Scope of Sonority Information in Statistical Parametric Speech Synthesis

Bidisha Sharma

Dept. of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati-781039,

s.bidisha@iitg.ernet.in

Abstract

Statistical parametric speech synthesis (SPSS) has been widely used due to its advantages like flexibility in terms of adaptation of speakers, emotions, lower footprint, robustness. However, the muffled quality of synthesized speech remains as one of the major drawbacks of SPSS. In this work, some of the shortcomings are attempted to compensate by integrating additional sonority feature into the SPSS framework. The scope of sonority information to reduce over-smoothing and improve voicing decision during excitation source generation is presented. It is analyzed that, the spectral tilt associated with the frames of synthesized speech is more compared to that of natural. A novel spectral tilt modification method is explored to minimize the deviation between natural and synthesized speech. The contributions made to these different modules of SPSS are expected to bring significant improvement to the synthesized speech.
Index Terms: Sonority, Post-filtering, Spectral tilt, Voicing.

1. Introduction

With the increase in contribution of technology in day-to-day life of human being, a natural sounding compact text-to-speech (TTS) synthesis system becomes of paramount significance. As the practical applicability demands a low footprint TTS system with sufficient intelligible synthesized speech, the research in speech synthesis has been moving from concatenation of speech segments towards use of statistical generative models since the last decade. Instead of directly storing the waveforms corresponding to training speech corpora, statistical parametric speech synthesis (SPSS) preserves the generative models trained using source and spectral features derived from the speech corpora with reference to corresponding labels. During synthesis, the models are used to generate the same features which is further passed through a vocoder to render the synthesized speech signal.

The averaging during modeling may lead to loose prosodical as well as other higher level detailed information intact in the natural speech signal. This results several developments in different modules of SPSS with a common goal to bring the synthesized speech quality to the level of natural speech. The three basic reasons behind the muffled quality of SPSS synthesized speech are simple vocoder architecture, inaccuracy in statistical modeling of parameter sequences and over-smoothing of generated parameter contours. Apart from these three primary reasons, the other facts include poor representation of speech signal in parametric form, inadequate parameter generation algorithm, error in voicing decision algorithm, simple excitation source generation method, lack of additional prosodic information in the synthesized speech.

In this report, we mainly focus on poor parametric representation of speech signal, over-smoothing of generated parameter sequences and voicing decision during excitation source generation. The conventional parameters used as representation of speech signal in SPSS are fundamental frequency (F_0), mel frequency cepstral coefficients (MFCCs) and their dynamic counterparts. These features may not preserve enough information to reconstruct back the speech signal with all characteristics intact. To bring out different modifications in the SPSS framework at different levels, it is required to incorporate some additional features, that can be also modeled using HMM. Several additional parameters along with their subsequent applications are used in the literature as mentioned in [1–4]. As the sonority notion is associated with formant prominence and excitation source strength that effect our perception, therefore an attempt is made to extract sonority information from the speech signal. To overcome the issue of over-smoothing, post-filtering [2,5,6] and improved parameter generation [7] methods are widely used. The sonority feature can be incorporated in the SPSS framework and employed to develop better post-filtering method. Another factor that significantly contributes to synthesized speech quality is excitation source generation. Voiced/unvoiced decision is an essential component for generation of excitation source [8]. It is obtained from F_0 and other excitation source evidences in the existing literature [9, 10]. Incorporation of improved voicing decision using sonority may substantially increase the naturalness of synthesized speech. Combination of the individual contributions to each of the above modules may bring significant improvement to the quality of SPSS synthesized speech.

The rest of the paper is organized as follows, in Section 2 an efficient representation of sonority information along with its significance is explained. The scope of the sonority information to employ in post-filtering along with a method for modification of spectral-tilt is presented in Section 3. Section 4 describes the significance of sonority feature in voicing decision. The direction of future work is presented in Section 5.

