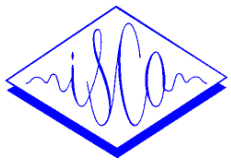

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

SUBBAND ARRAY PROCESSING FOR SPEECH ENHANCEMENT

Kristian Kroschel and Keld Lange

Institut fuer Nachrichtensysteme, Universitaet Karlsruhe
Kaiserstrasse 12, 76128 Karlsruhe, Germany

10.21437/Eurospeech.1993-151

ABSTRACT

Classical array processing systems for speech enhancement include three components. The first one is used for delay compensation of the speech signal in the different microphone channels, the second component is based on spectral subtraction or Wiener filtering to enhance the signal–to–noise–ratio, and the third component has to compensate residual noise components like musical tones. In this paper a new system based on this principal approach is presented. Instead of pure delay compensation an equalizer is used which compensates the delay of the speech signals and the differences of the transfer function of the different microphone channels. Depending on their position the microphones are related to subsections of the speech spectrum to avoid a dynamic delay compensation caused by the movement of the head of the speaker. A third improvement over the classical approach is given by the fact that instead of a classical FFT algorithm the Short–Time Fourier Transform (STFT) proposed by Portnoff is used which has been implemented using the FFT. Since the speech signal and the noise are instationary processes this transform is favourable. With this configuration the post processing is a simple addition of the partial results because the musical tones have been significantly removed by the other components of the system. It has been shown that the method presented in this paper can be realized in real time using an Intel i860 processor.

Keywords: *Array processing, noise reduction, speech enhancement, Short–Time Fourier Transform.*

1. INTRODUCTION

If speech communication systems with a handsetfree terminal are operated in a noisy environment so that the listener at the far end has difficulties to follow the conversation, noise reduction systems have to be implemented. Such a system may use a microphone array at the front end. The microphones have to be positioned in such a way that the correlation between the speech signals at their output is maximum and the correlation of the noise is minimum. Of course, only a compromise can be found for these contradictive conditions.

The structure for noise reduction systems using microphone arrays consists in principle of three components which are shown in Fig. 1. The first component with the input vector $\mathbf{r}$ consisting of M speech signals picked up from the microphones is used for delay compensation. For synchronisation purposes the delay of these signals is estimated and compensated. The vector $\mathbf{x}$ of the delay–compensated signals is fed into the second component for noise reduction using classical techniques of estimation theory modified by heuristic approaches such as spectral subtraction. This component is based on a modified Wiener filter. The third component with the input vector $\mathbf{y}$ is used for post processing to get rid of residual noise tones. Finally an estimate s of the speech signal is generated. Systems of this kind have been described in literature [1,2,3]. Most of these systems are realized in the frequency domain to simplify the transformation of the signals by the different components, especially if spectral subtraction is used for noise reduction. In this case the input signals are sampled and transformed into the frequency domain and at the output the inverse operation is executed.

Fig.1 Structure of the array–based noise reduction system

In this paper improvements of the system described in more detail in [4] are presented. This system which yields a higher degree of signal enhancement than noise suppression or noise compensation [5] lacks from two major disadvantages: First, the algorithm for delay

compensation is complicated and therefore leads to high hardware cost and second, the speech quality processed at the output is not too satisfactory due to a high content of residual musical tones. In the following improvements to overcome these problems concerning the front end of the system and the transformation in the frequency domain are presented. Further details can be taken from [6].

2. THE MICROPHONE ARRAY

Because no specified application of the array was assumed, 8 microphones were arranged in two square patterns as shown in Fig.2.

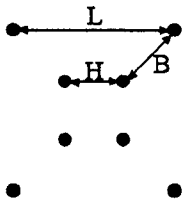

Fig.2 Geometry of the microphone array for low (L), intermediate (B) and high (H) frequency bands

The distance of the outer microphones is 30 cm, the inner microphones are 6 cm apart from each other, so that the closest distance of two microphones from the inner and the outer square measures 17 cm. This arrangement was taken to subdivide the whole frequency spectrum into three subbands with the cutoff frequencies 500 Hz and 1200 Hz. The low frequency components were taken from the paired microphones at the outer square (L), the intermediate frequency components from pairs of the diagonal (B), and the high frequency components from the inner square (H). The idea behind this division of the frequency band is as follows: The difference between the travel time of the speech signal to a pair of microphones divided by the appropriate wavelength is approximately the same in each frequency band assigned to the microphone pairs L, B, and H. If the mouth of the speaker moves in the range of a few centimeters with respect to the microphone array this difference can be ignored so that the dynamic delay compensation becomes obsolete which reduces the computation power requirements of the system significantly.

