Large Deviation First Formant Demodulation Via Empirical Mode Decomposition And Multirate Frequency Transformations

Wenjing Liu and Balu Santhanam
Dept. of ECE, University of New Mexico
Albuquerque, NM, 87131

Email: wenjing@unm.edu, bsanthan@ece.unm.edu

***Abstract*—Existing formant estimation approaches are generally based on the linear predictive model assuming that each formant is a narrowband AM component. However, the first formant of most vowels contains a significant amount of frequency modulation, resulting in a large bandwidth compared to its formant frequency. In this paper, we first apply the empirical mode decomposition to extract the first formant, and then perform the demodulation to obtain its instantaneous amplitude and frequency using energy separation algorithm via multirate frequency transformations as proposed in prior work for wideband AM-FM demodulation. Finally we demonstrate that the estimates of the first formant based on the proposed approach are more precise than the LPC estimates or Teager-Kaiser energy operator based demodulation that assumes narrowband AM-FM components.**

***Index Terms*—Formant frequency and bandwidth estimation, linear predictive coding, empirical mode decomposition, energy separation algorithm, multirate frequency transformations**

I. Introduction

Formants are natural resonances of the vocal tract that are closely related to the vocal tract geometry as a function of the velum, the lips, the jaw and the tongue. They are visually observed as the resonance peaks in the spectrum of the voiced speech. The center-frequency and bandwidth of the formant associated with different vowels differ in a number of ways. The formant estimation of its center-frequency and bandwidth has significant implications in various speech applications. Existing formant estimation approaches are generally based on the *linear predictive coding* (LPC) [1] assuming that each formant is merely a narrowband AM component. In this regard, LPC is a parametric approach that does not model the spectral valleys properly, hence incapable of handling formants with considerable amount of frequency modulation. For example, the center-frequency of the first formant of many vowels is only around 500 Hz. Such a formant is expected to have a large *bandwidth-to-center-frequency ratio* (BW/CF) due to its inherent small *carrier-to-information-bandwidth ratio* (CR/IB) and *carrier-to-frequency-deviation ratio* (CR/FD) as defined in [2], thus resulting in a significant amount of frequency modulation.

This research is supported by the Airforce Research Laboratory through FA9453-14-1-0234.

II. Formant Estimation Via LPC

Formant estimation based on LPC [1] is widely used in acoustics and speech processing. LPC has been the dominant approach for parameter estimation of the discrete-time speech model such as pitch, short-time spectra and formant. Based on the source-filter theory, the basic discrete-time model for speech production is an all-pole filter representing the composite spectrum effects of the vocal tract

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{1}$$

the speech sample $\hat{s}_n$ can be predicted via an autoregressive (AR) predictor given by

$$\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k}, \tag{2}$$

where p is the order of the predictor and a_k denote the coefficients. The prediction error is defined as

$$e_n = s_n - \hat{s}_n = s_n - \sum_{k=1}^{p} a_k s_{n-k}. \tag{3}$$

The coefficients a_k of the LPC polynomial can be obtained by solving the set of linear equations via the matrix form $\mathbf{Ra} = \mathbf{r}$, where the autocorrelation matrix $\mathbf{R}$ is defined as

$$\mathbf{R} = \begin{bmatrix} r_0 & r_1 & r_2 & \dots & r_{p-1} \\ r_1 & r_2 & r_3 & \dots & r_{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p-1} & r_{p-2} & r_{p-3} & \dots & r_0 \end{bmatrix}. \tag{4}$$

The term r_{k-i} can be computed from the windowed speech signal over a finite interval given by

$$r_{k-i} = \sum_{n=-\infty}^{\infty} s_{n-k} s_{n-i} = \sum_{n=-\infty}^{\infty} s_n s_{n-(k-i)}. \tag{5}$$

Solving for the LPC coefficients requires the inversion of the matrix $\mathbf{R}$, which is in general computationally complex. As for

the autocorrelation matrix, since it is symmetric and Toeplitz, the inversion can be simplified using the Levinson-Durbin recursion method that is more efficient.

