BEAMFORMER FOR DRIVING BINAURAL SPEECH ENHANCEMENT

Heinrich W. Löllmann and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, 52056 Aachen, Germany
loellmann@ind.rwth-aachen.de

ABSTRACT

A beamformer for binaural speech enhancement systems in digital hearing aids is proposed. Its single modules for the estimation of the time-difference-of-arrival (TDOA) and time-alignment operate in the frequency-domain and have a low computational complexity. The TDOA estimation is performed efficiently by a generalized cross-correlation with phase transform weighting. The estimation accuracy for filter-banks with a limited number of subbands, which are needed for hearing aids to meet tight delay constraints, is improved by a histogram-based TDOA estimation. The subsequent time-alignment is accomplished by a simple multiplication with spectral phase factors.

A primary application of the proposed system are binaural cue preserving speech enhancement systems based on spectral weighting. The proposed beamformer can be used as delay-and-sum and/or delay-and-subtract beamformer to provide subband signals from which the power spectral densities of interfering sources can be estimated to drive the spectral weight calculation.

Index Terms— delay-and-sum beamformer, TDOA estimation, binaural hearing aids, speech enhancement

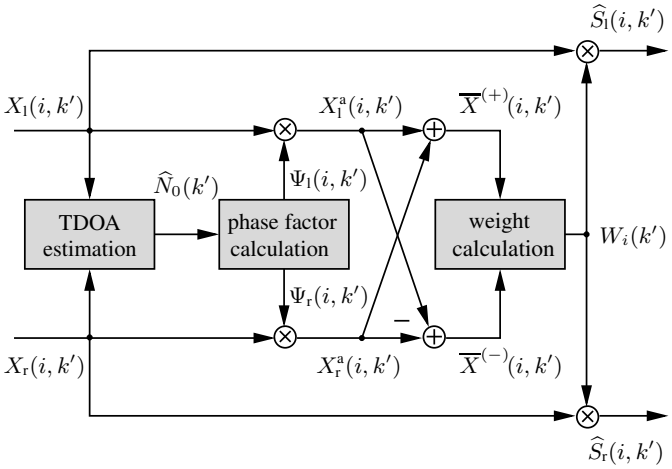

Fig. 1. *Binaural cue preserving speech enhancement by spectral weighting with spectral weight calculation driven by a delay-and-sum beamformer and/or delay-and-subtract beamformer (DS BF).*

1. INTRODUCTION

Modern hearing aids (HAs) are mostly equipped with multiple microphones and, more and more, with a wireless binaural data link between left and right HA device. This allows to execute multi-channel algorithms for signal classification or speech enhancement, e.g., [1]. Accordingly, the microphone array and beamformer design for HAs has received a lot of research interest, e.g., [2]. The design goal is usually to obtain a fixed or adaptive beamformer with a high directivity towards a desired source to achieve a high suppression of interfering sound sources, noise or room reverberation.

Another approach is to perform binaural speech enhancement by spectral weighting. In such a case, a beamformer is not employed for speech enhancement by beamsteering, but to *drive* the spectral weight calculation as proposed in [3] for binaural cue preserving speech dereverberation. A delay-and-sum beamformer (DS BF) is used to obtain an averaged 'reference signal' of the left and right HA signal from which the power spectral density (PSD) of late reverberant speech for the spectral weight calculation is estimated. However, the beamformer design, including the needed source localization, for such purposes is subject to challenging design constraints.

First, algorithms for HAs must have a low algorithmic complexity, mainly due to the limited battery size and capacity. Second, speech enhancement algorithms for HAs operate mostly in the frequency-domain and, at the same time, must feature a low system delay to avoid a disturbing comb-filter effect, cf., [4]. Different filter-bank designs have been proposed to enable subband processing with low signal delay, e.g., [5, 6]. They have in common that only a rather limited number of frequency channels can be used to meet the delay constraints for HAs. Accordingly, a beamformer implementation for HAs has to cope with a limited number of frequency bands and should operate exclusively in the subband-domain to avoid additional filter-banks.

