JOINT ESTIMATION OF FORMANT TRAJECTORIES VIA SPECTRO-TEMPORAL SMOOTHING AND BAYESIAN TECHNIQUES

C. Gläser[*†], *M. Heckmann*[*], *F. Joublin*[*], *C. Goerick*[*] *H. M. Groß* [†]

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

[†]Technical University of Ilmenau
Neuroinformatics and Cognitive Robotics
PO Box 10 05 65, D-98693 Ilmenau, Germany
horst-michael.gross@tu-ilmenau.de

ABSTRACT

We propose a method for the joint estimation of formant trajectories from spectrograms. Formants are enhanced in the spectrograms obtained from the application of a Gammatone filterbank via a smoothing along the frequency axis. In contrast to previously published approaches, the used tracking algorithm relies on the joint distribution of formants rather than using independent tracker instances. More precisely, Bayesian mixture filtering in conjunction with adaptive frequency range segmentation as well as Bayesian smoothing are used. The algorithm was evaluated on a publicly available database containing hand-labeled formant tracks. Experimental results show a significant performance improvement compared to a state of the art approach.

Index Terms— Speech processing, Bayes procedures, Tracking, Adaptive estimation, Dynamic programming

1. INTRODUCTION

Communication via speech is a key aspect in human-machine interaction. Current speech recognition system work well in idealized environments, but performance significantly drops when environments are characterized by variability. Recognition particularly becomes difficult for speech degraded due to large speaker-microphone distances and noise as in the interaction with a humanoid robot like Honda's ASIMO.

In contrast, humans perform marvelously well under such conditions. Designing a system based on findings on the functional principles of the human auditory system may lead a way to overcome the problems of state of the art systems. It is well known that human speech perception relies to a large extend on formant trajectories. Consequently, we propose a method for extracting formants which might ultimately be more robust to distortions than common feature extraction methods. 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. The results obtained on a large database with hand-labeled formant trajectories given in the final part of the paper show a significant improvement compared to a state of the art approach.

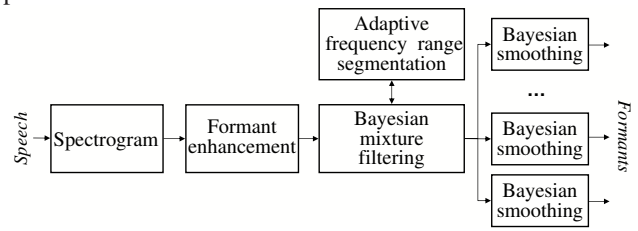

Fig. 1. The architecture of the formant estimation system.

2. FORMANT ENHANCEMENT

Firstly, the speech signal is transformed into the spectro-temporal domain by the application of a Gammatone filterbank with 128 channels covering the frequency range from 80 Hz to 8 kHz. Furthermore, the envelope of the filter responses is calculated via rectification and low-pass filtering.

According to Fant's linear source-filter theory speech is produced by a non-linear volume velocity source exciting a time-varying linear filter as well as radiation components. Thus, eliminating the spectral influence of excitation and radiation will significantly improve the extraction of formants from spectrograms.

At least for voiced sounds, the primary source is generated by the vibrating vocal folds converting the subpharyngeal steady airflow into a quasi-periodic train of flow pulses. In case of most common modal or creaky phonation a second-order low-pass filter can approximate the glottal flow spectrum [1],[2]. Hence, the glottal spectrum shows a monotonically decreasing characteristic of -12 dB/oct.

The principle opening from which speech is radiated is the mouth. A first-order high-pass filter approximates the relationship of lip volume velocity and sound pressure received at some distance [3]. For this reason, we model the spectral characteristics of the voiced excitation with a drop of -6 dB/oct and correct it via inverse filtering.

After this preemphasis is achieved, formants can be extracted by smoothing along the frequency axis which causes the harmonics to spread and form peaks at formant locations. Therefore a Laplacian kernel was used, where the kernel size was adjusted to the logarithmic arrangement of the Gammatone filterbank's channel center frequencies. This results in a constant kernel size on a linear frequency scale. Additionally, the filter responses were divided by the maximum at each sample so that formants become visible in signal parts with relatively low energy. A further contrast enhancement is achieved via a sigmoidal function.

3. FORMANT TRACKING

While tracking multiple formants, two general problems arise: Formant locations have to be estimated based on noisy observations and a data association problem has to be solved. That is, due to measurements being unlabeled, their allocation to one of the formants is a crucial step in order to break up ambiguities. In the case of tracking formants, focusing on one target yields only inferior results. Rather one has to consider the joint distribution of targets in conjunction with continuity constraints and target interactions.