Assume that p is an even integer, the z-transform of the vocal tract transfer function can be represented by

$$H(z) = \frac{b_0}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{b_0}{\prod_{k=1}^{p/2} \left(1 - p_k z^{-1}\right)\left(1 - p_k^* z^{-1}\right)}, \tag{6}$$

where b_0 is a constant gain factor, $p_k = r_k e^{j\omega_k}$ and p_k^* are the corresponding roots of the denominator or poles of the transfer function that are a pair of complex conjugates.

With a sufficiently large sampling frequency F_s, each formant is considered to be a discrete-time sinusoid modulated by a decaying exponential as given by

$$F_k(n) = e^{-\delta_k n} \cos(\omega_k n), \tag{7}$$

where k denotes the k^{th} formant, ω_i and δ_k can be computed from the roots of the LPC polynomials corresponding to the k^{th} formant. The formant frequency and the associated bandwidth are then determined by the computed ω_k and δ_k via

$$f_k = \frac{F_s \cdot \omega_k}{2\pi}, \tag{8}$$

$$B_k = \frac{F_s \cdot \delta_k}{\pi}. \tag{9}$$

The formant location can also be determined via other methods using the LPC coefficients such as peak-picking on the frequency response of the transfer function.

The limitations of LPC analysis for formant estimation are two-fold. First of all, the LPC model assumes merely narrowband amplitude-modulation (AM) for each formant that does not take frequency modulation into account. As a result, it is only suitable for estimation of formants with small BW/CF, such as the formants that lie in the high frequency range. Besides, crude estimation of the LPC coefficients is prone to incur significant error, leading to inaccurate poles location of the spectrum.

III. Formant Estimation Via AM-FM Demodulation

Frequency modulation has been taken into consideration in areas such as audio synthesis, as proposed in [3] using the sinusoidal FM model. In general, AM-FM signals are time-varying sinusoids of the form [4]:

$$s(t) = a(t) \cos\left[2\pi \int_{-\infty}^{t} q(\tau)\, d\tau + \theta\right], \tag{10}$$

where the *instantaneous amplitude* (IA) is denoted by $a(t)$ and the *instantaneous frequency* (IF) is denoted by $q(t)$. When applied to a speech formant, the center-frequency of the IF $q(t)$ is usually refered as the formant frequency. In fact, the AM-FM signal model has been widely used in speech synthesis.

In contrast to LPC analysis, AM-FM representation of speech retains the nonlinear nature of the resonance, which is evident for the first formant with a large deviation IF compared to its formant frequency. Other approaches using AM-FM representation of speech signals estimate the center-frequency and bandwidth of the formants from their demodulated IF and IA. According to Potamianos [5]–[7], the short-time estimates of the formant frequency and the bandwidth associated with the given formant can be obtained from the IA and IF using squared amplitude as weight given by

$$f_1 = \frac{\int_{t_0}^{t+T} q(t)[a(t)]^2 dt}{\int_{t_0}^{t+T} [a(t)]^2 dt}, \tag{11}$$

$$B_1 = \frac{\int_{t_0}^{t+T} \{(a(t)/2\pi)^2 + (q(t) - f_1)^2 [a(t)]^2\} dt}{\int_{t_0}^{t+T} [a(t)]^2 dt}. \tag{12}$$

A variety of demodulation techniques such as the Hilbert transform and the energy separation algorithm (ESA) can be applied to compute the IA and IF of the formant. However, most conventional demodulation techniques rely on the narrowband assumption for the signal and only work when the IA and the IF do not vary too fast or too greatly in value compared to its center-frequency.

Modeling the first formant as an amplitude-modulation frequency-modulation (AM-FM) signal with a large deviation FM component is more appropriate than using LPC with the narrowband AM assumption. Due to the significant amount of frequency modulation inherent in the first formant, the multirate frequency transformations (MFT) methodology as proposed in prior work [8], [9] can be combined with the ESA to achieve a better demodulation result for signals of this category. Therefore we can generate more precise center-frequency and bandwidth estimates for the large deviation first formant than the commonly used LPC estimates that are based on the narrowband AM assumption or the narrowband AM-FM constrained ESA.

