FP7-ICT-2013-C TWO!EARS Project 618075

Evaluation and progress report

Extension to a dynamic binaural model

WP2 *

May 31, 2016

* The TWO!EARS project (`http://www.twoears.eu`) has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 618075.

Project acronym:	Two!Ears
Project full title:	Reading the world with Two!Ears

Work package:	WP2
Document number:	D2.4
Document title:	Extension to a dynamic binaural model, evaluation and progress report
Version:	1

Delivery date:	31th May 2016
Actual publication date:	31th May 2016
Dissemination level:	Restricted
Nature:	Report

Editor(s)/lead beneficiary:	Tobias May
Author(s):	Tobias May, Chungeun Kim, Armin Kohlrausch
Reviewer(s):	Bruno Gas

Contents

1 Executive summary

The goal of the Two!Ears project is to develop an intelligent, active computational model of auditory perception and experience in a multi-modal context. The auditory front-end (AFE) represents the first stage of the system architecture and concerns bottom-up auditory signal processing, which transforms binaural signals into multi-dimensional auditory representations. The deliverable D2.4 documents the final steps performed in WP2 in the first half of year 3 of Two!Ears. These comprise the implementation, documentation and initial evaluation of additional binaural processors which are needed in the analysis of dynamic spatial sound fields. On the one hand, a preprocessing stage is now provided within the AFE which mimics the precedence effect processing. Another processor derives the spatial sound field parameter apparent source width (ASW). Next to this software development and implementation a number of evaluations, focusing on AFE functionality, are included in D2.4. Next to the evaluation of the two new processors, three additional evaluations and analyses are described. First, we demonstrate that replacing the linear peripheral preprocessing by the nonlinear basilar-membrane processor has considerable consequences for the binaural parameters interaural level difference (ILD) and interaural cross-correlation (IACC). Furthermore, in terms of binaural localization, the influence of training, and the use of head movements and the applied strategy are investigated and are shown to be in good agreement with perceptual data. Finally, the use of AFE features for pre-segmentation is analyzed for their ability to enhance noisy speech and is compared with other state-of-the-art algorithms.

2 The auditory front-end framework

This chapter summarizes the newest extensions of the AFE front-end with respect to the two previous deliverables (D2.1, D2.2 and D2.3).

2.1 Apparent source width processor

The functional ASW model consists of various processing stages, including Gammatone filtering, inner hair-cell (IHC) transduction and absolute threshold of hearing (ATH). Given the binaural signal, the model extracts interaural time differences (ITDs), ILDs and interaural coherence (IC), in order to predict ASW. A schematic diagram of the model is shown in Fig. 2.1.

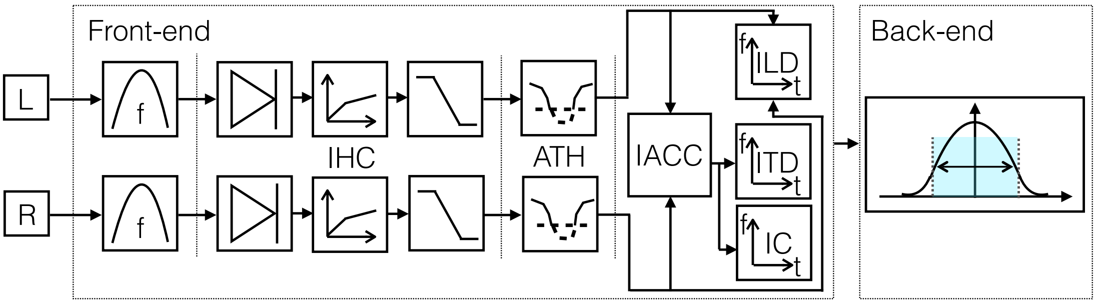


Figure 2.1: Schematic diagram of the binaural ASW model.

2.1.1 Front-end

The binaural signals were first analyzed by a Gammatone filterbank to represent the frequency selectivity of the basilar membrane. The 35 filters were set to a bandwidth of one equivalent rectangular bandwidth (ERB) in the frequency range between 80 to 11891 Hz. In the second stage, the IHC transduction was simulated, i.e. the loss of phase locking to the stimulus fine structure at high frequencies. The IHC processing was performed according to Bernstein *et al.* (1999), suggesting a cut-off frequency of 425 Hz and simulating basilar-membrane compression. In a following stage, the activity in each frequency band was estimated. The signals had been calibrated to a root mean square (RMS) value corresponding to the 70 dB sound pressure level (SPL) of the experimental stimuli. Frequency bands with an SPL below the ATH as defined in Terhardt (1979) were not considered further in the processing. In the last stage, ITDs, ILDs and IC were calculated per time-frequency (T-F) units. The signals of both ears were analyzed in short-time hanning windows of 20 ms duration, with an overlap of 50 %, which resulted in a T-F representation of each ear signal.

2.1.2 Back-end

The ASW estimation was based on the statistical distribution of the binaural cues. The width of this distribution was represented by percentiles and resembled the ASW. Hereby, the left- and right-most boundary of the sound source corresponded to the lower and upper percentile from the distribution's median. The final prediction of the left and right boundaries was then obtained by calculating the mean value across all frequency channels of the lower and upper percentile, respectively.

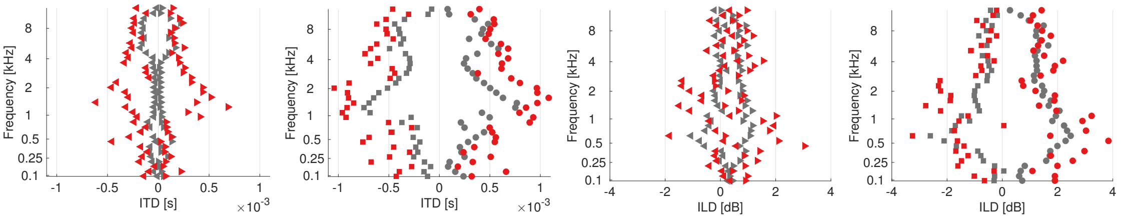


Figure 2.2: ITD-percentiles (left panels) and ILD-percentiles (right panels) as a function of frequency for a pink noise source. Shown are the [30 70] % percentiles (left and right pointing triangles, respectively) and the the [10 90] % percentiles (squares and circles, respectively) for a narrow (gray) and a wide source (red).

Figure 2.2 shows an example of the percentiles [30 70] % (left and right pointing triangles, respectively) per frequency channel for ITDs (left panel) and ILDs (third panel) for a noise source. The percentiles increase from a narrow source (narrow distribution in gray) to a wide source (wide distribution in red), especially for the ITDs. Choosing percentiles that are further away from the median, here illustrated for percentiles [10 90] % (squares and circles, respectively), the values of the ITDs (top right panel) and ILDs (bottom right panel) increase, but their dynamic range, i.e. the difference between the narrow and the wide source, is similar. For the following analysis, the [30 70] % percentiles were chosen to obtain a higher outlier rejection.

2.2 Reconstruction

Given the output of the Gammatone filterbank, it is possible to reconstruct the original input signal by compensating for the frequency-specific delay of the individual subband signals. First, the peak in the envelope domain can be aligned across all subband channels by introducing a frequency-specific time lead (Brown and Cooke, 1994). In addition, a phase compensation factor is necessary to align the peak in the fine structure across channels (Brown and Cooke, 1994). The Gammatone processor in the AFE has been extended to support this phase compensation strategy, which can be activated by the flag `fb_bAlign`. Such a reconstruction stage is particularly relevant in order to evaluate pre-segmentation strategies, as described in section 3.5 and Deliverable D3.2.

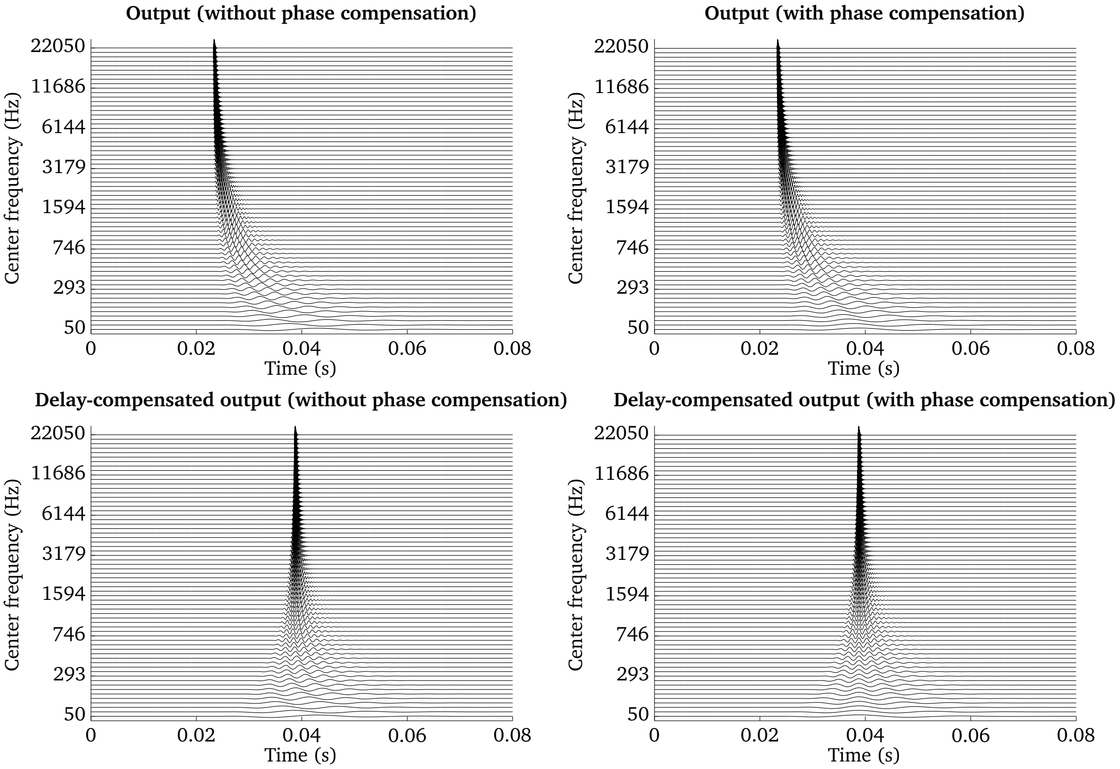


Figure 2.3: Frequency-specific output of a Gammatone filterbank: without phase and delay compensation (top left panel), with phase compensation (top right panel), with delay compensation (bottom left panel) and with delay and phase compensation (bottom right panel).

The impact of these different strategies is visualized in Fig. 2.3 and Fig. 2.4, which have been produced by the script `Demo_Gammatone_Reconstruction.m`. The four panels in Fig. 2.3 show the frequency-specific output of a Gammatone filterbank consisting of 64 filters spaced between 50 and 22050 Hz in response to an impulse located at 23.2 ms. The top left panel shows the output without delay and phase compensation. The impact of the phase compensation can be seen in the top right panel, which ensures that the fine structure of the subband signals is aligned across channels. The additional effect of the delay compensation can be seen in the bottom left and bottom right panels without and with phase compensation.

Finally, the compensation of the time delay can be combined with a frequency-specific gain factor to ensure a flat frequency response of the analysis-synthesis system (Hohmann, 2002). The impulse response and frequency response of three Gammatone-based analysis-synthesis systems is presented in Fig. 2.4. The output refers to a delay-compensated Gammatone filterbank without phase compensation, which produces the largest deviations when comparing it to the response of the input signal. The addition of the phase compensation allows for a reasonably flat frequency response, although the absolute magnitude is shifted with respect to the original input. When combining all three stages, namely the delay, phase and gain compensation, a flat frequency response close to 0 dB can be achieved.

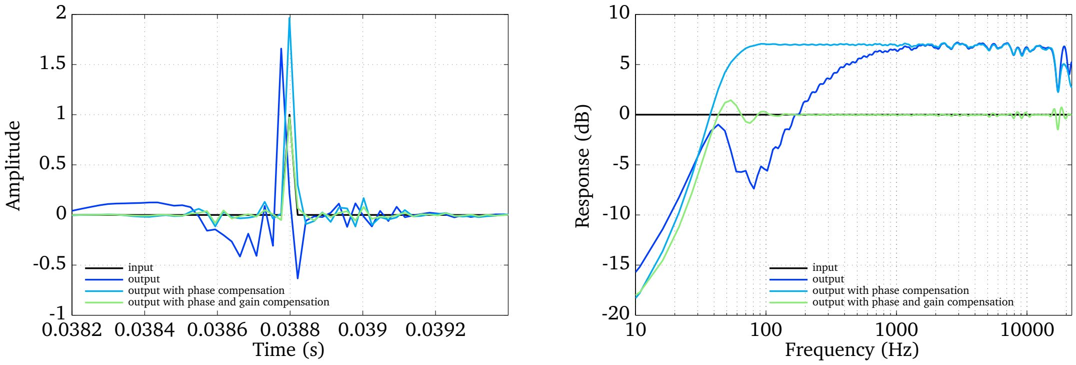


Figure 2.4: Impulse response (left panel) and frequency response (right panel) of an impulse (input signal) and three Gammatone-based analysis-synthesis systems.

2.3 Modification of the precedence effect processor

The precedence effect model in the AFE, integrated from a stand-alone version based on the work of Braasch (2013), was initially described in Deliverable D2.3. In attempts to enable the use of this model in conjunction with the localization Knowledge Sources of the Blackboard system, the initial AFE precedence effect model has undergone a modification. The following subsections describe the details of the modification.

2.3.1 Overview of initial version

The operation of the initial version of the processor was described in D2.3. Assuming that the input to the AFE comprises a delayed repetition (lag) of the original signal, the input to the precedence effect processor is a binaural T-F signal chunk from the Gammatone filterbank. The processor detects and removes the lags from the individual input signals, by means of an autocorrelation mechanism and deconvolution. Then the ITD is derived at all frequency channels from the cross-correlation between the lag-removed pair, and the ILD is also derived at all frequency channels. Afterwards a pair of single ITD and ILD values is calculated as the output, by integrating the ITDs and ILDs across the frequency channels according to the weighted-image model (Stern *et al.*, 1988), and through amplitude-weighted summation.