Bayes filters offer an excellent framework for the necessary joint tracking. 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 $Bel(x_t)$. Their purpose is the sequential estimation of such beliefs over the state space conditioned on all information contained in the sensor data z_t [4]:

$$Bel(x_t) = p(x_t|z_1, z_2, \ldots, z_t) \tag{1}$$

Let $Bel^-(x_t)$ denote the predicted belief at time t which is corrected according to the preprocessed spectral energy distribution $p(z_t|x_t)$ and normalized by α, then the standard Bayes filter recursion can be written as follows:

$$Bel^-(x_t) \;=\; \int p(x_t|x_{t-1}) \cdot Bel(x_{t-1})\, dx_{t-1} \tag{2}$$

$$Bel(x_t) \;=\; \alpha \cdot p(z_t|x_t) \cdot Bel^-(x_t) \tag{3}$$

One crucial requirement for tracking multiple formants is the maintenance of multimodality. Standard Bayes filters allow the pursuit of multiple hypotheses. Nevertheless, in practical implementations these 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. Thus, the standard Bayes filters are not suitable for the joint estimation of formants. In order to avoid this problem the mixture filtering technique [5] was adopted. The key issue of this approach is the formulation of the joint distribution $Bel(x_t)$ 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 one mixture component:

$$Bel(x_t) \;=\; \sum_{m=1}^{M} \pi_{m,t} \cdot Bel_m(x_t) \tag{4}$$

Hence, the two-stage standard Bayes recursion can be reformulated with respect to the mixture modeling approach. Furthermore, since the application of the Gammatone filterbank already discretized the frequency domain, a grid-based approximation will be an adequate belief representation. Assuming N filter channels are used, the state space can be written as $X = \{x_1, x_2, \ldots, x_N\}$. Thus, the resulting prediction and update formulas are:

$$Bel^-(x_{k,t}) = \sum_{m=1}^{M} \pi_{m,t-1} \cdot Bel_m^-(x_{k,t}) \tag{5}$$

$$Bel(x_{k,t}) = \sum_{m=1}^{M} \pi_{m,t} \cdot Bel_m(x_{k,t}) \tag{6}$$

$$Bel_m^-(x_{k,t}) = \sum_{l=1}^{N} p(x_{k,t}|x_{l,t-1}) Bel_m(x_{l,t-1}) \tag{7}$$

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

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

Consequently, the new belief can be obtained straightforwardly by independently computing the belief of each component. An interaction of mixture components only takes place during the calculation of the new mixture weights.

However, the more timesteps would be computed the more diffuse component beliefs would become. Thus, from time to time a procedure for merging, splitting, and reclustering components has to be applied. Assuming such a function exists and returns sets $R_1, R_2, \ldots, R_{M'}$ which divide the frequency range into contiguous formant-specific regions, then the belief can be recomputed, so that the mixture approximation of Eq. (4) before and after the reclustering procedure are equal in distribution. Furthermore, the probabilistic character of mixture weights as well as component beliefs is maintained, since both still sum up to 1. This is achieved by updating the mixture weights according to Eq. (10) and recalculating the beliefs according to Eq. (11).

$$\pi'_{m,t} = \sum_{x_{k,t} \in R_m} \sum_{n=1}^{M} \pi_{n,t} \cdot Bel_n(x_{k,t}) \tag{10}$$

$$Bel'_m(x_{k,t}) = \begin{cases} \frac{\sum_{n=1}^{M} \pi_{n,t} \cdot Bel_n(x_{k,t})}{\pi'_{m,t}} & , \forall x_{k,t} \in R_m \\ 0 & , \forall x_{k,t} \notin R_m \end{cases} \tag{11}$$

In this way, previously overlapping beliefs are separated via rearranging their component affiliation depending on the mixture weights. Furthermore, mixture weights change according to the amount of beliefs a component gave off and got. This results in a mixture of consecutive but separated components by which multimodality is preserved.

In order to find optimum component boundaries a new variable $x_{k,t}^{(m)}$ encoding the assignment of state x_k to component m at time t is introduced. Therewith the trellis shown

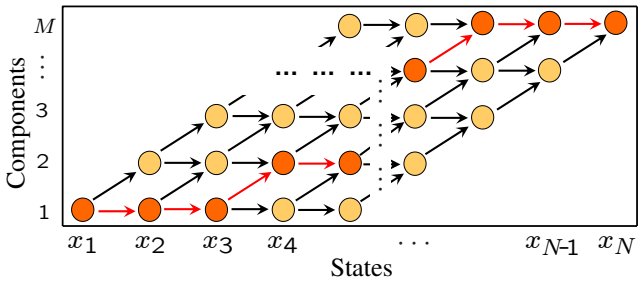

Fig. 2. The trellis used for frequency range segmentation.

