Digital Signal Processing 71 (2017) 131–143

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Improved voicing decision using glottal activity features for statistical parametric speech synthesis

Nagaraj Adiga *, Banriskhem K. Khonglah, S.R. Mahadeva Prasanna

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, 781039, India

ARTICLE INFO

Article history:
Available online 28 September 2017

Keywords:
Glottal activity features
Statistical parametric speech synthesis
Voicing decision
Support vector machine

ABSTRACT

A method to improve voicing decision using glottal activity features proposed for statistical parametric speech synthesis. In existing methods, voicing decision relies mostly on fundamental frequency F0, which may result in errors when the prediction is inaccurate. Even though F0 is a glottal activity feature, other features that characterize this activity may help in improving the voicing decision. The glottal activity features used in this work are the strength of excitation (SoE), normalized autocorrelation peak strength (NAPS), and higher-order statistics (HOS). These features obtained from approximated source signals like zero-frequency filtered signal and integrated linear prediction residual. To improve voicing decision and to avoid heuristic threshold for classification, glottal activity features are trained using different statistical learning methods such as the k-nearest neighbor, support vector machine (SVM), and deep belief network. The voicing decision works best with SVM classifier, and its effectiveness is tested using the statistical parametric speech synthesis. The glottal activity features SoE, NAPS, and HOS modeled along with F0 and Mel-cepstral coefficients in Hidden Markov model and deep neural network to get the voicing decision. The objective and subjective evaluations demonstrate that the proposed method improves the naturalness of synthetic speech.

1. Introduction

Statistical parametric speech synthesis (SPSS) became a state-of-the-art synthesis approach over the past few years [1]. The merit of this method is its capability to produce a reasonably intelligible speech with a small footprint [2]. The statistical models in SPSS give the flexibility to change the speaking style and prosody of speech [3]. In recent days, SPSS using Hidden Markov models (HMM) and deep neural network (DNN) is extensively used. Statistical models using HMM are commonly implemented using software known as HMM based speech synthesis system (HTS) [1–4]. To build the DNN based speech synthesis system, recently Merlin toolkit is introduced [5]. In Merlin system, linguistic features are taken as input and used to predict acoustic features, which are then passed to a vocoder to produce the speech waveform. However, the synthesis quality of speech is poor concerning naturalness or speaker individuality when compared to unit selection speech synthesis [3,6]. Many factors result in the unnaturalness of synthetic speech. These include poor representation of acoustic features, over-smoothing of acoustic features in statistical modeling, and a simplified vocoder model to generate synthetic speech [7]. This work focuses on the better extraction of acoustic features. In particular, the components present in glottal activity region like aperiodicity, phase information, variation in the strength of excitation, glottal pulse shape etc. have a positive impact on the naturalness of synthetic speech [8,9].

Speech segment can be categorized mainly into voiced and unvoiced regions, based on whether the glottal activity or glottal vibration is present or not. The glottal activity region characterized by variations in the locations of glottal closure instant or epoch due to quasi-periodic vibration of vocal folds, epoch strength due to change in the strength of excitation signal, and an aperiodic component due to the turbulence noise generated in voiced speech. In existing literature [10,11], epochs correspond to glottal closure instants (GCIs), glottal opening instants (GOIs), and onset of bursts [10]. However, GCIs are the major excitation present in speech production. Therefore, in this work, GCIs are referred as epochs. The glottal pulse is characterized by the duration of closing and opening phase of glottal cycle, skewness or glottal pulse shape etc [8,9]. In this work, the significance of glottal activity features illustrated for speech synthesis. Some of these glottal activity parameters are modeled and used to improve the naturalness of SPSS.

* Corresponding author.
E-mail addresses: nagaraj@iitg.ernet.in (N. Adiga), banriskhem@iitg.ernet.in (B.K. Khonglah), prasanna@iitg.ernet.in (S.R. Mahadeva Prasanna).

https://doi.org/10.1016/j.dsp.2017.09.007
1051-2004/© 2017 Elsevier Inc. All rights reserved.

In conventional SPSS using hidden Markov model (HMM), Mel-cepstral coefficients (MCEP) are used to model the vocal-tract transfer function, and the fundamental frequency (F0) or pitch parameter is used to model excitation feature [12]. In the base version of SPSS, voicing decision is computed from F0, which represents only quasi-periodic characterization of glottal activity. For voicing decision, other source parameters of glottal activity are ignored. F0 along with voicing decision plays a critical role in the quality of synthetic speech. The errors in F0 come from the pitch computation algorithm as well as from the average statistical model. In HMM modeling, due to pitch estimation errors, there may be a chance of occurrence of unvoiced frame for a voiced sound. If it occurs within the middle of voiced region, degradation in synthesis quality will be higher. Hence, the synthesis quality of SPSS is not natural like unit selection method [1]. Besides, when the speech sounds are weakly periodic or strength of excitation is low [13,14], there is a chance of missing the classification of some portions of voiced region. For instance, misclassification may happen around voiced–unvoiced (V–UV) transitions and UV–V transitions due to the low amplitude of speech signal [13,14].

