Speech Synthesis Based on Hidden Markov Models and Deep Learning

Marvin Coto-Jiménez[1,2], John Goddard-Close[2]

[1] University of Costa Rica, San José, Costa Rica
marvin.coto@ucr.ac.cr
[2] Autonomous Metropolitan University, México D.F., México
jgc@xanum.uam.mx

Abstract. Speech synthesis based on Hidden Markov Models (HMM) and other statistical parametric techniques have been a hot topic for some time. Using this techniques, speech synthesizers are able to produce intelligible and flexible voices. Despite progress, the quality of the voices produced using statistical parametric synthesis has not yet reached the level of the current predominant unit-selection approaches, that select and concatenate recordings of real speech. Researchers now strive to create models that more accurately mimic human voices. In this paper, we present our proposal to incorporate recent deep learning algorithms, specially the use of Long Short-term Memory (LSTM) to improve the quality of HMM-based speech synthesis. Thus far, the results indicate that HMM-voices can be improved using this approach in its spectral characteristics, but additional research should be conducted to improve other parameters of the voice signal, such as energy and fundamental frequency, to obtain more natural sounding voices.

Keywords: LSTM, HMM, Speech Synthesis, Statistical Parametric Speech Synthesis, Deep Learning.

1 Introduction

Speech synthesis is the process of creating artificial intelligible speech from a text, to propitiate human-machine interaction. Through this process, text-based information originated from computers, cell phones, and other technological devices can be transferred into speech. Since speech is the most common and natural way of communication for most human beings, this process is invaluable.

Due to its variation in speaking styles and potential for customization, speech synthesis based on HMM (also called statistical parametric speech synthesis) has been established as a viable alternative to dominant models, that are based on the concatenation of audio units [1].

In dominant models, audio waves of recording speech are handled directly, concatenating the wave segments to form new sentences. On the contrary, statistical parametric synthesis is based on the use of mathematical models that can learn and then reproduce parameters representing speech, for example HMM.

pp. 19–28; rec. 2016-04-04; acc. 2016-05-25

Recently, a small group of researchers [2–6], motivated by the success of deep learning algorithms in speech recognition, have reported encouraging results in adapting such algorithms in speech synthesis. As a result, a new possibility of developing artificial voices emerged, such that the original parameters could be kept while using new mathematical models that more accurately reproduce the characteristics of a human voice.

The reports made with these new techniques are preliminary due to restricted access to audio samples and to programs that reproduce these experiments and verify the results. This proposal aims to analyze such deep learning algorithms at different levels in statistical parametric speech synthesis, and develop techniques for using them in the process of creating higher quality artificial voices to extend their use and create new applications.

The rest of this paper is organized as follows: Section II describes the Research Problem; Section III presents the Research Methodology; Section IV discusses the expected Main Contribution; Section V reports the Results achieved so far; and finally Section VI presents the conclusions.

2 Background

Hidden Markov Models 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 speech synthesis and recognition, O_k are vectors of parameters representing speech, with fundamental frequency, spectral and duration information.

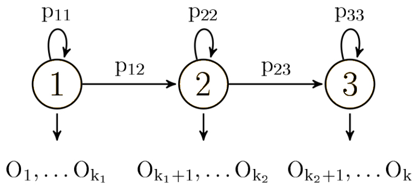
p_{11}
p_{22}
p_{33}
p_{12}
p_{23}
1
2
3
$\mathrm{O}_1, \ldots \mathrm{O}_{\mathrm{k}_1}$
$\mathrm{O}_{\mathrm{k}_1+1}, \ldots \mathrm{O}_{\mathrm{k}_2}$
$\mathrm{O}_{\mathrm{k}_2+1}, \ldots \mathrm{O}_{\mathrm{k}}$

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

HTS [7] is the only consolidated tool for using HMM in speech synthesis. The HTS automates the processes of extracting parameters (fundamental frequency, spectral and duration), training the HMM, and creating decision trees for search. Thus, it can search and find the most appropriate HMM to generate the parameters of the new phrases and synthesize a speech waveform.

