# Real-time audio signal enhancement for hands-free speech applications

*Thomas Fehér, Michael Freitag, Christian Gruber*

voice INTER connect GmbH

{thomas.feher,michael.freitag,christian.gruber}@voiceinterconnect.de

## Abstract

Automatic speech recognition in a home environment is currently of large interest for many practical applications. It is also challenging for signal processing, since several problems have to be solved satisfactorily. First, a large degradation of speech recognition performance is caused by room reverberations. Second, environmental noises like a radio, television or kitchen devices may further degrade speech recognition. And third, the problem of concurrent and possibly moving speakers has to be addressed.

In this paper we present our results of an extensive investigation of state-of-the-art multi-channel signal processing algorithms used for dereverberation, noise reduction and speaker localization. The investigation is based on a microphone array with 16 microphones which are part of a fixed ceiling-mounted device. All signal processing algorithms are implemented on a PC platform to run in real-time and produce an enhanced signal, which is finally used for speech recognition.

**Index Terms**: dereverberation, noise reduction, source localization, speech recognition

## 1. Introduction

To implement multichannel speech enhancement algorithms within a real-time setup, at first some practical questions have to be answered:

- number of microphones
- distributed vs. concentrated microphone array
- placement of microphones in a room

In this investigation we focused on using a large number of microphones. A concentrated microphone array was chosen in order to avoid long cables to connect the microphones. Due to the fact, that in a home environment the possible speaker positions differ mainly in horizontal direction and less in vertical direction, a ceiling mounted microphone array was chosen for this investigation. To utilize the advantage of known microphone positions, all microphones were placed on a stable mounting device.

Furthermore we decided to stay strictly in the signal domain, generating an enhanced output signal. This inhibits using algorithms in feature and model domain for dereverberation and noise reduction. However, our focus was to be independent of subsequent processing steps. Therefore the system could be used not only for speech recognition but also for handsfree communication or other applications.

The selection of algorithms was done under consideration of the large number of microphones, the known microphone positions and the design concept of producing an enhanced output signal.

In related works often distributed microphones are used. The signals are either joined in feature domain [1] or simply selected by certain criteria like SNR or speech content [2, 3, 4]. In [1] an additional video camera is used to improve tracking of persons.

A elaborate overview of the current state of similar approaches is given in [5].

## 2. System architecture

Figure 1 shows an overview of the implemented system as an audio signal pre-processing unit for hands-free speech applications. The sound acquisition unit is based on digital MEMS microphones with $I^2S$ output. For ease of development and debugging an off-the-shelf PC with multi-core processor and SSD is used for algorithm execution. Audio data is acquired by a 16 channel $I^2S$-to-USB sound card. The acquired digital audio data is processed in real-time by cooperation of concurrently executed processes implemented in either MATLAB or C. The software architecture is shown in figure 1 as well. The processed audio signal can then be used as input for automatic speech recognition or hands-free telephony applications, for example.

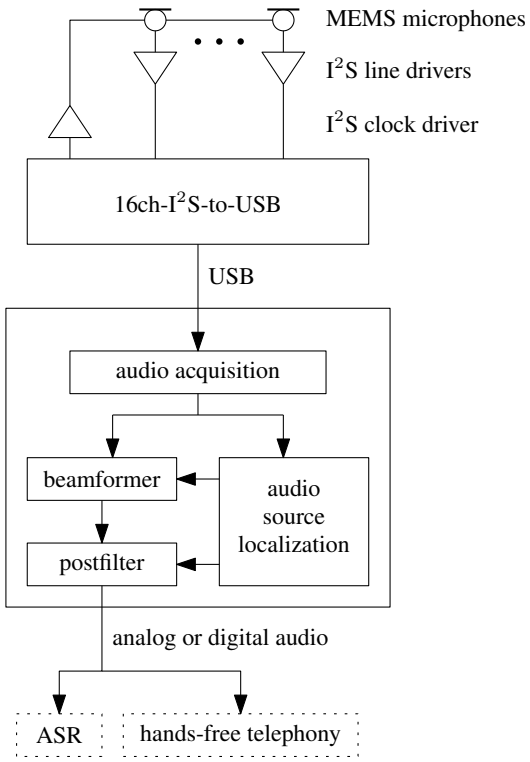


Figure 1: Hardware and software architecture

# 3. Array design

The arrangement of the microphones is based on the well known spiral geometry [6]. This geometry is often used in measurement setups due to its good side lobe suppression. For generating the position $(x_i, y_i)$ of microphone $i$ of one spiral arm, the following equation from [6] was utilized.

$$x_i = \alpha \cos\left(g(i,n)\right) \exp\left(\beta g(i,n)\right) \quad (1)$$

$$y_i = \alpha \sin\left(g(i,n)\right) \exp\left(\beta g(i,n)\right) \quad (2)$$

$$g(i,n) = \frac{(i-1)\,2\pi}{n} \quad (3)$$

Here $\alpha$ and $\beta$ are arbitrary parameters that can be adapted to fit the design requirements (room size, directivity, etc.). Specifying three microphones per spiral arm ($n$) and one center microphone resulted in an array of five arms with 16 microphones in total.