in Fig. 2 can be build. By stipulating that $x_{k,t}^{(m)}$ becomes true only if it's corresponding node is part of a path from the lower left to the upper right all possible frequency range segmentations are considered by the trellis. Thus, optimum component boundaries can be found by calculating the most likely path. What remains is an appropriate choice of node and transition probabilities. Here the following formula was used:

$$p(x_{k,t}^{(m)}) = p_m(x_{k,0}) \cdot Bel_m(x_{k,t}) \tag{12}$$

Since the belief represents the past segmentation updated according to the Bayesian filtering recursion, this formula applies a data-driven segment continuity constraint. Furthermore, the a priori probability density function (pdf) antagonizes segment degeneration by application of long-term constraints. The transition probabilities can not easily be obtained. Thus, they were set to an empirically chosen value of 0.5. Finally, the most likely path can be computed by the Viterbi algorithm.

Having such an algorithm for finding optimum component boundaries at hand, we are able to apply the Bayesian mixture filtering technique. This method does not just yield the filtering distribution, it rather adaptively divides the frequency range into formant-specific segments represented by mixture components. Thus, any further processing can be restricted to these segments.

Uncertainties already included in spectrograms can not be completely resolved by Bayesian filtering. They rather result in diffuse beliefs at these locations. This limit is reasonable, because Bayes filters rely on the assumption of the underlying process to be Markovian. Thus, the beliefs only depend on past observations. In order to achieve continuous trajectories also future observations have to be considered. That is why the Bayesian smoothing technique was used additionally [6].

Let $\widehat{Bel}(x_{k,t})$ denote the belief of $x_{k,t}$ regarding both past and future observations. Then the smoothed component belief can be obtained by:

$$\widehat{Bel}(x_{k,t}) = p(x_{k,t}|z_1, z_2, \ldots, z_t, \ldots, z_{T-1}, z_T) \tag{13}$$

$$\widehat{Bel}^{-}_m(x_{k,t}) = \sum_{l=1}^{N} \widehat{Bel}_m(x_{l,t+1}) \cdot p(x_{l,t+1}|x_{k,t}) \tag{14}$$

$$\widehat{Bel}_m(x_{k,t}) = \frac{Bel_m(x_{k,t}) \cdot \widehat{Bel}^{-}_m(x_{k,t})}{\sum_{l=1}^{N} Bel_m(x_{l,t}) \cdot \widehat{Bel}^{-}_m(x_{l,t})} \tag{15}$$

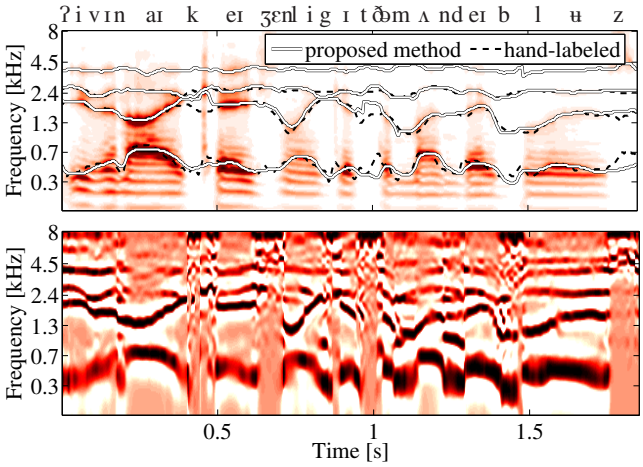

Fig. 3. The obtained results for the utterance "Even I occasionally get the Monday blues!" spoken by a male speaker. The estimated and hand-labeled formant trajectories superimposed to the original spectrogram (top) as well as the formant enhanced spectrogram (bottom) are shown.

This smoothing technique works in a very similar fashion as the Bayes filters, but in the reverse time direction. It recursively estimates the smoothing distribution based on system dynamics $p(x_{t+1}|x_t)$ as well as previously obtained filtering distributions $Bel(x_t)$. Therewith uncertainties by multiple hypotheses or diffuse filtering beliefs can be resolved.

What remains is the calculation of exact formant locations. Since the beliefs obtained are unimodal, this can be easily done via peak picking, such that the location of the m-th formant at time t equals the peak location in the smoothing distribution of component m:

$$F_m(t) = \arg\max_{x_k} \left[\widehat{Bel}_m(x_{k,t})\right] \tag{16}$$

4. EXPERIMENTAL RESULTS

The experimental setup comprises four mixture components corresponding to F1-F4. Additionally, one component covering the frequency range above F4 was used. Each of them offers gender-dependent Gaussian pdfs which can be immediately switched according to the decision of a gender detection system. A block processing based implementation of the system was used to make it applicable for an online system.

