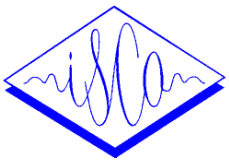

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

SPEECH ENHANCEMENT USING TWO VERSIONS OF THE NOISY SPEECH SIGNAL

Klaus Linhard
e-mail: linhard@dbag.ulm.DaimlerBenz.COM
Daimler Benz AG, Research and Technology
Wilhelm-Runge-Str. 11
D-89081 Ulm
Germany

ABSTRACT

In this paper we describe a two-channel speech enhancement system. Two noisy versions of a speech signal are picked up with closely spaced microphones. The system systematically exploits temporal and spatial characteristics of the sound field and fully operates in the frequency domain. The system is based on the spectral subtraction technique and adaptive beamforming. It may be used to improve speech quality for hands-free telephone as well as to preprocess speech for speech operated systems.

1. INTRODUCTION

There are many approaches for one-channel and multichannel noise reduction systems. A two-channel system offers the potential of both signal averaging and spatial resolution of the sound field and will need only a moderate amount of hardware and computation time for real time applications. - Man is a good example for a two-channel speech enhancement system.

Several noise reduction systems using two noisy speech channels have been suggested, for example [1] and [2]. The first paper describes a two-channel system which reduces reverberation (diffuse noise). The phase of one channel is modified using a short-term all-pass filter, the channels are added and diffuse noise is suppressed in a postprocessing stage. In the second paper a simple linear phase correction is used and noise is also suppressed in a postprocessing stage.

2. SYSTEM OVERVIEW

The system may be divided into five stages: two-channel spectral subtraction with smoothing and correction of the subtraction filter, time delay estimation, adaptive beamforming, post-processing and frequency response correction. Spectral subtraction itself is a widely used technique to reduce slowly time variing noise. Time delay estimation is used to detect the direction of the signal source, the speaker, and to allow phase matched addition of the two channels. Once the direction of the speaker is known, an adaptive beamfomer adjusts its directivity pattern to get the desired speech signal, and, at the same time suppress stationary or unstationary signals coming from other directions. Diffuse noise is additionally reduced in the postprocessing stage. Optionally a fifth stage may be used for frequency correction to combat a loss at low frequencies depending on the performance of the beamformer. The basic structure of the system is shown in Fig.1. The following sections give a description of the single stages.

3. SYSTEM DESCRIPTION

3.1. Two-Channel Spectral Subtraction

Spectral subtraction, in its simplest form, is multiplying the real filter coefficients H_{SPS} with the Fourier transformed input segments X. The time segments of the input signal x are usually half overlapped and weighted with a Hanning window. At the system output after the inverse Fourier transform the time signal is

reconstructed with the overlapp add technique. It can been shown that filtering is equivalent to the spectral subtraction of the estimated mean spectral noise magnitude $N(i) = \sqrt{E[|X(i)|^2]}$:

$$Y(i) = H_{SPS}(i)\, X(i) \ , \qquad (1)$$

$$H_{SPS}(i) = 1 - \sqrt{\frac{E[|X(i)|^2]}{|X(i)|^2}} = 1 - \frac{N(i)}{|X(i)|} \ . \qquad (2)$$

i denotes the discrete frequency index. The phase of the input signal is not processed. $N^2(i)$ is usually approximated by recursive time averaging in speech pauses. Speech pauses are detected with an energy criteria and some additional speech features like pitch and stationarity.

Spectral subtraction is a block-processing frequency domain filtering technique, thus, actual spectral differences, $X(i) - N(i)$, appear and disappear with the block processing rate. The differences are heard as musical tones. This artifacts are partially masked using a lower bound b on H_{SPS}, $H_{SPS} > b$, with $0 < b < .3$. Often a lot of musical tones are left. It is therefore usual to subtract an overestimated noise $a\ N(i)$ $(1 < a < 4)$, but $a > 1$ increases speech distortion.

The two-channel spectral subtraction consits of two single subtractions on each input channel. The addition of the channels in a subsequent stage yields averaged musical tones and because musical tones in each channel are at least partially uncorrelated they are reduced. We further examine the filter coefficients in each channel and suppress peaks (musical tones) appearing in one channel only. We do not need an overestimated noise and thus avoid speech distortion.

