Speech Understanding Performance of Cochlear Implant Subjects using Time-Frequency Masking based Noise Reduction

Obaid ur Rehman Qazi, Bas van Dijk, Marc Moonen, and Jan Wouters

***Abstract*—Cochlear Implant (CI) recipients report severe degradation of speech understanding in noisy conditions. Most CI recipients typically can require about 10-25 dB higher signal-to-noise ratio (SNR) than Normal Hearing (NH) listeners in order to achieve similar speech understanding performance. In recent years significant emphasis has been put on binaural algorithms, which not only make use of the head shadow effect, but also have two or more microphone signals at their disposal to generate binaural inputs. Most of the CI recipients today are unilaterally implanted but they can still benefit from the binaural processing utilizing a contra-lateral microphone. The Phase Error Filtering (PEF) algorithm tries to minimize the phase error variance utilizing a Time-Frequency (TF) mask for noise reduction. Potential improvement in speech intelligibility offered by the algorithm is evaluated with four different kinds of mask functions. The study reveals that the PEF algorithm which uses a contra-lateral microphone but unilateral presentation provides considerable improvement in intelligibility for both NH and CI subjects. Further, preference rating test suggests that CI subjects can tolerate higher levels of distortions than NH subjects, and therefore, more aggressive noise reduction for CI recipients is possible.**

***Index Terms*—Speech processing, cochlear implants, time frequency masking, phase error variance.**

I. Introduction

A Cochlear Implant (CI) is a surgically implanted auditory prosthesis which electrically stimulates the auditory nerve using an electrode array in the cochlea to produce sound sensation in people with profound or severe hearing loss. It enables hearing impaired people to perceive environmental sounds and speech [1], [2]. The effectiveness of such devices is dependent upon individual differences in auditory performance which are not always clearly understood. However, in general a satisfactory speech understanding performance with such devices is reported in noise-free environments. Whereas users report severe degradation in speech understanding in noisy conditions, especially in a cocktail party environment, where a nonsteady multi-talker noise overlaps in time and frequency with the target speech signal [3]. CI encodes sound by decomposing the input signal into different frequency bands using Discrete Fourier Transform (DFT) or a bank of bandpass filters and consequently modulating biphasic trains of electrical pulses on intra-cochlea electrodes using the filterbank output envelopes [4]. The envelope signals extracted from the bandpass filters are compressed with a nonlinear mapping function prior to the modulation in order to map the wide dynamic range of sound in the environment to the narrow dynamic range of electrically evoked hearing. The number of envelopes and electrodes selected for stimulation at each cycle differs for different coding strategies[4]. Around 60 % of CI recipients worldwide are using the Advanced Combinational Encoder (ACE™ offered in the nucleus™ system by Cochlear Ltd.) strategy, which is a so-called maxima selection strategy. ACE analyses the signal in 22 bands, but does not stimulate 22 electrodes on each analysis frame. Instead, the N largest bands are selected and stimulated, while the others are skipped. N is the 'number of maxima' and can be varied theoretically between 2 and 22; however, in clinical practice it varies between 8 and 12. In Continuous Interleave Sampling (CIS) strategy a fixed number of envelopes are computed and all corresponding electrodes are used for stimulation in every analysis frame [5]. Most CI recipients typically can require about 10-25 dB higher SNR than Normal Hearing (NH) listeners in order to achieve similar speech understanding performance [6], [12]. As a result various single channel speech enhancement techniques have been investigated in the past for cochlear implants[3], [7]-[11]. In contrast to the single channel noise reduction algorithms in hearing aids, where only comfort benefits are usually demonstrated but no improvement in speech intelligibility, CI does show some improvement in speech intelligibility. However, intelligibility improvement for CI remains small due to the estimation errors in the parameters for filtering and therefore cant not bridge the speech understanding performance gap between NH and CI recipients. In recent years significant emphasis has been put on multi-channel speech enhancement algorithms [12]-[15], which not only have two or more microphone signals at their disposal to generate an output, but also make use of the "best ear" signal in the binaural case.

A Time-Frequency (TF) mask is based on a TF representation of the signal, which can be obtained either by a short time Fourier transform (STFT) or a windowed auditory

Manuscript received October 24, 2011; revised January 3, 2012; accepted January 31, 2012. Date of publication; date of current version. This work is supported by EU project AUDIS and Cochlear Technology Centre Belgium.

O.R Qazi is with Cochlear Technology Center Belgium and Department of Neurosciences, Katholieke Universiteit Leuven 3000, Belgium. Phone: +32 15 79-5517, e-mail: oqazi@cochlear.com

B.v.Dijk is with Cochlear Technology Center Belgium. Phone: +32 15 79-5528, e-mail: BVanDijk@cochlear.com

M.Moonen is with Department of Electrical Engineering, Katholieke Universiteit Leuven 3000, Belgium. e-mail: Marc.Moonen@esat.kuleuven.be

J.Wouters is with ExpORL, Department of Neurosciences, Katholieke Universiteit Leuven 3000, Belgium. e-mail: jan.wouters@med.kuleuven.be

©2012 IEEE

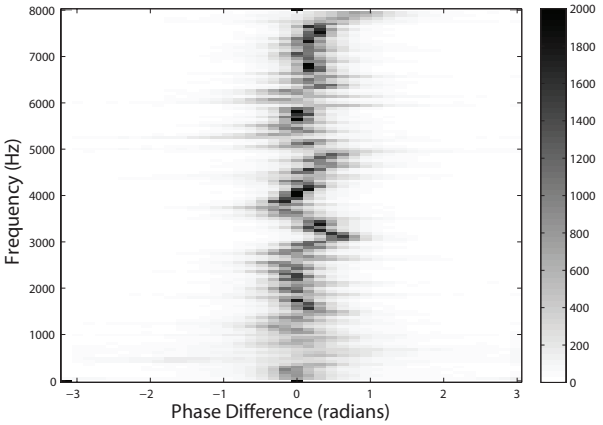


Figure 1. Phase error histogram for a target speech signal (10 sentences of LIST) in front of a KEMAR (45BA) manikin in low reverberation with RT60=0.1 seconds. Phase error is wrapped between -π to π radians and is shown along abscissa. The color bar represents the number of time frequency bins having depicted frequency and phase error.

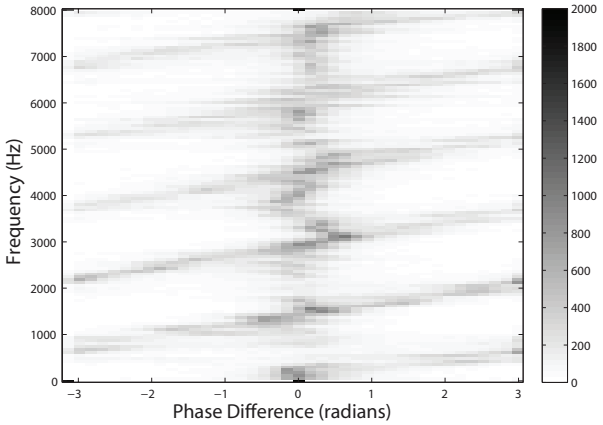


Figure 2. Phase error histogram for a noisy signal in the S0N90 scenario in low reverberation and an input SNR of 20 dB. The color bar represents the time-frequency bins having depicted frequency and phase error. The target and the noise source signals have different phase errors distributions depending on the spatial location of the two sources.

filterbank in the form of a cochleagram [16]. A detail review of single channel and multi channel time frequency masking based systems is given in [17]. TF masking systems for speech separation can be divided into feature-based [18], [19] and model-based [16], [20] systems. The feature- and model-based systems produce either binary or ratio (soft) masks. The Ideal Binary Mask (IBM) [21], [3] is often used as a measure of ceiling performance.

