Combining Auditory Preprocessing and Bayesian Estimation for Robust Formant Tracking

Claudius Gläser, Martin Heckmann, Frank Joublin, Christian Goerick

2010

Preprint:

This is an accepted article published in IEEE Transactions on Audio, Speech, and Language Processing. The final authenticated version is available online at: https://doi.org/[DOI not available]

Combining Auditory Preprocessing and Bayesian Estimation for Robust Formant Tracking

Claudius Gläser, Martin Heckmann, Frank Joublin, and Christian Goerick

***Abstract*—We present a framework for estimating formant trajectories. Its focus is to achieve high robustness in noisy environments. Our approach combines a preprocessing based on functional principles of the human auditory system and a probabilistic tracking scheme. For enhancing the formant structure in spectrograms we use a Gammatone filterbank, a spectral preemphasis, as well as a spectral filtering using Difference-of-Gaussians (DoG) operators. Finally, a contrast enhancement mimicking a competition between filter responses is applied. The probabilistic tracking scheme adopts the mixture modeling technique for estimating the joint distribution of formants. In conjunction with an algorithm for adaptive frequency range segmentation as well as Bayesian smoothing an efficient framework for estimating formant trajectories is derived. Comprehensive evaluations of our method on the VTR–Formant database emphasize its high precision and robustness. We obtained superior performance compared to existing approaches for clean as well as echoic noisy speech. Finally, an implementation of the framework within the scope of an online system using instantaneous feature-based resynthesis demonstrates its applicability to real-world scenarios.**

***Index Terms*—Speech analysis, Bayes procedures, Tracking, Adaptive estimation, Dynamic programming, Speech synthesis**

I. Introduction

HUMAN speech perception relies to a large extend on vocal tract resonance frequencies and their variation in time [1]. These resonance frequencies manifest themselves as energy concentrations in the spectro-temporal domain and are referred to as formants. Despite their expected advantages only very few automatic speech recognition systems try to use formant trajectories. The main reason is that common methods for their extraction lack in precision, robustness, or computational efficiency. Formant extraction becomes particularly difficult for speech degraded by large speaker-microphone distances and background noise. In contrast, humans perform marvelously well under such conditions [2]. Consequently, designing a system based on functional principles of the human auditory system is expected to overcome these problems.

Traditional approaches for extracting formants are characterized by a segregation into two processing stages: formant candidate estimation and a subsequent selection and allocation of candidates to the vocal tract resonance frequencies. In its simplest form this is achieved by performing spectral analysis based on *cross-channel correlation* [3] or *linear predictive coding (LPC)* [4] followed by peak picking in the resulting spectrogram. However, such approaches are error-prone, especially for speech degraded by noise. Consequently, they are not applicable for real-world scenarios.

The authors are with the Honda Research Institute Europe, Carl-Legien-Strasse 30, 63073 Offenbach, Germany, phone: +49-69-89011-702, fax: +49-69-89011-749, e-mail: {firstname.lastname}@honda-ri.de.

Methods targeting the improvement of spectral analysis techniques for the estimation of formant candidates are rare [5], [6]. In contrast, there has been considerable effort in developing more sophisticated algorithms for selecting candidates to obtain formant trajectories. Aspects covered by these methods vary from restricting formant extraction to certain frequency ranges [7], [8], [9], through imposing continuity constraints [10], [11], to incorporating contextual information [12], [13], [14]. In recent years, probabilistic techniques for estimating formant trajectories have become popular, resulting in numerous methods relying on *Bayesian filtering* [15], [16], [17], [18], [19] or *Hidden Markov Models (HMM)* [17], [20], [21], [22], [23].

In this paper we present our framework for the extraction of formant trajectories [24]. It differs in many aspects from common algorithms. First, with the aim of achieving higher robustness against speech degradations compared to the commonly used LPC analysis we use an auditory-inspired preprocessing for enhancing the formant structure in spectrograms. Thereby, we follow the spirit of [6], but contrary we do not rely on Fast Fourier Transformation (FFT) to mimic auditory filters. Rather the application of a Gammatone filterbank transforms the speech signal into the spectro-temporal domain in which we compensate for the spectral tilt. Next, spectral filtering using Difference-of-Gaussians (DoG) operators and a contrast sharpening mimicking a competition between filter responses enhance formants in spectrograms.

Second, we perform formant tracking by using Bayesian estimation. In contrast to previous approaches, we estimate the joint distribution of formants by adopting the mixture tracking technique [25] to the problem of tracking formants and apply it in conjunction with an algorithm for adaptive frequency range segmentation. Therewith, we solve the problem of tracking multiple formants neither by using single tracker instances for each formant separately [18] nor by extending the state space [15], two commonly used techniques for multi-target tracking. We rather model the joint distribution of formants via a mixture of component distributions sharing the same state space and let them evolve in a data-driven adaptive manner. Within the mixture tracking framework a component-specific probabilistic modeling of formant dynamics can be used by which differences in dynamic behavior or even correlations between formants [26] can be taken into account. Additionally, we do not assume Gaussian probability distributions by relying on Kalman filters [16], [17], [19]; instead we use a grid-based approximation of posteriors, so that multiple hypotheses can be evaluated in parallel. This is particularly important when operating in noisy environments (see [15] for another application of grid-based belief approximation to formant

tracking). We will show, that the application of Bayesian smoothing [27] on the obtained filtering distributions further enhances noise robustness.

Lastly, our framework incorporates algorithms for pitch, voicing, and gender extraction [28], [29]. This is advantageous as formant profiles of male, female, and children's voices differ significantly [30]. We use this gender decision to modulate the probabilistic tracking regime insofar as we switch gender-dependent probabilistic formant models according to the detected gender. We performed an evaluation on the VTR–Formant database [31] and further applied our framework within the scope of an online system using instantaneous feature-based resynthesis.

The remainder of the contribution is organized as follows. An overview of our formant estimation framework is given in section II. Section III focuses on the auditory-based preprocessing followed by a detailed description of the probabilistic tracking regime in section IV. The extraction of pitch, voicing, and gender as well as their application to tracking formants is highlighted in section V. A comprehensive evaluation of our method is presented in section VI. The application of the framework within the scope of an online system is highlighted in section VII. Finally, section VIII summarizes the paper.

II. System Overview

The architecture of our system for formant trajectory estimation is presented in Fig. 1. It can be divided into three main processing blocks: an auditory-based preprocessing for enhancing formant structure in the spectro-temporal domain (top left), a probabilistic tracking framework for estimating formant trajectories (bottom), as well as a gender extraction (top right), whose decision modulates the tracking of formants.

The following sections will focus on the individual parts of the architecture and provide insights into the detailed processing carried out.

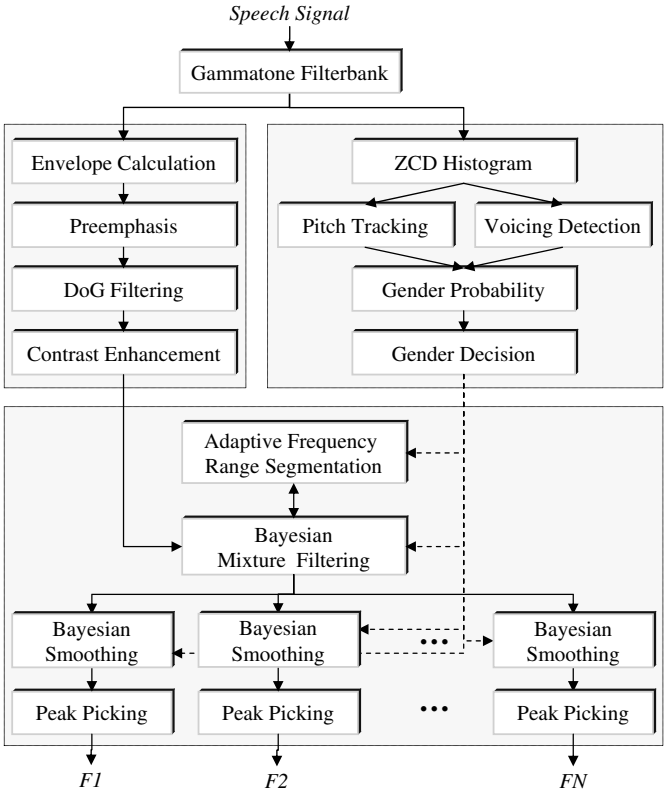

Fig. 1. The architecture of the formant estimation system.

III. Preprocessing

According to the linear source-filter theory speech is produced by a non-linear volume velocity source followed by a time-varying linear filter and radiation components [32]. Because formants are the resonance frequencies of the vocal tract, their extraction can be improved by eliminating the spectral influence of excitation and radiation. More precisely, the formant structure in a spectrogram is mainly impaired by two causes: First, excitation and radiation introduce a spectral tilt that has to be corrected via a preemphasis. Second, for voiced sounds the glottis convertes the steady airflow produced by the lungs into a quasi-periodic train of flow pulses by which the transfer function of the vocal tract is sampled at multiples of the fundamental frequency. Consequently, spectrograms feature spectral peaks at the harmonics rather than the vocal tract resonance frequencies. The preprocessing we present in the following aims at compensating for these effects.