The goal of this paper is to present a delay-and-sum / delay-and-subtract beamformer system, including the needed time-difference-of-arrival (TDOA) estimation, which addresses the outlined needs for speech enhancement in binaural HAs.

2. OVERALL SYSTEM

A diagram of the presented system for beamformer driven binaural speech enhancement is provided by Fig. 1.

The input signals of left and right HA channel, $X_\mathrm{l}(i,k')$ and $X_\mathrm{r}(i,k')$, are subband signals of two M-channel discrete Fourier transform (DFT) analysis filter-banks (not shown in Fig. 1). The index $i \in \{0,1,\ldots,M-1\}$ marks the frequency (channel) index and $k' = \lfloor k/R \rfloor$ the discrete time index after downsampling (frame index), where a downsampling rate of $R = M/4$ is used here. The complex input subband signals of both HA devices, $X_\mathrm{l}(i,k')$ and $X_\mathrm{r}(i,k')$, are multiplied with real weights $W_i(k')$ to perform speech enhancement by spectral weighting. Identical weights are applied to both channels to preserve the binaural cues, e.g., [3].

The enhanced subband signals, $\widehat{S}_\mathrm{l}(i,k')$ and $\widehat{S}_\mathrm{r}(i,k')$, are transformed into the time-domain by a DFT synthesis filter-bank (which

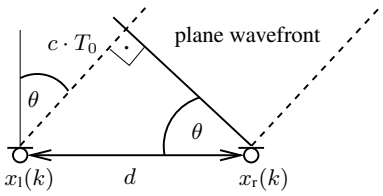

Fig. 2. *Setup and notation for TDOA estimation.*

is not depicted in Fig. 1). An alternative is to use the concept of the filter-bank equalizer [6] where the weights $W_i(k')$ are transformed into the time-domain to perform time-domain filtering instead of spectral weighting to achieve a lower signal delay. The following prototype lowpass filter for the DFT analysis filter-bank has been found suitable for the considered beamforming system

$$h_0(n) = \frac{1}{M}\,\mathrm{si}\left(\frac{\pi}{M}\left(n - \frac{L}{2}\right)\right), \quad n \in \{\,0, 1, \ldots, L-1\,\} \tag{1}$$

with prototype filter length $L = 2\,M$ and $\mathrm{si}(x) = \sin(x)/x$. It provides a nearly perfect signal reconstruction for a DFT analysis-synthesis filter-bank (AS FB) with the same synthesis prototype filter and perfect reconstruction for the filter-bank equalizer [6].

The calculation of the spectral weights $W_i(k')$ is based on the subband signals $\overline{X}^{(+)}(i,k')$ and/or $\overline{X}^{(-)}(i,k')$ obtained by a delay-and-sum and/or delay-and-subtract beamformer, both abbreviated as DS BF. The difference $\overline{X}^{(-)}(i,k')$ can be used for the PSD estimation of non-coherent signal components such as diffuse noise. The added signals $\overline{X}^{(+)}(i,k')$ can be used for the PSD estimation from coherent signal components. As an example, a delay-and-sum beamformer is used in [3] for binaural speech dereverberation to obtain a 'reference signal' from both HA signals by which the PSD of late reverberant speech components for the calculation of the common weights $W_i(k')$ can be estimated. For such purpose, a delay-and-sum beamformer is preferable to a superdirective beamformer as it performs a lower dereverberation and requires a lower complexity.

In both cases, the subband signals $X_\mathrm{l}(i,k')$ and $X_\mathrm{r}(i,k')$ need to be time-aligned before being subtracted or added. For this, an estimation of the *time-difference-of-arrival* (TDOA) for the desired source signal is needed. The efficient realization of TDOA estimation and time-alignment is treated in the following.

