Applying production knowledge for speech signal processing

Bhanu Teja Nellore

Speech Processing Laboratory,
International Institute of Information Technology, Hyderabad, India
bhanu.nellore@research.iiit.ac.in

1. Introduction

Knowledge of speech production process has been useful in several areas such as phonetics, linguistics, speech pathology and speech technology. In fact, the availability of the knowledge of the production process of speech signal, distinguishes it from other natuaral signals like radar, image etc.. This allows the speech signal to be analyzed or processed in non-statistical methods as well.

The focus of my thesis is to reliably extract production related information from a given speech signal and use this information to address certain tasks of speech technology.

This report is organized as follows. Certain acoustic features that reflect underlying speech production processes are described in Sec. 2. These features have been ussed in some of my previous studies. These studies are briefed in Sec. 3. Sec. 4 discusses some potential areas where the information described in Sec.s 2 and 3 can be useful, followed by acknowledgements in Sec. 5.

2. Acoustic features for extracting speech production information

Acoustic features described in this section are extracted using methods such as Zero Frequency Filtering, Zero Time Windowing and Single Frequency Filtering. Detailed descriptions of these methods can be found in [1], [2] and [3] respectively.

2.1. Excitation source based acoustic features

2.1.1. Glottal Closure Instants (Epochs)

Epoch locations are extracted using zero frequency filtering (ZFF) method [1]. This method involves passing the differenced speech signal through a cascade of two zero frequency resonators (ZFR). The ZFF signal clearly shows sharper zero crossings around the epoch locations. Hence the negative to positive zero crossing instants in ZFF signal are called epochs. The features of the glottal source of excitation derived from ZFF signal are as follows:

2.1.2. Strength of excitation (α)

Slope of ZFF signal around epochs gives a measure of the strength of impulse-like excitation (α). α corresponds to the rate of glottal closure [4]. Sharper the glottal closure, higher is the value α and vice-versa.

2.1.3. Energy of excitation (β)

Energy of excitation (β) is computed as the energy of the ZFF signal within a window length of 3 msec, centered at every epoch location (1.5 msec on each side of epoch). Window length of 3 msec is considered around each epoch to capture the predominant excitation source information around the epoch locations.

Epochs extracted from a given utterance using ZFF signal, the values of α and β for that utterance are shown in Figure 1.

2.2. Vocal tract system based acoustic features

Spectral characteristics of the vocal tract system are extracted using Zero-time windowing (ZTW) method [2]. Using ZTW, spectral information can be obtained with high spectral and temporal resolution at any instant of time, even for speech segments less than 5 msec. Spectral features derived from the Hilbert Evelope of Numerator Group Delay (HNGD) spectrum are described below:

2.2.1. Dominant resonance frequency (DRF)

Dominant resonance frequency (DRF) refers to the frequency of the dominant peaks in the obtained HNGD spectrum, as they represent the dominant resonances of the vocal tract system [5].

2.2.2. Dominant resonance strength (γ)

Dominant resonance strength (γ) is measured as the magnitude of the HNGD spectrum at DRF.

2.2.3. Slope of dominant resonance frequency (F_{sl})

Slope of dominant resonance frequency (F_{sl}) refers to the first order difference of the DRF values at epochs. Here absolute value of slope is considered.

2.2.4. Slope of dominant resonance strength (γ_{sl})

Slope of dominant resonance strength (γ_{sl}) refers to the first order difference of the γ values at epochs. Here absolute value of slope is considered.

The values of DRF, γ, F_{sl} and γ_{sl} extracted for a given speech utterance are shown in Figure 2 respectively.

2.3. Single Frequency Filter envelope and phase only signal

The objective in Single Frequency Filter (SFF) [3] is to derive the amplitude envelope of the signal as a function of time at a desired frequency component. The SFF is performed at $f_s/2$ (f_s=sampling frequency) for each frequency component, after frequency shifting the signal. This ensures that the filter characteristics remains same for all the frequency components. Since the SFF is performed using a resonator at $f_s/2$, whose pole is located close to the unit circle, the effect of other frequency components are reduced significantly.

We use the phase information $\phi_k[n]$ obtained from the SFF spectrum to reconstruct the speech signal. We replace the envelope values $e_k[n]$ with one and then use the phase to reconstruct the speech signal. This signal is called as phase only signal. For

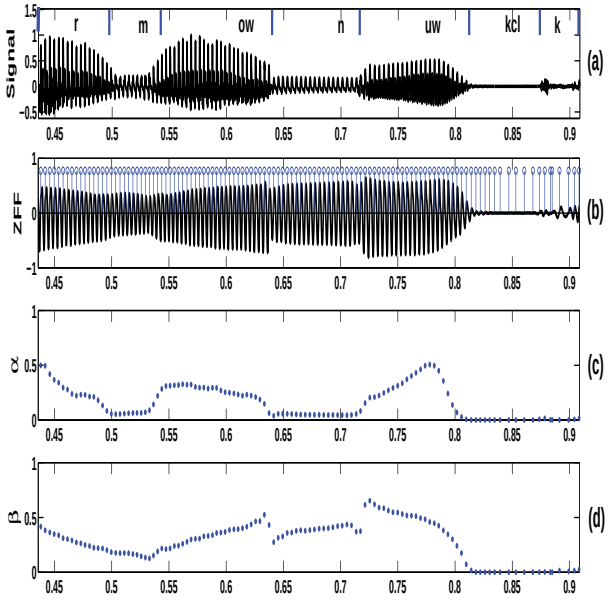

Figure 1: *Acoustic features from ZFF signal (a) Speech waveform for an utterance " ..ma/ nucl...". Manually marked phoneme labels are given above the signal. (b) ZFF signal along with hypothesized epochs at the positive zero crossings of the ZFF signal, (c) strength of excitation (α) values around epochs, and (d) energy of excitation values (β) around epochs. X-axes represent time in seconds.*