A. Gammatone Filterbank

We transform the speech signal into the spectro-temporal domain via the Patterson-Holdsworth auditory filterbank [33]. This filterbank is based on neurophysiological findings on the human auditory system and models the peripheral processing as carried out by the cochlea, where sound is transformed into spatio-temporal response patterns on the auditory nerve. The filterbank is implemented as a set of Gammatone filters, each of them being tuned to a different frequency range. Thereby, we follow the implementation suggested by Slaney [34]. Our filterbank is composed of 128 Gammatone filters covering the frequency range from 80 Hz to 8 kHz. Subsequently, the spectral envelope is calculated via rectification and low-pass filtering. The logarithmic envelope of the filter responses to an exemplary speech signal chosen from the TIMIT database [35] is shown in Fig. 2 (a).

B. Preemphasis

For voiced sounds, the voice source signal is produced at the glottis. By constituting relations between glottal spectra and the frequency response of linear filters, Fant suggested the use of a second-order low-pass filter for the approximation of the glottal flow spectrum [36]. This is a valid approximation when assuming the most common phonation types, modal and creaky phonation [37]. Thus, voiced excitation changes the spectral characteristic by -12 dB/oct.

Furthermore, it can be derived that a first-order high-pass filter models the lip impedance, which corresponds to the transfer function of the radiation component [38]. Thus, radiation changes the spectral characteristics by +6 dB/oct. Overall this means that a preemphasis via amplification of

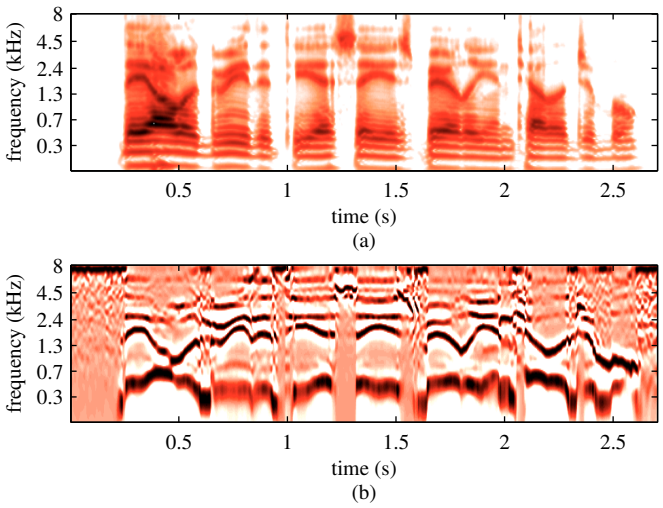

Fig. 2. In (a) the original spectrogram for the utterance "They all agree that the essay is barely intelligible." spoken by a male speaker is shown. The spectrogram after the application of the preprocessing for enhancing the formant structure is shown in (b).

frequency magnitudes by +6 dB/oct adequately eliminates the spectral influence of excitation and radiation.

C. Spectral Filtering

After the above mentioned preemphasis we enhance the formant structure in the spectrogram by smoothing along the frequency axis following the same spirit as [6]. Therefore, we use channel-dependent Difference-of-Gaussians (DoG) operators with standard deviations of the negative Gaussian components being twice as large as that of the corresponding positive ones:

$$DoG_n(f) = \frac{1}{\sqrt{(2\pi)}}\left(\exp\left(-\frac{(f-fc_n)^2}{2\sigma_n^2}\right) - \frac{1}{2}\exp\left(-\frac{(f-fc_n)^2}{8\sigma_n^2}\right)\right) \quad (1)$$

Here, DoG_n is the DoG operator of channel n featuring a center frequency fc_n. We set the standard deviations σ_n to 70 Hz. For filter channels with center frequencies in the range from 5 to 8 kHz we further increased the standard deviations linearly up to 400 Hz in order to suppress formants higher than F4. Additionally, the logarithmic arrangement of the Gammatone filterbank's channel center frequencies is taken into account, insofar as the DoGs are discretized by sampling and normalizing them accordingly. Fig. 3 exemplarily depicts the DoGs of 8 filter channels. The application of the DoG operators on the emphasized spectrogram spreads harmonics and forms peaks at formant locations, whereas regions between formants are suppressed.

D. Contrast Enhancement

In our experiments an additional contrast enhancement mimicking a competition between filter responses has been proven to be useful. At first a normalization of the values to the maximum at each sample is performed. By doing so,

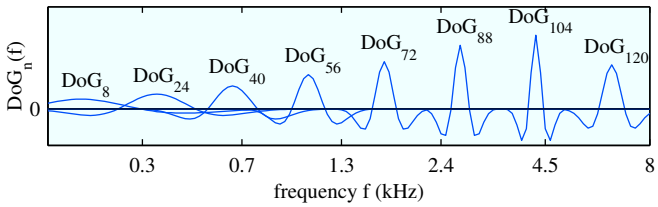

Fig. 3. The DoG operators of the filter channels used for enhancing formant structure in spectrograms vary in bandwidth and frequency resolution dependent on the channels' center frequencies. Here, the DoG operators of 8 exemplarily chosen filter channels are shown.

formant structures also become visible in signal parts where the energy is relatively low. An application of a sigmoidal function r to the normalized spectrogram S further enhances the spectral contrast.

$$r(S) = \frac{1}{1+\exp(-\alpha(S-\theta))} \quad (2)$$

We use $\alpha = 3.5$ and $\theta = -0.2$ for the slope and threshold parameters, respectively. Fig. 2 (b) continues the example depicted in (a) by showing the spectrogram after the application of the preprocessing. As can be seen, the proposed method significantly enhances the formant structure (see section VI-C for a thorough performance evaluation).

IV. Formant Tracking

The preprocessing algorithm described in the previous section enhances the formant structure in spectrograms. Nevertheless, it neither detects exact formant locations nor their trajectories. To obtain complete formant trajectories, a tracking algorithm is applied. Formant tracking has been investigated already for a long time. Yet it is still a rather challenging task as multiple formants have to be tracked at the same time.

While tracking multiple formants two general problems arise. The first one is widely known as the data association problem, i.e. the assignment of spectral peaks to formants. This cannot be achieved by focusing on only one formant; rather one has to look at the joint distribution of formants in conjunction with spectro-temporal constraints. The second problem is the sequential estimation of formant locations based on noisy observations. For this reason the robustness of the used tracking algorithm against noise and clutter is of particular interest. In the following a tracking framework meeting these requirements is proposed.

A. Bayesian Mixture Filtering

Because no sensor is perfect, handling uncertainty introduced by noise and clutter is one of the key issues in any dynamical system. Over the past years particularly Bayesian filter techniques have become popular, since they provide a powerful framework of probabilistically estimating a dynamic system's state. Their benefit in handling noisy observations was shown in different applications [39].

Bayesian filters represent the state at time t by random variables x_t, whereas uncertainty is introduced by a probabilistic distribution over x_t, called the *belief* $Bel(x_t)$. Bayesian filters

target the sequential estimation of such beliefs over the state space conditioned on all information contained in the sensor data [40]. Let z_t denote the observation at time t, then the standard Bayesian filter recursion can be written as follows:

$$Bel(x_0) = p(x_0) \tag{3}$$

$$Bel(x_t) = \alpha \cdot p(z_t|x_t) \int p(x_t|x_{t-1}) Bel(x_{t-1})\, dx_{t-1} \tag{4}$$

Here $p(x_0)$ is some *a priori distribution* used for initialization and α a normalization constant ensuring the belief's probabilistic character. Furthermore, the so called *motion model* $p(x_t|x_{t-1})$ and the *observation model* $p(z_t|x_t)$ are used for the description of the system dynamics as well as the state-dependent likelihood of perceiving observations, respectively. According to this, the standard Bayesian filter recursion can be interpreted as a two-stage process. Every time a sensor provides a new observation z_t the system calculates the *predictive belief* $Bel^-(x_t)$ using (5) and subsequently corrects the prediction according to (6):

$$Bel^-(x_t) = \int p(x_t|x_{t-1}) \cdot Bel(x_{t-1})\, dx_{t-1} \tag{5}$$

$$Bel(x_t) = \frac{p(z_t|x_t) \cdot Bel^-(x_t)}{\int p(z_t|x_t) \cdot Bel^-(x_t)\, dx_t} \tag{6}$$

Several implementations of Bayesian filters were proposed, which mainly differ in the used representation of beliefs [40]. The most famous one is the Kalman filter, which recently has become popular in the domain of formant tracking as well [16], [17], [19]. Kalman filters approximate beliefs by their first and second moment, which is identical to a unimodal Gaussian representation. In this way they are optimal estimators, assuming the initial uncertainty is Gaussian and the observation model and system dynamics are linear functions of the state. However, these assumptions are too restrictive in most cases, especially for tracking formants.

