Auditory-based Formant Estimation in Noise using a Probabilistic Framework

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

2008

Preprint:

This is an accepted article published in Proceedings of the INTERSPEECH. The final authenticated version is available online at: https://doi.org/[DOI not available]

Auditory-based Formant Estimation in Noise using a Probabilistic Framework

Claudius Gläser, Martin Heckmann, Frank Joublin, Christian Goerick

Honda Research Institute Europe
Carl-Legien-Str. 30, D-63073 Offenbach/Main, Germany
{claudius.glaeser, martin.heckmann, frank.joublin, christian.goerick}@honda-ri.de

Abstract

We recently introduced a computationally efficient framework for tracking formants which combines a biologically inspired preprocessing for enhancing formants in spectrograms with a probabilistic framework for estimating formant trajectories. In contrast to previously published approaches our tracking scheme relies on the joint distribution of formants rather than using independent tracking instances for each formant separately. Therewith more precise formant estimates could be obtained. In this paper we will briefly review our algorithm and extend it by using more sophisticated models of the formants underlying dynamics. Furthermore, we will dwell on the robustness of our method for speech degraded by various types of noise. A comprehensive evaluation on a large publicly available database containing hand-labeled formant trajectories shows significant performance improvements in both clean and noisy speech compared to state of the art approaches.

Index Terms: speech processing, formant extraction, tracking, robustness, Bayes procedures

1. Introduction

Human speech perception relies to a large extend on vocal tract resonance frequencies and their variation in time. These resonances are commonly referred to as formants and manifest themselves as energy maxima in the spectro-temporal domain. Formants have shown some potential for noise robustness in automatic speech recognition and further are of particular interest for speech synthesis, speech coding, and the development of hearing aids.

Despite their advantages automatic speech recognition systems do not rely on formants. This originates from the fact that common methods for their extraction lack in precision, robustness, and computational efficiency.

In this paper we demonstrate the robustness of our previously proposed method for formant extraction [1]. As shown in Fig. 1, the method involves a biologically inspired preprocessing for the enhancement of formants in spectrograms and a subsequent noise robust tracking via a Bayesian framework in order to extract formant trajectories.

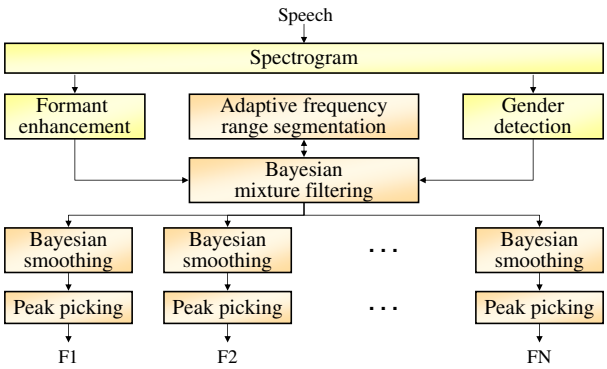

Figure 1: The architecture of the formant estimation system.

Firstly, we will briefly review the previously proposed framework and extend it by an enhanced modeling of formant dynamics. After that, we will present experimental results of extensive tests performed on a large publicly available database considering both clean and noisy conditions. The results will demonstrate the robustness of combining auditory-based preprocessing with probabilistic tracking techniques insofar as significant performance improvements compared to state of the art approaches are obtained.

2. Formant enhancement

The speech signal is initially transformed into the spectro-temporal domain by using the Patterson-Holdsworth auditory filterbank [2, 3] which is based on neurophysiological findings on the human auditory system, specifically the cochlea. We use a filterbank composed of 128 Gammatone filters covering the frequency range from 80 Hz to 8 kHz. A subsequent rectification and low-pass filtering calculates the 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. It has been shown that this influence can be adequately approximated by a first-order low-pass filter [4] which is valid at least for modal or creaky phonations being by far the most common ones [5]. For this reason, we used inverse filtering in order to correct the spectral tilt. More precisely, we emphasized 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 filterbanks 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

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 and in realistic scenarios the speech signal is degraded by noise. Previously published approaches usually try to overcome the former problem by using independent tracker instances which limits the success of

these methods. In contrast, we suggested the consideration of the joint distribution of formants in conjunction with continuity constraints and formant interactions.

