A Waveform Representation Framework for High-quality Statistical Parametric Speech Synthesis

Bo Fan*, Siu Wa Lee†, Xiaohai Tian‡§, Lei Xie* and Minghui Dong†
* School of Computer Science, Northwestern Polytechnical University, Xi'an, China
† Human Language Technology Department, Institute for Infocomm Research, Singapore
‡ School of Computer Engineering, Nanyang Technological University (NTU), Singapore
§ Joint NTU-UBC Research Center of Excellence in Active Living for the Elderly, NTU, Singapore
Email: {bofan,lxie}@nwpu-aslp.org, {swylee,mhdong}@i2r.a-star.edu.sg, xhtian@ntu.edu.sg

***Abstract*—State-of-the-art statistical parametric speech synthesis (SPSS) generally uses a vocoder to represent speech signals and parameterize them into features for subsequent modeling. Magnitude spectrum has been a dominant feature over the years. Although perceptual studies have shown that phase spectrum is essential to the quality of synthesized speech, it is often ignored by using a minimum phase filter during synthesis and the speech quality suffers. To bypass this bottleneck in vocoded speech, this paper proposes a phase-embedded waveform representation framework and establishes a magnitude-phase joint modeling platform for high-quality SPSS. Our experiments on waveform reconstruction show that the performance is better than that of the widely-used STRAIGHT. Furthermore, the proposed modeling and synthesis platform outperforms a leading-edge, vocoded, deep bidirectional long short-term memory recurrent neural network (DBLSTM-RNN)-based baseline system in various objective evaluation metrics conducted.**

I. Introduction

Statistical parametric speech synthesis (SPSS) has been increasingly popular due to its compact and flexible representation of voice characteristics [1]. Conventionally, in an SPSS system, we firstly extract parametric representations of speech including spectral and excitation parameters from a speech database and then model them with a set of models [2]. Several statistical generative models have been applied to SPSS successfully, e.g., hidden Markov model (HMM)-based SPSS [2], deep neural network (DNN)-based SPSS [3] and deep bidirectional long short-term memory recurrent neural network (DBLSTM-RNN)-based SPSS [4].

To parameterize speech signals into features for subsequent synthesis processes, vocoder has been typically used. It is based on the source-filter model [5], which assumes a stationary speech segment is generated by passing a sound source through a vocal tract filter. By using a vocoder, the resultant speech features are regular and suitable for modeling. However, in [6], their subjective listening test shows clear degradation of quality in vocoded speech. It further indicates that the source and filter parameters have to be jointly modelled for high-quality synthesis. Besides, to assure interframe coherence [7], a minimum phase hypothesis [7] has been used in most vocoders, which ignores the natural mixed-phase characteristics of speech signals, resulting in apparent degradation of the speech waveform quality.

More and more works have reported the importance of phase information in different speech processing applications, such as speech synthesis [8, 9], iterative signal reconstruction [10], automatic speech recognition [11, 12], speech coding [13] and pitch extraction [14]. Paliwal et al. [15] have investigated the relative importance of short-time magnitude and phase spectra on speech perception through human perception listening test. Results show that phase spectrum clearly contributes to the speech intelligibility. Sometimes its contribution is as much as the magnitude spectrum. Koutsogiannaki et al. [16] have proposed the phase distortion deviation feature, enabling to capture voice irregularities and highlights the importance of the phase spectrum in voice quality assessment. These two works indicate that phase information is important for both human perception and voice quality assessment. Combining phase spectrum with magnitude spectrum in frequency domain is equivalent to the speech waveform in time-domain. Therefore, the phase information is focused in our speech waveform representation framework.

There are some approaches of waveform representation directly in the time domain. Time domain pitch-synchronous overlap-add (TD-PSOLA) [17] performs pitch-synchronous analysis, modification and synthesis. During synthesis, speech frames are summed up. The quality of the reconstructed waveform with typical pitch or timing modification is similar to that of the original waveform. Multi-band re-synthesis pitch synchronous overlap add (MBR-PSOLA) [18] comments TD-PSOLA with three mismatches: phase mismatch, pitch mismatch, spectral envelope mismatch. It further suggests to solve these mismatches by re-synthesizing voiced parts of the speech database with constant phase and constant pitch. The artificial processing in MBR-PSOLA decreases the quality of speech and leads to buzzy sound [19]. Alternatively, there are a few recent works for SPSS directly in the time domain. Tokuda et al. [20] have proposed an approach to model cepstral coefficients to approximate the speech waveform. In their framework, periodic, voiced components have not been properly generated yet. In [21], complex cepstrum has been used to embed phase information for hidden semi-Markov models (HSMM) speech modelling.