In [22] and [23] a phase error based TF masking technique was presented, which rewards or punishes individual (TF) blocks based on the observed and expected phases of those blocks. It was shown that by masking the TF blocks in this way, noise components are suppressed beyond recognition while the speech source of interest maintains its perceptual quality. This was supported by software based speech recognition experiments where Phase Error Filtering (PEF) out performed super-directive beamforming and postfiltering [23]. They used a smooth mask to have fewer distortions. However, speech understanding can be greatly improved by a careful investigation of the mask functions. This especially applies for the CI recipients who not only have a limited frequency resolution and a much smaller dynamic range but also have a severely degraded access to the temporal fine structure [24]. These limitations may not only have the effect of reduced or increased speech understanding performance for a certain masks as compared to NH subjects, but may also have different intelligibility improvement and preference ratings for different masks. It is also important to note that the noise rejection requirements and speech distortion limitations for NH and CI subjects may be significantly different. Additionally, as CI users generally require higher SNR, the algorithms operating point may also be very different. Therefore it is worthwhile to explore the differences in speech understanding performance utilizing different mask functions for NH and CI subjects in realistic spatial scenarios.

This paper is organized into seven sections. In section II the signal model and previously reported PEF algorithm is discussed. Considering the limitations and possible requirements of CI recipients three mask functions are defined along with the originally proposed mask function in section III. Instrumental performance measures and simulation results are given in section IV. Section V describes the perceptual measure used, evaluation procedure and the subjective results. CI and NH subjective results are discussed and analyzed in sections VI and a conclusion is drawn in the last section.

II. SIGNAL MODEL AND PREVIOUS WORK

Let us assume we have two Behind The Ear (BTE) devices placed on each ear. There is one Omni-directional microphone on each BTE device. The received signal on each microphone can be modeled as

$$x_r(t) = s(t) \circledast h_{sr}(t) + n_1(t) \quad (1)$$

$$x_l(t) = s(t) \circledast h_{sl}(t) + n_2(t) \quad (2)$$

in discrete time domain, equivalently,

$$X_r(k,b) = S(k,b) \times H_{sr}(k,b) + N_1(k,b) \quad (3)$$

$$X_l(k,b) = S(k,b) \times H_{sl}(k,b) + N_2(k,b) \quad (4)$$

where h_{sr} and h_{sl} are the impulse responses associated with the target speech source for the right and left microphones, respectively. $x_r(t)$ and $x_l(t)$ are the signals obtained by the two microphones, $s(t)$ is the target speech signal and $n_1(t)$ and $n_2(t)$ are the noise source signals respectively. A spatial scenario with target speech signal (s) arriving from an angle x and one or more noise sources (N) arriving from angle(s) y is given as SxNy. The angles are defined clockwise with zero being in front of the subject. The received signals in such a spatial scenarios also incorporates the impulse responses associated with the noise source for the right and left microphones. In [22] and [23] a phase error based TF masking algorithm is presented. It is shown that Time Difference Of Arrival (TDOA) estimation for the target speech signal using phase transform (PHAT) is a phase error minimization technique in essence.

The phase error in b^{th} time frame and k^{th}frequency bin can be given as

$$\theta_\tau(k,b) = \angle X_r(k,b) - \angle X_l(k,b) - 2\pi k\tau \tag{5}$$

where τ is the target speech signal delay between the two microphones and can be estimated using the generalized cross correlation method. The value of τ is assumed to be zero for the target speech source in front at 0 degrees azimuth. The phase error is always wrapped between -π to π . The phase error histograms for a target speech signal at 0 degrees is shown in figure 1 in contrast to a noisy signal in the S0N90 spatial scenario, shown in figure 2. It is evident that in noise-free case almost all the phase error points are aligned along the zero radians, while in a noisy case the phase error is much more spread in the range -π to π . Ideally, in noise free environment, the phase error should be equal to zero. However, due to the presence of reverberations and effects of finite duration time window, the phase error is nonzero. The spread of the phase error can be estimated using the Mean Phase error Variance (MPV) and is given as

$$\Psi = \frac{1}{(B \times K)} \Sigma_{(k=1)}^{K} \Sigma_{(b=1)}^{B} \theta_{(k,b)}^2 \tag{6}$$

Where B is the total number of time frames and K is the total number of frequency bins. For the signal model given in equation 1 and 2 a simulation with the Leuven Intelligibility Sentence List (LIST) [25] was performed with white Gaussian noise (WGN) and multi-talker babble noise, under different values of mixing SNRs. Figure 3 clearly shows that MPV is a function of SNR, and increases when the SNR is decreased. In [22] and [23] a phase error based reward-punish criterion for noise reduction was proposed. TF blocks with large phase error are scaled down while keeping low phase error blocks intact using a mask function ζ. The processed signal at each ear can be obtained by the following equations

$$\hat{S}_r(k,b) = \zeta(k,b) X_r(k,b) \tag{7}$$

$$\hat{S}_l(k,b) = \zeta(k,b) X_l(k,b) \tag{8}$$

The algorithm provided significant improvement in a digit recognition task, in anechoic and low reverberant environment, over the super directive beamforming and postfiltering. Comparison of perceptual results with different mask functions for NH and CI subjects still needed to be investigated. For cochlear implant and hearing aid applications the complete binaural processing scheme is shown in figure 4. The mask calculation block needs only phase information from the DFT block and calculates a mask which is multiplied with the received microphone signals at the left and right ear.

III. MASKING FUNCTIONS

To study the effects of different types of masking functions we defined three different types of masks along with the originally proposed mask function. Descriptions and properties of these masks are given as follows.

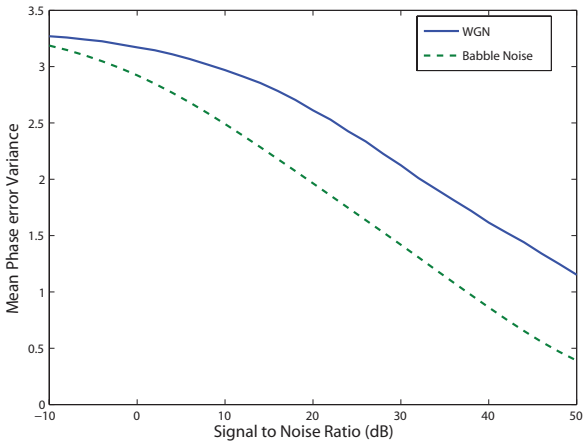


Figure 3. Relationship between Mean Phase Variance and input SNR. The solid line depicts the relationship for white Gaussian noise (WGN) whereas the dotted line is for multi-talker babble noise.

