INSTANTANEOUS FREQUENCY AND BANDWIDTH ESTIMATION USING FILTERBANK ARRAYS

Pirros Tsiakoulis[1], Alexandros Potamianos[2], Dimitrios Dimitriadis[3]

[1]Engineering Department, University of Cambridge, CB2 1PZ, U.K.
[2]Department of ECE, Technical University of Crete, Chania 73100, Greece
[3]AT&T Labs-Research, 180 Park Ave, Florham Park, NJ, USA
[1]pt344@cam.ac.uk, [2]potam@telecom.tuc.gr, [3]ddim@research.att.com

ABSTRACT

Accurate estimation of the instantaneous frequency of speech resonances is a hard problem mainly due to phase discontinuities in the speech signal associated with excitation instants. We review a variety of approaches for enhanced frequency and bandwidth estimation in the time-domain and propose a new cognitively motivated approach using *filterbank arrays*. We show that by filtering speech resonances using filters of different center frequency, bandwidth and shape, the ambiguity in instantaneous frequency estimation associated with amplitude envelope minima and phase discontinuities can be significantly reduced. The novel estimators are shown to perform well on synthetic speech signals with frequency and bandwidth micro-modulations (i.e., modulations within a pitch period), as well as on real speech signals. Filterbank arrays, when applied to frequency and bandwidth modulation index estimation, are shown to reduce the estimation error variance by 85% and 70% respectively.

Index Terms: speech analysis, time-frequency analysis, filterbank arrays, instantaneous frequency, micro-modulations

1. INTRODUCTION

Speech signals are non-stationary due to the rapid movement of the articulators and the complex dynamics of airflow in the vocal tract. Even for the so-called steady-state phonation, various second-order phenomena are observed in speech resonances that cannot be predicted by the linear source-filter model as derived from the simplified 1-D Navier-Stokes equations [1, 2, 3]. Deviations from the model are often manifested as micro-modulations in the amplitude envelope and instantaneous frequency of speech resonances often within a single pitch period. Some of these phenomena are well-known and have been documented over the past decades, e.g., secondary excitations at glottal openings, bandwidth modulation between the open and closed phase [4]; others are still under investigation. Some of the causes of such micro-modulations have been identified as the excitation of secondary resonator modes at the vocal folds, the non-linear dynamics of the airflow, as well as the non-linear source and vocal tract interaction (especially during the open phase).

A set of tools has been devised by the speech analysis research community to explore and analyze such phenomena. The need for such tools goes beyond the scientific curiosity of better understanding speech production; they are also motivated by speech applications such as speech synthesis, e.g., for pitch synchronous analysis, it is important to accurately identify primary excitation instants (and if possible, secondary excitations) and speech/speaker recognition. The main approaches for the estimation of excitation instants and associated micro-modulation in speech resonances are: (i) inverse filtering, where the effect of the vocal tract is assumed to be that of a linear filter that is estimated and removed to obtain the excitation signal, (ii) instantaneous phase or group delay processing [5] where the excitation instants are identified via the phase discontinuities they create in the instantaneous frequency estimates and (iii) analysis by synthesis [6]. Once the excitation instants have been identified, micro-modulations (amplitude, frequency, and bandwidth modulations within a pitch period) can be measured. These estimates are associated with specific frequency ranges (often focusing just on the speech resonances) and thus a filterbank is also employed here. Instantaneous amplitude and frequency signals can be used for formant frequency and bandwidth estimation [7, 8], or for directly measuring modulations in speech resonances [9, 6, 10].

An important factor that hinders accurate instantaneous frequency estimation in a speech resonance is the phase discontinuity at excitation instants. Such discontinuities appear as single or double spikes in the instantaneous frequency estimated after phase unwrapping and differentiation. The discontinuities are additionally being band-passed in order to isolate a speech resonance, spilling their effect in time and increasing the variance of (instantaneous and short-time) frequency and bandwidth estimates. We propose to use a set of filters that are varied in center frequency, bandwidth, and shape in order to reduce the effect of time-frequency estimation uncertainty and significantly improve the statistics of estimators. The process is motivated by auditory cognition, where auditory filters are averaged across frequency and adapted depending on the audio conditions [11].