In this paper, we propose a phase-embedded waveform representation framework, and establish a magnitude-phase

joint modeling platform for SPSS. This work uses glottal-synchronous overlap add approach for speech analysis and synthesis where glottal closure instants (GCIs) are employed. GCIs refer to the moments of most significant excitation that occur at the level of the vocal folds during each glottal period [22]. Short-term segments are defined as any two consecutive GCI periods. In order to produce smooth trajectories of our features which are required in SPSS, we design a cost function with a global smoothness constraint. The GCI locations selected are finally determined by conducting dynamic programming over a list of probable GCI candidates. Consequently, these segments will be very regular with stable magnitude and matched phase spectrum. With this waveform representation framework, the bottleneck suffered from vocoded speech is thus bypassed. This framework is hence capable of delivering better quality speech over the vocoded speech. Then we propose an approach for magnitude-phase joint spectrum modeling. Full spectrum is used in this framework, which is in line with the satisfactory performance in recent deep learning-based TTS [23]. To leverage on the modeling power of deep learning, we use DBLSTM-RNN to learn magnitude and phase spectrum simultaneously. Bidirectional recurrent connections can fully exploit the speech contextual information in both forward and backward directions. With purpose-built memory cells to store information, the long short-term memory (LSTM) architecture does better in finding and taking advantage of the long range context.

II. TD-PSOLA

Time domain pitch-synchronous overlap add (TD-PSOLA) is used for pitch and timing modification of speech signals [17], [24]. It is also popular for concatenation-based TTS. As no source-filter decomposition or vocoding is performed, the quality of resultant speech after analysis and reconstruction is highly similar to the original speech.

Given an arbitrary speech waveform signal $x(n)$, TD-PSOLA is carried out in the time domain. It first decomposes $x(n)$ into a sequence of overlapping, pitch-synchronized segments. Each segment $x_s(n)$ lasts for two pitch periods, running from a pitch period before and another pitch period after the segment centre. Then a window function $h_s(n)$, such as hanning window, will be applied to each segment. Assuming S denotes the total number of the segments, where $s = 1, 2, ..., S$,

$$x_s(n) = h_s(n)x(n) \tag{1}$$

$h_s(n)$ is non-zero during the above two-pitch period. This is how $x_s(n)$ is extracted for voiced speech; for unvoiced speech, the segment length is set to a constant. Any modification in pitch or timing can then be performed on these extracted segments. Finally, modified segments are overlapped and added to produce the speech output [24].

Although TD-PSOLA generates pitch- and timing-modified output signals with satisfactory speech quality, using TD-PSOLA in speech synthesis where statistical averaging, modeling or signal modification are common, is not sufficient. This is because matched attributes on phase and pitch are needed [18].

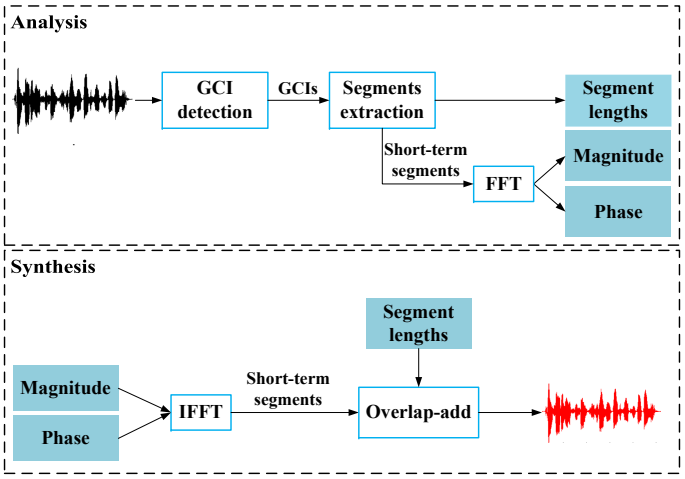

Fig. 1. Our proposed waveform representation framework.