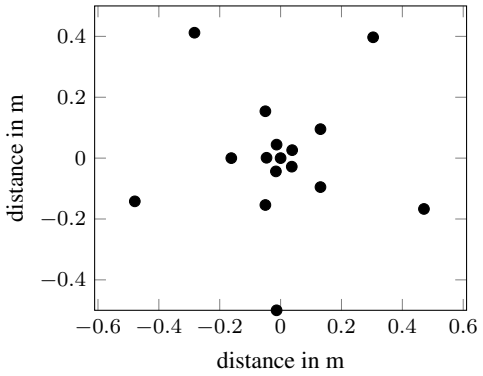


Figure 2: Array geometry

The parameters $\alpha$ and $\beta$ were optimized for maximum directivity with a differential evolution approach [7]. Additionally a third parameter describing the behaviour of the constant directivity beamformer (CDB) (see section 4) was optimized in this step. The overall diameter of the array was bound to 1 m. Calculation of the fitness was done by simulating the array response under free-field conditions and deriving the directivity index thereof. The resulting array geometry is shown in figure 2.

Figure 3 shows the simulated directivity index as well as the attenuation of the side lobes with respect to the main lobe.

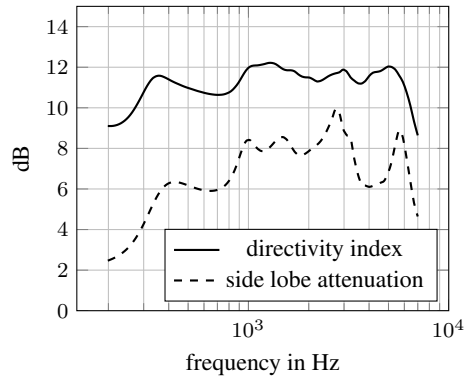


Figure 3: Simulated directivity index and side lobe attenuation

# 4. Algorithms

## 4.1. Beamforming

The beamforming algorithm performs two tasks: derverberation and noise suppression. Using the directivity feature, diffuse sound approaching from other than the preferred direction is attenuated. This leads to a significant dereverberation of the focused speech signal. Diffuse environmental noise is attenuated as well. Furthermore disturbing sources of directed background noise can be suppressed using the beamformer. By implementing a null steering even larger attenuation values compared to diffuse sound attenuation can be reached.

The beamforming algorithm is a classical constrained broadband MVDR beamformer [8] combined with a constant directivity beamformer (CDB) [9]. The MVDR approach leads to a constant beam width for low and mid-range frequencies. By constraining the MVDR, strong amplification of uncorrelated noise (self-noise of the microphone capsules) at very low frequencies is avoided. In the higher frequency range the CDB provides for a constant beam width.

Calculation of the frequency domain MVDR weights $\boldsymbol{W} = [W_1, W_2, \ldots, W_p, \ldots]^T$ for each microphone $p$ is achieved by

$$\boldsymbol{W}_{\mathrm{MVDR}} = \frac{\boldsymbol{\Gamma}_{vv}^{-1}\boldsymbol{d}}{\boldsymbol{d}^H\boldsymbol{\Gamma}_{vv}^{-1}\boldsymbol{d}}, \quad (4)$$

with $\boldsymbol{\Gamma}_{vv}$ being the coherence matrix of the noise field and the direction vector of the sound source $\boldsymbol{d} = [a_1 e^{j\omega\varphi_1}, a_2 e^{j\omega\varphi_2}, \ldots, a_p e^{j\omega\varphi_p}, \ldots]^T$ containing the phase shift $\varphi_p$ and attenuation $a_p$ of a sound wave from the desired direction. Notation of the frequency dependence is omitted for all terms.

The CDB is calculated as described in [9] and the resulting weights $\boldsymbol{W}_{\mathrm{CDB}}$ are combined with the weights of the MVDR $\boldsymbol{W}_{\mathrm{MVDR}}$ to yield a flat response in steering direction.

$$\boldsymbol{W} = \boldsymbol{W}_{\mathrm{MVDR}} \circ \boldsymbol{W}_{\mathrm{CDB}} \circ \frac{1}{\left|\left(\boldsymbol{W}_{\mathrm{MVDR}} \circ \boldsymbol{W}_{\mathrm{CDB}}\right)^H \boldsymbol{d}\right|} \quad (5)$$

The operator $\circ$ denotes the hadamard product.

