Binaural Scene Analysis with Multidimensional Statistical Filters

C. Spille, B. T. Meyer, M. Dietz and V. Hohmann

1 Introduction and Overview

Binaural hearing in humans has been investigated for more than a century. Thompson [63] and Rayleigh [52] identified differences in arrival time and in intensity between the left and the right ear to be the dominating cues for direction estimation. These cues are commonly termed as interaural time differences (ITD) and interaural level differences (ILD). Thompson [62] suggested that binaural sensitivity is not caused by acoustic interference, for instance, via the Eustachian tubes, but rather by neural processing in the brain. It took more than 50 years before the first conceptual model of neural ITD coding was suggested by von Békésy [65]. He suggested a model, where a population of neurons is excited by signals from one ear and inhibited by those from the other. The total population response then codes the interaural differences; this concept is referred to as *rate code*. Another two decades later Jeffress [28] suggested an alternative coding concept stating that the neural signals from each ear are delayed on counterdirected pathways, which act as *delay lines*. Along the pathways the two differently delayed signals are compared by coincidence neurons, which are activated if the signals arrive in coincidence. Due to the increase in relative delay along the delay lines, the position of the active coincidence neuron along the line indicates the ITD, what is known as *place coding*. Coincidence detection along counterdirected delay lines mathematically resembles cross-correlating the left and right signals. Probably because of these conceptual and mathematical simplicities the Jeffress model became the standard model type for developing and evaluating binaural processing models—see, for example, [35, 59, 60]—which became computationally tractable with the advent of digital computer. These models were able to explain a vast range of the binaural phenomena known experimentally from binaural psychoacoustics in humans. Variants of this model concept, for instance a subtraction of left

C. Spille · B. T. Meyer · M. Dietz · V. Hohmann (✉)
Department of Medical Physics and Acoustics, University of Oldenburg,
26111 Oldenburg, Germany
e-mail: volker.hohmann@uni-oldenburg.de

J. Blauert (ed.), *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-37762-4_6,

and right input along counterdirected delay lines have also been applied successfully [5] but all of these models are based on delay lines.

Physiological evidence for the existence of axonal delay lines was first found by Carr and Konishi [8] in the brainstem of barn owls. Although the auditory system of mammals was known to be significantly different from that of birds, the Jeffress model remained the standard approach for modeling binaural processing in mammals—see [41] for a review. However, recent evidence from physiology [42], functional magnetic resonance imaging [61], evoked potential measurements [53] and psychoacoustics [50] indicates that it is difficult to explain the mammalian data with the Jeffress model of place coding. Physiological data [4, 42] suggest that ITD might be coded in terms of the rate of firing of binaurally sensitive neurons—so-called *rate coding*—thus rather supporting the original hypothesis of von Békésy [65]. Based on these findings, Dietz et al. [17] developed a computational rate-coding model of binaural processing in humans that did not use any delay lines and that was based on the interaural phase difference (IPD) instead of the ITD. In line with recent models—for example, [20]—and psychoacoustical findings, but in contrast to earlier models that included an explicit temporal integration to model *binaural sluggishness*, this model reflects the high temporal resolution of the binaural system. Furthermore, it processes envelope and carrier IPDs in different channels, explaining psychoacoustic experiments that traded binaural cues of the envelope and carrier, respectively [14]. A further binaural processing model with the same motivation of rate coding has recently been suggested by Pulkki and Hirvonen [51].

All models of binaural processing in humans simulate the frequency selectivity initially provided by the cochlea by processing binaural information in frequency subbands. For this, many models use a linear Gammatone filterbank with auditory frequency and time resolution—see [27]. From a signal processing point of view, this renders the distinction between delay-line models and the IPD model by Dietz et al. [17] difficult: The subband signals are, on short time scales, almost sinusoidal, which means that ITD, as measured by cross-correlation, and IPD, as measured by the IPD model, are almost indistinguishable. A recent study designed to disambiguate these two approaches by studying the ability to lateralize stimuli with an ITD only in the second order envelope [16] does indeed hint towards the non-existence of long internal delays. Van der Heijden and Trahiotis [64] deduced from human psychoacoustic data of tone detection in double delayed noise stimuli that short internal delays up to 750 μs are indeed in operation. However, neither of these psychoacoustic experiments can distinguish between the general concepts of rate and/or place coding.

In the experimental study presented below, the IPD model [17] is used as a front-end, because it provides the required high temporal resolution for detecting short *glimpses* of robust binaural information. Similar results, however, can also be achieved with a cross-correlation model [20, 37, 54] or with the subtraction-based Breebaart model—[34], this volume.

Whereas binaural models mimic many of the distinguished capabilities of binaural hearing and have extensively been applied to investigate and simulate psychoacoustic data of binaural perception of artificial stimuli—see references above—as well as of speech—for example, [2]—their technical application to audio processing in hearing

aids, mobile audio devices, robotics, hands-free audio communication and speech-based computer interfaces is still limited. The hypothesis underlying this chapter is that this limitation is due to a *missing link* between basic binaural capabilities, as modeled by the binaural models mentioned above, and audio processing and interpretation of difficult acoustic conditions characterized by superposed speech, noise and reverberation. The latter requires inference about the causes of the observed auditory input in order to be able to *decode* the acoustic scene, that is, to identify and segregate different sources, or to select desired sound sources. This *cognitive* part of the processing is the missing link and constitutes enabling technologies for technical applications of binaural models. In this study, the possibilities of filling the missing gap by extending a binaural model towards interpreting the acoustic scene, that is, *computational binaural scene analysis*, are demonstrated by improving automatic speech recognition (ASR), of superposed spatially-moving speech signals.

Auditory scene analysis (ASA), in the sense of low-level cognitive processing of acoustic input in humans, has extensively been investigated in the literature, and many inference principles of the auditory system have been identified [6]. The most important acoustic cues used by the hearing system were identified to be harmonicity, periodicity, common onset—namley, synchronous increase in level across several auditory frequency bands—and the frequency-dependent binaural cues. Each of these cues only provides a small part of the information needed to decode the acoustic scene, namely, regarding sources being present, spatial configuration of the sources, room characteristics etc. Therefore, evidence from many of these different cues has to be integrated for scene interpretation, including integration across auditory frequency subbands and time. Evidence from electrophysiological, EEG data and functional magnetic resonance imaging (fMRI) in humans has led to the interpretation that the cognitive system, including the hearing system, performs cue integration by comparing the current sensory input to hypotheses about the expected observation [47, 70]. The system adapts via neural adaption to the expected input and codes only deviations of the input from the expectation—see [45]. This means that basically only *novelty* is processed by the nervous system, that is, sensory information that deviates from the hypotheses. Deriving hypotheses about the expected observation from earlier observations of the input requires a dynamic predictive model of the auditory scene and is thus part of recent cognitive inference models—see [66] for a recent approach of modeling novelty processing in auditory evoked EEG responses, and [21] for modeling cortical inference circuits.

Recent evidence from fMRI experiments in humans show that the premotor cortex is active during the perception of distorted speech, but not active when music is played. This suggests that premotor activity may facilitate interpreting speech when the input is sparse [46]. In the light of the proposed hypothesis-driven inference model, results are consistent with the notion that the premotor areas responsible for speech production might be used to generate the hypotheses when perceiving speech in difficult conditions—compare the motor theory of speech perception [69]. In other fMRI experiments, Bushara et al. [7] examined the neural binding between hearing and vision. Correlation of auditory and visual stimuli was found to be correlated with reduced brain activity when binding occurred. This suggested that the cognitive

system explores several hypotheses about the expected observation across modalities at a time, that is, generates *competing hypotheses*, and that their representations have mutually inhibitory interactions.

From a signal processing point of view, the principle of *novelty processing* agrees well with Bayesian inference methods, which are frequently used in computational-auditory-scene analysis (CASA) approaches—compare [68]. In particular, inference from competing hypotheses can be implemented numerically by sequential Monte-Carlo methods, which will be briefly introduced in the framework of computational binaural scene analysis in the next section. It is argued that CASA based on hypothesis-driven computing using predictive models is a key to successfully applying binaural models to decoding the acoustic scene. Three major limitations requiring this approach will be briefly outlined in the following.

