani et al., 1995; Lin et al., 1996). This contribution deals with the problem of a second speaker in the same room. Therefore, the interference signal is colored and non-stationary. Both signals the main speaker and the interferer are assumed to be in the far-field.

For a two-channel system a possible solution is published in (Shamsoddini and Denbigh, 1996). This algorithm is a derivation of a two-channel generalized sidelobe canceller, GSC (Griffiths and Jim, 1982). It can be shown, however, that this kind of algorithm does not perform very well in reverberant environments (Bitzer et al., 1998). In this contribution, we will focus on non-adaptive beamformer and adaptive post-filter solutions. In Section 2, 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. Furthermore, the concept of adaptive post-filters is shortly revised and a new technique is introduced to combine superdirective beamformers with adaptive post-filters. Sections 3 and 4 show the results of the noise reduction experiments.

2. Noise reduction algorithms

A general description of a non-adaptive broadband beamformer in the time-domain 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

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

The design of the filter A_i depends on the given or assumed noise-field and on the selected optimization criterion such as maximum likelihood (ML) or minimum mean square error (MMSE). An often used criterion is based on the minimum variance solution under the constraint of an unmodified and unfiltered look-direction. This solution is called minimum variance distortionless response (MVDR). The design of the MVDR is given by

$$\boldsymbol{A} = \frac{\boldsymbol{\Phi}^{-1}\boldsymbol{d}}{\boldsymbol{d}^{*}\boldsymbol{\Phi}^{-1}\boldsymbol{d}}, \tag{3}$$

where $\boldsymbol{\Phi}$ denotes the $(N \times N)$ power spectral density (PSD) matrix of the noise.

$$\boldsymbol{\Phi} = \begin{pmatrix} \Phi_{X_0X_0} & \Phi_{X_0X_1} & \dots & \Phi_{X_0X_{N-1}} \\ \Phi_{X_1X_0} & \Phi_{X_1X_1} & \dots & \Phi_{X_1X_{N-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{X_{N-1}X_0} & \Phi_{X_{N-1}X_1} & \dots & \Phi_{X_{N-1}X_{N-1}} \end{pmatrix}$$

and $\boldsymbol{d}$ represents the propagation vector between the sensors. For a linear array $\boldsymbol{d}$ is given by

$$\boldsymbol{d} = \left[1, \mathrm{e}^{-\mathrm{j}(\omega/c)d\cos\theta}, \mathrm{e}^{-\mathrm{j}(\omega/c)2d\cos\theta}, \dots, \mathrm{e}^{-\mathrm{j}(\omega/c)(N-1)d\cos\theta}\right]^{\mathrm{T}},$$

where d denotes the distance of adjacent sensors, c the speed of sound, and θ represents the direction of arrival ($\theta = 0$ in the endfire case and $\theta = \pi/2$ for broadside operation).

Eq. (3) can be interpreted as a spatial prewhitening of the noise, a matching of the desired signal, and a division by a normalization term.

For a better understanding of the design procedure our focus in this work lies on a unified description in terms of the complex coherence,

$$\Gamma_{X_0X_1}(\omega) = \frac{\Phi_{X_0X_1}(\omega)}{\sqrt{\Phi_{X_0X_0}(\omega)\Phi_{X_1X_1}(\omega)}}. \tag{4}$$

We assume that the noise is stationary and spatially homogeneous in the room to be examined. The noise has the PSD $\Phi_{NN}(\omega)$. Therefore, we can express the PSDs between the sensors in terms of the noise PSD and the coherence function,

$$\Phi_{X_0X_1}(\omega) = \Phi_{NN}(\omega)\Gamma_{X_0X_1}(\omega). \tag{5}$$

The design equation (3) reduces to

$$\boldsymbol{A} = \frac{\boldsymbol{\Gamma}^{-1}\boldsymbol{d}}{\boldsymbol{d}^{*}\boldsymbol{\Gamma}^{-1}\boldsymbol{d}}, \tag{6}$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1 & \Gamma_{X_0X_1} & \dots & \Gamma_{X_0X_{N-1}} \\ \Gamma_{X_1X_0} & 1 & \dots & \Gamma_{X_1X_{N-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{X_{N-1}X_0} & \Gamma_{X_{N-1}X_1} & \dots & 1 \end{pmatrix}$$

is the coherence matrix. Therefore, the design is independent of the actual noise power, only the spatial characteristic determines the behavior of the beamformer. Now, we can use all theoretically defined noise-fields in order to get optimal beamformers.