III. Waveform Representation Framework

In this work, a glottal-synchronous based waveform representation framework is proposed for speech modelling. Similar to TD-PSOLA, glottal closure instants (GCIs) represent both the pitch contours and the boundaries of individual cycles of speech. Existing GCI detection approaches generally estimate the GCI locations in a local manner, ignoring the resultant trajectories of various acoustic attributes, i.e. segment length (representing fundamental frequency ($F0$)), magnitude and phase spectrum, exhibited in the utterance. As smooth trajectories of these attributes are necessary for SPSS, we revise a state-of-the-art GCI detection approach, so as to facilitate satisfactory modelling of these attributes.

A. System Overview

The proposed framework, as shown in Fig. 1, consists of two parts: analysis and synthesis. In the analysis stage, given an arbitrary waveform, firstly, the GCI locations are detected by the following revised GCI detection module. Then, the waveform is decomposed into overlapping short-term segments. Each segment is defined by any two consecutive GCI periods. Finally, segment lengths, magnitude and phase spectrum are used to represent these segments.

In the synthesis stage, given corresponding segment lengths, magnitude and phase spectrum, we convert them into overlapping short-term segments. Then, the waveform is reconstructed using the similar technique as TD-PSOLA [17].

B. Glottal Closure Instant Detection

The GCI positions determine the features including segment lengths, magnitude and phase spectrum. Thus, the GCI detection method is of great importance.

Among the present GCIs detection techniques, the Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) algorithm [25] is widely used. In

[26], SEDREAMS was shown to have the highest robustness and reliability. During the detection, SEDREAMS outputs only one GCI location for each GCI segment [25]. This is a local estimation process, without considering the GCI detection results in the neighborhood. However, SPSS requires smooth trajectories of speech features, which are defined once GCI locations are determined. By considering lists of probable GCI candidates and estimating the optimal GCI locations in a global manner, the trajectories of these features are stabilized.

Based on SEDREAMS, our modified GCI detection method contains the following steps:

a) Given a waveform $x(n)$ (Fig. 2(a)), calculate the moving average signal (Fig. 2(b)).

b) Determine the intervals for possible GCI locations[1] (Fig. 2(c)).

c) M candidates are chosen, based on the top M highest linear predictive coding (LPC) residual values in the LPC residual signal (Fig. 2(d)), as the possible GCI locations in each interval. Suppose there are N intervals, the k-th candidate of i-th interval denoted as $g_{i,k}$.

d) Transfer all the possible segment lengths into $F0$. For the i-th segment, the j-th $F0$ is expressed as

$$F0_{i,j} = Fs/(g_{(i+1),s} - g_{i,t}), \tag{2}$$

where F_s is the sampling frequency, $i = 1, 2, ..., N$, $j = 1, 2, ..., M \times M$, $s = 1, 2, ..., M$ and $t = 1, 2, ..., M$.

e) Given the reference $F0^{\text{ref}}$, the optimal segment lengths are determined by dynamic programming with the following constraint,

$$E = \arg\min_{j} \sum_{i=1}^{N} \left\| F0^{\text{ref}} - F0_{i,j} \right\|. \tag{3}$$

f) Finally, the GCI locations are deduced accordingly (Fig. 2(e)).

In our implementation, M is five and the reference $F0$ is extracted by STRAIGHT [27]. STRAIGHT is robust for $F0$ tracking and can generate a highly accurate and smooth $F0$ trajectory. The $F0$ trajectory extracted from STRAIGHT is robust The dynamic programming process is implemented by the Viterbi algorithm. In a voiced segment, the pitch located in the middle is more stable compared to the rest. Consequently, Viterbi search starts at this middle position to both ends.

A comparison of $F0$ trajectory between our GCI detection and SEDREAMS is depicted in Fig. 3. From Fig. 3(a), it is observed that the $F0$ given by our GCI detection is smoother than the one from SEDREAMS. And from Fig. 3(b), it is clear that our GCI detection approach removes some abnormal jumps (around the 247-th frame) of the $F0$ trajectory occurred in the SEDREAMS.

IV. Waveform Modeling

State-of-the-art SPSS usually models the magnitude spectrum of speech signals and discards the phase spectrum.