3. TDOA ESTIMATION

The most challenging part of the beamforming system in Fig. 1 is the estimation of the TDOA by which the desired signal arrives at the microphones of left and right HA. The considered setup is depicted in Fig. 2. A plane wavefront reaches the microphone pair with distance d. The direction-of-arrival (DOA) marked by the angle θ and the TDOA T_0 in seconds are related by $\sin\theta = c \cdot T_0/d$ with $c = 340\,\mathrm{m/s}$ marking the speed of sound. In the following, the TDOA in sample instants is considered given by $N_0 = T_0\,f_s$ with f_s marking the sampling frequency.

The model of Fig. 2 is of course rather coarse as the head between the microphones of both HAs is not considered as, e.g., in [7], but this simple model facilitates on the other hand the use TDOA estimation techniques with a low complexity.

A comprehensive treatment of algorithms for TDOA estimation can be found e.g., in [2, 8]. A very efficient and widely used method to estimate the TDOA out of two sensor signals is the use of the generalized cross-correlation (GCC) function [9]. Different frequency-domain weighting functions are proposed in [9] for the evaluation of the GCC function where the so-called phase transform (PHAT) is of special interest for speech signals (see also [8]).

This approach, termed shortly as GCC-PHAT algorithm in the following, forms the basis for the employed method. In a first step, the cross-power spectral densities (CPSDs) of both input signals are calculated by recursive averaging

$$\widehat{\Phi}_{x_\mathrm{l},x_\mathrm{r}}(i,k') = \alpha\,\widehat{\Phi}_{x_\mathrm{l},x_\mathrm{r}}(i,k'-1) + (1-\alpha)\,X_\mathrm{l}(i,k')\,X_\mathrm{r}^*(i,k') \tag{2}$$

where the asterisk denotes the conjugate complex value. Only the values for $i \in \{\,0, 1, \ldots, M/2\,\}$ need to be calculated since

$$\widehat{\Phi}_{x_\mathrm{l},x_\mathrm{r}}(i,k') = \widehat{\Phi}^*_{x_\mathrm{l},x_\mathrm{r}}(M-i,k')\,. \tag{3}$$

The phase transform (PHAT) provides the weighted CPSDs

$$\widehat{\Phi}^{(\mathrm{phat})}_{x_\mathrm{l},x_\mathrm{r}}(i,k') = \frac{\widehat{\Phi}_{x_\mathrm{l},x_\mathrm{r}}(i,k')}{\max\left\{\,\left|\widehat{\Phi}_{x_\mathrm{l},x_\mathrm{r}}(i,k')\right|,\,\epsilon\,\right\}} \tag{4}$$

where the threshold $\epsilon > 0$ avoids a division by zero. The inverse fast Fourier transform (IFFT) of the weighted CPSDs yields the GCC function $\hat{\varphi}^{(\mathrm{gcc})}_{x_\mathrm{l},x_\mathrm{r}}(n,k')$. The index n for the maximum peak of this GCC function provides the preliminary TDOA estimate $\widehat{N}^{(\mathrm{pre})}_\mathrm{o}(k')$. For a sampling frequency of $f_s = 24\,\mathrm{kHz}$ and ear spacing of $d = 17\,\mathrm{cm}$, the TDOA estimate is supposed to lie within the range

$$-12 \le \widehat{N}^{(\mathrm{pre})}_\mathrm{o}(k') \le 12 \quad \text{for} \quad \widehat{N}^{(\mathrm{pre})}_\mathrm{o}(k') \in \mathbb{Z} \tag{5}$$

and estimates being outside this range are rejected, i.e., $\widehat{N}^{(\mathrm{pre})}_\mathrm{o}(k') = \widehat{N}^{(\mathrm{pre})}_\mathrm{o}(k'-1)$ in this case. The final TDOA estimate of the GCC-PHAT algorithm is obtained by recursive smoothing where a smoothing factor of 0.99 is used here. The GCC-PHAT algorithm can also be used to calculate a fractional value for $\widehat{N}^{(\mathrm{pre})}_\mathrm{o}(k')$ by means of interpolation which, however, is not considered here due to the increased computational burden.