This work builds on our recent studies on measuring amplitude and frequency modulations in speech resonant signals [6, 9]. The main contribution of this paper is a novel technique for processing instantaneous amplitude and frequency signals using filterbank arrays in order to reduce the effect of phase discontinuities. Here, we focus on frequency and bandwidth modulation indexes that quantify the micro-modulation depth. We show that applying the filterbank array technique to the frequency and bandwidth estimation results in unbiased estimators with reduced error variance. Moreover, the proposed method can be used for a variety of problems that require accurate mean instantaneous frequency estimates, such as formant and pitch tracking [7] or harmonic analysis [12]. This work is also motivated by recent research on time domain processing of speech. More specifically, instantaneous frequency and bandwidth modulation features have been successfully applied for speech recognition [13, 10, 14, 15], speaker identification [16, 17, 18]. This reveals that modulation patterns carry useful information for both speaker/speech recognition, and can also be robust to noise

[19, 10, 15, 20]. Our goal is to enhance the AM–FM analysis tools in order to improve the use of modulations in speech applications.

The remainder of the paper is organized as follows. In Section 2, the frequency and bandwidth estimators based on the AM–FM model are described. The *Frequency Modulation Index* (FMI) and the *Bandwidth Modulation Index* (BMI) are also introduced. Section 3 describes the proposed filterbank array technique. Section 4 reports on the performance of FMI and BMI metrics using the baseline estimators. Section 5 then compares the performance of FMI and BMI using the novel filterbank array estimators. Finally, Section 6 concludes this work.

2. FREQUENCY AND BANDWIDTH ESTIMATORS

This work is motivated by the AM–FM model [21] that describes a speech resonance as a signal $r(t)$ with a combined amplitude modulation (AM) and frequency modulation (FM) structure

$$r(t) = a(t)\cos(2\pi[f_c t + \int_0^t q(\tau)d\tau] + \theta) \qquad (1)$$

where f_c is the "center value" frequency, $q(t)$ is the frequency modulating signal, and $a(t)$ is the time-varying amplitude. The instantaneous frequency signal is defined as $f(t) = f_c + q(t)$. The speech signal $s(t)$ is then modeled as the sum of N AM–FM signals. The AM–FM signal can be demodulated into $a(t)$, $f(t)$ using the energy separation algorithm or the Hilbert transform demodulation algorithm. The speech resonant signals are extracted via filtering. Usually, the Gabor filter is preferred because it is maximally smooth and optimally concentrated both in time and frequency domain. The resulting band passed signals are further demodulated into instantaneous frequency and amplitude components. The whole process is known as *Multiband Demodulation Analysis* (MDA) [22].

Short-time analysis is performed on the instantaneous amplitude and frequency signals in order to extract three main estimates: *amplitude*, *frequency* and *bandwidth*. The time-frequency distribution of amplitude, frequency and bandwidth estimates are often used as features for speech recognition and speaker identification, as well as in other speech processing applications. Amplitude is an energy measure which can be derived directly via short-time integration of the squared instantaneous amplitude $A = \int_{t_0}^{t_0+T} a^2(t)dt$, where t_0 is the starting time index, and T is the integrating period. For frequency estimation, amplitude weighting is performed in order to eliminate the effect of low energy singularities, as follows:

$$F = \frac{\int_{t_0}^{t_0+T} f(t)[a(t)]^2 dt}{\int_{t_0}^{t_0+T} [a(t)]^2 dt}. \qquad (2)$$

F is an estimate of the mean frequency in the analysis frame and a good estimator of the formant frequency. Moreover, it is equivalent to the first spectral moment (frequency domain estimate). The bandwidth of an AM–FM signal is usually defined using two components to account for both instantaneous frequency and amplitude envelope variations [23, p. 534], as follows:

$$[B]^2 = \frac{\int_{t_0}^{t_0+T} \left[(\dot{a}(t)/2\pi)^2 + (f(t) - F)^2 a^2(t)\right] dt}{\int_{t_0}^{t_0+T} a^2(t)dt}. \qquad (3)$$

The amplitude component is the term $(\dot{a}(t)/2\pi)^2$, and can be thought of as the AM contribution to the bandwidth

$$\left[B^{AM}\right]^2 = \frac{\int_{t_0}^{t_0+T} (\dot{a}(t)/2\pi)^2 dt}{\int_{t_0}^{t_0+T} a^2(t)dt} \qquad (4)$$

It describes the decay rate of the amplitude envelope which is closely related to the formant bandwidth [24].