[1]For the detailed implementations of the moving average filter and interval determination, please refer to [25]

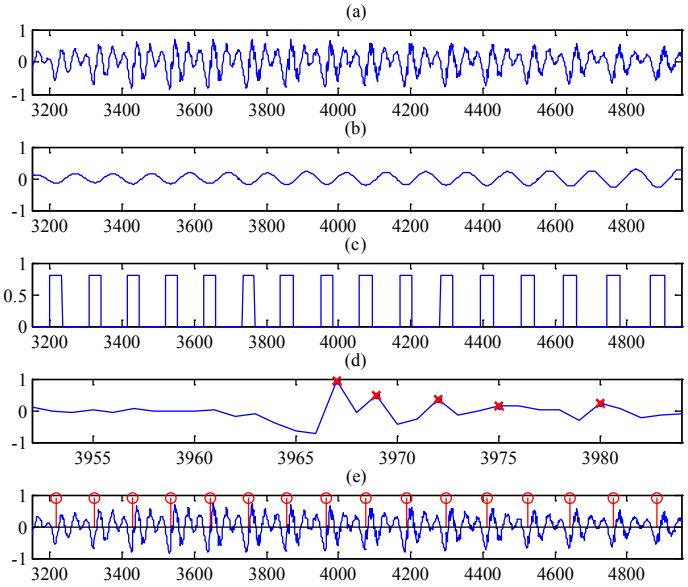

Fig. 2. (a) A section of voiced waveform; (b) The corresponding moving average signal; (c) Short intervals in the moving average signal; (d) LPC residual signal in one interval with candidates marked with red cross; (e) The final GCI locations marked with red stem.

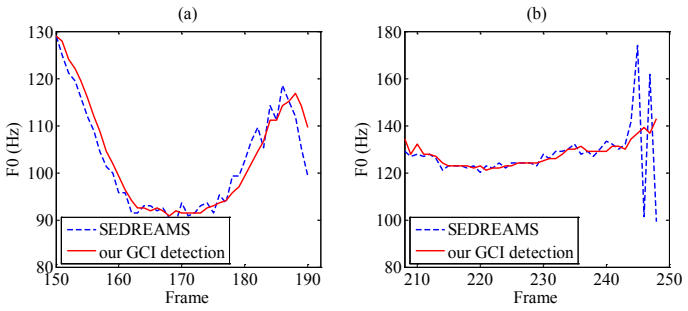

Fig. 3. A comparison of $F0$ trajectory between our GCI detection and SEDREAMS. (a) and (b) are two segments in the voiced parts.

During synthesis, a vocoder based on minimum-phase or zero-phase filter is often used together with the generated magnitude spectra to produce the synthesized output. Nevertheless, phase spectrum has been recently found to be essential for speech perception. The speech quality of vocoded outputs are found to be degraded from the original speech recordings [6]. This may shed light on SPSS, where speech waveform with phase information in addition to the existing magnitude spectrum, is modeled.

In our work, speech signals are modeled by the corresponding magnitude and phase spectra, without the use of a vocoder. Consequently, reconstruction of speech waveform is facilitated. We use a recently-emerging learning technique, DBLSTM-RNN, to jointly model the two spectra. DBLSTM-RNN is well-suited for learning sequential events apart from long time lags of unknown size [28]. Promising performance in various speech applications is observed [29], [4].

