Robust Formant Tracking in Echoic and Noisy Environments

Claudius Gläser, Martin Heckmann, Frank Joublin, Christian Goerick

2010

Preprint:

This is an accepted article published in Proceedings of the 9th ITG Conference on Speech Communication. The final authenticated version is available online at: https://doi.org/[DOI not available]

Robust Formant Tracking in Echoic and Noisy Environments

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

Honda Research Institute Europe, Carl-Legien-Strasse 30, 63073 Offenbach, Germany
{firstname.lastname}@honda-ri.de
www.honda-ri.de

Abstract

Despite the fact that formant extraction has been investigated for a long time it still remains a challenging task. Particularly in real-world environments, where noise and echoes are detrimental factors for speech processing, existing methods for formant extraction yield unfavorable results. Here, we present a framework for formant tracking which is specifically tailored for application in such difficult settings. Keys to our method are, firstly, an auditory inspired preprocessing which enhances formants in spectrograms and, secondly, a probabilistic scheme which estimates the joint distribution of formants. Especially the latter contributes to the robustness of our system as it naturally considers the uncertainty inherent to the speech data. We demonstrate the favorable performance of our framework by a comprehensive evaluation on a publicly available database as well as in form of an online system operating under real-world conditions.

1 Introduction

Formants are the resonance frequencies of the vocal tract and appear as energy concentrations in the spectral domain. Formant trajectories are of primary interest in the areas of speech recognition and speech synthesis. However, their use in current systems is limited, since common methods for formant extraction lack in precision, robustness, and computational efficiency.

Here, we propose a framework for formant extraction [1] which is specifically suited to operate in noisy and echoic environments. As illustrated in Fig. 1, the system comprises an auditory inspired preprocessing to enhance formants in spectrograms and a subsequent probabilistic tracking scheme which extracts continuous formant trajectories. We further incorporate a gender detection based on pitch extraction and voiced-unvoiced classification. The gender decision is used as additional information which modulates the probabilistic tracking.

In the following we give a detailed description of the processing blocks. Next, we demonstrate the superior performance of our approach compared to existing approaches. Therefore, results of extensive tests on a publicly available database are presented for both clean speech as well as speech degraded by noise and echoes. Finally, we present an online system which verifies the suitability of the framework to operate in real-world environments.

2 Formant Enhancement

We initially transform the speech signal into the spectro-temporal domain by using the Patterson-Holdsworth auditory filterbank [2]. This filterbank resembles neurophysiological findings on the human auditory system, specifically the cochlea. Our setup comprises 128 Gammatone filters covering the frequency range from 80 Hz to 8 kHz. A subsequent rectification and low-pass filtering calculates the

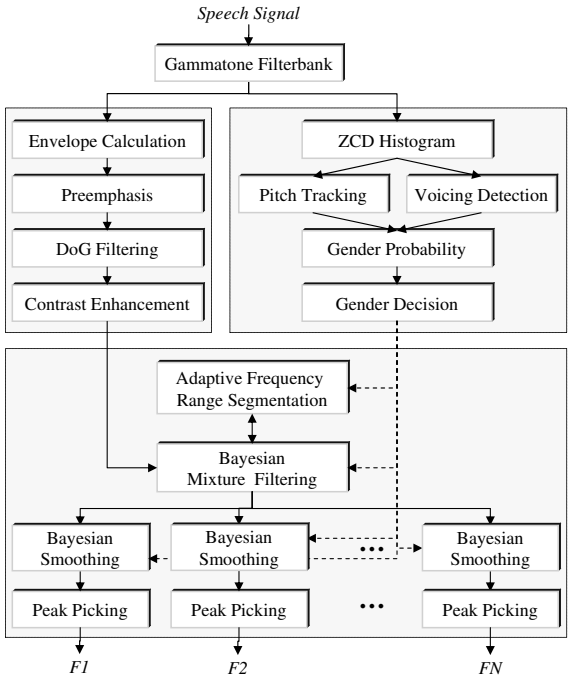

Figure 1: The framework for formant extraction.

envelope of the filter responses. Since formants are the resonance frequencies of the vocal tract, their extraction can be improved by eliminating the spectral influence of excitation and radiation contributing to human speech production. For doing so, we correct the spectral tilt via an emphasize of the spectral energy by +6 dB/oct.

