LSTM Deep Neural Networks Postfiltering for Improving the Quality of Synthetic Voices

Marvin Coto-Jiménez and John Goddard-Close

Abstract—Recent developments in speech synthesis have produced systems capable of outcome intelligible speech, but now researchers strive to create models that more accurately mimic human voices. One such development is the incorporation of multiple linguistic styles in various languages and accents.

HMM-based Speech Synthesis is of great interest to many researchers, due to its ability to produce sophisticated features with small footprint. Despite such progress, its quality has not yet reached the level of the predominant unit-selection approaches that choose and concatenate recordings of real speech. Recent efforts have been made in the direction of improving these systems.

In this paper we present the application of Long-Short Term Memory Deep Neural Networks as a Postfiltering step of HMM-based speech synthesis, in order to obtain closer spectral characteristics to those of natural speech. The results show how HMM-voices could be improved using this approach.

Index Terms—LSTM, HMM, Speech Synthesis, Statistical Parametric Speech Synthesis, Postfiltering, Deep Learning

I. INTRODUCTION

Text-to-speech (TTS) synthesis is the technique of generating intelligible speech from a given text. Applications of TTS have expanded from early supporting artifacts for the visually impaired, to in-car navigation systems, e-book readers, spoken dialog systems, communicative robots, singing speech synthesizers, and speech-to speech-translation systems [1].
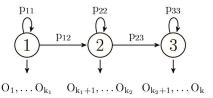
More recently, TTS systems have evolved from the sole production of intelligible voices to pursuit more sophisticated production of voices in multiple languages, with different styles and emotions [2]. Despite these trends, there are challenges, for example the overall quality of the voices. Researchers are striving to improve TTS systems by more closely mimicking natural human voices

The statistical methods for TTS, which arise in the late 1990s, have grown in popularity since then [3], particularly those based on Hidden Markov Models (HMM), for their flexibility in changing speaker characteristics and low footprint, including capacities to produce average voices. HMM have been utilized extensively in speech recognition since about 30 years ago, as they provide a robust representation of the main events in which speech can be segmented [4], with efficient parameter estimation algorithms.

More than 30 reports of HMM-based Speech Synthesis implementations (also called Statistical Parametric Speech Synthesis) can be found for several languages around the

M. Coto-Jiménez, Department of Electrical Engineering, University of Costa Rica, San José, Costa Rica and the Metropolitan Autonomous University, México D.F., México. e-mail: marvin.coto@ucr.ac.cr

John Goddard-Close, Department of Electrical Engineering, Metropolitan Autonomous University, México D.F., México. email:jgc@xanum.uam.mx

Fig. 1: Left to right example of an HMM with three states

world. For example [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], are a few of the recent ones. Every implementation in a new language or its variants requires the adoption of HMM-related algorithms, incorporating their own linguistic specifications and making a series of decisions regarding the multiple definitions related to HMM, decision tress, and training conditions.

In this paper, we present our implementation of a statistical parametric speech synthesis system based on HMM with Long Short-Term Memory Postfilter Neural Networks for improving its spectral quality.

The rest of this paper is organized as follows: Section 2 provides some details of the HMM-based speech synthesis system; Section 3 presents the Long Short-Term Memory Neural Networks; and Section 4 describes the system and experiments carried out in order to test the Postfilter. Section 5 presents the results and analysis of objective evaluations, and the conclusions are in Section 6.

II. SPEECH SYNTHESIS BASED ON HMM

HMM can be described from a Markov process, in which state transitions are given by a stochastic process. A second stochastic process models the emission of symbols when it comes to each state.

In Figure 1, a representation of a left to right HMM is shown, where there is a first state to the left from which transitions can occur to the same state or to the next on the right, but not in reverse direction. In this p_{ij} represents the probability of transition from state i to state j, and O_k represents the observation emitted in state k.

In HMM-based Speech Synthesis, the speech waveforms can be reasonably reconstructed from a sequence of acoustic parameters learned and emitted as vectors from the HMM states [1]. Typical implementation of this model includes vectors of observations with $f0$, MFCC and their delta and delta delta features for the adequate modeling of dynamic features of speech.