3.2. Time Delay Estimation

Fig.2 shows the simplified structure of the time delay estimation stage. The two signals after spectral subtraction are the inputs of this stage. A predetermined number of maxima values (MAX) of the short-term cross power spectral density function (KKF) are used to estimate the linear phase difference of the two channels. Pre-emphasis (PRE) is used to account for decreasing higher frequency components of the speech signal. The maxima values of the cross power spectra are monitored to destinguish between background noise and the beginning of speech (impulse). Only the maxima values labelled as beginning of speech contribute to the calculation of the short-term averaged phase rise $\overline{PHI'}$. If the speaker is only allowed to move in a predetermined sector signal information coming from outside this sector is considered as noise and not averaged to $\overline{PHI'}$. All-pass function $H_{ALL}(i)$ is used for phase shifting of one channel,

$$H_{ALL}(i) = \cos(i\ \overline{PHI'}) + j\ \sin(i\ \overline{PHI'}) \ . \qquad (3)$$

3.3. Adaptive Beamformer

Adaptive beamforming may be performed with several algorithms, for example [4], [5] and [6]. Generally a constraint on the allowed distortion on the desired speech signal is given and the filter is adjusted adaptively to minimize the output power subject to this constraint. [5] uses a constraint on the magnitude of the transfer function of the speech signal in the desired direction. [5] has the potential of high low frequency noise suppression with close microphone spacing, but has the disadvantage that a frequency response correction filter H_{INV} at the system output is neccessary. The magnitude distortion with algorithms [4] and [6] has the value zero. [6] is the so called generalized sidelobe canceller which we implemented in the frequency domain. In this case only one filter H_R is needed instead of two filters H_{R1} and H_{R2} of the other mentioned algorithms. The input of H_R is the difference of the two channels (reference input). The primary input is the sum of the two channels. The adaptive adjustment is performed with a standard frequency domain LMS-algorithm (least mean squares). Adaptation is only allowed in speech pauses. The speech-pause detector already available for spectral subtraction may be used.

3.4. Postprocessing

Postprocessing reduces diffuse noise. A modified short-term coherence function is calculated ([1]),

$$H_{KKF}(i) = \frac{|\, S_{xy}(i)\,|}{S_{xx}(i) + S_{yy}(i)} \quad . \qquad (4)$$

$S_{xx}(i)$ and $S_{yy}(i)$ denote the estimated power spectral desity of the preprocessed Signals x and y and $S_{xy}(i)$ denotes the estimated cross power spectral density of x and y.

3.5 Frequency Correction

This frequency correction is only neccessary for a beamformer with a magnitude constraint. The filter has a frequency response H_{INV} which is the inverse of the adapted magnitude response in the desired direction.

4. RESULTS

Fig.3 shows a comparison of one-channel and two-channel spectral subtraction in the situation of broadband stationary noise. The filter coefficients of spectral subtraction are shown in a speech pause (noise only). First with $b = .2$ and no overestimation ($a = 1$) many coefficient peaks may be seen which correspond to musical tones. There are always two sets of filter coefficients displayed overlayed at a time distance of 10ms to show its random (musical) character. With $a = 1.2$ musical tones are reduced but they are still present. The next examples show the coefficients of a two-channel system. No overestimation is used. If we suppress peaks (musical tones) appearing in one channel only, most of the peaks disappear. If we additionally suppress occasionally appearing isolated peaks no more peaks are left. This way we do not need an overestimated noise and thus avoid speech distortion.

Fig.4 shows an example of the spatial characteristics of the adapted directivity pattern of a generalized sidelobe canceller. Broadband noise from an open car window caused a high noise suppression in the direction of this noise. With this beamformer included into the two-channel system an additional gain of 5dB of noise reduction was achived.

5. CONCLUSIONS

We use a two-channel noise suppression system yielding good speech quality and yet need only a moderate amount of computation. The presented two-channel system may be divided into several stages each exploiting temporal or spatial characteristics of the sound field. The distance of the two microphones was only several cm (ex. 10cm). Such a small distance offers good possibilities for placing of the microphones. The system is well suited for speech enhancement applications in moderate or high noise environments. Serveral subjektive listening tests confirmed the high quality of the noise reduced speech.

REFERENCES

[1] Allen, J.B.; Berkley D.A.; und Blauert, J.:
Multimicrophone signal-processing technique to remove room reverberation from speech signals
J.Acoust.Soc.Am., Vol.62, No.4, p.912-915, 1977

[2] Kaneda, Y; Tohyama, M.:
Noise Suppression Signal Processing Using 2-Point Received Signals
Electronics and Communication in Japan, Vol.67-A, No.12, p.19-28, 1984

[3] Boll, S.F.:
Suppression of Acoustic Noise in Speech Using Spectral Subtraction
IEEE Trans.Acoust., Speech, Signal Processing, Vol. ASSP-27, No.2, p.113-120, 1979

[4] Frost, O.L.:
An Algorithm for Linearly Constrained Adaptive Array Processing
Proc. IEEE, Vol.60., No.8, p.926-935, 1972