2.3.2 Details of modification

However, the localization processes implemented within the blackboard system use either ITD and ILD in T-F representation (Gaussian mixture model (GMM)-based localization, see D3.2, section 5.1.1), or cross-correlation (CC) and ILD in T-F-lag and T-F representations (deep neural network (DNN)-based localization, see D6.1.2, section 4.2.1, and D3.4, section 4.1.1). This prohibited the integration of the initial precedence effect processor with the established localization models. Therefore, modifications were made such that:

- the ITD and ILD outputs are now in T-F format, before the integration over frequency channels
- a third parameter CC is added as an output, in the AFE-standardized format (time-frequency-lag signal)

The rest, including the input signal configuration and parameters remain the same. Due to the fact that the precedence effect processor has its own processing strategies to detect and remove lags, the AFE-generic auto- and cross-correlation, ITD and ILD processors could not directly be employed. Instead, as with the initial version, the processor supports three binaural cues as the output, now in different formats.

Fig. 2.5 shows an example of deriving the binaural cues using the modified precedence effect processor, from a binaural narrowband noise signal (800 Hz wide around 500 Hz center frequency, applied ITD = 0.4 ms, ILD = 0 dB, and 3 ms inter-stimulus interval), whose waveform is shown in Fig. 2.5a. Now the three output cues ITD, ILD and CC are in the same format as those of the individual corresponding AFE-generic processors. Although the output format such as in Fig. 2.5b is not supported any more, this can be easily derived from the CC, by finding the maxima over the lag values.

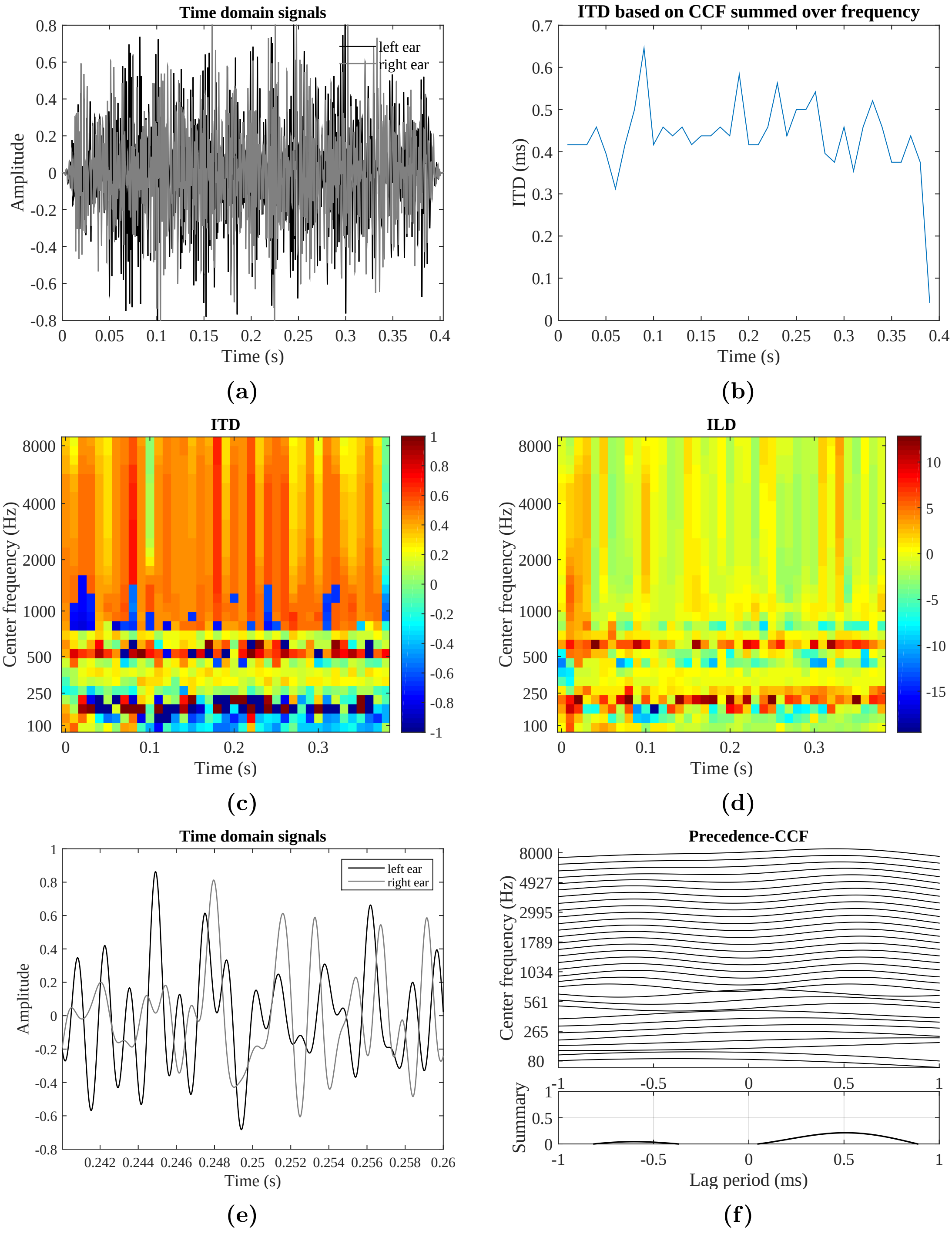


Figure 2.5: Input and output of modified precedence effect processor demo: (a) input signal, (b) ITD integrated over frequency channels over time frames, (c) ITD output as T-F representation, (d) ILD output as T-F representation, (e) input signal for a single frame used for the CCF extraction in (f).

3 Evaluation

In this chapter the AFE is evaluated with respect to the two application scenarios, namely dynamic auditory-scene analysis and quality of experience assessment.

3.1 Influence of level-dependent nonlinearity

Nonlinearity of the basilar membrane operation in response to changing input sound levels is a well-known finding in peripheral auditory processing, as is reflected in the operation of the dual-resonance non-linear (DRNL) filterbank in AFE. This motivates investigations into potential consequences of the level-dependent nonlinearity in terms of variations in binaural cues typically related to spatial perception. Therefore, simulations have been conducted using the AFE software, to derive some binaural cues known to be dependent upon the basilar membrane responses, with a view to revealing the effects of input level caused by the nonlinear basilar membrane operation.

3.1.1 Simulation procedure

Three types of binaural stimuli were used for the simulation: 1-kHz tone, narrowband diotic noise with 1-ERB bandwidth centered at 500 Hz, and the speech signal used for the AFE processor examples in Deliverable D2.3. Various signal levels were applied by means of controlling the presented ILD (referred to as "stimulus ILD" hereafter) around a range of common reference level - 20 to 90 dB reference levels were introduced in 10 dB steps, and for each reference level, 0 to 30 dB stimulus ILDs were applied in 3 dB steps. The ILD was applied such that for example, reference level and ILD pair of (50 dB, 12 dB) means that the left channel signal level is 56 dB and the right channel signal is 44 dB RMS. As the output, ILD and IC were derived using the AFE framework, with the DRNL filterbank model in place of the linear Gammatone filterbank model. These are the internal representations, and therefore to be distinguished from those of the input signals. The following subsections describe the results for the different input stimuli, examined in various ways.

3.1.2 Effect of nonlinearity: 1 kHz tone

Figure 3.1 shows an example of the input stimuli used (in the left panel), and the result of the simulation for various reference levels and stimulus ILDs (in the right panel).

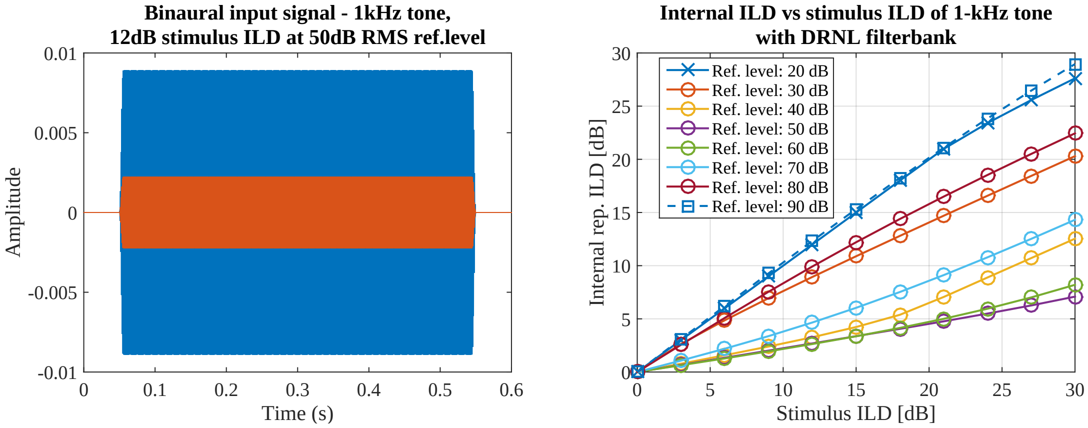


Figure 3.1: Binaural input signal for the simulation with 1 kHz tone (left panel), for the case of 12 dB stimulus ILD applied at 50 dB RMS reference level, and the ILDs derived from the AFE internal representaion for the tested stimulus ILDs and reference levels (right panel).

The effect of the compressive nonlinearity of the DRNL filterbank is seen, in that the internal representation ILD becomes reduced, compared to the stimulus ILD, which would not happen if the Gammatone filterbank is used. More specifically, when the reference level is low at 20 dB, the compression of ILD from the stimulus to the internal representation is rarely seen, indicating that both the left and right signal levels are not yet large enough to reach the range where the basilar membrane is compressive. As the reference level increases from 20 dB to about 50 dB, the internal representation ILD becomes more compressed from the stimulus ILD, because both signal levels eventually reach the basilar membrane compressive region. However, when the reference level increases further from 50 dB, a release from the compression is observed. This is because depending on the applied ILD, either or both signal levels increase outside the mid-level compressive nonlinear region. Overall these findings imply that the ILDs actually perceived internally can be different from the one presented at the ears, because of the nonlinear operation of the basilar membrane.

3.1.3 Effect of nonlinearity: Bandpass noise

Figure 3.2 shows an example of the input stimuli used (in the left panel), and the result of the simulation (in the right panel), in terms of the IC instead of the ILD as in the previous section which showed the same tendency. Here, the IC was initially derived over time for the 500 Hz frequency channel, and then averaged to display a single value per a (reference level - stimulus ILD) pair for easy visual comparison.

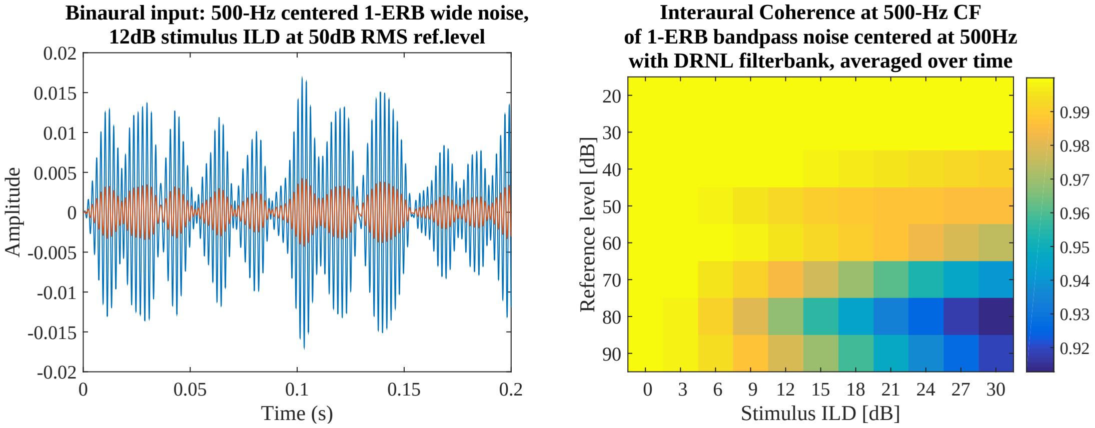


Figure 3.2: Binaural input signal for the simulation with the 500 Hz centered 1-ERB wide bandpass noise (left panel), for the case of 12 dB stimulus ILD applied at 50 dB RMS reference level, and the IC derived from the AFE internal representation for the tested stimulus ILDs and reference levels (right panel).

It can be seen that for low reference levels, increasing the stimulus ILD does not result in any noticeable change in the IC. For reference levels above 30 dB, however, increase in the stimulus ILD with a given reference level results in an overall decrease in the IC. At the highest reference level, the overall IC values increase slightly from those for the 80 dB reference level, implying the release from the basilar membrane compression at high signal levels. It is known from a number of previous studies that the just noticeable differences (JNDs) of interaural correlation can be as low as in the order of 0.01, although they vary widely depending on the source type, frequency, level, and the starting reference value of correlation for comparison (Kim *et al.*, 2008). Therefore, the decrease in the IC, as a result of simply varying the signal levels, may not be perceptually negligible.

3.1.4 Effect of nonlinearity: Speech

Finally, Fig. 3.3 shows an example of the speech input stimuli, which have been used for the AFE processor example scripts in D2.3. Following the level convention (RMS value of 1 in the signal corresponds to 100 dB SPL), the RMS levels of both channels are 51.3 and 55.4 dB for the left and right channels, respectively. For this simulation, the raw IC is displayed in Fig. 3.4 rather than the time-averaged version, considering the irregular fluctuation of the signal over time. In particular, the simulation output with the DRNL filterbank in the left panel is drawn against the simulation output with the Gammatone filterbank in the right panel, for direct comparison of the two models in terms of the internal IC.