This paper focuses on showing the significance of glottal activity features for voicing decision and then training these features in HMM and DNN framework for improving the naturalness of synthetic speech. The focus is on voiced speech as it is perceptually essential and relatively easier to model. The glottal activity features used as a voicing decision in speech synthesis task to generate the excitation signal. However, for voicing decision, some heuristic threshold has to be applied to the final evidence, which can be avoided by using the classifiers. The advantage of using classifiers is transforming features into a higher dimensional space, which may provide better separability for classification. The classifiers such as k-nearest neighbor (k-NN), support vector machine (SVM), and deep belief network (DBN) are studied [15–17]. The classifiers from simple to complex ones have been explored to examine their capability for the task of voicing classification and the classifier that performs best for using these features.

1.1. Contributions of this paper

The main contributions of the work are:

- To show the significance of different glottal activity features for voicing decision.
- Improving the voicing decision from glottal activity features like strength of excitation (SoE), normalized autocorrelation peak strength (NAPS), and higher-order statistics (HOS) using classifiers such as k-NN, DBN, and SVM.
- Modeling the glottal activity features SoE, NAPS, and HOS as a separate stream in HMM with the continuous distribution.
- Using the voicing decision for HMM and DNN synthesis framework to improve the naturalness of synthetic speech.

This paper is organized as follows: Section 2 gives an overview of voicing decision in the context of SPSS. The issues present in conventional voiced/unvoiced decision used in SPSS and behavior of glottal activity features for a better voicing decision are described in Section 3. The refinement of voicing decision using different classifiers is described in Section 4. The integration of proposed voicing decision for SPSS is explained in Section 5. The experimental evaluation and effectiveness of proposed voicing classification for synthesis is described in Section 6. The work is finally concluded in Section 7.

2. Related literature

2.1. Voicing decision

In the literature, different features have been used for voicing decision. They can be classified into time-domain and frequency-domain features. Typically, time-domain features measure the acoustic nature of voiced sound such as energy, periodicity, zero crossing rate, and short-term correlation [18,19]. Frequency-domain features include the spectral peaks and harmonic measure by decomposing the speech signal using the Fourier transform or wavelet transform [20,21]. Besides, various refinements have been introduced for voicing decision from degraded speech recorded in realistic conditions. These cover extraction of voicing decision in adverse situations by utilizing autocorrelation of the Hilbert envelope of linear prediction (LP) residual [22,23]. Further, some of the recent algorithms like YIN, PRAAT, Get_F0, SWIPE etc. gave very good voicing estimation [24–27]. Also, some of the algorithms focused on the instantaneous F0 estimation, which in turn gives voicing decision-based on the event processing [19,28]. In all these methods, voicing decisions are taken by some threshold which was chosen empirically. The performance of these methods depends critically on the threshold. To improve the accuracy and to avoid threshold, statistical methods such as HMM, Gaussian mixture model (GMM), neural network model, and deep neural network also used for voicing decision [17,18,29]. These methods do not depend on the threshold. However, they require discriminative features for training.

2.2. Voicing decision in HMM

In the existing literature of HMM, voicing decision along with fundamental frequency (F0) is modeled as Multi-space distribution (MSD) [30]. F0 is modeled as a continuous Gaussian distribution in the voiced region and discrete symbols in the unvoiced region. It is modeled for a state (S) as follows

$$P(F_+ = F|L, S) = \begin{cases} \mathcal{N}(F; \mu_S, \sigma_S), & L = \mathrm{V} \\ 0, & L = \mathrm{U} \end{cases} \tag{1}$$

$$P(F_+ = \mathrm{NULL}|L, S) = \begin{cases} 0, & L = \mathrm{V} \\ 1, & L = \mathrm{U} \end{cases} \tag{2}$$

where $\mathcal{N}$ is a Gaussian density with mean μ_S and variance σ_S, $F \in (-\infty, \infty)$ represents the real F0 value, and $L \in \{\mathrm{U}, \mathrm{V}\}$ is the voicing label. According to the hypothesis of continuous F0, the voicing label V and the NULL symbol value cannot be observed at the same time. Similarly, the unvoicing label U and the real F0 value cannot occur together. F0 model using MSD can provide good quality speech if the voicing decision is accurate. However, there are failures like voiced region being classified as unvoiced region (false unvoiced) and gives hoarseness to voice quality. Sometimes, there is also a case wherein unvoiced region is classified as a voiced region (false voiced) providing buzziness, mainly in a higher frequency of synthesized speech [31].

Another major approach to model F0 and voicing labels for HMM is using continuous F0 model, instead of doing discontinuous MSD F0 model [32,33]. In continuous model, F0 observations are assumed to be always available, and the voicing decision modeled separately. Yu et al. [32] modeled F0 and voicing labels independently in separate streams to make the voicing decision independent of F0. In globally tied distribution method [33], pitch values for unvoiced frames are extended from neighboring voiced frames by interpolation and smoothing. Hence, F0 and its derivatives are modeled in a single stream. Besides, two mixture GMM for each state is modeled to represent one mixture for voiced frames, and