In order to evaluate the proposed method tests on the VTR–Formant database [7], a subset of the widely-used TIMIT corpus [8] with hand-labeled trajectories for F1-F3, were performed. Thereby we counted a total of 516 utterances composed of 322 male and 194 female speech sequences, respectively. The results obtained on a typical example drawn from this database are shown in Fig. 3.

Furthermore, a comparison to a state of the art approach proposed in [9] was carried out. This method uses a gender detection system, too. For this reason we also used the algorithm proposed in [9] for detecting gender in order to obtain comparable results. Both algorithms were applied to the

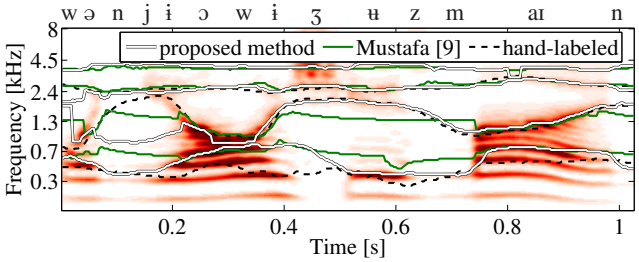

Fig. 4. The typical algorithm-specific errors shown on the exemplary phrase "when you always use mine".

complete VTR–Formant database. The estimated tracks were visually inspected which revealed differences in the characteristics of the two algorithms. This is illustrated by the example shown in Fig. 4 where the reader's focus should be set on F2.

As can be seen, the proposed method almost always correctly estimates formant locations, but sometimes loses track. This particularly happens when fast formant transitions between phonemes occur. In contrast, the method proposed in [9] works in a much more static fashion by limiting formant estimation to rigid frequency bands. Hence, it more often misses exact formant locations but does not achieve such large relative errors as our method sometimes does. Nevertheless, we consider the behavior of our method as preferable as it captures the correct formant position more often.

In the following absolute errors normalized by exact formant locations were calculated at time steps equally spaced by 10 ms. Segments without speech were excluded from the calculation of the error. Additionally, we introduced formant-specific thresholds by which relative errors were bounded. This technique is motivated by the fact that formant-based speech recognition systems need precise estimates. Thus, the impact of an error on the correct formant configuration estimation is similar if the error was 50 %, 100 % or even larger. In all cases a confusion with the neighboring formant is likely and therefore spoils the recognition process. Hence, the thresholds denote maximum acceptable errors. Errors above the threshold were set to identical values. The thresholds were set to the standard deviations of formants normalized by the mean formant frequencies which were calculated from the VTR–Formant database. The therewith obtained results are summarized in Table 1. Additionally, Table 2 shows the amount of errors bounded for each algorithm.

These results demonstrate the efficiency of our method. It consistently outperforms the state of the art approach by an improvement of up to 15 %. For male speech an even larger relative improvement of up to 25 % is achieved. Nevertheless, some problems regarding female speech were identified which are caused by the used gender detection system. This correctly detects male gender by 96 % in contrast to 56 % for female gender. Thus, the performance of our method will further improve when using a more powerful gender detection system.

	proposed method		Mustafa [9]		relative
formant	mean	(std)	mean	(std)	improvement
F1	15.62	(10.35)	18.40	(10.57)	+ 15.11 %
F2	8.28	(7.68)	9.40	(7.87)	+ 11.96 %
F3	5.75	(4.93)	6.68	(5.00)	+ 13.86 %

Table 1. The mean and standard deviation of bounded relative errors (in [%]) obtained via application of both methods.

method	F1	F2	F3
proposed method	23.84 %	12.81 %	19.64 %
Mustafa [9]	35.98 %	12.92 %	23.14 %

Table 2. The amount of relative errors bounded.

5. CONCLUSION

In this paper a method for the estimation of formant trajectories was proposed. As preprocessing we suggested the use of a Gammatone filterbank with subsequent smoothing along the frequency axis. In contrast to previously published approaches, the used tracking algorithm relies on 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. Experiments showed that the proposed method consistently outperforms a state of the art approach.

However, until now simplified Gaussian formant dynamics were used. The usage of context-dependent dynamics covering complex interactions of formants, which might be learned from data, would be preferable. In this way the exchange of beliefs between components could be extended leading to an improved interaction of them.

6. REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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," Tech. Rep. NISTIR 4930, NIST, 1993.

[9] 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.