Speech Communication 99 (2018) 201–210

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Significance of sonority information for voiced/unvoiced decision in speech synthesis

Bidisha Sharma[*], S.R. Mahadeva Prasanna

Indian Institute of Technology Guwahati, Guwahati 781039, India

ARTICLE INFO

Keywords:
SPSS
Excitation source
Voicing decision
TTS

ABSTRACT

The quality of synthesized speech obtained from statistical parametric speech synthesis (SPSS) significantly relies on excitation source generation. Voiced/unvoiced decision is an essential component for generation of excitation source. It is obtained from fundamental frequency and other excitation source evidence in the existing literature. The discontinuity at the point of contact in the vocal-folds excites energy into the vocal-tract resulting voicing effect in the produced speech signal. The perceptual reflection of voicing over the sound produced is correlated with the sonority information which is related to less vocal-tract constriction and significant glottal vibration. Therefore, the possible variation in voicing with the change in supraglottal pressure due to vocal-tract constriction, rate of closing of vocal folds and regularity in structure of the signal are intact in the sonority associated with a sound unit. Voicing and degree of opening of vocal-tract are the two most effective correlates of sonority, that potentially contribute to the sonority hierarchy for sonorants and obstruents uniformly. Therefore, the voicing effect can be captured by the sonority measurement derived from system, source and suprasegmental information in the speech signal. In this work, a novel voiced/unvoiced decision method using sonority information is proposed and integrated in the SPSS framework for generation of excitation source. It leads to better voicing decision compared to the existing methods resulting in synthesized speech of improved quality, which is assured from objective and subjective analysis.

1. Introduction

The statistical parametric speech synthesis (SPSS) is the state-of-the-art speech synthesis technique in the recent literature that generates synthesized speech with sufficient intelligibility. It provides flexibility in terms of adaptation of statistical behavior of different speaking styles, emotions, speakers, languages (Tamura et al., 2001; Yoshimura et al., 2000); compression factor over the other speech synthesizer like unit selection based speech synthesis (USS) (Kim et al., 2006; Gutkin et al., 2010) and robustness (Yamagishi et al., 2009). However, synthesized speech obtained from SPSS lacks naturalness compared to that of USS due to poor vocoder, deficient excitation source generation, inaccuracy in acoustic modeling and over-smoothing of the generated parameter sequences (Zen et al., 2009). The rich characteristics of the speech signal intact in natural speech may not be adequately represented by using only limited acoustic features modeled in the statistical environment. The naturalness in synthesized speech is mostly governed by the excitation source component. In the SPSS framework, along with other factors voiced/unvoiced decision plays a key role in excitation source generation module, which is basically impulse train for voiced frames and random noise for unvoiced frames. Generally, F_0 information in employed to know voiced and unvoiced frames in the utterance to be synthesized. Improving the voicing decision may absolutely improve the excitation source and synthesis quality. There are several successful efforts in the literature towards this direction.

Various time and frequency domain approaches are proposed in the literature for voiced/unvoiced detection (Dhananjaya and Yegnanarayana, 2010). These methods include the features related to speech production based characteristics of voiced sounds such as energy, periodicity, short term autocorrelation, zero crossing rate, autocorrelation peak strength, harmonic measure from the instantaneous frequency amplitude spectrum (Atal and Rabiner, 1976; Arifianto, 2007). These methods use some threshold obtained from empirical observation. To avoid such thresholding, statistical modeling based approaches have gained popularity using hidden Markov model (HMM), Gaussian mixture model (GMM), deep neural network model (Zhang and Wu, 2013). These methods aim to explore better modeling techniques using existing features and requires substantial amount of training data with manual segmentation. In conventional SPSS or HMM based speech synthesis, mel-generalized cepstral (MGC) coefficients are

[*] Corresponding author.
E-mail addresses: s.bidisha@iitg.ernet.in (B. Sharma), prasanna@iitg.ernet.in (S.R.M. Prasanna).

https://doi.org/10.1016/j.specom.2018.04.002
Received 16 June 2017; Received in revised form 6 February 2018; Accepted 6 April 2018
Available online 07 April 2018

used to represent vocal-tract spectrum (VTS) information and fundamental frequency (F_0) is used to model the excitation source aspect. F_0 is modeled along with voicing decision using multi space distribution (MSD)-HMM and consequently error in F_0 estimation is propagated to the excitation source generation module (Tokuda et al., 1999; 2013). In the voiced regions F_0 is modeled as continuous Gaussian distribution and discrete symbol in unvoiced regions (Tokuda et al., 2013). As the voicing decision is dependent on F_0, errors in F_0 estimation leads to error in MSD-HMM. This results in misdetection of voiced and unvoiced frames. In case of weakly voice sounds with lower energy like voiced fricatives, voiced stops and voiced affricates, the corresponding frames may get classified as unvoiced which will be propagated to the MSD-HMM in training.

During synthesis, for the frames corresponding to these sounds erroneous voicing information may be generated that results in creation of random noise as excitation. If these misclassification are repeating for several such frames in the same utterance the naturalness of the synthesized speech severely degrades. On the other hand, for false voiced regions impulse sequence is used as excitation instead of random noise resulting in buzziness in synthesized speech. In Yu and Young (2011); Yu et al. (2009) continuous F_0 model is explored instead of MSD, where F_0 is always available for both voiced and unvoiced regions. In this case, the voicing decision is modeled in an independent stream. There are other attempts to integrate GMM and multilayer perceptron based voicing decision to make improvement in SPSS (Kang et al., 2009; Ogbureke et al., 2012). In Narendra and Rao (2015) MSD-HMM modeling of voiced/unvoiced detection in SPSS is proposed by exploiting zero frequency filter (ZFF) in F_0 estimation. The voicing detection is performed using heuristic threshold over the strength of excitation (SoE) and modeled by MSD-HMM along with F_0. In the current version of SPSS voicing decision is pitch dependent and pitch estimation is made using robust algorithm for pitch tracking (RAPT) approach (Talkin, 2018), which performs frame by frame autocorrelation analysis to capture the periodicity information.

