Analysis of dynamics of vocal tract system using zero time windowing method

RaviShankar Prasad

International Institute of Information Technology, Hyderabad, India

ravishankar.prasad@research.iiit.ac.in

Introduction to the research problem

Speech signals are output of a dynamic production mechanism varying continuously with time. The process of speech production is dictated by the linguistic and para–linguistic information being conveyed by speaker. The resulting speech signal captures the characteristics of the time–varying vocal tract acoustic system. Consistent attempts have been made to understand and model the dynamics of the speech production characteristics. Almost all of these modeling analogies assume a stationary behavior of the vocal tract over an interval of time, owing to a limitation with their resolution. The spectral analysis based on discrete Fourier transform (DFT) or filterbank also study the behavior of speech averaged over temporal and spectral bands. These analysis and modeling methods do not incorporate the knowledge of the production characteristics and its transient nature.

The scope of the present research spans across studying the dynamics of the acoustic vocal tract system response during the production of speech. The opening and closing of the vocal folds at the glottis and the lowering of velum are events which impart a dynamic nature to the acoustic system response. These articulatory movements result in the production of voiced, nasal and nasalized vowel segments, which comprise a significant volume of speech. The average behavior of the production system characteristics for these segments has been studied and incorporated to develop systems and applications. The present research attempts to study the behavior of acoustic system response during the production of these segments with an improved resolution. The research also proposes new methods to address the identification of the glottal open phase, identification of presence of nasalization and detection of the extent of nasalization in vowels.

Related studies

The study of the glottal activity is carried out based on the source estimates obtained after canceling the vocal tract system components. Determination of the glottal opening instant (GOI) rely mostly on identification of GCI and further deciding on a suitable duration for the open phase [1]. The glottal inverse filtering (GIF) method to compute vocal tract response is usually implemented using closed phase inverse filtering [2], the digital all–pole (DAP) modeling technique [3, 4], or the auto–regressive moving average modeling of speech signals [5]. The iterative adaptive inverse filtering estimates the source with a two step iterative procedure using LP analysis of different orders in cascade [6]. The Dynamic programming projected Phase Slope Algorithm uses the phase slope function of the LP residual signal and a N–based dynamic programming to identify the GCIs and GOIs from speech [7]. The Yet Another GCI/GOI Algorithm demarcates GCI/GOI using a wavelet analysis along with group delay function and N–based DP over the glottal source signal estimated using IAIF [8]. Other GOI identification methods primarily explore a significant singularity between GCIs in the glottal source (EGG or LP residual) using wavelets and multilevel decomposition techniques [9, 10, 11, 12].

Study of vowel nasalization also has been carried out using spectral and temporal characteristics of the vowels occurring in different context with nasals. The presence of a low frequency spectral peak (in 250–300 Hz range) in the vowel spectrum and a spectral zero in the 700–1800 Hz range is understood as a characteristic signature of the coupling of oral and nasal cavities [13, 14]. Widening of the first formant ($F1$) bandwidth also indicates the coupling of the nasal and oral cavities [15, 16, 17]. The effects of synthetic introduction of a pole–zero pair in the low frequency region for vowels are studied for perception of nasalization [13, 18, 19]. A spectral correlate ($A1$–$P1$) to identify the presence of nasalization in vowels was introduced [20]. $A1$ is the amplitude of the peak harmonic closest to $F1$, and $P1$ is the amplitude of the nasal peak in the vicinity of $F1$. Another correlate ($A1$–$P0$), where $P0$ is the amplitude of the first resonance peak at low frequencies, was also suggested [21]. The average values of both these correlates are lower for vowel segments occurring in nasal context. Another study proposed a set of nine acoustic correlates aimed to capture low frequency behavior for detection of nasalization [22].

These studies highlight the production system behavior in a gross manner. The research uses new methods to present an improved understanding of different phenomena in the acoustic system response during the production of voiced and nasalized segments.

Studies of dynamics of speech production

Proposed methods

The studies in present research use the zero time windowing (ZTW) based analysis of speech signals. ZTW has capability to provide a good spectral resolution at a high temporal resolution, for analysis segments. ZTW segments speech signal using a heavily decaying window, which results in an integration operation performed twice in the frequency domain. The spectrum is therefore represented by computing the Hilbert envelope of the successive differentiation of the numerator of group delay function (HNGD). HNGD has an improved resolution around the spectral peaks [23], and is computed at every sampling instant. The HNGD spectrum is parameterized using the dominant resonance frequency (DRF) and the second dominant resonance (DRF_2). The spectra is smoothed using a 3–point median filtering to highlight the peak location, identified at the zero–crossings of the differenced spectra. The DRF can be seen in correspondence to the dominant cavity involved in the production of the segment. A change in the spectral characteristics of speech owing to the change in configuration of vocal tract acoustic system, reflects as a shift in their dominant resonance. The DRF contour serves as a concise representation of dominant characteristics of the production system. ZTW can be performed over a range of window lengths ($\sim$ 3 ms to 30 ms) to obtain acoustic system response. A smaller window helps in study of short–time phenomena in speech.