Additionally, the emphasized spectrogram is smoothed along the frequency axis using a Laplacian kernel adjusted to the logarithmic arrangement of the Gammatone filterbank's channel center frequencies. By doing so, the harmonics spread and peaks are formed at formant locations. A subsequent normalization of the filter responses to the maximum at each sample as well as an application of a sigmoidal function further enhances the spectral contrast.

3 Formant Tracking

Based on the formant enhanced spectrogram we next extract formant trajectories using Bayesian filtering – a probabilistic technique for estimating a dynamic system's state. Bayesian filters represent the state at time t by a probability distribution over random variables x_t, called the belief $Bel(x_t)$. Assuming the filterbank is composed of N channels, the state space at time t can be written as $x_t = \{x_{1,t}, x_{2,t}, \ldots, x_{N,t}\}$. We model the target distribution $Bel(x_{k,t})$ by a weighted mixture of M filtering distributions $Bel_m(x_{k,t})$, such that each formant is represented by one mixture component (see Fig. 2):

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

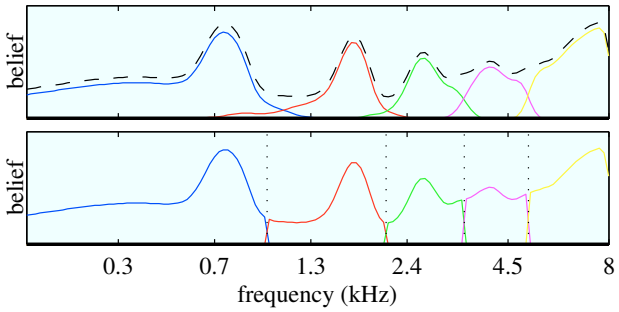

Figure 2: (top) The joint belief (dashed line) is represented as a mixture of components, each of them covering one formant. (bottom) A frequency segmentation and belief reclustering results in non-overlapping components.

Bayesian filtering targets the sequential estimation of the belief conditioned on all information contained in the sensor data $z = (z_1, \ldots, z_t)$. Let $p_m(x_{k,t}|x_{l,t-1})$ denote a model for the m-th formant's dynamics and $p(z_t|x_{k,t})$ an observation model, then the Bayesian filter recursion is

$$Bel_m^-(x_{k,t}) = \sum_{l=1}^{N} p_m(x_{k,t}|x_{l,t-1}) Bel_m(x_{l,t-1}) \quad (2)$$

$$Bel_m(x_{k,t}) = \frac{p(z_t|x_{k,t}) Bel_m^-(x_{k,t})}{\sum_{l=1}^{N} p(z_t|x_{l,t}) Bel_m^-(x_{l,t})} \quad (3)$$

$$\pi_{m,t} = \frac{\pi_{m,t-1} \sum_{k=1}^{N} p(z_t|x_{k,t}) Bel_m^-(x_{k,t})}{\sum_{n=1}^{M} \pi_{n,t-1} \sum_{l=1}^{N} p(z_t|x_{l,t}) Bel_n^-(x_{l,t})}. \quad (4)$$

We choose $p_m(x_{k,t}|x_{l,t-1})$ to be Gaussian and $p(z_t|x_{k,t})$ is given by the preprocessed spectral vector at time t.