2.1. Uncorrelated noise

In the uncorrelated noise-field, the coherence is zero at all frequencies and $\boldsymbol{\Gamma} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix of order N. Computing the coefficients we get $A_i(\omega) = 1/N$ plus a complex phase term, which controls the steering direction of the array. In the broadside design the phase term is 1, whereas in the endfire direction the phase term amounts to $\boldsymbol{d}$. This solution corresponds to the standard delay-and-sum beamformer without shading coefficients, which is the optimal solution for uncorrelated noise.

2.2. Isotropic noise in three dimensions

Another well-defined noise-field is given under the assumption that infinite noise sources arrive from all directions of a hypothetical sphere with an infinite radius. This kind of noise is called diffuse, and it is an approximation of noise-fields in reverberant environments. The coherence between two sensors in a perfectly diffuse noise-field is given by (Piersol, 1978)

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

where d_{ij} denotes the distance between the sensors i and j, and c represents the speed of sound. The result of Eq. (6) yields the standard superdirective beamformer. Unfortunately, with this design procedure low frequencies are amplified, which means that uncorrelated noise will be boosted. Therefore, a constrained design is necessary and can be explained in two ways. Usually, a small scalar is added to the main diagonal of the coherence or the PSD matrix (Gilbert and Morgan, 1955). This leads to a stable design, but the scalar cannot be interpreted directly. Our solution is slightly different, since we want to retain the interpretation as a coherence matrix. The variance of the uncorrelated noise at the sensors, caused by the self-noise of the microphones and the amplifiers, can be included in the coherence function. For example, in

a diffuse noise-field, with an uncorrelated noise with variance σ_n^2, the coherence is

$$\Gamma_{X_iX_j}(\omega) = \begin{cases} \sin\left(\frac{\omega d_{ij}}{c}\right) \Big/ \frac{\omega d_{ij}}{c}\left(1 + \frac{\sigma_n^2}{\Phi_{NN}(\omega)}\right) & \text{for } i \neq j, \\ 1 & \text{for } i = j, \end{cases} \tag{8}$$

where $\Phi_{NN}(\omega)$ is the assumed noise PSD 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 of about −20 to −40 dB.

2.3. Isotropic noise in two dimensions

If we reduce the three dimensions to two dimensions, we obtain a noise-field which is defined by infinite noise sources of a circle with an infinite radius. This kind of noise can arise when a lot of people are speaking in large rooms with well-damped ceilings and floors or in the free-field (cocktail-party noise). The coherence between two sensors is given by (Cron and Sherman, 1962)

$$\Gamma_{X_iX_j}(\omega) = J_0\left(\frac{\omega d_{ij}}{c}\right), \tag{9}$$

where J_0 is the zeroth-order bessel function of the first kind. This leads to the solution of Doerbecker (1997) as an improved superdirective design for speech enhancement. In order to constrain the coefficients, the same technique as in Eq. (8) may be used.

2.4. Optimal beamforming with a priori information

An optimal solution for the design problem can be found, if a priori information is available, e.g. a pre-described direction (θ = angle) of an incoming noise source. Assuming the noise source to be in the far-field of the microphone array, the complex coherence function is given by

$$\Re\{\Gamma_{X_iX_j}(\omega)\} = \cos\left(\frac{\omega \cos(\theta) d_{ij}}{c}\right), \tag{10}$$

$$\Im\{\Gamma_{X_iX_j}(\omega)\} = -\sin\left(\frac{\omega \cos(\theta) d_{ij}}{c}\right). \tag{11}$$

Another source of a priori information can be an actual measurement of the noise-field and its coherence. Both techniques will be used in the following sections to compute the coefficients for the speech recognition experiments.

