AN IMPROVED CHIRP GROUP DELAY BASED ALGORITHM FOR ESTIMATING THE VOCAL TRACT RESPONSE

M.K. Jayesh and C.S. Ramalingam

Department of Electrical Engineering
IIT Madras, Chennai–600 036, INDIA
ee11d040@ee.iitm.ac.in, csr@ee.iitm.ac.in

ABSTRACT

We propose a method for vocal tract estimation that is better than Bozkurt's chirp group delay method [1] and its zero-phase variant [2]. The chirp group delay method works only for voiced speech, is critically dependent on finding the glottal closure instants (GCI), deteriorates in performance when more than two pitch cycles are included for analysis, and does not work for unvoiced speech. The zero-phase variant eliminates these drawbacks but works poorly for nasal sounds. In our proposed method all outside-unit-circle zeros are reflected inside before computing the chirp group delay. The advantages are: (a) GCI knowledge not required, (b) the vocal tract estimate is far less sensitive to the location and duration of the analysis window, (c) works for unvoiced sounds, and (d) captures the spectral valleys well for nasals, which in turn leads to better recognition accuracy.

Index Terms— vocal tract estimation, group delay

1. INTRODUCTION

In conventional speech processing methods, feature vectors are extracted from the Fourier transform magnitude, with the phase spectrum being completely neglected. However, the phase of the Fourier transform is increasingly viewed as being not only containing important information but also robust [3]. The most commonly used phase-based feature is the group delay, which is defined as the derivative of continuous Fourier transform phase of the signal. It is well-known that the group delay plot of a minimum phase signal looks similar to its squared magnitude response around resonant frequencies, and hence called as the "group delay spectrum" (GDS) [3]. In this paper we propose a group delay based method for vocal tract estimation.

Vocal tract estimation is a well studied problem [4, and references therein]. The major challenge associated with any group delay based estimation method is the presence of zeros on or very close to the unit circle. A simplified model for speech is to consider it as the output of an all-pole filter excited by either a periodic signal or random noise. If we denote the excitation by $u[n]$, in the case of both periodic excitation and random noise, its z-transform $U(z)$ has zeros either on the unit circle or close to it. These zeros produce high amplitude spikes in the GDS, thereby masking any information about the vocal tract. It is for reducing these spikes that Bozkurt [2] suggested evaluating the group delay around a circle of radius 1.12 (an empirically derived value; see [1] for more details), the so-called chirp GDS, computed after removing all the roots that fall outside the unit circle.

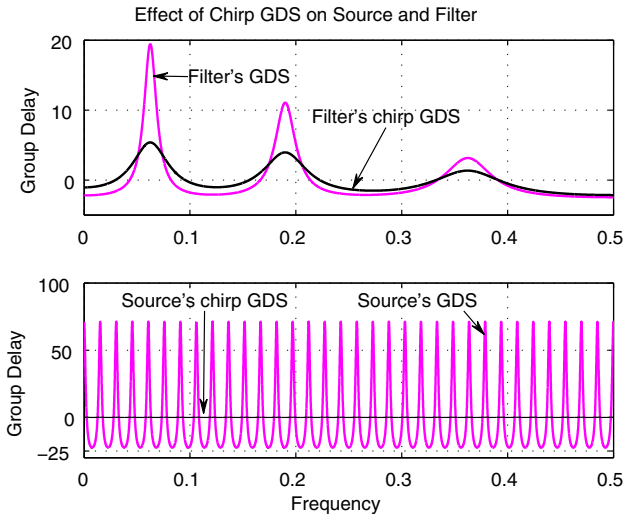

Fig. 1. GDS and chirp GDS of vocal tract filter (top) and source signal (bottom). The chirp GDS of the source has a significantly lower amplitude than that of the filter.

We wish to highlight the following advantage of chirp GDS: for the filter, the amplitude range of chirp GDS is only slightly smaller than that of GDS, whereas for the source the reduction is drastic. In Fig. 1 the GDS and chirp GDS of a 30 ms segment of an impulse train (with pitch of 120 Hz) are shown (note: we have chosen a radius of 1.01 for the GDS because of the presence of unit circle zeros). Also shown in that figure are the GDS and chirp GDS of an all-pole filter. The amplitude of the chirp GDS of the source is so greatly reduced that it appears to be nearly zero in Fig. 1 (the fluctuations are in the range $\approx \pm 0.12$); in contrast, the filter's chirp GDS is

only slightly smaller when compared with its corresponding GDS. We exploit this property in our method to suppress the presence of the source.