- *Ambiguity of the source separation problem* If the acoustic scene is composed of more than two sources and is received by only two sensors, left and right ear, the separation problem becomes ambiguous. In other words, the source signals cannot be reconstructed from the superposition using linear methods, such as linear microphone-array processing. Disambiguation is possible, however, by predictive models that limit the number of possible explanations of the scene—see, for example, [44].
- *Random fluctuations of signal-derived parameters* Diffuse background noise and reverberation impose statistical fluctuations on all signal-derived features, strongly reducing the statistical evidence provided by a single observation of the respective feature—see, for example, [43]—for a quantification of the fluctuations of binaural features in real acoustic conditions. Predictive statistical models explicitly model the noise and perform a statistical combination of several noise-deteriorated parameters, allowing a noise-robust extraction of information [43].
- *Missing information* Superposed daily-life signals overlap significantly in the time-frequency domain. Thus, a significant part of the information on the different sources is completely masked in the time-frequency domain. In order to separate the sources, predictive models are required to fill the missing information—compare [11].

Every natural or artificial cognitive system has to deal with these limitations and thus requires some structure that collects evidence from noisy, ambiguous and partly missing information. In this study, the principle of competing hypotheses based on a predictive model is applied, in order to achieve this task.

The remainder of this chapter is organized as follows: First, the basic approach to implement principles of computational binaural scene analysis is described in Sect. 2, and existing studies on the subject are reviewed in Sect. 3. Then, the approach to improve ASR using computational binaural scene analysis is introduced in Sect. 4. For this, the binaural model of Sect. 4.1, the statistical properties of its output with respect to sound source localization, see Sect. 4.3, the multidimensional statistical filter that tracks the location of superposed moving speakers, Sect. 4.4, the beamformer that is directed to the desired speaker and its adaptation by the location

tracks, Sect. 4.5 and, finally, the ASR system that recognizes speech at the output of the beamformer, Sect. 4.6, are described. Results from an ASR task are detailed in Sect. 4.7. Last, the chapter is summarized and conclusions are given in Sect. 5.

2 Computational Binaural Scene Analysis

In this section the focus is on computational binaural scene analysis systems that are based on competing hypothesis. Figure 1 shows the principle processing blocks of the proposed system that builds upon the model proposed by Nix and Hohmann [44]. Key to this approach is the use of *a priori* knowledge about the sound sources S that compose the current acoustic scene to generate a set of hypotheses H. Each hypothesis develops in time and represents a possible state of the sound sources, that is, a set of parameters that describe the exact configuration of all sources, such as, source position, pitch, formant frequencies or vocal tract parameters—depending on the type of source. In each time frame, hypotheses are checked against the observation, which is composed of a number of signal features O computed from the binaural audio input.[1] The likelihood of the observation to occur under the assumption that the hypothesis is true is computed and assigned to each hypothesis. This means that

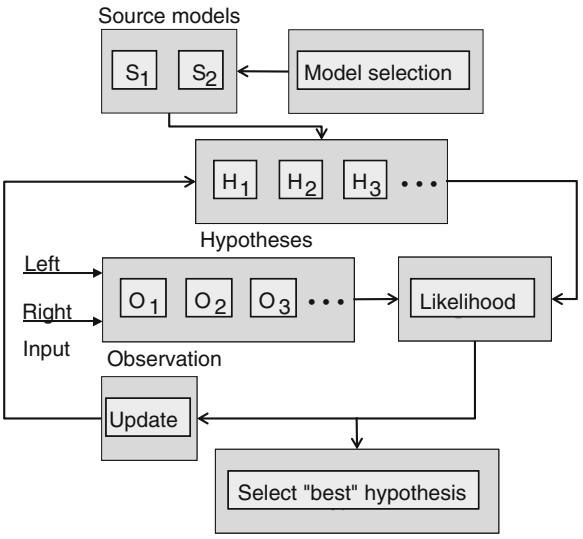

Fig. 1 Block diagram of a computational binaural scene analysis system based on competing hypothesis. See the text for details

[1] Note that this approach can be extended to more inputs, for example, multiple microphones or audiovisual input, or might be restricted to a single input.The current study covers its application to binaural input signals like recordings from a dummy head.

evidence from many features O is integrated and merged into a single likelihood value, such as feature integration across frequency. The definition of this likelihood function is challenging and its complexity depends on the number of state parameters and the number of observed features. It should reflect the relation between source parameters and observation as good as possible. Note that the likelihoods are stochastically distributed, because the observation is generally fuzzy, even if the state of the sound source is fixed, for instance, binaural features fluctuate even for fixed location of the source, as pointed out above. Note also that assigning likelihoods can be described as extracting the novelty about the sound sources embedded in the observation. If the novelty is high for a specific hypothesis, the observation does not match the expectation set by the hypothesis, and it will be assigned a low likelihood. If the novelty for a specific hypothesis is low, the hypothesis will be assigned a high likelihood. The set of hypotheses and their assigned likelihoods represents the distribution of possible states and thus the current estimate of the auditory scene.

Finally, based on the likelihood and *a priori* known dynamics of the sound sources, which is restricted by physical constraints, such as smooth pitch and location contours or limited rate of change of vocal tract parameters, each hypothesis is updated, that is, the parameter set associated with the hypothesis is changed. By this, the expectation about the state present in the next time step is established. This update function also employs a stochastic factor in most applications, for instance, the location of a sound source might be updated according to a random-walk process. Note that two identical hypotheses develop differently in time due to this random component.

The set of hypotheses and their assigned likelihoods represents the inference about the causes of the sensory input. In many applications, the hypothesis that was assigned the maximum likelihood is taken as the *best* hypothesis. The functioning of the approach very much depends on whether the applied source models S match the sources present in the audio input. A mismatch automatically means a fundamental misinterpretation of the scene. For example, if the system would erroneously select a speech signal to be present in the scene, parameters like formant frequencies, pitch and the temporal evolution of these parameters would be estimated and interpreted as the state of a speech signal—which is actually not present. Therefore, the *model-selection* block in Fig. 1 is most relevant. A biological system has to select or estimate the appropriate models based on information present in the sensory input. Many technical applications assume that the appropriate source models are *a priori* known, that is to say, they omit the model-selection block. In this case, models are fixed and only the hypotheses are being updated dynamically.

Common mathematical approaches to implementing a system according to Fig. 1 are so-called sequential Monte-Carlo methods, in particular particle filtering. Arulampalam et al. [1] provide a tutorial on generic particle filters, which shall be introduced briefly here—for mathematical details the reader is referred to the literature. Figure 2 shows a block diagram of a generic particle filter. The circle of processing blocks is performed for each time instance. A *state* is a mathematical description of the current configuration of each sound source and corresponds to the

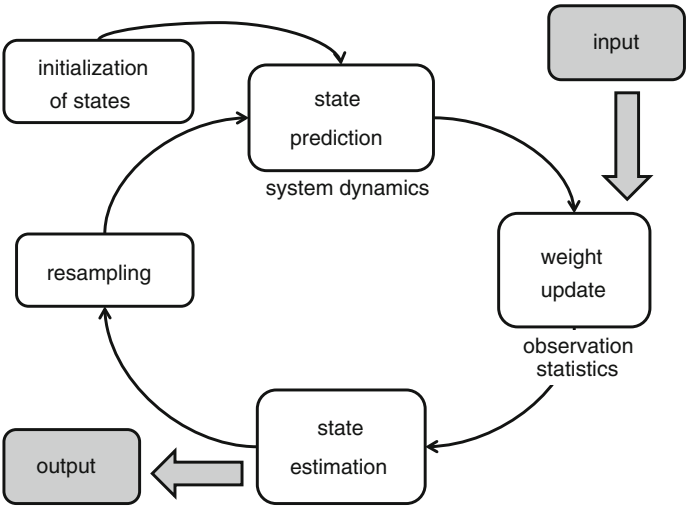

Fig. 2 Block diagram of a particle filtering algorithm. See the text for details

hypothesis from Fig. 1. It is assigned a *weight* that corresponds to the likelihood from Fig. 1. The combination of a state and its weight is called a *particle*. At system onset, before the first input is taken, a set of particles is initialized, for instance, by sampling the states from an equal distribution across all possible states and assigning random weights. *System dynamics* implements a dynamic model of each sound source by a random mathematical function and represents the update function from Fig. 1. Updating the hypotheses—Fig. 1—is implemented by predicting the future state from the current state and the system dynamics separately for each particle. In the next step, the weight of each particle is updated by an *observation statistics*, which corresponds to the likelihood computation from Fig. 1. The observation statistics links the input, namely, the observation, and the weights by increasing the weight for states that are likely given the input, and vice versa. Even if the observation statistics is not identical to the *true* likelihood function, the set of particles, namely, the states and their normalized assigned weights, represents a sampled version of the *true* likelihood function under very general constraints [19]. *State estimation* means the selection of the particle with the highest weight, which denotes the filter output. Some implementations output the expected value across the set of particles instead. The last step in the processing chain is the resampling step, which does not find a correspondence in Fig. 1. Resampling means that particles are discarded if their weights fall below a certain minimum and are replaced by randomly selected particles that are similar to the particles with the highest weights. The processing chain then begins again with the prediction step, that is, time is taken one step forward.