2.5. Noise reduction with post-filter

In order to increase the noise reduction behavior of arrays some algorithms for reverberant enclosures – the so-called adaptive post-filters – have been developed (Zelinski, 1988; Simmer and Wasiljeff, 1992). A study on this kind of algorithms can be found in (Marro et al., 1998). One typical estimation of the transfer function of the post-filter is

$$\widehat{W}_0 = \frac{1}{\Re\left\{\sum_{i=0}^{N-2}\sum_{j=i+1}^{N-1} A_i A_j^*\right\}} \times \frac{\Re\left\{\sum_{i=0}^{N-2}\sum_{j=i+1}^{N-1} A_i A_j^* \Phi_{X_iX_j}\right\}}{\Phi_{Y_bY_b}}, \tag{12}$$

where the numerator is the real-part of the sum over all cross-power spectral densities (CPSDs) of the filtered input signals. Assuming a coherent speech signal with uncorrelated noise at each microphone, this will give an estimation of the speech signal's PSD. The denominator is the PSD $\Phi_{Y_bY_b}$ of the beamformer output, including speech and noise. Therefore, the transfer function is an optimal Wiener-filter for uncorrelated noise. However, in a diffuse noise-field the noise is not correlated at lower frequencies and, therefore, the noise reduction is low compared to the optimal solution. It can be shown (Marro et al., 1998) that in this structure the noise reduction behavior of this post-filter is directly linked to the noise reduction performance of the non-adaptive beamformer. Unfortunately, $\widehat{W}_0(\omega)$ tends to be unstable if a superdirective design has been used for the computation of the coefficients $A_i(\omega)$.

Our new approach explains the design of post-filter structures, where the properties of superdirective beamformers can be used for further noise

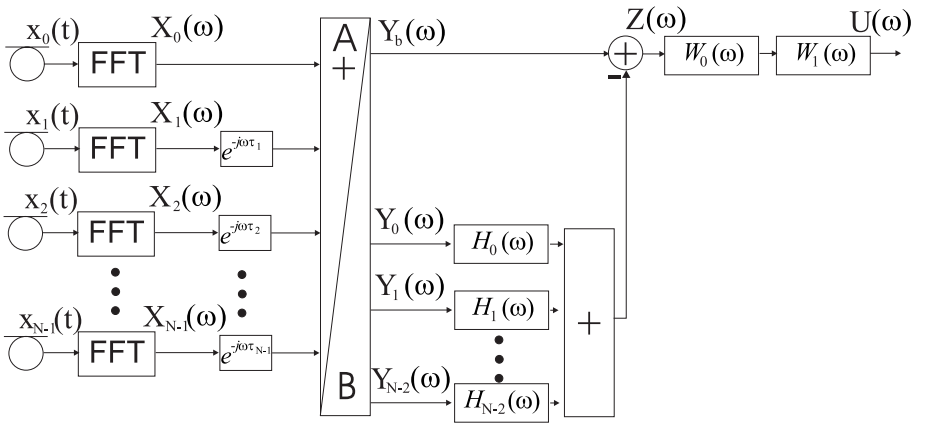

Fig. 1. Block diagram of a GSC-like superdirective and post-filter beamformer in a frequency-domain implementation.

reduction in the post-filter transfer function. The algorithm consists of three parts (see Fig. 1):

- A standard delay-and-sum beamformer with shading coefficients A and a standard post-filter $W_0(\omega)$.
- A superdirective extension (lower path) with a blocking matrix B.
- A second post-filter $W_1(\omega)$.

The post-filter transfer function in the first part $W_0(\omega)$ is estimated according to Eq. (12), and $A_i = 1/N$,

$$\widehat{W}_0 = \frac{2}{N^2 - N} \frac{\Re\left\{ \sum_{i=0}^{N-2} \sum_{j=i+1}^{N-1} \Phi_{X_i X_j} \right\}}{\Phi_{Y_b Y_b}}. \tag{13}$$

The second part of the algorithm is a superdirective extension of the delay-and-sum beamformer. The mathematical background and the design procedure of this new scheme can be found in (Bitzer et al., 1999). [1] The idea can be explained as follows: Cox et al. (1987) have shown that the design of superdirective beamformers and Frost's adaptive algorithm (Frost, 1972) are based on the same optimization criterion. Thus, the Frost algorithm converges to the superdirective beamformer in an isotropic (diffuse) noise-field. Furthermore, Buckley (1986) has shown that the Frost algorithm is equivalent to the GSC (Griffiths and Jim, 1982), if the $(N-1) \times N$ blocking matrix $\boldsymbol{B}$ has the following properties:

- The sum of all values in one row is zero.
- The matrix has to be of rank $N-1$.
- The rows have to be mutually orthogonal.