2. PROPOSED ALGORITHM

It is well known that for an LTI system that the GDS is additive. It follows that the chirp GDS is also additive, since the only change is the radius about which the delay spectrum is computed. Preserving this additivity property is worthwhile, as we shall see.

In our approach we propose that any zero of the z-transform of the analysis segment that falls outside the unit circle be reflected inside. This, coupled with the chirp GDS, gives a good estimate of the filter spectrum while suppressing the oscillations due to the source.

In the case of unvoiced signals, the excitation signal is random noise. The noise zeros tend to fall randomly on both sides of the unit circle and even on it. For such a source signal, it was observed that the chirp group delay of the corresponding minimum phase signal produces a chirp GDS that has no significant fluctuations. Hence, when this is added to the chirp GDS of the filter, the latter's shape remains practically the same. An example of the chirp GDS of noise is shown in Fig. 2. The noise segments were 30 ms in duration and a Blackman window was used for analysis.

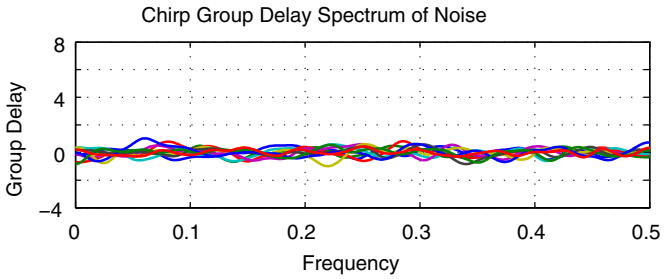

Fig. 2. Chirp GDS of a random noise segment for 10 realizations. It is seen that the proposed method gives a reasonably flat spectrum for noise.

The resemblance between GDS and the squared magnitude spectrum is true only for minimum phase signals [3]. In the mixed-phase example shown in Fig. 3, the GDS shown on the top right plot has a *peak* around $f = 0.33$ rather than a valley. However, if the outside root is reflected, the valley is restored (bottom left). Since both the log magnitude and the phase are additive for product factors, it is customary to compare GDS with log magnitude. This comparison is show in the bottom right plot of Fig. 3; also shown in that figure is the estimate obtained using the zero-phase chirp GDS method, which is not able to capture the spectral valley. In these figures we have normalized the plots for ease of comparison.

We also observed that the additivity property and suppression of the source fluctuations are approximately preserved

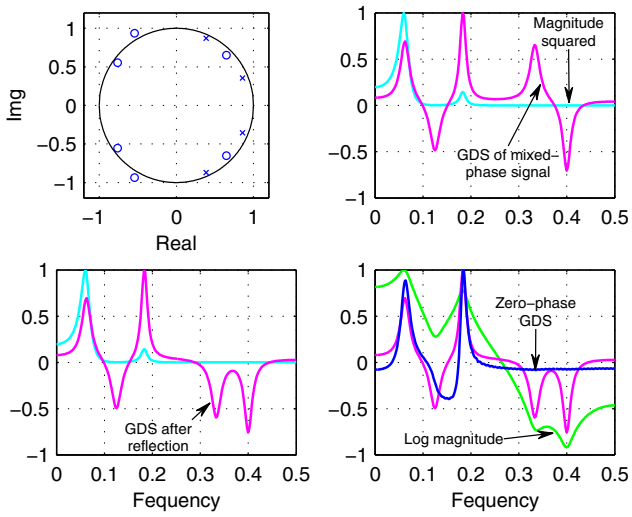

Fig. 3. The GDS of a mixed-phase signal has a peak at $f = 0.33$ instead of a valley. The valley is restored once the zero is reflected (bottom left). GDS and log magnitude spectra are similar (bottom right). Note that the zero-phase chirp GDS is not able to capture spectral valleys.

when we window the output speech. The Blackman window was found to give good results, which confirms Bozkurt's observation in [1, 5]. Thus, our algorithm is as follows:

1. Take a segment of windowed speech $y[n] = w[n] \cdot x[n]$, where $w[n]$ is the analysis window (typically a Blackman window of 20–30 ms in duration).
2. Compute its z-transform $Y(z)$.
3. Find the zeros of $Y(z)$, i.e., $z_i = r_i e^{j\theta_i}$.
4. Reflect outside-unit-circle zeros inside, i.e., if $r_i > 1$, then replace $r_i e^{j\theta_i}$ by $\frac{1}{r_i} e^{j\theta_i}$.
5. Calculate the chirp GDS of each zero using the easily derivable formula $\dfrac{r_i^2 - \rho_0 r_i \cos(\omega - \theta_i)}{\rho_0^2 + r_i^2 - 2\rho_0 r_i \cos(\omega - \theta_i)}$, where ρ_0 is the radius about which the group delay is evaluated (we have used $\rho_0 = 1.12$ in our work).
6. Add the chirp GDS of each zero to get the final overall chirp GDS.

3. RESULTS

We used both synthetic and natural speech signals for testing our proposed method. For synthetic voiced speech we used an all-pole filter with formants located at 300 Hz, 1500 Hz, and 2400 Hz having bandwidths 45, 60, and 180 Hz respectively, and excited it with an impulse train whose f_0 was 120 Hz. For the unvoiced case, an all-pole filter with the same formant locations was used but the bandwidths were changed to 90 Hz, 30 Hz, and 100 Hz (these bandwidth changes were

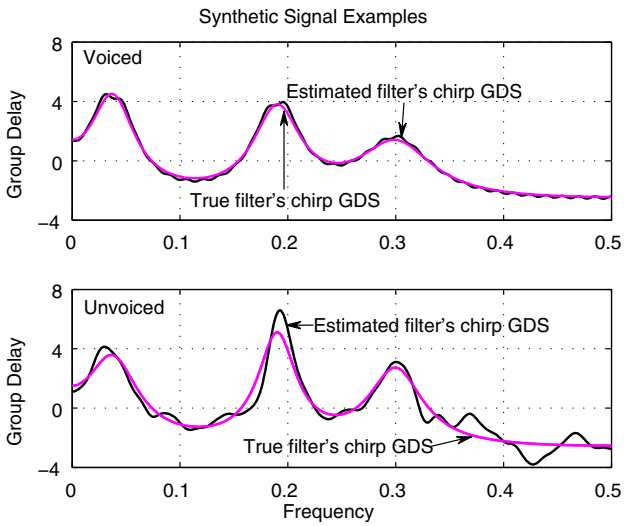

Fig. 4. Chirp GDS of true vocal tract filter and its estimate for a voiced segment (top) and an unvoiced segment (bottom). It is seen that the proposed method gives a good estimate for both voiced and unvoiced speech segments.

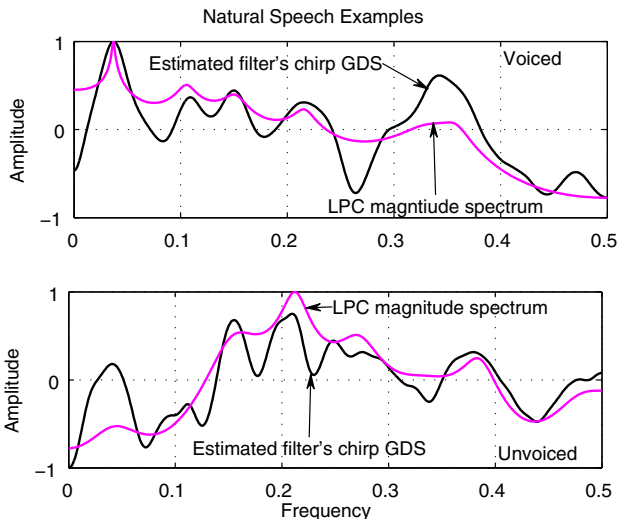

Fig. 5. Chirp GDS of estimated vocal tract filter for natural speech: voiced segment (top) and an unvoiced segment (bottom). For comparison, the LPC magnitude spectra are also shown (the curves have been scaled to facilitate comparison).

made just to introduce variation and have no significance otherwise). The sampling frequency was 8 kHz.

The results of estimating the vocal tract filter using our method of reflecting the zeros and computing the chirp GDS is shown in Fig. 4 for the synthetic speech examples. The top and bottom panels show the results for voiced and unvoiced speech segments. It is seen that the proposed method gives a good estimate of the vocal tract filter in both the voiced and unvoiced cases.