The probabilistic tracking technique we proposed is based upon Bayes filters which provide an excellent framework for handling noisy observations [6]. They 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 $Bel(x_t) = p(x_t|z_1, \ldots, z_t)$. Their purpose is the sequential estimation of such beliefs over the state space conditioned on all information contained in the sensor data z_t [7].

Let $Bel^-(x_t)$ denote the predicted belief at time t which can be obtained via the application of the formants' underlying dynamics $p(x_t|x_{t-1})$. Then the belief at time t is calculated by correcting the predicted belief according to the preprocessed spectral energy distribution $p(z_t|x_t)$ and a normalization factor α. Thus, the standard Bayesian filter recursion can be written as follows:

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

$$Bel(x_t) = \alpha \cdot p(z_t|x_t) \cdot Bel^-(x_t) \quad (2)$$

Since standard Bayesian filtering is not an appropriate technique for tracking multiple formants at the same time, we adopted a mixture filtering approach recently introduced in the computer vision community [8]. Thereby the joint distribution $Bel(x_t)$ is modeled through a non-parametric mixture of M component beliefs $Bel_m(x_t)$ with associated weights $\pi_{m,t}$, so that each target is covered by exactly one mixture component:

$$Bel(x_t) = \sum_{m=1}^{M} \pi_{m,t} \cdot Bel_m(x_t) \quad (3)$$

Hence, by substituting the beliefs in Eq. (1) and (2) with Eq. (3) the Bayesian filter recursion can be rewritten with respect to the mixture modeling approach. Furthermore, since we want to estimate formant locations on a discrete grid defined by the channels of the Gammatone filterbank, a grid-based approximation of the belief is chosen. Thus, assuming that the filterbank is composed of N channels, the state space at time t can be written as $X_t = \{x_{1,t}, x_{2,t}, \ldots, x_{N,t}\}$ which leads to the following Bayesian filter recursion:

$$Bel(x_{k,t}) = \sum_{m=1}^{M} \pi_{m,t} \cdot Bel_m(x_{k,t}) \quad (4)$$

$$Bel_m^-(x_{k,t}) = \sum_{l=1}^{N} p_m(x_{k,t}|x_{l,t-1}) Bel_m(x_{l,t-1}) \quad (5)$$

$$Bel_m(x_{k,t}) = \frac{p(z_t|x_{k,t}) Bel_m^-(x_{k,t})}{\sum_{l=1}^{N} p(z_t|x_{l,t}) Bel_m^-(x_{l,t})} \quad (6)$$

$$\pi_{m,t} = \frac{\pi_{m,t-1} \sum_{k=1}^{N} p(z_t|x_{k,t}) Bel_m^-(x_{k,t})}{\sum_{n=1}^{M} \pi_{n,t-1} \sum_{l=1}^{N} p(z_t|x_{l,t}) Bel_n^-(x_{l,t})} \quad (7)$$

The formulas obtained are quite elegant, since mixture components evolve independently over time. But consequently, belief degenerations (i.e. component distributions becoming more and more diffuse) might occur and cause loosing track of the formants. For this reason, another algorithm which reclusters the beliefs at each time step is needed to ensure the maintenance of multimodality. Assuming such a function exists and returns sets $R_{1,t}, R_{2,t}, \ldots, R_{M,t}$ dividing the frequency range into contiguous formant-specific regions at each time step t, then the belief can be recomputed, so that the mixture approximations of (4) before and after the reclustering procedure are equal in distribution:

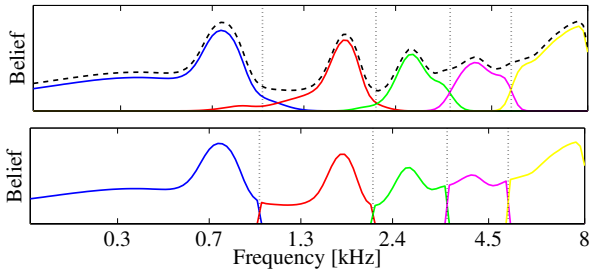

Figure 2: The mixture modeling approach. Top: the joint belief (dashed) is modeled by overlapping mixture components (solid); Bottom: the calculated cluster boundaries (dotted) are used for the recalculation of beliefs resulting in non-overlapping mixture components (solid).

$$\pi'_{m,t} = \sum_{x_{k,t} \in R_m} \sum_{n=1}^{M} \pi_{n,t} \cdot Bel_n(x_{k,t}) \quad (8)$$

$$Bel'_m(x_{k,t}) = \begin{cases} \frac{\sum_{n=1}^{M} \pi_{n,t} \cdot Bel_n(x_{k,t})}{\pi'_{m,t}} & , \forall x_{k,t} \in R_{m,t} \\ 0 & , \forall x_{k,t} \notin R_{m,t} \end{cases} \quad (9)$$

In this way, previously overlapping beliefs are separated via rearranging their component affiliation depending on associated mixture weights which results in a mixture of consecutive but separated components as shown in Fig. 2.

For the necessary segmentation of the frequency range into formant-specific non-overlapping regions we suggested a dynamic programming approach [1]. More precisely, a trellis was build up by which the former problem could be reformulated to the problem of finding the most likely path through the trellis for which the Viterbi algorithm offers an elegant solution. Since 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. However, when operating in noisy conditions a subsequent backward pass on the already obtained filtering distributions $Bel_m(x_{k,t})$ is recommended since it significantly enhances the noise robustness of the algorithm. Bayesian smoothing provides such a mechanism. It aims to recursively estimate a smoothed version $\widehat{Bel}(x_{k,t})$ of the belief, thereby depending on both past and future observations [9]:

$$\widehat{Bel}(x_{k,t}) = p(x_{k,t}|z_1, z_2, \ldots, z_t, \ldots, z_{T-1}, z_T) \quad (10)$$

$$\widehat{Bel}_m^-(x_{k,t}) = \sum_{l=1}^{N} \widehat{Bel}_m(x_{l,t+1}) \cdot p_m(x_{l,t+1}|x_{k,t}) \quad (11)$$

$$\widehat{Bel}_m(x_{k,t}) = \frac{Bel_m(x_{k,t}) \cdot \widehat{Bel}_m^-(x_{k,t})}{\sum_{l=1}^{N} Bel_m(x_{l,t}) \cdot \widehat{Bel}_m^-(x_{l,t})} \quad (12)$$

The final calculation of exact formant locations $F_m(t)$ can easily 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:

$$F_m(t) = \arg\max_{x_{k,t}} \left[\widehat{Bel}_m(x_{k,t})\right] \quad (13)$$

The probabilistic tracking scheme described above offers many advantages over common approaches. Firstly, it aims at estimating the joint distribution of formants which is a significant improvement since it enables us to adaptively solve the problem of resolving ambiguities emerging from measurements being unlabeled. Secondly, the segmentation of the frequency