For the evaluation of the TDOA estimation with realistic setups, a speech signal were convolved with different binaural room impulse responses (BRIRs) at 24 kHz sampling frequency. The BRIRs were measured with HA dummies (Siemens Life 500) mounted on a dummy head. Setups with different microphone-loudspeaker distances D, DOAs θ and reverberation times (RTs) T_{60} were considered: a cellar room ($D = 4.4$ m, $\theta = 90°$, $T_{60} = 1.6$ s), a foyer room ($D = 2.6$ m, $\theta = 60°$, $T_{60} = 1.2$ s) and a meeting room ($D = 1.1$ m, $\theta = 15°$, $T_{60} = 0.4$ s). The stated T_{60} value is given by the rounded average value of the RTs calculated for each BRIR.

Fig. 3 exemplifies the performance of the described GCC-PHAT algorithm for the setup with the foyer room. It can be observed that the estimation accuracy of the GCC-PHAT algorithm improves with an increasing DFT size M. However, only a limited DFT size ($M \le 128$) can be usually tolerated for filter-banks used in HAs, cf., [5]. Therefore, a modification of the GCC-PHAT algorithm is devised to improve the estimation accuracy for such cases.

In a first step, the last L_h preliminary estimates $\widehat{N}^{(\mathrm{pre})}_\mathrm{o}(k')$ which fulfill Eq. (5) are buffered. The first and last value of this buffer are used in a second step to update the histogram for all buffered estimates. The TDOA value corresponding to the maximum of this histogram is taken as new TDOA estimate $\widehat{N}^{(\mathrm{peak})}_\mathrm{o}(k')$. It should be noted these operations can be realized with only a few if-statements, assignments and a for-loop. In addition, each preliminary TDOA estimate $\widehat{N}^{(\mathrm{pre})}_\mathrm{o}(k')$ is integer-valued and bounded according to Eq. (5).

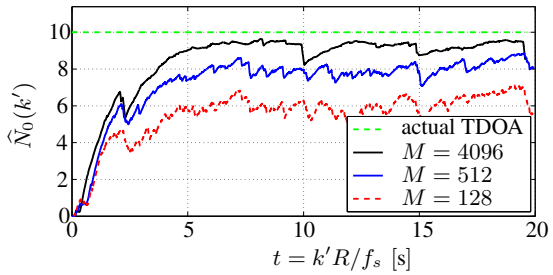

Fig. 3. *TDOA estimation by GCC-PHAT algorithm with different DFT sizes M in a foyer room with $T_{60} = 1.2$ s and $\theta = 60°$, which equals an integer-valued TDOA of 10 sample instants.*

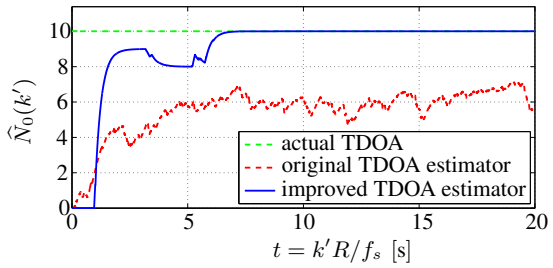

Fig. 4. *TDOA estimation with modified and original GCC-PHAT algorithm for $M = 128$ subbands. The other parameters are the same as for Fig. 3.*

Therefore, the memory consumption for the buffer as well as the histogram can be kept rather low as each preliminary estimate can be represented by 5 bit for the considered setup.