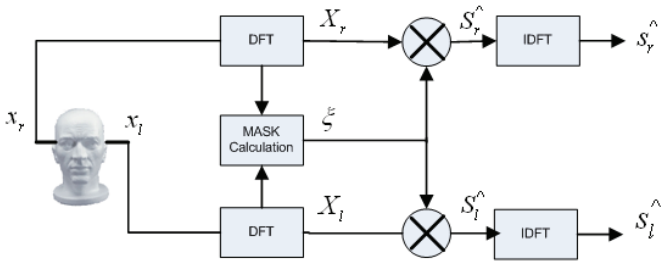


Figure 4. Processing scheme for a Binaural system for bilateral input. The mask is estimated based on the phase information from the DFT blocks and is applied to both left and right ear signals.

A. Soft Mask (SMASK)

A parametrized soft mask function was originally proposed by the Aarabi et al., [22] and is given by

$$\eta_{(k,b)} = \frac{1}{(1 + \gamma\theta_{(k,b)}^2)} \tag{9}$$

where larger values of γ punish large phase error time frequency bins more severely whereas smaller values insure minimum distortion. Depending upon the value of γ the mask allows some level of noise as well and tries to maximize the output SNR. The optimum value of the γ changes with the input SNR. The influence of γ is studied in detail in [23] The value of γ is fixed and chosen to be one for this study. This value insures fewer distortions in the speech segments but also provides less noise reduction. This value of γ also insures that even at very high input SNR the output signal is not degraded due to the masking. With the defined value of γ the mask function becomes robust against head movements and impulse response variations.

B. Binary Mask (BMASK)

The binary mask is defined below

$$\beta_{(k,b)} = 1 \quad for \;\; |\theta_{(k,b)}| <= \Delta_\beta \;\; radians$$

$$\beta_{(k,b)} = 0 \quad for \ \ |\theta_{(k,b)}| \ \ > \Delta_\beta \ \ radians \tag{10}$$

Where Δ_β is the allowed deviation of phase error. Larger values of the Δ_β parameter provide less noise reduction but also introduce fewer distortions in the target speech. Simulations in realistic reverberant environments show that around 90 % of the TF bins have a phase error less than 0.7 radians for target speech signals in front, therefore Δ_β is chosen to be 0.7 radians.

C. Ternary Mask (TMASK)

The ternary mask not only rewards the smaller phase error TF bins but also allows some level of noise with the parameter α . Alpha was chosen as 0.1, so that large phase error TF bins are attenuated by 20 dB's at maximum. Δ_{T1} was set to 0.5 radians to have more noise reduction and Δ_{T2} was set to π radians.

$$\begin{aligned} T_{(k,b)} &= 1 \ \ for \qquad \theta_{(k,b)} <= \Delta_{T_1} \ \ radians \\ T_{(k,b)} &= \alpha \ \ for \ \ \Delta_{T1} < \theta_{(k,b)} <= \Delta_{T2} \ radians \\ T_{(k,b)} &= 0 \ \ for \qquad \theta_{(k,b)} \ > \ \Delta_{T2} \ \ radians \end{aligned} \tag{11}$$

D. Gaussian Mask (GMASK)

Binary and ternary mask functions involve discrete steps in the gain function. In order to avoid such steps and at the same time to completely remove all the large phase error TF bins, a Gaussian mask function was used which is given as.

$$\begin{aligned} G_{(k,b)}(n) &= e^{-\frac{1}{2}(\kappa \frac{n}{N/2})^2} \\ where -\frac{N}{2} &<= n <= \frac{N}{2} \\ \kappa >= 2 \quad &, \quad L = N+1 \end{aligned} \tag{12}$$

Value of κ is set to 9. This value of the parameter provides similar noise reduction as the ternary mask, but completely removes the large phase error points. In contrast ternary mask does allow some level of all the phase error points. All masking functions with the defined parameters are plotted in figure 5. Among these fours mask, SMASK is the smoothest mask and provides minimum noise rejection, whereas GMASK and TMASK are expected to provide maximum noise reduction but have different characteristics. BMASK is a compromise between the two extreme conditions i.e maximum noise rejection and minimum distortion.

IV. Simulations

Simulations were carried out to evaluate and compare the noise reduction and speech enhancement capabilities of different mask functions. Leuven Intelligibility Sentence Test (LIST) material [25] is used as the target speech material. Time shifted versions of multi-talker babble noise (Auditec, St. Louis, MO, USA) are used as the noise source signal n_1 and n_2. The room environment conditions are simulated with Head Related Transfer Functions (HRTFs) measured with Cochlear™ Nucleus 5 BTE devices (CP810) worn by a KEMAR manikin in low reverberant environment with

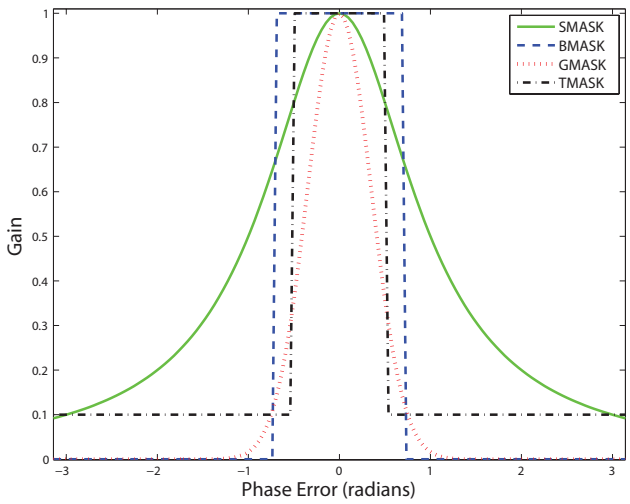


Figure 5. Relationship between different masks and the phase error. Large phase error time frequency bins are attenuated according to these four masks.

RT60=0.1 sec. The impulse responses are convolved with the target speech and multi-talker babble noise (n_1 and n_2) is added to generate two microphone signals at the left and right ear. A window length of 128 samples with 75% overlap is used with a sampling frequency of 16 kHz for PEF.

A. Performance Measures

To quantify the noise reduction offered by different mask functions, an SNR measure is used. SNR is measured by the energy ratio of the target speech signal to the noise signal and is given by

$$SNR_{in} = 10log10(\frac{\Sigma_n|s(n)|^2}{\Sigma_n|n_1(n)|^2}) \tag{13}$$

$$SNR_{out} = 10log10(\frac{\Sigma_n|\hat{s_r}(n)|^2}{\Sigma_n|\hat{n_1}(n)|^2}) \tag{14}$$

where n is the time index. In SNR measurement the target speech and noise components are respectively calculated when only target speech or noise signal is active. In equation 14, $\hat{s_r}(n)$ is the estimated target speech signal obtained using equation 7 when no noise signal is present while the $\hat{n_1}$ is the estimated signal when no target speech is present. Signal to Noise Rejection Ratio (SNRR) can be given as the difference of equation 13 and 14 and quantifies how much noise power is reduced with respect to the signal power. Distortions in the enhanced signal are evaluated through the Signal to Distortion Ratio (SDR) measure, and is given by

$$SDR = 10log10(\frac{\Sigma_n|s(n)|^2}{\Sigma_n|s(n) - \rho\hat{s_r}(n-L)|^2}) \tag{15}$$

where

$$\rho = \frac{\Sigma_n|s(n)|^2}{\Sigma|\hat{s_r}(n-L)|^2} \ , \ L = argmax_L \frac{1}{N}\Phi s\hat{s_r}(L)$$