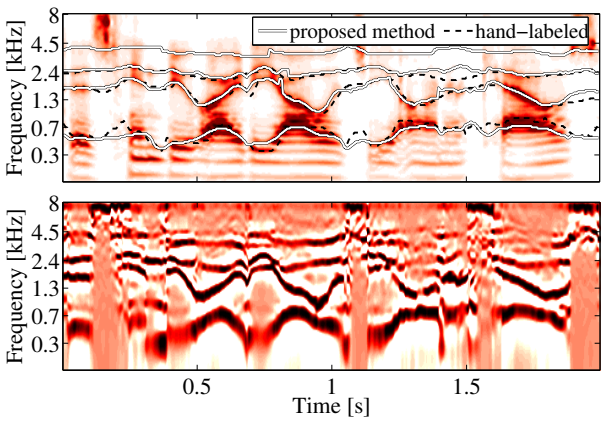

Figure 3: The obtained results for the utterance "Does Hindu ideology honor cows?" spoken by a male speaker. The estimated and hand-labeled formant trajectories superimposed to the original spectrogram (top) as well as the formant enhanced spectrogram (bottom) are shown.

range into formant-specific regions, which commonly limits the success of methods using independent tracking instances, is adaptively solved here, thereby taking the interaction of formants into account. Last, due to mixture components evolving independently over time, models of the formants underlying dynamics $p_m(x_{k,t}|x_{l,t-1})$ as well as a priori distributions of formant frequencies $p_m(x_{k,0})$ can be chosen for each formant individually. Furthermore, they can be adapted to different conditions such as gender, voicing, or context. Here we used gender-dependent pdfs which can be immediately switched according to the decision of a gender detection system. We assumed that the pdfs can be appropriately modeled by a normal distribution as is shown in Eq. (14) and (15).

$$p_m(x_{k,0}) \propto \mathcal{N}(f(x_k), \mu_m, \sigma_1^{(m)}) \quad (14)$$

$$p_m(x_{k,t}|x_{l,t-1}) \propto \mathcal{N}(f(x_k), f(x_l), \sigma_2^{(m)}) \cdot p_m(x_{k,0}) \quad (15)$$

Here $f(x_k)$ denotes the center frequency of the k-th filter channel. But in contrast to our proposal in [1] we added a mean tendency to $p_m(x_{k,t}|x_{l,t-1})$ which is reasonable since the probability that a formant performs a rising slope is much higher when the formant is actually located at a low than a high frequency. Additionally, an enhanced normalization on an extended grid was used for calculating the probabilities. These mechanisms further improved the precision of our method.

4. Experimental results

In order to evaluate the proposed method tests on the publicly available VTR–Formant database [10] were performed. As a subset of the widely-used TIMIT corpus, this database comprises a total of 516 utterances with additional hand-labeled trajectories for the first three formants. Thereby, 322 and 194 utterances were spoken by male and female speakers, respectively, covering a wide variety in dialects. For testing the robustness of our method we further added white noise, babble noise, and car noise at 7 different signal-to-noise ratios (SNRs) to the clean speech signals where non-speech samples were excluded from the energy calculation.

The setup of our algorithm comprises four mixture components corresponding to the first four formants (F1–F4). One additional component was used in order to cover the frequency range above F4. A block processing-based implementation of the system was used in order to make it applicable for an online system. Fig. 3 shows the formant trajectories obtained on a typical example drawn from the VTR–Formant database.

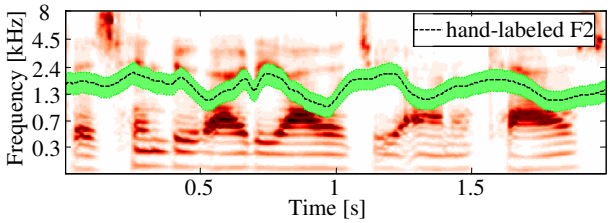

Figure 4: The region centered at hand-labeled formant trajectories (here shown exemplarily for F2) defines the range of formant estimates considered to be moderate errors. Estimates lying outside this region were treated as mistakes.

Furthermore, a comparison to the Snack Sound Toolkit used in WaveSurfer [11] and a recently proposed approach [12] also targeting noise robust tracking was carried out. Since the latter method uses a gender detection system as well, we used the gender decision of [12] in order to obtain comparable results. The algorithms were applied to the complete VTR–Formant database for both the clean and noisy conditions.