The estimate obtained by the peak of the histogram is smoothed recursively to obtain the final TDOA estimate

$$\widehat{N}_0(k') = (1-\beta) \cdot \widehat{N}_0^{(\text{peak})}(k') + \beta \cdot \widehat{N}_0(k'-1) \qquad (6)$$

where a factor of $\beta = 0.99$ is employed for the later simulations.

The effect of the proposed modification is illustrated in Fig. 4. It can be seen that the improved system provides a much higher estimation accuracy than the original GCC-PHAT algorithm. The new TDOA estimation requires a short adaption period as the histogram has to be built up. Table 1 reveals that the improved approach achieves for all considered setups significantly better results than the original GCC-PHAT algorithm. The deviations from the actual TDOA value even for an environment with low reverberation are due to the shadowing of the head, since this effect has not been observed for setups without a dummy head.

It is important to notice that the employed prototype lowpass filter or filter-bank design, respectively, has a strong influence on the performance of the TDOA estimation. As exemplified by Fig. 5, the use of a simple rectangular prototype filter, $h_0(n) = 1$ for $n = 0, 1, \ldots M-1$, is less suitable for the TDOA estimation due to its low spectral resolution.

Table 1. *Average TDOA estimates (in sample instants) for the GCC-PHAT algorithm and the new approach for different setups.*

	cellar room	foyer room	meeting room
actual TDOA	12.0	10.4	3.1
new approach	10.0	9.7	2.0
GCC-PHAT	2.8	5.9	1.9

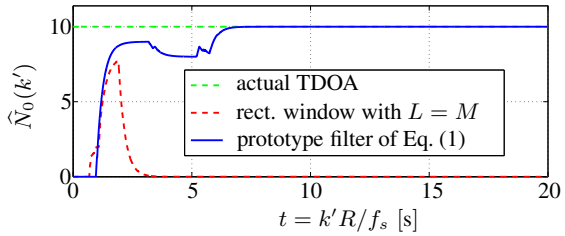

Fig. 5. *TDOA estimation by new GCC-PHAT algorithm using a DFT analysis filter-bank with $M = 128$ subbands and different prototype lowpass filters. The other parameters are the same as for Fig. 3.*

The GCC-PHAT algorithm and its improved version have of course their limitations regarding scenarios with multiple sources or strong distortions by noise. However, the TDOA estimation in such cases with only two sensors *and* low complexity is still a challenging problem. The devised TDOA estimation has the beneficial property that it tends to a value of zero in the case of strong interference (see also Fig. 5), which corresponds to the common default assumption that the desired source is usually in the look direction of the HA user. It should be noted thereby that the beamformer of Fig. 1 is used for PSD calculations such that an erroneous TDOA estimation causes a less severe degradation of the speech quality than, e.g., for speech enhancement systems based on beamsteering.

4. TIME-ALIGNMENT

The TDOA estimate is needed for the time-alignment of the input signals according to Fig. 1 to ensure that the signals are added (or subtracted) in phase. Here, this delay compensation is not performed in the time-domain by a fractional delay filter [10] (as suggested in [3]), but in the frequency-domain. For this, the subband signals to be delayed are multiplied with the phase factors

$$\Psi(i,k') = \exp\left\{\frac{-j\,2\,\pi\,i\,|\widehat{N}_0(k')|}{M}\right\} \qquad (7)$$

for $i = 0, 1, \ldots, M/2$, where the remaining values are given by the symmetry relation of Eq. (3). In order to ensure causality, only the magnitude of the TDOA value $\widehat{N}_0(k')$ is used and the sign of the TDOA estimate determines whether the subband signals of the left or right analysis filter-bank are delayed. Hence

$$\Psi_\text{l}(i,k') = 1 \quad \wedge \quad \Psi_\text{r}(i,k') = \Psi(i,k') \quad \text{for} \quad \widehat{N}_0(k') \geq 0 \qquad (8a)$$

$$\Psi_\text{l}(i,k') = \Psi(i,k') \quad \wedge \quad \Psi_\text{r}(i,k') = 1 \quad \text{for} \quad \widehat{N}_0(k') < 0 \qquad (8b)$$

for the setup according to Fig. 1 and Fig. 2.