The formulas show that the component distributions $Bel_m(x_{k,t})$ evolve independently over time. An interaction between the components only takes place during the calculation of the mixture weights $\pi_{m,t}$. To prevent belief degeneration, which may result in loosing track of formants, our framework additionally relies on a dynamic programming-based algorithm [1] which adaptively segments the frequency range into consecutive formant-specific regions $R_{1,t}, R_{2,t}, \ldots, R_{M,t}$ (see Fig. 2). 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 covering a certain formant. This reclustering of component beliefs incorporates short-term continuity constraints as well as long-term constraints on valid formant locations. We consequently recalculate the component beliefs, such that the mixture approximations of (1) before and after the reclustering procedure are equal in distribution:

$$\pi'_{m,t} = \sum_{x_{k,t} \in R_m} \sum_{n=1}^{M} \pi_{n,t} \cdot Bel_n(x_{k,t}) \quad (5)$$

$$Bel'_m(x_{k,t}) = \begin{cases} \frac{\sum_{n=1}^{M} \pi_{n,t} \cdot Bel_n(x_{k,t})}{\pi'_{m,t}} & , \forall x_{k,t} \in R_{m,t} \\ 0 & , \forall x_{k,t} \notin R_{m,t} \end{cases} \quad (6)$$

We next apply Bayesian smoothing on the obtained filtering distributions. In contrast to Bayesian filtering, this technique recursively estimates a smoothed distribution which relies on both past and future observations. Thereby, it works in the reverse time direction and uses the already obtained filtering distributions $Bel_m(x_t)$:

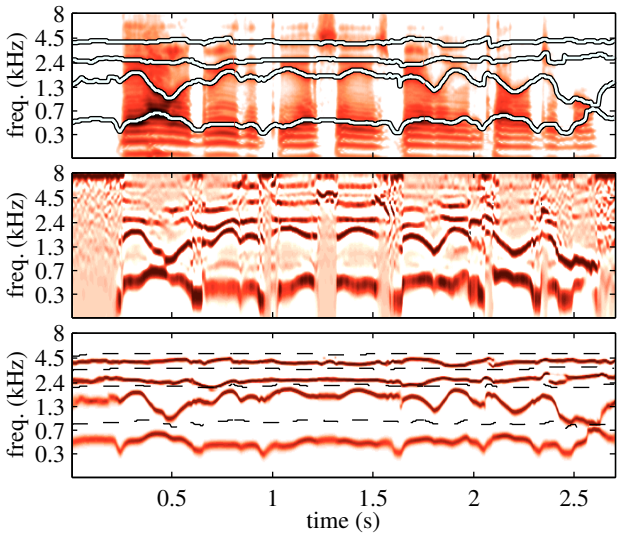

Figure 3: Example utterance "They all agree that the essay is barely intelligible.": (top) extracted formants overlaid to the original spectrogram, (middle) preprocessed spectrogram, and (bottom) smoothed component beliefs.

$$\widehat{Bel}_m^-(x_{k,t}) = \sum_{l=1}^{N} \widehat{Bel}_m(x_{l,t+1}) \cdot p_m(x_{l,t+1}|x_{k,t}) \quad (7)$$

$$\widehat{Bel}_m(x_{k,t}) = \frac{Bel_m(x_{k,t}) \cdot \widehat{Bel}_m^-(x_{k,t})}{\sum_{l=1}^{N} Bel_m(x_{l,t}) \cdot \widehat{Bel}_m^-(x_{l,t})} \quad (8)$$

For an online operation of our framework (see section 6) we consider a finite time horizon and apply a sliding window technique to implement Bayesian smoothing.

Finally, the m-th formant equals the peak location in the smoothed distribution of component m (see Fig. 3):

$$F_m(t) = \arg\max_{x_{k,t}} \left[\widehat{Bel}_m(x_{k,t}) \right] \quad (9)$$

4 Gender Extraction

Our probabilistic framework for tracking formants additionally uses information on the gender of a speaker. This is reasonable as formant profiles of female and male speech differ significantly. More precisely, female formant patterns are on average scaled to 20% higher frequencies than corresponding male patterns [3]. We incorporate gender information by relying on gender-specific models of both the formant dynamics $p_m(x_{k,t}|x_{l,t-1})$ and the formant locations $p_m(x_k)$. These models are instantaneously switched according to the decision provided by a gender detection.