An example for $N = 4$ is given by

$$\boldsymbol{B} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix}.$$

Therefore, it is possible to implement the superdirective beamformer in a GSC-like structure having fixed filters in the sidelobe path. These filters can be computed in advance by using the Wiener optimization criterion. The new structure has the advantage of reduced complexity, since the fixed filters H_i are real- or imaginary-valued only. Furthermore, the delay-and-sum beamformer output is available without the additional superdirective part.

The second post-filter can be estimated by

$$\widehat{W}_1 = \frac{\Phi_{ZZ}(\omega)}{\Phi_{Y_b Y_b}(\omega)}, \tag{14}$$

where Φ_{ZZ} is the PSD of the output of the superdirective beamformer. This transfer function leads to 1 at higher frequencies, as the superdirective beamformer and the delay-and-sum beamformer perform equally at higher frequencies. On the other hand, at lower frequencies the transfer function tends to zero, since the superdirective

[1] Can be retrieved at http://www.comm.uni-bremen.de/

beamformer suppresses the spatially correlated diffuse noise-field in contrast to the delay-and-sum beamformer. Furthermore, the estimation depends on the SNR: if the SNR is high, the transfer function tends to 1, whereas at low SNRs it is close to zero.

The two post-filter transfer functions can be combined by multiplication, and the result should be restricted to

$$0.05 \leqslant \widehat{W}_0 \cdot \widehat{W}_1 \leqslant 1$$

for a better speech quality.

Therefore, the complete structure depicted in Fig. 1 has three outputs for further extensions: a delay-and-sum beamformer signal, a superdirective beamformer output, and an adaptive broadband noise reduced output.

3. Noise reduction performance

In this section, the different non-adaptive beamformer designs will be examined in terms of their noise reduction behavior for coherent sources. In order to evaluate the performance, the beam-pattern for all designs will be given. The beam-pattern describes the spatial-frequency transfer function of the beamformer. Fig. 2 shows this pattern for the delay-and-sum beamformer (four microphones, distance $d = 5$ cm). We can see that the desired look-direction or main-lobe is distortionless, and that no spatial filtering is reached at frequencies below 800 Hz. Grating-lobes occur above 3 kHz, as the spatial sampling theorem is not fulfilled.

Figs. 3 and 4 depict the beam-patterns for the superdirective design in three dimensions and two dimensions, respectively. The spatial filtering is extended to lower frequencies compared to the delay-and-sum beamformer. The differences between the two designs lead to a broader main-lobe and a better reduction for sources opposite the look-direction ($\theta = 180°$) in the two-dimensional case.

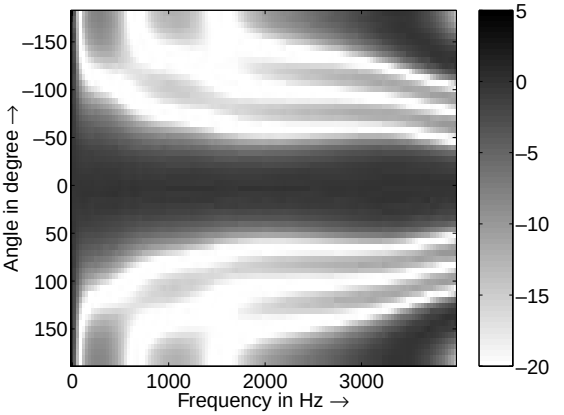

Fig. 3. Beam-pattern for superdirective beamformers (three-dimensional isotropic noise).

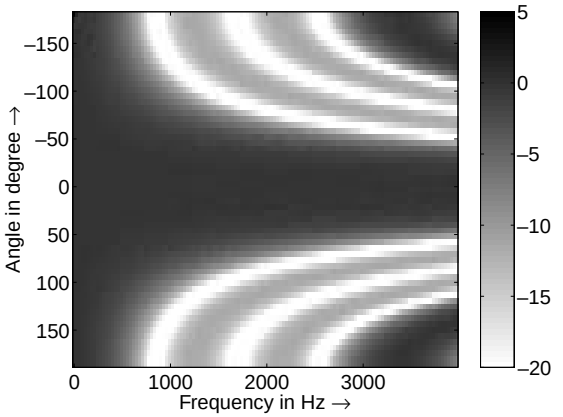

Fig. 2. Beam-pattern for delay-and-sum beamformers ($A_i(\omega) = 1/N$).