IV. First Formant Extraction via EMD

Prior to demodulating the first formant to obtain the IA and IF estimates, we first need to extract it from the original speech signal by separating out the different formants. This separation of formants is usually achieved by multiband filtering, for example the Garbor filterbanks as proposed by Potamianos [5]. But filtering the wideband first formant may require particularly accurate center-frequency and bandwidth. Instead of multiband filtering, in this paper we propose the use of the empirical mode decomposition (EMD) to extract the first formant, due to the following reasons: 1) EMD does not require precise center-frequency and bandwidth information, which are hardly accessible. 2) EMD allows for more sidelobes of the large deviation first formant since it does not have a fixed passband that will directly cut out the spectral components which locate outside the passband.

Initially proposed in [10], the EMD is an intuitive method that peforms the decomposition process adaptively with an aposteriori defined basis derived from the data itself. It generally involves two constituent procedures, namely the sifting process and decomposition. A function is called an intrinsic mode function (IMF) if the following conditions are satisfied:

1) The number of extremas and the number of zero-crossings equals or differs at most by one; 2) the average of the upper envelope defined by local maximas and the lower envelope defined by local minimas at any point is zero. The IMF reflects the oscillation mode inherent in the signal and can be modeled as AM-FM monocomponent signal.

The sifting process is a systematic way to extract the IMF from the input data $x(t)$ and can be sumarized via

- Initialize $d_0(t) = x(t)$
- Identify the local extremas of $d_n(t)$.
- Interpolate the local maximas and local minimas to form the the upper envelope $u_n(t)$ and lower envelope $v_n(t)$ respetively.
- Determine the local mean of the upper and lower envelopes $m_n(t) = [u_n(t) + v_n(t)]/2$.
- Extract the detail: $d_{n+1}(t) = d_n(t) - m_n(t)$.
- Repeat from step 2 to step 5 until $d_{n+1}(t)$ is an IMF (zero mean or stopping criterion met).

Assume that the speech signal $S(t)$ is composed of oscillatory modes that can be modeled as IMFs. Decomposition is a procedure that keeps repeating the sifting process to decompose the original signal as the sum of IMFs plus the residue, as given by

$$S(t) = \sum_{k=1}^{n} c_k(t) + r_n(t), \tag{13}$$

where $c_k(t)$ denotes the corresponding IMF and $r_n(t)$ denotes the final residue. The decompsition procedure is summarized via

- Initialize $r_0(t) = S(t)$.
- Apply the sifting process on $r_n(t)$ to obtain the corresponding IMF $c_{n+1}(t)$ and the residue $r_{n+1}(t) = r_n(t) - c_{n+1}(t)$.
- Repeat the previous step until the residue $r_{n+1}(t)$ has no more extremas or meets the stopping criterion.

Note that the number of extremas associated with the extracted IMF is gradually reduced iterating from one residue to the next in the decomposition procedure, the EMD hence functions as a filterbank with the subbands changing from high frequency range to low frequency range. However, it is different from any predetermined subband filtering, since the frequency range and resolution associated with each subband is adaptively time-varying. It offers more flexibility than the conventional multiband filtering approach in capturing features that are nonstationary.

To extract the large deviation first formant, we select a number m, ignore the first m consecutive IMFs that oscillate at high-frequency range and sum up the rest of the IMFs and the residue as given by

$$F_1(t) = \sum_{k=m+1}^{n} c_k(t) + r_n(t). \tag{14}$$

By adjusting the parameters associated with the stopping criterion and observing the oscillation modes of the IMFs, we can determine an appropriate number m to obtain the first formant. In this paper, we adopt the *Complete Ensemble Empirical Mode Decomposition with Adaptive Noise* (CEEMDAN) approach [11] to extract the large deviation first formant of the vowels.