2.1. Estimating frequency and bandwidth modulations

In this work we focus on micro-modulations of the instantaneous frequency and bandwidth signals within a pitch period. More specifically, the *Frequency Modulation Index* (FMI) and the *Bandwidth Modulation Index* (BMI) are introduced for instantaneous frequency and bandwidth respectively, as metrics for their modulation depth, i.e. the degree of divergence from their average level.

FMI is an estimate of the degree of divergence of the instantaneous frequency $f(t)$ from its mean value f_c (roughly corresponding to the formant frequency). To quantify this, we define two regions in the pitch period, the primary region roughly corresponding to the closed phase and the secondary one, roughly corresponding to the open phase of phonation. Identifying the primary and secondary regions is beyond the scope of this paper – for a method that uses an analysis-by-synthesis loop see [6]. Eq. (2) is used to estimate F_p and F_s, i.e., the weighted mean instantaneous frequency estimate for the primary and secondary region. FMI is hence defined as FMI $= |F_p - F_s|/F$. The presence of a spike in the instantaneous frequency signal in one of the two regions results in unbalanced estimation errors between F_p and F_s. This does not only affect the estimation, it also increases the variance of the estimator (see Sec. 4).

BMI is defined in a similar way. The same partition of the pitch period is used and the bandwidth is estimated separately for the primary and secondary regions. The relative difference between the two estimates is used as the modulation index. In order to investigate the effect of the proposed techique on instantaneous amplitude estimate, we focus only on the AM contribution to the bandwidth and define BMI$^{AM} = |B_p^{AM} - B_s^{AM}|/B^{AM}$ using the B^{AM} definition in Eq. (4) to estimate bandwidth separately for the primary and secondary region of a pitch period. The estimate of the instantaneous amplitude signal is usually more stable and does not exhibit spike-like singularities, however, subtle differences in its dynamics can have significant effect on bandwidth estimation.

3. FILTERBANK ARRAYS

A variety of techniques has been used to eliminate or filter-out the spikes in the $f(t)$ signal associated with phase discontinuities and $a(t)$ minima. These include:

- Post-processing on the $f(t)$ signal including peak-to-peak pruning, median filtering, smoothing [19].
- Alternative estimates of the mean frequency such as, pseudo-instantaneous frequency [10], the model-based approach in [25] or heavier amplitude envelope weighting in (2), (3) [19].

Here instead, we propose robust estimation of the instantaneous frequency and amplitude signals using filterbank arrays. The term *robust* refers to robustness to phase discontinuities and energy singularities, and should not be confused with noise robustness. The latter will be investigated in future work.

We start from a baseline MDA scheme, where a resonance signal $r(t)$ is isolated via filtering and then demodulated to $a(t)$ and $f(t)$ signals. Next, for each filter with impulse response $h_n(t)$ and center frequency f_l, $l = 1..L$, we create an array of $2K + 1$ filters by varying its center frequency in the vicinity of f_l as follows: $f_{l,k} = f_l + k\Delta f$, with $k = -K... - 1, 0, 1...K$ and Δf being the distance (in frequency) between adjacent filters. Alternatively, one could also modify the bandwidths and shapes of the filters. We have experimentally observed that varying the center frequency of the filter provides the best performance (reduction in estimation variance). Then we proceed to filter and demodulate the resonance signal via each of the filters in the array and obtain a time-frequency (TF) distribution of amplitude envelopes $a(t,k)$ and instantaneous frequency $f(t,k)$ signals for each resonance. The TF distributions are then combined to obtain robust estimates of the short-time resonance frequency and bandwidth either (i) by simply averaging their outputs in time before estimating F, i.e.,

$$F_A = \frac{\sum_k \left(\int_{t_0}^{t_0+T} f(t,k)[a(t,k)]^2 dt \right)}{\sum_k \left(\int_{t_0}^{t_0+T} [a(t,k)]^2 dt \right)} \tag{5}$$