Figure 3 shows that this approach results in a constant directivity of about 11 dB for a frequency range from 200 Hz to 6000 Hz.

## 4.2. Post filtering

Further improvement is realized by applying a post-filter. Using the approach by Leukimmiatis presented in [10] a significant improvement in noise reduction and dereverberation was accomplished. The coefficients $H$ of the post-filter are calculated by

$$H = \frac{\boldsymbol{\Phi}_{ss}}{\boldsymbol{\Phi}_{ss} + \boldsymbol{\Phi}_{vv}\boldsymbol{W}^H\boldsymbol{\Gamma}_{vv}\boldsymbol{W}}, \quad (6)$$

containing the estimated noise power spectral density

$$\Phi_{vv} = \frac{2}{M(M-1)} \sum_{p=0}^{M-2} \sum_{q=p+1}^{M-1} \Phi_{vv}^{(pq)} \quad (7)$$

$$\Phi_{vv}^{(pq)} = \frac{\frac{1}{2}(\Phi_{Y_pY_p} + \Phi_{Y_qY_q}) - \Re\{\Phi_{Y_pY_q}\}}{1 - \Re\{\Gamma_{v_pv_q}\}} \quad (8)$$

and the estimated signal power spectral density

$$\Phi_{ss} = \frac{2}{M(M-1)} \sum_{p=0}^{M-2} \sum_{q=p+1}^{M-1} \Phi_{ss}^{(pq)} \tag{9}$$

$$\Phi_{ss}^{(pq)} = \frac{\Re\{\Phi_{Y_pY_q}\} - \frac{1}{2}(\Phi_{Y_pY_p} + \Phi_{Y_qY_q})\Re\{\Gamma_{v_pv_q}\}}{1 - \Re\{\Gamma_{v_pv_q}\}} \tag{10}$$

for microphones $p$ and $q$ with a total of $M$. Operator $\Re$ denotes the real part of a complex number.

### 4.3. Source localization

The beamforming algorithm needs directional information about the source of interest. Therefore a source localization algorithm is essential. This algorithm must be able to find different audio sources in a room, to distinguish and even to track them. We decided to use a steered-response-power algorithm (SRP) to calculate a map of the acoustic power $P$ for a grid of points $r$ and a vector of input signals $\boldsymbol{X} = [X_1, X_2, \ldots, X_p, \ldots]^T$.

$$P(r) = \left| \boldsymbol{W}_r^H \boldsymbol{X} \right|^2 \tag{11}$$

To avoid recalculating equation (5) and specifically the computationally expensive inversion of $\boldsymbol{\Gamma}$ in equation (4) all filter-coefficients $\boldsymbol{W}$ are pre-calculated and stored in the random-access memory (RAM). This reduced the computational load significantly and allowed calculating the SRP in real-time. Using the normalized arithmetic mean (NAM) from [11] resulted in further improvements of the localization. In order to formulate the NAM it is necessary to partition equation (11) into the power for each frequency bin

$$P_f(r) = |\boldsymbol{W}_r \circ \boldsymbol{X}|^2 \tag{12}$$

and the broadband power

$$P(r) = \sum_f P_f(r). \tag{13}$$

Applying the NAM algorithm to (13) results in

$$P(r) = \sum_f \frac{P_f(r)}{\max_r[P_f(r)]}, \tag{14}$$

which effectively normalizes the energy in each frequency bin separately to the maximum value found at all positions.

The selection of the preferred audio source is realized by heuristic methods, which take loudness, stationarity and movement into account.

## 5. Implementation details

For sound acquisition digital MEMS microphones are used due to their superior characteristics compared to electret microphones in terms of

- magnitude and phase mismatch of the microphone sensitivity among different samples
- reduced susceptibility for electromagnetic interferences due to digital transmission
- cheaper cabling and no additional components like preamplifiers

Especially the generally lower magnitude and phase variations of these microphones present advantages for beamforming. When no complete calibration can be accomplished, magnitude variations could cause degradation of the side lobe structure whereas phase deviations could disturb the direction of the main lobe.

To bridge the long distance between the microphones and the I$^2$S-to-USB sound card, proprietary I$^2$S clock and data line drivers (see figure 1) have been developed which allow for a maximum distance of 16 m between standard MEMS microphones and the corresponding receivers using standard flat ribbon cable. With this setup a fully synchronous sampling of all 16 microphones can be achieved.