In contrast to traditional techniques for artificial voice production, statistical parametric speech synthesis is based on mathematical models that learn to generate parameters which can build sound waves of speech, allowing for the pronunciation of any arbitrary text. This requires the extraction of an adequate representation of parameters that can model a person's voice and the combination and generation of new phrases that preserve the features of the original voice.

To complete this task, the process requires input into a database of natural voice recordings and the text of the new phrases that requires pronunciation using the artificial voice. Statistical parametric speech synthesis has many advantages over concatenative approaches, such as:

- Greater flexibility, because it is possible to modify characteristics of the synthesized speech, combining voices and create new ones manipulating the parameters.
- Small footprint, because only the parametric representations have to be stored and processed.
- Small requirements of database size.

Despite these advantages, its quality is not yet comparable to the best systems based on concatenation of speech units, so its use in applications is still not as widespread. The main reasons are [1]:

- Model of speech production: The reconstruction of speech waveform is based on a source-filter model, where the sounds with fundamental frequency are generated from a pulse train, and sounds without fundamental frequency are generated from white noise. This is considered a simplistic model in comparison with the production of human speech.
- Over smoothing: HMMs are trained with examples of speech that are segmented into phonemes and then averaged, which results in some loss of information. This results in smaller variability in contours of fundamental frequency and spectra.
- Mathematical modeling: It has been questioned whether the HMM may suitably represent natural speech. Some extensions of the technique as Global variance [8], Trajectory HMM [9] and autoregressive HMM [10] have presented improvements, but still need to prove its usefulness in the context of more natural and expressive speech.

3 Research Problem

The adoption of HMMs with Gaussian distributions to model parameters has emerged from its proven efficiency in the topic of speech recognition, previously developed since the 1980s in works like [11, 12]. Subsequently, there have been some improvements in the area of recognition. For example, the Adaptation Technique, where average models of Gaussian distributions are linearly transformed to improve recognition in new speakers [13, 14], produced benefits in its equivalence in speech synthesis, such as the creation of new voices. Recently, researchers have explored the use of new models and strategies for speech recognition, with encouraging results. Specifically, deep learning algorithms, based on deep neural network architectures stacking many hidden layers, have been highlighted. These implementations can be classified into two types:

1. Type I: Replacing the HMM with deep neural networks, so that all the acoustic representation of speech is done with the new models. Experiences using models such as: Deep Artificial Neural Networks (DNN) [15, 16] and recurrent artificial neural networks (RNN) [17], including the special case of LSTM [18] networks.

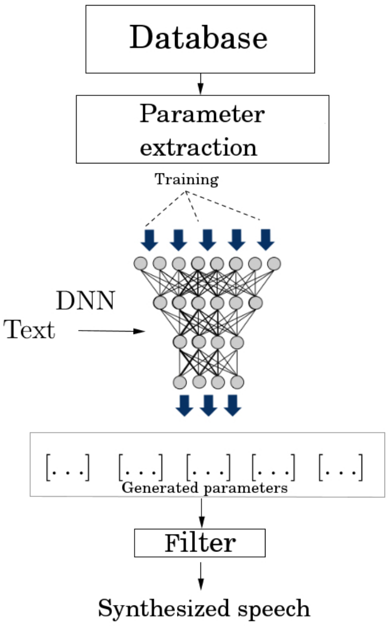

Fig. 2. Replacing the HMM by deep learning networks for parametric speech synthesis (Type I)

2. Type II: Adaptation or modification of the HMM models to improve recognition: Similar to the technique of adaptation, deep learning algorithms to improve HMM in recognition has been explored, and thereby increase the recognition rate. This is an addition to the models that have been tested for decades, without completely replacing the HMM. For example, in [19–21], where some of the elements of the voice are transformed.