3 Examples from Literature

In the following, some examples of computational binaural scene analysis are briefly reviewed. Note that this review is far from complete; the studies presented here, however, will ease the access to the large body of literature on this topic. A comprehensive overview of CASA techniques—not necessarily binaural—can be found in the book edited by Wang and Brown [68].

A very early approach of binaural scene analysis was introduced by Lyon [36]. Peaks of the subband interaural cross-correlation function identified the location of sounds and time-varying filters steered by these peaks segregated a directional source. The approach pursued here is similar, but relies on a different binaural model and includes particle filtering for modeling source continuity.

Nix and Hohmann [44] presented a binaural-scene-analysis algorithm that tracked the azimuthal direction of arrival (DOA), and the power spectrogram of each of two superposed and moving speech signals. As the observation, short-term FFT-spectra and frequency-specific binaural parameters, ILD and ITD, derived from the spectra were used. As a source model, a first-order Markov process was used that could generate a plausible succession of speech spectra from a random process. For this, typical speech spectra and their transition probabilities were derived using a cluster analysis method from a large speech database that contained many hours of speech. $N = 10{,}000$ typical spectra were used in the Markov process, that is, the transition matrix contained $N \times N$ entries. For tracking the time-course of source location and spectra, a particle filtering approach was used. 100,000–1,000,000 particles were used in the different experiments. Each particle contained the source configuration, namely, azimuth directions and current short-time spectra of both sources. The authors reported that, on the same signal, some runs of the filter succeeded in tracking the sources correctly, whereas other runs failed. This shows the dependence of the system on the random initialization of the particles and shows that the particle filter may fail even if the source model perfectly matches the sources present in the input signal. For the successful runs of the filter, the algorithm was able to track azimuth and magnitude spectra from two superposed speech signals. Evaluations of the signal-to-noise ratio (SNR) showed that the algorithm was able to improve the SNR at input SNRs around zero or below, which is difficult to achieve with algorithms that do not use speech models—compare [44]. The computational effort, however, was very high.

Dietz et al. [15] used the perceptually and physiologically inspired IPD model for estimating the azimuthal DOA of superposed directional sound sources, including free field conditions with up to five concurrent speakers, three concurrent speakers in background noise and one speaker in reverberation. Key to the IPD model is that only those time-frequency segments contribute to the DOA estimate that have a high interaural coherence, similar to Faller and Merimaa [20] and [34] this volume. Those segments usually occur during short instances of time, often in the order of a few tens of milliseconds, when one sound source dominates the binaural input. By processing each of these high-coherence segments as a single event called *glimpse*, a sparse

representation of the binaural features is generated, which is in accordance with recent physiological evidence. A glimpsing representation is especially reasonable in strongly modulated signals such a speech [11]. Tracking of the sound sources, that is, estimating its DOA from the sequences of *glimpses* was achieved using a particle filter by Särkkä et al. [55] that handles sparse input. This filter implementation solves the linear parts of the estimation process with a Kalman filter and leaves the nonlinear parts to the particle filter. This approach is called *Rao-Blackwellized particle filter*. The IPD model in combination with the particle filter by [55] is used as the basis for the experimental study presented in this chapter. An elaborate analysis of DOA estimates from the model is given below.

Woodruff and Wang [71] describe a binaural localization framework for multiple sources in noisy and reverberant conditions. Monaural source segregation was used to increase the robustness of azimuth estimates from a binaural input and was shown to improve performance relative to binaural-only methods. This framework also allows model selection or adaption in the sense that an azimuth-dependent model of binaural features allows for adaptation to new environments.

Christensen et al. [10] introduce a speech fragment approach to localizing multiple speakers in reverberant environments. Key to this approach is that binaural and pitch information is sampled from time-frequency regions, so-called *fragments*, that are likely to be dominated by one of the speakers. This method is reported to improve localization performance by up to 24 % compared to a state-of-the-art localizer.

3.1 Application to Automatic Speech Recognition

Mel-frequency cepstral coefficients (MFCCs) are one of the standard features for today's ASR systems [13]. They effectively encode the spectral envelope of short-time segments of speech, perform well for acoustically clean conditions and reflect properties of the auditory system only to a limited extent. Auditory-inspired pre-processing of speech signals, however, has also been shown to be a useful approach in automated speech processing tasks. Applications include the identification of speakers [40], this volume and [38, 39] as well as automatic speech recognition, ASR, for which auditory frontends have been shown to increase the robustness in the presence of noise and reverberation [58]. Examples of the large number of studies following this approach range from the integration of signal processing strategies known to be employed in the inner ear [26] to the application of filters resembling pattern observed in the primary auditory cortex of mammals [31].

Across-frequency binaural processing has also been investigated in the framework of binaural speech recognition. Palomäki and Brown [48] compare across-frequency and within-frequency processing in combination with internal noise in a computational model of binaural speech recognition. Palomäki et al. [49] use the statistics of binaural features to identify unreliable spectro-temporal segments. Unreliable segments are treated as *missing data* by the speech recognition system. In other words, no evidence is provided by this segment and the system's speech model re-generates

the missing data in the estimation process. These missing data techniques have been elaborated further [23, 32] and have been shown to be very successful in rendering ASR more robust in noisy conditions. In this chapter, these techniques are not employed, but it would be possible to use a missing data recognizer directly on the output of the binaural model, which establishes a means to define missing data due to its sparseness. Instead, following the philosophy of the AABBA project, auditory-inspired processing is employed in form of the binaural model described in the next section. Note that [40], this volume, elaborate further on missing data techniques.

4 ASR in Multi-Speaker Conditions Using Binaural Scene Analysis

Figure 3 shows a block diagram of the whole processing chain from the raw speech data to the ASR system. Speech data is used to generate moving speakers by convolving it with recorded 8-channel HRIRs—2 in-ear channels and 3 channels from each of two behind-the-ear (BTE) hearing aids. The in-ear signals are fed into the binaural model that is employed to estimate the direction of arrival of spatially distributed speakers—compare Fig. 1 block *Observation*. A particle filter is then used to keep track of the positions of the moving speakers. This relates to the blocks *Source Model*, *Hypotheses*, *Likelihood*, *Update* and *Select best hypotheses* in Fig. 1. The particle filter's output is used to steer a beamformer, enhancing the 6-channel speech signal that is to be transcribed by an ASR system. In the following sections each of these processing steps is described in more detail.

4.1 Binaural Model Structure

The main aim of this study was to apply an auditory binaural model, the IPD model [17], to automatic speech recognition. The IPD model has previously been extended and applied to direction of arrival (DOA) estimation [15] which, in turn, has been

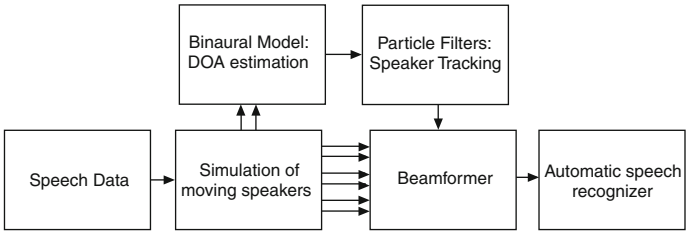

Fig. 3 Block diagram of the experimental setup. See the text for details

applied for binaural synthesis [57]. The model extracts binaural parameters such as IPD and ILD in a way that mimics the performance of the human auditory system. Four specific aspects of temporal auditory processing were of specially in the focus of the IPD model, particularly,

- High temporal resolution.
- Limited phase-locking range.
- Use of temporal envelope disparities.
- A limited internal ITD range.

For the sake of consistency the implementation was kept unchanged from Dietz et al. [15], even though further improvements such as DC-offset free modulation filters [16] and aspects of pre-binaural adaptation—for example, [18, 30]—have recently been suggested to model psychoacoustic performance of envelope ITD sensitivity and binaural tuning of single cells more realistically.

Figure 4 gives an overview of the processing stages of the IPD model. For the stages up to the extraction of the interaural transfer function, ITF, the IPD and the ILD were adopted from [17]. Later stages of Fig. 4, from interaural vector strength (IVS) to DOA glimpse extraction, belong to the binaural cue selection. Both the IPD model and the binaural cue selection are described in [15]. In the following only the conceptually relevant aspects are briefly reviewed.

