Reducing mismatch in training of DNN-based glottal excitation models in a statistical parametric text-to-speech system

Citation for published version:

Digital Object Identifier (DOI):
10.21437/Interspeech.2017-848

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings Interspeech 2017

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Reducing mismatch in training of DNN-based glottal excitation models in a statistical parametric text-to-speech system

Lauri Juvela¹, Bajibabu Bollepalli³, Junichi Yamagishi²,³, Paavo Alku¹

¹Aalto University, Department of Signal Processing and Acoustics, Finland
²National Institute of Informatics, Japan
³The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
{lauri.juvela,bajibabu.bollepalli,paavo.alku}@aalto.fi, jyamagis@nii.ac.jp

Abstract

Neural network-based models that generate glottal excitation waveforms from acoustic features have been found to give improved quality in statistical parametric speech synthesis. Until now, however, these models have been trained separately from the acoustic model. This creates mismatch between training and synthesis, as the synthesized acoustic features used for the excitation model input differ from the original inputs, with which the model was trained on. Furthermore, due to the errors in predicting the vocal tract filter, the original excitation waveforms do not provide perfect reconstruction of the speech waveform even if predicted without error. To address these issues and to make the excitation model more robust against errors in acoustic modeling, this paper proposes two modifications to the excitation model training scheme. First, the excitation model is trained in a connected manner, with inputs generated by the acoustic model. Second, the target glottal waveforms are re-estimated by performing glottal inverse filtering with the predicted vocal tract filters. The results show that both of these modifications improve performance measured in MSE and MFCC distortion, and slightly improve the subjective quality of the synthetic speech.

Index Terms: statistical parametric speech synthesis, excitation modeling

1. Introduction

Statistical parametric speech synthesis (SPSS) ¹ has gained popularity due to its several favorable properties, such as good generalization to unseen text, flexible model adaptation to new speakers with relatively small amount of data, and small runtime resource requirements compared to unit-selection synthesis ². The emergence of the neural network-based acoustic models has improved the quality of SPSS systems ³, ⁴, and the use of sequence models, such as Long Short-Term Memory (LSTM) networks ⁵, ⁶ has become widely adopted in SPSS.

Traditionally, the back-end of a SPSS system consists of two separate parts: speech parametrization and waveform generation are done with a vocoder, such as STRAIGHT ⁷, and an acoustic model is trained to map linguistic features onto the speech parameters. In conjunction, the two major issues in SPSS, over-smoothness of the generated acoustic parameters, and "buzzy" synthetic sound quality have been attributed to the acoustic model and vocoder, respectively. Recent efforts have improved the acoustic model performance resulting in more natural synthetic parameter trajectories using, for example, autoregressive mixture density networks ⁸, or generative adversarial network-based post-filtering ⁹. However, the performance of these systems is still upper-bounded by the analysis-synthesis quality of the vocoder.

Recently, there has been growing interest in joint optimization for the acoustic model and speech parametrization. One such approach was proposed in ¹⁰, ¹¹, where a network mapping text features to a cepstrum was trained directly on speech waveforms. Another viable approach to direct speech modeling operates in the spectrogram domain: deep auto encoders for spectral envelope prediction were proposed in ¹², and full spectrogram prediction with phase recovery-based synthesis was presented in ¹³. The emergence of advanced neural net-based waveform generation methods ¹⁴, ¹⁵ also seems to lead towards a closer integration between acoustic models and waveform synthesis. Indeed, success in training end-to-end speech synthesis systems with these methods has been reported ¹⁶, ¹⁷. Although connected end-to-end, these systems are still initially trained to map text to pre-estimated low-dimensional acoustic features, i.e. mel-generalized cepstrum (MGC) ¹⁸ and filter-bank mel-frequency cepstral coefficients (MFCC) ¹⁹. Overall, these two proposed end-to-end systems retain the general structure of parametric TTS systems, while all the system parts are now interconnected neural nets.