Our joint model of magnitude and phase is constructed

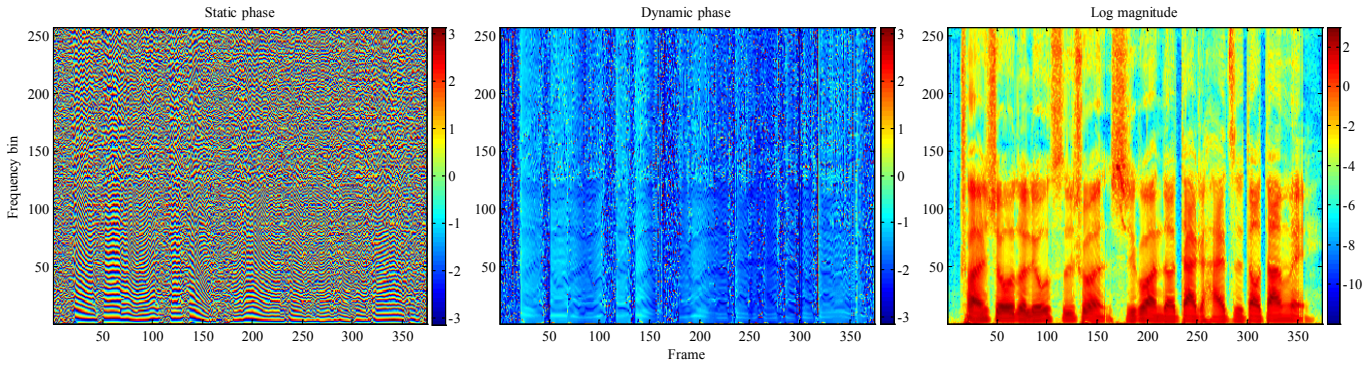

Fig. 4. Static and dynamic phase spectrum as well as their corresponding magnitude spectrum.

as follows. We employ line spectrum pair (LSP) as the feature representation of magnitude spectrum. LSP, being an alternative LPC spectral representation, is robust and suitable for interpolation and modeling [30], [31].

For phase spectrum, we propose to use the dynamic phase spectrum for this waveform learning TTS framework. It is also called group delay: the group delay $\tau_k(n)$ at time n and frequency bin k is calculated as the frequency derivative of the instantaneous phase $\theta_k(n)$, i.e.

$$\tau_k(n) = \theta_k(n) - \theta_{k-1}(n). \quad (4)$$

To enable reconstruction of the phase spectrum after DBLSTM-RNN modeling, the instantaneous phase at the first frequency bin is kept, together with the group delays of the remaining frequency bins. In other words, our phase representation consists of $\theta_1(n), \tau_2(n), \tau_3(n), ..., \tau_K(n)$, where K is the total number of frequency bins.

This group-delay-based phase representation is found to be stable and facilitates statistical modeling in subsequent TTS process, as shown in Fig. 4. Comparing the spectra of static phase and dynamic phase, the distribution of the dynamic phase often exhibits a smaller range. Comparing the log magnitude spectrum with the dynamic phase spectrum, patterns of voiced and unvoiced portions are consistent and spectral patterns of individual speech sounds are quite similar in the log magnitude spectrum and the dynamic phase spectrum. This is important and useful for our joint modeling. On the contrary, there is no clear difference in the static phase spectrum for individual speech sounds. When moving along the time-axis, the static phase spectra look like the same.

V. Experiments

We conducted two experiments to assess the efficacy of our waveform representation framework. In the experiment on waveform reconstruction, objective and subjective evaluations were carried out to compare the performance between our framework and other three vocoders: STRAIGHT, Tandem-STRAIGHT [32] and AHOCoder [33] respectively. As we know, STRAIGHT is a very popular vocoder used for speech analysis and reconstruction, and Tandem-STRAIGHT is the upgrade version of STRAIGHT. AHOCoder is reported to be of similar quality compared with STRAIGHT. In the experiment on waveform modeling, we trained a text-to-speech (TTS) system based on our framework and also a baseline TTS system [4] as a comparison. This baseline is a leading-edge approach based on DBLSTM-RNN and generates high-quality synthesized speech. It uses STRAIGHT as its vocoder.

A corpus with 4,936 Chinese utterances (around 6 hours) spoken by a native male speaker in a neutral style was used in our experiments. Speech waveform signals are sampled at 16kHz. The contextual labels are both phonetically and prosodically rich, including quin-phone, prosody, tone and syllable information. For TTS systems, the training, validation and test data consist of 3,949, 494 and 493 utterances, respectively.

A. Experiment on Waveform Reconstruction

Speech waveform in the test set of the corpus was analyzed and re-synthesized using our waveform representation framework and the three vocoders. The reconstructed speech waveform was then used for objective and subjective evaluations.

1) Objective Evaluation: In the objective evaluation, we calculated the root mean square error (RMSE) between the reconstructed and original speech waveform signals in the voiced parts (RMSE_voiced), the unvoiced parts (RMSE_unvoiced) and the entire waveform (RMSE), respectively. The results are shown in Table I. These voiced/unvoiced results from our framework and the three vocoders generally represent the performance on vowels/consonants respectively.