Natural speech examples are shown in Fig. 5. The top panel shows the chirp GDS obtained by analyzing a 30 ms segment of natural speech; the segment corresponds to the phoneme `/ae/` occurring in the word "had", taken from the TIMIT sentence, "She had your dark suit in greasy wash water all year". The bottom panel is for the phoneme `/sh/` occurring in the word "she" of the same sentence. Since we don't have the ground truth for natural speech, we have compared the proposed method's estimate with the LPC magnitude spectrum. Since the range of the LPC magnitude spectrum and chirp GDS can be quite different, to facilitate comparison, each curve was scaled such that the maximum absolute value became unity after scaling. The formant peak locations of the proposed method and that of the LPC are in good agreement, lending credence to this estimate.

To establish the ability of the proposed method to capture spectral valleys that characterize nasal sounds, a pole-zero filter with formants at 290 Hz, 2400 Hz, and 2900 Hz, and a spectral valley at 1200 Hz was used. The formant bandwidths were 75, 85, and 100 Hz; the spectral valley's bandwidth was 80 Hz. This filter was excited by an impulse train with an f_0 of 120 Hz. As before, a Blackman window of 30 ms was used for analysis. The results are shown in Fig. 6. For comparison, the LPC magnitude spectrum and the zero-phase chirp group delay spectrum are shown. LPC's inability to capture valleys is well-known and does not require elaboration. Whereas, this example illustrates that the zero-phase chirp GDS method also suffers from the same drawback as LPC. Note, however, that Bozkurt's GCI-based method [6] is also able to capture spectral valleys similar to our method, but requires accurate GCI information.

The proposed method is fairly insensitive to the starting point of the analysis window. This means that one need not estimate the glottal closure instants (GCIs) that is crucially needed for Bozkurt's method. The insensitivity of our estimates to the starting point of the analysis window is illustrated in Fig. 7. Overlay of filter estimates with different starting points for a 30 ms segment of synthetic voiced speech (top) and unvoiced speech (bottom) are shown. Also, we observed that for a rectangular window the variability of the filter estimates is much more, unlike for the Blackman window.

Our method is also less sensitive to the length of the analysis window. In Fig. 8, the resulting of varying the analysis window from 20 ms to 50 ms is shown when analyzing a voiced segment. As can be seen, the variability in the estimates is acceptably small.

We used our method to track the formants in the sentences "Should we chase those cowboys?" (`m0127s.dat` and "We were away a year ago" (`m0125s.dat`), both taken from [7]. A 30 ms Blackman window was used with an overlap of 22.5 ms; a simple peak picking algorithm was used for locating the formant peaks in each frame (without any further post-processing). The estimated tracks agree very well with the peaks of the wideband spectrograms (Fig. 9).

The proposed method's ability to capture the valleys

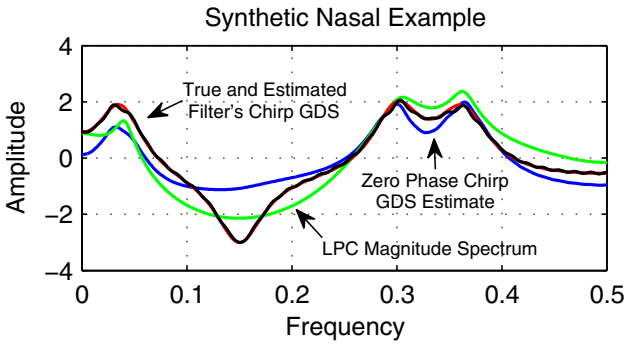

Fig. 6. Chirp GDS of true and estimated vocal tract filters for a synthetic nasal sound. Also shown are the LPC magnitude spectrum and the estimate obtained using zero-phase chirp GDS. Both methods are unable to capture the spectral valley. However, the proposed method captures the spectral valley quite accurately.

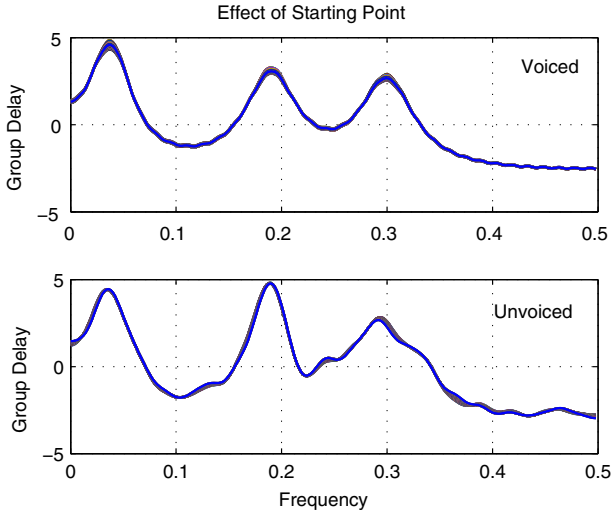

Fig. 7. Overlay plots of chirp GDS of estimated filter for a 30 ms segment of synthetic voiced speech (top) and unvoiced speech (bottom) as a function of the starting point. The estimates are fairly insensitive to the starting point of the analysis window.