2. Extraction of Sonortiy Information

Sonority refers to relative loudness of sound units resulting from higher energy and periodicity [11]. As found in the study made in [12, 13], the relative sonority notion between two adjacent sound has a strong impact on human speech perception. It influences the perception of syllable, word structure and therefore may have impact on speech intelligibility which is less studied in the literature of speech processing [14]. The variation in sonority associated with different sound units is due to the change in the behavior of different articulators during production, that is manifested in the speech signal. It is associated with spectral sharpness, strength of excitation (SoE) and periodicity, which greatly effects the perception of speech signal. As a perceptual representation of speech signal, an effort has been made in [15] to bring out a feature from system, source and suprasegmental aspects of speech signal, having capability to represent degree of sonority associated with a sound unit.

To extract the sonority feature, as a representation of vocal-tract spectrum Hilbert envelope (HE) of differenced numerator

of group delay (HNGD) spectrum is derived for 5 ms window around each epoch location [16]. The HNGD spectrum is used to derive 5-dimensional system feature, which are : (a) mean of first three spectral peak values, (b) mean of relative deviation between amplitudes of first three spectral peaks, (c) mean value of amplitude of formant valleys preceding to each spectral peak, (d) mean slope associated with each of first three formant peaks and (e) mean bandwidth of the first three formants [15]. The excitation source feature for sonority is defined as the ratio between the value of central peak of HE of linear prediction (LP) residual at each epoch location to mean of sample values from 2 ms to 3 ms duration in the 3 ms HE of LP residual segment (1.5 ms left and 1.5 ms right of each epoch). The periodicity aspect of sonority can be quantized in terms of correlation over the speech samples corresponding to successive pitch periods of the speech signal as explained in [15]. This results in representation of sonority in terms of a 7-dimensional feature, that can be integrated to the SPSS framework, for its further utilization in improvement in synthesized speech quality. The sonority feature is also found to give improved performance in vowel onset point detection [17].

3. Post-filtering to Reduce Over-smoothing

The post-filtering methods are employed to enhance the generated source and spectral parameters before passing through the vocoder for synthesis. The existing post-filtering methods do not consider the fact that, the characteristics of the speech parameters may extensively vary based on broad categories of the sound units. Representation of these parameters using single mean and variance may not be able to reflect the intact variabilities and fine structure. The variation in source parameters like F_0, SoE and spectral parameters related to formant peaks and valleys amplitudes between natural and synthesized speech can be analyzed separately for broad categories of sound units. From the analysis post-filtering factors can be derived. This categorization can be performed based on sonority associated with. Based on the sonority associated with a sound unit, its spectral prominence and SoE changes. Therefore, the source and spectral attributes related to different sonorant classes can be post-filtered by different factors. This may help in improving dynamic range as well as reducing deviation from the natural counterparts. To accomplish this the sonority feature needs to be integrated in the SPSS framework and used in the classification of frames into different sonorant categories.

Another aspect of speech signal that effects in perception is spectral tilt [18, 19]. In our previous work [20], it has been explored that the spectral tilt associated with the frames of synthesized speech is more negative compared to that of natural. The suppression of higher harmonics may be one factor behind muffled quality of synthesized speech. To reduce this deviation between natural and synthesized speech, a method of spectral tilt modification is proposed in [20]. In this method, the class specific error vectors are obtained from difference of the average power spectrum of first order LP coefficients corresponding to same set of natural and synthesized speech frames. In the modification step, the first order LP power spectrum is obtained for each test frame and the corresponding class specific error vector is added. From this modified first order power spectrum, corresponding LP filter coefficients are derived. The first order LP residual is then passed through this modified filter to derive tilt-modified speech signal corresponding to each frame. In this case the classification is done based different sonorant categories. The experiments performed in [20] ensured that, modification in spectral tilt achieves improvement in terms of naturalness, intelligibility and speaker similarity.

Another method of enhancement of synthesized speech is proposed in [21], which is applied to synthesized speech obtained from USS. In this method, both LP spectrum and residual signals are enhanced to improve the intelligibility. The samples adjacent to each epoch location corresponding to the LP residual is multiplied by a Gaussian window to improve SoE. The formants of the LP spectrum is also enhanced to improve the formant prominence that effects intelligibility. The synthesized speech obtained from the enhanced LP residual and spectrum is found to be more intelligible and usable in noisy environment. The same method can be also applied in case of SPSS.