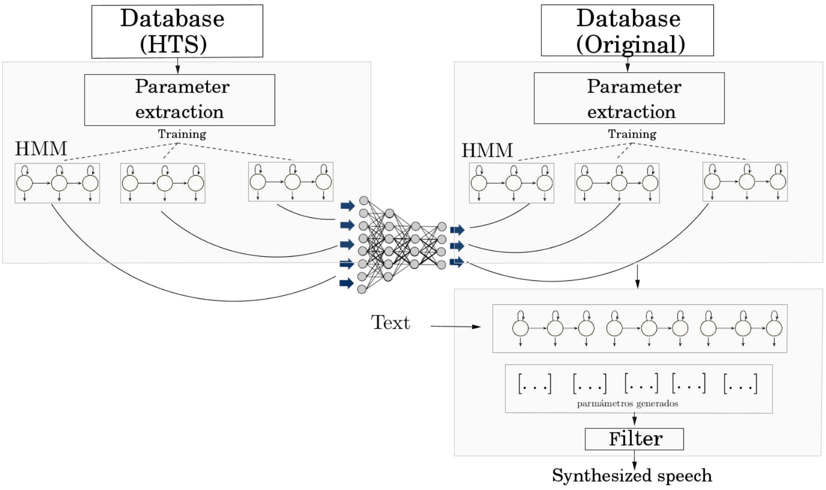

Fig. 3. Transforming HMM with deep learning algorithms for parametric speech synthesis (Type II)

From these results and following the path of incorporating mathematical models in speech synthesis that have been successful previously in recognition, emerges the possibility of improving the quality of artificial voices with the addition of deep learning algorithms. The two types of strategies are shown in Figures 2 and 3, in which HMM are substituted with deep learning networks (Type I) or are incorporated into transformation models (Type II).

Improvements that these algorithms can bring to speech synthesis are in at least one of the quality characteristics of speech synthesis based on HMM, for example:

- Greater naturalness and intelligibility
- Greater preference by users
- Greater capacity to produce emotive voices
- Lesser oversmoothing in spectral and f0
- Smaller requirements on database size
- Less complexity in the processes of creating voices

4 Research Methodology

Multiple steps are required to produce synthesized speech from an arbitrary text, using deep learning algorithms. From language modeling to evaluating results, arises the need to address the following stages:

4.1 Creation of Voice Databases

The first experiments carried out to synthesize voice, taking into account HMM based speech synthesis and deep learning, have been conducted on large databases with more than five hours of recordings. Therefore, it is necessary to analyze possible sources of data and appropriate phrases for use them in synthesis.

After the selection and processing of the voice recordings, it is necessary to separate sentences and carefully transcribe each one of them. The next step is to codify the features of the text that are necessary and sufficient enough to predict prosodic aspects and produce more natural speech across various contexts.

The most common features extracted from the speech signal arises from those used in speech recognition: Mel Frequency Cepstral Coecients (MFCC) and $f0$. A vector of parameters, and their first and second derivative discrete-approximation are extracted from each frame of 10 ms.

4.2 Segmentation of Audio and Parameters Extraction

When processing the phonetic units of speech that can be used in statistical parametric speech synthesis with deep learning algorithms, it is necessary to have methods to establish their boundaries within the recorded phrases, and then extract them into the corresponding parametric representation. There are several possibilities to use parameters that must be analyzed according to the viability of using them on the algorithms to implement.

In both Type I and Type II systems, coding can make a noticeable difference. It should be noted that there are also restrictions on the coding complexity raised from processing time and computer memory.

4.3 Algorithm Implementation

After having the appropriate representation of the speech signal in parameters, extensive experimentation should be performed on available systems to implement some of the proven deep learning algorithms in speech recognition: DNN, LSTM, BLSTM and RBM, and determine its feasibility to be incorporated into the development speech synthesis, in both Type I and Type II systems.

4.4 Evaluation of Results

There are different approaches to solving the problem of evaluating objectively synthesized speech. Since software is unavailable for this purpose, theory should be studied thoroughly and then implemented based on the proposed measures.

The subjective assessment, based on listeners who evaluated according to certain scales of perception, are often used to complement and validate the final audio results.

5 Main Contribution

This will be the first time that implementation of deep learning algorithms for both Type I and Type II speech synthesis systems will be conducted and documented in detail. With the results, we can provide a comparative analysis between both systems and propose full text to speech conversion systems with those models that have better functionality.

The required experimentation includes modeling phonetic units that have not previously been considered in parametric speech synthesis for the case of Spanish, such as syllables or groups of phonemes, which may constitute a contribution even for the synthesis based on HMM.