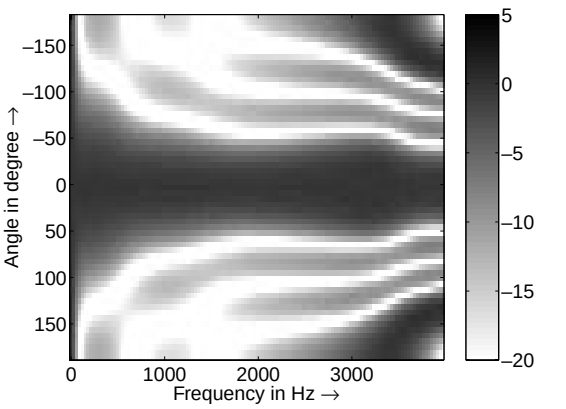

Fig. 4. Beam-pattern for superdirective beamformers (two-dimensional isotropic noise).

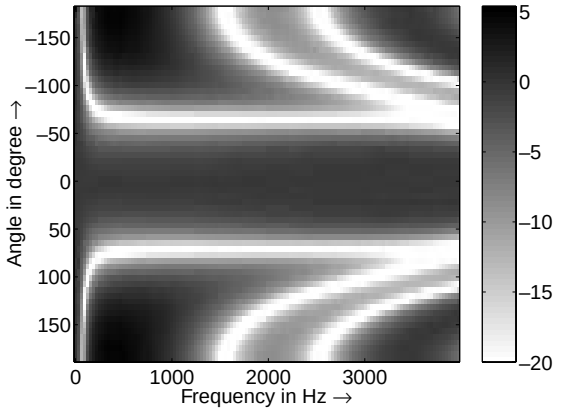

Fig. 5. Beam-pattern for optimized beamformers, where the angle of the interferer is known a priori ($\theta_1 = 66°$).

In order to compute the values in Fig. 5, only one interferer at 66° in the far-field without any reflections is assumed. The design procedure sets a null in this direction. At lower frequencies, however, the constraining factor becomes the determinant parameter. Therefore, the null cannot be designed. In a free-field scenario this design would suppress the interferer completely at frequencies above 500 Hz. On the other hand, an interferer at 150° would be increased by up to 5 dB between 300 and 500 Hz.

The last design example with a priori information is based on an actual measurement of the coherence in a simulated room with reverberation times of $\tau_{60} = 100$ ms and $\tau_{60} = 300$ ms, respectively. The room scenario is depicted in Fig. 8. We can see that the direct path component at 66° is still dominant, but the large number of reflections lead to a more complex beam-pattern. This behavior is strengthened by higher reverberation times (see Figs. 6 and 7).

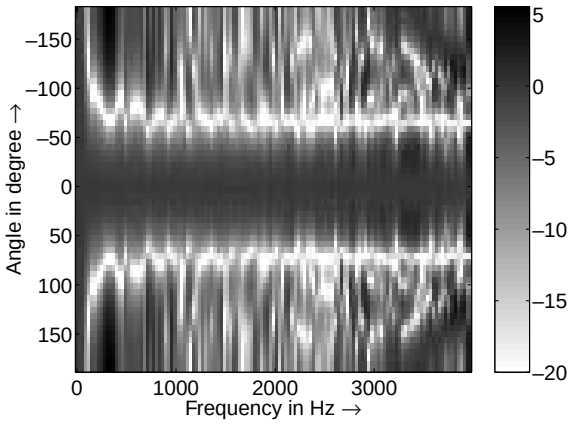

Fig. 6. Beam-pattern for optimized beamformers, where the measured coherence for white Gaussian noise at the position of the interferer is known a priori ($\tau_{60} = 100$ ms).

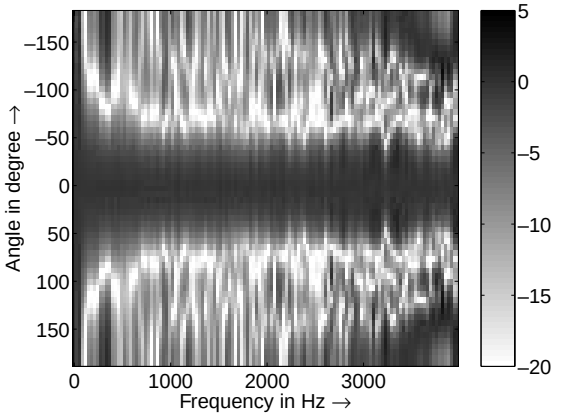

Fig. 7. Beam-pattern for optimized beamformers, where the measured coherence for white Gaussian noise at the position of the interferer is known a priori ($\tau_{60} = 300$ ms).

