ISSN 2007-9737

Testing Three Different Speech Synthesizers to Acknowledge the Advantages of DNN Systems against HMM Methods

Carlos Ángel Franco-Galván[1,*], José Abel Herrera-Camacho[2]

[1] Facultad de Artes BUAP,
Mexico

[2] Universidad Nacional Autónoma de México,
Laboratorio de Tecnologías del Lenguaje,
Mexico

carlosangel.franco@correo.buap.mx, abelherrerac1@gmail.com

Abstract. This document reports MOS results after testing naturalness and expressiveness in three different speech synthesis systems. A first system is based on HMM, the second one combines HMM and DNN and the third one is solely based on DNN. According to the results, DNN systems outperform HMM systems in synthetic speech quality.

Keywords. Speech synthesis, voice parameterization, line spectral pair, hidden Markov models, deep neural networks.

1 Introduction

From 2014 to date, DNN based speech synthesis is the most accepted method in terms of efficiency, naturalness and even expression. This statement is supported by a study (Meyer, 2021) on speech synthesis with systems from 2014 to 2021. It was discovered, after looking at multiple MOS test results from different systems based on HMM and DNN phoneme selection methods that HMM based systems are becoming obsolete and giving up space to DNN synthesizers.

This tendency obliges current HMM based systems to be reviewed, analyzed and either updated or abandoned depending how they perform compared to the top ranked systems based on DNNs. The authors performed a series of MOS tests on three different systems: An LSP-HMM based synthesizer, An LSP-DNN based synthesizer and a solely DNN based synthesizer.

This paper is organized as follows: Section 1 exposed the arguments on how DNN gains terrain in speech synthesis, section 2 offers a brief overview on current systems, section 3 describes the first experiment, section 4 discusses the DNN performance and section 5 closes with a final argument.

2 Overview on Speech Synthesizers

The authors have been working on speech synthesis for the last 5 five years. Their work is mostly based on Hidden Markov Models as Text to Speech Synthesis (Zen et al., 2007). Extensively worked by (Keiichi Tokuda et al., 2013). This type of synthesizers require the voice to be parameterized using a mathematical representation of the frequency spectrum coefficients, most of such parameterizations are based on the Mel scale using cepstral coefficients (Ganchev, 2011).

To advance on this research, it was decided to use Linear Spectral Pair LSP parameterization (Backstrom, 2004) as an alternative to the traditional Mel-cepstral parameterizations (C. Franco-Galván & Herrera-Camacho, 2020) and (C. A. Franco-Galván, 2021). LSP parameterization

doi: 10.13053/CyS-29-3-4517

Table 1. MOS Results

Phrase 1 HMM	3.17
Phrase 2 HMM	3.16
Phrase 3 HMM	3.2
Phrase 1 DNN	3.56
Phrase 2 DNN	3.68
Phrase 3 DNN	3.6

was chosen because it economizes computational resources. The results were acceptable but not optimal and looking at the level of acceptation of DNN synthesis in other works (Watts et al., 2016), it was an academic necessity to compare the LSP-HMM system with the recent speech synthesizers based on Deep Neural Networks.

DNN are the new standard on speech synthesis since they perform a better parameter selection in the speech synthesis scheme. It must be considered that a DNN system requires to process multiple data in the shortest possible time. The parallel processing and recurrent information learning features of a Neural Network achieve what traditional concatenative synthesis methods could not.

Deep learning systems have demonstrated to be powerful tools in speech processing as stated by (Mehrish, et. al., 2023) these authors reviewed besides DNN, Convolutional Neural Networks CNN and Recurrent Neural networks RNN. Nevertheless, in their research, they claimed that deep learning systems “still faces certain challenges that need to be addressed”.

A remarkable system based on DNNs known as Tacotron (Shen et al., 2018) outperforms most of the systems at 2018 and has evolved into a synthesizer very well ranked in terms of naturalness and expressiveness. Google sponsors Tacotron and allows certain versions for experimentation proposes. According to Meyer (2021) this system’s last version is the best ranked in MOS tests with 4.54 out of 5.

More recently, a DNN system called FastSpeech (Ren et al., 2019) is a faster system and comparable quality, FastSpeech2 had added more quality and a faster system (Ren et al, 2021). Ther are other remarkable systems (Chen et al-, 2020; Luo et al.,2021; Ping et al., 2019; Prenger et al., 2019). The technique Zero-shot obtained highly natural speech for multiple speakers (Tan et al.,2021).