Most importantly, the signals were analyzed in 23 auditory filters in the range of 200 Hz to 5.0 kHz. Considering the human limit to binaurally exploit fine-structure information above ∼1.4 kHz, the fine-structure filter is only implemented in the 12 lowest auditory filters below 1.4 kHz. Envelope IPDs are derived from all 23 filters, but are not exploited in the current study.

A problem occurring especially for fine-structure IPDs in filters above 700 Hz is that their corresponding ITDs do no longer cover the whole range of possible interaural delays, resulting in an ambiguity of direction. Inspired by psychoacoustic findings such as time–intensity trading—for instance, [33]—the sign of the ILD is employed here to extend the unambiguous range of IPDs from $[-\pi,\ \pi]$ to $[-2\pi,\ 2\pi]$. Accordingly, the frequency range for unambiguous fine-structure IPD-to-azimuth mapping is extended from ∼700 to 1400 Hz. IPD-to-azimuth mapping itself is performed with a previously learned mapping function.

As argued in [15], the IPD model does not rely on cross-correlation, and, thus, interaural coherence (IC) is not directly assessable. However, Goupell and Hartmann [22] have shown that the temporal fluctuations of the interaural functions are possibly an even better measure for psychoacoustic decorrelation sensitivity. Therefore, in the IPD model, the IPD fluctuations are directly accessible and are specified in the form of the interaural vector strength (IVS). The IVS was used to derive a filter mask which consists of a binary weighting of the interaural parameters based on a threshold value of $\mathrm{IVS}_0 = 0.98$.

By processing each of these high-coherence segments as a single event called *glimpse*, a sparse representation of the binaural features is generated from the median value of the azimuth estimation of this segment. If the IVS constantly exceeds IVS_0 for more than 20 ms, a new glimpse is assigned from the same segment. Depending

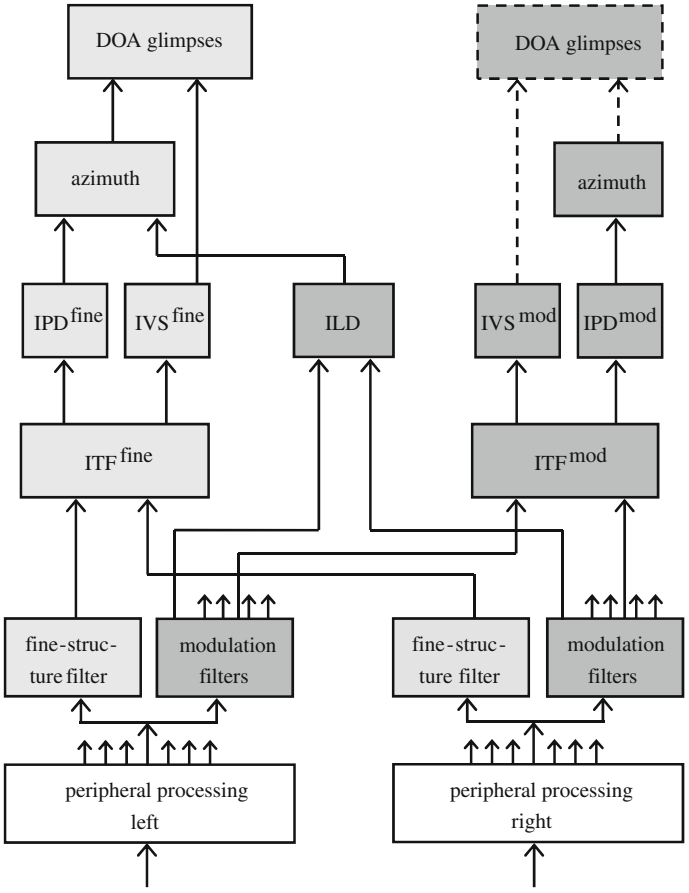

Fig. 4 Processing stages of the IPD model from [15]. Peripheral processing splits the input signal up into 23 frequency channels from 200 to 5000 Hz. Only one of these channels is drawn for the further processing blocks. IPDs and interaural vector strength (IVS) are derived from one fine-structure and from several modulation filters. Fine-structure information is only derived in the 12 lowest frequency channels from 200 to 1400 Hz. In addition, the ILD is derived at the output of an envelope low-pass filter. The azimuth is derived from the IPDs with a previously stored frequency dependent mapping function. For fine-structure channels from 700 to 1400 Hz additional ILD information is employed to unwrap the IPD—see main text. Those azimuth estimates that occur during IVS > 0.98, result in so-called *glimpses*, which represent expectedly salient and sparse estimates of the direction of arrival. Within the current study only glimpses from fine-structure channels are considered

on the application it is decided whether the segments are grouped together to form a glimpse, or not. Here, for DOA estimation of stationary sounds, Sect. 4.3, it is not necessary, while it is highly beneficial for tracking applications such as the tracking of moving speakers, Sect. 4.4, in order to reduce the computational load.

4.2 Multi-Channel Speech Material

This section describes the multi-channel speech material used for the experiments. The first and second subsection describe the monaural speech corpus and the generation of the spatial multi-channel signals, respectively. The simulation of moving sources is presented in the last subsection.

Speech Data

The speech data used for the experiments consists of sentences produced by ten speakers—four male, six female. The syntactical structure and the vocabulary were adapted from the Oldenburg Sentence Test (OLSA) [67], where each sentence contains five words with ten alternatives for each word and a syntax that follows the pattern <name><verb><number><adjective><object>, which results in a vocabulary size of 50 words. The original recordings with a sampling rate of 44.1 kHz were downsampled to 16 kHz and concatenated—using three sentences from the same speaker. This resulted in sentences with a mean duration of 6.44 s, suitable for speaker tracking. For ASR experiments, the speech material was split into training and test sets with a total duration of 30 and 88 min, respectively. With this amount of speech data, a good ASR performance can be expected in relatively clean acoustics, whereas the estimation of acoustic models from noisy observations usually requires a larger database even for a relatively small vocabulary. Hence, the experiments presented in this chapter concentrate on the performance with one competing, moving speaker. The generation of training and test material is based on processing with a beamformer and is described in more detail in Sect. 4.6.

Generation of Multi-Channel Signals

Spatially localized and diffuse sound sources are simulated using a database of head-related impulse responses, the HRIR database, which features impulse responses recorded with three microphones from each of two behind-the-ear (BTE) hearing aids attached to left and the right ear and two in-ear microphones. The HRIRs used in this study are a subset of the database described in [29]: Anechoic free-field HRIRs from the frontal horizontal half-plane measured at a distance of 3 m between microphones and loudspeaker were selected. The HRIRs from the database were measured with a 5° resolution for the azimuth angles, which was interpolated to obtain a 0.5° resolution. The coordinate system is illustrated in Fig. 5.

Moving Speakers

The signals used throughout the experiments contain data of two moving speakers without interfering noise sources. Initial and final speaker positions were randomly drawn from a $-90°$ to $+90°$ azimuth interval, which represents the valid azimuth

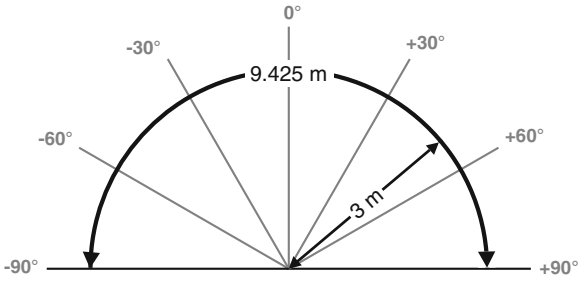

Fig. 5 Available azimuth range of the generated signals

range of the binaural model. The speakers moved linearly from the start to the end point for the duration of the respective stimulus and crossed their tracks with a 50 % probability. A frame-wise processing scheme was employed by applying 64 ms Hann windows with 50 % overlap and convolving each time frame with the respective HRIR. Since a source separation cannot be performed with a beamformer when the signals come from the same direction, boundary conditions were defined that guaranteed an average angle difference of at least 10°. Additionally, the minimal distance between the start and end points was set to 10 and 20° for non-crossing and crossing speakers, respectively.

