# LOMBARD SPEECH SYNTHESIS USING LONG SHORT-TERM MEMORY RECURRENT NEURAL NETWORKS

*Bajibabu Bollepalli, Manu Airaksinen, Paavo Alku*

Department of Signal Processing and Acoustics, Aalto University, Finland

**ABSTRACT**

In statistical parametric speech synthesis (SPSS), a few studies have investigated the Lombard effect, specifically by using hidden Markov model (HMM) -based systems. Recently, artificial neural networks have demonstrated promising results in SPSS, specifically by using long short-term memory recurrent neural networks (LSTMs). The Lombard effect, however, has not been studied in the LSTM-based speech synthesis systems. In this study, we propose three methods for Lombard speech adaptation in LSTM-based speech synthesis. In particular, (1) we augment Lombard specific information with the linguistic features as input, (2) scale the hidden activations using the learning hidden unit contributions (LHUC) method, and (3) fine-tune the LSTMs trained on normal speech with a small Lombard speech data. To investigate the effectiveness of the proposed methods, we carry out experiments using small (10 utterances) and large (500 utterances) Lombard speech data. Experimental results confirm the adaptability of the LSTMs, and similarity tests show that the LSTMs can achieve significantly better adaptation performance than the HMMs in both small and large data conditions.

***Index Terms***— Lombard speech synthesis, adaptation, LSTM-TTS

## 1. INTRODUCTION

For seamless communication, humans involuntarily modify their voice based on the acoustic and auditory environment. The Lombard effect is a speaking style that humans typically take advantage of when talking in noisy surroundings. Speech produced in such conditions is called as Lombard speech or speech-in-noise. Compared to speech produced in quiet environment, Lombard speech shows increase in loudness (and intensity) and pitch, modification in spectral tilt, duration and prosody [1]. From now on, speech produced in quiet environment is referred to as normal speech in the current paper.

Despite vast progress in current text-to-speech (TTS) systems, the intelligibility of synthetic speech is typically below that of natural speech in noisy conditions [2]. Therefore, there is a great need for technologies to improve the intelligibility of synthetic speech, particularly by trying to incorporate the Lombard effect in a similar manner as natural talkers do in noisy environments. In this work, we study synthesis of Lombard speech by particularly focusing on various adaptation approaches to be used in artificial neural network -based speech synthesis.

## 2. RELATION TO PRIOR WORK

Intelligibility enhancement of synthetic speech in noise has been studied in a number of previous studies. Some of these studies have applied signal processing techniques on synthetic speech to mimic the acoustic changes observed in production of Lombard speech. The methods utilized are, for example, cepstral modification using the Glimpse proportion measure [3], as well as spectral shaping and dynamic range compression [4, 5]. These techniques do not require Lombard speech to modify the synthetic speech. However, only a few studies have explicitly used Lombard speech to enhance the intelligibility of synthetic speech by employing either voice conversion [6], or adaptation techniques [7, 8, 9]. These previous adaptation studies, however, are all based on statistical parametric speech synthesis (SPSS) systems utilizing hidden Markov model (HMM)-based speech synthesis, due to its adaptation abilities and flexibility in changing voice characteristics (e.g speaker, speaking style, and emotion state), and small memory footprint [10]. The HMMs trained on normal speech are adapted by a small amount of Lombard speech with the constrained structural maximum a posteriori linear regression combined with maximum a posteriori (CSMAPLR + MAP) adaptation technique [11]. These previous studies show that the intelligibility of synthetic speech generated by the Lombard adapted TTS system is significantly higher in noise environments than the corresponding synthetic speech generated from normal speech [2, 7, 12].

