Advances in Statistical Estimation and Tracking of AM-FM Speech Components

Athanassios Katsamanis and Petros Maragos

School of Electrical and Computer Engineering
National Technical University of Athens, Athens, Greece
{nkatsam, maragos}@cs.ntua.gr

Abstract

In this paper we present two extensions of a statistical framework to demodulate speech resonances, which are modeled as AM-FM signals. The first approach utilizes bandpass filtering and a standard demodulation algorithm which regularizes instantaneous amplitude and frequency estimates. The second employs particle filtering techniques to allow temporal variations of the parameters that are connected with spectral characteristics of the analyzed signal. Results are presented on both synthetic and real speech signals and improved performance is demonstrated. Both approaches appear to cope quite satisfactorily with the nonstationarity of speech signals.

1. Introduction

"Speaking and hearing" machines is a prospect that is gaining increasing popularity nowadays. Expectations of this kind seem to be well justified by the rapid progress of speech technologies in the last decade. At the same time however, in such a context, it is obvious that research in the area is brought up against various challenges. For example, robustness in adverse conditions, e.g. in varying and noisy environments, is still an open issue. The need to efficiently cope with such problems and the failure of approaches based on the traditional linear speech model to provide a uniform solution, have motivated the exploration of alternative speech representations. Such representations may be used in speech recognition for extracting features that successfully capture speech nonstationarity. Other potential applications are in speech enhancement or speech synthesis.

1.1. Previous work

In [1] Maragos et al. proposed modeling each speech resonance as an amplitude and frequency modulated signal. Speech could then be considered to be a sum of such signals:

$$y(k) = \sum_i \alpha_i(k) cos[\phi_i(k)] \tag{1}$$

This model has been inspired by related experimental evidence and may account for the local nonstationarity of the speech signal. In [1] they also described the Energy Separation Algorithm (ESA) and provided an initial framework to allow the extraction of instantaneous frequency (IF) and amplitude (IA) from a mono-component AM-FM signal.

Based on the same model, Lu and Doerschuk [2], proposed a statistical formulation of the IF and IA estimation of speech resonances. They describe the dynamics of each resonance i by the following state-space (time-update) equations:

$$\alpha_i(k+1) = \beta_{\alpha_i}\alpha_i(k) + q_{\alpha_i} w_{\alpha_i}(k) \tag{2}$$

$$\nu_i(k+1) = \beta_{\nu_i}\nu_i(k) + q_{\nu_i} w_{\nu_i}(k) \tag{3}$$

$$f_i(k+1) = f_i(k) + q_{f_i} w_{f_i}(k) \tag{4}$$

In Eq. 4, f_i is the slowly varying part of the instantaneous frequency, that roughly corresponds to the widely accepted notion of formant frequency. It is modeled as random walk. In Eqs. 2 and 3, α_i is the instantaneous amplitude and ν_i is the frequency modulating signal of the resonance. They both appear as first order autoregressive (AR) processes. Their power and bandwidth may be controlled independently by the parameters q_{α_i}, q_{ν_i} and $\beta_{\alpha_i}, \beta_{\nu_i}$ respectively. The observation (measurement-update) equation is:

$$y(k) = \sum_{i=1}^{K} \alpha_i(k) \cos(\varphi_i(k)) + ru(k) \tag{5}$$

where y is the observed signal and K is the number of resonances. In Eq. 5 φ_i is the instantaneous phase of the resonance and can be readily expressed in terms of f_i and ν_i as:

$$\varphi_i(k) = \varphi_i(0) + 2\pi\, T_s \sum_{m=0}^{k-1} [f_i(m) + \nu_i(m)] \tag{6}$$

where T_s is the sampling frequency. The signals $u, w_{\alpha_i}, w_{\nu_i}, w_{f_i}$ are independent, identically distributed $N(0,1)$ stochastic processes. The parameter vector $\boldsymbol{\theta} = (\beta_{\alpha_i}, \beta_{\nu_i}, q_{\alpha_i}, q_{\nu_i}, q_{f_i}, u)$, $i = 1 \ldots K$, is determined mainly by spectral analysis of the model. In [2], the Model Based Demodulation Algorithm (MBDA) is proposed to estimate the instantaneous amplitudes and frequencies of speech signal using Extended Kalman Filtering (EKF) after proper initialization.