4.3 Statistical Analysis of Binaural Features

As demonstrated in [15], the IPD model can be employed to localize several concurrent speakers. While the model suffers stronger from reverberation than normal hearing human listeners, its accuracy and performance is very good in free field multi speaker conditions. Even three speakers in noise at −6 dB SNR with same frequency characteristics as speech were robustly localized. While the number of speakers was only increased up to five in this previous study, Fig. 6 shows that even 6 concurrent speakers can be localized by analyzing the azimuth distribution in the 12 fine-structure channels over a few seconds. The time course of the azimuth estimate of an exemplary channel, $f_c = 1000$ Hz, is plotted in Panel (b). It can be seen that the estimate quickly oscillates between the six speaker positions. Over ten groups of glimpses per second indicate robust DOA estimates while transition periods that can contain any azimuth value are reliably suppressed by the IVS filter.[2]

[2] A demo folder containing the file *exp_spille2013* used to run the IPD model and to generate Fig. 6 is available in the *AMToolbox* [56].

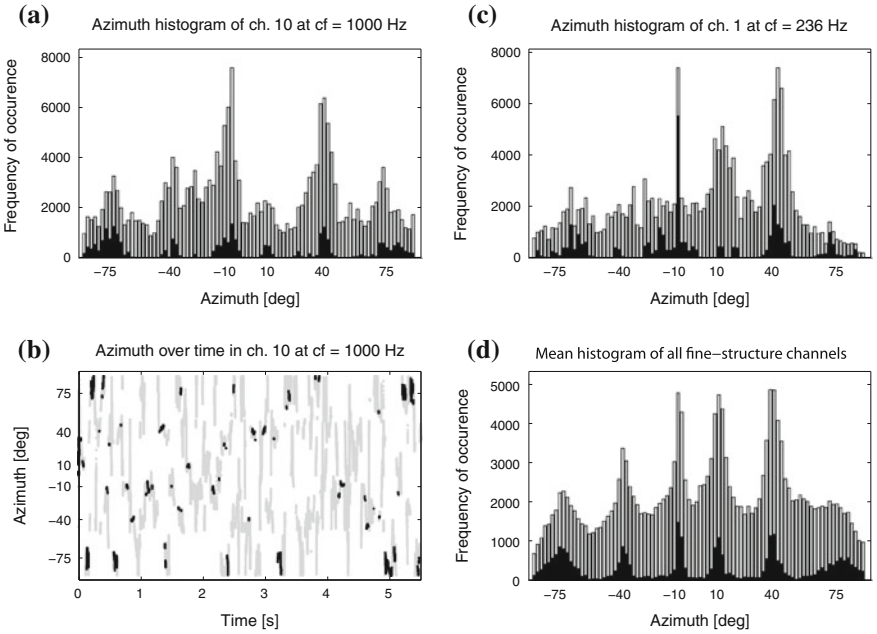

Fig. 6 Model output for six simultaneously speaking stationary speakers at −75, −45, −10, +10, +45, and +75°. The speech duration was 5.5 s. *Grey color* indicates all DOA estimates without IVS filtering, *black color* indicates DOA estimates with IVS filtering, $\mathrm{IVS}_0 = 0.98$. **a** Azimuth histogram of the fine-structure channel centered at $f_c = 1000$ Hz. **b** Time course of the azimuth estimation for the same channel and input signal as in panel (a). **c** Azimuth histogram of the fine-structure channel centered at $f_c = 236$ Hz with the same format and input signal as in (a). **d** Mean azimuth histogram of the twelve fine-structure channels. Same format and same input signal as in (a). It can be seen that the position of the speaker at +75° can only be determined with IVS filtering

4.4 Tracking Superposed Speakers

In the framework of the current application, the knowledge of the speaker positions is used to steer a beamformer that enhances the selected speaker by spatial filtering of the six BTE-microphone signals, that is, a binaural multi-channel beamformer. A tracking algorithm for multiple-speaker conditions was already implemented in [15], but only as a proof of concept. Here a more elaborate version is presented and its precision in several two-speaker scenarios is demonstrated.

Particle Filters and Monte-Carlo Data Association

The main challenge in the tracking of multiple targets is the mapping from observations—in this case, of DOA glimpses—to a specific target, which is a prerequisite for the actual tracking. In this chapter, an algorithm provided by

Särkkä et al. [55] is applied to solve this problem.[3] The main idea of the algorithm is to split up the problem into two parts—so-called *Rao-Blackwellization*. First, the posterior distribution of the data association is calculated using a equential-importance resampling SIR, particle-filtering algorithm. Second, the single targets are tracked by an extended Kalman filter that depends on the data associations. Rao-Blackwellization exploits the fact that it is often possible to calculate the filtering equations in closed form. This leads to estimators with less variance compared to the method using particle filtering alone [9]. For more details of the algorithms see [25, 55].

Application to Speaker Tracking

The tracking toolbox described in the previous section was applied to tracking the speakers from the DOA *glimpses* given by the IPD model. To apply the filter to the signals, the so-called dynamic model and the measurement model have to be defined. The dynamic model defines the temporal dynamics of the system and implements the block *state prediction* of Fig. 2. The state x of the system is determined by the actual position of the target, α, and velocity, v. x is a vector consisting of the elements α and v. The dynamic model is then given by

$$x_k = A_{k-1}x_{k-1} + q_{k-1}. \tag{1}$$

The matrix, A, is the transition matrix of the dynamic model and reflects the dynamics of the system. In this case it is given as

$$A_k = \begin{pmatrix} 1 & \Delta t_k \\ 0 & 1 \end{pmatrix}, \tag{2}$$

where $\Delta t_k = t_{k+1} - t_k$ is the time step between two states of the system. This means that the system's state at time step k is a linear progression of the system at time $k-1$ with constant speed plus some process noise, q_k, which is introduced to account for uncertainties in the system's development

$$x_k = \begin{pmatrix} \alpha_{k-1} + \Delta t_{k-1}\, v_{k-1} \\ v_{k-1} \end{pmatrix} + q_{k-1}. \tag{3}$$

The process noise is assumed to be a multivariate Gaussian with zero mean and covariance matrix

$$q_k = \begin{pmatrix} \frac{1}{3}\Delta t_k^3 & \frac{1}{2}\Delta t_k^2 \\ \frac{1}{2}\Delta t_k^2 & \Delta t_k \end{pmatrix} q_f, \tag{4}$$

which is calculated using the previously mentioned toolbox [24]. q_f is a process-noise factor that was set to 0.1 in this case.The prior distribution of the state x_0 (see

[3] The algorithm is part of a Matlab-Toolbox provided by [25].

block *initialization of states* in Fig. 2), is also a multivariate Gaussian of the form

$$x_0 \sim N(m_0, P_0),$$

where m_0 denotes the prior mean of the state and P_0 its prior covariance matrix containing the variances of the system's position and velocity that is set to

$$P_0 = \begin{pmatrix} 50 & 0 \\ 0 & 15 \end{pmatrix},$$

In other words, the actual position α has a variance of 50 deg^2 and the variance of the velocity is 15 m^2/s^2. The Kalman filter predicts the mean and the covariance of the state using the prior values together with the transition matrix, A, and the covariance matrix of the process noise, q. The equations for the predicted mean $\overline{m_k}$ and the predicted covariance $\overline{P_k}$ are as follows,

$$\begin{aligned} \overline{m_k} &= A_{k-1} m_{k-1} \\ \overline{P_k} &= A_{k-1} P_{k-1} A_{k-1}^T + q_{k-1}. \end{aligned} \tag{5}$$

Note that the process noise is only used for predicting the new covariance matrix. During the update step—block *weight update* in Fig. 2—these predictions are updated using the actual measurement, that is, *glimpses*, at time step k as well as the measurement model which describes the relation between the measurement and the state of the system. The measurement model is given by

$$y_k = H_k x_k + r_k, \tag{6}$$

where y_k is the actual measurement at time k, H_k is the measurement model matrix and r_k is the Gaussian measurement noise, $r_k \sim N(0, R)$. In the measurement model used here, only the position of the target is measured. This measurement can be corrupted by some noise reflecting the variance of the DOA estimation. Thus, the measurement model matrix, H, and the noise variance, R, are given by

$$H = (1 \quad 0)\ R = 50 \text{ deg}^2.$$

As the glimpses are sparse and occur with varying distance in time, the choice of the sampling interval is crucial. A sampling frequency equal to $1/\Delta t$ was chosen for the tracking algorithm and each glimpse was assigned to the nearest sampling point at this sampling rate. Glimpses were sampled at the original rate of the speech material. In the very rare cases that more than one glimpse fell in one bin, all but one glimpse were discarded.