To achieve synchronous and low delay audio signal processing in the context of cooperating parallel processes, the inter-process-communication library ZeroMQ is utilized. A rigid request-reply scheme allows for fast processing and reliable detection of any buffer overflows and underruns.

## 6. Test framework

In order to systematically test and optimize the implemented algorithms, a test framework was designed that allows for offline processing of audio data. Figure 4 shows the structure. All components are realized in software. A test case starts with the definition of a representative real-life audio scene that can comprise an arbitrary number of (possibly moving) human speakers and stationary or nonstationary noise sources as well as a reverberant environment. A continuous audio data stream is then either recorded with the afore-mentioned audio acquisition device or generated via convolution on the basis of pre-recorded clean speech data, room impulse responses and noise sources.

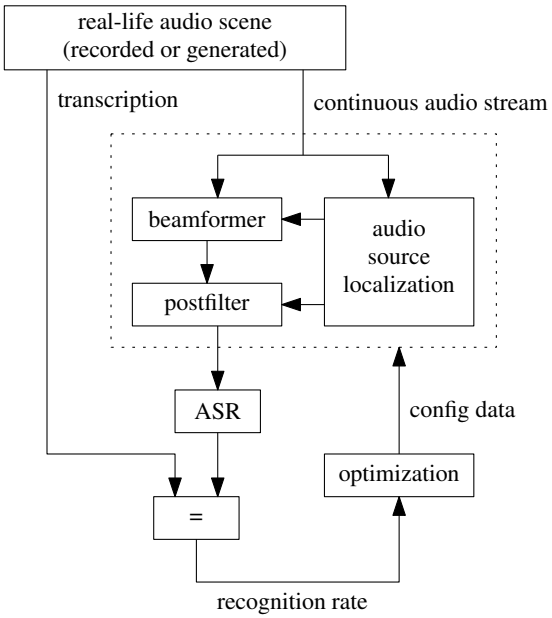


Figure 4: Test framework for offline audio processing

An appropriate transcription of the audio stream is used to score the ASR results of the processed speech. The resulting measures like utterance recognition rate provide the basis for an automatic optimization of the audio algorithm parameters. Since all software components support multiple instances, a module layer concurrency can be realized, which leads to real

Table 1: ASR scenarios

| # | distance between speaker and microphone array | distance between speaker and noise source | noise type |
|---|---|---|---|
| 1 | 1.60 m | 3.70 m | radio |
| 2 | 1.90 m | 1.20 m | water tap |
| 3 | 1.60 m | 2.40 m | hoover |
| 4 | 1.60 m | 0.40 m | kitchen device |
| 5 | 1.90 m | - | - |
| 6 | 2.80 m | - | - |

time factors in the order of only one percent with state-of-the-art recognizer technology on ordinary desktop PCs.

## 7. Results

The performance of the complete signal enhancement system was evaluated using the test framework described in section 6. The speech signal database used for evaluation was recorded with 37 different male and female speakers in a natural home environment. This room exhibits a base area of about $30\,\mathrm{m}^2$ and a height of $3.15\,\mathrm{m}$. The reverberation time $T_{60}$ is ca. $550\,\mathrm{ms}$. The recorded utterances are focused on command-and-control utterance recognition. The database comprises 103 different utterances recorded from each speaker.

A total of 500 randomly selected utterances spread across all speakers were used for evaluating the utterance recognition rate (RR). Each utterance is a sequence of predefined commands, which can include filling words. The utterance recognition rate is defined as the amount of completely correct recognized utterances out of the total corpus. Any erroneous, missing or inserted command in an utterance leads to rejection.

Additionally four different noise signals were recorded in this room. The recorded speech and noise signals were arranged to six different scenarios (see table 1). The different scenarios differ in the distance of the speaker to the microphone array, the distance of the noise source to the speaker and in the noise type. Especially in scenario 4 the distance of the noise source to the speaker is quite short. Scenarios 5 and 6 are undisturbed scenarios without any noise. In return the speaker distance to the microphone array is quite large in scenario 6.

Finally the RR was evaluated first with the unprocessed output signal of the mid microphone of the array and second with the processed output signal of the full microphone array. The same corpus was used for both experiments. Figure 5 shows the results of ASR tests.