Other efforts to achieve decomposition of speech into modulated components include [3, 6, 4, 5] . Potamianos et al. apply a Gabor filterbank to isolate the resonances. Pai and Doerschuk [6] also use bandpass signals and extend the work of [2]. Rao and Kumaresan [4, 5] estimate instantaneous modulations based on a different model for speech. From a different point of view, Vermaak et al. in [7], regard speech as a time varying AR (TVAR) process in order again to account for its local nonstationarity. They apply particle filtering methods and achieve enhancement of the signal.

1.2. Contribution

The contribution of our paper is twofold. Firstly, we enhance instantaneous amplitude and frequency estimation by indirectly introducing constraints to the model (Eqs. 2-5) based

on a proper application of the Energy Separation Algorithm. Secondly, we consider the model parameters $\{q_{\alpha_i}, q_{\nu_i}\}$ to be slowly varying in time and we apply Particle Filtering methods in order to achieve *joint estimation* of both state and parameters. Both approaches allow the extraction of more robust estimates, even in the case of inexact initialization or wide fluctuations of the spectral properties of the speech signal.

The remainder of the paper is organized as follows: Details concerning the first proposed model enhancement along with some necessary background information and results are presented in section 2. We elaborate on the second direction in section 3. The paper ends with a discussion on the presented approaches and future research directions.

2. Combining Gabor-ESA and MBDA

2.1. Gabor-ESA

The ESA may be applied to efficiently compute the instantaneous amplitude and frequency of an AM-FM signal $y = a(t)\cos(\varphi(t))$. However, when the signal contains more than one modulated components, the algorithm does not apply directly. In such cases, the common approach is to isolate the components by Gabor bandpass filtering and demodulate each component separately [1]. However, for nonstationary signals, positioning the Gabor filters in the frequency domain properly is not straightforward.

2.2. Introducing Gabor-ESA in the state-space model

On the other hand, MBDA allows parallel extraction of the amplitude and frequency modulations, even for multicomponent signals, without bandpass filtering. As a downside, thorough experimentation with synthetic signals has demonstrated that the estimates obtained in this way may be extremely sensitive to the initial configuration of the Extended Kalman Filter (e.g. initial state, parameter choice) or spectral variations of the signal and may diverge in the presence of outliers .

To improve performance we constrain α_i and ν_i to evolve closely to the corresponding initial estimates obtained by applying the ESA to each component separated by Gabor bandpass filtering. Specifically, we modify the original model by considering two additional observation equations for each component:

$$y_{\alpha_i}(k) = |\alpha_i(k)| + r_{\alpha_i} u_{\alpha_i}(k) \quad (7)$$

$$y_{\nu_i}(k) = \nu_i(k) + f_i(k) + r_{\nu_i} u_{\nu_i}(k) \quad (8)$$

The values y_{α_i} and y_{ν_i} are pointwise-determined estimates of the instantaneous frequency and amplitude of component i at moment k. These are attained by applying the ESA to the output of a Gabor bandpass filtered window of the signal centered at moment k (Gabor-ESA). The Gabor filter is centered at frequency $\hat{f}_i(k|k-1)$ which is the current estimate of frequency f_i as given by Eq. 4 (prediction estimate of the EKF). The stochastic processes u_{α_i} and u_{ν_i} are independent and normally $(N(0,1))$ distributed and represent possible uncertainty in the validity of the Eqs. 7, 8. The amount of uncertainty accepted may be imposed by suitably choosing the r_{α_i} and r_{ν_i} parameters.