We calculated the absolute errors normalized by the formant locations given in the manual labels at time steps equally spaced by 10 ms. Thereby non-speech samples were excluded from the calculation. The relative errors were additionally bounded by formant-specific thresholds in order to account for large errors. Therefore we introduced a region centered at the exact formant frequency as illustrated in Fig. 4. This region mimics the range of moderate estimation errors. Thus, formant estimates lying outside this region were treated as mistakes by setting their errors to the corresponding threshold value. This is reasonable since formant-based speech recognition systems necessitate precise estimates. Hence, large errors have to be treated as mistakes regardless of their exact error values. Here the thresholds were set to the standard deviations of formant locations normalized by the mean formant frequencies. More precisely, an estimate was considered to be a mistake, if its deviation from the correct formant frequency exceeded 29.46 %, 23.76 %, or 13.74 % for F1, F2, and F3, respectively.

Plots of the mean relative errors achieved by our method under various noisy conditions are shown in the top row of Fig. 5. The results mirror the typical performance curves where the error continuously increases when SNR decreases. However, the performance of our algorithm never breaks down, rather it yields suitable estimates under all conditions. 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.

Fig. 5 further shows the relative performance improvements achieved by our method with respect to the algorithms used in [11] (middle row) and [12] (bottom row). It consistently outperforms the other methods in clean speech, thereby reaching a relative improvement in the range of 8 % up to 23 %. The improvement becomes even more significant in noisy conditions, except for the tracking of the first formant in babble noise, when our algorithm and the one proposed in [11] show identical behavior. Overall the results demonstrate that our method is significantly more robust against backround noise than the other methods tested. It is interesting to note that the number of formant estimates classified to be mistakes is much higher for the algorithms proposed in [11, 12] than for our method as can be seen in Table 1. This further stresses the high precision achieved by the proposed algorithm.