The main difference between the ASR results for the unprocessed and the processed signals can be seen in scenario 1. The radio is a non-stationary noise source and even worse the radio signal is a news broadcast, which means the radio behaves like a second speaker. Using the microphone array the radio signal could be drastically attenuated, which makes speech recognition possible in the first place. The same holds for scenarios 2 and 3. The microphone array leads to significant better recognition results. In scenario 4 no significant improvement could be reached due to the small distance of the noise source to the speaker. In scenarios 5 and 6, where no noise is present, RR is already relatively high even without processing. But due to the dereverberation effect of the microphone array, the RR could be further improved from approximately $85\,\%$ to $93\,\%$ (scenario 5) and $96\,\%$ (scenario 6) respectively. That means the utterance error rate is reduced from approximately $15\,\%$ to $7\,\%$ and $4\,\%$.

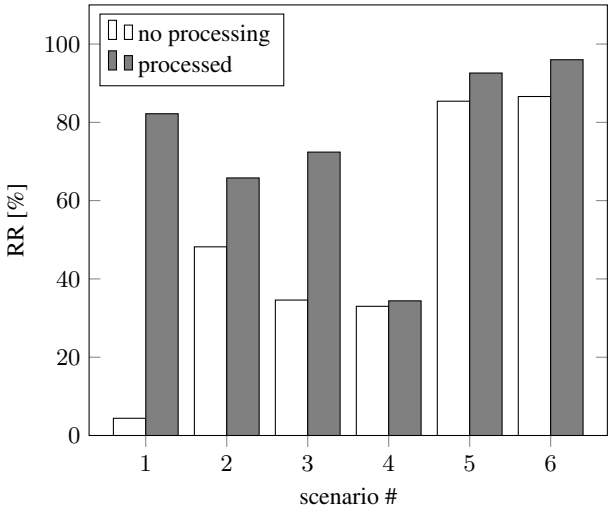


Figure 5: ASR results as recognition rates on utterance level

## 8. Conclusions

The investigation showed, that a ceiling mounted microphone array with a large number of microphones and fixed microphone positions is a promising concept for practical speech recognition applications in a home environment. Using beamforming and post filtering techniques, significant signal dereverberation as well as noise reduction can be achieved. This in turn considerably improves speech recognition performance in disturbed and undisturbed situations. Additionally using audio source localization techniques several speakers can be distinguished, localized and tracked.

## 9. References


[1] A. Karpov, L. Akarun, H. Yalçın, A. Ronzhin, B. E. Demiröz, A. Çoban, and M. Železnỳ, "Audio-visual signal processing in a multimodal assisted living environment," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[2] M. Vacher, B. Lecouteux, and F. Portet, "Multichannel Automatic Recognition of Voice Command in a Multi-Room Smart Home : an Experiment involving Seniors and Users with Visual Impairment," in *Interspeech 2014*, Singapour, Singapore, Sep. 2014, pp. 1008–1012. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01003492

[3] E. Principi, S. Squartini, F. Piazza, D. Fuselli, and M. Bonifazi, "A distributed system for recognizing home automation commands and distress calls in the italian language." in *INTERSPEECH*, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, Eds. ISCA, 2013, pp. 2049–2053. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2013.html#PrincipiSPFB13

[4] M. Vacher, B. Lecouteux, D. Istrate, T. Joubert, F. Portet, M. Sehili, and P. Chahuara, "Evaluation of a real-time voice order recognition system from multiple audio channels in a home," in *the 14rd Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013, pp. 2062–2064.

[5] M. Chan, D. Estève, C. Escriba, and E. Campo, "A review of smart homespresent state and future challenges," *Computer methods and programs in biomedicine*, vol. 91, no. 1, pp. 55–81, 2008.

[6] E. J. Arcondoulis, C. J. Doolan, A. C. Zander, and L. A. Brooks, "Design and calibration of a small aeroacoustic beamformer," in

*International Congress on Acoustics (ICA)(20th: 2010: Sydney, New South Wales)*, 2010.

[7] S. Das, A. Abraham, U. Chakraborty, and A. Konar, "Differential evolution using a neighborhood-based mutation operator," *Evolutionary Computation, IEEE Transactions on*, vol. 13, no. 3, pp. 526–553, June 2009.

[8] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*. Springer, 2001, pp. 39–60.

[9] D. B. Ward, R. A. Kennedy, and R. C. Williamson, "Constant directivity beamforming," in *Microphone Arrays*. Springer, 2001, pp. 39–60.

[10] S. Leukimmiatis and P. Maragos, "Optimum post-filter estimation for noise reduction in multichannel speech processing," in *Proc. 14th European Signal Processing Conference*, 2006.

[11] D. Salvati, C. Drioli, and G. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," *Signal Processing Letters, IEEE*, vol. 21, no. 5, pp. 581–585, May 2014.