Fig. 1(a) and (b) demonstrate speech signal and corresponding differential electro-glotto-graph (DEGG) with reference voiced/unvoiced marking respectively. Fig. 1(c) shows corresponding voicing decision obtained from RAPT algorithm, where some false voiced frames (corresponding to /hh/) can be observed. Corresponding excitation source and synthesized speech is shown in Fig. 1(d) and (e) respectively. Similarly, in Fig. 1(h) falsely detected unvoiced frames can be seen corresponding to the sound /ae/ for the speech and DEGG signals shown in Fig. 1(f) and (g) respectively. This error is propagated to the synthesized speech as shown in Fig. 1(i) and (j). In most of these cases, we have observed that, if the periodicity of the speech signal is not very distinctly evident the autocorrelation based methods tend to give wrong voicing decision.

As voicing strength is generally governed by the movement of vocal-folds, most of the approaches use excitation source information as dominant feature to extract the voicing information. 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 supra-glottal pressure due to vocal-tract constriction, the shape of glottal waveform specifically the amplitude changes (Stevens, 2000). Despite this change is not much significant in case of moderate constriction, as the constriction increases resulting in higher supra-glottal 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. The sonority associated with a sound unit can be defined in terms of degree of vocal-tract constriction and voicing strength (Parker, 2002). Based on voicing associated with the sonorant and obstruent sounds the degree of associated sonority value changes. The sonority hierarchy can be seen in the increasing order of sonority as : *Voiceless stops, voiceless fricatives, voiced stops, voiced fricatives, (voiced) nasals, voiced laterals, voiced r-sounds, (voiced) high vowels, (voiced) mid vowels, (voiced) low vowels* (Parker, 2002). This correlation between voicing and sonority motivated us to explore the sonority information in the task of effective voicing detection.

In our previous work, a set of features that reflects the behavior of

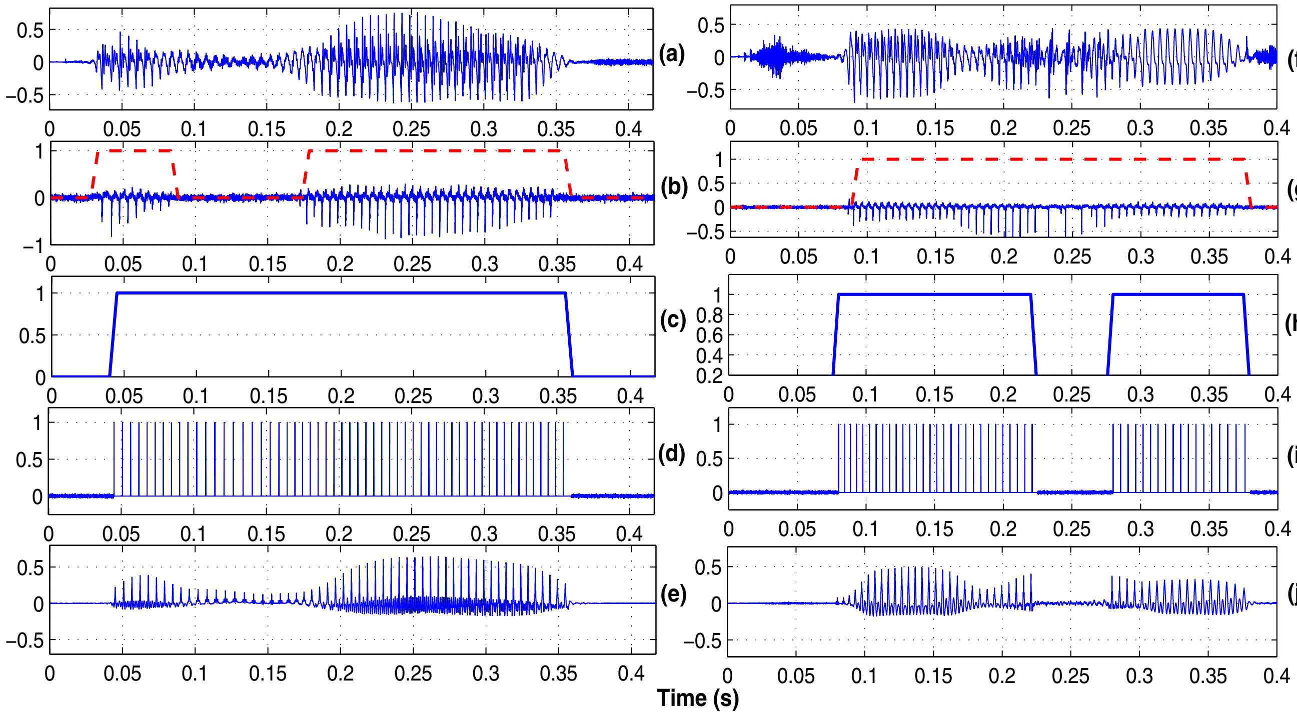

Fig. 1. (a),(f) Natural speech signal; (b),(g) corresponding DEGG with reference voiced/unvoiced marking; (c),(h) voiced/unvoiced decision obtained from RAPT; (d),(i) generated excitation from voicing decision in (c); (e),(j) synthesized speech signal with the excitation shown in (d) and (i) respectively, for SLT speaker.