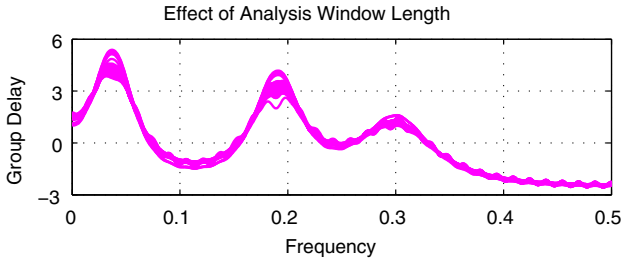

Fig. 8. Overlay plots of chirp GDS of estimated filter when varying the length of the analysis window from 20 ms to 50 ms. The variability is acceptably small.

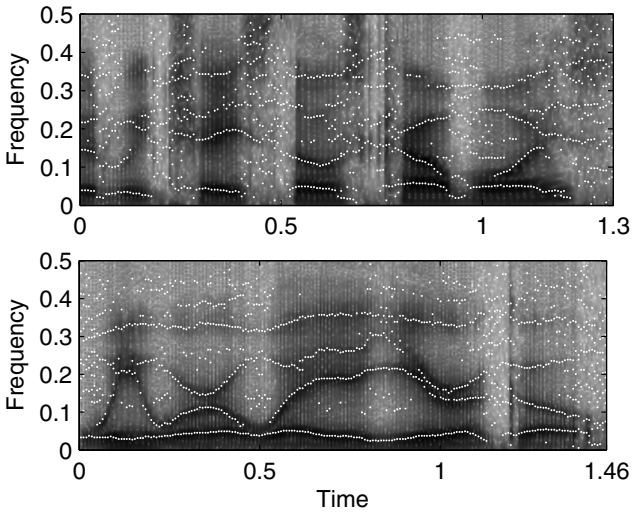

Fig. 9. Formant tracks from our method overlaid on the spectrograms of the sentences, "Should we chase those cowboys?"(top) and "We were away a year ago"(bottom).

present in the nasal spectrum in turn led to improved recognition performance when we tried it on the UCLA nasal sounds database. This database consists of the nasal consonants /m/ and /n/ in pre-stressed syllable-initial position with the vowels /a/, /i/, and /u/ [8,9]. The speech examples were spoken by 2 male and 2 female speakers with 8 repetitions per syllable (192 tokens in total, 16 kHz sampling rate). Training and testing each used 96 sentences. A standard Hidden Markov Model (HMM) based automatic speech recognition system (ASR) was used in our experiments. Each HMM had 6 states and each state was represented by a mixture of 3 Gaussians. We used the same set of triangular filters that are standard fare in MFCC feature extraction. The analysis window size was 30 ms with an overlap of 15 ms.

The recognition accuracy was tested using features derived from (a) standard MFCC, (b) zero-phase chirp GDS, and (c) the proposed method. The recognition accuracy is given in Table 1. The proposed method gives the best performance for nasals compared to the other two. Note that our baseline MFCC result of 95.8% is better than the 90% reported in [9].

Table 1. Recognition accuracy for different methods.

Features	**Number of syllables accurately recognized**
MFCC	92 (95.8%)
Zero phase method	92 (95.8%)
Proposed method	96 (100%)

4. DISCUSSION

When compared with Bozkurt's method [6] our approach is superior in the following respects: (a) it does not require the estimation of glottal closure instants; (b) it is fairly insensi-

tive to the analysis window position and length; and (c) it is applicable to unvoiced speech segments also. In Fig. 10 we show the results of using Bozkurt's GCI-based method when analyzing a voiced segment by varying the starting point of the analysis window: the unacceptable variation due to lack of proper alignment is clearly evident (in our method the variability is very minimal—see Fig. 7). In addition, the method in [6] requires the window to be two cycles long; if more cycles are taken, the performance deteriorates [1].