The authors propose a synthesizer using Linear Spectral Pair. This system combines DNN and HMM during the training stage. It is based on the model proposed by (Keiichi Tokuda, 2017), this system is denominated LSP-DNN. Its naturalness was tested the same way it was done with previous systems, using MOS tests. It is a scaled version of the LSP-HMM synthesizer (cite) the parameterization is consistent with former software versions, since LSP offers good quality in voice.

3 Advantages and Disadvantages of Both Systems

A MOS test was carried out comparing a previous LSP-HMM based synthesizer (Franco-Galvan et al., 2019) with the aforementioned system LSP-DNN. The aim for this experiment was to investigate the naturalness of the newer synthesizer. Three phrases were synthesized from text using both systems: the LSP-HMM synthesizer and the DNN-HMM synthesizer.

The phrases were played by a group of listeners and each listener validated the phrase on a scale from 0 to 5. Being 5 very natural and 0 no natural at all. the DNN based synthesizer overcame the voice produced by the LSP-HMM synthesizer. Graphical results are shown in figure 1. It can be seen from the results that LSP-DNN is over 0.3 points above LSP-HMM, the MOS on the phrases.

The results show that the system based solely on HMMs was outrun by the system combining DNNs. This result is consistent with other researchers’ results.

4 Validating a DNN System

An MOS test was carried out using a third system based on Tacotron, implemented in Laboratorio de Tecnologías del Lenguaje at UNAM (Herrera &

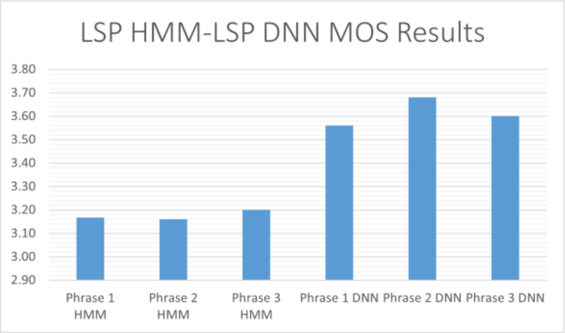

Fig. 1. MOS Results for LSP HMM and LSP DNN synthesized speech

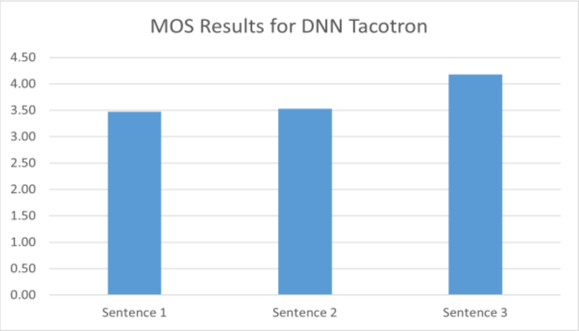

Fig. 2. MOS results for DNN Speech Synthesis

Table 2. MOS Results on DNN

Phrase 1	3.47
Phrase 2	3.53
Phrase 3	4.18

Morales, 2022). This experiment was conducted to make sure that DNN is much better than HMM to dismiss the last one and take a new route on research.

Just as it was stated in section 3 on this document, when comparing LSP-HMM and LSP-DNN, a MOS test was carried out. Three phrases were produced and validated in its naturalness from 0 to 5 by different users. The graphics are shown on figure 2. The results were:

The three phrases are 70% of the maximum mark, this condition helps us probe that DNN based synthesizers produce more natural phrases than their HMM counterpart. As it was discussed earlier in the document.

5 Final Discussion

As it is reported in other investigations (Kaur & Singh, 2022) HMM based synthesis is about to become obsolete, giving way to DNN based synthesis as the new standard in Text to Speech. Furthermore, the authors of original HTS system (K Tokuda et al., 2002) stopped updating their system since 2021. The last version HTS 2.3.2 uses DNN to complement to the HMM selection system and recently offered a system named DNN HSMM in which the HMM play a smaller role. Zen and colleagues (Zen et al., 2013) performed a series of experiments to mention the shortcomings of HMM synthesis and propose DNN synthesis as the direction to take.

The MOS scores obtain in this study support the opinions of previous researchers. The authors' system has yet to overcome certain elements to improve its quality. Nevertheless, it is worth mentioning that future efforts will be on the DNN based system leaving the HMM based system as historical references.