TABLE I
RECONSTRUCTION PERFORMANCE: OUR FRAMEWORK VS. THE THREE VOCODERS

Methods \ Measures	RMSE_voiced	RMSE_unvoiced	RMSE
Our framework_voiced	**0.026**	**0.042**	**0.031**
STRAIGHT [27]	0.173	0.044	0.152
Tandem-STRAIGHT [32]	0.177	0.044	0.156
AHOCoder [33]	0.182	0.049	0.160

The objective evaluation result shows that the performance of our framework is much better than that of the three vocoders

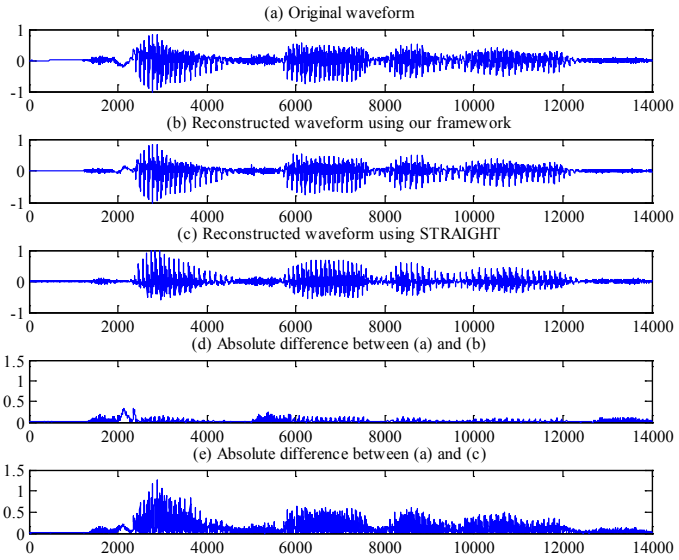

Fig. 5. The reconstructed waveform using our framework and STRAIGHT.

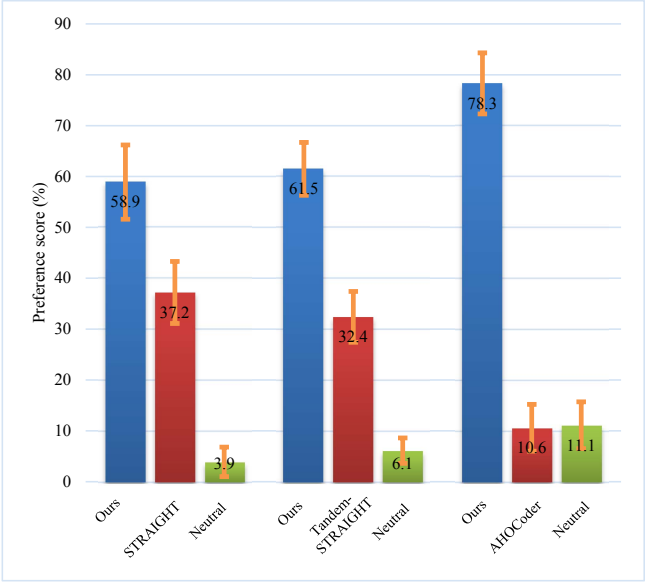

Fig. 6. The ABX result of the reconstructed speech waveform using our framework and the three vocoders. We conducted t-test using a significance level of $p < 0.05$ which is depicted with the error bars in yellow.

especially in the voiced parts. The short-term segments are extracted at a constant rate in the unvoiced parts from our framework which is similar to STRAIGHT. Taking the waveform in Fig. 5 around the 5000-th sample as an example, the absolute difference between (a) and (b) is very close to that between (a) and (c). In the voiced parts, our framework performs much better than STRAIGHT does. It is because our framework retains the full phase spectrum, while STRAIGHT discards it and uses a minimum-phase setting instead. We can see clearly from Fig. 5 that the absolute difference between (a) and (b) is much smaller than that between (a) and (c) in the voiced parts.

2) Subjective Evaluation: 20 pairs of speech waveform are randomly selected from the reconstructed waveforms. Then a group of 20 subjects were asked to perform the ABX preference test. We put the original waveform into **X**, while we put the waveform reconstructed using our framework and each of the three vocoders into **A** and **B** randomly. Each subject was asked to answer which one(**A** or **B**) is more similar to **X**. The third option **Neutral** means the subject has no preference on A or B. The ABX result is shown in Fig. 6. We can clearly see that the reconstructed speech waveform using our framework is significantly preferred as compared with all of the three vocoders.