[5] Sondhi, M.M.; Elko, G.W.:
Adaptive Optimization of Microphone Arrays under a Nonlinear Contraint
Int. Conf. on ASSP, Tokyo, p.981-984, 1986

[6] Griffiths, L.J.; Jim, C.W.:
An Alternativ Approach to Linearly Constrained Adaptive Beamforming
IEEE Trans.Ant.a.Prop., Vol.Ap-30, No.1, p.27-34, 1982

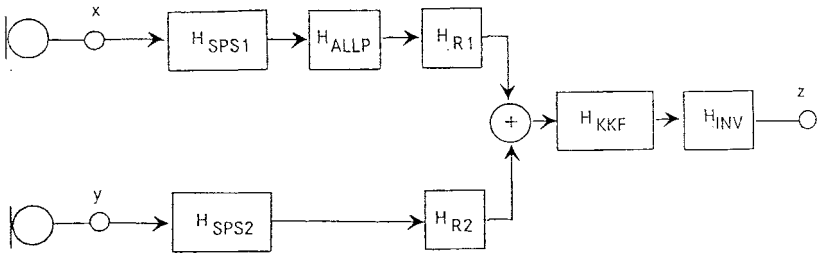

Fig.1.
Two-channel noise reduction system

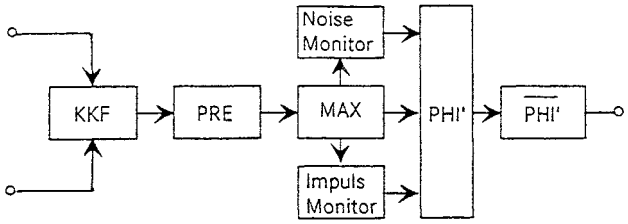

Fig. 2.
Time delay estimation

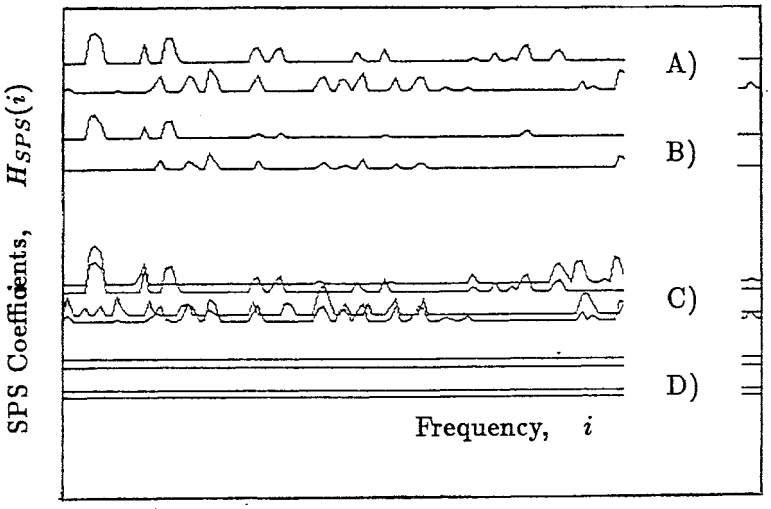

Fig. 3.
Example of spectral subtraction coefficients in a speech pause. Two sets of coefficients with time distance 10ms are overlayed.
A) one-channel SPS with $a = 1$;
B) one-channel SPS, $a = 1.2$;
C) two-channel SPS, $a = 1$;
D) two-channel SPS, $a = 1$, coeff. corr.

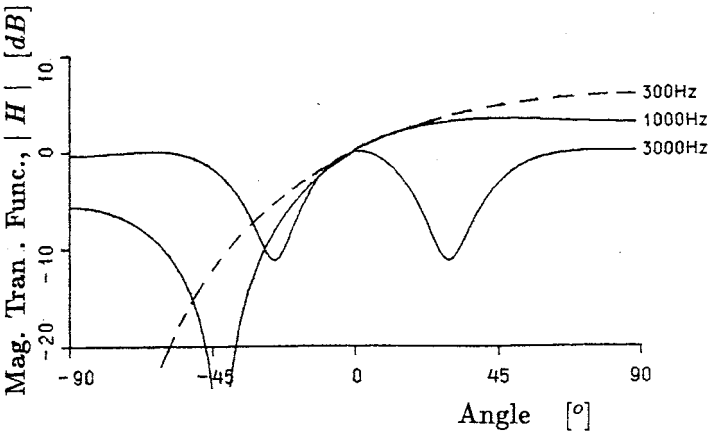

Fig. 4.
Example of a directivity pattern (high noise coming from left, 0° is the desired speech direction)