The target speech and processed signals are time aligned and scaled to the same level before the calculation of SDR. In equation (15) ρ is the scaling factor, L is the time lag and

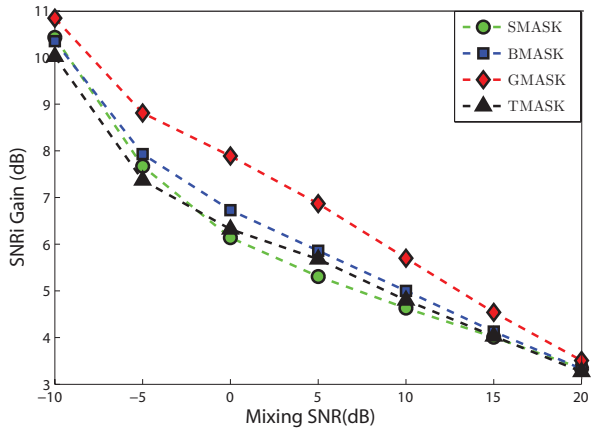


Figure 6. Intelligibility weighted SNR improvement for different values of input SNR.

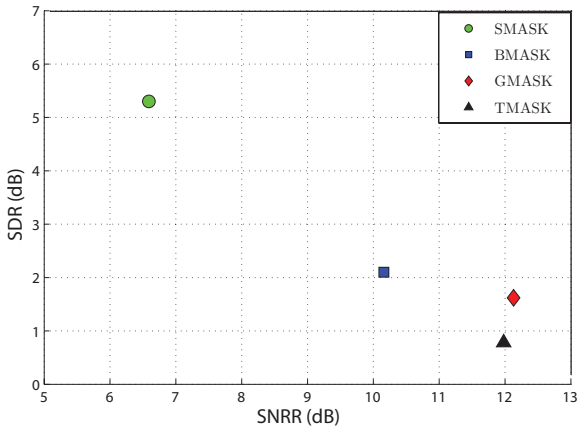


Figure 7. Signal to noise rejection ratio vs signal to distortion ratio for different masks at an input SNR of 0 dB.

$\Phi_{s\hat{s}}$ is the cross correlation between the processed and the reference target speech signal.To access speech intelligibility enhancement capabilities, speech intelligibility weighted signal-to-noise ratio (SNRi) as defined in [26] is used. The weight w_j emphasizes the importance of the j^{th} frequency band's overall contribution to intelligibility, and SNR_j is the signal to noise ratio of the j^{th} frequency band. Band definitions and individual weights of the J frequency bands are given in [27]. Improvement in SNRi can be given as below.

$$\Delta SNR_i = \Sigma_{j=1}^{J} w_j (SNR_{j,out} - SNR_{j,in}) \qquad (16)$$

B. Simulation results

Intelligibility weighted SNR gain for different masks versus input SNR are plotted in figure 6 for the signal model described in section II. The simulation results are averaged over 10 lists of the target speech material. Although all four mask functions provide similar improvement in intelligibility weighted SNR, noise rejection performance is significantly different. SNRR and SDR is plotted in figure 7 at an input SNR level of 0 dB. Below this input SNR level an increased SDR is observed, however the trend among different mask functions remains similar. Simulation results clearly show that the BMASK provides higher SNR gain but at the same time introduces more distortions compared to the SMASK. TMASK and GMASK attenuate the larger phase error TF units more severely than the BMASK as shown in figure 5. This may give high SNR improvements but not necessarily contribute towards intelligibility improvement due to the more distortions they introduce.

V. PERCEPTUAL EVALUATIONS

A. Speech Intelligibility

1) Materials and Methods: For the target speech corpus we again selected the Dutch LIST sentence material which consists of 35 lists of 10 sentences each, from a female speaker. LIST sentence material is particularly suitable for severe hearing impaired and CI subjects. The multi-talker babble noise (Auditec, St. Louis, MO, USA) was used as the

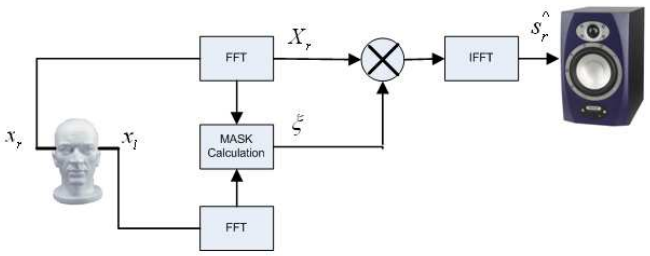


Figure 8. Left and right ear signals are obtained using HRTF's. Only noisy or processed right ear signal is presented to the subject through a loudspeaker.

noise signal. For multiple noise source scenarios, time shifted versions of the same noise source signal were used.

Eight adults with NH thresholds and five CI recipients (S1-S5) using unilateral implants with the ACE strategy as their preferred setting participated in this study. All subjects were native Dutch speakers and CI subjects were paid for their participation. The detailed biographical data for the CI subjects is given in Table I. Unilaterally implanted subjects are targeted in this study, where the phase information from the contra-lateral microphone is used for improved noise reduction. The noisy or processed right ear signal is only presented to the subjects through a loudspeaker in front at 1 meter distance, which is within the critical distance of the test room. This ensures that improvement comes from the processing of the right ear signal and the information from the contra-lateral ear is only used for the calculation of the mask. The testing

Table I
DETAILS ABOUT THE CI SUBJECTS WHO PARTICIPATED IN THIS STUDY

Subject	Age (yr)	Processor	CI Experience
S1	53	Freedom™	15 years
S2	62	Freedom™	11 years
S3	59	Freedom™	07 years
S4	22	CP810™	04 months
S5	60	Freedom™	14 years
S6	53	Freedom™	10 years
S7	62	Freedom™	05 years
S8	59	Freedom™	07 years

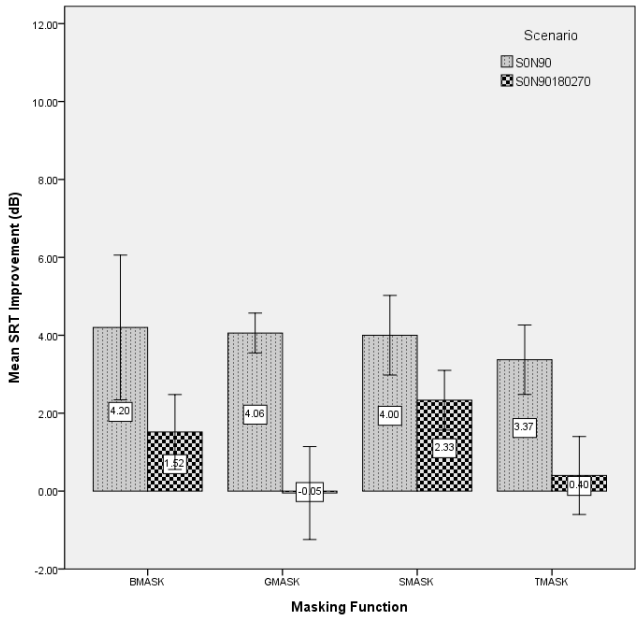


Figure 9. SRT improvements for 8 NH subjects in different spatial scenarios. The error bars represent the 95% confidence intervals. The mean SRT improvements are also shown for each condition.