It is expected to obtain the first artificial voices in Spanish language using such algorithms, including the possibility of incorporating unused algorithms in other languages, such as Convolutional Networks, Deep Auto Encoders and LSTM.

We believe that the application of these algorithms can improve at least one aspect of quality of the voices that are currently obtained with statistical parametric speech synthesis based on HMM. Thereby enabling the widespread use of this technology in innovative applications, such as changing a voice accent and speech synthesis with emotions, for use in applications such as speech to speech translation in real time between different languages.

6 Results Achieved

Our first results with statistical parametric speech synthesis using HTS to produce voices in Mexican Spanish would have shown the advantages of using this technique, and have been documented in [22–24]. Those results were reported without any deep learning algorithms involved, and showed the improvement opportunities in que quality of the voices. Later, we experimented with Type II systems using Long Short-term Memory (LSTM) deep neural networks. The input of the network is the artificial voice produced with HTS, and the output is the corresponding natural voice, thus the network is trained to bring the sound of the synthesized speech closer to the natural voice. Figure 4 shows the current system development.

This is equivalent to find a mapping $y = f(x)$ between a degraded set of $n-$dimensional parameters x, from HTS voice, to a clean $n-$ dimensional parameters y of the natural voice. This is a similar approach to denoising in speech enhancement and speech recognition, with recent implementations of LSTM with successful results [25].

With this scheme, there have been notable improvements in the parameters corresponding to the spectrum of the signals. For example, Figure 5 shows the trajectory of one of the spectral parameters used, and shows how the LSTM improved artificial speech more closely resembles natural speech that the same parameter in the HTS voice.

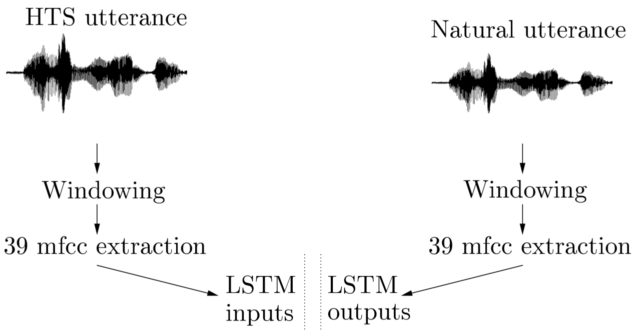

Fig. 4. Proposed system. HTS and Natural utterances are aligned frame by frame

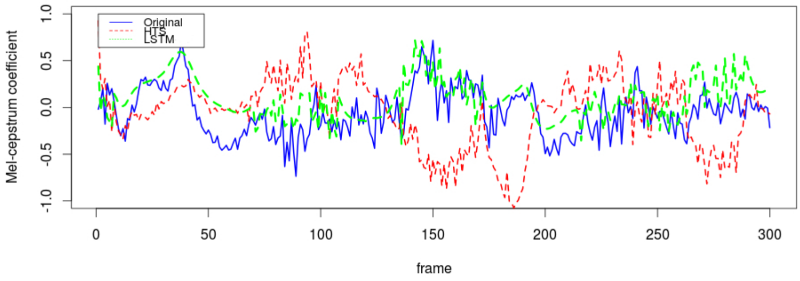

Fig. 5. Illustration of enhancing the 5th mel-cepstral coefficient trajectory by LSTM enhancement

Observations on other objective measures such as Mel-Cepstral Distance and similarity of spectrograms, also show significant improvements, indicating the improvement of the voice processed with LSTM.

7 Conclusions

We have presented a proposal for the implementation of deep learning algorithms to improve the results achieved so far in statistical parametric speech synthesis based on Hidden Markov Models.

The strategy is based on preparing frame aligned data of synthesized and natural speech. Further, we propose to use the algorithms to model the transformation of parameters of the artificial voice generated from the HTS system, the corresponding voices of the natural speech.

The results show improvement of spectral features of the voice in objective evaluation, but have to be tested with subjective assessments. New strategies for data modeling, algorithms and configurations have to be explored in order to improve other aspects of the speech signal, such as the fundamental frequency, energy and aperiodic coefficients that can further improve the sound of the artificial voice to a natural voice.