However, the naturalness of synthetic speech rendered through HMM-based synthesis system is not as good as that of the best samples from unit-selection speech synthesizers. This is mainly caused by three factors: 1) quality of vocoder, 2) accuracy of acoustic model, and 3) effect of over-smoothing. To address the issue No. 2), the use of deep neural networks (DNNs) has been proposed after their success in speech recognition. Study [13] has demonstrated that the quality of synthetic speech generated by DNNs is significantly better than that of HMM-based systems. One reason for the success of DNNs compared to HMMs is that they can provide a better and more efficient representation of complex dependencies between linguistic and acoustic features. To model the sequential nature of speech, the DNNs are extended to recurrent neural networks especially long short-term memory networks (LSTMs), which capture the correlations among consecutive frames [14, 15].

A few studies have explored DNNs for speaker adaptation in TTS [16, 17], despite the fact that DNNs have shown promising results in speaker adaptation in the area of speech recognition [18, 19]. In DNNs, the adaptation techniques have been applied at three different levels: at input level [16, 17, 19], at model level [16, 18], and at output level [16, 17]. The DNN-based speaker adapted systems outperformed the HMM-based systems in terms of naturalness and speaker similarity [16]. However, to the best of our knowledge, there are no previous studies on speaking style adaptation in DNN-based TTS.

In this work, we investigate the adaptation of LSTMs to a specific speaking style, Lombard speech. We study three LSTM adaptation methods. The first one relies on style specific information at

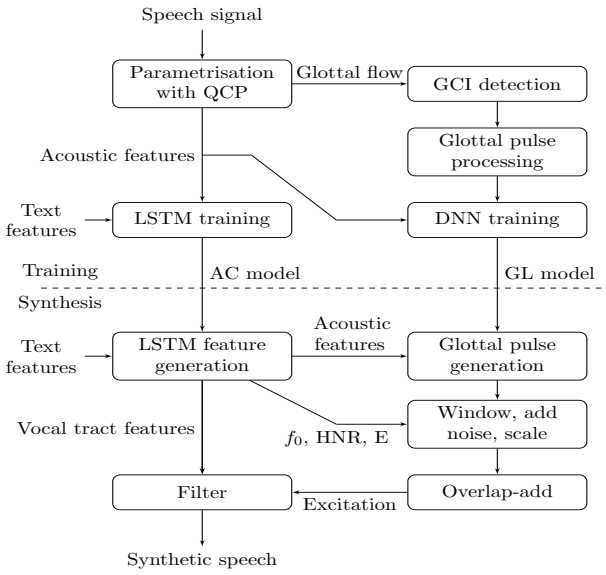



**Fig. 1**. Block diagram of the LSTM-based speech synthesis system using the GlottDNN vocoder.

the input level, the second method depends on learning Lombard-dependent amplitudes of the hidden unit contributions (LHUC). The third approach directly updates the parameters of the normal speech LSTM-TTS through fine-tuning.

## 3. LSTM-BASED SPEECH SYNTHESIS SYSTEM

Figure 1 illustrates the block diagram of our LSTM-based speech synthesis system. In this study, we employed our new vocoder, denoted as GlottDNN [20], to model both normal and Lombard speech. The GlottDNN vocoder is built on the principles of its predecessor, GlottHMM [21], but the new vocoder introduces three main improvements: GlottDNN (1) takes advantage of a new more accurate glottal inverse filtering method, quasi-closed phase analysis (QCP) [22], (2) uses a new method of deep neural network (DNN) -based glottal excitation generation [23], and (3) proposes a new approach of band-wise processing of full-band speech. Studies [20] and [23] clearly showed that the new vocoder performs better than GlottHMM and the widely-used STRAIGHT vocoder.

In the training, we extract parameters from speech using the QCP inverse filtering method, which decomposes the speech signal into the vocal tract filter and voice source signal. This enables the further parametrisation of the voice source and the segmentation of the glottal flow waveforms. Table 1 describes the acoustic parameters extracted from both normal and Lombard speech. The vocal tract and the glottal flow pulses are modeled separately as 1) Acoustic (AC) model, and 2) glottal (GL) model.