We want to focus on belief representations which are able to represent arbitrary distributions, thus allowing a multi-hypotheses tracking. Such a multi-hypotheses tracking is particularly important for achieving noise robustness. Particle filters provide the possibility to represent arbitrary beliefs. Their inherent property of focusing on the most important regions of the state space let particle filtering become a powerful technique when operating in high-dimensional (continuous) state spaces. However, since we want to estimate formant locations on a low-dimensional discrete grid defined by the channels of the Gammatone filterbank, we choose a grid-based approximation of the belief. Thus, assuming that the filterbank is composed of N channels, the state space at time t can be written as $x_t = \{x_{1,t}, x_{2,t}, \ldots, x_{N,t}\}$.

Nevertheless, in practical implementations Bayesian filters can maintain multimodality only over a defined time-window. Longer durations cause the belief to migrate to one of the modes, subsequently discarding all other modes. As a consequence the standard Bayesian filters are not suited for multi-target tracking as in the case of formant tracking. For this reason, we adopt the mixture filtering technique which was recently introduced in the computer vision community [25]. More precisely, we model the target distribution $Bel(x_{k,t})$ by a non-parametric mixture of M filtering distributions $Bel_m(x_{k,t})$, such that each formant is represented by one mixture component (see Fig. 4 (a)). In addition a mixture weight $\pi_{m,t}$ is assigned to each component belief to maintain the correct target distribution:

$$Bel(x_{k,t}) = \sum_{m=1}^{M} \pi_{m,t} \cdot Bel_m(x_{k,t}) \tag{7}$$

Hence, the Bayesian filter recursion can be rewritten with respect to the mixture modeling technique by substituting the beliefs in (5) and (6) with (7):

$$Bel_m^-(x_{k,t}) = \sum_{l=1}^{N} p_m(x_{k,t}|x_{l,t-1}) Bel_m(x_{l,t-1}) \tag{8}$$

$$Bel_m(x_{k,t}) = \frac{p(z_t|x_{k,t}) Bel_m^-(x_{k,t})}{\sum_{l=1}^{N} p(z_t|x_{l,t}) Bel_m^-(x_{l,t})} \tag{9}$$

$$\pi_{m,t} = \frac{\pi_{m,t-1} \sum_{k=1}^{N} p(z_t|x_{k,t}) Bel_m^-(x_{k,t})}{\sum_{n=1}^{M} \pi_{n,t-1} \sum_{l=1}^{N} p(z_t|x_{l,t}) Bel_n^-(x_{l,t})} \tag{10}$$

The formulas show that the incorporation of the mixture modeling technique results in the standard Bayesian recursion at the level of individual mixture components. This means that the component filtering distributions $Bel_m(x_{k,t})$ evolve independently over time, whereas an interaction between the components only takes place during the calculation of the new mixture weights $\pi_{m,t}$. This nice result unfortunately also entails that the mixture filtering distributions are not immune against belief degeneration. More precisely, component beliefs will become more and more diffuse over time which may result in loosing track of formants. This is exactly the opposit of what we want to achieve. In fact, components assigned to formants should be clearly separated in order to avoid ambiguities and maintain multimodality.

To prevent mixture components from belief degeneration, a procedure which reclusters the component beliefs has to be applied from time to time. Assuming such a function exists and returns sets $R_{1,t}, R_{2,t}, \ldots, R_{M,t}$ for M components which segment the frequency range into consecutive formant-specific regions. This means that each frequency channel $x_{k,t}$ at each instance in time is element of exactly one set $R_{m,t}$ and therewith assigned to exactly one non-empty mixture component m covering a certain formant. What remains is the recalculation of the component beliefs and the associated mixture weights, such that the joint beliefs before and after the reclustering procedure are equal. Because the segmentation algorithm results in disjunct state sets and the joint filtering distribution before and after the reclustering has to be equal, the new component beliefs $Bel'_m(x_{k,t})$ and associated mixture weights $\pi'_{m,t}$ have to be of the form:

$$\pi'_{m,t} = \sum_{k \in R_{m,t}} \sum_{n=1}^{M} \pi_{n,t} \cdot Bel_n(x_{k,t}) \tag{11}$$

$$Bel'_m(x_{k,t}) = \begin{cases} \frac{\sum_{n=1}^{M} \pi_{n,t} \cdot Bel_n(x_{k,t})}{\pi'_{m,t}} & , \forall x_{k,t} \in R_{m,t} \\ 0 & , \forall x_{k,t} \notin R_{m,t} \end{cases} \tag{12}$$

Fig. 4 illustrates the proposed method of recalculating component beliefs and mixture weights. As shown in (a) the

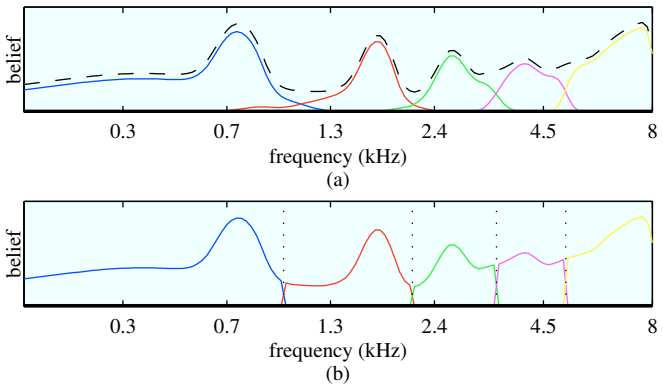

Fig. 4. The mixture modeling approach: (a) the joint belief (dashed line) is modeled by overlapping mixture components (solid lines); (b) the calculated cluster boundaries (dotted lines) are used for the recalculation of beliefs resulting in non-overlapping mixture components.

component beliefs overlap significantly close to the boundaries. After the reclustering algorithm has been applied and optimal component boundaries, which divide the frequency range into disjunct sets of frequency channels, are known, the recalculation can be done. Fig. 4 (b) shows that previously overlapping probabilities are separated. Thereby, consecutive components exchange parts of their probabilities in a mixture weight dependent manner. With this method, a mixture of consecutive but separated components is achieved by which multimodality can be maintained.

B. Adaptive Frequency Range Segmentation

To find the optimal segment boundaries we introduce a new variable $x_{k,t}^{(m)}$ that specifies the assignment of state x_k to segment m at time t. With the help of this variable the trellis shown in Fig. 5 can be build, where each $x_{k,t}^{(m)}$ is represented by one node. Now the problem of frequency range segmentation can be reformulated to finding an optimum path from $x_{1,t}^{(1)}$ to $x_{N,t}^{(M)}$ through the trellis. This can be seen by postulating that each $x_{k,t}^{(m)}$ becomes true only if its corresponding node is part of the path.

If all possible paths from $x_{1,t}^{(1)}$ to $x_{N,t}^{(M)}$ are considered, it can be seen that each path has length N and exactly one node corresponding to state $x_{k,t}$ is part of each path. This means, that all paths encode one possible segmentation by assigning exactly one component label m to each state $x_{k,t}$. Furthermore, by choosing the topology shown in Fig. 5 all possible segmentations are covered by paths through the trellis. Additionally, the sequential order of components is maintained by using only vertical and diagonal transitions between nodes. Finally, this algorithm produces non-empty components since at least one node encoding the assignment of a state to each component is part of a path.

What remains is an appropriate choice of necessary parameters as well as an algorithm for calculating the most likely path through the trellis. The latter can be done by using Viterbi decoding. For the former three parameters can be identified: the likelihood of each state $p(x_{k,t}^{(m)})$ as well as the transition probabilities $p(x_{k,t}^{(m)} \mid x_{k-1,t}^{(m-1)})$ and $p(x_{k,t}^{(m)} \mid x_{k-1,t}^{(m)})$.

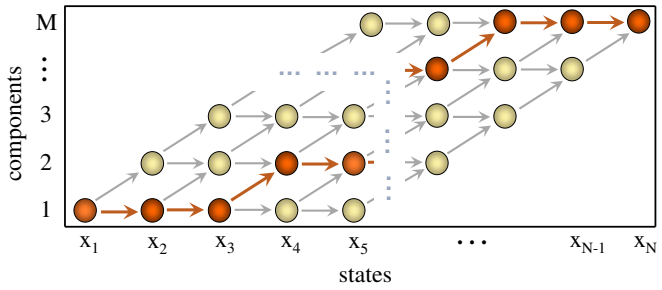

Fig. 5. The trellis used for frequency range segmentation. The colored path through the trellis encodes one possible segmentation.

Unfortunately, these likelihoods are not available, rather they have to be set according to some heuristics. Following the principle of maximum entropy we use uniformly distributed transition probabilities, whereas the following formula is used to model the priors:

$$p(x_{k,t}^{(m)}) = p_m(x_{k,0}) \cdot Bel_m(x_{k,t}) \tag{13}$$

According to (13) the likelihood of state $x_{k,t}^{(m)}$ depends on the a priori probability distribution of component m as well as the actual m-th component belief. Since the belief represents the segmentation at the last timestep updated according to the motion and observation models, this formula applies some data-driven segment continuity constraint. Furthermore, the used a priori probability distribution $p_m(x_{k,0})$ antagonizes segment degeneration by application of long-term constraints.