Having filtered the microphone signals in the appropriate lowpass, bandpass, and highpass filters, the output components are synthesized (S) to complete the frequency band, and the sum and the difference of each synthesized signal is calculated as shown in Fig.3. For p=8 microphones q=4 of these filter blocks are realized. The sums and differences of the outputs are used in the noise reduction component.

If the speech signals picked up from the paired micro

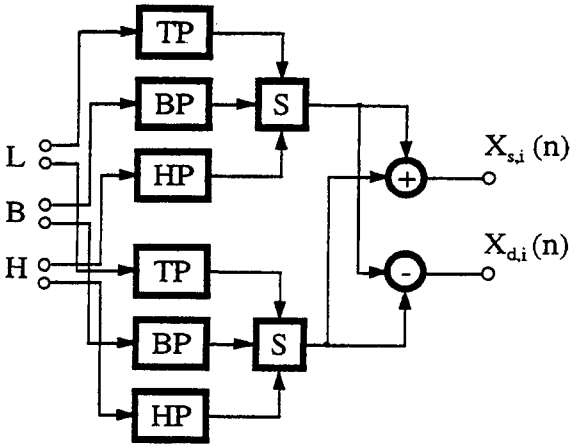

Fig.3 Filter block of a subarray

phones are in perfect synchronisation, i.e. are not delayed with respect to each other, their sum will yield a 3 dB signal enhancement, and the difference containes pure noise. For these desired results a static delay compensation is required. Furthermore the accoustic paths between the speakers mouth and the microphones, the microphones themselves and the filter channels may have different transfer functions so that furthermore these transfer functions have to be equalized. For this purpose a broadband test signal generated at the position of the speakers mouth is applied to the array to measure the unknown transfer function between the mouth of the speaker and the output of each subarray filter block.

3. NOISE REDUCTION

The component for noise reduction is based on spectral subtraction and consists of q=4 channels. The transfer function of each of these channels is given by

$$H_i(n) = \begin{cases} 1 - a \cdot S_{nn,i}(n)/S_{rr,i}(n) \\ b \end{cases} \qquad 1 \leq i \leq 4 \qquad (1)$$

with n the frequency index, $S_{rr,i}(n)$ and $S_{nn,i}(n)$ the estimates of the power spectra of the corrupted speech signal and the noise, respectively. The parameter a is the overestimate factor in the range $1 \leq a \leq 4$ and b is the spectral floor, e.g. b=0,1. The upper relation in (1) is valid if the given expression is greater than b, otherwise b is taken. The nonlinearity of the transfer function is caused by the fact that the estimates of the power spectra might be incorrect so that their weighted difference becomes negative which would contradict the definition of a power spectrum. Modifications of the transfer functions $H_i(n)$ use the square root of

the whole expression or the square root of the ratio of the power spectra. The estimates of the power spectra are functions of the spectra $X_{s,i}(n)$ and $X_{d,i}(n)$, e.g.:

$$S_{rr,i}(n) = |X_{s,i}(n)|^2 \quad (2)$$

$$S_{nn,i}(n) = |X_{d,i}(n)|^2 \quad . \quad (3)$$

The transfer function of the sum channel and the difference channel is not identical so that the estimate of the noise power spectrum (3) might be modified [4]: In speech pauses the transfer function $H_{ds,i}(n)$ between the difference and the sum channel is calculated and during speech activity the noise power spectrum is given by

$$S_{nn,i}(n) = |H_{ds,i}(n)|^2 \, |X_{d,i}(n)|^2 \quad . \quad (4)$$

To improve the estimates of the power spectra $S_{rr,i}(n)$ and $S_{nn,i}(n)$, they are exponentially smoothed using a time constant which depends on the correlation time of the speech and noise processes, respectively.
Informal listening tests have shown that the transform according to (4) does not improve the enhanced speech signal significantly so that only the averaging by an update procedure is recommended. Thus the speech pause detector and the calculation of the transfer functions $H_{ds,i}(n)$ can be omitted. Two reasons can be given why the approach (4) does not improve the quality of the speech over the approach (3): First, the equalization of the input channels reduces the deviation between the sum and difference channels and second, the standard FFT is replaced by the short time Fourier transform (STFT) proposed by Portnoff [7].