Several sampling frequencies were tested and the minimum median-squared-error of the tracking was derived. For this, a dataset consisting of 71 sentences was used. A final sampling frequency of 500 Hz was chosen based on the results in Table 1.

Table 1 Median-squared-errors and their roots for the different sampling frequencies

Sampling frequency	Median-squared-error	Root median-squared-error
50	6.6615	2.5810
100	2.8361	1.6841
200	2.1954	1.4817
400	1.9353	1.3912
500	1.4857	1.2189
1000	1.5297	1.2368
1600	1.8244	1.3507

Speaker Tracking

The particle filter was initialized with a set of 20 particles using a known starting position of the first speaker, that is, the location variable of the first target was set to the position for all particles. The location variable of the second target was altered for each particle in equidistant steps throughout the whole azimuth range. Initial velocities were set randomly between $\pm 2\,\mathrm{m/s}$ for each target in each particle. The covariance matrix was equal for both targets and was set to P_0 as above.

If no glimpse is observed at time step t, the update step of the Kalman filter was skipped for this time step and the prediction was made based on the internal particle states. The range of the predicted angles was limited to the interval $[-90, 90]$ by setting all predictions outside that range to $-90°$ or $90°$, respectively.

Figure 7 presents two exemplary tracking results. The figure shows that the particle filter is able to track speakers even when they cross tracks, *left panel*. The tracking algorithm was evaluated by calculating the root median-squared error for each of the 9 data sets. On average the error was below 1.5°.

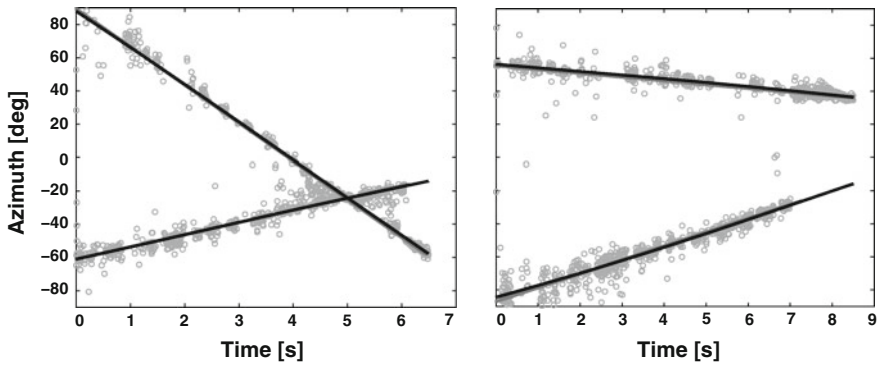

Fig. 7 Tracking results of a two-speaker scenario. *Light-grey circles* represent the glimpses produced by the binaural model—see text. *Dark-grey lines* represent the real azimuth angles of the speakers. *Solid black lines* show the smoothed estimates obtained by tracking

4.5 Steerable Beamformer for Source Selection

In the proposed application, a position estimate for both the target and concurrent speaker are required to control the beamformer parameters to either enhance the speech of a certain speaker or block out a concurrent speaker, thereby increasing the overall signal-to-noise ratio and subsequently lower the word error rates of an automatic speech recognizer. The beamformer employed here is a super-directive beamformer based on the minimum-variance distortionless-response principle [12] that used the six BTE microphone inputs jointly—three channels left and three channel from the right ear. In general, it suppresses the noise coming from all directions while not affecting the speech of the desired speaker. Additionally, the beamformer strongly suppresses the speech of the concurrent speaker which, in this setup, is has a higher impact than the enhancement of the desired source. Let W be the matrix containing the frequency domain filter coefficients of the beamformer, d_1 and d_2 the vectors containing the transfer functions to the microphones of speakers one and two respectively and Φ_{VV} the noise power-spectral density, PSD, matrix. Then, the following minimization problem has to be solved,

$$\min_{W} \; W^H \Phi_{VV} W, \qquad \text{with } W^H d_1 = 1 \text{ and } W^H d_2 = 0. \tag{7}$$

The solution to this is the minimum-variance distortionless-response beamformer [3]. The transfer functions in vectors d_1 and d_2 result from the impulse responses that are chosen based on the angle estimation of the tracking algorithm. The coherence matrix which is required to solve (7) is also estimated using the impulse responses used for generating the signals. Note that relying on the true impulse responses implies the use of *a-priori* knowledge not available in a real-world application, for which the impulse responses need to be estimated. The beamforming by itself therefore represents an upper bound, and will be extended to be used with estimated impulse responses in future work. However, since the IPD model, the tracking algorithm and the ASR system do not use such *a-priori* knowledge in reflecting realistic conditions, and robust methods for estimation of impulse responses exist, the results should still be transferable to real-world applications.

4.6 ASR System

Feature Extraction and Classifier

The benefits of the proposed processing chain for speech processing are analyzed by performing automatic speech recognition (ASR) on the output signals of the beamformer. The ASR system consists of a feature extraction and a classification stage.

The features extracted from speech should represent the information required to transcribe the spoken message and ideally suppress unwanted signal components.

For the experiments, the feature type most commonly applied in ASR, namely, *Mel-frequency cepstral coefficients* (MFCCs) [13] was chosen. These features effectively encode the smoothed short-time Fourier transform (STFT) magnitude, which is computed every 10 ms using overlapping analysis windows of 25 ms duration. Each frame of the STFT is processed by a mel-filterbank that approximates the frequency sensitivity of the human ear, compressed with the logarithm and transformed to cepstral parameters using a discrete cosine transformation. By selecting twelve lower cepstral coefficients, only the coarse spectral structure is retained. By adding an energy value and calculating an estimate for the first and second derivative, the so-called delta and double-delta features, to include some information about temporal dynamics on the feature level, 39-dimensional feature vectors were finally obtained.

The feature vectors are used without normalization to train and test the Hidden-Markov model (HMM) classifier, which has been set up as word model with each word of the vocabulary corresponding to a single HMM. During testing, the likelihoods of each HMM generating the observed sequence of feature vectors are compared and the word with the highest likelihood is selected. A grammar reflecting the fixed syntax of OLSA sentences is used to ensure a transcription with a valid OLSA sentence structure, in particular the following, <name><verb><number> adjective><object>, repeated three times due to the concatenation of sentences. The HMM used ten states per word model and six Gaussians per mixture and was implemented using the Hidden-Markov Toolkit (HTK) decribed in [72].

Training and Test Material

ASR training was carried out using sentences with one moving speaker, which were processed with the beamformer. The steering vectors of the beamformer were set to the true azimuth angles of the desired speaker instead of using the output of the complete processing chain including DOA estimation. This resulted in signals containing some beamforming artifacts, that is, the classifier was able to adapt to the resulting feature distortions and still carried the relevant information to create proper word models. The effects of speaker-dependent (SD) versus speaker-independent (SI) recognition was investigated by creating two training sets with the test speaker being either included in the training data, SD, or excluded from training, SI. The original data contained 71 long sentences each of which was used several times for the simulation of moving speakers, thereby increasing the amount of training material. Each sentence was processed four times with random start and end positions of the speakers, which resulted in 284 training sentences or 30 min, repectively, for the SD system and approximately 250 training sentences or 27 min for the SI system.

For *testing*, signals with two moving speakers were processed by the complete chain depicted in Fig. 3, one being the target source and the other one the suppressed source, and the recognition rate for the words uttered by the target speaker was obtained. To increase the number of test items, each speaker was selected as the target speaker once and the training/testing procedure was carried out ten times. As for the training set, the original 71 sentences were used for movement simulation

several times to further increase the number of test items and hence the significance of the results. For testing, a factor of 11 was chosen due to computational constraints. This resulted in a total number of 781 sentences or 88 min with randomised start and end positions for two speakers.

4.7 ASR Results

When using the complete processing chain that included the DOA estimation, tracking, beamforming, and ASR, a word-recognition rate (WRR) of 88.4 % was obtained for the speaker-dependent ASR system. When using a speaker-independent system, a word-recognition rate of 72.6 % was achieved. The data presented in the following were obtained with the speaker-dependent ASR system. When the ASR system cannot operate on beamformed signals, but is limited to speech that was converted to mono signals by selecting one of the eight channels from the behind-the-ear or in-ear recordings, the average WRR was 29.4 %. The variations of WRRs between channels were relatively small, ranging from 28.1 to 30.8 %. When the best channel for each sentence was selected, that is, the channel that resulted in the highest WRR for that specific sentence to simulate the best performance when limited to one channel, the average WRR was increased to 38.8 %.