5. PSD ESTIMATION

The use of the DS BF of Fig. 1 for the estimation of the spectral variance of late reverberant speech is now investigated. The late reverberant spectral variance (LRSV) is needed for speech dereverberation by spectral subtraction [3, 11]. The LRSV can be determined by means of a statistical model for late reverberation as proposed in [11]. This method (as described in [3]) is used to estimate the LRSV either from the subband signals $X_\text{l}(i,k')$, $X_\text{r}(i,k')$ or $\overline{X}^{(+)}(i,k')$ according to Fig. 1. The reverberant binaural speech signal ($T_{60} \approx 1.6\,\text{s}$) for the foyer room is used with DOAs of $0°$ and

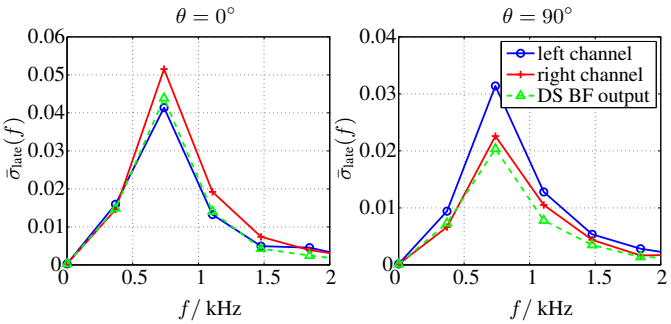

Fig. 6. *Average LRSV $\bar{\sigma}_{late}(f)$ estimated from binaural speech signals of foyer room (20 s duration, $T_{60} \approx 1.6$ s) at different DOAs θ using either $X_l(i,k')$, $X_r(i,k')$ or $\overline{X}^{(+)}(i,k')$ according to Fig. 1.*

90° as described Sec. 3. The averaged LRSV estimates are shown in Fig. 6. It can be observed that the average LRSV obtained from the DS BF is mostly equal to or lower than the lowest LRSV estimate of both channels and less a quantity in-between. (Similar observations can also be made for other binaural signals.) This mitigates the effect of signal distortions due to LRSV overestimation and justifies the assumption that a DS BF provides a suitable 'reference signal' of left and right channel to drive the spectral weight calculation.

6. COMPLEXITY AND DELAY

The algorithmic complexity of the DS BF of Fig. 1 in dependence of the number of subbands M is listed in Table 2.[1] The actual complexity for the exponential operation to determine the phase factors of Eq. (7) depends on the used procedure (e.g., look-up tables or direct calculation). The beamformer can operate with a higher downsampling rate than the analysis filter-bank, i.e., $R' = r\,R$ with $r \in \mathbb{N}$. For example, the used filter-bank with $M = 128$ subbands and $R' = 64$ requires only 38 real multiplications, 47 summations, 2 divisions and 2 exponential operations on average per input sample pair to obtain the M spectral coefficients $\overline{X}^{(+)}(i,k')$ at instant k'.

The overall complexity and delay of the system depends on the employed filter-bank, which is not considered in Table 2. For a DFT AS FB, the overall signal delay amounts to $L-1$ sample instants, where a delay of only $L/2$ samples can be achieved by the concept of the filter-bank equalizer, which corresponds to delays of 10.34 ms and 5.17 ms for the employed DFT filter-bank with prototype filter given by Eq. (1). A more detailed description of these filter-banks and a discussion of their algorithmic complexity can be found in [6].

7. CONCLUSIONS

This contribution addresses the practical problems of a DS BF implementation for binaural speech enhancement in HAs. The devised system is especially designed for subband processing systems in HAs, which can only operate with a rather restricted number of subbands ($M \leq 128$) to meet tight delay constraints.