4. FREQUENCY TRANSFORMATION

Speech is in principle an unstationary process but within an interval of 10 to 30 ms it is assumed to be quasi–stationary. Therefore usually the fast Fourier transform (FFT) is taken for frequency transformation. With a sampling frequency of f_S=8 kHz a block length of N=128 or N=256 samples corresponds to the given interval of stationarity. Since the assumption of stationarity is critical, the FFT has been replaced by the STFT with the definition

$$X(n,l) = \sum_{k=-\infty}^{\infty} x(k)\, w(l-k)\, e^{-j2\pi kn/N} \qquad 0 \le n \le N-1 \quad . \quad (5)$$

In this definition n describes the frequency parameter, l is the actual time parameter, and w(k) is a window function, e.g. the Hamming window. At each time instant l the STFT yields N spectral lines. In principle the length of the window w(k) may be chosen arbitrarily but since the STFT is implemented using the FFT to reduce the calculation load, the window w(k) consists of N values, too:

$$w(k) \begin{cases} \neq 0 & -N/2 \le k \le N/2-1 \\ = 0 & \text{elsewhere} \end{cases} \quad . \quad (6)$$

The inverse STFT or ISTFT is given by

$$x(l) = \frac{1}{N} \sum_{n=0}^{N-1} X(n,l)\, e^{j2\pi nl/N} \quad . \quad (7)$$

To guarantee that, if the STFT and the ISTFT are applied to a signal x(k) one behind the other, x(l) remains unchanged, the window w(k) has to have the properties

$$w(0) = 1 \quad ; \quad w(iN) = 0 \,,\ i \neq 0,\ i \in \mathbb{Z} \quad (8)$$

which are fulfilled by all standard windows.
The STFT given in (5) can be interpreted as the convolution of x(l) weighted by exp (–j2πln/N) with the window w(l):

$$X(n,l) = \left[x(l) \cdot e^{-j2\pi ln/N} \right] * w(l) \quad . \quad (9)$$

Therefore X(n,l) will be bandlimited by the bandwidth of w(l). If w(k) is a Hamming window with N samples unequal zero the normalized bandwidth is

$$\Omega_{max,Ham} = 4\pi/N \quad . \quad (10)$$

On the other hand the sampling theorem requires that the highest normalized frequency component is equal to

$$\Omega_{max} = \pi \quad (11)$$

so that the sampling rate of the STFT X(n,l) can be reduced by a factor

$$R = \frac{\Omega_{max}}{\Omega_{max,Ham}} = \frac{N}{4} \quad (12)$$

which reduces the computation load of the STFT significantly. The reconstruction of the signal at the output of the system requires an interpolation of the samples before the ISTFT is applied. Without this reduction for each sampled value x(l) a set of N frequency values would have been calculated whereas with the FFT for a block of N values x(l) N frequency values are determined. Thus the factor of N comparing the calculation load of the STFT over the FFT is reduced to 4.
It has been shown [8] that the array system with 8 channels based on the STFT which has been realized using the FFT can be implemented for real–time

operation at f_S=8 kHz sampling frequency of the output signals on a signal processing board with the Intel i860 processor as the core unit.