Training end-to-end neural TTS systems can be difficult, and they require considerable amounts of training data and computational resources. Furthermore, when a system is fine-tuned end-to-end, any initial intermediate representations, such as filter parameters, lose their interpretability and can not be used directly in signal processing. From the perspective of human voice production, the source-filter model is still relevant, and the explicit use of auto-regressive filters to model the vocal tract is powerful and widespread in many speech applications. The glottal source, the excitation of voiced speech, is a more elementary signal than the speech waveform itself because the glottal source is generated at the level of vocal folds and therefore does not include resonances of the vocal tract. Hence, the glottal excitation is an attractive domain for generative waveform modeling. Indeed, glottal excitation generation with neural nets has been applied successfully to TTS in e.g. ¹⁸, ¹⁹, while still using relatively lightweight networks for the task. With this approach, glottal pulse excitation waveforms can be generated from acoustic features with a simple feedforward net.

Until now, however, the models for generating glottal excitation waveforms have been trained separately from the acoustic model, and it is therefore not known whether combining the two into an end-to-end framework would benefit the synthesis quality. Specifically, the separate training of excitation and acoustic models leads to mismatch between training and synthesis at both the excitation model input and the desired output. This paper addresses the mismatch by proposing connected training of the system: first the acoustic model is trained normally and
fixed, after which the excitation model training (both inputs and outputs) are changed to compensate for the errors in the acoustic model, as detailed in Section 2.2. Synthesis systems are trained with the suggested modifications for female and male voices in Section 3 and evaluated with objective measures and listening tests. Finally, conclusions are drawn in Section 4.

2. Synthesis system

This paper uses the same text features and the same acoustic model as in [20], while training of the glottal excitation model is modified. For details on using glottal pulses in DNN training, see [18]. Fig. 1 shows an overview of the synthesis system. There are three variants in training the excitation model (right side in figure): (1) the baseline training scheme corresponds to [20], whereas (2) uses generated acoustic inputs for training, and (3) additionally performs inverse filtering with generated vocal tract filters, as elaborated on in Sections 2.4 and 2.5 respectively. The acoustic model part (left side in figure) is shared in all the variants, and the variations marked with (1–3) only affect how the excitation model is trained.

![Diagram of the speech synthesis system](image)

**Figure 1: Overview of the speech synthesis system. The baseline approach (1) uses the original estimated vocal tract filter for both inverse filtering and training the excitation model, whereas systems (2) and (3) are trained with features generated by the acoustic model. Furthermore, the system (3) re-estimates the target waveforms by inverse filtering with the generated vocal tract filter.**

2.1. Text features

In this work, text features refer to the acoustic model neural network inputs, which are derived from full-context linguistic labels extracted from text. The full-context labels include phoneme, syllable, word, phrase, and sentence level information and are created using the Flite [21] speech synthesis frontend and the Comibilex [22] lexicon. The text and acoustic features are aligned with the HMM-based speech synthesis system (HTS) [23] and the resulting durations are combined with the labels to create 396-dimensional (per time-frame) text features.

2.2. Acoustic model

In the training phase, acoustic features are estimated from speech signal at a 5-ms frame rate similarly to [20]. The vocal tract filter is estimated with the Quasi Closed Phase (QCP) [24] algorithm, and parametrized with line spectral frequencies (LSFs). Fundamental frequency ($f_0$) and voicing decisions (VUV) are extracted with RAPT [25], and Reaper [26] is used for glottal closure instant (GCI) detection. Additionally, glottal source spectral envelope and harmonic-to-noise ratio (HNR) are estimated similarly to [27].

For training the acoustic model, deltas and delta-deltas of the acoustic features are included. The model is a bidirectional LSTM network that maps the frame-rate text features (see Section 2.1) to the dynamic acoustic features. The network configuration is given in Table 1. At synthesis stage, acoustic features are generated on given text inputs, after which the maximum likelihood parameter generation (MLPG) is applied to produce smooth feature trajectories. Additionally, the generated vocal tract parameters are post-filtered with the formant enhancement method presented in [28]. Voiced excitation signal is obtained from the excitation model by conditioning it on the generated acoustic features, while white noise excitation is used for unvoiced speech.