Acknowledgments

This work was supported by the SEP and CONACyT under the Program SEP-CONACyT, CB-2012-01, No.182432, in Mexico, as well as the University of Costa Rica in Costa Rica.

References

1. Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K.: Speech synthesis based on hidden markov models. In: Proceedings of the IEEE, vol. 101(5), pp. 1234–1252 (2013)
2. Zen, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7962–7966 (2013)
3. Zhen-Hua, L., Deng, L., Dong, Y.: Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. IEEE Transactions on Audio, Speech, and Language Processing 21(10), 2129-2139 (2013)
4. Shiyin, K., Qian, X., Hsiang-Yun, M.: Multi-distribution deep belief network for speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8012–8016 (2013)
5. Fan, Y., Qian, Y., Xie, F.L., Soong, F.K.: TTS synthesis with bidirectional LSTM based recurrent neural networks. Interspeech 1964–1968 (2014)
6. Zhizheng, W., Valentini-Botinhao, C., Watts, O., King, S.: Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4460–4464 (2015)
7. HMM-based Speech Synthesis System (HTS). `http://hts.sp.nitech.ac.jp/`
8. Tomoki, T., Tokuda, K.: A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE TRANSACTIONS on Information and Systems 90(5), 816–824 (2007)
9. Zen, H., Tokuda, K., Kitamura, T.: Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. Computer Speech & Language, 21(1), 153–173 (2007)
10. Shannon, M., Zen, H., Byrne, W.: Autoregressive models for statistical parametric speech synthesis. IEEE Transactions on Audio, Speech, and Language Processing 21(3), 587–597 (2013)
11. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. In: Proceedings of the IEEE, vol. 77(2), pp. 257–286 (1989)

12. Wellekens, C.: Explicit time correlation in hidden Markov models for speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 12, pp. 384–386 (1987)
13. Giuliani, D., Omologo, M., Svaizer, P.: Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation. In: Proceedings of the Fourth International Conference on Spoken Language, ICSLP 96, vol. 3, pp. 1329–1332 (1996)
14. Flores, J.A., Young, S.J.: Continuous speech recognition in noise using spectral subtraction and HMM adaptation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 409–412 (1994)
15. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine 29(6), 82–97 (2012)
16. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 20(1), 30–42 (2012)
17. Graves, A., Mohamed, A.R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649 (2013)
18. Graves, A., Eck, D., Beringer, N., Schmidhuber, J. : Biologically plausible speech recognition with LSTM neural nets. In: Biologically Inspired Approaches to Advanced Information Technology, Springer Berlin Heidelberg, pp. 127–136 (2004)
19. Seide, F., Gang, L., Dong, Y.: Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. Interspeech, 437–440 (2011)
20. Dong, Y., Chen, X., Deng, L.: Factorized deep neural networks for adaptive speech recognition. In: International workshop on statistical machine learning for speech processing (2012)
21. Dong, Y., Yao, K., Su, H., Li, G., Seide, F.: KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7893–7897 (2013)
22. Coto-Jiménez, M.: Síntesis estadística paramétrica de voz. Master thesis. Universidad Autónoma Metropolitana, Unidad Iztapalapa. México D.F., México (2014)
23. Coto-Jiménez, M., Goddard-Close, J. Martínez-Licona, F.M.: Quality Assessment of HMM-Based Speech Synthesis Using Acoustical Vowel Analysis. In: Speech and Computer. Springer International Publishing, pp. 368–375 (2014)
24. Coto-Jiménez, M., Martínez-Licona, F.M., Goddard-Close, J.: Acoustic Vowel Analysis in a Mexican Spanish HMM-based Speech Synthesis. Research in Computing Science 86, 53–62 (2014)
25. Coto-Jiménez, M., Goddard-Close, J., Martínez-Licona, F.M.: Improving Automatic Speech Recognition Containing Additive Noise Using Deep Denoising Autoencoders of LSTM Networks. In: 18th International Conference on Speech and Computer SPECOM (Forthcoming) (2016)