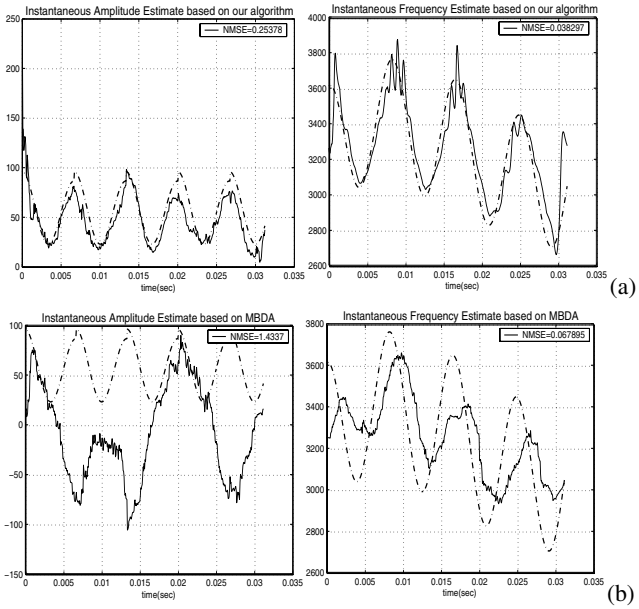

Figure 1: *Instantaneous Amplitude and Frequency estimates for the third component of the synthetic signal described in Subsection 2.3,(a) as obtained by the enhanced algorithm, (b) as obtained by the MBDA. The original amplitude and frequency signals are superimposed in the figures by dash-dotted lines.*

2.3. Experimental Results

We apply the enhanced demodulation algorithm to the synthetic signal $y_s(t) = \sum_{i=1}^{3} y_i(t)$, where:

$$y_i(t) = \gamma_i[1 + \kappa_i \cos(2\pi\, f_i^{AM} t)] \cos(2\pi \int_0^t f_{ci}(\tau) + f_i^{FM} \beta_i \cos(2\pi\, f_i^{FM} \tau) d\tau) \quad (9)$$

Each component is both amplitude and frequency modulated, with κ_i, β_i being the amplitude and frequency modulation indexes respectively.$\gamma_i's$ are scaling factors, $\boldsymbol{\gamma} = \begin{array}{ccc} 1500 & 150 & 60 \end{array}$. The values of the signal properties in Hz are:

$$\mathbf{f_c} = \begin{array}{ccc} 375 & 2312 & 3250 \end{array}$$

$$\mathbf{f^{AM}} = \begin{array}{ccc} 80 & 90 & 150 \end{array}$$

$$\mathbf{f^{FM}} = \begin{array}{ccc} 120 & 120 & 120 \end{array}$$

The model parameter vector $\boldsymbol{\theta}$ has been determined using the system identification procedure described in [2]. The parameters r_{α_i} and r_{ν_i} are set so that the estimates of ESA are trusted more for the less powerful components (2nd and 3rd) and much less for the strongest component (1st). For comparison, in Fig. 1(b) the instantaneous amplitude and frequency estimates of the third component as given by the MBDA are also presented. The Normalized Mean Square Error (NMSE) is also displayed on the graphs. Superior performance of the proposed approach can be observed.

In Fig. 2, results of speech analysis for the second speech resonance of the phoneme /ee/ are also presented, along with the corresponding estimates of the MBDA. It is worth mentioning that Unscented Kalman Filtering (UKF)[8] has also been tested

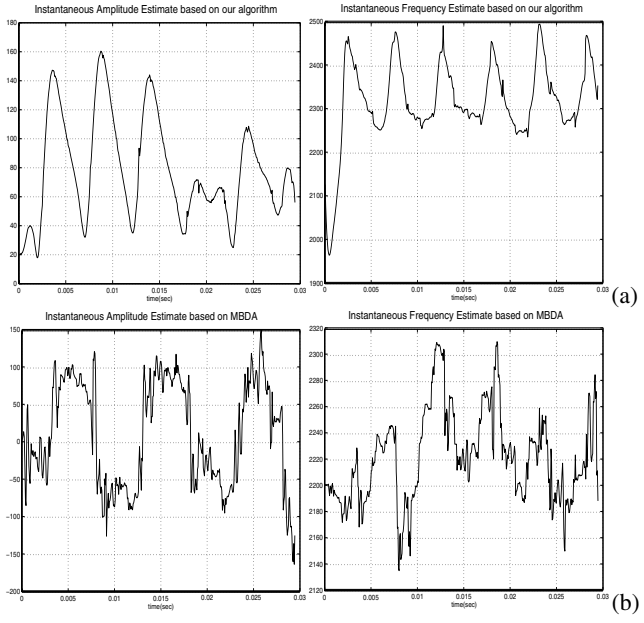

Figure 2: *Instantaneous Amplitude and Frequency estimates for the second resonance of the phoneme /ee/, (a) as obtained by the enhanced algorithm, (b) as obtained by the MBDA*

as an alternative to EKF in these experiments with quite similar results. An extensive presentation of experiments and results may be found at: http://cvsp.cs.ntua.gr/∼nassos.