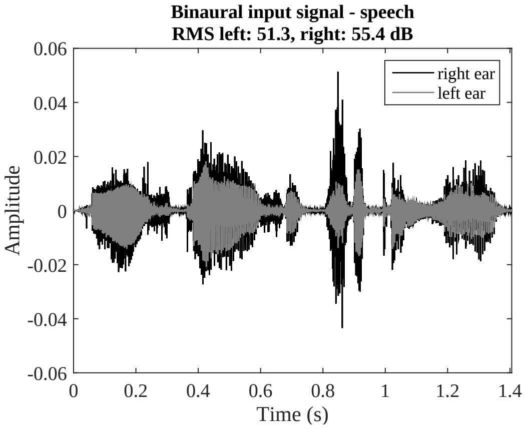


Figure 3.3: Binaural input signal for the simulation with the speech as used in the AFE examples in D2.3. The signal levels are calculated following the level convention of the framework.

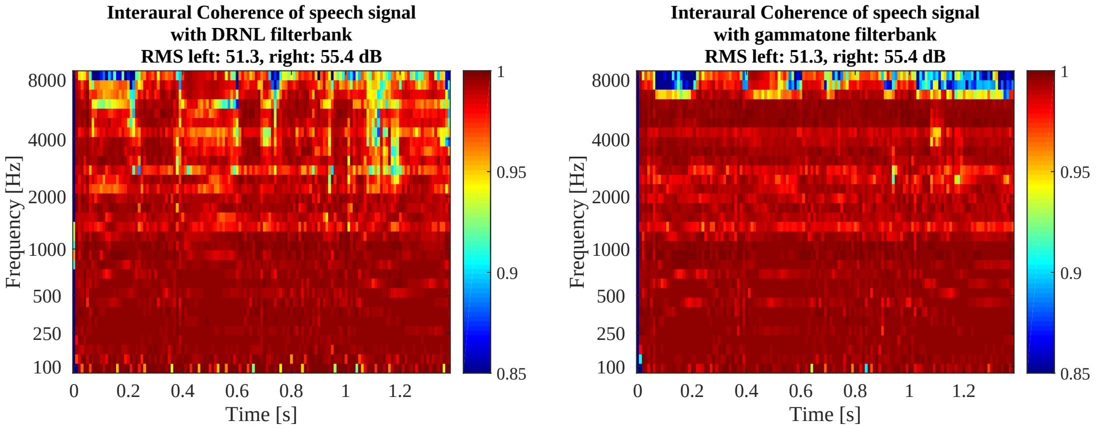


Figure 3.4: The ICs derived from the AFE internal representaion for the tested speech stimulus of Fig. 3.3, using the DRNL filterbank (left panel), and using the Gammatone filterbank (right pannel) as the basilar membrane model.

Comparing the two plots of Fig. 3.4 shows that again the IC fluctuates to a wider extent when the DRNL filterbank is used, in comparison to the Gammatone filterbank. The reduction of the IC over time is irregular due to the fluctuation of input signal which may or may not be compressed by the nonlinear basilar membrane operation. However, also in this case the difference in the IC between the two cases can be perceptually noticeable, considering the range of the JNDs.

3.1.5 Conclusions

As expected from the input-output characteristics of the DRNL filterbank, the ILDs derived from the internal representation were in general smaller than the stimulus ILDs. The amount of ILD reduction from the stimuli to the internal representation depended on the reference level and the stimulus ILD, which finally determine the amount of compression by the basilar membrane in either ear. It was also observed that the IC is affected by the stimulus ILDs alone, which is not observed with linear peripheral models. The findings suggest that the aspects of spatial perception which depend upon these binaural cues need to incorporate this signal-level dependence, either in the scene analysis or in the quality evaluation case. These simulation results motivate further perceptual investigations to reveal how the human auditory system actually copes with these level-dependent variations in binaural cues; whether these are perceived as such with potentially reduced resolution, or compensated at another stage of auditory processing. One such example in the literature is the level dependence of time-intensity trading ratios found with short pulses (Deatherage and Hirsh, 1959). A similar experiment is currently under progress to investigate the interaural time-intensity difference trading ratio in lateralization of auditory signals, at varying overall presentation level, whose outcomes will be reported later in the final deliverable under WP6.

3.2 Sound source localization

Sound source localization is achieved by learning a mapping between source direction and the statistical distribution of binaural cues, namely ITDs and ILDs, that are extracted by the AFE. To increase the robustness to room reverberation and competing sources, the following multi-conditional training (MCT) procedure is applied (May *et al.*, 2011, 2013, 2015b):

1. Mix an anechoic target signal located at a particular azimuth angle φ_k with diffuse noise from all directions

2. Extract a binaural feature vector $\vec{x}_{t,f}$ consisting of ITD and ILD at time frame t and frequency channel f

3. Model binaural feature vector by a GMM λ_{f,φ_k}

The dynamic localization model is equipped with a hypothesis-driven feedback stage which can trigger head movements in case when the sound source direction cannot be unambiguously estimated (May *et al.*, 2015b, Ma *et al.*, 2015a,b). The head movement strategy is illustrated in Fig. 3.5 and assumes that the number of active sound sources is known *a priori*. Given an initial posterior distribution of the sound source azimuth (top left panel), the number of local peaks across a pre-defined threshold are identified. If this number of local peaks is larger than the *a priori* known number of active sound sources, the azimuth information is assumed to be ambiguous, and consequently, a head movement strategy is performed. The second posterior distribution after a 20 ° head rotation produces again two prominent peaks at different azimuths (top right panel). Assuming stationary sound source positions, the initial posterior distribution and the head rotation azimuth can be used to predict the azimuth distribution after head movement. This is done by aligning the two posterior distributions according to the head rotation angle. If a peak in the initial posterior distribution corresponds to a true source positions, then it should have moved towards the opposite direction of the head rotation and will appear in the second posterior distribution. On the other hand, if a peak is due to a *phantom source* as a result of front-back confusion, it will not appear at the same position in the second posterior distribution. By exploiting this relationship, potential phantom peaks are eliminated from both posterior distributions, as illustrated in the bottom left panel of Fig. 3.5. The final localization is performed based on the average of both posterior distributions, as shown in the bottom right panel.

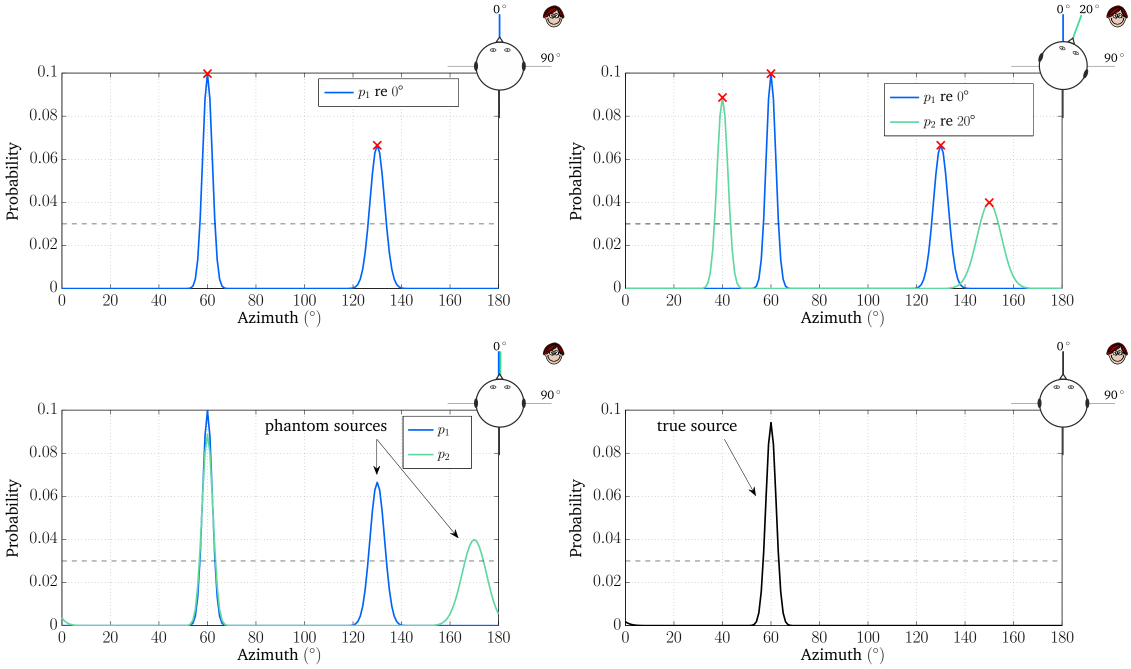


Figure 3.5: Illustration of the head movement strategy. Top left) posterior probability where two candidate azimuths are identified above the threshold. Top right) posterior probability after head rotation by 20 °. Bottom left) Align posterior distributions and remove phantom peaks. Bottom right) Averaged posterior indicating true sound source location.

3.2.1 Influence of multi-conditional training

The localization accuracy is shown in Fig.3.6 for various sets of binaural room impulse responses (BRIRs) when localizing one, two and three competing speakers. When the localization model was trained with clean ITDs, the gross accuracy was only around 30%. The joint evaluation of ITDs and ILDs improved performance considerably, in particular in anechoic conditions. When using the MCT approach, the system was substantially more robust in multi-talker scenarios and in the presence of room reverberation.

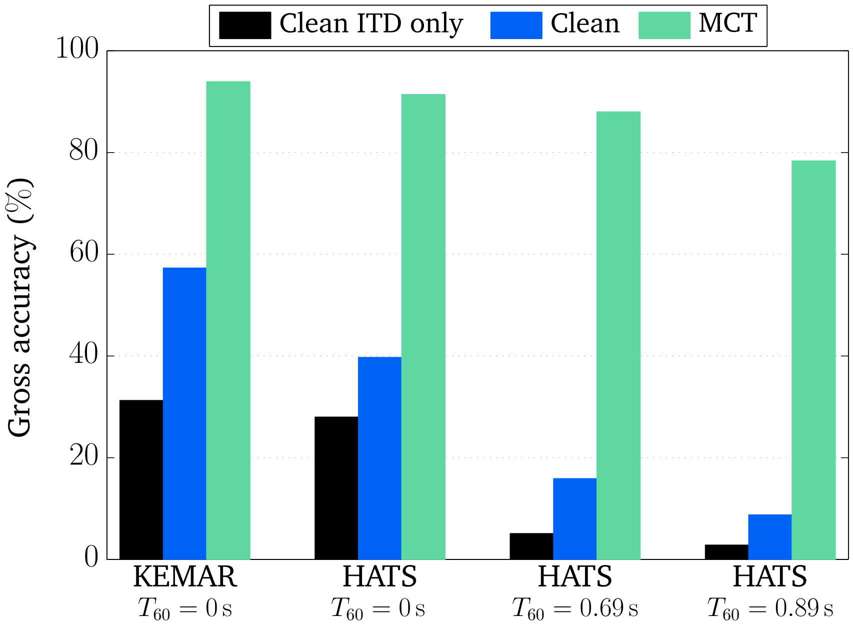


Figure 3.6: Gross accuracy in % for various sets of BRIRs when localizing one, two and three competing speakers.

3.2.2 Influence of random head movements

To quantify the influence of random head movements, the percentage of quadrant errors is presented in Fig. 3.7 for the three previously tested localization models. It is apparent that the head movement strategy systematically reduced the amount of front-back confusions for all models. This indicates that head rotations provide complementary cues that can be effectively exploited by the localization model to disambiguate sources positioned in the front and in the rear hemifield.

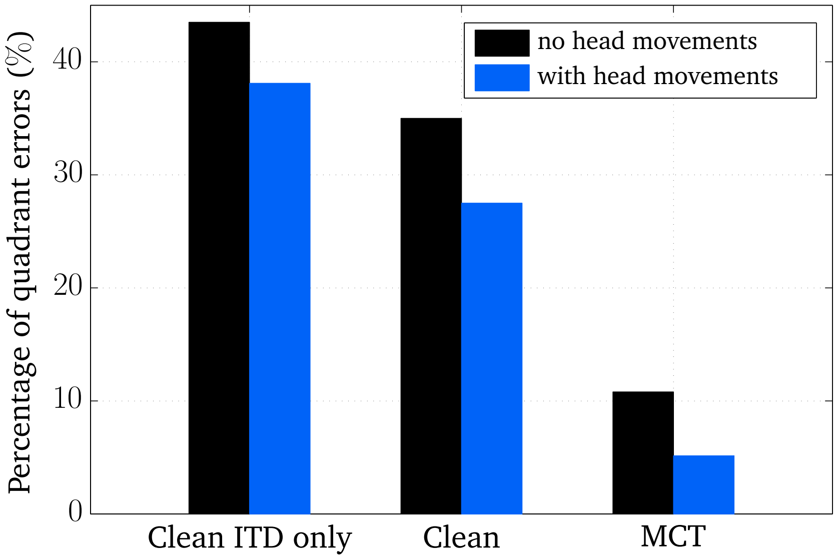


Figure 3.7: Percentage of quadrant errors for the three localization models with and without head movements averaged across rooms and the number of speakers.

3.2.3 Influence of head movement strategy

As opposed to random movements, listeners tend to move their head towards the position of a sound source for improved localization (Perrett and Noble, 1997). Therefore, the following strategies were evaluated: (1) no movement; (2) random movement; (3) rotate exactly to the location of the most likely (ML) source.

The RMS-error is shown in Fig. 3.8 for these three strategies for two different signal durations (0.5 and 2 s). It can be seen that the signal duration did not have a strong effect for the “No movement” baseline. However, both head rotation systems benefited greatly from having longer signals for localization. Clearly, the best performing strategy was to rotate the head towards the most likely source direction.