Because the suggested method relies on the component beliefs at the actual timesteps, the frequency range is sequentially segmented in an adaptive and computationally efficient manner. Therewith we are able to apply the Bayesian mixture filtering for tracking the joint distribution of formants while maintaining its multimodality. Fig. 6 continues the example of Fig. 2. In (a) the filtering distributions obtained by the application of Bayesian mixture filtering using 5 components are shown. Addtionally, the calculated segment boundaries are overlayed (dashed lines).

This example demonstrates the ability of the proposed approach of effectively maintaining multimodality through mixture modeling. Thereby, the frequency range segmentation algorithm plays an important role in dividing the frequency range into formant specific parts, thus, resolving the data association problem. The computed segment boundaries adapt to formant profiles in a data-driven manner. Even during closely contiguous formant frequencies this approach performs well. The reason for this is the application of continuity as well as long-term constraints.

Nevertheless, this example also reveals limits of Bayesian mixture filtering. Uncertainties already included in observations cannot be resolved completely. They rather result in a diffuse mixture filtering distribution at these locations. Hence some further processing is necessary in order to achieve robust formant trajectories even in these cases.

C. Bayesian Smoothing

If we keep in mind that Bayesian mixture filtering assumes the underlying process to be Markovian, its limits can readily

be understood. In a Markovian process the actual state only depends on the previous state as well as the actual observation. Thus the belief of a state $x_{k,t}$ depends on all observations up to time t as shown by (14). In order to robustly extract formant trajectories, particularly in noisy environments, also future observations have to be taken into account. Insofar (15), where $\widehat{Bel}(x_{k,t})$ denotes the belief regarding both past and future observations, would be a preferable measurement for tracking formants.

$$Bel(x_{k,t}) = p(x_{k,t} \,|\, z_1, z_2, \ldots, z_t) \tag{14}$$

$$\widehat{Bel}(x_{k,t}) = p(x_{k,t} \,|\, z_1, z_2, \ldots, z_t, \ldots, z_{T-1}, z_T) \tag{15}$$

Bayesian smoothing provides a way of estimating such a distribution over $x_{k,t}$ [27]. It works very similar to standard Bayesian filters, but in reverse time direction. It recursively estimates the smoothed distribution $\widehat{Bel}(x_{k,t})$ based on predefined system dynamics $p_m(x_{t+1}|x_t)$ as well as the already obtained filtering distributions $Bel_m(x_t)$:

$$\widehat{Bel}^{-}_m(x_{k,t}) = \sum_{l=1}^{N} \widehat{Bel}_m(x_{l,t+1}) \cdot p_m(x_{l,t+1}|x_{k,t}) \tag{16}$$

$$\widehat{Bel}_m(x_{k,t}) = \frac{Bel_m(x_{k,t}) \cdot \widehat{Bel}^{-}_m(x_{k,t})}{\sum_{l=1}^{N} Bel_m(x_{l,t}) \cdot \widehat{Bel}^{-}_m(x_{l,t})} \tag{17}$$

Our implementation of Bayesian smoothing incorporates the sliding window technique in order to make the algorithm suitable for online operation. For the experimental results reported in section VI the window size is set to 150 ms. However, in order to reduce the introduced signal delay the window size can be decreased considerably while obtaining similar performance (e.g. in our application presented in section VII a window size of 80 ms is used).

Fig. 6 (b) shows the result of Bayesian smoothing applied to the filtering distribution of (a) and demonstrates that the formant trajectories are significantly enhanced. Due to the continuity constraints regarding both past and future observations almost all ambiguities were resolved. Thus former diffuse filtering distributions were sharpened, likewise at locations where observations were characterized by uncertainty.

The final calculation of exact formant locations $F_m(t)$ can be done by picking the peaks of the smoothed component beliefs such that the location of the m-th formant equals the peak location in the smoothed distribution of component m (see (18)). In Fig. 6 (c) the resulting formant tracks are overlayed to the original spectrogram.

$$F_m(t) = \arg\max_{x_{k,t}} \left[\widehat{Bel}_m(x_{k,t})\right] \tag{18}$$

The formant tracking algorithm introduced in this section has several advantages over conventional approaches. Besides using Bayesian filtering to cope with noisy environments [40], our algorithm models the joint distribution of formants via a mixture of formant-specific component beliefs. These beliefs are represented using a grid-based approximation with the grid being defined by the filterbank's channels. Since mixture components evolve independently over time, formant-specific a priori distributions $p_m(x_{k,0})$ and motion models $p_m(x_{k,t}|x_{l,t-1})$ can be chosen for each component separately.

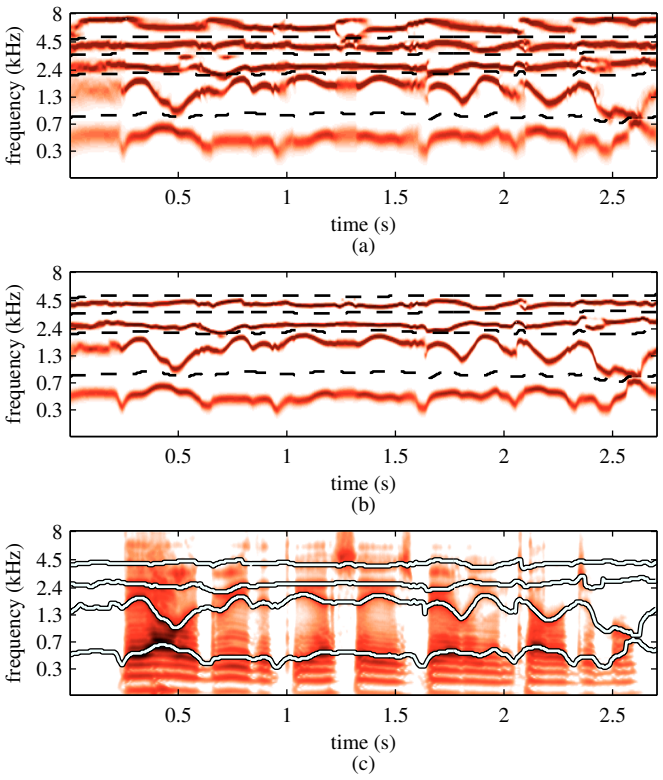

Fig. 6. The formant tracking algorithm applied to the example depicted in Fig. 2. The filtering distributions (a), the smoothing distributions (b), as well as the resulting formant trajectories overlaid to the original spectrogram (c) are shown.

V. Gender Extraction

Beyond the utilization of formant-specific probability distributions, the extraction of formants can further be improved by taking additional information into account. Here, we incorporate knowledge about the gender of speakers via gender-dependent formant-specific probability distributions $p^G_m(x_{k,0})$ and $p^G_m(x_{k,t}|x_{l,t-1})$. This is reasonable since it is well-known that formant profiles of children, women, and men differ significantly [30]. More precisely, there is a close relation between formant locations and vocal tract size, which in turn differs from male to female subjects. As a result, female formant patterns are on average scaled to about 20 % higher frequencies than corresponding male formant patterns [41]. These results can be extended to vocal tract resonance frequencies of children by further introducing an age-dependent scaling factor [42]. It is additionally known that a subject's mean pitch is correlated with its vocal fold length [43]. This renders pitch an excellent feature for detecting gender.

For this reason, we extract pitch and voicing information and subsequently judge a speaker's gender. Based on this information different gender-dependent formant-specific probability distributions are used in the formant tracking framework.

A. Pitch & Voicing extraction

In the following we will sketch our algorithms for pitch and voicing extraction in order to give the reader a comprehensive

overview of our framework. Nevertheless, a detailed description of the used algorithms is beyond the scope of this paper. For more details refer to [28] and [29].

In our pitch extraction algorithm we combine information residing in the temporal and spectral representation [28]. On the one hand the algorithm captures the temporal aspects by relying on a histogram of *Zero Crossing Distances (ZCDs)*. Such a histogram is very similar to the so called *all order interspike histogram* modeling the phase locked firing of neurons in the auditory system [44]. On the other hand spectral aspects were incorporated by using comb filters. Most importantly, combining both cues allows a suppression of spurious side peaks within the histogram of ZCDs by which a significant increase in noise robustness can be achieved.

In order to obtain continuous pitch trajectories we apply a tracking algorithm on the final ZCD histogram. Here, we use the same algorithm as for the estimation of the formant trajectories (see section IV), with the exception that only one mixture component is used, since no multi-target tracking has to be done. In this case, the algorithm resembles the standard Bayesian filtering and smoothing. After that, the maximum at each sample in time is picked and converted from a distance measure to a frequency.

Next, we carry out a voiced-unvoiced classification of speech samples based on two cues [29]. One is the ratio of the energy in a high frequency band to that in a low frequency band. The other cue is the harmonicity of the signal which is estimated by calculating the variance of the ZCD histogram. Finally, a multidimensional hypothesis test integrating both features yields a voicing decision.