4. Experiments and results

In order to demonstrate the benefits of arrays with a second disturbing speaker we examined the

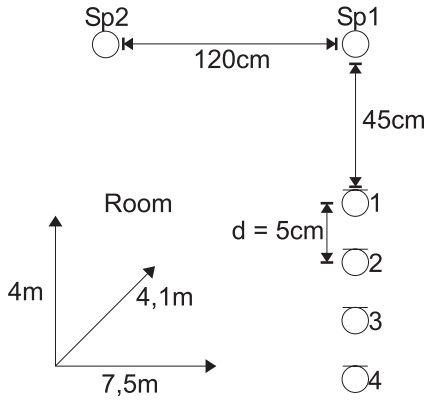

Fig. 8. Configuration for the speech recognition experiment.

following situation. A main speaker is in line with the endfire array of four omnidirectional microphones. The interfering speaker is relatively close to the microphone array at 66° (see Fig. 8). This situation could occur in an office with several people using speech recognition in order to control their computers.

The reverberation is simulated by the image method in the frequency domain in order to obtain fractional delays. The reverberation time τ_{60} is set to 10 ms, 100 ms and 300 ms. Our speech-recognition system is a speaker-independent isolated word recognizer. The feature vector consists either of short-time modified coherence (SMC) coefficients (Mansour and Juang, 1989) of order 12 or cepstral coefficients of the same order. 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 four different speakers (not included in the training phase) (4 utterances and 40 words = 640 test words/experiment), convolved with the room impulse response for the endfire direction. Additionally, the interfering signal was created by convolving the speech signal with the impulse response of the interference direction. The overlapping region between the two words in each test signal is random and it depends only on the speech file. Therefore, there are test files with full overlapping and test files where only a small amount of the desired signal is disturbed by the interferer. In our opinion, this comes closest to a real situation, where there will be either one or two people speaking. The voice activity detection (VAD) is assumed to be perfect (hand-labeled) for the desired speaker (see Fig. 9(a)). However, a real VAD for this situation is still an open topic. Fig. 9(c) shows an example of a mixed signal. The recognition rate of our system reaches 96% for the cepstral coefficients and 93% for SMC in the undisturbed case.

We tested five different algorithms as noise reduction front-ends:

1. Delay-and-sum beamformer (D&S, optimal for uncorrelated noise).
2. Standard superdirective beamformer (SB, optimal for diffuse noise).

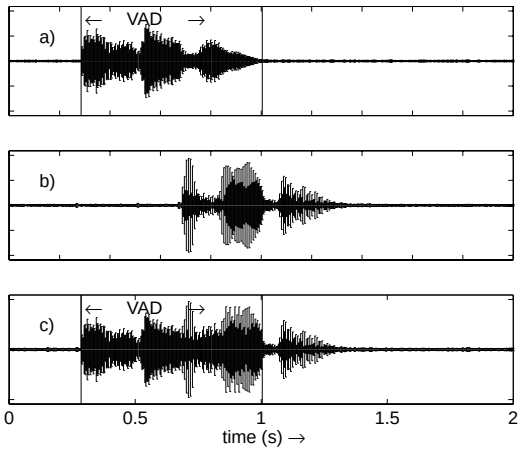

Fig. 9. Example for a speech signal: (a) original; (b) interference; (c) sum of desired and interfering speaker.

3. Superdirective beamformer with a priori information (SB Opt, optimized for this situation).
4. D&S in conjunction with an adaptive post-filter (APF).
5. Adaptive post-filter for superdirective beamformers (APES).

In a first experiment, we compared the recognition rates for the two sets of feature vectors. Figs. 10(a) and 10(b) depicts the results of the experiment in the endfire case, based on cepstral coefficients and SMC respectively. In order to compare the results, we included the recognition rates in the single channel case. The recognition rate was dominated by the position of the single microphone. The distance of only 15 cm decreases the rate by 6–8% between the first and the fourth microphone. Therefore, as a first result, we can say that the microphone should be put as close to the mouth as possible, which is not very surprising. If we compare the results of SMC (Fig. 10(b)) to the cepstral coefficients (Fig. 10(a)), we can see that the cepstral coefficients are more robust subject to the interfering speaker. Furthermore, the SMC coefficients appear to be sensitive against modifications of the noise spectrum. Especially the results of the D&S and of the APF, when using the SMC, are very interesting, as there is no improvement at all. Both algorithms are unable to suppress low-frequency noise, which results in a strong low-pass effect subject to the interfering signal.