To judge a speaker's gender we first extract pitch using an algorithm which combines information residing in the temporal and spectral representation of the speech signal [4]. Based on the harmonicity of the speech signal as well as the energy ratio between a high and a low frequency band we further perform a voiced-unvoiced classification [5]. By relying on a reference of typical fundamental frequencies of male and female speech, each pitch estimate in a voiced region consequently produces a gender probability. A temporal smoothing of these probabilities yields the final gender decision. Thereby, the smoothing suppresses fluctuations in gender decision and extends the result from voiced to unvoiced speech regions [1].

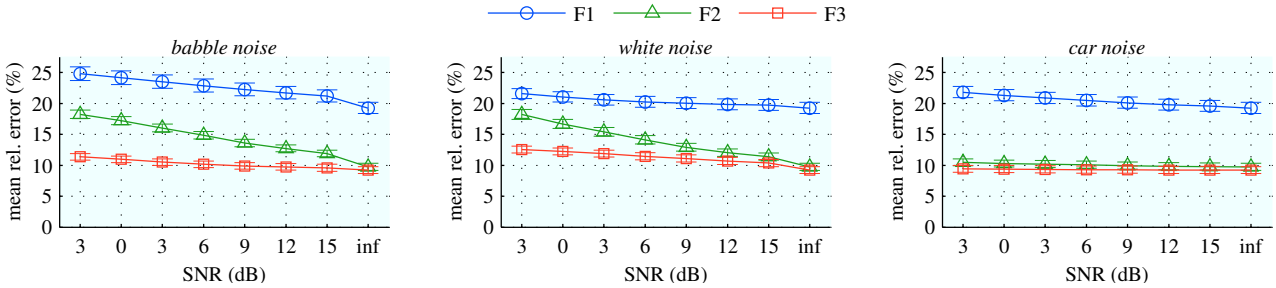

Figure 4: The plots depict the mean relative errors of our method when speech was degraded by various types of noise at different SNRs. Bars mark 95 % confidence intervals.

Formant	Mustafa [6]			Snack [7]			Praat [8]		
	babble	*white*	*car*	*babble*	*white*	*car*	*babble*	*white*	*car*
F1	50.1 (+2.1;-2.2)	39.0 (+2.6;-2.6)	64.2 (+2.5;-2.7)	8.3 (+1.8;-1.8)	52.5 (+2.1;-2.2)	45.3 (+2.7;-2.8)	50.5 (+2.1;-2.3)	79.4 (+0.9;-1.0)	74.0 (+1.4;-1.5)
F2	-0.1 (+2.3;-2.4)	19.8 (+2.8;-2.9)	28.4 (+4.0;-4.2)	11.0 (+2.8;-2.9)	39.1 (+2.9;-3.0)	33.4 (+4.6;-4.8)	29.4 (+3.8;-3.9)	63.0 (+2.3;-2.4)	64.0 (+2.8;-2.9)
F3	30.2 (+4.6;-4.9)	34.2 (+3.3;-3.6)	35.4 (+4.9;-5.3)	31.4 (+3.4;-3.5)	40.8 (+2.5;-2.6)	30.8 (+4.5;-4.7)	34.3 (+3.8;-4.1)	56.4 (+2.4;-2.6)	55.1 (+2.9;-3.0)

Table 1: Mean relative improvements (and 95 % confidence intervals) in % of our method compared to [6, 7, 8]

5 Results

To evaluate the proposed method we used the publicly available VTR–Formant database [9]. This database comprises utterances spoken by male and female speakers and additionally provides hand-labeled trajectories for the first three formants. For testing the robustness of our method we further added white noise, babble noise, and car noise to the clean speech signals. This was done for signal-to-noise ratios (SNRs) of -3 ... 15 dB. The performance of our method was measured by means of the relative deviation of the extracted formant locations with respect to the manual labels. The results depicted in Fig. 4 show that the error continuously increases when SNR decreases. Nevertheless, our method yields suitable estimates in all conditions without any significant drop in performance.