B. Gender Detection

In order to judge a speaker's gender, we introduce variables $g \in \{\text{'male'}, \text{'female'}\}$, $v \in \{\text{'voiced'}, \text{'unvoiced'}\}$, and F_0 denoting gender, voicing, and pitch, respectively. Next, we set $p(g|F_0, v = \text{'unvoiced'}) = 0$ and estimate the posterior $p(g|F_0, v = \text{'voiced'})$ using the training set of the VTR–Formant database [31]. These posteriors are finally used to calculate the gender $G(t)$ of a speaker at each timestep t:

$$G(t) = \begin{cases} \text{'male'} & \text{, if } h(\text{'male'}, t) \geq h(\text{'female'}, t) \\ \text{'female'} & \text{, otherwise} \end{cases} \quad (19)$$

$$h(g, t) = (1 - \kappa) \cdot h(g, t) + \kappa \cdot p(g|F_0(t), v(t)) \quad (20)$$

As stated in (20), we perform a smoothing on the posteriors with time constant κ. This yields two desirable properties: fluctuations in gender decision are suppressed and gender decisions are extended from voiced to unvoiced speech segments.

VI. Results

To evaluate the proposed method, we performed tests on the VTR–Formant database [31]. As a subset of the widely-used TIMIT corpus, this database comprises a total of 516 utterances spoken by male and female speakers. It additionally provides formant trajectories which have been initially derived by an automatic formant tracker [45] and subsequently hand-corrected for the first three formants. However, the difficulty of the VTR–Formant database in providing a ground truth is worth noting. First, methods related to [45] may benefit from the employed semi-automatic labeling and, second, in some cases even visual inspection may not provide means to identify real formant locations. We nevertheless think that this database provides a reasonable basis for deriving quantitative performance measures. For this reason we used it in order to compare our algorithm to existing approaches with respect to precision and robustness in noisy echoic environments.

A. Experimental Setup

For all experiments our algorithm contained four mixture components corresponding to the first four formants (F1–F4) as well as an additional component covering the frequency range above F4. The gender-dependent priors $p_m^G(x_{k,0})$ and motion models $p_m^G(x_{k,t}|x_{l,t-1})$ for each mixture component m, which describe the spectro-temporal behavior of the corresponding formants, were estimated based on the training set of the VTR–Formant database. The same holds for the posterior $p(g|F_0, v)$ used for detecting gender g based on extracted pitch F_0 and voicing v.

The evaluation was carried out on the test set of the VTR–Formant database, which consist of 34 and 56 utterances spoken by male and female speakers, respectively. We further added white noise, babble noise, and car noise at 7 different signal-to-noise ratios (SNRs) to the clean speech signal. For estimating SNRs, non-speech samples were excluded from energy calculation.

We applied our algorithm to clean and noisy utterances and calculated the absolute errors normalized by the formant locations given in the manual labels at time steps equally spaced by 10 ms. The obtained mean relative errors are depicted by the plots in Fig. 7 where 95 % confidence intervals are additionally shown. For estimating confidence intervals we applied the bootstrap technique using sentence-wise data sampling [46], [47]. The results mirror the typical performance curves where the error continuously increases when SNR decreases. Overall our algorithm yields suitable estimates under all conditions without showing any significant drops in performance. The plots additionally illustrate that car noise, with its energy concentrated on low frequencies, has only a small influence on the estimation of high formants.

B. Comparison to Existing Approaches

To judge the quality of the obtained results we compared them to existing approaches. More precisely, we applied the formant estimation algorithms provided by two widely-used speech processing tools (*Praat* [48] and the *Snack Sound Toolkit* used in *WaveSurfer* [49]) as well as that of a recently proposed approach [50] also targeting noise robust tracking.

Both *Praat's* and *Snack's* formant extraction rely on LPC analysis. After an initial preemphasis they estimate formant candidates by solving for the roots of the linear predictor polynomial. Whereas *Praat* considers formant candidates to already be the final formant estimates, *Snack* additionally applies a dynamic programming based cost minimization of connecting formant candidates to obtain complete trajectories.

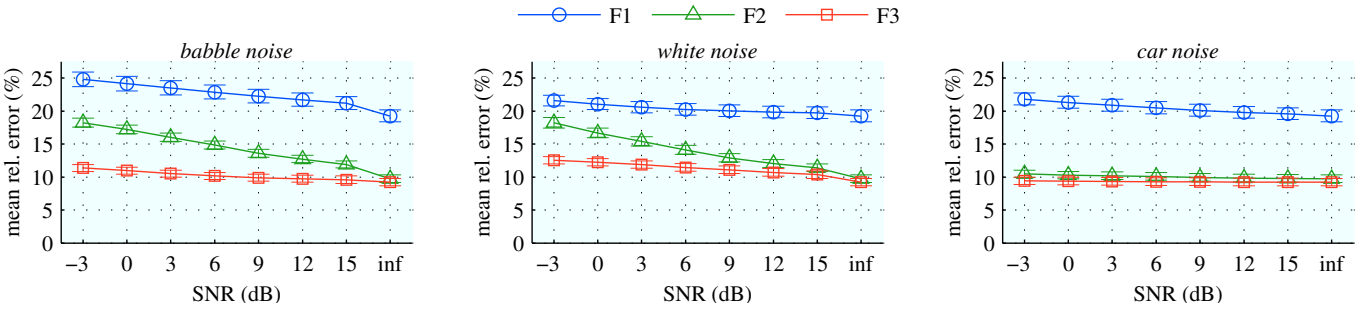

Fig. 7. The plots depict the mean relative errors of our method when speech was degraded by various types of noise at different SNRs. Bars mark 95 % confidence intervals.

TABLE I
MEAN RELATIVE IMPROVEMENTS (AND 95 % CONFIDENCE INTERVALS) IN % OF OUR METHOD COMPARED TO [48], [49], [50]

Formant	Praat [48]			Snack [49]			Mustafa [50]		
	babble	*white*	*car*	*babble*	*white*	*car*	*babble*	*white*	*car*
F1	50.5 (+2.1;-2.3)	79.4 (+0.9;-1.0)	74.0 (+1.4;-1.5)	8.3 (+1.8;-1.8)	52.5 (+2.1;-2.2)	45.3 (+2.7;-2.8)	50.1 (+2.1;-2.2)	39.0 (+2.6;-2.6)	64.2 (+2.5;-2.7)
F2	29.4 (+3.8;-3.9)	63.0 (+2.3;-2.4)	64.0 (+2.8;-2.9)	11.0 (+2.8;-2.9)	39.1 (+2.9;-3.0)	33.4 (+4.6;-4.8)	-0.1 (+2.3;-2.4)	19.8 (+2.8;-2.9)	28.4 (+4.0;-4.2)
F3	34.3 (+3.8;-4.1)	56.4 (+2.4;-2.6)	55.1 (+2.9;-3.0)	31.4 (+3.4;-3.5)	40.8 (+2.5;-2.6)	30.8 (+4.5;-4.7)	30.2 (+4.6;-4.9)	34.2 (+3.3;-3.6)	35.4 (+4.9;-5.3)

Therefore, it is hypothesized that *Snack* yields superior performance to *Praat*. The algorithm presented in [50] follows another approach. After a preemphasis and Hilbert transformation the signal is filtered by 4 *Formant Filters*. These are adaptive bandpass filters whose zeros and poles are updated based on the formant frequency estimates at the previous timestep by which a separation of formants into different channels can be achieved. A first-order LPC analysis on each of the 4 filter channels finally estimates F1-F4.

The relative performance improvements achieved by our framework with respect to these methods are summarized in Table I where the relative improvements are averaged over all SNRs for each type of noise, respectively. Table I additionally shows 95 % confidence intervals. As can be seen, our approach significantly outperforms the other methods in all cases tested, except for speech degraded by babble noise where the algorithm presented in [50] reaches similar performance for F2. However, in all other cases we achieve relative performance enhancements ranging from 20 % to 60 %. In some cases, we even obtain improvements of 80 %.

Finally, we evaluated the influence of echoic environments on the precision of the different formant tracking algorithms. For doing so, we measured impulse responses of a loudspeaker-enclosure-microphone (LEM) system using loudspeaker-microphone distances of 1 and 3 meters in a room with an echo constant of $\tau_{60} = 1100$ ms. We convolved clean speech signals with the obtained impulse responses and additionally added babble noise, white noise, and car noise at an SNR of 6 dB. The mean relative errors as obtained by averaging over all noise types are plotted in Fig. 8. As shown, the incorporation of echoes impairs the performance of the algorithms, particularly for the extraction of F2. Moreover, for our algorithm there is just a minor effect of echoic environments with respect to the extraction of F1 and F3. Overall our algorithm reaches superior performance compared to the other approaches in all cases tested.

C. Relative Contributions of the System Components

Given the compelling results, in the following we separate out the individual contributions of the system components with respect to their effects on the performance of the overall framework. Therefore, we performed an additional test in which we applied the proposed formant tracking algorithm to a spectrogram as obtained via LPC analysis. More precisely, we used the same preprocessing as *Snack* does, that is a signal resampling to 10 kHz followed by a preemphasis (factor 0.7) and a $\cos^4$-windowed (length = 49 ms, increment = 10 ms) 12th-order LPC analysis. Finally, a spectrogram is constructed from the LPC coefficients, which serves as input to the formant tracking algorithm.