It is interesting to note that the WRRs were very similar when analyzing crossing and non-crossing speaker tracks separately, namely, 88.3 and 88.4 %, respectively. An analysis of the average separation of speakers in ° showed that the overall accuracy was nearly constant for spatial distances ranging from 40 to 100°—Fig. 8a—but will definitely drop down for smaller distances. The average distance was, of course, significantly higher for non-crossing speakers, namely, 64.9°, than for crossing speak-

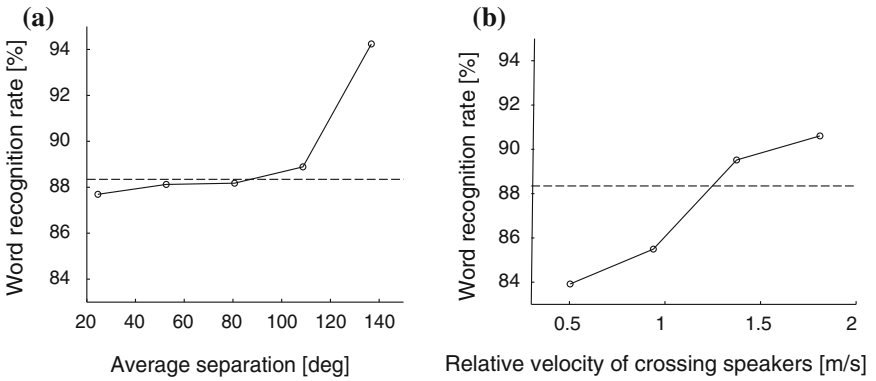

Fig. 8 **a** Word-recognition rate, WRR, of the ASR system to be dependent on the average separation of sources. The *dashed line* denotes the average WRR for all speaker tracks. **b** WRR for crossing speakers to depend on the difference speed of competing speakers. The *dashed black line* shows the recognition rate that was obtained for crossing speaker tracks

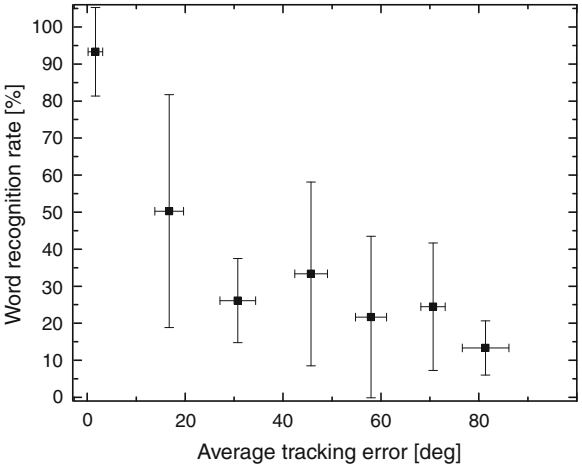

Fig. 9 Word-recognition rate versus average tracking error. The range of the tracking error was devided in equidistant bins. Data points denote the mean tracking error and the mean word-recognition rate. *Error bars* show the corresponding standard deviations

ers, 41.0°. But, due to the constraints in the case of crossing speakers, the average distance was high enough to not reduce the WRR. The parameter determining the average distance of crossing speakers is the relative velocity of speakers, where high relative velocities correspond to short durations of spatially close speakers—Fig. 8b. The constraints for starting and end positions that were chosen for crossing speakers resulted in an average relative speed difference of 1.2 m/s. Hence, the comparable high WRRs for crossing tracks can be attributed to the high relative velocity of speakers ensured by the contraints in signal generation.

The word-recognition rate also depends strongly on the localization accuracy: The overall localization accuracy was quantified by calculating the average tracking error, which is the root median squared error between the smoothed tracking estimates and the real azimuth angles of the speakers—see Table 1. Figure 9 shows that the WRR is highly dependent on the average tracking error where higher traking errors cause significantly lower WRRs.

5 Summary and Conclusions

This study provided an overview of computational auditory scene analysis based on binaural information and its application to a speech recognition task. The usability of the IPD model in automated speech processing was demonstrated by performing a DOA estimation for stationary and moving speakers. For the moving-speaker scenario, it was also shown that the binaural model enables efficient tracking and greatly increases the performance of an automatic speech recognition system in situ-

ations with one interfering speaker. The word-recognition rate (WRR) was increased from 30.8 to 88.4 %, which shows the potential of integrating models of binaural hearing into speech processing systems. It remains to be seen if this performance gain in anechoic conditions can be validated in real-world scenarios, that is, in acoustic conditions with strong reverberation, several localized noise sources embedded in a 3D-environment compared to the 2D simulation presented here or with a changing number of speakers. Follow-up studies are suggested that explore a combination of a binaural model, a tracking system and beamforming for other problems in speech and hearing research, such as speaker identification, speaker diarization or the improvement of noise reduction in hearing aids.

Acknowledgments Supported by the DFG—SFB/TRR 31 *The active auditory system*, URL: http://www.uni-oldenburg.de/sfbtr31. The authors would like to thank M. Klein-Hennig for casting the IPD model code in the AMToolbox format, D. Marquardt and G. Coleman for their contributions to the beamforming algorithm, M. R. Schädler for sharing the code of the OLSA recognition system, H. Kayser for support with the HRIR database, and two anonymous reviewers for constructive suggestions.

References

1. M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear / non-Gaussian bayesian tracking. *IEEE Trans.* Signal Process., 50:174–188, 2002.
2. R. Beutelmann and T. Brand. Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 120:331–342, 2006.
3. J. Bitzer and K. U. Simmer. Superdirective microphone arrays. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 2. Springer, 2001.
4. A. Brand, O. Behrend, T. Marquardt, D. McAlpine, and B. Grothe. Precise inhibition is essential for microsecond interaural time difference coding. *Nature*, 417:543–547, 2002.
5. J. Breebaart, S. van de Par, and A. Kohlrausch. Binaural processing model based on contralateral inhibition. I. Model structure. *J. Acoust. Soc. Am.*, 110:1074–1088, 2001.
6. A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT Press, 1990.
7. K. O. Bushara, T. Hanakawa, I. Immisch, K. Toma, K. Kansaku, and M. Hallett. Neural correlates of cross-modal binding. *Nat. Neurosci.*, 6:190–195, 2003.
8. C. E. Carr and M. Konishi. Axonal delay lines for time measurement in the owl's brainstem. *Proc. Natl. Acad. Sci. U. S. A.*, 85:8311–8315, 1988.
9. G. Casella and C. Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83:81–94, 1996.
10. H. Christensen, N. M. N. Ma, S. N. Wrigley, and J. Barker. A speech fragment approach to localising multiple speakers in reverberant environments. In *IEEE ICASSP*, 2009.
11. M. Cooke. Glimpsing speech. *Journal of Phonetics*, 31:579–584, 2003.
12. H. Cox, R. Zeskind, and M. Owen. Robust adaptive beamforming. *IEEE Trans. Acoust., Speech, Signal Process.*, 35:1365–1376, 1987.
13. S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Process.*, 28:357–366,1980.