V. Wideband First Formant Demodulation Via MFT-ESA

A. Energy Separtion Algorithm

The energy separation algorithm (ESA) as proposed in the work [12] by Maragos *et al*., based on the Teager-Kaiser energy operator $\Psi[x(t)] = \dot{x}^2(t) - x(t)\ddot{x}(t)$, is widely used for monocomponent AM-FM demodulation, for example, to analyze the oscillation of signals with time-varying amplitude and frequency. The IA $a(t)$ and the IF $q(t)$ of an AM-FM signal $x(t)$ can be estimated via the *continuous ESA* (CESA) summarized by

$$\frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \approx |a(t)|, \tag{15}$$

$$\sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \approx q(t), \tag{16}$$

where we assume that the IA $a(t)$ and the IF $q(t)$ do not vary too fast or too greatly in value compared to its center-frequency. The performance of the ESA and other demodulation techniques can be significantly diminished when the IF associated with the signal is in the large deviation regime.

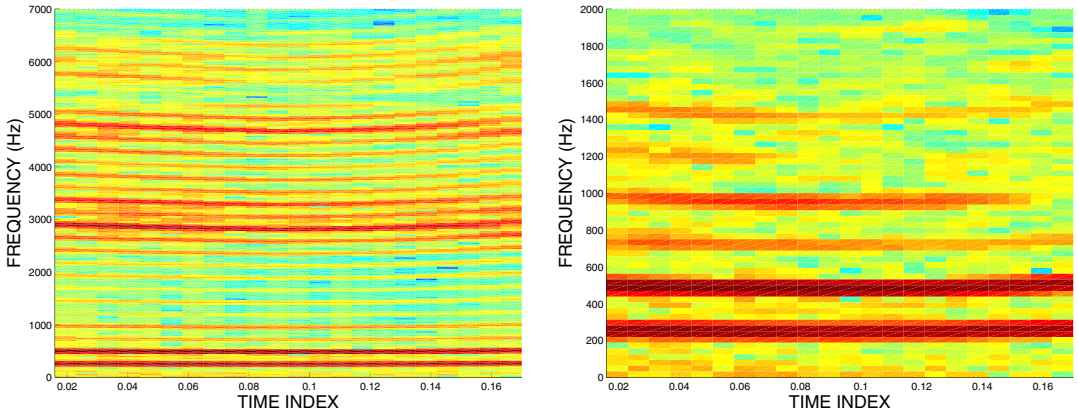
B. Multirate Frequency Transformations

The performance of conventional demodulation techniques such as ESA when directly applied to AM-FM signals with a large deviation FM component is poor due to the narrowband constraint. In recent work of the authors [8], [9], frequency transformations enacted via multirate signal processing were used for wideband FM to narrowband FM conversion to enable a wider range of wideband FM signals, and were also extended to AM-FM signals and two-dimensional images [13]. The goal of the multirate processing module is to compress the bandwidth of the FM signal by a factor R, however this is accompanied by a reduction in the carrier frequency of the FM signal. To compensate, a heterodyning module that translates the FM signal in frequency with an upshift of ω_d is introduced. After the multirate heterodyne combination, the CR/IB and the CR/FD of the transformed signal is constrained to a range, where standard narrowband monocomponent FM demodulation algorithms work optimally. The MFT framework was demonstrated in prior work to provide a solid demodulation result for wideband signals, and will be employed in this paper in combination with the ESA for demodulation of large deviation first formant.

VI. Simulation Results

The short-time fourier transform spectrum of a women's vowel /i:/ [1] and that of the first formant extracted via the EMD

[1]The experimental data is based on the source in the vowel database of Hillenbrand, Getty, Clark & Wheeler (1995)



(a) (b)

Fig. 1. Short-time frequency spectrum of a women's vowel /i:/. (a) Original speech segment. (b) First formant extracted via the EMD.

are illustrated in Fig. 1. As we can observe, the first formant has a large BW/CF and non-negligible sidelobes induced by its significant amount of frequency modulation. Note that the first formant extracted by EMD retains these sidelobes that are usually ignored in the LPC analysis.