(the bandwidth estimator B_A is defined in a similar fashion) or (ii) by using the variance of the instantaneous frequency estimates $v_f(t) = E_k\{(f(t,k) - E_k\{f(t,k)\})^2\}$ as an additional weighting in (2), (3), leading to the following generalized estimator:

$$F_{n,m} = \frac{\int_{t_0}^{t_0+T} f(t)[a(t)]^n [v_f(t)]^{-m} dt}{\int_{t_0}^{t_0+T} [a(t)]^n [v_f(t)]^{-m} dt} \tag{6}$$

where n, m are non-negative integers. For $n = 2, m = 0$ we get the estimator in (2), while for $n = 0, m = 1$ we get the inverse variance weighting estimator. Finally, for $n = 2, m = 1$ we get the inverse variance amplitude square weighted estimator.

We expect the instantaneous frequency estimates averaged over the filterbank array to be robust to phase discontinuities at excitation instants. The argument is that phase discontinuities (and the corresponding area under the pulses in $f(t)$ estimates) at primary excitation instants roughly correspond to the difference between the phase of the formant frequency and the phase of the highest energy harmonic in the speech resonance (for a theoretical explanation see the difference between unweighted and weighted estimates of short-time frequency in [7]). An experimental demonstration of this phenomenon can be found in Section 5.2 of [22]. Moving around the center frequency of a filter with a shaped pass-band (e.g. Gabor) in the vicinity of the formant leads to a modified formant frequency estimate, while the highest energy harmonic usually remains the same. Thus, the amount of phase discontinuity and the shape/direction of spikes is different in each $f(t,k)$ (see Fig. 1). As a result, averaging $f(t,k)$ over an array of filters centered in the formant vicinity can reduce the effect of phase discontinuities and lead to more robust instantaneous and short-time frequency estimates.

4. FMI AND BMI ERROR ANALYSIS

Next, we investigate the frequency FMI and bandwidth BMI^{AM} micro-modulation estimation errors as a function of the following parameters: 1) amount of frequency or bandwidth micro-modulation, 2) fundamental frequency (F0), 3) formant proximity, 4) center frequency of the Gabor band-pass filter f_l and 5) the bandwidth parameter α of the Gabor band-pass filter. The bias and standard deviation of the FMI and BMI^{AM} estimators are computed on synthetic speech signals that are generated using a cascade formant synthesizer. Frequency and bandwidth modulations are added to the speech resonances by appropriately modulating the parameters of the resonators in the first half and second half of the pitch period using step functions. Gabor filtering followed by Hilbert demodulation is used to estimate the $a(t)$, $f(t)$ signals for each resonance. The parameters used are: $F0 = 120$ Hz, $(F1, F2, F3) = (500, 1500, 2500)$ Hz, $(BW1, BW2, BW3) = (70, 100, 100)$ Hz, FMI = 0.05, BMI^{AM} = 0.5. Gabor filters with center frequency equal to the formant and bandwidth parameter $\alpha = 1000$ are used.

Table 1. Mean and standard deviation of error of FMI/BMI estimators for various parameter specifications.

	Mean Error		STD of Estimator	
Param.–Value	FMI ($\times 10^{-4}$)	BMI^{AM} ($\times 10^{-2}$)	FMI ($\times 10^{-4}$)	BMI^{AM} ($\times 10^{-2}$)
FMI = 0.02	-6		73	
FMI = 0.05	-7		72	
FMI = 0.1	**-23**		73	
BMI = 0.2		7		18
BMI = 0.5		5		15
BMI = 1		3		9
$F0 = 120$ Hz	-13	5	74	14
$F0 = 180$ Hz	9	8	263	55
$F0 = 240$ Hz	**-64**	**33**	**472**	**75**
$F2 = 900$ Hz	**141**	**75**	**877**	**76**
$F2 = 1200$ Hz	42	-6	117	21
$F2 = 1500$ Hz	-9	6	71	14
$f_l = 1400$ Hz	-18	0	96	16
$f_l = 1500$ Hz	-9	5	72	16
$f_l = 1700$ Hz	20	-6	79	16
$\alpha = 800$	**-36**	**11**	**210**	**35**
$\alpha = 1000$	-9	5	72	14
$\alpha = 1200$	-4	1	21	4