In order to improve the quality of the results, some researchers have recently experienced Postfiltering stages in

which the parameters obtained with HTS voices have enhanced deep generative architectures [17], [18], [19], [20], for example DBM, RMB, BAM and recurrent neural networks.

In the next section, we present our proposal to incorporate Long Short-Term Memory Recurrent Neural Networks in the improvement of the quality of HMM-based speech synthesis.

III. Long Short-Term Memory Recurrent Neural Networks

mong the new algorithms to improve some tasks related to speech, such as speech recognition, groups of researchers have explored the use of Deep Neural Networks (DNN), with encouraging results. Deep learning, based on several kinds of neural networks with many hidden layers, have achieved great results in many machine learning and pattern recognition tasks. The disadvantage of using such networks is they cannot directly model the dependent nature of each sequence of parameters with the former, which is desirable to mimic the production of human speech. To solve this problem, it has been suggested to include RNN [21] [22] in which there is feedback from some of the neurons in the network, backwards or to themselves, forming a kind of memory that retains previous states.

An extended kind of RNN, which can store information over long or short time intervals, has been presented in [23], called Long Short-Term Memory (LSTM). LSTM was recently introduced to speech recognition, giving the lowest recorded error rates on the TIMIT database [24], among other successful applications of speech recognition [25]. The storage and use of long-term and short-term information is potentially significant for many applications, including speech processing, non-Markovian control, and music composition [23].

In a RNN, output vector sequences $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ are computed from input vector sequences $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ and hidden vector sequences $\mathbf{h} = (h_1, h_2, \ldots, h_T)$ iterating equations 1 and 2 from 1 to T [21]:

$$h_t = \mathcal{H}\left(\mathbf{W}_{xh}x_t + \mathbf{W}_{hh}h_{t-1} + b_h\right) \quad (1)$$

$$y_y = \mathbf{W}_{hy}h_t + b_y \quad (2)$$

where $\mathbf{W}_{ij}$ is the weight matrix between layer i and j, b_k is the bias vector for layer k and $\mathcal{H}$ is the activation function for hidden nodes, usually a sigmoid function $f : \mathbb{R} \to \mathbb{R}$, $f(t) = \frac{1}{1+e^{-t}}$.

Each cell in the hidden layers of a LSTM, has some extra gates to store values: an input gate, forget gate, output gate and cell activation, so values can be stored in the long or short term. These gates are implemented following the equations:

$$i_t = \sigma\left(\mathbf{W}_{xi}x_t + \mathbf{W}_{hi}h_{t-1} + \mathbf{W}_{ci}c_{t-1} + b_i\right) \quad (3)$$

$$f_t = \sigma\left(\mathbf{W}_{xf}x_t + \mathbf{W}_{hf}h_{t-1} + \mathbf{W}_{cf}c_{t-1} + b_f\right) \quad (4)$$

$$c_t = f_t c_{t-1} + i_t \tanh\left(\mathbf{W}_{xc}x_t + \mathbf{W}_{hc}h_{t-1} + b_c\right) \quad (5)$$

$$o_t = \sigma\left(\mathbf{W}_{xo}x_t + \mathbf{W}_{ho}h_{t-1} + \mathbf{W}_{co}c_t + b_o\right) \quad (6)$$

$$h_t = i_t \tanh\left(c_t\right) \quad (7)$$

where σ is the sigmoid function, i is the input gate activation vector, f the forget gate activation function, o is the output gate activation function, and c the cell activation function. $\mathbf{W}_{mn}$ are the weight matrices from each cell to gate vector.

IV. Description of the system

The resulting voices from the HTS system have notable differences with the original voices used in its production. Reducing the gap between natural and artificial voices can be learned directly from data [17]. In our proposal, we use aligned utterances from natural and synthetic voices produced in the HTS system to establish correspondence between each frame.