The demodulated IF of this large deviation first formant via different approaches are compared in Fig. 2(a). The estimated IF of the LPC approach varies slowly like a straight line, since the LPC only picks up the pole and cuts off the sidelobes induced by frequency modulation. The estimated IF by the ESA varies too sharply in some range due to the large deviation nature of the first formant. The IF estimated via the MFT-ESA combination varies smoothly within the frequency range of the first formant, turning out to be better than the prior estimates. The demodulated IA estimates via the ESA and the MFT-ESA are also compared in Fig. 2(b), from which we can see that the IA estimate by the MFT-ESA is varying slowly and smoother than that of the ESA. From Eq. 7 we know that the envelope for the formant modeled by the LPC approach is a decaying exponential function, which is diffierent from the IA estimates of both the ESA and the MFT-ESA, thus not compared in Fig. 2(b). According to Eq. 12, the square of the IA estimates serve as the weight for computing the bandwidth for the first formant.

The center-frequency and bandwidth estimates of the first formants extracted via the EMD based on three approaches for different female and male vowels in the database are compared in Table I and Table II respectively. The bandwidth estimates of the MFT-ESA turn out to be greater than those of the LPC approach assuming only amplitude modulation, matching the spectrum better as indicated in Fig. 1. The estimated BW/CF of each first formant via the MFT-ESA lies in the large-deviation regime while that of the LPC approach is too small to characterize the frequency modulation inherent in the first formant. Without the MFT, the ESA, however, incurs significant error in the large-deviation regime [5], the bandwidth estimates are too large, leading to erroneous BW/CF estimates as well. Therefore we conclude that the formant estimates by the MFT-ESA are more precise than the LPC estimates or the stand-alone ESA for large deviation first formants.

VII. Conclusion

In this paper, we have presented an approach that applies IF demodulation via the MFT-ESA combination developed by the authors to the first formant of vowels extracted by the EMD and then computed the formant frequency and bandwidth estimates based on the demodulated IF and IA. By taking into account the large deviation nature of the first formant, which usually has a large BW/CF, the formant estimates via the MFT-ESA are demonstrated to be more reasonable than the LPC estimates that assumes narrowband AM formant or the narrowband AM-FM constrained ESA estimates.

References

[1] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2010.

[2] B. Santhanam and P. Maragos, "Energy demodulation of two-component AM-FM signal mixtures," *IEEE Signal Process. Lett.*, vol. 3,no. 11, pp. 294–298,1996.

[3] Chowning, John M., "The Synthesis of Complex Audio Spectra by Means of Frequency Modulation," *J. Audio Eng. Soc*, vol. 21, no. 7, pp. 526–534, 1973.

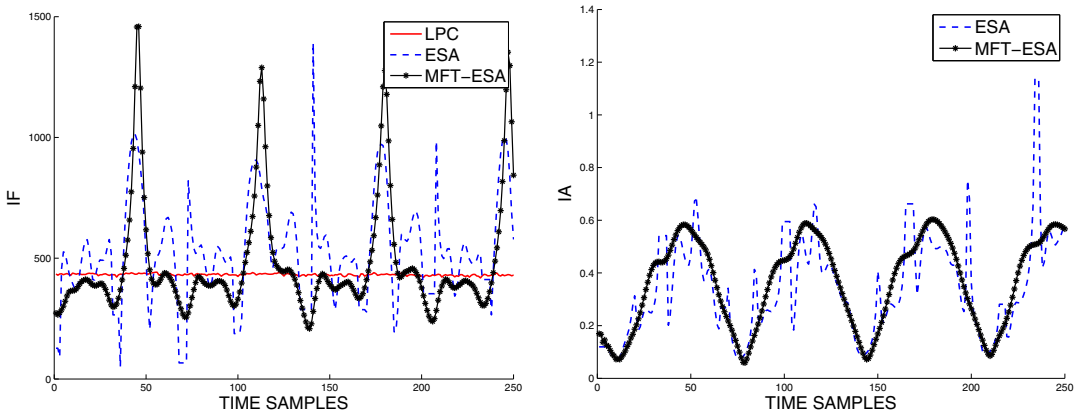
[4] Maragos, P., Kaiser, J. F., and Quatieri, T. F., "On amplitude and frequency demodulation using energy operators"*IEEE Trans. Signal Process.*, vol. 41, no. 4, pp. 1532–1550, 1993.