The proposed beamformer including the needed TDOA estimation operates entirely in the subband-domain and has a low computational complexity. The TDOA estimation is performed by a histogram-based GCC-PHAT algorithm. This modified approach achieves a significantly higher estimation accuracy than the original GCC-PHAT algorithm with only a moderate increase of its computational complexity. The time-alignment of the subband signals prior to the summation is performed by spectral multiplication with phase factors, which causes no noticeable signal distortions.

A primary application of the proposed beamformer is binaural cue preserving speech enhancement in HAs, which is investigated in [12] in more detail.

[1] The algorithmic complexity can (only) serve as a rough indicator for the actual computational load, e.g., on a digital signal processor, which depends on the instruction set, processor architecture etc. Besides, the complexity for assignments, for-loops, if-statements etc. is not considered here.

Table 2. *Algorithmic complexity of the proposed DS BF according to Fig. 1 in terms of real multiplications, real summations, real divisions and exponential operations per input sample pair to calculate M spectral coefficients $\overline{X}^{(+)}(i,k')$ at instant k'.*

	TDOA estimation	time alignment & summation
mult.	$\frac{1}{R'}(4\,M + 14 + 2\,M\,\log_2 M)$	$\frac{1}{R'}(\frac{M}{2}+1)$
sum.	$\frac{1}{R'}(2M + 10 + 3\,M\,\log_2 M)$	$\frac{1}{R'}(\frac{M}{2}+1)$
div.	$\frac{1}{R'}(M+2)$	0
exp.	0	$\frac{1}{R'}(M+2)$

8. REFERENCES

[1] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal Processing in High-End Hearing Aids: State of the Art, Challenges, and Future Trends," *EURASIP Journal on Applied Signal Processing, Special Issue on DSP in Hearing Aids and Cochlear Implants*, , no. 18, pp. 2915–2929, Oct. 2005.

[2] M. Brandstein and D. Ward (Eds.), *Microphone Arrays*, Springer, Berlin, Heidelberg, 2001.

[3] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-Based Dereverberation Preserving Binaural Cues," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1732–1745, Sept. 2010.

[4] M. A. Stone and B. C. J. Moore, "Tolerable Hearing Aid Delays II: Estimation of Limits Imposed During Speech Production," *Ear and Hearing*, vol. 23, no. 4, pp. 325–338, 2002.

[5] R. W. Bäuml and W. Sörgel, "Uniform Polyphase Filter Banks for Use in Hearing Aids: Design and Constraints," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[6] H. W. Löllmann and P. Vary, "Low Delay Filter-Banks for Speech and Audio Processing," in *Speech and Audio Processing in Adverse Environments*, E. Hänsler and G. Schmidt, Eds., chapter 2, pp. 13–61. Springer, Berlin, Heidelberg, 2008.

[7] S. Goetze, T. Rohdenburg, V. Hohmann, B. Kollmeier, and K.-D. Kammeyer, "Direction of Arrival Estimation Based on the Dual Line Approach for Binaural Hearing Aid Microphone Arrays," in *Proc. of Intl. Symposium on Intelligent Signal Processing and Communication Systems*, Xiamen, China, Nov. 2007, pp. 84–87.

[8] J. Chen, J. Benesty, and Y. Huang, "Time Delay Estimation in Room Acoustic Environments: An Overview," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–19, 2006.

[9] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[10] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the Unit Delay," *IEEE Signal Processing Magazine*, vol. 13, no. 1, pp. 30–60, Jan. 1996.

[11] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A New Method Based on Spectral Subtraction for Speech Dereverberation," *acta acoustica - ACOUSTICA*, vol. 87, no. 3, pp. 359–366, 2001.

[12] H. W. Löllmann and P. Vary, "Efficient Speech Dereverberation for Binaural Hearing Aids," in *Proc. of ITG Conference on Speech Communication*, Braunschweig, Germany, Sept. 2012.