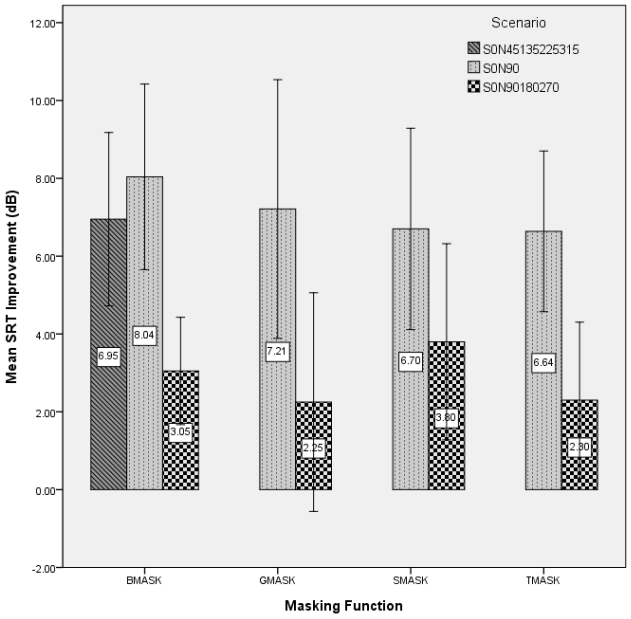


Figure 10. SRT improvements for 5 CI subjects in three different spatial scenarios. The error bars represent the 95% confidence intervals. The mean SRT improvements are also shown for each condition.

scenario is depicted in figure 8. As shown in the figure, the processed signal is re-synthesized and is presented through the loudspeaker in front for both NH and CI subjects.

The tests were carried out with NH and CI subjects to access the intelligibility gain offered by the algorithm with different mask functions. Speech Reception Threshold (SRT) in noise is used as a perceptual measure which is obtained using an adaptive test procedure [28], [29],in which the overall level is fixed at 65dB SPL. Both target speech and noise levels were adjusted to obtain the desired SNR level. The SNR of the first sentence of each list was increased in steps of 2 dB until the sentence was identified correctly. Subsequently, the SNR is varied in steps of 2 dB adaptively in a one-down, one-up procedure to target the 50% intercept. Scoring was done based on the sentence level accuracy. The lists were not repeated during the test. The presentation order was randomized across the sessions and subjects. A training list was presented to the CI subjects prior to the actual tests, to familiarize them with the test procedure.

Three spatial scenarios namely S0N90, S0N90/180/270, S0N45/125/225/315 were implemented using the same HRTFs as used in the simulation section. The noisy signals were processed with the algorithm utilizing all four masks and results were saved for all SNR levels in steps of 2 dB in the range -20 to +20 dB. The SNR at the right ear is defined as the reference SNR.

For processing of the signals a window length of 128 samples with 75% overlap was used to ensure a low latency in real time implementations. A square root hanning window is used for analysis and resynthesis. The signals are presented to the subjects, using a RME Hammerfall Multiface II sound card mounted in a Pentium PC. The output of the sound card drives the Genelec loudspeaker (type 8020A). The room is equipped with curtains to reduce the reverberation time to 0.21 sec. The PC and sound-cards are located in a control room adjacent to the test room. The presentation level of the signals is always 65 dB SPL i.e. the processed signals are scaled to this level. The sentences are presented at a sampling frequency of 16 kHz. One unprocessed and four processed signals are presented to the subject in two spatial scenarios namely S0N90, S0N90/180/270. Number of sources may also have a critical effect on the performance of the algorithm. Therefore Binary mask is also tested with CI subjects only, in the S0N45/125/225/315 scenario. So in total 11 SRTs were calculated per subject per session. Test and retest in two sessions were conducted with CI subjects to access the consistency of the experiments. NH subjects were tested only once with all the 11 test conditions.

2) Results: The mean SRT for NH subjects in the unprocessed noisy condition is -1.9 dB with a standard deviation of 1.5 dB. This is consistent with the previously reported results for the LIST material in babble noise[30]. SRT improvements for NH subjects are shown in figure 9. An ANOVA with repeated measures indicated a significant effect ($F[5,40]=26.11$, $p<0.01$) of the mask functions in the S0N90 scenario. The effect size is significantly reduced ($F[5,40]=3.18$, $p<0.05$) in the S0N90/180/270 scenario. Post-hoc comparisons (according to Scheffe's test) were run to assess significant differences between different mask functions. All four mask functions provide a significant improvement in the S0N90 scenario and there is no significant performance difference. However, in the S0N90/180/270 scenario only the SMASK provides a significant improvement over the unprocessed condition.

The SRT's for the CI subjects in noisy condition with ACE strategy range from 7dB to 14 dB with a mean SRT of 10.43 dB with a standard deviation of 2.1 dB. SRT

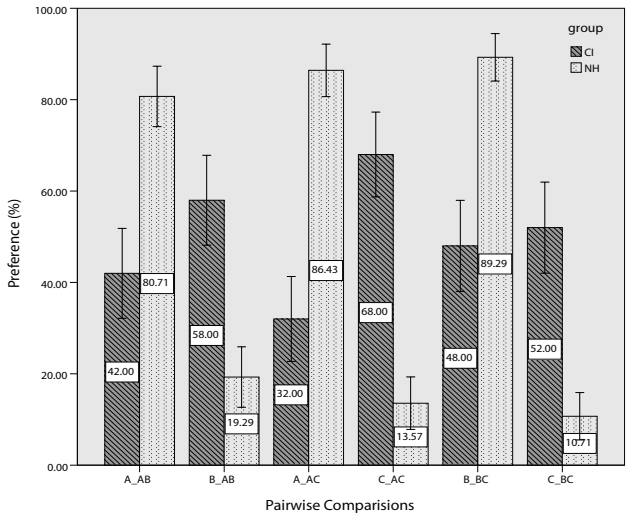


Figure 11. Percentage win-scores for the pair-wise comparison of A) unprocessed, B) SMASK and C) BMASK for normal hearing and cochlear implant test subjects. For each pair-wise comparison the preference for each condition is indicated by the letters A, B and C.