To judge the quality of our system we compared our results to those of existing approaches, i.e. to a recently proposed method also targeting noise robust tracking [6] as well as two widely-used speech processing tools (the *Snack Sound Toolkit* [7] and *Praat* [8]). The relative performance improvements achieved by our framework with respect to these methods are summarized in Table 1, where the relative improvements are averaged over all SNRs for each type of noise, respectively. 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 [6] reaches similar performance for F2. However, in all other cases we achieve relative performance enhancements ranging from 20 % to 60 %. In some cases, the improvements even exceed 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 a reverberation time $RT_{60} = 1100$ 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 results shown in Fig. 5 demonstrate that the incorporation of echoes impairs the performance of the algorithms, particularly for the extraction of F2. However, 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.

6 Application to Speech Synthesis

Formants are of primary interest for speech synthesis. However, formant-based synthesizers necessitate accurate information on the trajectories of the formants in order to produce natural sounds. To assess the quality of our framework we consequently implemented an online system which resynthesizes speech solely based on the extracted parameters, i.e. pitch and formants [10]. At the end the system reminds one of a parrot which repeats everything it hears. Here, our aim was to use the intelligibility of the resynthesized speech as a subjective measure for the performance of the feature extraction in a real-world environment. This is reasonable, since an erroneous extraction of formants and pitch will result in the generation of unnatural sounds or deviating pitch trajectories, respectively. Fig. 6 shows the architecture of the system, which has been implemented using the *ToolBOS* framework [11]. Overall, the system runs on one computer with an Intel Quad Core processor (Q6600 @ 2.4 GHz). Thereby, the processing introduces a signal delay of 124.5 ms.

We tested the system in rooms featuring reverberation times of 625 ms, 810 ms, and 975 ms. Due to additional noise sources (e.g. computers and air conditioning) the scenarios resulted in SNR levels ranging from 15 dB to 0 dB. The most difficult setup with an 8 m speaker-

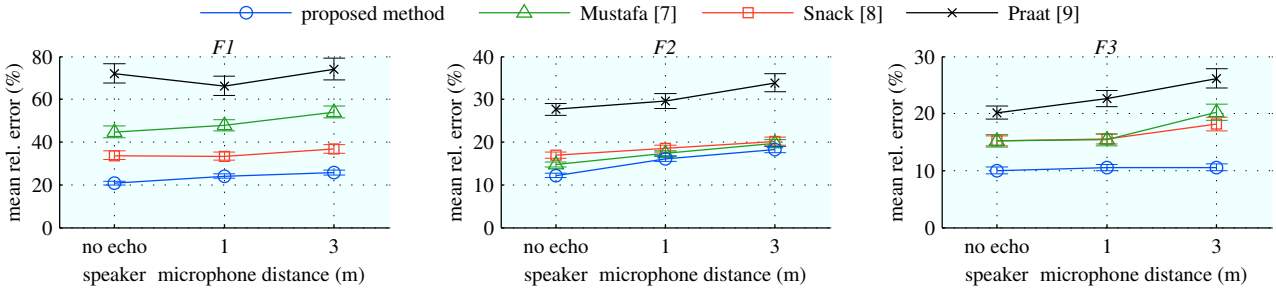

Figure 5: Evaluation in a room with echo constant $\tau_{60} = 1100\,\text{ms}$ using speaker-microphone distances of 1 and 3 meters. Plots of the mean relative errors as obtained by averaging over various types of additionally added noise (6 dB SNR) are shown. Bars mark 95 % confidence intervals.