5. RESULTS

It is always a difficulty to evaluate a system with an output which is applied to a human. He always will use a subjective measure and no objective figure of merit for evaluation which complicates the optimization of such a system. Therefore enhancement of the signal to noise ratio is not a very satisfying figure of merit but it is easily calculated and therefore it has been used for evaluation. It is defined by

$$\mathrm{SNRE} = \frac{\mathrm{SNR}_{\mathrm{out}}}{\mathrm{SNR}_{\mathrm{in}}} = \frac{\mathrm{E}\{s^2_{\mathrm{out}}(k)\} \cdot \mathrm{E}\{n^2_{\mathrm{in}}(k)\}}{\mathrm{E}\{n^2_{\mathrm{out}}(k)\} \cdot \mathrm{E}\{s^2_{\mathrm{in}}(k)\}} \quad (13)$$

where $s_{\mathrm{out}}(k)$=s is the estimated speech signal s at the output of the block diagram given in Fig. 1. The environmental noise is named $n_{\mathrm{in}}(k)$ whereas $n_{\mathrm{out}}(k)$ is the residual noise including musical tones measured at the output of the system.

The tests have been executed in a laboratory environment with the microphones set up as given in Fig. 2. A loudspeaker was used as the noise source for traffic noise in a car, furthermore printers and a vacuum cleaner represented noise sources typical for office environment and a workshop. Two speakers, a female and a male, spoke prepared sentences in a distance of about 50 cm from the microphone array. The input signal to noise ratio $\mathrm{SNR}_{\mathrm{in}}$ was in the range of 2.0 to 8.2 dB and the output signal to noise ratio $\mathrm{SNR}_{\mathrm{out}}$ ranged from 10.4 to 16.3 dB, the minimum enhancement was 8.1 and the maximum was 13.8 dB.

These figures are not too representative because only a few tests have been executed. Compared to other results reported in literature these figures are nevertheless very promising.

It was further noticed that the movement of the speaker does not influence the result significantly. Another effect was observed during the tests: If the input signal to noise ratio falls below a specific level the noise reduction does not operate satisfactorily any more because the distortion of the processed speech is not acceptable even if there is a high value of the SNRE. This again demonstrates that the SNRE as a figure of merit is not too satisfactory.

6. CONCLUSION

A laboratory version of an array based noise reduction system has been presented which can be implemented for real time operation. There is no question that the system needs further improvement and investigation. It is not too user friendly that a setup procedere for the adaptation of the equalizers to a new speaker has to be executed. Furthermore it is open how much the different microphones contribute to the noise reduction individually and how they should be positioned in a specified environment like a car or an office. But it is no question that systems for noise reduction of this structure will be of importance in the future because the market for handset free communication systems like the mobile phone is growing dramatically with the introduction of digital services.

ACKNOWLEDGEMENT

The investigations published in this paper have been supported by the Deutsche Bundespost TELEKOM under contract No. FI—St—3 B 1392/5.

REFERENCES

[1] *Zelinski, R.*: A Microphone Array with Adaptive Post—Filtering for Noise Reduction in Reverberant Rooms. Proc. Int. Conf. on ASSP, pp. 2578—2581, ICASSP 1988

[2] *Kroschel, K.*: Enhancement of Speech Signals Using Microphone Arrays. Proc. Int. Conf. on Digital Signal Processing, pp. 223—228, Florence 1991

[3] *Grenier, Y.; Xu, M.*: An adaptive Microphone Array for Speech Input in Cars. Proc. 22nd Intl. Symposium on Automotive Technology and Automation, pp. 485—492, Florence 1990

[4] *Gierl, S.*: Noise Reduction for Speech Transmission Using Microphone Array Systems. PhD dissertation, Karlsruhe 1990 (in German)

[5] *Kroschel, K.; Reich, W.*: A Comparison of Noise Reduction Systems for Speech Transmission. Proc. Europ. Conf. on Circuit Theory and Design, pp. 565—568, Prague 1985

[6] *Kroschel, K.; Lange, K.*: A Real—time System for Noise Reduction Using Microphone Subarrays. Nachrichten—Elektronik, to be published (in German)

[7] *Portnoff, M.R.*: Time—Frequency Representation of Digital Signals and Systems Based on Short—Time Fourier Analysis of Sampled Speech. Trans. on Acoustics, Speech, and Signal Processing, vol. ASSP—28, no. 1, pp 56—59, February 1980

[8] *Kroschel, K.; Lange, K.*: Rapid Prototyping for Digital Signal Processing. GME—Fachbericht 11 Mikroelektronik, Dresden 1993, pp. 317—322 (in German)