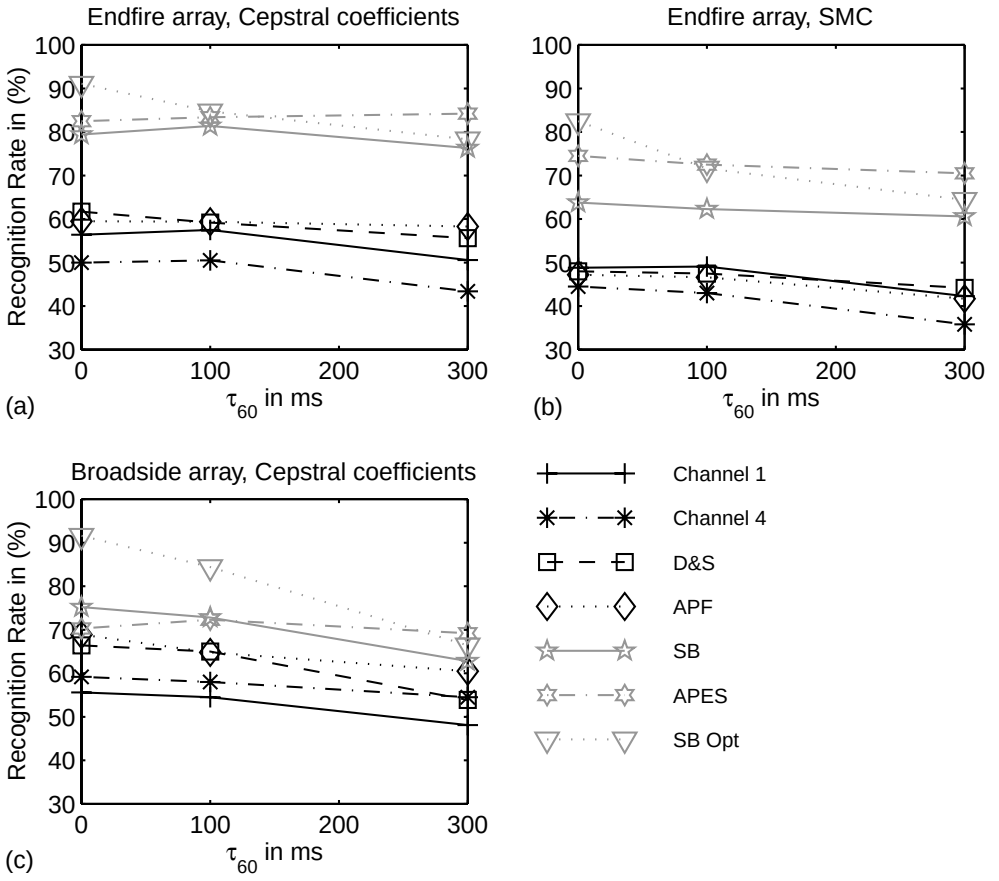

Fig. 10. Recognition rates versus reverberation time: (a) endfire array, cepstral coefficients; (b) endfire array, SMC coefficients; (c) broadside array, cepstral coefficients.

On the other hand, the superdirective beamformers (SB, SB Opt, and APES) perform very well. Especially if a priori knowledge is available, a recognition rate close to the optimum can be achieved for small reverberation times. Usually, this information is not available and the reverberation time in office rooms is larger than 200 ms. The only two algorithms which can handle all cases are the superdirective beamformer and our new adaptive algorithm APES. APES increases the recognition rate up to 8% compared to all other algorithms. However, this requires a more complex algorithm.

In a second experiment, we rotated the array by 90°, in order to get a broadside array aperture. Since the interferer is in the far-field we can compare these results to the endfire steering direction if the reverberation time is large enough. In this case, the dominant part of the interference are the reflections and not the direct path of the interferer. Hence, the noise-field is more diffuse than coherent. On the other hand, if the reverberation time is very small the rotation of the array aperture will lead to a different geometry and, therefore, the direct comparison is not valid.