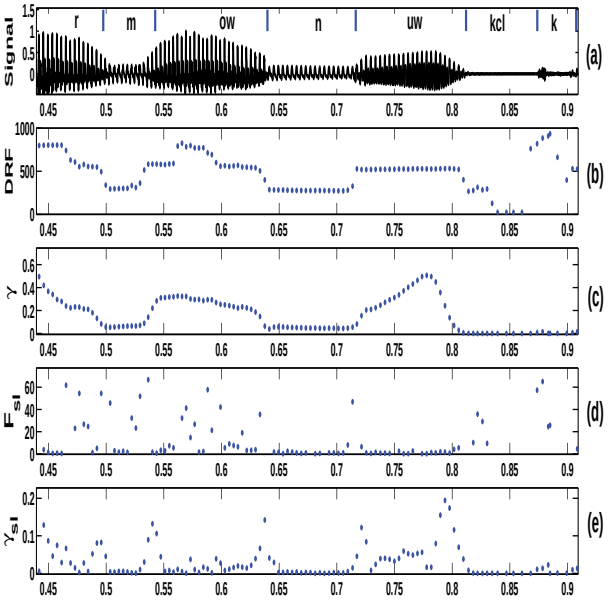

Figure 2: *Dominant resonance frequency based acoustic features. (a) Speech signal with manually marked phoneme labels, (b) dominant resonance frequency (DRF) values, (c) dominant resonance strength (γ) values, (d) slope of DRF (F_{sl}) values, and (e) Slope of dominant resonance strength (γ_{sl}) values at epochs. X-axes represent time in seconds.*

detailed analysis please refer to [6]. Figure 3(a) shows a speech signal $s[n]$ along with phase only signal (Fig. 3(b)).

3. Studies previously done

α and β are generally high in sonorant regions. In the range of 0 - 900 Hz, DRF and γ values are higher for vowels than other sounds. DRF values are lower for nasals compared to other speech sounds. SFF based phase only reconstructed signal enhances, in the temporal domain, every segment in a speech signal which can be used to locate sounds such as weak bursts. Therefore by analyzing acoustic properties of different speech sounds using features described in Sec. 2, the following problems were addressed:

- sonorant segmentation [7],
- vowel landmark detection [8],
- locating burst onsets [6] and
- nasal detection [9].

A snapshot of results obtained in these studies are presented in Table 1. For detailed information, please refer to the papers of respective studies mentioned above.

4. Future work

Knowledge of category of speech sound is critical in many applications. For example, noise affects each sound category in a unique manner. Therefore one potential area to use above knowledge is near end speech enhancement where we are looking to modify clean speech based on sound categories in order to make it more intelligible in noisy conditions.

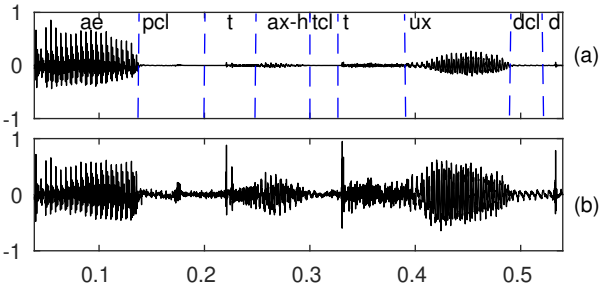

Figure 3: *(a) Speech signal for the utterance 'aptitude' (phone boundaries are marked using dashed line). (b) Signal reconstructed using phase only information.*

Table 1: *Previous studies made by the author along with the corresponding results obtained.*

Task	Evaluation Measure	Result (%)
Vowel Landmark Detection	Precision	91.71
	Recall	94.77
	F Measure	93.21
Sonorant Segmentation	Accuracy	93.95
	True Positive Rate (TPR)	94.47
	False Alarm Rate (FAR)	7.53
Locating Burst Onsets	% identified within a deviation of 10ms	79.2
Nasal Detection	TPR	91.49
	FAR	16.58

5. Acknowledgements

The author would like to thank his PhD advisors Prof. B. Yegnanarayana and Dr. Suryakanth V Gangashetty.

6. References

[1] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[2] B. Yegnanarayana and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013.

[3] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 705–717, 2015.

[4] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE signal processing letters*, vol. 16, no. 6, pp. 469–472, 2009.

[5] R. Prasad and B. Yegnanarayana, "Acoustic segmentation of speech using zero time liftering (ZTL)," in *Proc. of Interspeech*, 2013, pp. 2292–2296.

[6] B. T. Nellore, R. Prasad, S. R. Kadiri, S. V. Gangashetty, and B. Yegnanarayana, "Locating burst onsets using sff envelope and phase information," in *Proc. Interspeech 2017*, Hyderabad, India, 2017, pp. 3023–3027.

[7] S. H. Dumpala, B. T. Nellore, R. Nevali, and B. Yegnanarayana, "Robust features for sonorant segmentation in continous speech," in *Proceedings of Interspeech*, Dresden, Germany, May 2015, pp. 1987–1991.

[8] S. H. Dumpala, B. T. Nellore, R. Nevali, S. V. Gangashetty, and B. Yegnanarayana, "Robust vowel landmark detection using epoch-based features," in *Proceedings of Interspeech*, San Francisco, USA, Sep. 2016, pp. 160–164.

[9] B. T. Nellore, S. H. Dumpala, K. Nathwani, and S. V. Gangashetty, "Excitation source and vocal tract system based acoustic features for detection of nasals in continuous speech," in *Proc. Interspeech 2019*, Graz, Austria, 2019, p. to appear.