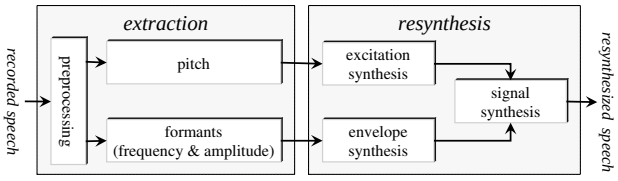

Figure 6: The online system first extracts pitch and formants and subsequently resynthesizes the speech based on the extracted parameters.

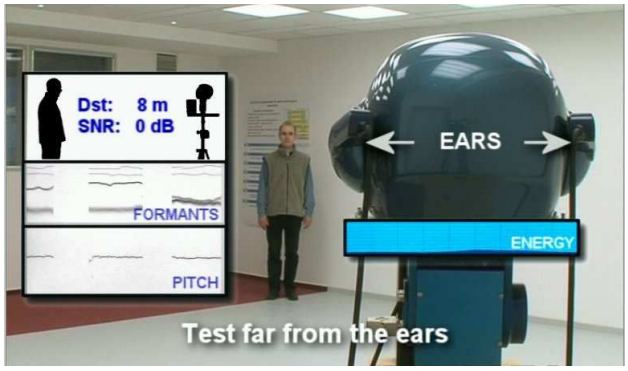

Figure 7: The most difficult experimental setting for testing the online system is shown.

microphone distance and a rather low SNR of $\approx 0\,\text{dB}$ is shown in Fig. 7. Our experiments with the system revealed that the resynthesized speech is highly intelligible when we talk close to the microphone. For larger speaker-microphone distances the speech intelligibility only drops a little bit. From the results we conclude that our framework shows a large amount of robustness against noise and echoes as they occur in real-world environments.

7 Summary

Noisy and echoic environments pose serious problems to common methods for formant extraction. In the design of our framework we explicitly considered these aspects. Firstly, we implemented a preprocessing which is inspired by the processing carried out in the human auditory system. Since humans perform marvelously well in such difficult conditions, this may lead a way to overcome the problems of existing approaches. In fact, additional tests [1] (whose results are not shown here) revealed that our auditory inspired preprocessing significantly contributes to the robustness of our framework as compared to using Linear Predictive Coding (LPC). Secondly, the probabilistic treatment of measurements (as it is inherent to our tracking scheme) extracts formant locations by integrating multiple individually ambiguous observations. The tight coupling between Bayesian filtering, Bayesian smoothing, and an adaptive frequency range segmentation estimates the joint distribution of formants, thereby taking possible interactions between neighboring formants into account. In our experiments we could show that the combination of both aspects yields a framework which significantly outperforms state of the art methods and is suitable to be applied in real-world scenarios.

References

[1] C. Gläser, M. Heckmann, F. Joublin, and C. Goerick. Combining auditory preprocessing and bayesian estimation for robust formant tracking. *IEEE Trans. Audio, Speech, and Lang. Process.*, 18(2):224–236, 2010.

[2] R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and Allerhand M. Complex sounds and auditory images. In *Proc. Symp. Hearing, Auditory Physiology and Perception*, pages 429–446, 1992.

[3] G. Fant. A note on vocal tract size factors and non-uniform f-pattern scaling. *STL-QPSR*, 7(4):22–30, 1966.

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

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

[6] K. Mustafa and I.C. Bruce. Robust formant tracking for continuous speech with speaker variability. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(2):435–444, 2006.

[7] K. Sjölander and J. Beskow. WaveSurfer - an open source speech tool. In *Proc. ICSLP*, volume 4, pages 464–467, 2000.

[8] P. Boersma. PRAAT, a system for doing phonetics by computer. *Glot. Int.*, 5(9/10):341–345, 2001.

[9] 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*, pages I – 369–372, 2006.

[10] 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.

[11] 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.*, 56(1):14–28, 2008.