Fig. 9 shows plots of the mean relative errors including 95% confidence intervals for three systems applied to noisy speech:

- ABP+BMT: the proposed auditory-based preprocessing (ABP) followed by the proposed Bayesian mixture tracking (BMT)
- LPC+BMT: the LPC analysis (LPC) followed by the proposed Bayesian mixture tracking (BMT)
- LPC+DPT: the LPC analysis (LPC) followed by dynamic programming-based tracking (DPT) which is the framework of *Snack* [49]

By comparing the performance curves of the different methods the relative contributions of the auditory-based preprocessing and the Bayesian mixture tracking to the overall performance of our framework can be assessed.

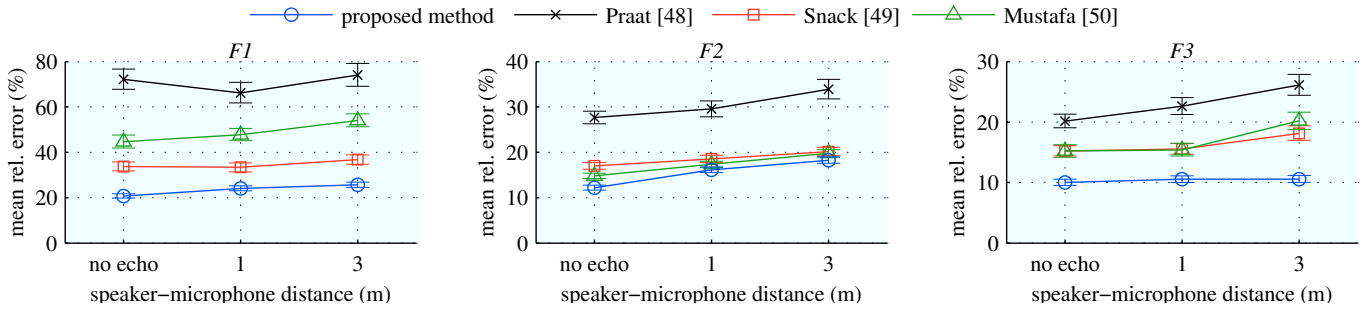

Fig. 8. Evaluation in a room with echo constant $\tau_{60} = 1100\,\text{ms}$ using speaker-microphone distances of 1 and 3 meters. Plots of the mean relative errors as obtained by averaging over various types of additionally added noise (6 dB SNR) are shown. Bars mark 95 % confidence intervals.

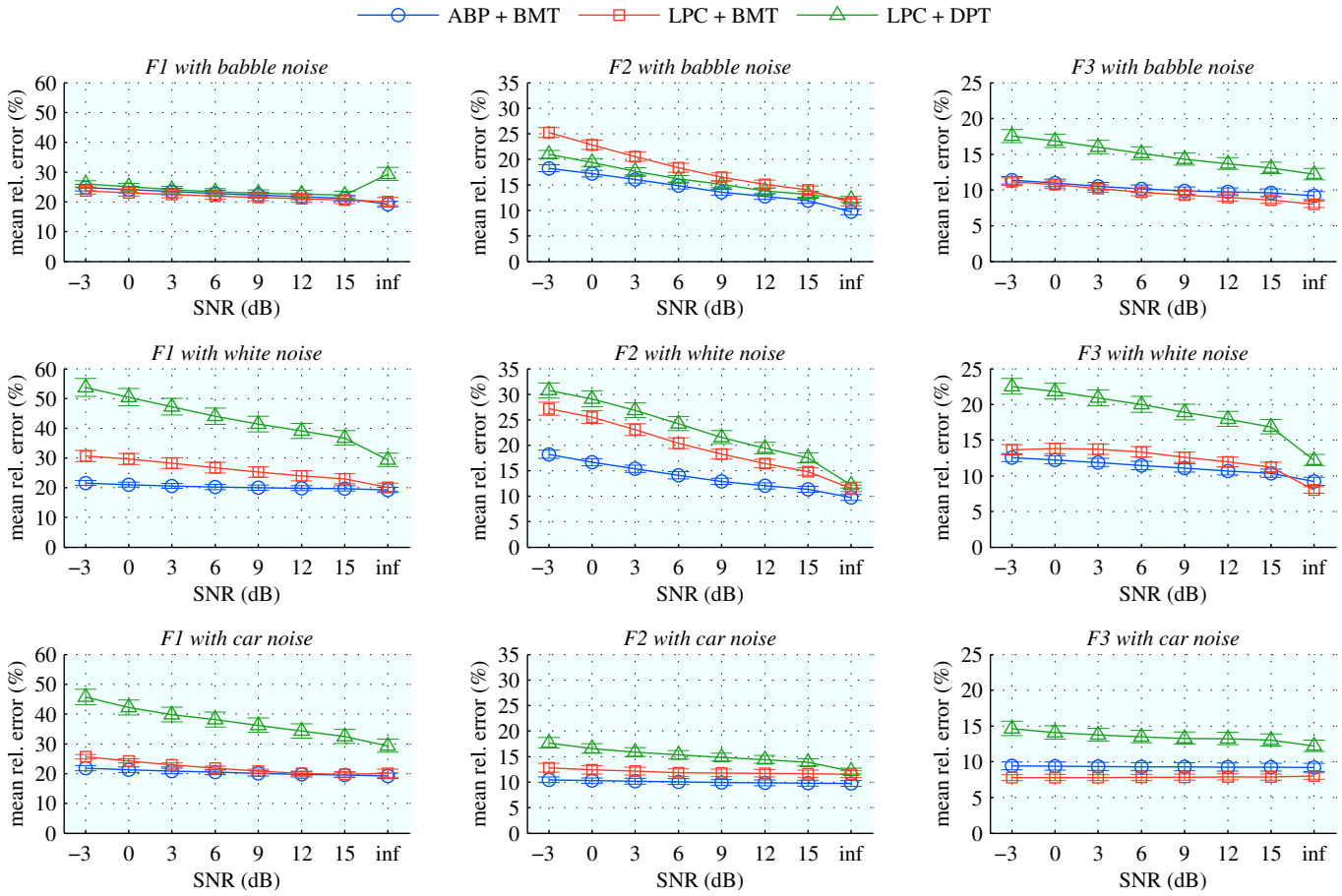

Fig. 9. Different combinations of two methods for enhancing formants in spectrograms (the proposed auditory-based preprocessing (ABP) and LPC analysis (LPC)) as well as two methods for formant tracking (the proposed Bayesian mixture tracking (BMT) and dynamic programming-based tracking (DPT)) were applied to noisy speech. The plots depict the obtained mean relative errors including 95 % confidence intervals. By comparing the performance curves the relative contributions of the auditory-based preprocessing and the Bayesian mixture tracking to the overall performance of our framework can be judged.

First, a comparison between the results for LPC+BMT and LPC+DPT highlights that the proposed formant tracking significantly improves performance compared to the tracking algorithm used by *Snack*, except for the estimation of F2 when speech is degraded by babble noise. In this case performance slightly decreases. Next, by comparing the performance of ABP+BMT to that of LPC+BMT we see that the proposed auditory-based preprocessing is superior to LPC analysis, particularly for F1 and F2. The most likely reason for minor problems in the extraction of F3 is that the used preemphasis by 6 dB/oct may not be a valid compensation of the spectral tilt in high-frequency regions. The hypothesis that the reduced frequency resolution (logarithmic compared to linear arrangement of channel center frequencies) may limit performance for F3 could be ruled out via additional tests. Overall the results illustrate that both the preprocessing and the formant tracking substantially contribute to the superior performance of our framework compared to existing approaches. Nevertheless, the relative contribution of the tracking algorithm exceeds that of the preprocessing.

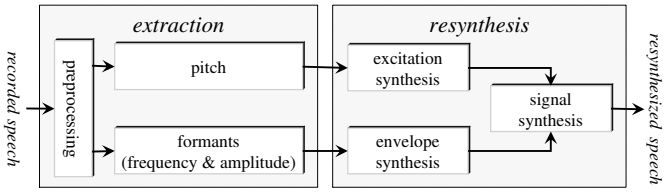

Fig. 10. The architecture of the system used for assessing the quality of pitch and formant extraction via instantaneous feature-based resynthesis.

Lastly, we investigated the effect of incorporating gender information. Therefore, we performed an additional test without using gender information. Irrespective of using gender or not there is a general decline in performance from low to high pitch-valued speech. However, the incorporation of the gender decision yielded significant relative improvements for the extraction of F2 and F3 of 4.7 % (+1.9 %;-1.8 %) and 5.4 % (+2.7 %;-2.8 %), respectively. More precisely, improvements for F2 are particularly present for high pitch-valued speech (female voices), whereas the extraction of F3 is improved for low pitch-valued speech (male voices). In contrast, no significant effect on the extraction of F1 (0.3 % (+1.3 %;-1.4 %)) could be observed.