3.2.4 Conclusions

The MCT of binaural cues was shown to allow for robust localization in the presence of multiple competing sound sources. Furthermore, the approach generalizes to unseen head-related transfer functions (HRTFs), unseen rooms and unseen number of target sources. In addition, an effective head movement strategy was implemented which substantially

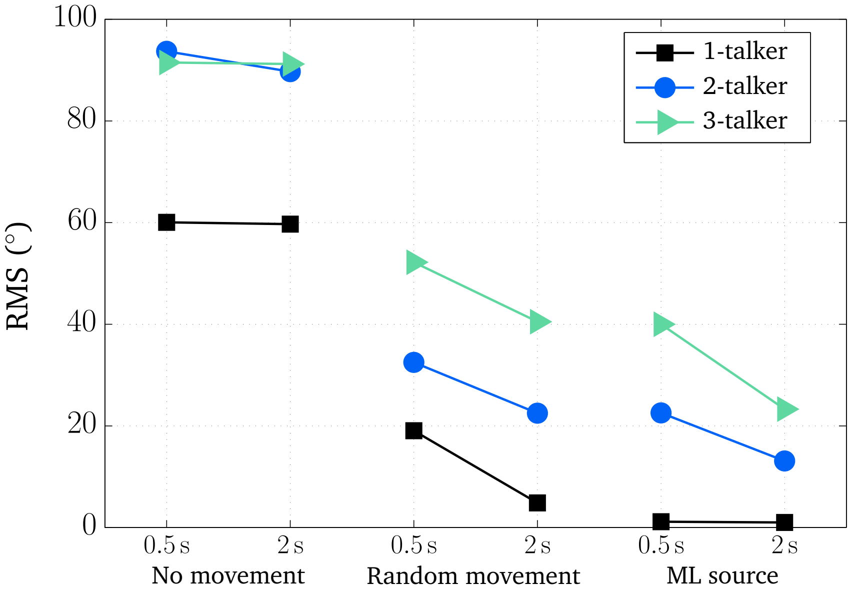


Figure 3.8: RMS-error in degrees of three localization systems that exploit either no rotation, a random rotation or a rotation towards the most likely (ML) source position by 60 °. Results are shown for two different signal durations (0.5 s and 2 s).

reduced the amount of front-back confusions. Finally, it was shown that rotating the head towards the most likely source position is most effective strategy and consistent with perceptual studies.

3.3 Precedence effect

The performance of the modified precedence effect processor was tested further using other types of input signals than described earlier in chapter 2.

3.3.1 Test with anechoic speech and synthesized lag

Firstly, the precedence effect processor was run with a speech signal with an artificially created lag. More specifically, the anechoic speech signal in the AFE IC example of D2.3 was used as the direct sound. Then a lag was created by shifting the left and right signals, suppressing and delaying them by 3 dB and by 4 ms. This lag was added to the original signal to create the input signal. Figure 3.9a shows the resultant input to AFE. The plots of Fig. 3.9 are drawn in the same manner as Fig. 2.5. Although the ITD and ILD plots as T-F representation vary over time and frequency, making it difficult to observe a clear tendency, Figs. 3.9b and 3.9f show that the processor detects the ITD to be around 0.3 ms.

3.3.2 Test with reverberant speech

Secondly, the precedence effect processor was run with the reverberant speech signal in the AFE IC example of D2.3. Figure 3.10a shows the input to AFE, and the rest of Fig. 3.10 are drawn in the same manner as Fig. 3.9. It is now seen that the ITD and ILD in T-F representation fluctuate more. Nevertheless, Figs. 3.10b and 3.10f show that the the ITD is around 0.3 ms, similar to the results of Figs. 3.9b and 3.9f. Although this is based on frame-based direct calculation, adequate low-pass filtering can be applied to remove the fluctuations for better localization.

3.3.3 Conclusions

The modified precedence effect processor was tested with two versions of speech-based input signals, in the same manner as in the example in chapter 2. In both cases, the ITD based on the CCF summed over frequency channels seems to have been detected in the right range, although with some errors. The T-F representations in general made it less clear to conclude single ITD and ILD over time, due to the variation of the input signal across time and frequency. The precedence effect processor seems to be reasonably robust for these two cases, which can be considered possibly more realistic than the initial example using bandpass noise. However, further validation is desirable with a variety of more realistic signals - with larger ranges of number of reflections, lag arrival time, and amplitudes for instance. Also, more tests in conjunction with the localization modules in the blackboard system will reveal the usefulness of this modified precedence effect processor in various scene configurations including room characteristics, number of sources, and tasks to be performed.

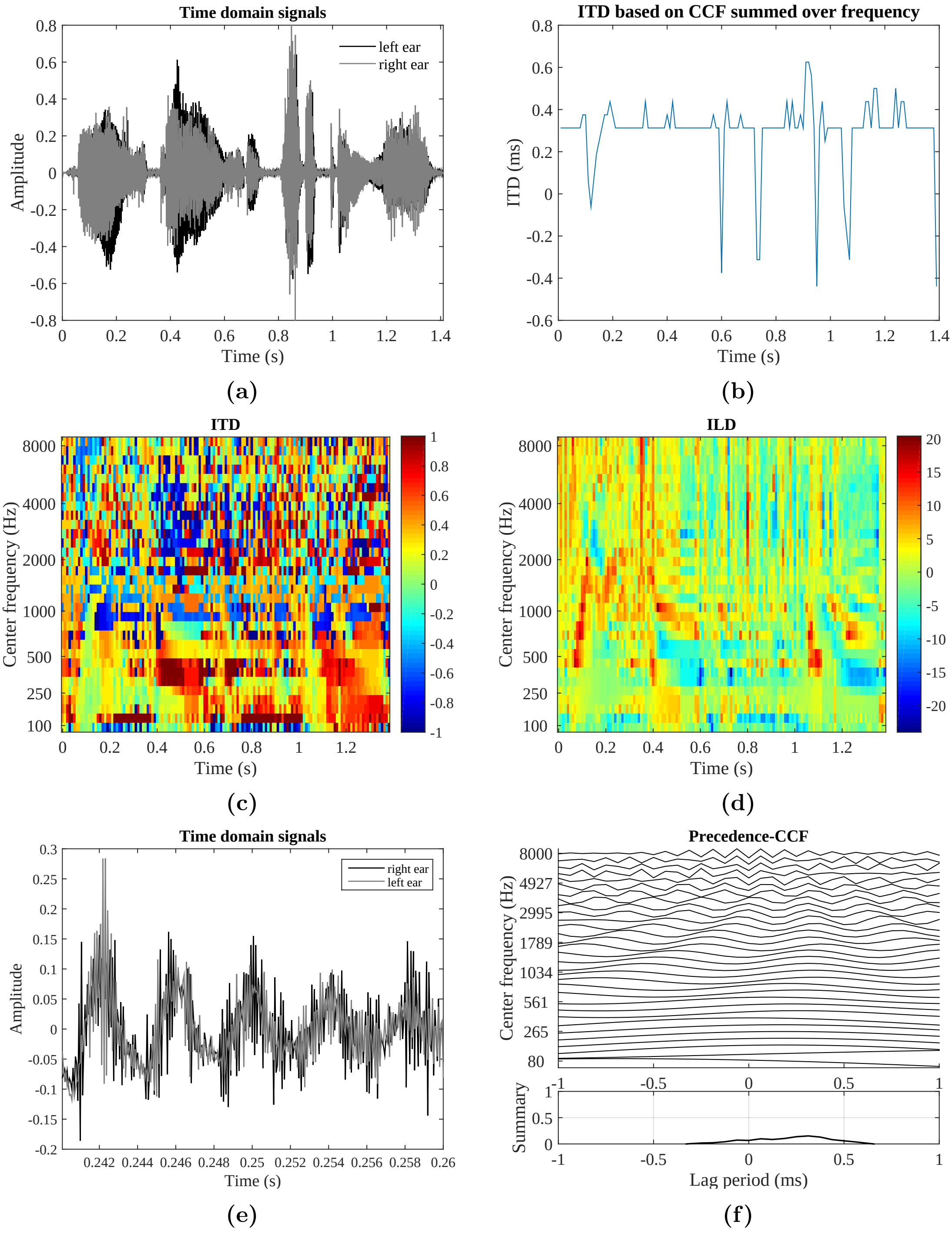
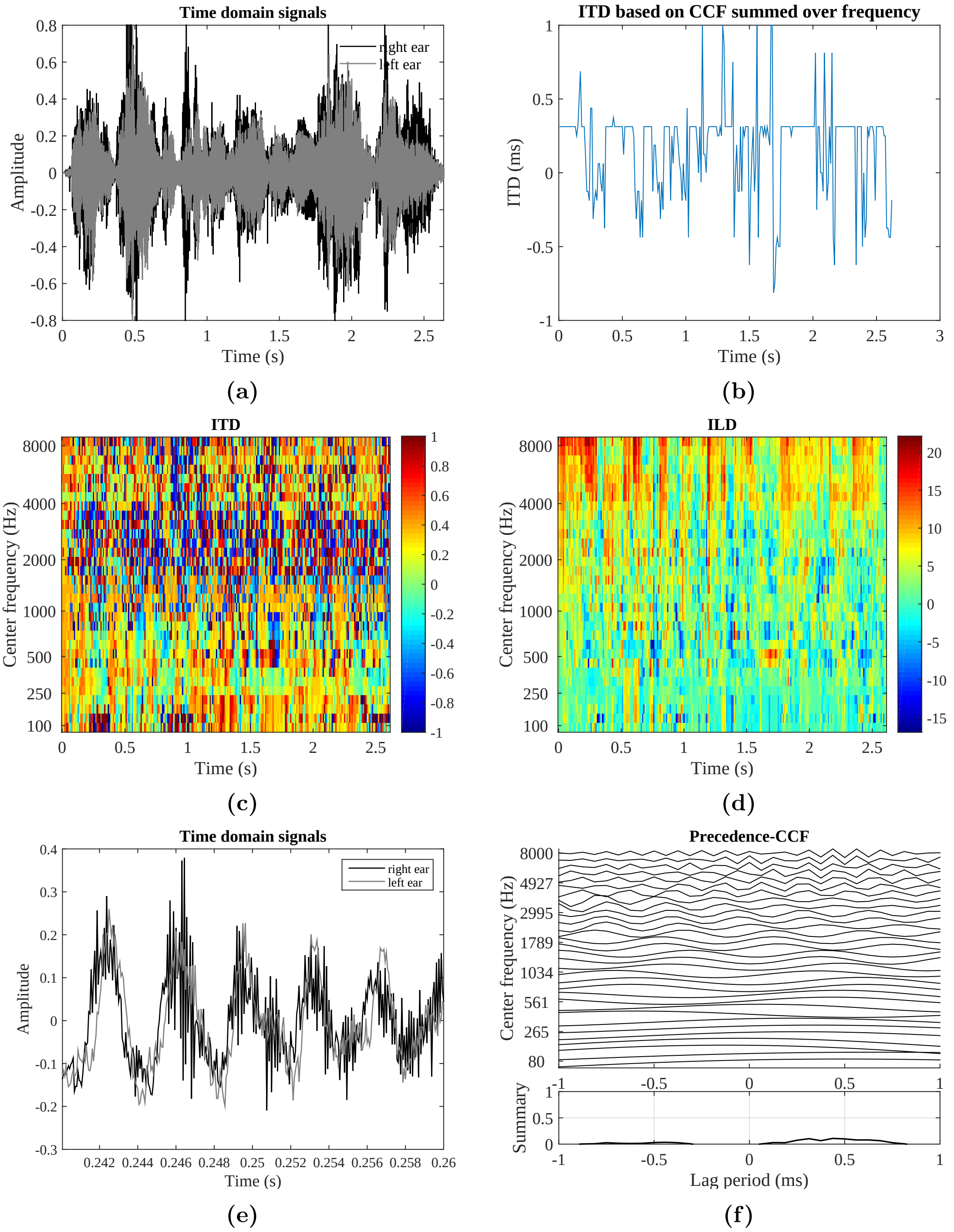


Figure 3.9: Input and output of modified precedence effect processor with an anechoic speech signal and a synthesized lag: (a) input signal, (b) ITD integrated over frequency channels, (c) ITD output as T-F representation, (d) ILD output as T-F representation, (e) input signal for a single frame used for the CCF extraction in (f).



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.10: Input and output of modified precedence effect processor with a reverberant speech signal: (a) input signal, (b) ITD integrated over frequency channels, (c) ITD output as T-F representation, (d) ILD output as T-F representation, (e) input signal for a single frame used for the CCF extraction in (f).

3.4 Prediction of apparent source width

The perceived horizontal extent of sound sources is typically described by the apparent source width (ASW). According to literature, three binaural cues are mainly contributing to ASW: The ITDs and the ILDs which are also important for determining the location of a sound source in the horizontal plane (see section 3.2), and the IC. Due to reflections in rooms and from the head and torso of the listener, all three cues fluctuate over time. With increasing amount of room reflections, the IC decreases and larger variations in ITDs and ILDs occur, leading to an increased ASW. The psychophysical relation between these three binaural cues and ASW can be exploited by binaural auditory models.

Traditional models of ASW have been used to evaluate the quality of concert halls by analyzing the IACC function (Ando, 2007). Based on the IACC, the IC is extracted as the absolute maximum value normalized by the RMS value of the left- and right-ear signals. Hereby, an inverse relation between IC and ASW exists. Okano and colleagues proposed a frequency-specific weighting of the IC, termed $\mathrm{IACC}_{\mathrm{E3}}$, that averages the IC in three octave bands 0.5, 1 and 2 kHz (Okano *et al.*, 1995).

Blauert and Lindemann suggested that both, ITD and ILD fluctuations, contribute to ASW (Blauert and Lindemann, 1986). They combined the standard deviation of both cues with equal weights and reported a higher correlation with perceptual data ($r = 0.75$) as opposed to an IC-based model ($r = 0.61$). Later, Mason *et al.* (2005) developed an ASW model that combined both ITDs and ILDs according to the duplex theory, by using ITDs at low frequencies and ILDs at high frequencies (Mason *et al.*, 2005).

The ability of the ASW processor described in section 2.1 to predict the perceived horizontal extent of sound sources has been evaluated (Käsbach *et al.*, 2016). Specifically, the generalizability was assessed by comparing the model performance across two experimental datasets that were obtained for band-limited and broadband noise, as well as speech and music signals. In addition, it was investigated whether (i) correlation-based approaches, i.e. using IC or ITDs are sufficient for the estimation of ASW, (ii) their suggested frequency regions, i.e. three octave bands at 0.5, 1 and 2 kHz or below 2 kHz, are optimal in such approaches or whether high-frequency IC or ITDs also contribute to ASW and (iii) a model combining ITDs and ILDs (as suggested by Blauert and Lindemann (1986), Mason *et al.* (2005)) is feasible.