References

1. **Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., Poria, S. (2023).** A review of deep learning techniques for speech processing. Information Fusion, Vol. 99, 101869. Doi: 10.1016/j.inffus.2023.101869.
2. **Backstrom, T. (2004).** Linear predictive modelling of speech - constraints and line spectrum pair decomposition. Matrix. https://aaltodoc.aalto.fi/bitstream/handle/123456789/2392/isbn9512269473.pdf?sequence=1.
3. **Chen, M., Tan, X., Ren, Y., Xu, J., Sun, H., Zhao, S., Qin, T. (2020).** Multispeech: Multi-speaker text to speech with transformer. INTERSPEECH, pp. 4024–4028.
4. **Franco-Galván, C. A. (2021).** An Advanced Study to Validate Synthesized Speech Parameterized by Cepstral Coefficients and LSP. New Visions in Science and Technology, Vol. 9, pp. 35–44. Doi: 10.9734/bpi/nvst/v9/5329f.
5. **Franco-Galvan, C., Franco-Galvan, C. A., Herrera-Camacho, J. A., Escalante-Ramirez, B. (2019).** Application of Different Statistical Tests for Validation

of Synthesized Speech Parameterized by Cepstral Coefficients and LSP. Computación y Sistemas, Vol. 23, No. 2, pp. 461–468. Doi: 10.13053/cys-23-2-2977.

6. **Franco-Galván, C., Herrera-Camacho, J. A. (2020).** MOS validation on Synthesized Speech parameterized by Cepstral Coefficients and LSP. Vol. 22, No. 6, pp. 14–18. Doi: 10.9790/0661-2206011418.
7. **Ganchev, T. (2011).** Contemporary methods for speech parameterization. Springer.
8. **Herrera, A., Morales, E. (2022).** DNN Synthetizer for Mexican Spanish Language. In: Procceding 30ª Reunión de Otoño de Comunicaciones, Computación, Electrónica y Exposición Industrial, ROC&C'2021, Vol. 1, No. 6, pp. 4. http://rvp.ieee.org.mx/assets/ponencia-rocc-96.pdf
9. **Kaur, N., Singh, P. (2022).** Conventional and contemporary approaches used in text to speech synthesis: a review. In: Artificial Intelligence Review Springer Netherlands. Doi: 10.1007/s10462-022-10315-0.
10. **Meyer, P. (2021).** State of the Art of Speech Synthesis at the End of May 2021. Towards Data Science. https://towardsdatascience.com/state-of-the-art-of-speech-synthesis-at-the-end-of-may-2021.
11. **Ping, W., Peng, K., Chen, J. (2019).** Clarinet: Parallel wave generation in end-to-end text-to speech. International Conference on Learning Representations.
12. **Prenger, R., Valle, R., Catanzaro, B. (2019).** Waveglow: A flow-based generative network for speech synthesis. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3617–3621.
13. **Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T-Y (2019).** Fastspeech: Fast, robust and controllabe text to speech. Advances in Neural Information Processing, pp. 3165–3174.
14. **Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y. (2021).** Fastspeech 2: Fast and high-quality end-to-end text to speech. International Conference on Learning Representations, https://arxiv.org/pdf/2006.04558.
15. **Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., Wu, Y. (2018).** Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 4779–4783. Doi: 10.1109/ICASSP.2018.8461368.
16. **Tan, X., Qin, T., Soong, F., Liu, T.-Y. (2021).** A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561.
17. **Tokuda, K., Zen, H., Black, A. (2002).** An HMM-based speech synthesis system applied to English. In: IEEE Speech Synthesis Workshop. http://www.scs.cmu.edu/afs/cs.cmu.edu/Web/People/awb/papers/IEEE2002/hmmenglish.pdf.
18. **Tokuda, K. (2017).** HMM/DNN-based Speech Synthesis System (HTS). http://hts.sp.nitech.ac.jp/.
19. **Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K. (2013).** Speech Synthesis Based on Hidden Markov Models. In: Proceedings of the IEEE, Vol. 101, No. 5, pp. 1234–1252. Doi: 10.1109/JPROC.2013.2251852.
20. **Watts, O., Henter, G. E., Merritt, T., Wu, Z., King, S. (2016).** From HMMS to DNNS: Where do the improvements come from?. In: Proceedings (ICASSP) IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5505–5509. Doi: 10.1109/ICASSP.2016.7472730.
21. **Zen, H., Senior, A., Schuster, M. (2013).** Statistical parametric speech synthesis using deep neural networks. In: Proceedings ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7962–7966. Doi: 10.1109/ICASSP.2013.6639215.
22. **Zen, H., Tokuda, K., Black, A. (2007).** Statistical parametric speech synthesis. Speech Communication, pp. 1229–1232. Doi: 10.1016/j.specom.2009.04.004.

Article received on 17/02/2023; accepted 13/01/2025.
**Corresponding author is Carlos Ángel Franco-Galván.*