2.3. DNN-based excitation model

Glottal excitation pulses are processed by first performing inverse filtering, and then extracting a two pitch-period segment with one GCI in the middle and two at the both ends (see left side of Fig. 2 for illustration). These pulses are then cosine-windowed and zero-padded to a constant length of 400 samples. Pulses are associated with acoustic features by assigning the pulse nearest to the mid-point of the 5-ms frame, and unvoiced frames are excluded from the training sequences to accommodate LSTM. The network structure used for the excitation models is listed in Table 1. At synthesis time, generated acoustic features are fed into the excitation model, after which the generated pulses are windowed, scaled to the desired energy, and joined with pitch-synchronous overlap-add (PSOLA) [29]. In the baseline excitation model training scheme [18] [20], the original estimated acoustic features are used as input, and the corresponding original glottal pulses are set as the target outputs.

2.4. Training with generated acoustic features

The first issue to be observed in the excitation model training is the difference between how the model is trained and how it is used in synthesis: if the model is trained with the original acoustic features as the input, a mismatch occurs because only synthetic acoustic features will be available at test time. This problem is straightforward to fix by using the acoustic model to generate the training set acoustic features from the text input, and further feed them into the excitation model inputs at training time. This modification is denoted in Fig. 1 by number (2).

Since we want to use the acoustic model outputs for signal processing, we fix the acoustic model after training, and only alter the excitation model to retain connectivity at the input. The acoustic model will inevitably produce some errors, and the synthetic acoustic features will be corrupted versions of the
original ones. As a result, this mismatch may cause the wave-
form model to overfit the natural acoustic features, while gener-
alizing poorly with the synthetic acoustic input. This behavior
is illustrated in Fig. 2 (training details in section 3.2). The use
of generated acoustic features for excitation model input can
also been seen as a form of regularization, where training with
corrupted input helps to prevent overfitting.

2.5. Re-estimating glottal excitation waveforms
The second mismatch issue is related to the output of the
excitation model: since the acoustic model is fixed after training,
synthesized vocal tract filters will differ from the original (see
Fig. 2), and can no longer give perfect reconstruction even with
the original excitation waveform. However, we can re-estimate
the excitation model target waveforms by performing inverse
filtering with the predicted filter—this aims to compensate for
the errors made in acoustic modeling. A similar idea has been
adopted in speech coding, where the encoder uses quantized fil-
ter parameters (that are always utilized by the decoder), rather
than using the more accurate non-quantized filter parameters,
e.g. [30]. This change is denoted in Fig. 1 by number (3).

3. Experiments
3.1. Speech material
Two speaker-specific systems were trained for the experiments.
Both speakers (one male, one female) are professional British
English voice talents. The dataset for the male speaker “Nick”
comprises 2542 utterances, totaling 1.8 hours, and the dataset
for the female speaker “Jenny” comprises 4314 utterances, to-
taling 4.8 hours. For both speakers, 100 utterances were ran-
domly selected for both validation and testing, while the rest
were used for training the systems. The material was downsam-
ples to 16 kHz sample rate from the original 48 kHz rate.

3.2. Training
The acoustic model network consist of two feedforward (FF)
layers with logistic activation function, with two bidirectional
LSTM layers stacked on top of them. The model was trained
with stochastic gradient descent and early stopping was applied
after 5 epochs of no improvement on validation set. Dynamic
features were included for the acoustic model outputs. The
excitation models were trained similarly, but now only static acoustic features were used at the model input. This is
because we want to exactly match the domain of the acoustic

Table 1: Two types of networks are used in the system: acoustic model maps text features (TXT) to acoustic (AC) delta-features, while excitation model maps acoustic features to glottal pulse waveforms (GL).