B. Experiment on Waveform Modeling

In the baseline DBLSTM-RNN-based TTS [4], STRAIGHT is used to vocode the speech waveform by a 25-ms moving window, and shifted every 5-ms. The generated magnitude spectrum from STRAIGHT was converted into LSP. The dimensionality of the input contextual label is 427. The output feature contains voiced/unvoiced flag (1 dimension), log F0 (1 dimension), LSP (40 dimensions) and gain (1 dimension), totally 43 dimensions. As suggested in [4], a neural network with two BLSTM layers sitting on two feed forward layers with 256 nodes in each layer is employed to train the DBLSTM-RNN-based TTS.

For our TTS system, features were extracted from the short-term segments specified by GCI locations. The format of the input label is the same as the baseline. The segment length is transformed into $F0$. The output feature comprises several components: voice/unvoiced flag (1 dimension), log $F0$ (1 dimension), LSP (40 dimensions), gain (1 dimension) and dynamic phase feature (257 dimensions), totally 300 dimensions. The same network topology as baseline is used to train our TTS system.

To evaluate the performance of these two TTS systems, five metrics are used for objective evaluation:

- RMSE_$F0$: root mean square error in $F0$ estimation;
- Voiced/unvoiced (V/U) error rate;
- Log spectral distance (LSD):

$$LSD(\mathbf{S}_p, \mathbf{S}_g) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\sum_{k=1}^{M_s}[10log_{10}\, s_p(j,k) - 10log_{10}\, s_g(j,k)]^2)}, \quad (5)$$

  where $\mathbf{S}_p$ and $\mathbf{S}_g$ are the predicted and ground-truth magnitude spectrum, respectively. N is the total number of frames in the voiced parts and M_s refers to the dimensionality of magnitude spectrum. $s_p(j,k)$ is the the k-th value of magnitude in j-th frame;
- Mel cepstral distance (MCD):

$$MCD(\mathbf{c}_p, \mathbf{c}_g) = \frac{10}{ln\,10}\sqrt{2\sum_{k=1}^{M_c}[c_p(k) - c_g(k)]^2}, \quad (6)$$

where $\mathbf{c}_p$ and $\mathbf{c}_g$ are the predicted and ground-truth Mel cepstrum coefficient vectors, respectively, and M_c refers to the dimensionality of Mel cepstrum coefficients;

- Dynamic phase distance (DPD):

$$DPD(\mathbf{d}_p, \mathbf{d}_g) = \sqrt{\sum_{k=1}^{M_d} [d_p(k) - d_g(k)]^2}, \qquad (7)$$

where $\mathbf{d}_p$ and $\mathbf{d}_g$ are the predicted and ground-truth dynamic phase feature vectors, respectively, and M_d refers to the dimensionality of the dynamic phase feature.

The synthesized speech waveform from the labels in the test set uses the ground-truth durations. These five metrics are calculated at the GCIs level, i.e., the short-term segments are specified by the GCIs locations. In order to make the systems comparable, GCI detection is required for all speech waveforms synthesized from any system under comparison. And after the GCI detection, it should be aligned to the ground-truth GCIs by finding out the closest one.

The objective evaluation result is shown in Table II. It shows that our TTS system is better than the baseline in terms of all the five metrics. In particular, for DPD, the average absolute difference in one frequency bin is about 0.70rad in our TTS system while 0.91rad for the baseline TTS system.

TABLE II
OBJECTIVE EVALUATION ON WAVEFORM MODELING WITH t-TEST USING A SIGNIFICANCE LEVEL OF $p < 0.05$.

Methods / Measures	Our TTS system	Baseline [4]
RMSE_F0 (Hz)	**23.6 ± 1.2**	28.0 ± 2.3
V/U error rate (%)	**5.9 ± 0.6**	8.6 ± 1.0
LSD (dB)	**59.2 ± 0.8**	63.9 ± 1.1
MCD (dB)	**4.5 ± 0.1**	4.8 ± 0.1
DPD (rad)	**11.4 ± 0.2**	14.7 ± 0.3

VI. CONCLUSIONS AND FUTURE WORK

This paper proposed a glottal-synchronous based waveform representation framework for high-quality statistical parametric speech synthesis. Speech signal was represented by magnitude and phase full-spectral components, without the the use of a vocoder. We revised the SEDREAMS GCI detection approach to improve the feature stability for statistical modelling. Both objective and subjective evaluations were conducted to assess the reconstruction performance of our framework. Results indicate that, comparing to the reconstructed signal obtained by three popular vocoders, the proposed framework achieves promising results in RMSE in time domain speech waveform and preference score.