3.4.1 Summary of the perceptual studies

Two previously conducted studies on ASW perception (Käsbach *et al.*, 2014, 2015), in the following referred to as Exp. A and B, were considered here to develop and evaluate models of ASW perception. Distinct sensations of ASW were generated by using stereo loudspeaker setups. In such a setup, the listener perceives a phantom sound image in the center of the two loudspeakers.

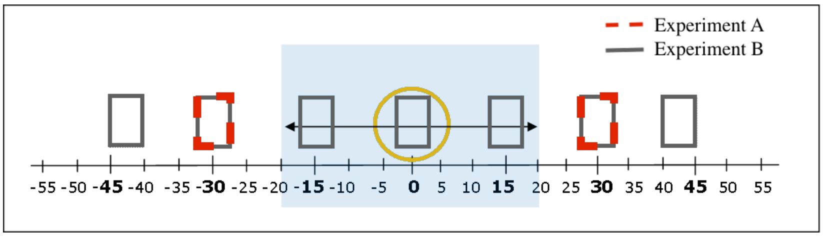


Figure 3.11: Sketch of the experimental set-up. The loudspeaker pairs generate a phantom source at 0 °. Listeners were asked to indicate the ASW in degree, for both boundaries of the source image. For further details, see Käsbach *et al.* (2014) and Käsbach *et al.* (2015).

The ASW was measured as a function of the physical source width (PSW) which was controlled by two experiment-specific settings, the loudspeaker layout and applied signal processing. In the measurement procedure, listeners indicated the perceived ASW on a degree scale as illustrated in Fig. 3.11. In Exp. B, listeners could indicate the left and right most boundary of the sound source separately, whereas in Exp A, the response had to be given symmetrically. In the present study, only 3 source signals per experiment were used. In Exp. A, the stereo setup at an angle of ±30 ° was used indicated by the red dashed rectangles in Fig. 3.11. Five distinct PSW values, denoted by PSW #1 to PSW #5, were generated by varying the coherence between the two loudspeaker channels accordingly to $\mathrm{IC}_{\mathrm{LS}} = [1\ 0.8\ 0.6\ 0.3\ 0]$. The source signal was either Gaussian white noise, band-pass filtered with a bandwidth of 2 octaves at a center frequency of 0.25 kHz or high-pass (HP) filtered at 8 kHz. The stimuli had a duration of 4 s and were presented at 70 dB SPL. In Exp. B, the PSW was controlled by varying the angle between the stereo speakers. In addition, a source widening algorithm was applied as described in Zotter and Frank (2013). Specifically, a line-array of 3 stereo loudspeaker pairs (Type Dynaudio BM6) plus an additional loudspeaker in the center of the array was used as indicated by the gray rectangles in Fig. 3.11. In total, five distinct PSW values were generated. The source signals were pink noise, male speech and a guitar sample. The stimuli had a duration of 6 s and were presented at 70 dB SPL.

In Fig. 3.12, the perceived ASW as a function of PSW averaged across listeners is shown for Exp. A (left panel) and Exp. B (right panel). The error bars represent the standard deviation across listeners. It can be seen that ASW increases with increasing PSW. In Exp. A (left panel), the different signal types (represented by the different symbols and line styles) show similar results with a tendency that the bandpass-filtered signal at 250 Hz and the white noise signal were perceived with larger ASW than the HP filtered signal at 8 kHz. In a statistical analysis with a linear mixed-effects model, the factor PSW showed a similar effect size ($F(4, 48) = 113.6$, $p < 0.001$) compared to the factor source signal (F(4,

42) = 97.2, $p < 0.001$) which was larger than the interaction of both (F(16, 2436) = 21, p < 0.001). In Exp. B (right panel), it can be seen that ASW increases as well with PSW in a similar manner as in Exp. A. Small differences can be seen between the source signals, such that the noise source was generally perceived to have a larger ASW than the speech and guitar signals. In a statistical analysis with a linear mixed-effects model, the factor PSW showed a dominating effect size (F(4, 20) = 110, $p < 0.001$) compared to the factor source signal (F(2, 719) = 31.8, $p < 0.001$) and the interaction of both (F(8, 718) = 9.5, p < 0.001).

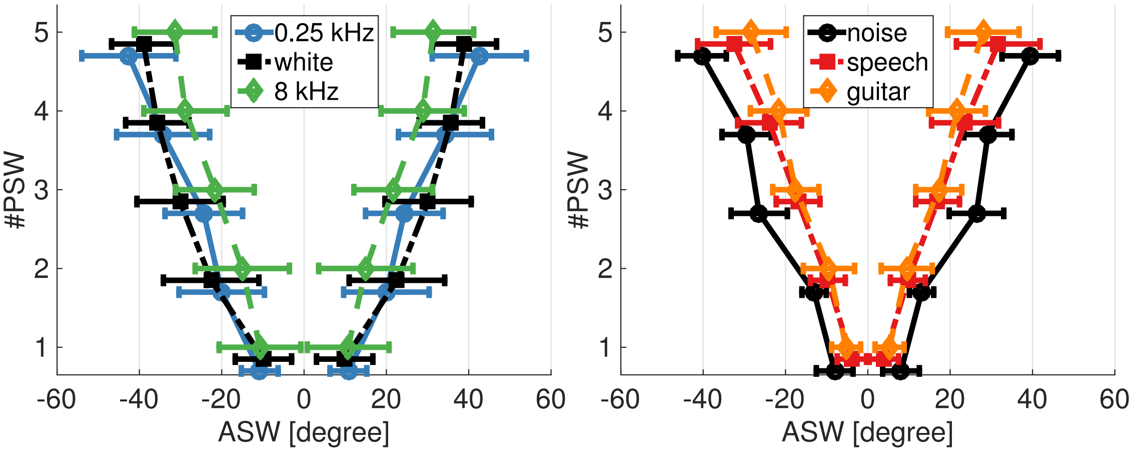


Figure 3.12: Perceptual results of ASW for Exp. A (left panel) and Exp. B (right panel) in degrees. ASW is shown as a function of the PSW, denoted by ASW #1 (narrow) to #5 (wide). Plotted are the mean and standard deviation. The different symbols and line styles represent the different source signals.

3.4.2 Model configurations

A variety of different configurations of the back-end (see section 2.1.2) was considered. The first back-end, termed DUPLEX, combined the percentiles of the ITDs and ILDs according to the duplex theory (Macpherson and Middlebrooks, 2002) which was motivated by Blauert and Lindemann (1986) and Mason *et al.* (2005). The combination of both binaural cues required the normalization of each cue. ITDs were normalized by 1.1 ms and ILDs by 12 dB SPL, which corresponded to the observed maxima, respectively, in the percentiles across stimuli. According to the duplex theory, ITDs contribute up to 1.5 kHz and ILDs contribute above this frequency value. The final prediction of the left and right boundaries was then obtained by calculating the mean value across all frequency channels of the lower and upper percentile, respectively. In a second back-end, termed $\mathrm{ITD}_{\mathrm{low}}$, only the ITD-percentiles were analyzed with an upper frequency limit of 2 kHz according to van Dorp Schuitman *et al.* (2013). The third back-end used the IC for the

ASW prediction, termed IC_{E3}, resembling a short-term analysis of the $IACC_{E3}$. In total, 16 Gammatone filters of the front-end were selected corresponding to the frequency range between 0.35 to 2.83 kHz, defined by the octave-wide filters in $IACC_{E3}$ at 0.5, 1 and 2 kHz. The frame-based values of IC were averaged with equal weights across all frames and frequency channels. The $IACC_{E3}$ according to Okano *et al.* (1995) served as a reference. A calibration stage was required to map the output of each model to ASW in degrees. Using a linear fitting approach, the calibrated model output was $y_{cal} = ay + b$, where a is a sensitivity parameter, b an offset and y the uncalibrated model output. For the calibration two data points were used, PSW #1 and PSW #5 of the white noise stimulus in Exp. A.

3.4.3 Modeling results

The individual model performance was accessed by calculating Pearson's correlation coefficient r^2 and the RMS-error between the calibrated model outputs and all experimental data (left and right boundaries), i.e. for Exps. A and B including all source signals. The corresponding values are displayed in Tab. 3.1. In general, all four models provided a high correlation with the perceptual data (ranging from $r^2 = 0.92$ to $r^2 = 0.97$). This is due to the fact that PSW is the dominating factor compared to the source stimulus which is captured correctly by all models.

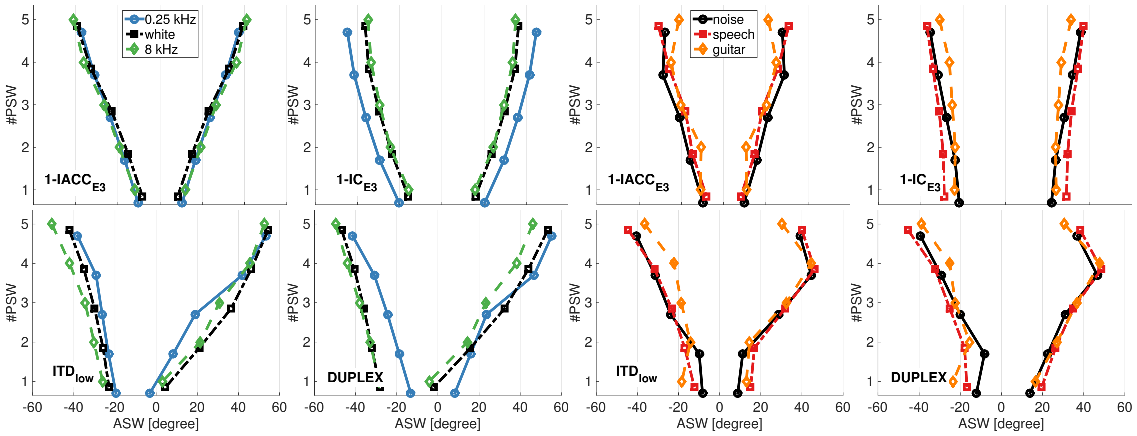


Figure 3.13: Modeling results of ASW for Exp A (left panels) and Exp B (right panels) in degrees. From top to bottom: $1 - IACC_{E3}$, $1 - IC_{E3}$, ITD_{low} and DUPLEX. ASW is shown as a function of the PSW, denoted by ASW #1 (narrow) to #5 (wide). The different symbols and line styles represent the different source signals.

In Fig. 3.13, the outputs of the four tested models, $IACC_{E3}$, IC_{E3}, ITD_{low} and DUPLEX are presented for Experiment A (left panels) and for Exp B (right panels). Note that the first two models are inversely proportional to ASW and are therefore shown as $1 - IACC_{E3}$

and $1 - \mathrm{IC}_{\mathrm{E3}}$, respectively. Further, both models produced a single output value and are therefore shown with a symmetric ASW. It can be seen that all models are able to predict the general trend in the data, i.e. that the perceived ASW increases with PSW. Differences occur with respect to the slopes of the predicted boundaries of the ASW and between source signals. The $\mathrm{IACC}_{\mathrm{E3}}$ model achieves the highest correlation of the considered models with $r^2 = 0.97$ ($r = 0.98$ which corresponds to the findings in Zotter and Frank (2013)) due to the fact that it captures the dynamic range in ASW correctly, i.e. the difference between smallest and largest ASW, for both experiments. However, the model does not capture the increase in ASW for PSW #5 in Exp. B and does only reveal minor differences between the source signals. Considering the model denoted by $\mathrm{IC}_{\mathrm{E3}}$, the performance decreases to $r^2 = 0.92$. This indicates that a short-term analysis of the IC (including the IHC and ATH model stages) and a higher frequency resolution (16 Gammatone filters as opposed to 3 octave-wide filters in $\mathrm{IACC}_{\mathrm{E3}}$) are not required to account for the perceptual data. The $\mathrm{IC}_{\mathrm{E3}}$ predictor has a reduced sensitivity, i.e. a more shallow slope of the boundaries. However, it partially captures source signal differences in Exp. A, e.g. larger ASW for low frequencies (blue circles) compared to high frequencies (green diamonds), but contradicts the data for the noise source (black rectangles). The $\mathrm{ITD}_{\mathrm{low}}$ model's performance is with $r^2 = 0.94$ between the $\mathrm{IC}_{\mathrm{E3}}$ and $\mathrm{IACC}_{\mathrm{E3}}$ models. Since both the IC-based models and $\mathrm{ITD}_{\mathrm{low}}$ are extracted from the IACC, this result is plausible. Its output shows a dynamic range similar to that in the data and is also more asymmetric due to the fact that the boundaries are estimated separately by the corresponding percentiles, such that a potentially asymmetric head and torso simulator (HATS) positioning becomes more crucial. Hence, prediction errors are caused by the asymmetric output and an overestimation in case of the speech and guitar source signals in Exp B. In Tab. 3.1, the performance of the $\mathrm{IACC}_{\mathrm{E3}}$, $\mathrm{IC}_{\mathrm{E3}}$ and $\mathrm{ITD}_{\mathrm{low}}$ models is also shown for the case when including the entire bandwidth for the analysis (denoted with the subscript 'broad'). The corresponding performance is decreased compared to their low frequency estimates. Interestingly, the $\mathrm{IACC}_{\mathrm{broad}}$ and $\mathrm{IC}_{\mathrm{broad}}$ result both in $r^2 = 0.88$, indicating that it becomes irrelevant whether a long- or

Table 3.1: Model performances in terms of correlation coefficient r^2, r, RMS-error and AIC.