3. Particle Filtering for Varying Parameters

3.1. Modeling Parameter Variations

Allowing for time-varying model parameters is another promising approach in the effort to enhance demodulation of speech. After spectral analysis of the speech representation in Eq. 5, Lu and Doerschuk [2] connect the parameters q_{α_i} and q_{ν_i} to the formant power and bandwidth respectively. However, these formant properties may vary significantly in an utterance and it seems too restrictive to keep the relative parameters fixed.

In our approach, we allow slow variations of q_{α_i} and q_{ν_i}. As in [7] we assume that the logarithms of their squares $\lambda_{\alpha_i} = \log q_{\alpha_i}^2$, $\lambda_{\nu_i} = \log q_{\nu_i}^2$ evolve according to a first-order Markov process fully specified by its initial state and state transition distributions:

$$p(\lambda_{\alpha_i}(0)) = N(2\log q_{\alpha_i}^0, \delta_{\lambda_{\alpha_i}^0}^2) \tag{10}$$

$$p(\lambda_{\alpha_i}(k)|\lambda_{\alpha_i}(k-1)) = N(\lambda_{\alpha_i}(k-1), \delta_{\lambda_{\alpha_i}}^2) \tag{11}$$

$$p(\lambda_{\nu_i}(0)) = N(2\log q_{\nu_i}^0, \delta_{\lambda_{\nu_i}^0}^2) \tag{12}$$

$$p(\lambda_{\nu_i}(k)|\lambda_{\nu_i}(k-1)) = N(\lambda_{\nu_i}(k-1), \delta_{\lambda_{\nu_i}}^2) \tag{13}$$

The initial parameters $q_{\alpha_i}^0$ and $q_{\nu_i}^0$ are chosen as indicated in [2] and $\delta_{\lambda_{\alpha_i}^0}, \delta_{\lambda_{\alpha_i}}, \delta_{\lambda_{\nu_i}^0}, \delta_{\lambda_{\nu_i}}$ are fixed at small values.

3.2. Particle Filtering

In order to achieve state estimation in case of varying parameters it is common to augment the state vector with the parameters and perform joint estimation based on the complete state-space representation. The state-space system defined in this way by Eqs. 2-5 and 10-13 is highly nonlinear, so, applying the EKF or UKF for estimation is inadequate. As an alternative, we suggest applying Particle Filtering [9, 10] which allows the representation of the probability distributions of interest by a number of properly sampled and weighted particles. The advantage is that estimation and tracking become possible even when the assumed model is highly nonlinear or non-Gaussian.

However, naive application of the generic Particle Filter, that is the Sequential Importance Sampling Algorithm, could be inefficient mainly due to high dimensionality of the augmented state vector. The problem may be reduced if we exploit the fact that, conditional on the parameters $\mathbf{q}$, the state-space equations are as assumed in the MBDA. Indeed, we may write:

$$\mathbf{x} = (\alpha_i, \nu_i, f_i | i = 1 \dots K) \tag{14}$$

$$\mathbf{q} = (q_{\alpha_i}, q_{\nu_i} | i = 1 \dots K) \tag{15}$$

$$p(\mathbf{x}_k, \mathbf{q}_{0:k}|y_{1:k}) = p(\mathbf{x}_k|\mathbf{q}_{0:k}, y_{1:k}) p(\mathbf{q}_{0:k}|y_{1:k}) \tag{16}$$