The AC model is trained with a LSTM which takes textual features as input and predicts acoustic parameters as output. The GL model is trained with a simple DNN, which takes acoustic parameters as input and predicts the two-pitch period glottal waveform as output. The reason we use the DNN instead of LSTM for the GL model is that unvoiced sounds do not have glottal pulse excitation, thus it is hard to capture sequential information. The acoustic parameters of normal speech are used for training an LSTM-based voice, after which it can be adapted to Lombard speech.

In synthesis, acoustic parameters are first predicted from the AC model using textual features and later the predicted acoustic parameters are employed to predict the time-domain glottal waveform from the GL model. The excitation signal is generated separately for voiced and unvoiced frames. The excitation signal is further processed by the predicted acoustic parameters and combined into a single excitation signal [23]. The vocal tract LSFs are also processed to reduce over-smoothing caused by LSTM modelling. Finally, the LSFs are converted back to linear prediction (LP) coefficients and the obtained filter is finally used to filter the combined excitation signal.

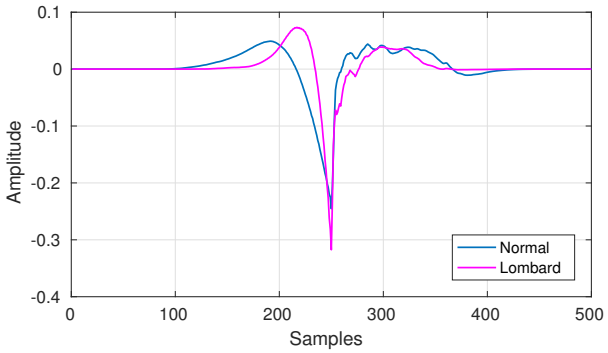



**Fig. 2**. Illustration of the mean of the windowed two-period glottal flow derivative waveforms for normal and Lombard speech of a male speaker.

## 4. ADAPTATION TO LOMBARD SPEECH

We outline three methods using: (a) auxiliary features, (b) learning hidden unit contributions (LHUC) and (c) fine-tuning to explicitly adapt the normal speech AC model to Lombard speech and also outline the GL model for Lombard speech.

### 4.1. Auxiliary features

The use of augmented or auxiliary features is an approach in speaker-adaptive training in which the linguistic features are augmented with additional speaker-specific features computed for each speaker at both training and test stages. Studies in [16, 17, 19] have successfully used the auxiliary information such as gender, speaker identity, age or i-vector for speaker adaptation in DNN-based speech recognition and synthesis. In this work, we augment the speaking style specific information as auxiliary features to linguistic features at input level. We use two binary values for normal and Lombard speech in one-hot representation. The augmented values enable distinguishing the same linguistic content spoken in the normal speaking style from that produced using Lombard speech.

### 4.2. Learning hidden unit contributions (LHUC)

The LHUC method has been proposed in DNN-based speech recognition [24] for unsupervised speaker adaptation and it was later applied in speech synthesis also for speaker adaptation [16]. In principle this method can be used in any acoustic model adaptation [24], which is the reason why we employed it in the current study for Lombard speech adaptation. It has several advantages including lower number of parameters and robustness against overfitting. For the implementation, we followed in detail the procedure described in [24].

### 4.3. Fine-tuning

In this method, the parameters of a LSTM trained on normal speech are fine-tuned by Lombard speech. We fine-tune all the layers of

the LSTM network, the acoustic differences between normal and Lombard speech can act as regularization thereby preventing overfitting. The parameters are fine-tuned by the standard back propagation through time algorithm. This fine-tuning is motivated by the assumption that the LSTM trained on normal speech contain generic features, which are more robust against data variation and therefore useful in Lombard speech adaptation [25].

### 4.4. Lombard glottal (GL) model

The Lombard effect is known to manifest itself in the excitation of voiced speech, the glottal flow. In the production of Lombard speech, natural talkers tend to decrease both the glottal pulse length (thereby increasing the fundamental frequency, F0) as well as its spectral tilt [26]. Figure 2 shows examples of the mean glottal flow derivative waveforms for both normal and Lombard speech. It can be clearly seen that the glottal flow derivative of Lombard speech is more skewed than the corresponding wave in normal speech hence suggesting that the Lombard sound has been produced using a larger vocal effort.