[5] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3795–3806, 1996.

[6] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an am–fm modulation model," *Speech Communication*, vol. 28, no. 3, pp. 195–209, 1999.

[7] P. Tsiakoulis and A. Potamianos, "On the effect of fundamental frequency on amplitude and frequency modulation patterns in speech resonances," in *Proc. INTERSPEECH*, pp.649–652, 2010.

[8] L. Wenjing and B. Santhanam, "Wideband-FM demodulation for large wideband to narrowband conversion factors via multirate frequency transformations," *Signal Processing and Signal Processing Education Workshop (SP/SPE), 2015 IEEE*, pp. 7–12, 2015.



(a) (b)

Fig. 2. Large deviation first formant estimation of /i:/ extracted by the EMD. (a) IF estimates via the LPC (red line), ESA (blue dashed line), MFT-ESA (black dotted line). Note that the ESA-2 algorithm with second order binomial smoothing, a large multirate factor $R = 32$ and a normalized heterdyning frequency $f_d = 0.2$ is applied in the MFT-ESA for this example. (b) IA estimates via both ESA (blue dashed line) and MFT-ESA (black dotted line).

TABLE I
COMPARISON OF THE FORMANT ESTIMATES FOR DIFFERENT FEMALE FIRST FORMANTS.

	/i:/			/uw/		
	Formant	Bandwidth	BW/CF	Formant	Bandwidth	BW/CF
LPC	433 Hz	71 Hz	16.4%	438 Hz	12 Hz	2.7%
ESA	523 Hz	412 Hz	78.8%	453 Hz	240 Hz	53.0%
MFT-ESA	387 Hz	157 Hz	40.6%	435 Hz	90 Hz	20.7%
	/ei/			/ae/		
	Formant	Bandwidth	BW/CF	Formant	Bandwidth	BW/CF
LPC	458 Hz	29 Hz	6.3%	627 Hz	75 Hz	12.0%
ESA	479 Hz	378 Hz	78.9%	682 Hz	433 Hz	63.5%
MFT-ESA	451 Hz	102 Hz	22.6%	588 Hz	167 Hz	28.4%

TABLE II
COMPARISON OF THE FORMANT ESTIMATES FOR DIFFERENT MALE FIRST FORMANTS.

	/i:/			/uw/		
	Formant	BW	BW/CF	Formant	BW	BW/CF
LPC	290 Hz	33 Hz	11.4%	332 Hz	34 Hz	10.2%
ESA	516 Hz	1012 Hz	196.1%	464 Hz	338 Hz	72.8%
MFT-ESA	280 Hz	115 Hz	41.1%	338 Hz	139 Hz	41.1%
	/ei/			/ae/		
	Formant	BW	BW/CF	Formant	BW	BW/CF
LPC	396 Hz	31 Hz	7.8%	633 Hz	56 Hz	8.8%
ESA	455 Hz	482 Hz	105.9%	753 Hz	667 Hz	88.6%
MFT-ESA	393 Hz	151 Hz	38.4%	622 Hz	191 Hz	30.7%

[9] B. Santhanam, "Generalized energy demodulation for large frequency deviations and wideband signals," *IEEE Signal Process. Lett.*, vol. 11, no. 3, pp. 341–344, 2004.

[10] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[11] M. A. Colominas, G. Schlotthauer, and M.E. Torres, "Improved complete ensemble EMD: A suitable tool for biomedical signal processing," *Biomed. Sig. Process. and Control*, vol. 14, pp. 19–29, 2014.

[12] Maragos, P., Kaiser, J. F., and Quatieri, T. F., "Energy separation in signal modulations with application to speech analysis,"*IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, 1993.

[13] L. Wenjing and B. Santhanam, "Wideband image demodulation via bi-dimensional multirate frequency transformations,"*Journal of Optical Society of America A*, vol. 33, pp. 1668–1678, 2016.