<table>
<thead>
<tr>
<th>Network layer</th>
<th>Acoustic model</th>
<th>Excitation model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input (size)</td>
<td>TXT (396)</td>
<td>AC (47)</td>
</tr>
<tr>
<td>Hidden (size)</td>
<td>FF (512)</td>
<td>LSTM (128)</td>
</tr>
<tr>
<td></td>
<td>FF (512)</td>
<td>FF (512)</td>
</tr>
<tr>
<td>Output (size)</td>
<td>AC (142)</td>
<td>GL (400)</td>
</tr>
</tbody>
</table>

features being used at waveform synthesis, which for generated
features includes MLPG and the vocal tract formant enhance-
ment post-filter. To illustrate the mismatch issue, Fig. 3 shows
the excitation model training and test set errors as a function of
ePOCHS. Direct comparison is applicable between the two
systems which attempt to predict the original glottal waveform,
either from the original (AC-GL) or the generated (GEN-AC-
GL) acoustic features. In the first case, poor generalization on
test set is evident for both voices: excitation model fits tightly
to the original acoustic features of the training set, while the test
set performance does not improve. On the other hand, training
with the generated acoustic features does not reach as low train-
ing error, but performs considerably better on the test set. The
third system using re-estimated target waveforms was trained
with network structure and generated acoustic inputs similar to
GEN-AC-GL. The CURRENNT toolkit [31] was utilized for
training all the networks.

3.3. Objective evaluation
For objective evaluation, we aim to measure the difference be-
tween the target speech signal and the synthesised speech af-
ter filtering the generated excitation. Since the systems were
trained with the MSE criterion, the same metric should be used
for the evaluation. However, measuring point-wise errors di-
rectly on the final synthetic speech waveform is difficult, since
synchronism is broken in overlap-add due to the original and
generated pitch being different. Nevertheless, the final result
can be closely approximated by individually filtering the gener-
ated pulses frame-by-frame, as illustrated in Fig. 2. The target
waveform is created by filtering the original estimated glottal
pulse with the original vocal tract filter, whereas the generated
pulses from various systems are filtered with the filter given by
the acoustic model.

In addition to MSE, MFCC distortion is calculated, as this
measure is commonly used in speech applications and roughly
correlates with perceptual differences. The MFCCs were cal-
culated with a filterbank size of 24, 13 cepstrum coefficients were
used, and the distortion based on squared MFCC error is given
in decibels. Fig. 3 shows the average MSE and MFCC distor-
tions over the voiced frames in the test set. The trend is clear for
both voices: training with generated acoustic features reduces the
objective errors, and using re-estimated target waveforms
further improves performance.

3.4. Listening experiment
For the listening test samples, text features with forced align-
ment durations were used as the acoustic model input in or-
der to focus on the combined performance of the acoustic and
excitation generation models. An A-B preference test was per-
This study addresses two sources of mismatch that occur when glottal excitation models and acoustical models are trained.

4. Conclusions
This study addresses two sources of mismatch that occur when glottal excitation models and acoustical models are trained and used for parameter generation in TTS. The results show that training the excitation model with inputs generated by the acoustic model improved generalization to test set, reduced the objective error metrics and slightly improved the perceived quality of synthetic speech. Additionally, training with excitation waveforms re-estimated by glottal inverse filtering with the generated vocal tract filters further improved both objective and subjective performance.

Apart from the proposed training scheme, another way to reduce mismatch between original and generated acoustic features is to improve the acoustic model. Nevertheless, the ideas presented here still apply while the acoustic features are used to control generative neural models. The shortcomings in the acoustic model can be addressed in the future with the use of more advanced models, such as mixture density network LSTMs [8].

From a waveform generation perspective, the current DNN-based excitation models trained with the squared error criterion are inherently limited to generating conditional averages. As such, they cannot reproduce the stochastic properties desirable in excitation signals. Potential extensions to our excitation modeling framework include the use of more powerful generative methods, such as generative adversarial networks, or some form of the recent powerful sample-by-sample generative methods, e.g. [15, 14].

5. Acknowledgements
This work was supported by the Academy of Finland (proj. no. 284671) and MEXT KAKENHI Grant Numbers (26280066, 15H01686, 15K12071, 16H06302).
6. References