Study of glottal activity

The glottal activity is generally studied based on source characteristics and doesn't clearly explain the response across different phases. Identification of the glottal opening and glottal open phase is difficult due to the weak excitation behavior of GOI. The research studies glottal activity for its effect on the characteristics of production system response across different glottal phases.

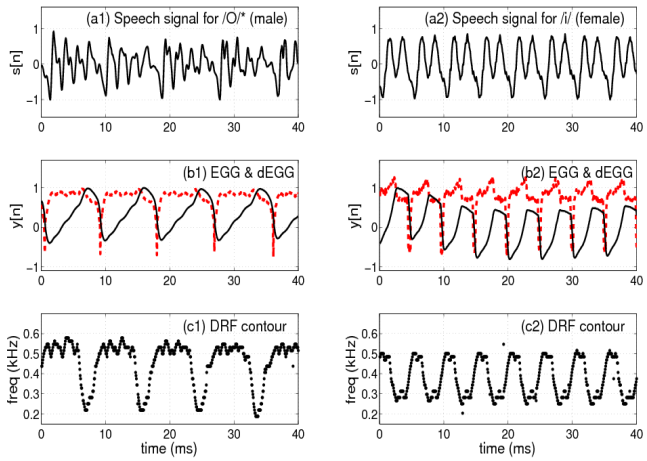

Figure 1: *DRF contours for vowel segments, obtained from HNGD. (a1), (a2) Speech signals from male and female speakers. (b1), (b2) EGG (solid line) and dEGG (dotted line) signals. (c1), (c2) DRF contours.*

Figure 1 illustrates the DRF based representation of glottal activity, for voiced segments obtained from male and female speakers. The DRF contour is obtained from the HNGD spectra computed using ZTW with a window length 4 ms. The opening and closing of the vocal folds leads to a periodic change in the length of the production cavity. An increment in length of the production tract at GOI can be seen as a transition in DRF to low frequency range, across successive glottal cycles. As the analysis window approaches the GCI, the DRFs transit back to relatively higher frequency. The transitions in the DRF contour corresponding to different glottal phases appear in contrast to singularities in the EGG signal, for both the speakers. An algorithm is developed to extract the glottal opening and the glottal open phases using the present representation [24]. The study also helps in analyzing the behavior of factors such as changes in formant locations and F_1 with respect to the glottal open phase. The results obtained from the proposed study over speech signals appear close to those obtained over EGG signals.

Study of vowel nasalization

Nasalization of vowels result in a distinct category of sound where the air flow takes place through both oral and nasal cavities. The extent of coupling called degree of nasalization, depends on the extent of opening of the velopharyngeal section, and hence the volume of airflow through each of these cavities. The research studies co–articulatory nasalization for vowels appearing in a phonetic proximity to nasal consonants. The DRF contours for nasal and vowel segments appear in distinct frequency range. The nasal DRFs are weaker than the oral DRFs due to lossy nature of the nasal tract. The dominant spectral characteristics for a coupled oral and nasal cavities alternate between oral and nasal DRF regions.

Figure 2 shows DRF and DRF_2 contours to illustrate the presence and extent of coupling of oral and nasal cavities across

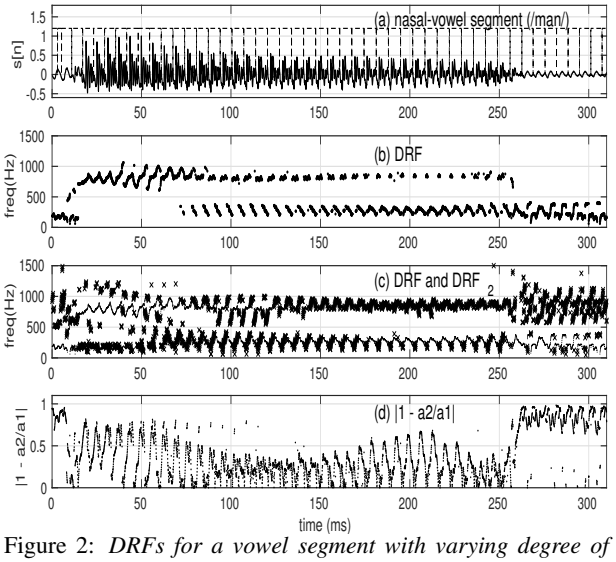

Figure 2: *DRFs for a vowel segment with varying degree of coupling of nasal and oral tracts. (a) Speech segment with the GCI locations (dotted). (b) DRF contour and (c) DRF (–) along with DRF_2 (**). (d) Relative resonance strength difference.*

a vowel segment. The speech signal, corresponding to utterance 'man', exhibits three segments with a distinct DRF contour. For the non–nasalized segment, the DRF remain in high frequency range, characteristic to vowels. For a partial degree of nasalization, the DRF contour fluctuates between low and high frequency range, characteristic to nasal and vowel segments, respectively. This signifies the present of both cavities during the production of the segment. For a full degree of coupling, the DRFs for vowel segments appear in low frequency range, corresponding to nasals. The DRF_2 contour indicates the presence of both cavities across the length of nasalized segment, with varying dominance. A normalized difference between the strengths of DRF and DRF_2 explain that glottal open regions help in highlighting the low frequency resonances in nasalized vowels.