Model	r^2	r	RMS-error [°]	AIC ($dof = 13$)
$\mathrm{IACC}_{\mathrm{E3}}$	0.97	0.98	4.3	159
$\mathrm{IACC}_{\mathrm{broad}}$	0.88	0.94	8.7	-
$\mathrm{IC}_{\mathrm{E3}}$	0.92	0.95	10.5	128
$\mathrm{IC}_{\mathrm{broad}}$	0.88	0.94	14.7	-
$\mathrm{ITD}_{\mathrm{low}}$	0.94	0.96	6.4	136
$\mathrm{ITD}_{\mathrm{broad}}$	0.91	0.95	7.9	-
DUPLEX	0.92	0.95	7.9	137
$\mathrm{DUPLEX}_{\mathrm{short}}$	0.87	0.93	13.1	-
ILD	0,77	0.88	12.7	-

short-term analysis is performed in this case. The $\mathrm{ITD_{broad}}$ model results in $r^2 = 0.91$. This suggests that high-frequency components in IACC-based measures do not provide useful information for ASW. The DUPLEX model provides a similar output as the $\mathrm{ITD_{low}}$ model, but performance decreases to $r^2 = 0.92$. Therefore, adding ILDs in the analysis does not provide a further benefit.

The presented ASW models, $\mathrm{IACC_{E3}}$, $\mathrm{IC_{E3}}$, $\mathrm{ITD_{low}}$ and DUPLEX were compared in a statistical analysis. A 3-way analysis of variance (ANOVA) was performed using the model type, PSW and source signal as factors. In contrast to the correlation coefficient r^2, this allowed for a more detailed model analysis across both factors PSW and source signal. The evaluation was based on the AIC (using 13 degrees of freedom, dof = 13) which is a relative criterion, whereby a lower AIC indicates a better model performance. In such an analysis, listed in Tab. 3.1, the $\mathrm{IC_{E3}}$ model performed best (AIC = 128), the $\mathrm{ITD_{low}}$ and the DUPLEX provided similar performance (AIC = 136 and 137, respectively) and the $\mathrm{IACC_{E3}}$ (AIC = 159) model performed less well. However, in a post-hoc analysis with Bonferroni correction (correction factor of 4), no significant differences ($p_{posthoc} < 0.05$) between the models could be revealed.

3.4.4 Conclusions

In this study, two experiments were presented where the ASW was measured as a function of the PSW. The stimuli were analyzed by four binaural functional models to predict ASW. A model that combines ITDs and ILDs according to the duplex theory (DUPLEX) was compared to other existing approaches in the literature, i.e. $\mathrm{IACC_{E3}}$, $\mathrm{IC_{E3}}$, and $\mathrm{ITD_{low}}$. Models based on the interaural cross-correlation function (either extracting IC or ITD) produced similar results for the estimation of ASW. The best performance was obtained by a long-term analysis of the binaural signals using the $\mathrm{IACC_{E3}}$. Apparently, the signals were stationary enough such that a long-term analysis was sufficient. The previously suggested frequency regions for the analysis with cross-correlation based models seems optimal, i.e. averaging across three octave bands at 0.5, 1 and 2 kHz for the $\mathrm{IACC_{E3}}$ and $\mathrm{IC_{E3}}$ models and considering frequencies only below 2 kHz for the $\mathrm{ITD_{low}}$ model. Adding higher frequency components deteriorated the ASW estimation in all models. The DUPLEX model that also included ILDs could not provide any further benefit in the ASW estimation, possibly due to the stationary character of the chosen stimuli.

3.5 Pre-segmentation based on amplitude modulation features

The pre-segmentation stage based on amplitude modulation spectrogram (AMS) features, as described in Deliverable D3.2 (see section 4.1.2), has been evaluated with respect to its ability to enhance noisy speech in challenging acoustic conditions (May *et al.*, 2015a, Bentsen *et al.*, submitted).

Despite substantial research efforts that focused on the development of noise reduction algorithms over the past decades, the improvement of speech intelligibility in noisy conditions remains a challenging task (Hu and Loizou, 2007, Hilkhuysen *et al.*, 2012). Assuming *a priori* knowledge about the target speech and the interfering noise, it is possible to construct an ideal binary mask (IBM) which separates the T-F representation of noisy speech into target-dominated and masker-dominated T-F units. The IBM has been shown to significantly improve speech perception in noisy conditions (Anzalone *et al.*, 2006, Wang *et al.*, 2008, Kjems *et al.*, 2009). The IBM produces intelligible speech when a resolution of about 12 - 16 frequency channels is used (Wang *et al.*, 2008, Li and Loizou, 2008). At the same time, the manipulation of individual T-F units should be performed with a temporal resolution of at least 15 ms, in order to produce significant speech reception threshold (SRT) improvements (Anzalone *et al.*, 2006).

Unfortunately, the IBM is not available in practice and, hence, needs to be estimated based on the noisy speech. In that context, the aforementioned requirements regarding the spectral and temporal resolution determine the bandwidth and the window size with which an estimated binary mask (EBM) should be obtained. In contrast to IBM processing, where the T-F manipulation can be performed at an arbitrarily high temporal resolution (e.g. on a sample-by-sample basis (Anzalone *et al.*, 2006)), algorithms which try to derive an EBM typically operate on window durations between 20 ms Han and Wang (2011) and 90 ms Gonzalez and Brookes (2014).

Several previous studies have employed the extraction of AMS features with linearly-scaled modulation filters (Han and Wang, 2011, Kim *et al.*, 2009, May and Dau, 2013, May and Gerkmann, 2014). Recently, it has been shown that a speech segregation system based on logarithmically-scaled AMS features, inspired by auditory processing principles, is superior to the linear AMS feature representation and can estimate the IBM with high accuracy (May and Dau, 2014a). One critical parameter is the window duration in the AMS feature representation. Modulation-based processing commonly involves longer analysis windows to fully resolve a period of low-frequency modulations within a single analysis window (e.g. 250 ms to analyze one period of 4 Hz modulations). This seems important for the ability to estimate speech-dominated T-F units, since it is known that low-frequency modulations are important for speech perception in the presence of stationary background noise (Drullman *et al.*, 1994). In addition, a longer analysis window may also improve the accuracy of the EBM, since more information can be extracted from the noisy speech. However, a longer analysis window will introduce temporal smearing, which, in turn, may limit the effectiveness of manipulating individual T-F units.

Despite high levels of interfering noise, speech-dominated T-F units tend to cluster in spectro-temporal regions, forming so-called *glimpses*, and the size of these glimpses has been shown to correlate well with speech intelligibility scores from normal-hearing listeners (Cooke, 2006). Consequently, many computational segregation systems exploit contextual information, either implicitly through the use of *delta* features (Han and Wang, 2011, Kim *et al.*, 2009), or explicitly, by incorporating a spectro-temporal integration stage (May and Dau, 2013, 2014a, Healy *et al.*, 2013). However, the interaction between the window duration and the spectro-temporal integration stage and its impact on speech segregation performance has not yet been clarified.

The goal of the present study is, therefore, to investigate the influence of the window duration on computational speech segregation based on auditory-inspired modulation features. Specifically, the interaction between window duration, estimation accuracy of the EBM and predicted speech intelligibility is analyzed. Moreover, the influence of a spectro-temporal integration stage is examined. The estimation accuracy of the EBM is measured using a technical classification measure (the hit rate minus false alarm rate) (Kim *et al.*, 2009). In addition, the predicted intelligibility of the reconstructed target speech is evaluated using the short-time objective intelligibility (STOI) metric (Taal *et al.*, 2011).

3.5.1 Computational speech segregation

The segregation system consisted of a Gammatone-based analysis and synthesis stage. In the analysis stage, the noisy speech was sampled at a rate of 16 kHz and decomposed into 31 frequency channels using a Gammatone filterbank. The center frequencies were equally spaced on the ERB scale between 80 and 7642 Hz. The envelope in each frequency channel was extracted by half-wave rectification and further smoothed by a second-order low-pass filter with a cutoff frequency of 1 kHz to roughly simulate the loss of phase-locking in the auditory system towards higher frequencies. Based on this auditory spectrogram-like representation, a set of AMS features was extracted. A two-layer segregation stage was trained to discriminate between speech-dominated and noise-dominated T-F units by exploiting *a priori* knowledge about the AMS feature distribution corresponding to speech and noise activity (May and Dau, 2014a). This segregation stage produced an EBM that was applied to the individual subbands of the noisy speech in the synthesis stage in order to attenuate noise-dominated T-F units.

AMS feature extraction

Prior to the AMS feature extraction, each subband envelope was normalized by its median computed over the entire sentence. This normalization stage was shown to be crucial in order to deal with effects of room reverberation, spectral distortions and unseen signal-to-noise ratios (SNRs) (May and Gerkmann, 2014, May and Dau, 2014a).

Each normalized subband was then analyzed by a modulation filterbank, consisting of a first-order low-pass filter and second-order band-pass filters whose center frequencies were logarithmically spaced up to 1024 Hz (May and Dau, 2014a). The bandpass filters were assumed to have a constant Q-factor of 1, inspired by findings in auditory modeling (Ewert and Dau, 2000). The cutoff frequency of the modulation low-pass filter f_{LP} was set to the inverse of the window duration T_w, to ensure that at least one period of the modulation frequency was included in the analysis window. The modulation power was measured for each frequency channel by computing the RMS value within each time window at the output of each modulation filter.

Segregation stage

In order to discriminate between speech-dominated and noise-dominated T-F units, a two-layer segregation stage was employed, which consisted of a GMM classifier combined with a spectro-temporal integration stage based on a support vector machine (SVM) classifier (May and Dau, 2014a). First, a GMM classifier was trained for each individual frequency channel f to model the AMS feature distribution of speech-dominated and noise-dominated T-F units, denoted by $\lambda_{1,f}$ and $\lambda_{0,f}$. Given the AMS feature vector $\mathbf{X}(t,f)$ for a particular time frame t and frequency channel f, the *a posteriori* probability of speech and noise presence was computed by

$$P\left(\lambda_{1,f}|\mathbf{X}(t,f)\right) = \frac{P\left(\lambda_{1,f}\right)P\left(\mathbf{X}(t,f)|\lambda_{1,f}\right)}{P(\mathbf{X}(t,f))}, \tag{3.1}$$

$$P\left(\lambda_{0,f}|\mathbf{X}(t,f)\right) = \frac{P\left(\lambda_{0,f}\right)P\left(\mathbf{X}(t,f)|\lambda_{0,f}\right)}{P(\mathbf{X}(t,f))}, \tag{3.2}$$

where the two *a priori* probabilities $P\left(\lambda_{0,f}\right)$ and $P\left(\lambda_{1,f}\right)$ were computed by counting the number of feature vectors during training. The EBM without spectro-temporal integration was estimated by comparing the two *a posteriori* probabilities of speech and noise presence for each individual T-F unit

$$\mathcal{M}(t,f) = \begin{cases} 1 & \text{if } P\left(\lambda_{1,f}|\mathbf{X}(t,f)\right) > P\left(\lambda_{0,f}|\mathbf{X}(t,f)\right) \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

In the second layer, the *a posteriori* probability of speech presence $P\left(\lambda_{1,f}\right)$ was considered as a new feature spanning across a spectro-temporal integration window, and subsequently learned by a SVM classifier (May and Dau, 2014a). The output of this second classification layer represented the EBM with spectro-temporal integration.

Waveform synthesis

Before the EBM was applied to the noisy speech, a lower limit β was incorporated. This flooring limited the amount of noise attenuation, but reduced the impact of distortions (musical noise) caused by the binary processing (Anzalone *et al.*, 2006). A flooring value of $\beta = 0.1$, corresponding to 20 dB attenuation, was considered appropriate. This frame-based EBM was then interpolated to a sample-based EBM. Transitions in the EBM from speech-dominated to noise-dominated units or noise-dominated to speech-dominated units were smoothed by a raised-cosine window (Wang and Brown, 2006). Then, the sample-based EBM was applied to the subband signals of the noisy speech. To remove across-frequency phase differences, the weighted subband signals were time-reversed, passed through the corresponding Gammatone filter, and time reversed again (Wang and Brown, 2006, Brown and Cooke, 1994). Finally, the target signal was reconstructed by summing up the weighted and phase-aligned subband signals across all frequency channels.

3.5.2 Evaluation

Stimuli

Noisy speech was created by corrupting randomly selected male and female sentences from the TIMIT corpus with one of four different noise signals, from which a random segment was selected for each sentence. The noise was switched on 250 ms before the speech onset and was switched off 250 ms after the speech offset. The following noise types were used: two types of speech-shaped noise (SSN) (stationary ICRA1-noise and non-stationary, speech-modulated ICRA5-noise; (Dreschler *et al.*, 2001)), 8-Hz amplitude-modulated pink noise and a recording of a cracking oak tree with wind noise[1]. The noise signals were split in two halves of equal size to prevent any overlap between the signals used during training and testing, which would result in an overly optimistic segregation performance (May and Dau, 2014b).

[1] Recording taken from `www.freesound.org/people/klankbeeld/sounds/211776/`

Model training

The GMM classifier described in section 3.5.1 was trained with randomly selected sentences from the training set of the TIMIT corpus (Garofolo *et al.*, 1993) that were corrupted with one of the four background noises at $-5, 0$ and 5 dB SNR. As explained in section 3.5.2, the number of sentences involved in the training depends on the AMS feature configuration (see Tab. 3.2). A local criterion (LC) of -5 dB was applied to the *a priori* SNR in order to separate the AMS feature distribution into speech-dominated and noise-dominated T-F units. The SVM-based spectro-temporal integration stage consisted of a causal, plus-shaped integration window spanning across 9 adjacent frequency channels and 3 time frames (May and Dau, 2014a). A linear SVM classifier (Chang and Lin, 2001) was trained with 10 sentences mixed at $-5, 0$ and 5 dB SNR. Afterwards, new SVM decision thresholds were obtained that maximized the hit minus false alarm (HIT-FA) rate (Han and Wang, 2011) on a validation set of 10 sentences mixed at $-5, 0$ and 5 dB SNR. A separate GMM and SVM classifier was trained for each noise type.