Given a sentence of natural speech and voice corresponding HTS, we extract a representation consisting of one coefficient for f0, one coefficient for energy, and 39 MFCC coefficients, using the system Ahocoder [26]. The inputs to the LSTM network correspond to the MFCC parameters of each frame of the sentences produced with the HTS voice, while the output corresponds to the MFCC parameters of the natural voice of the same sentence. In this case, we have an exact correspondence between the vector, representing each phrase from HTS voice and natural voice by the alignment between both.

In this way, each LSTM Network attempts to solve the regression problem of transforming the values of the artificial speech and natural voice. This allows further improvement of the quality of new synthesized utterances with HTS, using this neural network as a subsequent step to approach these synthetic parameters to those of natural voice. Figure 2 outlines the proposed system.
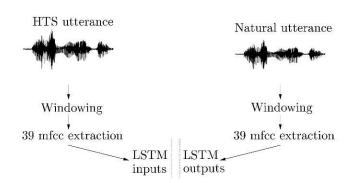

Fig. 2: Proposed system. HTS and Natural utterances are aligned frame by frame

A. Corpus description

The CMU_ARCTIC databases were constructed at the Language Technologies Institute at Carnegie Mellon University. They are phonetically balanced, with several US English speakers. It was designed for unit selection speech synthesis research.

The databases consist of around 1150 utterances selected from out-of-copyright texts from Project Gutenberg. The databases include US English male and female speakers. A

detailed report on the structure and content of the database and the recording conditions is available in the Language Technologies Institute Tech Report CMU-LTI-03-177 [27]. Four of the available voices were selected: BDL (male), CLB (female), RMS (male) and SLT (female).

B. Experiments

Each voice was parameterized, and the resulting set of vectors was divided into training, validation, and testing sets. The amount of data available for each voice are shown in Table I. Despite all voices uttering the same phrases, the length differences are due to variations in the speech rate of each speaker.

TABLE I: Amount of data (vectors) available for each voice in the databases

Database	Total	Train	Validation	Test
BDL	676554	473588	135311	67655
SLT	677970	474579	135594	67797
CLB	769161	538413	153832	76916
RMS	793067	555147	158613	79307

The LSTM networks for each voice had three hidden layers, with 200, 160 and 200 units in each one respectively.

To determine the improvement in the quality of the synthetic voices, several objective measures were used. These measures have been applied in recent speech synthesis experiments and were found to be reliable in measuring the quality of synthesized voices [28] [29]:

- Mel Cepstral Distortion (MCD): Excluding silent phonemes, between two waveforms v^{targ} and v^{ref} it can be measured following equation 8 [30]

$$\text{MCD}\left(v^{\text{targ}}, v^{\text{ref}}\right) = \frac{\alpha}{T} \sum_{t=0}^{T-1} \sqrt{\sum_{d=s}^{D} \left(v_d^{\text{targ}}(t) - v_d^{\text{ref}}(t)\right)^2}$$

(8)

where $\alpha = \frac{10\sqrt{2}}{\ln 10}$, T is the number of frames of each utterance, and D the total number of parameters of each vector.

- MFCC trajectory and spectrogram visualization: Simple observation of these elements by comparison permits to visualize quality in terms of similitude with those of the natural voice.

These measures were applied to the test set after being processed with the LSTM networks, and the results were compared with those of the HTS voices. The results and analysis are shown in the following section.

V. RESULTS AND ANALYSIS

For each synthesized voice produced with HTS and processed with LSTM networks, MCD results are shown in Table II. It can be seen how this parameter improved when all voices were processed with LSTM networks. This shows the ability of these networks to learn the particular regression problem of each voice.

TABLE II: MCD between HTS and Natural Voices, and between LSTM Postfiltering and Natural Voice

Database	HTS to Natural	HTS to LSTM-PF
BDL	8.46	7.98
CLB	7.46	6.87
SLT	7.03	6.65
RMS	7.66	7.60

The best result of MCD improvement with the LSTM Postfiltering is CLB (7.9%) and the least best was RMS(1%). Figure 3 shows how the MCD evolves with the training epochs for each voice. All HTS voices, except one, were improved by the LSTM Neural Network Postfilter in the MCD from the first 50 ephocs of training.