Conclusion and future road map

The present research intends towards a detailed study of the dynamics in the acoustic system response during the production of speech. The DRF contour obtained using a ZTW analysis efficiently represents the changes in dominant behavior of production system across a glottal cycle. The glottal activity derived from change in system characteristics is a unique and reliable way to study different glottal phases. Study of nasalization highlights the transient system response for a coupled oral and nasal cavity across successive glottal cycles. An extension to this analysis is planned to study the effect of presence of nasal sounds in different contexts to vowels, to study the contextual load behavior. Further studies are also planned towards analysis of production characteristics for other dynamic sounds such as stops, approximants, glides, etc.

Advents in speech application requires the development of sophisticated systems for real time and low resource scenarios. Such systems call for an evolved understanding of speech production mechanism, which can result in an improved performance. The present research can be utilized to provide contrastive details to build improved systems for applications spanning across speech recognition, synthesis, enhancement, coding etc.

References

[1] D. G. Childers and C. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[2] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice Hall, 1978.

[3] A. El-Jaroudi and J. Makhoul, "Discrete all–pole modeling," *IEEE Transactions on signal processing*, vol. 39, no. 2, pp. 411–423, 1991.

[4] P. Alku and E. Vilkman, "Estimation of the glottal pulseform based on discrete all–pole modeling," in *International Conference on Speech & Language Processing (Yokohama, Japan)*, 1994, pp. 1619–1622.

[5] Y. Ting and D. Childers, "Speech analysis using the weighted recursive least squares algorithm with a variable forgetting factor," in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-90.* IEEE, 1990, pp. 389–392.

[6] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[7] M. R. Thomas, J. Gudnason, and P. A. Naylor, "Detection of glottal closing and opening instants using an improved dypsa framework," in *Proc. 17th Eur. Signal Process. Conf, (Glasgow, Scotland)*, 2009, pp. 2191–2195.

[8] ——, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.

[9] M. R. Thomas and P. A. Naylor, "The sigma algorithm: A glottal activity detector for electroglottographic signals," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 8, pp. 1557–1566, 2009.

[10] A. Bouzid and N. Ellouze, "Local regularity analysis at glottal opening and closure instants in electroglottogram signal using wavelet transform modulus maxima," in *Eighth European Conference on Speech Communication and Technology (EUROSPEECH,03) (Geneva, Switzerland)*, 2003, pp. 2837–2840.

[11] ——, "Open quotient measurements based on multiscale product of speech signal wavelet transform," *Journal of Electrical and Computer Engineering*, vol. 7, pp. 1–5, 2007.

[12] C. d'Alessandro and N. Sturmel, "Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude," *Sadhana*, vol. 36, no. 5, pp. 601–622, 2011.

[13] A. S. House and K. N. Stevens, "Analog studies of the nasalization of vowels," *Journal of Speech and Hearing Disorders*, vol. 21, no. 2, pp. 218–232, 1956.

[14] M. K. Huffman, *Implementation of nasal: timing and articulatory landmarks*. Phonetics Laboratory, Department of Linguistics, UCLA, 1990.

[15] S. Hawkins and K. N. Stevens, "Acoustic and perceptual correlates of the non-nasal–nasal distinction for vowels," *The Journal of the Acoustical Society of America*, vol. 77, no. 4, pp. 1560–1575, 1985.

[16] J. Glass and V. Zue, "Detection of nasalized vowels in american english," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'85.*, vol. 10. IEEE, 1985, pp. 1569–1572.

[17] G. Fant, *Acoustic theory of speech production*, 1958.

[18] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[19] M. Båvegård, G. Fant, J. Gauffin, and J. Liljencrants, "Vocal tract sweeptone data and model simulations of vowels, laterals and nasals," *STL-QPSR*, vol. 4, pp. 43–76, 1993.

[20] M. Y. Chen, "Acoustic parameters of nasalized vowels in hearing-impaired and normal-hearing speakers," *The Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2443–2453, 1995.

[21] ——, "Acoustic correlates of english and french nasalized vowels," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2360–2370, 1997.

[22] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for the automatic detection of vowel nasalization." in *Proc. of Int. Conf. on Spoken Language Processing (INTERSPEECH, 2007) (Antwerp, Belgium)*. Citeseer, 2007, pp. 1925–1928.

[23] M. A. Joseph, S. Guruprasad, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *Proc. Int. Conf. Spoken Language Processing (INTERSPEECH)*, Pittsburgh PA, USA, Sept. 2006, pp. 1009–1012.

[24] R. S. Prasad and B. Yegnanarayana, "Determination of glottal open regions by exploiting changes in the vocal tract system characteristics," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 666–677, 2016.