Table 1 summarizes the performance of the estimators for each of the parameters. Specifically, we compute the mean error (ME) and the square root of the mean square error of FMI/BMI. The former measures the bias while the latter measures the standard deviation (STD) of each estimator. The ME and STD are reported for three characteristic values of each parameters under investigation. Errors were averaged over 1000 experiments (the synthetic signal parameters for each experiment are selected to vary randomly within $\pm 10\%$ of the values specified above). The baseline Eqs. (2),(4) were used to estimate F and B. Estimation was performed in the middle third of the primary region for F_p, B_p^{AM} and the middle third of the secondary region for F_s, B_s^{AM} using a single pitch period. When using the whole primary and secondary regions the error STD is approximately three times higher, but the relative performance of the estimators as a function of the parameters shown here is similar. Primary and secondary regions were defined as the first and second half of the duration between two adjacent excitation pulses.

Results presented in the table show that:

- High FMI values can be somewhat underestimated due to the effect of the band-pass filter (see for example ME for FMI = 0.1 and $\alpha = 800$); this is less of an issue for BMI^{AM}.
- The variance of both F and B estimation increases as a function of $F0$ since time-frequency resolution within the pitch period is reduced with fewer samples to estimate from.
- When two formants approach closer than 300-500 Hz (e.g., see $F2 = 900$ Hz closing on $F1 = 500$ Hz) cross-modulations appear in the F and B^{AM} estimates from neighboring formants causing large biases and increased estimation variance.
- Moving the band-pass filter center frequency in the proximity of the formant frequency does not significantly bias the estimates or increase estimation variance.

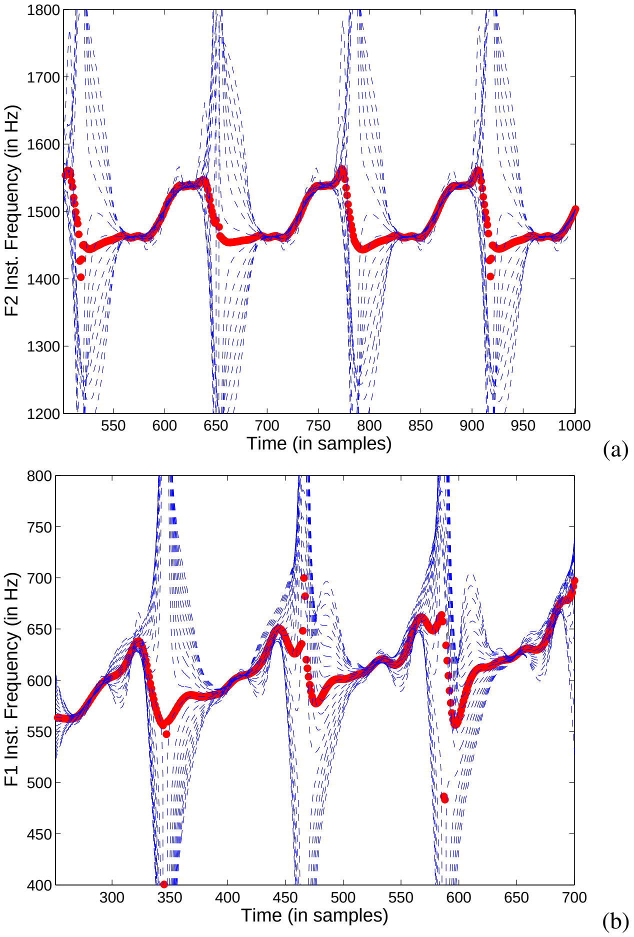

Fig. 1. *Average (red-dotted) $f(t)$ estimates using an array of eight filters vs. various $f(t)$ estimates using a single filter (filter locations ranging from 80% to 120% of formant value) for (a) synthetic resonance signal with 5% step-wise FM modulation and (b) real speech resonance F1 of phone /aa/ instance.*

- Smaller band-pass filter bandwidths cause undershoot for FMI (e.g., $\alpha = 800$), while BMI^{AM} is overestimated.