VII. Application

Even though the results obtained on the VTR–Formant database demonstrate the efficiency of our method compared to existing approaches, it is still difficult to assess its quality in terms of its applicability to real-world scenarios. Therefore, we investigated the behavior of our framework in an interactive setting. More precisely, we implemented the extraction of formants and pitch in an online system in conjunction with a resynthesis solely based on these parameters [51]. Consequently the system reminds one of a parrot which repeats everything it hears. However, it is important to note that the focus of this work is on the extraction of the parameters. The resynthesis solely enabled us to assess the quality of the parameter extraction, insofar as we could judge the intelligibility of the resynthesized speech.

Fig. 10 shows a sketch of the system architecture. The *ToolBOS* framework for the software design, execution, and monitoring of real-time applications was used for the implementation [52]. The system runs on one computer with an Intel Quad Core processor (Q6600 @ 2.4 GHz). The signal delays introduced by the different system components are as follows:

- 40 ms for enhancing formant structure in spectrograms (12 ms for the application of the Gammatone filterbank + 28 ms for the smoothing during envelope calculation)
- 80 ms for the estimation of formant trajectories via the tracking framework (the delay is introduced by the sliding window technique used during Bayesian smoothing)
- 112.5 ms for the optional extraction of gender (12.5 ms for the ZCD histogram calculation + 100 ms for the pitch tracking)

Due to the parallel processing employed the overall signal delay within the application is reduced to 124.5 ms.

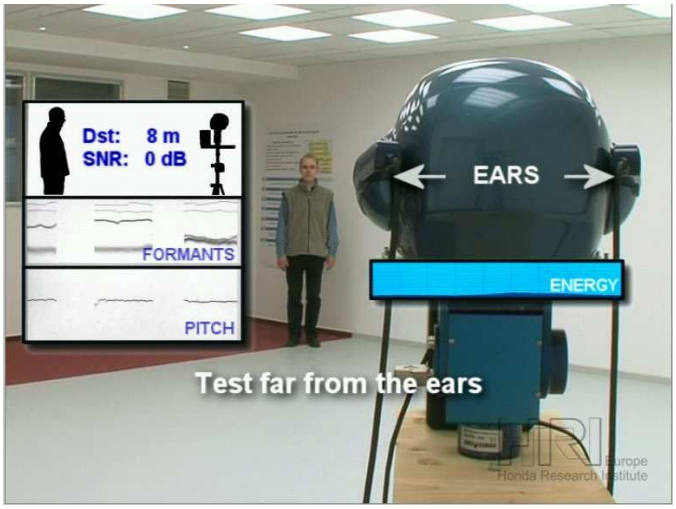

Fig. 11. Test of the system in a room with $\tau_{60} \approx 810$ ms at a distance of 8 m and an SNR of 0 dB. Only the left microphone (ear) is used in this experiment.

In contrast to using prerecorded sentences there is no ground truth when speaking to the online system and it is difficult to judge the correctness of the extraction. For this reason, we use the intelligibility of the resynthesized speech signal in order to subjectively assess the extraction performance of the complete system. This is reasonable, since an erroneous extraction of formants and pitch will result in the generation of unnatural sounds or deviating pitch trajectories, respectively. The capabilities of the system are demonstrated in the accompanying video (available at http://ieeexplore.ieee.org) where we first talk close to the microphone and then at a distance of 8 m, which is the limit we could test in our laboratory. An image of the video is shown in Fig. 11.

When evaluating the system, it has to be taken into account that we do not model unvoiced parts of speech and rather resynthesize all segments as voiced. Nevertheless, the resynthesized speech is highly intelligible in the case where we talk close to the microphone and only drops a little bit when we talk from far. This demonstrates the large amount of robustness against noise and echoes that our system achieves.

We tested the system in rooms featuring echo constants of 625 ms, 810 ms, and 975 ms. The scenarios resulted in SNR levels ranging from 15 dB to 0 dB (due to additional noise sources like computers and air conditioning) which were estimated based on recordings of the stationary noise signal and the speech signal plus noise. The most difficult setup with an 8 m speaker-microphone distance and a rather low SNR of $\approx$ 0 dB is shown in the accompanying video. Consequently, the system achieved better performance in all other scenarios.

VIII. Summary

In this paper, we proposed a framework for estimating formant trajectories from continuous speech. The focus of our work was to overcome the problems of existing approaches with respect to precision and robustness to narrow the gap between theoretical models for formant tracking and their applicability to real-world scenarios.

We believe that a processing following functional principles of the human auditory system may ultimately be more robust than common methods of spectral analysis. Therefore, the first building block of our system is a Gammatone filterbank which transforms the speech signal in the spectro-temporal domain. We illustrated that a subsequent spectral preemphasis and DoG filtering as well as a contrast enhancement mimicking a competition between filter responses enhances formants in spectrograms considerably. Experimental results showed a superior performance when using the proposed preprocessing compared to the commonly used LPC analysis.

The probabilistic framework for tracking formants constitutes the key innovation of our work. In contrast to previous approaches, we estimate the joint distribution of formants by adapting a mixture modeling approach. Thereby, the joint distribution is modeled via a non-parametric mixture of component distributions, each of them covering exactly one formant. We formulated an algorithm for the independent evolution of component distributions over time. Furthermore, we showed how a reclustering of component beliefs based on dynamic programming ensures the maintenance of multimodality, one of the key issues in multi-target tracking.

Our formant tracking algorithm contributes to the overall robustness of the framework in different ways. First, the tight coupling between Bayesian mixture filtering and adaptive frequency range segmentation incorporates interactions between neighboring formants. This particularly enhances performance when formants are close to each other. Second, the application of Bayesian smoothing on the obtained filtering distributions resolves ambiguities arising from uncertain noisy measurements by incorporating past as well as future observations in the process of estimating formant trajectories. Lastly, we do not rely on Kalman filtering/smoothing which would restrict estimation to unimodal Gaussian distributions. Rather we use a grid-based approximation of beliefs, thereby allowing an evaluation of multiple hypotheses in parallel. This is especially important when operating in noisy environments.

We incorporated a gender decision based on pitch and voicing information into the formant tracking regime via switching the probabilistic formant models based upon it.

Finally, results of comprehensive evaluations of our framework were presented. Tests on the VTR–Formant database yielded significant performance improvements compared to existing approaches. More importantly, evaluations on noisy speech considering different types of noise at various SNR levels highlighted the high robustness of our method against speech degradation. Even the incorporation of echoes did not disrupt the performance. For demonstrating the applicability of our method to real-world scenarios, we implemented the formant estimation framework within the scope of an online system. Thereby, an instantaneous feature-based resynthesis allowed us to assess the quality of the parameter extraction by judging the intelligibility of the resynthesized speech. The results demonstrated the high precision of our method even for large speaker-microphone distances of up to 8 m and SNR levels of ≈ 0 dB.

ACKNOWLEDGMENT

The authors would like to thank Horst-Michael Gross for helpful remarks regarding earlier versions of the framework, Miguel Vaz for his contributions to the feature-based resynthesis, as well as Tobias Rodemann and Christophe Lorin for their help concerning the implementation of the online system.

REFERENCES

[1] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Amer.*, vol. 80, no. 4, pp. 1016–1025, 1986.
[2] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Comm.*, vol. 22, no. 1, pp. 1–15, 1997.
[3] L. Deng and C. D. Geisler, "A composite auditory model for processing speech sounds." *J. Acoust. Soc. Am.*, vol. 82, no. 6, pp. 2001–2012, 1987.
[4] R. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 129–134, 1993.
[5] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Comm.*, vol. 27, no. 3-4, pp. 187–207, 1999.
[6] T. Baer, B. C. Moore, and S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times," *J Rehabil Res Dev*, vol. 30, no. 1, pp. 49–72, 1993.
[7] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 36–48, 1998.
[8] M. Padmanabhan, "Spectral peak tracking and its use in speech recognition," in *Proc. ICSLP*, 2000, pp. 604–607.
[9] B. Chen and P. Loizou, "Formant frequency estimation in noise," in *Proc. ICASSP*, vol. 1, 2004, pp. I – 581–584.
[10] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *J. Acoust. Soc. Amer.*, vol. 82, no. S1, p. 55, 1987.
[11] K. Xia and C. Espy-Wilson, "A new strategy of formant tracking based on dynamic programming," in *Proc. ICSLP*, vol. 3, 2000, pp. 55–58.
[12] Y. Zheng and M. Hasegawa-Johnson, "Formant tracking by mixture state particle filter," in *Proc. ICASSP*, vol. 1, 2004, pp. I – 565–568.
[13] M. Lee, J. vanSanten, B. Mobius, and J. Olive, "Formant tracking using context-dependent phonemic information," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 741–750, 2005.
[14] D. Toledano, J. Villardebo, and L. Gomez, "Initialization, training, and context-dependency in HMM-based formant tracking," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 14, no. 2, pp. 511–523, 2006.
[15] L. Deng, A. Acero, and I. Bazzi, "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 425–434, 2006.
[16] L. Deng, L. J. Lee, H. Attias, and A. Acero, "Adaptive kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 13–23, 2007.
[17] Q. Yan, S. Vaseghib, E. Zavareheib, B. Milnerc, J. Darchc, P. Whited, and I. Andrianakis, "Formant tracking linear prediction model using HMMs and kalman filters for noisy speech processing," *Comput. Speech Lang.*, vol. 21, no. 3, pp. 543–561, 2007.
[18] Y. Shi and E. Chang, "Spectrogram-based formanttracking via particle filters," in *Proc. ICASSP*, 2003, pp. I–168–171.
[19] I. Y. Özbek and M. Demirekler, "Vocal tract resonances tracking based on voiced and unvoiced speech classification using dynamic programming and fixed interval kalman smoother," in *Proc. ICASSP*, 2008, pp. 4217–4220.
[20] G. Kopec, "Formant tracking using hidden markov models and vector quantization," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 34, no. 4, pp. 709–729, 1986.
[21] A. Acero, "Formant analysis and synthesis using hidden markov models," in *Proc. EUROSPEECH*, 1999, pp. 1047–1050.
[22] K. Weber, S. Bengio, and H. Bourlard, "HMM2- extraction of formant structures and their use for robust ASR," in *Proc. EUROSPEECH*, 2001.
[23] J. Malkin, X. Li, and J. Bilmes, "A graphical model for formant tracking," in *Proc. ICASSP*, 2005, pp. 913– 916.
[24] C. Gläser, M. Heckmann, F. Joublin, C. Goerick, and H. M. Gross, "Joint estimation of formant trajectories via spectro-temporal smoothing and bayesian techniques," in *Proc. ICASSP*, 2007, pp. IV–477–480.