The differences in the amount of necessary epochs to reach convergence in each case are notable. This can be explained by the difference in the MCD between HTS and natural voices. The gap between them is variable and the LSTM network requires more epochs to model the regression function between them. An example of parameters generated by the HTS and the enhancement obtained by the LSTM Postfilter is shown in Figure 4. It can be seen how the LSTM Postfilter fits the trajectory of the MFCC better than the HTS base system. In Figure 5 a comparison of three spectrograms of the utterance "Will we ever forget it?" for the HTS voice (a), Original (b) and LSTM Postfilter enhanced (c) is shown. The HTS spectrogram usually shows bands in higher frequencies not present in the natural voice, and the LSTM-Postfilter helps to smooth it, making it closer to the original voice spectrogram.

VI. CONCLUSIONS

We have presented a new proposal to improve the quality of synthetic voices based on HMM with LSTM Postfiltering Networks. The LSTM have been able to learn directly from the data how to improve an artificial voice and make it mimic a more natural sound in its spectral characteristics.

We evaluated the proposed LSTM Postfilter in four voices, two masculine and two feminine, and the results show that all of them were improved in spectral features such as MCD measurement, spectrograms and mfcc trajectory generation.

The improvement of the HTS voices in MCD to the original voices were observed from the first training epochs of the LSTM neural network, but the convergence to a minimum distance took many more epochs. Due to the extensive amount of time required to train each epoch, further exploration should determine new network configurations or training conditions to reduce training time.

Future work will include the exploration of new representation of speech signals, hybrid neural networks and fundamental frequency enhancement with LSTM Postfilters.
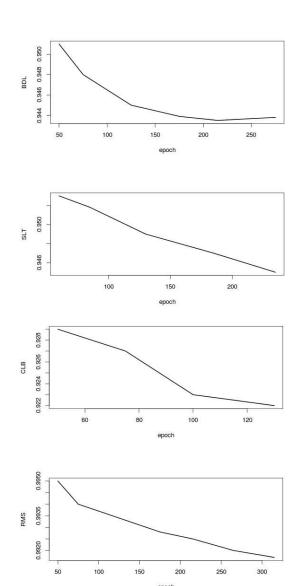
VII. ACKNOWLEDGEMENTS
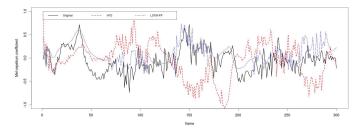
Fig. 3: Evolution of MCD improvement in LSTM Postfiltering during training epochs

Fig. 4: Illustration of enhancing the 5th mel-cepstral coefficient trajectory by LSTM Postfiltering

(a) Original

(b) HTS

(c) LSTM Postfiltering

Fig. 5: Comparison of spectrograms

REFERENCES

[1] Tokuda K, Nankaku Y, Toda T, Zen H, Yamagishi J, and Oura K (2013): Speech synthesis based on hidden markov models. In: Proceedings of the IEEE, 101(5):1234–1252.
[2] Black AW (2003): Unit selection and emotional speech. in *INTER-SPEECH*.
[3] Yoshimura T, Tokuda T, Masuko T, Kobayashi T and Kitamura T (1999): Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In Proc. Eurospeech:2347–2350.
[4] Falaschi A, Giustiniani M, and Verola M (1989): A Hidden Markov Model Approach to Speech Synthesis. In EUROSPEECH:2187–2190.
[5] Karabetsos S, Tsiakoulis P, Chalamandaris A, and Raptis S (2008): HMM-based Speech Synthesis for the Greek Language. In Text, Speech and Dialogue. Springer, p 349–356.
[6] Pucher M, Schabus D, Yamagishi Y, Neubarth F, and Strom V (2010): Modeling and interpolation of Austrian German and Viennese Dialect in HMM-based Speech Synthesis. Speech Communication 52(2):164–179.
[7] Erro D, Sainz I, Luengo I, Odriozola I, Sánchez J, Saratxaga I, Navas E, and Hernáez I (2010): HMM-based Speech Synthesis in Basque Language Using HTS. IN: Proceedings of the FALA.
[8] Stan A, Yamagishi Y, King S, and Aylett M (2011): The Romanian Speech Synthesis (RSS) Corpus: Building a High Quality HMM-based Speech Synthesis System Using a High Sampling Rate. In: Speech Communication, 53(3):442–450.
[9] Kuczmarski T (2010): HMM-based Speech Synthesis Applied to Polish. Speech and Language Technology 12:13.
[10] Hanzlíček Z (2010):Czech HMM-based speech synthesis. In: Text, Speech and Dialogue. Springer, p 291–298.
[11] Li Y, Pan S, and Tao J (2010): HMM-based Speech Synthesis with a