Given the essential role of the glottal pulse in the production of natural Lombard speech, the current study uses a simple DNN to model the GL model which maps acoustic parameters to corresponding glottal pulses in Lombard speech. The use of DNN-generated glottal pulses is also justified by our recent study indicating that the approach gives a significant quality improvement in synthesis of high-pitched speech compared to a baseline system based on using a single mean pulse [23].

## 5. EXPERIMENTS

### 5.1. Speech material

We employed the Hurricane challenge corpus [27] for our experiments. It contains both normal and Lombard speech data recorded by a male native British English professional voice talent. The Lombard speech data consists of 720 Harvard utterances and the normal speech data consists of 2542 utterances recorded in a hemi-anechoic chamber. To elicit the Lombard effect, a temporally-modulated speech-shaped noise masker was played over headphones at a calibrated level of 84 dB(A).

### 5.2. TTS systems

In training the normal speech TTS, the data were divided into 2490, 20, and 32 utterances as training set, development set and evaluation set, respectively. In Lombard speech adaptation, we considered two adaptation conditions: 500 utterances and 10 utterances. In both conditions, 20 utterances were used as a development set and 180 utterances were used as an evaluation set.

The sampling rate of the corpus was 16 kHz. The GlottDNN vocoder was used to extract both vocal tract and voice source parameters according to Table 1. The full contextual labels were generated from the text files, which were available along with speech, using the Festival toolkit. Since the study focuses on adapting the spectral properties of a glottal-vocoded LSTM-based TTS system, the state-level durations of the entire speech data were obtained by forced alignment based on HMMs.

We trained context-dependent hidden Semi-Markov models (HSMMs) as the baseline. The HSMMs have 5 states and each consisted of four streams: (1) the vocal tract spectrum LSFs combined with energy, (2) the voice source spectrum LSFs, (3) the harmonic-to-noise (HNR), and (4) the fundamental frequency in log scale, $logF_0$. A total of 141 acoustic parameters including delta, delta-delta features were employed in the training. To develop the Lombard voice models, we first trained the voice models of the normal speaking style and later the decision trees of these voice models were adapted with the constrained maximum likelihood linear regression (CMLLR) algorithm. The global variance (GV) of the original training data was taken into account in the speech parameters generation.

| Feature | Type/Unit | Dimension |
|---|---|---|
| Vocal tract spectrum | LSF | 30 |
| Energy | dB | 1 |
| Fundamental frequency | $logf_0$ | 1 |
| Harmonic-to-noise ratio | dB/ERB | 5 |
| Voice source spectrum | LSF | 10 |