Model evaluation

The segregation system was evaluated with 60 randomly selected sentences from the testing set of the TIMIT corpus mixed with the four different background noises at $-5, 0$ and 5 dB SNR. The segregation performance was assessed by comparing the EBM with the IBM. Specifically, the hit rate (HIT; percentage of correctly identified speech-dominated T-F units) minus the false alarm rate (FA; percentage of erroneously classified noise-dominated T-F units) was reported. In addition, the predicted intelligibility of the reconstructed speech signal was compared to the clean speech signal using the STOI metric (Taal *et al.*, 2011), which has been shown to correlate with subjectively-measured speech intelligibility scores. For the STOI evaluation, the 250 ms noise-only segments at the beginning and the end of each sentence were discarded.

Moreover, the segregation system was compared to an short-time discrete Fourier transform (STFT)-based speech enhancement algorithm. Specifically, the log-minimum mean square error (MMSE) noise reduction algorithm[2] (Ephraim and Malah, 1985) combined with the MMSE-based noise power estimation algorithm (Gerkmann and Hendriks, 2012) was used. The complete 250 ms noise-only segments before speech onset were used to properly initialize the noise power estimation.

[2] Matlab implementations were taken from the Voicebox toolbox provided by M. Brookes: `www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

Table 3.2: AMS feature settings.

T_w	T_s	f_{LP}	# dim.	# sentences
256 ms	64 ms	4 Hz	9	960
128 ms	32 ms	8 Hz	8	480
64 ms	16 ms	16 Hz	7	240
32 ms	8 ms	32 Hz	6	120
16 ms	4 ms	64 Hz	5	120
8 ms	2 ms	128 Hz	4	120
4 ms	1 ms	256 Hz	3	120

Experimental setup

The segregation system was trained with AMS features based on 7 different window durations T_w, as shown in Tab. 3.2. Accordingly, the cutoff frequency of the modulation low-pass filter f_{LP} varied between 4 Hz (9 AMS features) and 256 Hz (3 AMS features). The frame shift was always set to $T_s = T_w/4$. As a result, the number of feature vectors available during training was higher for the AMS features with shorter window durations compared to longer window durations. To compensate for this, the number of TIMIT sentences used to train the GMM classifier was adjusted for window durations above 32 ms according to Tab. 3.2. To investigate the influence of exploiting contextual information, two different segregation systems were trained: a single-layer GMM-based segregation system and a two-layer GMM-SVM segregation system including the spectro-temporal integration stage.

3.5.3 Experimental results

Effect of the window duration

The performance of the AMS-based segregation system is shown in Fig. 3.14 as a function of the window duration for the four different background noises. The top panel in each of the four subplots shows the STOI improvement relative to the unprocessed noisy speech for the IBM as well as the EBM with and without the SVM-based spectro-temporal integration stage. In addition, the corresponding HIT-FA rates of the two EBM systems are shown in the bottom panel.

It can be seen that the IBM produced the highest STOI improvements due to the availability of *a priori* information and the performance increased monotonically with increasing temporal resolution. Despite the fact that the HIT-FA rates of both EBM systems almost continuously increased with increasing window durations for all the noise types,

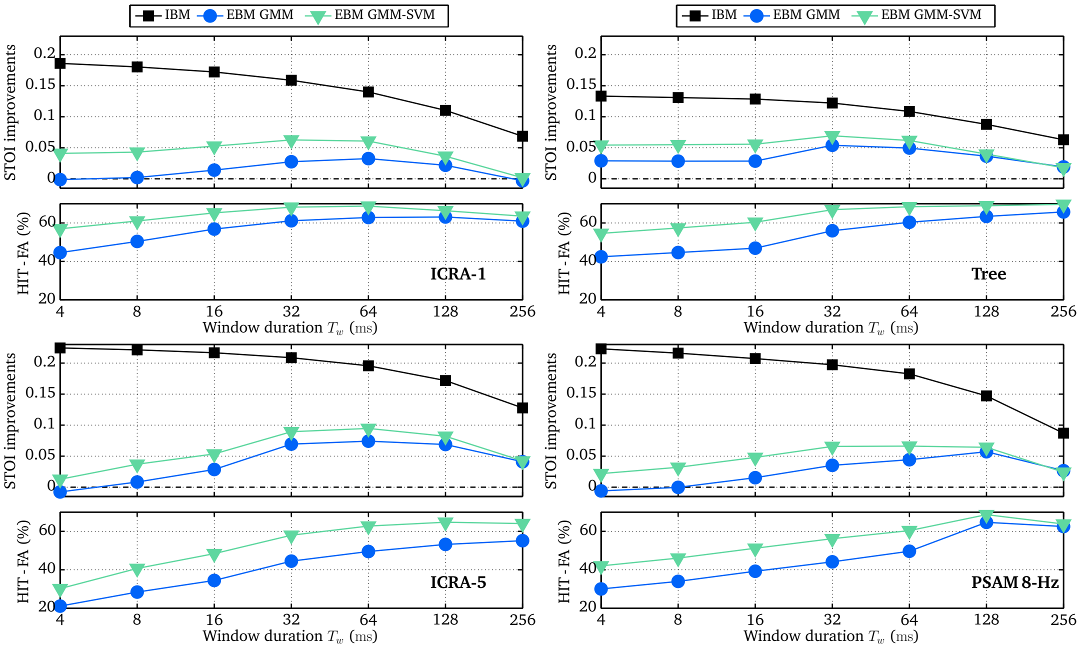


Figure 3.14: STOI improvements for the IBM and two EBM systems along with their corresponding HIT-FA rates averaged across all sentences and SNRs. The results are shown separately for each of the four noise types.

the STOI improvement showed a plateau for window durations between 32 − 64 ms, and the performance was lower for shorter and longer window durations. Considering the ICRA-5 noise, there was a considerable improvement in the HIT-FA rates when increasing the window duration from 16 ms to 32 ms, which also led to a larger STOI improvement.

Overall, the EBM system with the SVM-based spectro-temporal integration stage produced substantially higher HIT-FA rates, which was also reflected in larger STOI improvements. In addition, the SVM-based integration of contextual information seemed to reduce the required window size. This was most noticeable for the PSAM 8-Hz noise, for which the EBM-GMM system with a window duration of 128 ms, required to resolve a full period of 8 Hz, produced the largest STOI improvements. The same performance was obtained with the EBM with the spectro-temporal integration stage using a window size of 32 ms.

3.5.4 Comparison with noise reduction algorithm

Inspired by the analysis presented in Gonzalez and Brookes (2014), Fig. 3.15 shows the sentence-based STOI predictions for the unprocessed noisy speech in relation to the measured STOI improvement for the following three systems: a) the EBM with the spectro-temporal integration stage, b) the log-MMSE noise reduction algorithm and c) the IBM. In addition, a least-squares fit is shown for each noise type. Based on the analysis in the previous section, all algorithms operated on a window size of 32 ms.

As expected, the IBM-based system produced the largest STOI improvements across all noise types. Also the EBM system improved the predicted speech intelligibility, in particular for conditions where the STOI values of the noisy speech were below 0.7. Whereas the STOI improvements were moderate for the IRCA-1 noise and the PSAM 8-Hz, a larger benefit was observed for the ICRA-5 noise and the tree noise.

The log-MMSE-based noise reduction system showed minor improvement for the ICRA-1 noise, presumably because the stationary background noise could be reasonably well estimated. However, in case of the other non-stationary noises, it appeared that the rapid fluctuations could not be predicted by the noise estimation algorithm. As a consequence, the predicted intelligibility improvements were around zero or even negative, which is in line with previous studies (Hu and Loizou, 2007, Hilkhuysen *et al.*, 2012, Gonzalez and Brookes, 2014)

3.5.5 Conclusions

The choice of a window duration in modulation-based speech segregation constitutes a trade-off between the ability to resolve low-frequency modulations and the temporal resolution with which the segregation system can manipulate individual T-F units. This choice is only moderately affected by the modulation content of the interfering noise. In general, a window size of 32 ms seems to represent a good compromise. It is conceivable that the modulation analysis could be performed at multiple time constants, as implemented in Jørgensen *et al.* (2013), and that the decision about speech and noise activity is combined across various decision streams based on different time constants.

The spectro-temporal integration stage effectively improves the ability of the segregation system to analyze low-frequency modulations by combining contextual knowledge about the speech presence probability across neighboring T-F units, thereby reducing the required window duration. However, a high performance in terms of the frequently-used performance metric, the HIT-FA rate, does not necessarily lead to improvements in predicted speech intelligibility, if the T-F manipulation is not performed with a sufficiently high temporal resolution. Finally, the segregation system has been evaluated using a technical performance

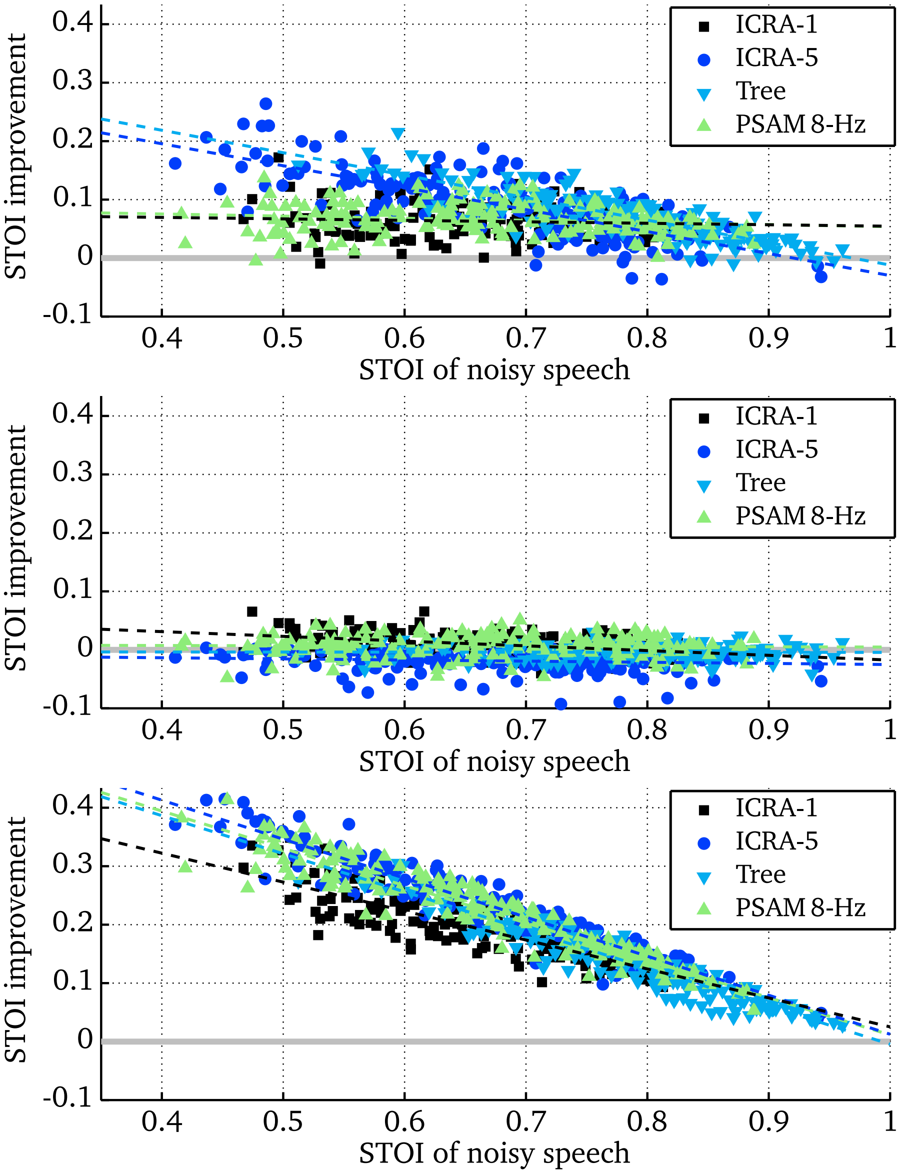


Figure 3.15: STOI predictions for the EBM including the spectro-temporal integration stage (top panel), log-MMSE noise reduction (middle panel) and IBM processing (bottom panel).

measure and model predictions. The next step is to confirm these findings with behavioral listening tests.

4 Summary and conclusion

4.1 Summary

This document summarizes the final extensions of the AFE framework which makes this preprocessing stage, and thus the whole Two!Ears model, more flexible in the evaluation of dynamic binaural scenes. For this purpose some additional processors have been developed and integrated in the software framework. The structure, content and user manual of this framework have been the focus of the earlier WP2 deliverables D2.1, D2.2 and D2.3. With these activities the bottom-up peripheral signal-processing part of the Two!Ears project has been finalized, as far as the development of new processors is being concerned.

Furthermore, we have evaluated the framework for a number of scenarios and binaural conditions for which specific properties of the AFE are critical. It has to be noted that the evaluation of the AFE in general is only possible in the framework of the whole Two!Ears processing system which is at the time of writing this deliverable (M30) still under development. Furthermore, for benchmarking of the Two!Ears system, perceptual data are being derived as part of the activities within WP6. The corresponding evaluation and comparison between experiment and model prediction will be in the focus of the project activities in the final months and will be incorporated in deliverables due at the project end (e.g., D4.3, D5.3, D6.1.3, D6.2.3).

The two additional processors are needed for evaluating dynamic spatial sound fields. The precedence effect processor is based on earlier work of the Two!Ears partner RPI, and its integration into the software structure of Two!Ears required major software engineering effort. As a result, the framework is now capable of a direct comparison of spatial listening performance with and without inclusion of such a module, which does not only allow the evaluations reported in section 3.3 and the study of, e.g., the combined effect of precedence processing module and active head movement, but enables a highly comprehensible spatial analysis tool for the international research community.

The ASW model connects a number of AFE processing stages to derive distributions of spatial parameters from the spatial sound field. ASW can then be estimated by choosing a certain percentile value of these distributions, and by including or excluding individual parameters and specific spectral subbands. In the comparison with a set of experimental data derived at partner DTU for stationary sound fields, a clear difference in agreement for the various combinations of binaural parameters distributions could be demonstrated. Again, this sections shows the applicability of the framework to perceptual data in spatial sound field analysis.