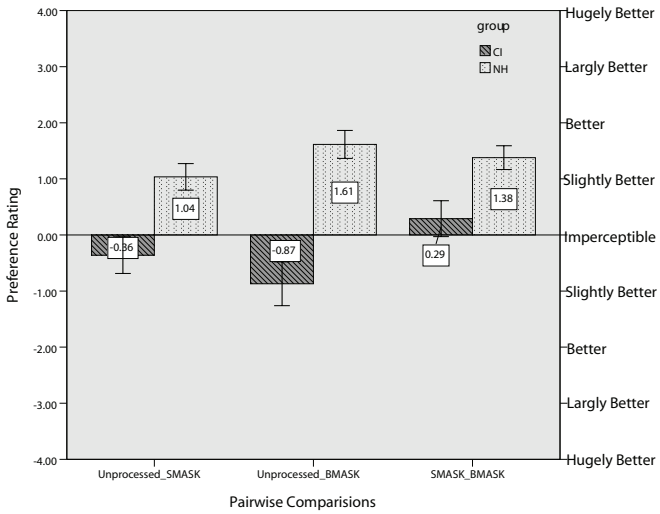


Figure 12. Preference rating scores for each condition for normal hearing and cochlear implant recipients. The preference is indicated by positive scores for the first condition and by negative scores for the second condition in a pair-wise comparison.

improvements for CI subjects are shown in figure 10. A significant effect ($F[5,50]=16.31$, $p<0.01$) of the masking function on the SRT is observed. There is no significant difference between different mask functions in the S0N90 scenario and all mask functions provide a significant improvement over the unprocessed condition. The masking functions also have a significant effect ($F[5,40]=5.17$, $p<0.05$) in the S0N90/180/270 scenario. In this case SMASK as well as BMASK provide a significant improvement over the unprocessed condition. BMASK also provides significant improvement ($p<0.01$) in the S0N45/135/225/315 spatial scenario.

B. Speech Quality

1) Material and methods: In addition to evaluating the intelligibility improvements offered by the different masks, the quality of the processed speech is also evaluated using a two stage preference rating test [31], [32]. The pair-wise comparisons are conducted across (i) unprocessed speech vs. SMASK, (ii) unprocessed speech vs. BMASK and (iii) SMASK vs. BMASK. In this pair-wise comparison analysis NH and CI test subjects are asked to select the condition that has a better sound quality or is less distorted. They further need to quantify their preference with a rating score ranging from 0 to 4 (Imperceptible, Slightly better, Better, Largely better, Hugely better). The unprocessed noisy speech in babble noise in S0N90 scenario is presented at 4 different (10, 15, 20, 40 dB) SNR levels. To reduce the interaction between quality and intelligibility the processed speech is presented at 5 dB below the unprocessed condition's SNR (5, 10, 15, 35 dB). Each evaluation sample consisted of 10 individual sentences from the LIST sentence material, concatenated and played in a continuous loop. For each pair-wise comparison, 5 samples were presented repeatedly, thus leading to (5 repetitions/comparison) $\times$ 3 (pair-wise comparisons) $\times$ 4(mixing conditions) = 60 trials. The subjects can listen to each sample for as long as they wanted before and during the rating process. The rating followed a two step procedure. The first stage is an absolute preference of one sample over the other, whereas the second stage is a subject-dependent qualitative measure of how much one signal is preferred over the other. 10 subjects with NH thresholds and 5 CI subjects (S3, S4, S6-S8) participated in this experiment. The 10 normal hearing test subjects had no self-reported hearing deficit and their age varied from 25 to 50.

2) Results: The percentage win-scores for the pair-wise comparison of unprocessed, SMASK and BMASK are given in figure 11 in which condition A represents the unprocessed noisy signal, B represents the processed speech with SMASK and C corresponds to the BMASK processed speech. This merely represents the percentage of times a given approach is preferred in the pair-wise comparison. The results for NH subjects indicate a clear preference for the unprocessed speech as compared to BMASK or SMASK over the range of tested SNRs. NH subjects also show a clear preference for the SMASK over the BMASK. In Contrast to NH subjects CI subjects show an opposite trend. CI subjects prefer SMASK or the BMASK over the unprocessed speech whereas when SMASK and BMASK are compared CI subjects don't show any preference. There is no significant effect of the SNR over the preference. Figure 12 shows the relative preference rating scores of NH and CI subjects for the two conditions, in which the preference is indicated by positive scores for the first condition and negative scores for the second condition.

VI. Discussion

Although the PEF algorithm provides a significant improvement in SRT for both NH and CI subjects, however, there are some significant differences in speech understanding performance for the two subject groups for the same masking functions. For example, in the S0N90/180/270 scenario only SMASK provides a significant ($p< 0.05$) improvement in

SRT for NH subjects but for CI recipients SMASK and BMASK both provide a significant improvement over the unprocessed condition. BMASK provides more noise rejection and therefore introduces a lot of distortions which seems to degrade the performance for NH subjects in this scenario but not for the CI recipients. No significant difference is found between the SMASK and BMASK for the CI subjects, which is possibly due to the lower susceptibility of CI listeners to the distortions. This less susceptibility to distortions can be attributed to limited frequency resolution and low dynamic range along with the maxima selection process. It is also important to note that the mean improvements among the two subject groups are significantly different. CI subjects on average have 3 dB or more improvement than the NH subjects in the S0N90 scenario. This difference in performance improvement can be due to two reasons. Firstly, because CI subjects have 10-25 dB higher SRT than NH subjects, and the algorithm provides more improvement due to the large variation in phase variance at these SNR levels, which can be seen in figure 3. Secondly, CI subject's insensitivity to distortions as compared to NH subjects.

The results demonstrate not only large improvement in intelligibility with real-time constraints in controlled environments for NH and unilaterally implanted CI subjects. The study also establishes the differences in the perception of speech intelligibility and quality for NH and CI subjects. The literature on TF masking generally shows that binary masking tends to lead to a higher SNR gain but poor quality [33]. However, speech preference rating experiment confirmed the hypothesis that CI subjects are less susceptible to distortions and therefore show no preference for the SMASK over the BMASK in contrast to the NH subjects who have a strong preference for the SMASK or the unprocessed speech where available. CI's less susceptibility to distortions can be due to a number of reasons. The major reasons being (i) a compromised frequency resolution due to a limited number of channels for stimulation and spread of excitation; (ii) severally degraded temporal fine structure information due to the envelope extraction algorithms; (iii) weak temporal pitch cues and (iv) impaired auditory system due to the damage or complete destruction of the sensory hair cells. This lower sensitivity to distortions also offers the opportunity to use more aggressive noise reduction algorithms. The research outcome has the potential implications for future algorithm research for CI. It is evident that the signal processing algorithms that NH prefer may not be ideal for CI recipients as well.

The major techniques for two-microphone (or binaural) speech separation were established by Roman et al. [34] and Yilmaz and Rickard [35]. They utilized characteristic clustering in feature space and used IBM as ground truth for performance evaluation. The segregation system by Roman et al. is motivated by binaural processing of the human auditory system. Yilmaz and Rickard's proposed system is based on the sparsity principle, and suggested a measure of approximate orthogonality. In [34] authors evaluated the performance of their model for two and three source configurations in anechoic conditions. They reported large improvements in percentage correct scores especially at low SNRs. Yilmaz and Rickard's system also gives large improvements in SNR. The direct comparison of the results is not possible due to different test setups; however, PEF provides comparable improvements in SNR and subjective intelligibility. The PEF algorithm can be viewed as a form of fixed beamforming with given directions of arrival where performance does not degrade substantially with the increase in number of sources as long as they are spatially separated. The Roman et al. algorithm uses adaptive beamforming / supervised learning which is somewhat more complex to implement than fixed beamforming in real-time systems [17].

In the S0N90/180/270 scenario the noise source at 180 degrees cannot be cancelled out, due to the same phase error as that of the target speech signal, and therefore limits the improvement for both the subject groups. The algorithm still attenuates the noise from the other two noise sources and therefore a significant improvement in the case of SMASK is observed. GMASK and TMASK do not improve in this scenario; possibly due to the reason that these mask functions introduce relatively more distortion. BMASK which is broader than GMASK and TMASK provides significant improvement in SRT for CI subjects in both spatial scenarios. Binary mask provides relatively less noise reduction but also introduces less distortion as compared to the other two mask functions, and therefore better results are achieved.