Based on arguments similar to those presented in [7], we may use the initial model to get an estimate of the state $\mathbf{x}$ by EKF (as in MBDA), given an approximation of $p(\mathbf{q}_{0:t}|y_{1:t})$. Such an approximation is obtained by using Sequential Importance Sampling of the *time-varying parameters* [10]: We update the approximation in time by sampling from an importance probability distribution π, resampling and then properly updating the weights w. The importance distribution and weights are defined as:

$$\pi(\mathbf{q}_k|\mathbf{q}_{0:k-1}, y_{1:k}) = p(\mathbf{q}_k|\mathbf{q}_{k-1}) \tag{17}$$

$$w(\mathbf{q}_{0:k}) = w(\mathbf{q}_{0:k-1}) w_k \tag{18}$$

$$w_t \propto p(y_k|\mathbf{q}_{0:k}, y_{1:k-1}) \tag{19}$$

The number of particles $\mathbf{q}_{0:k}^{(i)}$ drawn is N. An estimate of the augmented state is given by:

$$\hat{\mathbf{x}}_k = \sum_{i=1}^{N} \tilde{w}_{0:k}^{(i)} E_{p(\mathbf{x}_k|\mathbf{q}_k^{(i)}, y_{1:k})}\{\mathbf{x}_k\} \tag{20}$$

$$\hat{\mathbf{q}}_k = \sum_{i=1}^{N} \tilde{w}_{0:k}^{(i)} \mathbf{q}_k^{(i)} \tag{21}$$

$$\tilde{w}_{0:k}^{(i)} \triangleq \frac{w(\mathbf{q}_{0:k}^{(i)})}{\sum_{j=1}^{N} w(\mathbf{q}_{0:k}^{(j)})},\ i = 1, \dots, N. \tag{22}$$

$E_{p(\mathbf{x}_k|\mathbf{q}_k^{(i)}, y_{1:k})}\{\mathbf{x}_t\}$ is estimated by EKF, conditional on the last sample $\mathbf{q}_k^{(i)}$ of the particle trajectory i.

To avoid noisy fluctuations of the time-varying parameters and possible instabilities caused by outliers in the observation sequence, we update the time-varying parameters every L observed samples. To improve robustness, we estimate the weight updating factor $w_{k|L}$ of a particle as the likelihood of all L previously observed samples given the particle, $p(y_{(k-L+1):k}|\mathbf{q}_{k|L}^{(i)})$. Every L observed samples, the estimates of the states for every moment are finalized according to Eq.20 after all the particle weights have been computed and before resampling.

3.3. Experimental Results

Performance of the particle filtering approach to demodulation is demonstrated in Fig. 3 first for a random monocomponent

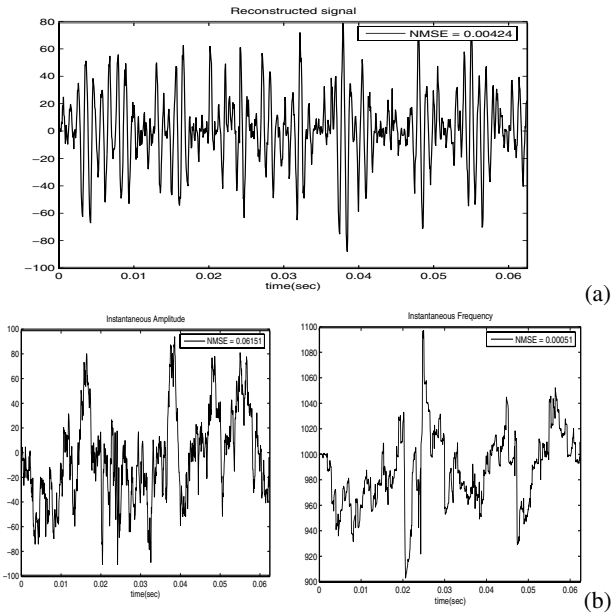

Figure 3: *The particle filtering approach tested on a random synthetic signal. The number of particles used is* $N = 100$ *and* $L = 100$*. (a) The reconstructed signal, (b)estimated instantaneous amplitude and frequency. Normalized mean square error is also displayed.*