4. Sonority in Voicing Decision

The sonority feature represents source, system and suprasegmental information, among which excitation source aspect is abundantly studied in the literature of voiced/unvoiced classification. For most of the voiced sounds the main excitation occurs at the closing of the vocal folds. This is followed by the closed phase, where formants are the most prominent with high amplitude, slope and less bandwidth. Although the mechanism of source generation is independent of the vocal-tract shape, many studies have shown that with the variation in supraglottal pressure due to vocal-tract constriction, the shape of glottal waveform specifically the amplitude changes [22]. Despite this change is not much significant in case of moderate constriction, as the constriction increases resulting in higher supraglottal pressure, its effect on glottal waveform also increases. Therefore the openness of vocal-tract may also play significant role in voicing strength as well as voicing decision. Looking into these aspects, we have carried out studies of voicing decision based on sonority feature, that shows improved performance. Therefore, the modeled sonority feature can be also employed in the voicing decision during excitation source generation in the SPSS framework.

5. Direction of Future Work

The focus of this work is to improve quality of synthesized speech using sonority information. In this regard, a potent representation of sonority feature using vocal-tract system, excitation source and suprasegmental aspect is extracted from speech signal. As a next level, this sonority feature can be modeled using HMM and incorporated in the SPSS framework. It has further scope for utilization in alleviating the over-smoothing of the generated parameter sequence. The sonorant class based dynamic post-filtering of F_0, SoE and spectral parameters will be developed by using the knowledge of the sonority feature. The spectral-tilt modification method explained also uses sonorant class specific error vector. However, the sonority feature is not incorporated with the tilt modification method, which will be our next future work. Some explorations has been done to develop a better voiced/unvoiced classifier using the sonority feature. The same will be integrated with SPSS framework. The modeling of sonority feature in SPSS along with conventional source, spectral features and its application to improve different modules can act as a useful means to assure improved synthesized speech quality.

6. Acknowledgment

I would like to acknowledge Prof. S. R. Mahadeva Prasanna for his inspiration and valuable suggestions in carrying out this work.

7. References

[1] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis." in *INTERSPEECH*, 2001, pp. 2263–2266.

[3] O. Abdel-Hamid, S. M. Abdou, and M. Rashwan, "Improving Arabic HMM based speech synthesis quality." in *INTERSPEECH*, 2006.

[4] C. Hemptinne, "Integration of the harmonic plus noise model into the hidden Markov model-based speech synthesis system," *Master thesis*, 2006.

[5] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method," in *Blizzard Challenge Workshop*, 2006.

[6] K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "Postfiltering for HMM-based speech synthesis using mel-LSPs," in *Proc. Autumn Meeting of ASJ*, 2007, pp. 367–368.

[7] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[8] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[9] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *ICASSP*, vol. 1. IEEE, 1999, pp. 229–232.

[10] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1071–1079, 2011.

[11] S. G. Parker, "Quantifying the sonority hierarchy," Ph.D. dissertation, University of Massachusetts Amherst [Published by the GLSA.], 2002.

[12] M. Miozzo and A. Buchwald, "On the nature of sonority in spoken word production: Evidence from neuropsychology," *Cognition*, vol. 128, no. 3, pp. 287–301, 2013.

[13] I. Deschamps, S. R. Baum, and V. L. Gracco, "Phonological processing in speech perception: What do sonority differences tell us?" *Brain and language*, vol. 149, pp. 77–83, 2015.

[14] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[15] Bidisha Sharma and S R M Prasanna, "Sonority measurement using system, source, and suprasegmental information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 505–518, March 2017.

[16] Y. Bayya 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.

[17] Bidisha Sharma and S R M Prasanna, "Vowel onset point detection using sonority information." in *INTERSPEECH*, 2017.

[18] J. M. Alexander and K. R. Kluender, "Spectral tilt change in stop consonant perception," *The Journal of the Acoustical Society of America*, vol. 123, no. 1, pp. 386–396, 2008.

[19] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, no. 12, pp. 1253–1262, 2009.

[20] Bidisha Sharma and S R M Prasanna, "Enhancement of spectral tilt in synthesized speech," *IEEE Signal Processing Letters*, vol. PP, no. 99, pp. 1–1, 2017.

[21] ——, "Speech synthesis in noisy environment by enhancing strength of excitation and formant prominence." in *INTERSPEECH*, 2016, pp. 131–135.

[22] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.