14. M. Dietz, S. D. Ewert, and V. Hohmann. Lateralization of stimuli with independent fine-structure and envelope-based temporal disparities. *J. Acoust. Soc. Am*., 125:1622–1635, 2009.
15. M. Dietz, S. D. Ewert, and V. Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Commun*., 53:592–605, 2011.
16. M. Dietz, S. D. Ewert, and V. Hohmann. Lateralization based on interaural differences in the second-order amplitude modulator. *J. Acoust. Soc. Am*., 131:398–408, 2012.
17. M. Dietz, S. D. Ewert, V. Hohmann, and B. Kollmeier. Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences. *Brain Res*., 1220:234–245, 2008.
18. M. Dietz, T. Marquardt, D. Greenberg, D. McAlpine. The influence of the envelope waveform on binaural tuning of neurons in the inferior colliculus and its relation to binaural perception. In B. C. J. Moore, R. Patterson, I. M. Winter, R. P. Carlyon, H. E. Gockel, editors, *Basic Aspects of Hearing: Physiology and Perception*, chapter 25. Springer, New York, 2013.
19. A. Doucet, N. de Freitas, and N. Gordon. An introduction to sequential Monte Carlo methods. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*. Springer, 2001.
20. C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Am*., 116:3075–3089, 2004.
21. K. Friston and S. Kiebel. Cortical circuits for perceptual inference. *Neural Networks*, 22:1093–1104, 2009.
22. M. J. Goupell and W. M. Hartmann. Interaural fluctuations and the detection of interaural incoherence: Bandwidth effects. *J. Acoust. Soc. Am*., 119:3971–3986, 2006.
23. S. Harding, J. P. Barker, and G. J. Brown. Mask estimation for missing data speech recognition based on statistics of binaural interaction. *IEEE T. Audio. Speech*., 14:58–67, 2006.
24. J. Hartikainen and S. Särkkä. Optimal filtering with Kalman filters and smoothersa Manual for Matlab toolbox EKF/UKF. Technical report, Department of Biomedical Engineering and Computational Science, Helsinki University of Technology, 2008.
25. J. Hartikainen and S. Särkkä. RBMCDAbox-Matlab tooolbox of rao-blackwellized data association particle filters. Technical report, Department of Biomedical Engineering and Computational Science, Helsinki University of Technology, 2008.
26. H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am*., 87:1738–1752, 1990.
27. V. Hohmann. Frequency analysis and synthesis using a Gammatone filterbank. *Acta Acustica united with Acustica*, 88:433–442, 2002.
28. L. a. Jeffress. A place theory of sound localization. *J. Comp. Physiol. Psychol*., 41:35–39, 1948.
29. H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier. Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP Journal on Advances in Signal Processing*, 2009:298605, 2009.
30. M. Klein-Hennig, M. Dietz, V. Hohmann, and S. D. Ewert. The influence of different segments of the ongoing envelope on sensitivity to interaural time delays. *J. Acoust. Soc. Am*., 129:3856–3872, 2011.
31. M. Kleinschmidt. Methods for capturing spectro-temporal modulations in automatic speech recognition. *Acta Acustica united with Acustica*, 88:416–422, 2002.
32. D. Kolossa, F. Astudillo, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, R. Martin. CHiME challenge : Approaches to robustness using beamforming and uncertainty-of-observation techniques. *Int. Workshop on Machine Listening in Multisource, Environments*, 1:6–11, 2011.
33. A.-G. Lang and A. Buchner. Relative influence of interaural time and intensity differences on lateralization is modulated by attention to one or the other cue: 500-Hz sine tones. *J. Acoust. Soc. Am*., 126:2536–2542, 2009.
34. N. Le Goff, J. Buchholz, and T. Dau. Modeling localization of complex sounds in the impaired and aided impaired auditory system. In J. Blauert, editor, *The technology of binaural listening*, chapter 5. Springer, Berlin-Heidelberg-New York NY, 2013.
35. W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals. *J. Acoust. Soc. Am*., 80:1608–1622, 1986.

36. R. F. Lyon. A computational model of binaural localization and separation. In *IEEE ICASSP*, volume 8, pages 1148–1151, 1983.
37. T. May, S. Van De Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE T. Audio. Speech*., 19:1–13, 2011.
38. T. May, S. Van De Par, and A. Kohlrausch. A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation. *IEEE T. Audio. Speech*., 20:1–15, 2012.
39. T. May, S. Van De Par, and A. Kohlrausch. Noise-robust speaker recognition combining missing data techniques and universal background modeling. *IEEE T. Audio. Speech*., 20:108–121, 2012.
40. T. May, S. van de Par, and A. Kohlrausch. Binaural localization and detection of speakers in complex acoustic scenes. In J. Blauert, editor, *The technology of binaural listening*, chapter 15. Springer, Berlin-Heidelberg-New York NY, 2013.
41. D. McAlpine and B. Grothe. Sound localization and delay lines-do mammals fit the model? *Trends Neurosci*., 26:347–350, 2003.
42. D. McAlpine, D. Jiang, and a. R. Palmer. A neural code for low-frequency sound localization in mammals. *Nat. Neurosci*., 4:396–401, 2001.
43. J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. *J. Acoust. Soc. Am*., 119:463–479, 2006.
44. J. Nix and V. Hohmann. Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering. *IEEE T. Audio. Speech*., 15:995–1008, 2007.
45. B. Opitz, A. Mecklinger, A. D. Friederici, and D. Y. Von Cramon. The functional neuroanatomy of novelty processing: integrating ERP and fMRI results. *Cereb. Cortex*, 9:379–391, 1999.
46. B. Osnes, K. Hugdahl, and K. Specht. Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception. *Neuroimage*, 54:2437–2445, 2011.
47. P. Paavilainen, M. Jaramillo, R. Näätänen, and I. Winkler. Neuronal populations in the human brain extracting invariant relationships from acoustic variance. *Neurosci. Lett*., 265:179–182, 1999.
48. K. Palomäki and G. J. Brown. A computational model of binaural speech recognition: Role of across-frequency vs. within-frequency processing and internal noise. *Speech Commun*., 53:924–940, 2011.
49. K. J. Palomäki, G. J. Brown, and D. Wang. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Commun*., 43:361–378, 2004.
50. D. P. Phillips. A perceptual architecture for sound lateralization in man. *Hear. Res*., 238:124–132, 2008.
51. V. Pulkki and T. Hirvonen. Functional count-comparison model for binaural decoding. *Acta Acustica united with Acustica*, 95:883–900, 2009.
52. L. Rayleigh. On our perception of sound direction. *Philos. Mag*., 13:214–232, 1907.
53. H. Riedel and B. Kollmeier. Interaural delay-dependent changes in the binaural difference potential of the human auditory brain stem response. *Hear. Res*., 218:5–19, 2006.
54. N. Roman, D. Wang, and G. J. Brown. Speech segregation based on sound localization. *J. Acoust. Soc. Am*., 114:2236–2252, 2003.
55. S. Särkkä, A. Vehtari, and J. Lampinen. Rao-Blackwellized particle filter for multiple target tracking. *Information Fusion*, 8:2–15, 2007.
56. P. Søndergaard and P. Majdak. The auditory-modeling toolbox.In J. Blauert, editor, *The technology of binaural listening*, chapter 2. Springer, Berlin-Heidelberg-New York NY, 2013.
57. S. Spors and H. Wierstorf. Evaluation of perceptual properties of phase-mode beamforming in the context of data-based binaural synthesis. In 5th *International Symposium on Communications Control and Signal Processing (ISCCSP)*, 2012, pages 1–4, 2012.
58. R. Stern and N. Morgan. Hearing is believing: Biologically-inspired feature extraction for robust automatic speech recognition. *IEEE Signal Processing Magazine*, 29:34–43, 2012.

59. R. Stern, A. Zeiberg, and C. Trahiotis. Lateralization of complex binaural stimuli: A weighted-image model. *J. Acoust. Soc. Am*., 84:156–165, 1988.
60. R. M. Stern and H. S. Colburn. Theory of binaural interaction based in auditory-nerve data. IV. A model for subjective lateral position. *J. Acoust. Soc. Am*., 64:127–140, 1978.
61. S. K. Thompson, K. von Kriegstein, A. Deane-Pratt, T. Marquardt, R. Deichmann, T. D. Griffiths, and D. McAlpine. Representation of interaural time delay in the human auditory midbrain. *Nat. Neurosci*., 9:1096–1098, 2006.
62. S. P. Thompson. On binaural audition. *Philos. Mag*., 4:274–276, 1877.
63. S. P. Thompson.On the function of the two ears in the perception of space. *Philos. Mag*., 13:406–416, 1882.
64. M. van der Heijden and C. Trahiotis. Masking with interaurally delayed stimuli: the use of "internal" delays in binaural detection. *J. Acoust. Soc. Am*., 105:388–399, 1999.
65. G. von Békésy. Zur Theorie des Hörens. Über das Richtungshören bei einer Zeitdifferenz oder Lautstärkenunggleichheit der beiderseitigen Schalleinwirkungen. *Phys. Z*., 31:824–835, 1930.
66. C. Wacongne, J. P. Changeux, and S. Dehaene. A neuronal model of predictive coding accounting for the mismatch negativity. *J. Neurosci*., 32:3665–3678, 2012.
67. K. C. Wagener and T. Brand. Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and masking parameters. *Int. J. Audiol*., 44:144–156, 2005.
68. D. Wang and G. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.
69. S. Wilson, A. Saygin, M. Sereno, and M. Iacoboni. Listening to speech activates motor areas involved in speech production. *Nat. Neurosci*., 7:701–702, 2004.
70. I. Winkler. Interpreting the Mismatch Negativity. *J. Psychophysiol*., 21:147–163, 2007.
71. J. Woodruff and D. Wang. Binaural localization of multiple sources in reverberant and noisy environments. *IEEE T. Audio. Speech*., 20:1503–1512, 2012.
72. S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. The HTK book. *Cambridge University Engineering Department*, 3, 2002.