One of the options of the AFE framework is to switch between linear and nonlinear basilar-membrane processing. While it is obvious that the level-nonlinear input-output function must influence binaural parameters which are mathematically dependent on the monaural excitation, our analysis shows which parameters and to which extent are affected by this nonlinearity. This includes ILDs, but also the IACC which will, for coherent signals with an interaural level difference, decrease from 1 as soon as the overall level of the stimulus lies in the range of about 40 to 80 dB. How strong these nonlinearities affect human localization behavior is currently unclear, because the literature provides conflicting results.

One of the top-down processing capabilities of the Two!Ears framework is the control of head movements, and the multi-conditional training (MCT). These have been evaluated in the context of the localization of a sound source both with and without reverberation, and in the presence of up to 3 competing sound sources. It could be shown that in such conditions a strategy to move the head towards the most likely sound source was the best to reduce front-back confusions and the localization error.

In the context of this WP2 evaluation, we focused on active head movements to improve source localization by reducing ambiguities. When evaluating head movements using the full model functionality, we will need to consider also other causes for moving the head, being it reflexive interactions like the "turn to" reflex, or intentional explorative movements such as turning the head towards interesting sources. The resulting need to prioritize such actions all asking for head movements will be addressed in the final evaluations to be reported in the M36 deliverables.

A final evaluation focused at the intersection between the processing in WP2 and WP3, where pre-segmentation of signals for, e.g., the purpose of noise reduction is realized. Here, the use of perception-inspired modulation features to steer signal segmentation was evaluated and compared to noise reduction algorithms from the literature. This analysis again demonstrated the flexibility of the Two!Ears software architecture and showed that the use of perceptually inspired features, here for pre-segmentation, forms an alternative for more signal-processing driven approaches.

4.2 Conclusion

Concluding the work on the AFE framework as first finalized part of the Two!Ears project allows to look back. The philosophy of this framework, both from a psychoacoustics and a software engineering perspective, has been formulated in one of the earlier deliverables D2.2: "The auditory front-end (AFE) represents the first stage of the system architecture and concerns bottom-up auditory signal processing, which transforms binaural signals into multi-dimensional auditory representations. The output provided by this AFE consists of several transformed versions of ear signals enriched by perception-based descriptors which form the input to the higher model stages. Specific emphasis is given on the modularity of the software framework, making this AFE more than just a collection of models documented in the literature. Bottom-up signal processing is implemented as a collection of processor modules, which are instantiated and routed by a manager object. A variety of processor modules is provided to compute auditory cues such as rate-maps, interaural time and level differences, interaural coherence, onsets and offsets. An object-oriented approach is used throughout, giving benefits of reusability, encapsulation and extensibility."

In fact, a considerable part of the work has been focused on the architecture and the software implementation, because basically, no new peripheral model stages had to be developed, but rather, a great variety of existing modules had to be redesigned to fit into the object-oriented structure. The gain of this software-engineering effort is only for a certain part visible in the evaluations which are done within the consortium. Through the public visibility and availability of the project and its software resources, we foresee a major step forward also in the future, when new projects initiated by the consortium members, but more importantly, the worldwide community of auditory, room acoustic and audio signal processing experts are using these modules.

List of Acronyms

Acronyms

AFE auditory front-end

AMS amplitude modulation spectrogram

ANOVA analysis of variance

ASW apparent source width

ATH absolute threshold of hearing

BRIR binaural room impulse response

CC cross-correlation

DNN deep neural network

DRNL dual-resonance non-linear

EBM estimated binary mask

ERB equivalent rectangular bandwidth

GMM Gaussian mixture model

HATS head and torso simulator

HIT-FA hit minus false alarm

HP high-pass

HRTF head-related transfer function

IACC interaural cross-correlation

IBM ideal binary mask

IC interaural coherence

IHC inner hair-cell

ILD interaural level difference

ITD interaural time difference

JND just noticeable difference

LC local criterion

MCT multi-conditional training

MMSE minimum mean square error

PSW physical source width

RMS root mean square

SNR signal-to-noise ratio

SPL sound pressure level

SRT speech reception threshold

SSN speech-shaped noise

STFT short-time discrete Fourier transform

STOI short-time objective intelligibility

SVM support vector machine

T-F time-frequency

Bibliography

Ando, Y. (**2007**), "Concert hall acoustics based on subjective preference theory," in *The Springer Handbook of Acoustics*, edited by T. D. Rossing, Springer Science + Business Media, New York NY, chap. 10, pp. 351–386. (Cited on page 23)

Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**), "Determination of the potential benefit of time-frequency gain manipulation," *Ear and Hearing* **27**(5), pp. 480–492. (Cited on pages 29 and 32)

Bentsen, T., May, T., Kressner, A. A., and Dau, T. (**submitted**), "Comparing the influence of spectro-temporal integration in computational speech segregation," in *Proceedings of Interspeech.* (Cited on page 29)

Bernstein, L. R., van de Par, S., and Trahiotis, C. (**1999**), "The normalized interaural correlation: Accounting for N_oS_π thresholds obtained with Gaussian and "low-noise" masking noise," *Journal of the Acoustical Society of America* **106**(2), pp. 870–876. (Cited on page 3)

Blauert, J. and Lindemann, W. (**1986**), "Auditory spaciousness: Some further psychoacoustic analyses," *The Journal of the Acoustical Society of America* **80**(2), pp. 533–542. (Cited on pages 23 and 25)

Braasch, J. (**2013**), "A precedence effect model to simulate localization dominance using an adaptive, stimulus parameter-based inhibition process." *The Journal of the Acoustical Society of America* **134**(1), pp. 420–35. (Cited on page 6)

Brown, G. J. and Cooke, M. P. (**1994**), "Computational auditory scene analysis," *Computer Speech and Language* **8**(4), pp. 297–336. (Cited on pages 5 and 32)

Chang, C. C. and Lin, C. J. (**2001**), "LIBSVM: A library for support vector machines," URL `www.csie.ntu.edu.tw/~cjlin/libsvm`. (Cited on page 33)

Cooke, M. (**2006**), "A glimpsing model of speech perception in noise," *Journal of the Acoustical Society of America* **119**(3), pp. 1562–1573. (Cited on page 30)

Deatherage, B. H. and Hirsh, I. J. (**1959**), "Auditory localization of clicks," *The Journal of the Acoustical Society of America* **31**(4), pp. 486–492. (Cited on page 15)

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (**2001**), "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Audiology* **40**(3), pp. 148–157. (Cited on page 32)

Drullman, R., Festen, J. M., and Plomp, R. (**1994**), "Effect of temporal envelope smearing on speech preception," *Journal of the Acoustical Society of America* **95**(2), pp. 1053–1064. (Cited on page 29)

Ephraim, Y. and Malah, D. (**1985**), "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing* **33**(2), pp. 443–445. (Cited on page 33)

Ewert, S. D. and Dau, T. (**2000**), "Characterizing frequency selectivity for envelope fluctuations," *Journal of the Acoustical Society of America* **108**(3), pp. 1181–1196. (Cited on page 31)

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1993**), "DARPA TIMIT Acoustic-phonetic continuous speech corpus CD-ROM," *National Inst. Standards and Technol. (NIST)* . (Cited on page 33)

Gerkmann, T. and Hendriks, R. C. (**2012**), "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech and Language Processing* **20**(4), pp. 1383–1393. (Cited on page 33)

Gonzalez, S. and Brookes, M. (**2014**), "Mask-based enhancement for very low quality speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7029–7033. (Cited on pages 29 and 36)

Han, K. and Wang, D. L. (**2011**), "An SVM based classification approach to speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4632–4635. (Cited on pages 29, 30, and 33)

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. L. (**2013**), "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *Journal of the Acoustical Society of America* **134**(6), pp. 3029–3038. (Cited on page 30)

Hilkhuysen, G., Gaubitch, N., Brookes, M., and Huckvale, M. (**2012**), "Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios," *Journal of the Acoustical Society of America* **131**(1), pp. 531–539. (Cited on pages 29 and 36)

Hohmann, V. (**2002**), "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica United With Acustica* **88**, pp. 433–442. (Cited on page 6)

Hu, Y. and Loizou, P. C. (**2007**), "A comparative intelligibility study of single-microphone noise reduction algorithms," *Journal of the Acoustical Society of America* **122**(3), pp.

1777–1786. (Cited on pages 29 and 36)

Jørgensen, S., Ewert, S. D., and Dau, T. (**2013**), “A multi-resolution envelope-power based model for speech intelligibility,” *Journal of the Acoustical Society of America* **134**(1), pp. 1–11. (Cited on page 36)

Käsbach, J., Hahmann, M., May, T., and Dau, T. (**2016**), “Assessing the contribution of binaural cues for apparent source width perception via a functional model,” in *Proceedings of the 22nd International Congress on Acoustics (ICA)*. (Cited on page 23)

Käsbach, J., May, T., Goff, N. L., and Dau, T. (**2014**), “The importance of binaural cues for the perception of ASW at different sound pressure levels,” in *Proceedings of DAGA*. (Cited on pages 23 and 24)

Käsbach, J., Wiinberg, A., May, T., Jepsen, M. L., and Dau, T. (**2015**), “Apparent source width perception in normal-hearing, hearing-impaired and aided listeners,” in *Proceedings of DAGA*. (Cited on pages 23 and 24)

Kim, C., Mason, R., and Tim, B. (**2008**), “Initial investigation of signal capture techniques for objective measurement of spatial impression considering head movement,” in *Audio Engineering Society Convention 124*. (Cited on page 13)

Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**), “An algorithm that improves speech intelligibility in noise for normal-hearing listeners,” *Journal of the Acoustical Society of America* **126**(3), pp. 1486–1494. (Cited on pages 29 and 30)

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. L. (**2009**), “Role of mask pattern in intelligibility of ideal binary-masked noisy speech,” *Journal of the Acoustical Society of America* **126**(3), pp. 1415–1426. (Cited on page 29)

Li, N. and Loizou, P. C. (**2008**), “Effect of spectral resolution on the intelligibility of ideal binary masked speech,” *Journal of the Acoustical Society of America* **123**(4), pp. EL59–EL64. (Cited on page 29)

Ma, N., May, T., and Brown, G. J. (**2015**a), “Robust localisation of multiple speakers exploiting deep neural networks and head movements,” in *Proceedings of Interspeech*, pp. 3302–3306. (Cited on page 15)

Ma, N., May, T., Wierstorf, H., and Brown, G. (**2015**b), “A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2699–2703. (Cited on page 15)

Macpherson, E. A. and Middlebrooks, J. C. (**2002**), “Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited,” *The Journal of the Acoustical*

Society of America **111**(5), pp. 2219–2236. (Cited on page 25)

Mason, R., Brookes, T., Rumsey, F., and Neher, T. (**2005**), "Perceptually motivated measurement of spatial sound attributes for audio-based information systems," *EPSRC Project Reference: GR/R55528/01* URL `http://iosr.uk/projects/PMMP/index.php`. (Cited on pages 23 and 25)

May, T., Bentsen, T., and Dau, T. (**2015**a), "The role of temporal resolution in modulation-based speech segregation," in *Proceedings of Interspeech*, pp. 170–174. (Cited on page 29)

May, T. and Dau, T. (**2013**), "Environment-aware ideal binary mask estimation using monaural cues," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–4. (Cited on pages 29 and 30)

May, T. and Dau, T. (**2014**a), "Computational speech segregation based on an auditory-inspired modulation analysis," *Journal of the Acoustical Society of America* **136**(6), pp. 3350–3359. (Cited on pages 29, 30, 31, 32, and 33)

May, T. and Dau, T. (**2014**b), "Requirements for the evaluation of computational speech segregation systems," *Journal of the Acoustical Society of America* **136**(6), pp. EL398–EL404. (Cited on page 32)

May, T. and Gerkmann, T. (**2014**), "Generalization of supervised learning for binary mask estimation," in *International Workshop on Acoustic Signal Enhancement*, Antibes, France. (Cited on pages 29 and 31)

May, T., Ma, N., and Brown, G. J. (**2015**b), "Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2679–2683. (Cited on page 15)

May, T., van de Par, S., and Kohlrausch, A. (**2011**), "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech and Language Processing* **19**(1), pp. 1–13. (Cited on page 15)

May, T., van de Par, S., and Kohlrausch, A. (**2013**), "Binaural Localization and Detection of Speakers in Complex Acoustic Scenes," in *The Technology of Binaural Listening*, edited by J. Blauert, Springer, Berlin–Heidelberg–New York NY, chap. 15, pp. 397–425. (Cited on page 15)

Okano, T., Beranek, L. L., and Hidaka, T. (**1995**), "Interaural cross-correlation, lateral fraction, and low- and high frequency sound levels as measures of acoustical quality in concert halls," *The Journal of the Acoustical Society of America* **98**(2), pp. 255–265. (Cited on pages 23 and 26)

Perrett, S. and Noble, W. (**1997**), “The effect of head rotations on vertical plane sound localization,” *Journal of the Acoustical Society of America* **102**(4), pp. 2325–2332. (Cited on page 18)

Stern, R. M., Zeiberg, A. S., and Trahiotis, C. (**1988**), “Lateralization of complex binaural stimuli: A weighted-image model,” *The Journal of the Acoustical Society of America* **84**(1), pp. 156–165. (Cited on page 7)

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**), “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech and Language Processing* **19**(7), pp. 2125–2136. (Cited on pages 30 and 33)

Terhardt, E. (**1979**), “Calculating virtual pitch,” *Hearing Research* **1**, pp. 155–182. (Cited on page 4)

van Dorp Schuitman, J., de Vries, D., and Lindau, A. (**2013**), “Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model,” *The Journal of the Acoustical Society of America* **133**(3), pp. 1572–1585. (Cited on page 25)

Wang, D. L. and Brown, G. J. (Eds.) (**2006**), *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley / IEEE Press. (Cited on page 32)

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2008**), “Speech perception of noise with binary gains,” *Journal of the Acoustical Society of America* **124**(4), pp. 2303–2307. (Cited on page 29)

Zotter, F. and Frank, M. (**2013**), “Efficient phantom source widening,” *Archives of Acoustics* **38**(1), pp. 27–37. (Cited on pages 24 and 27)