Flexible Mandarin Stress Adaptation Model. In: Proc. 10th ICSP2010 Proceedings, Beijing, p 625–628.

[12] Phan ST, Vu TT, Duong CT and Luong MC (2013): A study in Vietnamese Statistical Parametric Speech Synthesis Based on HMM. International Journal, 2(1):p 1-6.

[13] Boothalingam R, Sherlin Solomi V, Gladston AR, Christina SL, Vijay-alakshmi P, Thangavelu N, and Murthy HA (2013): Development and evaluation of unit selection and HMM-based speech synthesis systems for Tamil. In: National Conference on Communications (NCC), IEEE, p 1–5.

[14] Khalil KM and Adnan C (2015): Implementation of speech synthesis based on HMM using PADAS database. In: 12th International Multi-Conference on Systems, Signals & Devices (SSD). IEEE, p 1–6

[15] Nakamura K, Oura K, Nankaku Y, and Tokuda K (2014): HMM-Based Singing Voice Synthesis and its Application to Japanese and English. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p 265–269.

[16] Roekhaut S, Brognaux S, Beaufort R, and Dutoit T (2014): Elite-HTS: a NLP tool for French HMM-based speech synthesis. In: Interspeech, p 2136–2137.

[17] Chen LH, Raitio T, Valentini-Botinhao C, Ling ZH and Yamagishi J (2015): A deep generative architecture for postfiltering in statistical parametric speech synthesis. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 23(11):2003–2014.

[18] Takamichi S, Toda T, Neubig G, Sakti S and Nakamura S (2014): A postfilter to modify the modulation spectrum in HMM-based speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p 290-294.

[19] Takamichi S, Toda T, Black AW and Nakamura S (2014): Modified post-filter to recover modulation spectrum for HMM-based speech synthesis. In IEEE Global Conference on Signal and Information Processing (GlobalSIP), p 547–551.

[20] Prasanna Kumar M and Black AW (2016): Recurrent Neural Network Postfilters for Statistical Parametric Speech Synthesis. arXiv preprint arXiv:1601.07215.

[21] Fan Y, Qian Y, Xie FL and Soong FK (2014): TTS synthesis with bidirectional LSTM based recurrent neural networks. In Interspeech, p 1964–1968.

[22] Zen H and Sak H (2015): Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p 4470–4474.

[23] Hochreiter S and Schmidhuber J (1997): Long short-term memory. Neural computation 9(8): 1735–1780.

[24] Graves Alan, Jaitly N, Mohamed A (2013): Hybrid speech recognition with deep bidirectional LSTM. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU).

[25] Graves A, Fernández S and Schmidhuber J. (2005): Bidirectional LSTM networks for improved phoneme classification and recognition. Artificial Neural Networks: Formal Models and Their Applications–ICANN. Springer Berlin Heidelberg, p 799–804.

[26] Erro D, Sainz I, Navas E, Hernaez I (2011): Improved HNM-based Vocoder for Statistical Synthesizers. InterSpeech, p 1809–1812.

[27] Kominek J and Black AW (2004): The CMU Arctic speech databases. Fifth ISCA Workshop on Speech Synthesis.

[28] Zen H, Senior A, and Schuster M (2013): Statistical parametric speech synthesis using deep neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[29] Zen H and Senior A (2014): Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[30] Kominek J, Schultz T and Black AW (2008): Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. SLTU.