Under optimal conditions, conventional two-microphone adaptive beamformer can create one deep null which almost perfectly cancels out one interference [36]. The performance, however, degrades when the number of interfering sources increases. Whereas PEF algorithm also provides considerable improvement in the S0N45/135/225/315 scenario, which shows that performance is not significantly reduced with the number of noise sources as long as TDOA of target speech and noise signals is different. When target and noise signal have same TDOA then performance is significantly reduced which can be seen in the S0N90/180/270 scenario.

BMASK and SMASK provide similar SRT improvements and there is no significant difference between the two mask functions for the CI subjects. However, the performance of the masks is entirely dependent on the phase error distribution, which changes significantly with the head movements and the varying amount of room reverberation. Therefore it is possible that due to the binary decision involved, the BMASK may not be very robust against these two effects as compared to the SMASK.

In this study the left and right ear signals are generated using the HRTFs which were measured using the BTE devices on a KEMAR manikin in a low reverberant environment. The signals are processed off-line and are then presented through a loudspeaker to the subjects. It is also assumed that the target signal is coming from front which may not be true in all realistic listening environments. Therefore, the study does not include the effects of head movements, varying room impulse response, latency and any other issues related with the on-line processing. However, the study demonstrates the potential of the PEF algorithm for speech in noise enhancement specially for unilaterally implanted CI subjects. A future direction for this study is to validate a full on-line version of the algorithm

in realistic listening environments.

VII. Conclusions

The PEF algorithm is evaluated using four different kinds of masking functions in three different spatial scenarios. It is concluded that the PEF algorithm with real-time constraints provides considerable improvement in intelligibility for both NH and CI subjects. It is also found that the same masking functions have significantly different effects on the speech intelligibility and quality when tested with NH and CI subject groups. The study also suggests that CI subjects can accept slightly higher levels of distortions compared to NH subjects, and therefore, more aggressive noise reduction for CI recipients is possible.

Acknowledgment

This research is supported by EU-ITN project AUDIS - Digital Signal Processing in Audiology (http://www.audis-itn.eu) and Cochlear Technology Centre Belgium. The authors are thankful to the test subjects for their patient and enthusiastic participation. We also thank Astrid Van Wieringen, Wim Buyens and Sofie Jansen for their help during the tests.

References

[1] Wilson, B. S. and Dorman, M. F., "Cochlear implants: A remarkable past and a brilliant future", Hearing Research, vol. 242, pp. 3-21, 2008.

[2] Moore, B. C. J., "Coding of sounds in the auditory system and its relevance to signal processing and coding in CI's", Otology & Neurotology,vol. 24, issue 2, pp. 243-254, 2003.

[3] Hu, Y. and Loizou, P. C., "A new sound coding strategy for suppressing noise in cochlear implants", Journal of Acoustical Society of America, 124(1), pp. 498-509, July 2008.

[4] Loizou, P. C., "Signal-processing techniques for CI's", Engineering in Medicine and Biology Magazine, IEEE , vol.18, no.3, pp. 34-46, May-June 1999.

[5] Kiefer, J., Hohl, S., Stürzebecher, E., Pfennigdorff, T. and Gstöettner, W., "Comparison of Speech Recognition with Different Speech Coding Strategies (SPEAK, CIS, and ACE) and Their Relationship to Telemetric Measures of Compound Action Potentials in the Nucleus CI 24M CI System", Audiology, Vol. 40, No. 1, pp. 32-42, 2001.

[6] Parkinson, A. J., Parkinson, W., Tyler, R. S., Lowder, M. W., Gantz, B. J., "Speech perception performance in experienced cochlear-implant patients receiving the SPEAK processing strategy in the Nucleus Spectra-22 cochlear implant", Journal of Speech, Language and Hearing Research 41, 1073-1087, 1998.

[7] Weiss, M., "Effects of noise and noise reduction processing on the operation of the Nucleus-22 CI processor", J. Rehabil. Res.Dev. 30, pp. 117-128, 1993.

[8] Hochberg, I., Boothroyd, A., Weiss, M. and Hellman, S., "Effects of noise and noise suppression on speech perception by CI users", Ear Hear. 13, pp. 263-271, 1992.

[9] Loizou, P. C., Lobo, A. and Hu, Y., "Subspace algorithms for noise reduction in CI's", Journal of Acoustical Society of America, Vol. 118, No. 5, pp. 2791-2793, November 2005.

[10] Yang L. P. and Fu, Q. J., "Spectral subtraction-based speech enhancement for CI patients in background noise", Journal of Acoustical Society of America. Volume 117, Issue 3, pp. 1001-1004, March 2005.

[11] Verschuur, C. A., Lutman, M. E. and Wahat, N. H. A., "Evaluation of a non-linear spectral subtraction noise suppression scheme in CI users", Cochlear Implants International, 7, (4), pp. 193-196, 2006.

[12] J.Wouters and Vanden Berghe, "Speech recognition in noise for cochlear implantees with a two-microphone monaural adaptive noise reduction system", Ear Hear. 22, pp. 420-430, 2001.

[13] Spriet, A., Van Deun, L., Eftaxiadis, K., Laneau, J., Moonen, M., van Dijk, B., Van Wieringen, A., Wouters, J., "Speech understanding in background noise with the two microphone adaptive beamformer in the Nucleus Freedom CI's", Ear & Hearing. 28, pp. 62-72, 2007.

[14] Cochlear Ltd, "Introducing the Audallion BEAMformer digital noise reduction system", Cochlear clinical bulletin, pp. 1-5, April 1997.

[15] Van den Bogaert, T., Doclo, S., Wouters, J., Moonen, M., "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids", Journal of Acoustical Society of America, Vol. 125, No. 1, pp. 360-371, January 2009.

[16] Wang, D. L., Brown, G. J. "Computational auditory scene analysis: Principles, algorithms, and applications". Hoboken, NJ: Wiley/IEEE Press, 2006.

[17] Wang, D. L., "Time-Frequency Masking for Speech Separation and Its Potential for Hearing Aid Design". Trends in Amplification Volume 12 Number 4, 332-353, 2008

[18] Li, P., Guan, Y., Xu, B., Liu, W., "Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech". IEEE Transactions on Audio, Speech, and Language Processing, 14, 2014-2023, 2006

[19] Deshmukh, O. D., Espy-Wilson, C. Y., Carney, L. H., "Speech enhancement using the modified phase-opponency model ". Journal of the Acoustical Society of America, 121, 3886-3898, 2007.

[20] Reddy, A. M., Raj, B., "Soft mask methods for single-channel speaker separation". IEEE Transactions onAudio, Speech, and Language Processing, 15, 1766-1776, 2007.

[21] Hu, G., Wang, D.L., "Monaural speech segregation based on pitch tracking and amplitude modulation ". IEEE Trans. Neural Networks 15 (5), 1135-1150, 2004.

[22] Aarabi, P. and Shi, G., "Phase-based dual-microphone robust speech enhancement", IEEE Transactions on Systems, Man and Cybernetics, vol. 34(4), pp. 1763-1773, August 2004.