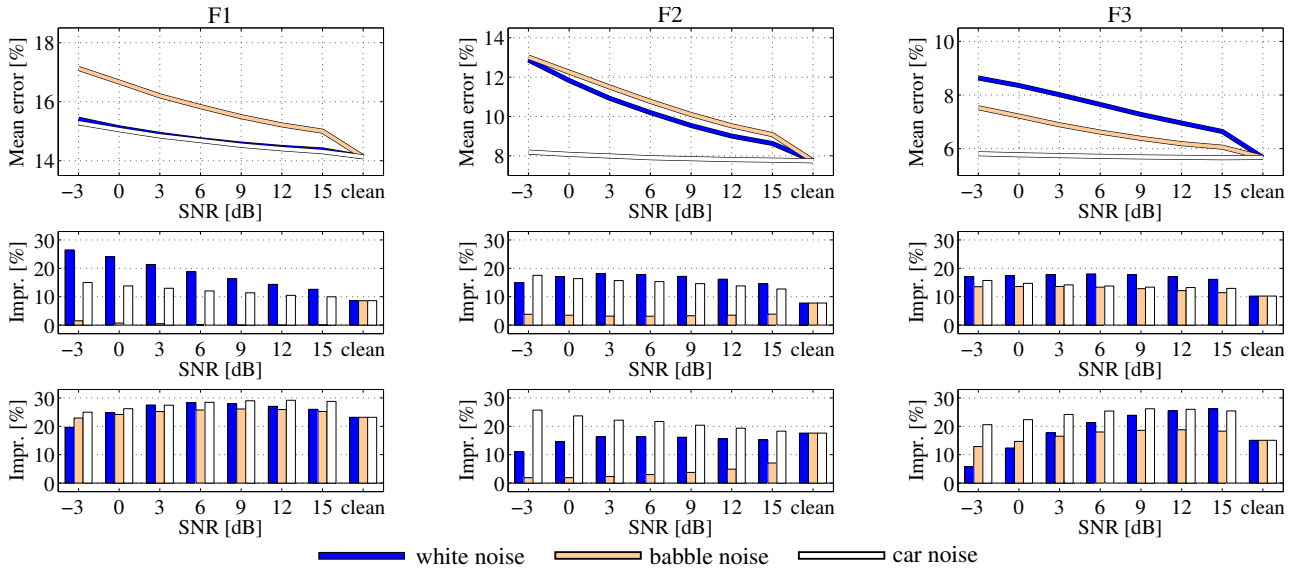

Figure 5: The results achieved by the proposed method (top row) as well as the relative improvement of our method with respect to the algorithms proposed in [11] (middle row) and [12] (bottom row).

formant	proposed method	Snack [11]	Mustafa [12]
F1	22.4 %	32.4 %	45.8 %
F2	17.7 %	21.0 %	19.5 %
F3	24.1 %	34.3 %	35.1 %

Table 1: The percentage of formant estimates classified to be mistakes as obtained via averaging over all noise types and SNRs.

However, we sometimes identified problems for female speech which are mainly caused by the used gender detection system being the same as in [12]. In clean speech it correctly detects male gender by 96 % in contrast to 56 % for female gender. The effect of this preference for male gender is best reflected by the results for F2, particulary when speech is degraded by babble noise. In such situations our method sometimes looses track of F2 by confusing it with neighboring formants, especially F3. Thus, by using a more powerful gender detection system, the performance of our method could further be improved.

5. Conclusion

We proposed a method for the robust estimation of formant trajectories combining an auditory-based preprocessing for enhancing formants in spectrograms with a probabilistic tracking framework. More precisely, as preprocessing the use of a Gammatone filterbank with a subsequent preemphasis and smoothing along the frequency axis is suggested. The Bayesian framework used for tracking formants sequentially estimates the joint distribution of formants rather than using independent tracker instances for each formant. By doing so, interactions of trajectories were considered which particularly improves the performance when the spectral gap between formants is small. We evaluated the proposed algorithm on a hand-labeled database in a wide range of noisy scenarios. In comparison to two other formant tracking algorithms our method showed superior performance in all cases tested, particularly when speech was degraded by noise.

To further improve our algorithm it should rely on an improved gender detector and more sophisticated models of the formant dynamics. These models might be learned from data, could include context-dependencies, and would cover more complex interactions of formants.

6. References

[1] 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. IEEE Int. Conf. on Audio, Speech and Signal Process. (ICASSP)*, 2007, pp. IV–477–480.

[2] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. H. Allerhand, "Complex sounds and auditory images," in *Proc. 9th Symp. on Hearing, Auditory Physiology and Perception*, 1992, pp. 429–446.

[3] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank," Tech. Rep. #35, Apple Computer Co., 1993.

[4] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, 2nd edition, 2000.

[5] D. G. Childers and C. K. Lee, "Vocal quality factors: analysis, synthesis, and perception," *J. of the Acoust. Soc. of America (JASA)*, vol. 90, no. 5, pp. 2394–2410, 1991.

[6] S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.

[7] D. Fox, J. Hightower, L. Liao, D. Schulz, and G. Borriello, "Bayesian filters for location estimation," *Pervasive Comput.*, vol. 2, no. 3, pp. 24–33, 2003.

[8] J. Vermaak, A. Doucet, and P. Pérez, "Maintaining multimodality through mixture tracking," in *Proc. IEEE Int. Conf. Comp. Vision (ICCV)*, 2003, vol. 2, pp. 1110–1116.

[9] S. J. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing for nonlinear time series," *J. of the American Stat. Assoc.*, vol. 99, no. 465, pp. 156–168, 2004.

[10] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. IEEE Int. Conf. on Audio, Speech and Signal Process. (ICASSP)*, 2006, pp. 60–63.

[11] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proc. IEEE Int. Conf. on Spoken Lang. Process. (ICSLP)*, 2000, pp. 464–467.

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