**Table 1**. Acoustic features used in training the LSTM-based AC model and the DNN-based GL model.

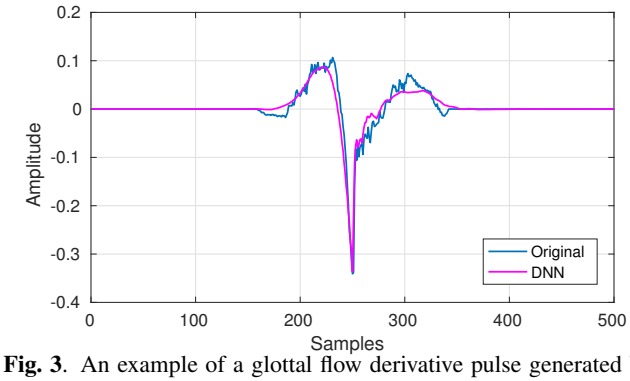



**Fig. 3**. An example of a glottal flow derivative pulse generated by the GL model on the Lombard speaking style.

The AC model using LSTMs was trained as described in [28]. The full-context labels were converted into binary and numerical features using the question file employed in the decision tree clustering in the HMM-based speech synthesis system. These features comprised information such as the phoneme identity, syllable location, part-of-speech, number of words in an utterance, and number of phrases in an utterance. Extra 9 numerical values were appended providing information about the position of frame within the HMM state and phoneme, the state position within the phoneme, and state and phoneme durations. In total, the input feature vector was 335 in dimension. The output parameters were the same as the acoustic parameters in the baseline system. The $F0$ was linearly interpolated and an extra V/UV feature was added to acquire the voice/unvoiced information at runtime synthesis. Thus, in total, the output feature was 142 dimensional. The input features were normalized to the range of [0.1, 0.99] by using the min-max method. The output features were normalized using the mean-variance normalization method. The development and evaluation set were normalized by the values derived from the training data. Similar normalization was applied to the adaptation data. To generate smooth parameter trajectories, the maximum likelihood parameter generation (MLPG) algorithm was applied on predicted acoustic parameters using the global variances.

The LSTM architecture used in the current study consisted of 3 hidden layers which were followed by a linear layer at the output. The 3 hidden layers consisted of 2 feed-forward network layers at bottom and 1 simplified LSTM layer on top. The bottom feed-forward layers were intended to act as feature extraction layers, with

| Adaptation method | 10 utterances adaptation | | | | 500 utterances adaptation | | | |
|---|---|---|---|---|---|---|---|---|
| | LSF | LSFsource | HNR (dB) | F0 (Hz) | LSF | LSFsource | HNR (dB) | F0 (Hz) |
| Auxiliary | 0.278 | 0.139 | 12.099 | 43.636 | 0.190 | 0.116 | 8.621 | 16.711 |
| LHUC | 0.237 | 0.133 | 10.638 | 25.729 | 0.215 | 0.124 | 9.437 | 18.518 |
| Fine-tuning | **0.223** | **0.129** | **9.882** | **24.437** | **0.184** | **0.112** | **8.359** | **15.868** |

**Table 2**. Objective results of LSTM adaptation methods. The mean square error (MSE) was calculated for all acoustic features on the evaluation data set in both the 10 and 500 utterances adaptation conditions. LSFsource denotes the LSFs of voice source spectrum.

512 hidden units using tangent activation function in each layer. The top layer had 256 LSTM blocks. The learning rate was tuned on the development set. The implementation was done with the Merlin toolkit [29].

The Lombard GL model using DNNs was developed as described in [23]. The input features were the same as described in Table 1 (i.e. 47 in dimension) and the output features were 500 time-domain samples of the duration normalized glottal flow waveform. The DNN architecture consisted of three feed-forward multilayer nets with 100, 200, and 300 units. The sigmoid and linear activations were used for hidden and output layers, respectively. An example of a glottal flow derivative pulse generated with the GL model for the Lombard speaking style is shown in Figure 3. It can be observed that the predicted glottal waveform is very close to the original glottal waveform estimated from Lombard speech.

### 5.3. Objective evaluation

An objective evaluation was conducted to analyze the performance of each adaptation approach in the 10 utterances and 500 utterances adaptation conditions. The mean square error (MSE) was computed between predicted and original acoustic parameters of the entire evaluation set. Results are presented in Table 2. As expected, the MSE values using 10 utterances were higher than those obtained using 500 utterances. The fine-tuning method approach achieved the lowest MSE, among the three proposed adaptation methods, across all acoustic parameters. The auxiliary features method performed second best in the 500 utterances condition, but in the 10 utterance condition it performed worst. The LHUC method performed worst in the 500 utterances condition whereas in the 10 utterances condition it performed second best.