[23] Lai, C. Y and Aarabi, P., "Multiple-microphone time-varying filters for robust speech recognition" IEEE International Conference on Acoustics Speech and Signal Processing, pp. I-233-6, 2004.

[24] Marro, C., Mahieux, Y. and Simmer, U., "Analysis of noise reduction and dereverberation techniques based on microphone arrays with post-filtering", IEEE Trans. Speech Audio Processing, vol. 6, pp. 240-259, May 1998.

[25] Van Wieringen, A., Wouters, J., "LIST and LINT: sentences and numbers for quantifying 29 speech understanding in severely impaired listeners for Flanders and The Netherlands",International Journal of Audiology, 30, 47(6), pp. 348-355, 2008.

[26] Greenberg, J. E., Peterson, P. M. and Zurek, P. M., "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance", Journal of Acoustical Society of America, vol. 94, no. 5, pp. 3009-3010, Nov 1993.

[27] Acoustical Society of America "American National Standard Methods for Calculation of the Speech Intelligibility Index", in ANSI S3.5-1997, 1997.

[28] Plomp, R. and Mimpen, A., "Improving the reliability of testing the speech reception threshold 19 for sentences ", Audiology, 18(1), pp. 43-52, 1979.

[29] Nilsson, M., Soli S. D. and Sullivan J. A., "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise", Journal of Acoustical Society of America, vol. 95, pp. 1085-1096, 1994.

[30] Francart, T., Van Wieringen, A. and Wouters, J., "Comparison of fluctuating maskers for speech recognition tests". International Journal of Audiology, 50 (1), pp. 2-13, 2010.

[31] International Telecommunication Union - Radiocommunication Sector, Recommendation, "Subjective assessment of sound quality", BS. 562-3, (1990)

[32] IEEE Subcommittee, "IEEE Recommended Practice for Speech Quality Measurements". IEEE Trans. Audio and Electroacoustics AU-17(3), 225-246, 1969.

[33] Araki, S., Makino, S., Sawada, H., Mukai, R., "Reducing musical noise by a fine-shift overlap-and-add method applied to source separation using a timefrequency mask". In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processings, Vol. III, pp. 81-84, 2005.

[34] Roman, N., Wang, D. L., Brown, G. J."Speech segregation based on sound localization ". Journal of the Acoustical Society of America, 114, 2236-2252, 2003.

[35] Yilmaz, O., Rickard, S., "Separation of speech mixtures via time-frequency masking". IEEE Transactions on Signal Processing, 52, 1830-1847, 2004.

[36] Greenberg, J. E., and Zurek, P. M., "Microphone - arrays hearing aids". in Microphone Arrays: Signal Processing Techniques and Application, edited by M. Brandstein and D.Ward, Springer, Berlin, pp 229-253, 2001.



Obaid ur Reman. Qazi, b.1981, comes from Haripur, Pakistan. He received a B.Eng in Electrical Engineering (2004) with distinction from University of Engineering and Technology, Peshawar, Pakistan, and an MS in Communications and Information Technology (2009) from Karlsruhe Institute of Technology, Germany and Politecnico de Torino, Italy on a European Union funded scholarship. He also received the Marie Curie Scholarship for Doctoral studies under the EU project AUDIS. His general research interests include speech enhancement, speech recognition, statistical signal processing and digital communications. He is currently working as a sound coding researcher at Cochlear Technology Center, Belgium and is also a doctoral candidate at Katholieke Universiteit Leuven, Belgium where he is researching aspects of speech enhancement for CIs under the supervision of Professor Marc Moonen and Professor Jan Wouters.



Marc Moonen (M'94, SM'06, F'07) received the electrical engineering degree and the PhD degree in applied sciences from Katholieke Universiteit Leuven, Belgium, in 1986 and 1990 respectively.Since 2004 he is a Full Professor at the Electrical Engineering Department of K.U.Leuven, where he is heading a research team working in the area of numerical algorithms, signal processing for digital communications, wireless communications and audio signal processing. He received the 1994 K.U.Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with Piet Vandaele), the 2004 Alcatel Bell (Belgium) Award (with Raphael Cendrillon), and was a 1997 "Laureate of the Belgium Royal Academy of Science". He received a journal best paper award from the IEEE Transactions on Signal Processing (with Geert Leus) and from Elsevier Signal Processing (with Simon Doclo). He has served as Editor-in-Chief for the "EURASIP Journal on Applied Signal Processing" (2003-2005), and has been a member of the editorial board of "IEEE Transactions on Circuits and Systems II" (2002-2003) and "IEEE Signal Processing Magazine" (2003-2005) and "Integration, the VLSI Journal". He is currently a member of the editorial board of "EURASIP Journal on Applied Signal Processing", "EURASIP Journal on Wireless Communications and Networking", and "Signal Processing".



Bas van Dijk was born in Amsterdam, the Netherlands in 1968. He received his masters degree in 'computer sciences and physics' in 1993 from the University of Amsterdam. In 1998 he received his Ph.D. from the Biophysics department of the University of Leiden (the Netherlands). From 1998 to 2001 Dr. Bas van Dijk worked as Research Audiologist in the audiology department of the University Medical Centre Utrecht (the Netherlands) were he worked on clinical and scientific aspects of cochlear implant fitting and signal processing. In 2001 he joined Cochlear Ltd, initially as clinical advisor to the R&D teams in the Cochlear Technology Centre Belgium (Mechelen). In recent years, he is responsible for coordinating research projects for Cochlear Ltd. on signal processing worldwide, and in addition he is leader of the research team in Mechelen, working mainly on signal processing, clinical validation, and development of research tools. He (co-)developed many innovations in the commercial cochlear implant system and is the (co-)inventor on 9 international patents and (co-)author of 6 peer-reviewed publications in the field of cochlear implants.



Jan Wouters was born in 1960 and received the physics degree and the Ph.D. degree in sciences / physics from the Katholieke Universiteit Leuven, Belgium, in 1982 and 1989, respectively. From 1989 till 1992 he was a Research Fellow with the Belgian National Fund for Scientific Research (FWO) at the Institute of Nuclear Physics (UCL Louvain-la-Neuve and K.U.Leuven) and at NASA Goddard Space Flight Center (USA). Since 1993 he is a Professor at the Neurosciences Department of the K.U.Leuven (Full Professor since 2005). His research activities center around audiology and the auditory system, signal processing for cochlear implants and hearing aids. He is author of about 190 articles in international peer-reviewed journals and is a reviewer for several international journals. Dr. Wouters received an Award of the Flemish Government in 1989, a Fulbright Award and a NATO Research Fellowship in 1992, the Flemish VVL Speech therapy - Audiology Award in 1996, and an award from the Deutsche Gesellschaft für Audiologie (DGA) in 2007. He is in the editorial board of the International Journal of Audiology, the Journal of Communication Disorders and the journal B-ENT. He is president-elect of the European Federation of Audiological Societies (EFAS), member of the International Collegium for ORL (CORLAS), a board member of the International Collegium for Rehabilitative Audiology (ICRA) and is responsible for the hearing research group at the Division of Experimental ORL (ExpORL), Department of Neurosciences, and teaches 5 courses in the audiology program at K.U.Leuven.