Fig. 10(c) shows the recognition rate for the cepstral coefficients. The results of the SMCs are worse, though basically similar. Thus, they are not shown here. The results of the single channel cases can be explained, if we take into account that the fourth channel is further apart from the interfering speaker and, therefore, the SNR of this channel is much better than that of the first sensor. The deterioration of the superdirective beamformer and

APES can easily be explained, if we take into account that significant superdirectivity and the inherent better performance at lower frequencies are only possible in the endfire steering case. This behavior is independent of the reverberation time, and it is a feature of these algorithms. Therefore, the endfire steering direction is the best choice for multi-microphone noise reduction techniques, since all other steering directions will worsen performances.

5. Conclusion

In this contribution, we have shown that superdirective beamformers and adaptive post-filter techniques are a good choice in order 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, special caution has to be exercised in the design of office speech-recognition systems. Additionally, a new algorithm has been proposed, which outperforms the related algorithms in terms of the speech recognition rate. Our results show that the steering direction of the array and the feature vector are important parameters in order to increase recognition rates when using multi-microphone noise reduction devices.

References

Bitzer, J., Simmer, K.U., Kammeyer, K.D., 1998. Multichannel noise reduction – algorithms and theoretical limits. In: Proceedings of the EURASIP European Signal Processing Conference (EUSIPCO), Vol. 1, Rhodes, Greece, pp. 105–108.

Bitzer, J., Simmer, K.U., Kammeyer, K.D., 1999. An alternative implementation of the superdirective beamformer. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, pp. 7–10.

Buckley, K.M., 1986. Broad-band beamforming and the generalized sidelobe canceller. IEEE Trans. Acoust. Speech Signal Process. 34 (5), 1322–1323.

Cox, H., Zeskind, R.M., Owen, M.M., 1987. Robust adaptive beamforming. IEEE Trans. Acoust. Speech Signal Process. 35 (10), 1365–1375.

Cron, B.F., Sherman, C., 1962. Spatial-correlation functions for various noise models. J. Acoust. Soc. Am. (JASA) 34 (11), 1732–1736.

Doerbecker, M., 1997. Speech enhancement using small microphone arrays with optimized directivity. In: Proceedings of the International Workshop on Acoustic Echo and Noise Control, London, UK, pp. 100–103.

Frost, O.L., 1972. An algorithm for linearly constrained adaptive array processing. Proc. IEEE 60 (8), 926–935.

Gilbert, E., Morgan, S., 1955. Optimum design of directive antenna arrays subject to random variations. Bell Syst. Tech. J., 637–663.

Giuliani, D., Matassoni, M., Omologo, M., Svaizer, P., 1995. Hands free continuous speech recognition in noisy environment using a four microphone array. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 860–863.

Griffiths, L.J., Jim, C.W., 1982. An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag. 30, 27–34.

Lin, Q., Che, C., Yuk, D.S., Jin, L., Vries, B., Pearson, J., Flanagan, J., 1996. Robust distant-talking speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 1, pp. 21–24.

Mansour, D., Juang, B.H., 1989. The short-time modified coherence representation and noisy speech recognition. IEEE Trans. Acoust. Speech Signal Process. 37 (6), 795–804.

Marro, C., Mahieux, Y., Simmer, K.U., 1998. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. IEEE Trans. Speech Audio Process. 6 (3), 240–259.

Mokbel, C.E., Chollet, G.F.A., 1995. Automatic word recognition in cars. IEEE Trans. Speech Audio Process. 3 (5), 346–356.

Piersol, A.G., 1978. Use of coherence and phase data between two receivers in evaluation of noise environments. J. Sound Vibration 56 (2), 215–228.

Rex, J.A., Elliot, S.J., 1994. An optimal microphone array for speech reception in a car. In: Proceedings of the EURASIP European Signal Processing Conference (EUSIPCO), Edingburgh, UK, pp. 1752–1755.

Shamsoddini, A., Denbigh, P., 1996. Enhancement of speech by suppression of interference. In: International Conference on Signal Processing, Beijing, People's Republic of China, pp. 753–756.

Simmer, K.U., Wasiljeff, A., 1992. Adaptive microphone arrays for noise suppression in the frequency domain. In: Second Cost 229th Workshop on adaptive Algorithms in Communications, Bordeaux, France, pp. 185–194.

Zelinski, R., 1988. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New York, USA, pp. 2578–2581.