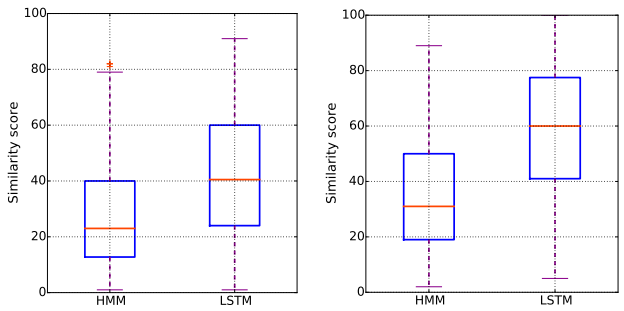
Results indicate that in the large data condition the LSTMs were able to discriminate better between normal and Lombard speech using auxiliary features. The change in the objective scores between the large and small data conditions was less for the LHUC compared to the other two methods. This indicates that the LHUC method is more robust to changes in data sizes. The fine-tuning method performed best most likely because the same speaker data were employed in both normal speech TTS training and Lombard speech adaption. We conducted an informal listening test on the three methods and selected the fine-tuning method to be used in formal subjective evaluations to compare the LSTM-based adaptation with a HMM-based system.

### 5.4. Subjective evaluation

In order to compare the performance of the HMM- and LSTM-based adaptation systems, a subjective evaluation was carried out using a similarity listening test. In this test, each subject first listened to the natural Lombard speech as a reference, and then listened to two samples generated by either the LSTM-based or the HMM-based system. The listener was asked to rate on a scale from 0 to 100 how close the synthetic sound is to the reference sample (0: sounds totally different from reference, 100: sounds exactly like reference). The reference and the synthetic samples had the same linguistic content and durations to make the listener focus only on the effects of acoustic feature adaptation. The listeners were able to listen to each sample as many times as they wished and the order of the test cases was randomized separately for each listener.

The listening test was conducted via a web-based interface implemented with the modified Beaqlejs application [30]. 19 synthesized sentences were selected from both adaptation systems. A total of 15 subjects (11 native English speakers and 4 international students at Aalto University) participated in the listening test. All subjects were included in the final analysis of the results.

The results of the subjective evaluation are presented in Figure 4. The left panel shows the results using 10 utterances as adaptation data and the right panel shows the results using 500 utterances as adaptation data. It is observed that in both adaptation conditions, the LSTM-based method achieves clearly better performance than the HMM-based baseline in terms of Lombard similarity. This confirms the adaptability of LSTMs, and shows the effectiveness of the fine-tuning based adaptation method.




(a) 10 utterances adaptation (b) 500 utterances adaptation

**Fig. 4**. Box plot of similarity test results between HMM and LSTM-based adapted systems. (a) Adaptation using 10 utterances. (b) Adaptation using 500 utterances.

## 6. CONCLUSIONS

In this study, a systematic experimental analysis was conducted on Lombard speech synthesis using long short-term memory recurrent neural networks (LSTMs). The experiments were conducted using 10 utterances (small) and 500 utterances (large) of Lombard speech data for adaptation. The subjective evaluation confirmed that the LSTMs are able to adapt better to the Lombard style than HMMs. We also found that the simple fine-tuning method was the best adaptation technique in the case when both normal and Lombard speech was spoken by the same speaker. In the future, we will study how the proposed LSTM-based adaptation method works for other speaking styles including breathy and shouted speech.

The samples and listening test results used in the experiments are available online via this link: `http://bit.ly/2c01a8O`

**Acknowledgment:** This work was supported by the Academy of Finland (project No. 256961, 284671).

## 7. REFERENCES


[1] Walter Van Summers, David B Pisoni, Robert H Bernacki, Robert I Pedlow, and Michael A Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[2] Martin Cooke, Catherine Mayo, Cassia Valentini-Botinhao, Yannis Stylianou, Bastian Sauert, and Yan Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[3] Cassia Valentini-Botinhao, Junichi Yamagishi, Simon King, and Ranniery Maia, "Intelligibility enhancement of hmm-generated speech in additive noise by modifying mel cepstral coefficients to increase the glimpse proportion," *Computer Speech & Language*, vol. 28, no. 2, pp. 665–686, 2014.

[4] D. Erro, T. C. Zoril, and Y. Stylianou, "Enhancing the intelligibility of statistically generated synthetic speech by means of noise-independent modifications," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 22, no. 12, pp. 2101–2111, Dec 2014.