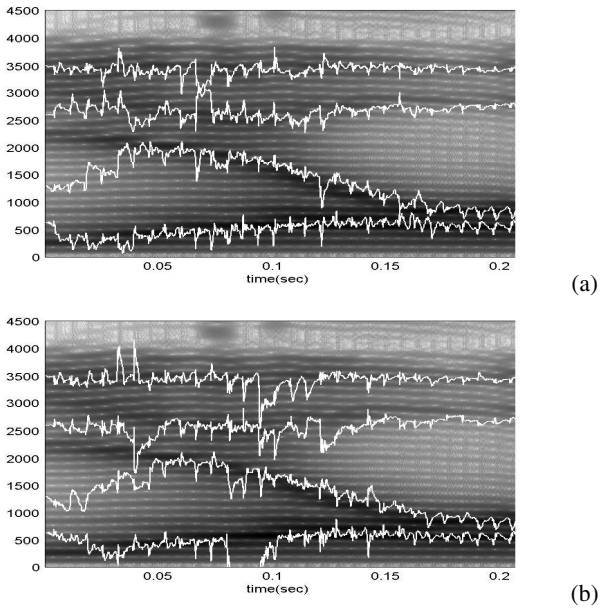

Figure 4: *Instantaneous Frequency estimates superimposed on the spectrogram of the word "yell". (a) Estimates by the particle filtering approach,* $N = 400$ *and* $L = 100$*, (b) estimates by MBDA appear more noisy and the estimated instantaneous frequency of the first resonance gets negative. In general, the proposed algorithm appears to perform better in tracking areas of high spectral density.*

amplitude and frequency modulated synthetic signal. This synthetic signal has been generated based on the model with varying parametes. We systematically observed that for such signals the MBDA failed to track instantaneous amplitude and frequency, while, with the proposed approach and the same initialization, the estimates were accurate.

In Fig. 4(a), real speech analysis results are also given. Instantaneous frequency estimates of four resonances are superimposed on the spectrogram of the word "yell". For comparison, estimates by the MBDA are given in 4(b). The latter estimates are worse in the sense that they exhibit erroneous variations. For example, the estimate corresponding to the first formant gets negative for a while. An extensive presentation of experiments and results may be found at the website indicated above.

4. Conclusions

In this paper, we present two directions along which the statistical framework proposed in [2] for the demodulation of speech exhibits improved performance. In the first approach, we combine Gabor-ESA and a statistical framework, which results in improved performance since robust estimates by the demodulation algorithm are properly utilized by the tracking algorithm. In the second approach, we allow time-variations of the spectrum-related parameters of the model. To handle the resulting complex model we apply particle filtering techniques, similar to the ones presented in [7]. Representative results concerning both synthetic and natural signals have been presented. In our on-going work we plan to incorporate the proposed techniques in speech recognition and speech enhancement applications and we expect them to facilitate robust speech processing in adverse conditions.

5. Acknowledgements

This work has been supported by the basic research program "Protagoras" of the National Technical University of Athens.

6. References

[1] P. Maragos, J. F. Kaiser and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis", IEEE Trans. on Signal Processing, Vol. 41, No. 10, October 1993

[2] S. Lu and P. C. Doerschuk, "Nonlinear Modeling and Processing of Speech Based on Sums of AM-FM Formant Models", IEEE Trans. on Signal Processing, Vol. 44, No. 4, April 1996

[3] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation", J. Acoust. Soc. Am., Vol. 99, No. 6, June 1996

[4] R. Kumaresan, A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications", J. Acoust. Soc. Am., Vol. 105, No. 3, March 1999

[5] A. Rao, R. Kumaresan, "On decomposing speech into modulated Components", IEEE Trans. on Speech and Audio Processing, Vol. 8, No. 3, May 2000

[6] W.-C. Pai and Peter C. Doerschuk, "Statistical AM-FM Models, Extended Kalman Filter Demodulation, Cramer-Rao Bounds, and Speech Analysis", IEEE Trans. on Signal Processing, Vol. 48, No. 8, August 2000

[7] J. Vermaak, C. Andrieu, A. Doucet and S. J. Godsill, "Particle Methods for Bayesian Speech Modeling and Enhancement of Speech Signals", IEEE Trans. Speech and Audio Processing, Vol. 10, No. 3, March 2002

[8] E. Wan and R. van der Merwe, "The Unscented Kalman Filter", in *Kalman Filtering and Neural Networks*, S. Haykin (editor), 2001

[9] J.S. Liu and R. Chen, "Sequential Monte Carlo Methods for Dynamical Systems", J. Amer. Statist. Assoc., vol.93, pp. 1032-1044, 1998

[10] A. Doucet, N. De Freitas, N.J. Gordon (editors), *Sequential Monte Carlo Methods in Practice*, New York: Springer-Verlag, May 2001