These trends have been verified on a large range of values for the parameters under investigation.

5. PERFORMANCE COMPARISON OF FILTERBANK ARRAY ESTIMATORS

Next, we investigate the performance of the average F_A and inverse variance weighted frequency estimators $F_{n,m}$ (using filterbank arrays), as well as their instantaneous bandwidth counterpart B_A. Specifically, we compare the performance of F_A, B_A with that of F, B on synthetic and real speech resonances (see Fig. 1), as well as compute the bias and standard deviation of FMI, BMI using filterbank arrays (see Table 2).

Fig. 1(a), (b), shows various instantaneous frequency $f(t)$ signals (blue-dotted) estimated via multi band demodulation (Gabor filtering, Hilbert demodulation) from 20 different filters that are centered within ±20% from the formant frequency ($\alpha = 1000$). The average $<\!f(t)\!>_k$ estimate of 8 filters (used in F_A) is also shown (red-dotted). Fig. 1(a) shows results for a synthetic signal $F0 = 120$ Hz, $(F1, F2, F3) = (500, 1500, 2500)$ Hz, with frequency micro-modulations of ±50 Hz in F2 (step function within a pitch period). It is clear from the figure that the $<\!f(t)\!>_k$ estimate accurately tracks the FM in F2, and that averaging effectively filters out most of the erratic spikes associated with phase discontinuities at excitation instants. Similarly, Fig. 1(b) shows the results for the first resonance of an instance of the phoneme /aa/. Again, averaging significantly improves estimation; the frequency modulation in the first formant between the open and closed phonation phase is now clear. Overall, moving the Gabor filter in the proximity of the formant frequency does not significantly bias the F_A estimate or increase its estimation error variance (see also previous section), while the actual estimates $f(t)$ do change. Averaging $f(t)$ can lead to significant reduction of estimation error (and variance), as shown graphically in Fig. 1.

Table 2. Mean and standard deviation of error using baseline (F, B^{AM}) and filterbank array (F_A, B_A^{AM}, $F_{n,m}$) estimators.

Estimator	Mean Error		STD of Estimator	
	FMI ($\times 10^{-4}$)	BMI^{AM} ($\times 10^{-2}$)	FMI ($\times 10^{-4}$)	BMI^{AM} ($\times 10^{-2}$)
F, B^{AM}	-15	7	74	21
F_A, B_A^{AM}	**-3**	-6	**10**	**6**
$F_{0,1}$	-14	–	87	–
$F_{2,1}$	-9	–	73	–

Table 2 summarizes the bias and estimator standard deviation when computing FMI and BMI^{AM} in synthetic speech resonances using estimators F, F_A, $F_{n,m}$ and B, B_A. The synthetic signals used have the same parameters as in the previous section. Results are averaged over 1000 examples (variants obtained by varying the value of the parameters within ±10 of these values) and computed from the middle third of the primary and secondary regions of a pitch period. Ten filters spaced within ±20% of the formant frequency are used to estimate F_A, B_A, $F_{n,m}$. The reduction in estimation bias is significant for FMI when using filterbank arrays and estimator F_A. The reduction in the standard deviation of the estimator is even more impressive; seven times for FMI and three to four times for BMI^{AM}. The inverse variance estimates $F_{0,1}$, $F_{2,1}$ do not achieve a large reduction in FMI estimation bias or STD. Overall, averaging the instantaneous frequency and amplitude over filterbank arrays can produce more robust F, B, FMI and BMI estimates, and significantly reduce estimator variance.

6. CONCLUSIONS

We have shown that averaging instantaneous amplitude and frequency signals across filterbank arrays significantly improves the robustness to phase discontinuities of the frequency and bandwidth estimators, mainly by reducing estimator variance. We have demonstrated this both on synthetic and real signals for measuring micro-modulation in speech resonances. Filterbank arrays significantly improve the frequency and bandwidth modulation index estimation; the estimator error variance is reduced by 85% and 70% respectively. Future work should include a more detailed theoretical analysis of filterbank arrays and their connection with human auditory cognition, fast and efficient implementations of array filtering, combination of F_A, B_A estimates with various weighting-based, model-based and post-filtering schemes to further improve estimation accuracy, as well as deriving enhanced micro-modulation features for speech applications.