[5] Cassia Valentini-Botinhao, Junichi Yamagishi, Simon King, and Yannis Stylianou, "Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of hmm-based synthetic speech in noise.," in *Interspeech*, 2013, pp. 3567–3571.

[6] Brian Langner and Alan W Black, "Improving the understandability of speech synthesis by modeling speech in noise," in *ICASSP*, 2005, pp. 265–268.

[7] Tuomo Raitio, Antti Suni, Martti Vainio, and Paavo Alku, "Analysis of hmm-based lombard speech synthesis.," in *Interspeech*, 2011, pp. 2781–2784.

[8] Antti Suni, Reima Karhila, Tuomo Raitio, Mikko Kurimo, Martti Vainio, and Paavo Alku, "Lombard modified text-to-speech synthesis for improved intelligibility: submission for the hurricane challenge 2013," in *Interspeech*, 2013, pp. 3562–3566.

[9] Benjamin Picart, Thomas Drugman, and Thierry Dutoit, "Analysis and hmm-based synthesis of hypo and hyperarticulated speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 687–707, 2014.

[10] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[11] Junichi Yamagishi, Takao Kobayashi, Yuji Nakano, Katsumi Ogata, and Juri Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 17, no. 1, pp. 66–83, 2009.

[12] Tuomo Raitio, Antti Suni, Martti Vainio, and Paavo Alku, "Synthesis and perception of breathy, normal, and lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, 2014.

[13] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *ICASSP*, 2013, pp. 7962–7966.

[14] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks.," in *Interspeech*, 2014, pp. 1964–1968.

[15] Heiga Zen and Haşim Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *ICASSP*, 2015, pp. 4470–4474.

[16] Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, Steve Renals, and Simon King, "A study of speaker adaptation for dnn-based speech synthesis," in *Interspeech*, 2015.

[17] Blaise Potard, Petr Motlicek, and David Imseng, "Preliminary work on speaker adaptation for dnn-based speech synthesis," Tech. Rep., Idiap, 2015.

[18] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *ICASSP*, 2013, pp. 7893–7897.

[19] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors.," in *ASRU*, 2013, pp. 55–59.

[20] Manu Airaksinen, Bajibabu Bollepalli, Lauri Juvela, Zhizheng Wu, Simon King, and Paavo Alku, "Glottdnna full-band glottal vocoder for statistical parametric speech synthesis," in *Interspeech*, 2016.

[21] Tuomo Raitio, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio, and Paavo Alku, "Hmm-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 19, no. 1, pp. 153–165, 2011.

[22] Manu Airaksinen, Tuomo Raitio, Brad Story, and Paavo Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 22, no. 3, pp. 596–607, 2014.

[23] Lauri Juvela, Bajibabu Bollepalli, Manu Airaksinen, and Paavo Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *ICASSP*, 2016, pp. 5120–5124.

[24] Pawel Swietojanski, Jinyu Li, and Steve Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 24, no. 8, pp. 1450–1463, 2016.

[25] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[26] Thomas Drugman and Thierry Dutoit, "Glottal-based analysis of the lombard effect.," in *Interspeech*, 2010, pp. 2610–2613.

[27] Martin Cooke, Catherine Mayo, and Cassia Valentini-Botinhao, "Hurricane natural speech corpus," [sound], 2013.

[28] Zhizheng Wu and Simon King, "Investigating gated recurrent networks for speech synthesis," in *ICASSP*, 2016, pp. 5140–5144.

[29] Simon King Zhizheng Wu, Oliver Watts, "Merlin: An open source neural network speech synthesis system," in *ISCA Speech Synthesis Workshop (SSW9)*, 2016.

[30] Sebastian Kraft and Udo Zölzer, "Beaqlejs: Html5 and javascript based framework for the subjective evaluation of audio quality," in *Linux Audio Conference, Karlsruhe, DE*, 2014.