[25] J. Vermaak, A. Doucet, and P. Perez, "Maintaining multimodality through mixture tracking," in *Proc. ICCV*, vol. 2, 2003, pp. 1110–1116.

[26] D. Rudoy, D. N. Spendley, and P. J. Wolfe, "Conditionally linear gaussian models for estimating vocal tract resonances," in *Proc. of INTERSPEECH*, 2007, pp. 526–529.

[27] S. J. Godsill, A. Doucet, and M. West, "Monte carlo smoothing for nonlinear time series," *J. Amer. Stat. Assoc.*, vol. 99, no. 465, pp. 156–168, 2004.

[28] M. Heckmann, F. Joublin, and C. Goerick, "Combining rate and place information for robust pitch extraction," in *Proc. INTERSPEECH*, 2007.

[29] M. Heckmann, M. Moebus, F. Joublin, and C. Goerick, "Speaker independent voiced-unvoiced detection evaluated in different speaking styles," in *Proc. INTERSPEECH*, 2006, pp. 1670–1673.

[30] S. P. Whiteside, "Sex-specific fundamental and formant frequency patterns in a cross-sectional study." *J. Acoust. Soc. Amer.*, vol. 110, no. 1, pp. 464–478, 2001.

[31] L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. ICASSP*, 2006, pp. I – 369–372.

[32] G. Fant, *Acoustic theory of speech production*. The Hague, Netherlands: Mouton, 1960.

[33] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and A. M., "Complex sounds and auditory images," in *Proc. Symp. Hearing, Auditory Physiology and Perception*, 1992, pp. 429–446.

[34] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," Apple Computer, Tech. Rep. No.35, 1993.

[35] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "DARPA TIMIT acoustic-phonetic continuous speech corpus," NIST, Tech. Rep. NISTIR 4930, 1993.

[36] G. Fant, "Glottal source and excitation analysis," *STL-QPSR*, vol. 20, no. 1, pp. 85–107, 1979.

[37] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[38] K. N. Stevens, *Acoustic Phonetics*. Cambridge, Massachusetts, USA: MIT Press, 2000.

[39] S. Thrun, "Probabilistic robotics," *Comm. ACM*, vol. 45, no. 3, pp. 52–57, 2002.

[40] D. Fox, J. Hightower, L. Liao, D. Schulz, and G. Borriello, "Bayesian filtering for location estimation," *IEEE Perv. Comp.*, vol. 2, no. 3, pp. 24–33, 2003.

[41] G. Fant, "A note on vocal tract size factors and non-uniform f-pattern scaling," *STL-QPSR*, vol. 7, no. 4, pp. 22–30, 1966.

[42] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech. pitch and formant frequency," *J. Acoust. Soc. Amer.*, vol. 101, no. 5, p. 3194, 1997.

[43] D. Rendall, S. Kollias, C. Ney, and P. Lloyd, "Pitch (f0) and formant profiles of human vowels and vowel-like baboon grunts: the role of vocalizer body size and voice-acoustic allometry." *J. Acoust. Soc. Amer.*, vol. 117, no. 2, pp. 944–955, 2005.

[44] P. A. Cariani, "Temporal codes and computations for sensory representation and scene analysis." *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1100–1111, 2004.

[45] L. Deng, L. J. Lee, H. Attias, and A. Acero, "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," in *Proc. ICASSP*, 2004.

[46] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, ser. Monographs on Statistics and Applied Probability. CRC Press, 1993.

[47] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. ICASSP*, 2004, pp. I – 409–412.

[48] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9/10, pp. 341–345, 2001.

[49] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proc. ICSLP*, vol. 4, 2000, pp. 464–467.

[50] K. Mustafa and I. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 435–444, 2006.

[51] M. Heckmann, C. Gläser, M. Vaz, T. Rodemann, F. Joublin, and C. Goerick, "Listen to the parrot: Demonstrating the quality of online pitch and formant extraction via feature-based resynthesis," in *Proc. IROS*, 2008.

[52] A. Ceravola, M. Stein, and C. Goerick, "Researching and developing a real-time infrastructure for intelligent systems evolution of an integrated approach," *Robot. Auton. Syst.*, vol. 56, no. 1, pp. 14–28, 2008.

Claudius Gläser was born in Gera, Germany, in 1982. He received a Dipl.-Inf. degree (with honors) in computer science from the Technische Universität Ilmenau, Ilmenau, Germany, in 2006.

Since 2006 he is a Scientist at the Honda Research Institute Europe, Offenbach, Germany. Since 2007 he is also a Ph.D. student at the Honda Research Institute Europe focusing on the acquisition of sensorimotor competencies and their use for grounding language during social interaction. His research interests include speech analysis, language acquisition, neural information processing, and cognitive behavior control.

Martin Heckmann studied electrical engineering at the Universität Karlsruhe, Karlsruhe, Germany, and the Institut National des Sciences Appliquées, Lyon, France. He received a Dipl.Ing. degree and a Ph.D. degree in electrical engineering from Universität Karlsruhe in 1997 and 2003.

In 1998 he worked for Carl Zeiss Meditec, Dublin, CA. While working was a Research Assistant at Universität Karlsruhe from 1998 to 2003 he was engaged in the Collaborative Research Center Humanoid Robots at Universität Karlsruhe, and a visiting researcher at gipsa-lab Grenoble, France, in the framework of the EC research programm SPeech, HEaring and Recognition. Since 2004 he is a Senior Scientist at Honda Research Insitute Europe, Offenbach, Germany. His research interests include speech and image processing, auditory scene analysis, robust speech recognition, and language acquisition.

Frank Joublin received the Dipl.Ing. degree in electrical engineering from the Ecole Universitaire D'Ingénieurs de Lille (EUDIL), Lille, France, in 1987 and the Ph.D. degree in neuroscience from the University of Rouen, Rouen, France, in 1993.

From 1994 to 1998 he was a Postdoctoral Research Fellow at the Institute for Neural Computation, Ruhr-Universität Bochum, Bochum, Germany. His research activity there focused on the modeling of cortical plasticity, stereo-vision algorithms, and place cell representations for robot navigation. From 1998 to 2001 he was a Customer Project Manager at Philips Speech Processing, Aachen, Germany, where he developed voice-activated telephony applications. Since 2001, he is a Principal Scientist at the Honda Research Institute Europe, Offenbach, Germany, and since 2008 a Board Member of the CoR-Lab Graduate School of the Bielefeld University, Bielefeld, Germany. His research interests include auditory scene analysis, language and semantic acquisition, and developmental robotics.

Christian Goerick studied electrical engineering at the Ruhr-Universität Bochum, Bochum, Germany, and at the Purdue University, West Lafayette, IN, USA. He received a Dipl.Ing. degree in electrical engineering and a Ph.D. degree in electrical engineering and information processing from the Ruhr-Universität Bochum in 1993 and 1998.

From 1993 to 2000 he was with the Institute for Neural Computation, Ruhr-Universität Bochum, where he became a Research Assistant, Doctoral Worker, and Project Leader in 1993, and a Post-Doctoral Worker, Project Leader, and Lecturer in 1998. The research was concerned with biologically motivated computer vision for autonomous systems and learning theory of neural networks. Since 2000 he is a Chief Scientist at the Honda Research Institute Europe, Offenbach, Germany. His special interest is in embodied brain-like intelligence covering research on behavior based vision, audition, behavior generation, cognitive robotics, car assistant systems, systems architecture as well as hard- and software environments.