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 approaches based on the traditional linear speech model 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, has 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:

\[ y(k) = \sum_{i} \alpha_i(k) \cos[\phi_i(k)] \]  

(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:

\[
\alpha_i(k+1) = \beta_{\alpha_i} \alpha_i(k) + q_{\alpha_i} \nu_{\alpha_i}(k)
\]  

(2)

\[
\nu_i(k+1) = \beta_{\nu_i} \nu_i(k) + q_{\nu_i} \nu_{\nu_i}(k)
\]  

(3)

\[
f_i(k+1) = f_i(k) + q_{f_i} \nu_{f_i}(k)
\]  

(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, \( \alpha_i \) is the instantaneous amplitude and \( \nu_i \) 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 \( q_{\alpha_i}, q_{\nu_i} \) and \( \beta_{\alpha_i}, \beta_{\nu_i} \), respectively. The observation (measurement-update) equation is:

\[
y(k) = \sum_{i=1}^{K} \alpha_i(k) \cos(\varphi_i(k)) + ru(k)
\]  

(5)

where \( y \) is the observed signal and \( K \) is the number of resonances. In Eq. 5, \( \varphi_i \) is the instantaneous phase of the resonance and can be readily expressed in terms of \( f_i \) and \( \nu_i \) as:

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

(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 \( \theta = (\beta_{\alpha_i}, \beta_{\nu_i}, q_{\alpha_i}, q_{\nu_i}, q_{f_i}, u) \), \( i = 1 \ldots K \), 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. Pau 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 \( y(t) = a(t) \cos(2\pi f_0 t + \varphi(t)) \). 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 \( \alpha_i \) and \( \nu_i \) 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 \( y_{\alpha i} \) and \( y_{\nu i} \) 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 \( f_i(k)k - 1 \) which is the current estimate of frequency \( f_i \) as given by Eq. 4 (prediction estimate of the EKF). The stochastic processes \( u_{\alpha i} \) and \( u_{\nu i} \) are independent and normally 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 \( r_{\alpha i} \) and \( r_{\nu i} \) parameters.

![Figure 1: Instantaneous Amplitude and Frequency estimates for the third component of the synthetic signal described in Subsection 2.3.1(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.](image)

### 2.3. Experimental Results

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

\[
y_i(t) = \gamma_i [1 + \kappa_{\epsilon} \cos(2\pi f_{\epsilon}^{AM} t)] \cos(2\pi \int_0^t f_{\epsilon i}(\tau) + f_{\epsilon i}^{FM} \beta_i \cos(2\pi f_{\epsilon i}^{FM} \tau) d\tau) \quad (9)
\]

Each component is both amplitude and frequency modulated, with \( \kappa_{\epsilon}, \beta_i \) being the amplitude and frequency modulation indexes respectively. \( \gamma_i \) are scaling factors, \( \gamma_1 = 1500 \quad 150 \quad 60 \). The values of the signal properties in Hz are:

\[
f_{\epsilon} = 375 \quad 2312 \quad 3250
\]

\[
f_{\epsilon}^{AM} = 80 \quad 90 \quad 150
\]

\[
f_{\epsilon}^{FM} = 120 \quad 120 \quad 120
\]

The model parameter vector \( \theta \) has been determined using the system identification procedure described in [2]. The parameters \( r_{\alpha i} \) and \( r_{\nu i} \) 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.
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 $q_{\alpha i}$ and $q_{\nu i}$ 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 $q_{\alpha i}$ and $q_{\nu i}$.

As in [7] we assume that the logarithms of their squares $\lambda_{\alpha i} = \log q_{\alpha i}, \lambda_{\nu i} = \log q_{\nu i}$ evolve according to a first-order Markov process fully specified by its initial state and state transition distributions:

$$p(\lambda_{\alpha i}(0)) = N(2 \log q_{\alpha i0}, \delta_{\lambda_{\alpha i0}})$$

$$p(\lambda_{\alpha i}(k)|\lambda_{\alpha i}(k-1)) = N(\lambda_{\alpha i}(k-1), \delta_{\lambda_{\alpha i}}^2)$$

$$p(\lambda_{\nu i}(0)) = N(2 \log q_{\nu i0}, \delta_{\lambda_{\nu i0}}^2)$$

$$p(\lambda_{\nu i}(k)|\lambda_{\nu i}(k-1)) = N(\lambda_{\nu i}(k-1), \delta_{\lambda_{\nu i}}^2)$$

The initial parameters $q_{\alpha i0}$ and $q_{\nu i0}$ are chosen as indicated in [2] and $\delta_{\lambda_{\alpha i0}}, \delta_{\lambda_{\nu i0}}, \delta_{\lambda_{\alpha i}}, \delta_{\lambda_{\nu i}}$ 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 $q$, the state-space equations are as assumed in the MBDA. Indeed, we may write:

$$x = (\alpha_i, \nu_i, f_i | i = 1 \ldots K)$$

$$q = (q_{\alpha i}, q_{\nu i} | i = 1 \ldots K)$$

$$p(x_{k}, q_{0:k} | y_{1:k}) = p(x_{k} | q_{0:k}, y_{1:k}) p(q_{0:k} | y_{1:k})$$

Based on arguments similar to those presented in [7], we may use the initial model to get an estimate of the state $x$ by EKF (as in MBDA), given an approximation of $p(q_{0:k} | y_{1:k})$. 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 $\pi$, resampling and then properly updating the weights $w$. The importance distribution and weights are defined as:

$$\pi(q_{k} | q_{0:k-1} | y_{1:k}) = p(q_{k} | q_{k-1})$$

$$w(q_{0:k}) = w(q_{k-1} | y_{1:k}) w_{k}$$

$$w_{i} \propto p(y_{1:k} | q_{0:k}, y_{1:k-1})$$

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

$$\bar{x}_{k} = \sum_{i=1}^{N} w_{0:k}^{(i)} \{ x_{k} | q_{0:k}^{(i)}, y_{1:k} \}$$

$$q_{0:k}^{(i)} = \sum_{i=1}^{N} w_{0:k}^{(i)} q_{0:k}^{(i)}$$

$$w_{0:k}^{(i)} = \frac{w(q_{0:k}^{(i)} | y_{1:k}^{(i)} | y_{1:k-1}^{(i)})}{\sum_{j=1}^{N} w(q_{0:k}^{(j)} | y_{1:k}^{(j)} | y_{1:k-1}^{(j)})}, i = 1, \ldots, N.$$
amplitude and frequency modulated synthetic signal. This synthetic signal has been generated based on the model with varying parameters. 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