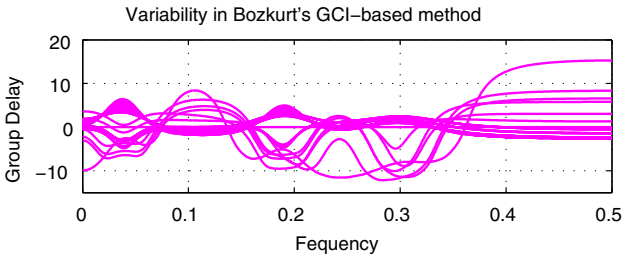

Fig. 10. Overlay plots of chirp GDS of estimated filter obtained using Bozkurt's GCI-based method when the window is not positioned synchronously with the GCI. The variability is unacceptably high.

Bozkurt's "zero-phase chirp group delay" method [2] shares the many advantages of our method, such as insensitivity to the starting point and duration of the analysis window, not requiring knowledge of GCIs, and the ability to give good estimates for unvoiced segments as well. However, it is not able to capture the valleys in nasal sounds, as was shown earlier. We also observed that methods such as MODGD [10] and Zhu and Paliwal's "product spectrum" [11] are unable to capture spectral valleys.

The main disadvantage of the proposed method is the need to compute the roots of a high order polynomial, which is, in general, a difficult task. Bozkurt's GCI-based method, too, requires computing the roots of a large order polynomial.

5. CONCLUSION

In this paper we proposed a method that combines reflection of outside-unit-circle zeros and evaluation of group delay about a circle whose radius is greater than unity. Because we do not discard any zero, the additivity property of source and filter GDS is preserved, which is also approximately true for windowed data. With the help of synthetic and natural speech examples we demonstrated that it gives good estimates of the vocal tract filter. Unlike Bozkurt's method [6], it does not require knowledge of GCIs, is much less sensitive to the starting point of the analysis window and its duration, and works well even for unvoiced speech. It is able to capture very well the spectral valleys that occur in nasal sounds. The ability to model nasals better results in improved recognition accuracy, as illustrated on the UCLA nasals database. The proposed method's main drawback is the computationally intensive and numerically sensitive task of finding the roots of a large order polynomial.

6. REFERENCES

[1] B. Bozkurt, *Zeros of the Z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals*, Ph.D. thesis, Polytech. Mons, Faculte Polytechnique de Mons, Octoberl 2005.

[2] Baris Bozkurt, Laurent Couvreur, and Thierry Dutoit, "Chirp group delay analysis of speech signals," *Speech Communication*, vol. 49, no. 3, pp. 159 – 176, 2007.

[3] Hema A. Murthy and B. Yegnanarayana, "Group delay functions and its applications in speech technology," *Sadhana, Indian Academy of Sciences*, vol. 36, no. 5, pp. 745–782, October 2011.

[4] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*, Prentice Hall, Upper Saddle River, NJ, 2011.

[5] Baris Bozkurt, Boris Doval, Chritophe D'Alessandro, and Thiery Dutoit, "Appropriate windowing for group delay analysis and roots of z-transform of speech signals," in *Proc. 12th European Sig. Proc. Conf. (EUSIPCO)*, Vienna, Austria, 2004, pp. 733–736.

[6] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Zeros of z-transform representation with application to source-filter separation in speech," *Signal Processing Letters, IEEE*, vol. 12, no. 4, pp. 344–347, April 2005.

[7] D. G. Childers, *Speech Processing and Synthesis Toolbox*, John Wiley & Sons, New York, NY, 2000.

[8] Abeer Alwan, Jeff Lo, and Qifeng Zhu, "Human and machine recognition of nasal consonants in noise," in *Proceedings of 14th ICPhS, San Francisco*, 1999, pp. 167–170.

[9] Qifeng Zhu and Abeer Alwan, "On the use of variable frame rate analysis in speech recognition," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. IEEE, 2000, vol. 3, pp. 1783–1786.

[10] Hema A. Murthy, *Algorithms for Processing Fourier Transform Phase of Signals*, Ph.D. thesis, Department of Computer Science & Engineering, IIT Madras, December 1991.

[11] Donglai Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *Proc. IEEE ICASSP–2004*, May 2004, vol. 1, pp. 125–128.