We also proposed a platform for speech modelling. DBLSTM-RNN is applied to jointly model the corresponding magnitude and phase spectra, and group delay-based phase representation is used to facilitate statistical modelling. Objective results show that, the TTS system based on the proposed framework generates the features, specifically the phase feature, with lower distortion as compared with a vocoder based system. Further works include studying the speech quality of synthesized speech and the associated factors and experiments on subjective evaluation.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (61175018 and 61571363).

REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

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

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*. IEEE, 2013, pp. 7962–7966.

[4] Y. Fan, Y. Qian, F. Xie, and F. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. Interspeech*, 2014, pp. 1964–1968.

[5] G. Fant, *Acoustic Theory of Speech Production*, The Hague: Mouton, 1960.

[6] T. Merritt, T. Raitio, and S. King, "Investigating source and filter contributions, and their interaction, to statistical parametric speech synthesis," in *Proc. Interspeech*, 2014, pp. 1509–1513.

[7] M. Crespo, P. Velasco, L. Serrano, and J. Sardina, "On the use of a sinusoidal model for speech synthesis in text-to-speech," in *Progress in Speech Synthesis*, pp. 57–70. Springer, 1997.

[8] H. Banno, K. Takeda, and F. Itakura, "The effect of group delay spectrum on timbre," *Acoustical Science and Technology*, vol. 23, no. 1, pp. 1–9, 2002.

[9] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1080–1090, 2011.

[10] L. Alsteris and K. Paliwal, "Iterative reconstruction of speech from short-time Fourier transform phase and magnitude spectra," *Computer Speech & Language*, vol. 21, no. 1, pp. 174–186, 2007.

[11] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proc. ICASSP*. IEEE, 2001, vol. 1, pp. 133–136.

[12] G. Shi, M. Shanechi, and P. Aarabi, "On the importance of phase in human speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1867–1874, 2006.

[13] H. Pobloth and W. Kleijn, "Squared error as a measure of perceived phase distortion," *The Journal of the Acoustical Society of America*, vol. 114, no. 2, pp. 1081–1094, 2003.

[14] T. Nakatani, T. Irino, and P. Zolfaghari, "Dominance spectrum based v/uv classification and F0 estimation," in *Proc. Eurospeech*, 2003, pp. 2313–2316.

[15] K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. Eurospeech*, 2003, pp. 2117–2120.

[16] M. Koutsogiannaki, O. Simantiraki, G. Degottex, and Y. Stylianou, "The importance of phase on voice quality assessment," in *Proc. Interspeech*, 2014, pp. 1653–1657.

[17] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5, pp. 453–467, 1990.

[18] T. Dutoit and H. Leich, "MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database," *Speech Communication*, vol. 13, no. 3, pp. 435–440, 1993.

[19] Y. Stylianou, "Removing linear phase mismatches in concatenative speech synthesis," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 232–239, 2001.

[20] K. Tokuda and H. Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," in *Proc. ICASSP*, 2015, pp. 4215–4219.

[21] R. Maia, M. Akamine, and M. Gales, "Complex cepstrum as phase information in statistical parametric speech synthesis," in *Proc. ICASSP*. IEEE, 2012, pp. 4581–4584.

[22] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 5, pp. 325–333, 1995.

[23] Z. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2129–2139, 2013.

[24] P. Taylor, *Text-to-Speech Synthesis*, United Kingdom: University of Cambridge, 2007.

[25] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. Interspeech*, 2009, pp. 2891–2894.

[26] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: a quantitative review," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 994–1006, 2012.

[27] 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.

[28] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer, 2012.

[29] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*. IEEE, 2013, pp. 6645–6649.

[30] F. K. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. ICASSP*. IEEE, 1984, pp. 37–40.

[31] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, pp. S35, 1975.

[32] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proc. ICASSP*, 2008, pp. 3933–3936.

[33] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, 2014.