7. REFERENCES

[1] T. V. Ananthapadmanabha and G. Fant, “Calculation of true glottal flow and its components,” *Speech Communication*, vol. 1, no. 3/4, pp. 167–184, 1982.

[2] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, “Modeling of the glottal flow derivative waveform with application to speaker identification,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, September 1999.

[3] I. R. Titze, “Nonlinear source–filter coupling in phonation: Theory,” *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 2733–2749, 2008.

[4] K. N. Stevens, *Acoustic Phonetics*. MIT press, 2000, vol. 30.

[5] M. Brookes, P. A. Naylor, and J. Gudnason, “A quantitative assessment of group delay methods for identifying glottal closures in voiced speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 456–466, March 2006.

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

[7] 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, pp. 3795–3806, June 1996.

[8] B. Yegnanarayana and K. S. R. Murty, “Event-based instantaneous fundamental frequency estimation from speech signals,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[9] P. Tsiakoulis and A. Potamianos, “Statistical analysis of amplitude modulation in speech signals using an AM–FM model,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3981–3984.

[10] Y. Kubo, S. Okawa, A. Kurematsu, and K. Shirai, “Temporal AM–FM combination for robust speech recognition,” *Speech Communication*, vol. 53, no. 5, pp. 716–725, 2011.

[11] S. Shamma, “On the role of space and time in auditory processing,” *Trends in Cognitive Sciences*, vol. 5, no. 8, pp. 340–348, 2001.

[12] Y. Pantazis, O. Rosec, and Y. Stylianou, “Adaptive AM–FM signal decomposition with application to speech analysis,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 290–300, February 2011.

[13] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, “Short-time instantaneous frequency and bandwidth features for speech recognition,” in *Proceedings of the IEEE Workshop on Automatic Speech Recognition Understanding*, 2009.

[14] H. Yin, V. Hohmann, and C. Nadeu, “Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency,” *Speech Communication*, vol. 53, no. 5, pp. 707–715, 2011.

[15] V. Tiwari and J. Singhai, “AM–FM features and their application to noise robust speech recognition: A review,” *The IUP Journal of Telecommunications*, vol. 2, no. 1, pp. 7–19, 2010.

[16] M. Grimaldi and F. Cummins, “Speaker identification using instantaneous frequencies,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1097–1111, August 2008.

[17] T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, “Computationally efficient frame-averaged FM feature extraction for speaker recognition,” *Electronics Letters*, vol. 45, no. 6, pp. 335–337, 12 2009.

[18] L. Hou, X. Hu, and J. Xie, “Application of formant instantaneous characteristics to speech recognition and speaker identification,” *Journal of Shanghai University*, vol. 15, no. 2, pp. 123–127, 2011.

[19] D. Dimitriadis, P. Maragos, and A. Potamianos, “Robust AM–FM features for speech recognition,” *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 621–624, September 2005.

[20] M. S. Deshpande and R. S. Holambe, “Speaker identification based on robust AM–FM features,” in *Proceedings of the 2nd International Conference on Emerging Trends in Engineering and Technology (ICETET)*, December 2009, pp. 880–884.

[21] P. Maragos, T. F. Quatieri, and J. F. Kaiser, “Speech nonlinearities, modulations and energy operators,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, May 1991, pp. 421–424.

[22] A. Potamianos, “Speech processing applications using an AM–FM modulation model,” Ph.D. dissertation, Harvard University, Cambridge, MA, April 1995.

[23] B. Boashash, “Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals,” *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520–538, April 1992.

[24] L. Cohen and C. Lee, “Instantaneous Bandwidth,” in *Time-Frequency Signal Analysis–Methods and Applications*, B. Boashash, Ed. Melbourne: Longman Cheshire, 1992, pp. 98–117.

[25] R. Kumaresan and A. Rao, “Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications,” *The Journal of the Acoustical Society of America*, vol. 105, pp. 1912–1924, 1999.