Measuring the Effect of Reverberation on Statistical Parametric Speech Synthesis

Marvin Coto-Jiménez(✉)

PRIS-Lab, Escuela de Ingeniería Eléctrica, Universidad de Costa Rica,
San Pedro, Costa Rica
marvin.coto@ucr.ac.cr

Abstract. Text-to-speech (TTS) synthesis is the technique of generating intelligible speech from a given text. The most recent techniques for TTS are based on machine learning, implementing systems which learn linguistic specifications and their corresponding parameters of the speech signal. Given the growing interest in implementing verbal communication systems in different devices, such as cell phones, car navigation system and personal assistants, it is important to use speech data from many sources. The speech recordings available for this purpose are not always generated with the best quality. For example, if an artificial voice is created from historical recordings, or a voice created from a person whom only a small set of recordings exists. In these cases, there is an additional challenge due to the adverse conditions in the data. Reverberation is one of the conditions that can be found in these cases, a product of the different trajectories that a speech signal can take in an environment before registering through a microphone. In the present work, we quantitatively explore the effect of different levels of reverberation on the quality of artificial voice generated with those references. The results show that the quality of the generated artificial speech is affected considerably with any level of reverberation. Thus, the application of algorithms for speech enhancement must be taken always into consideration before and after any process of TTS.

Keywords: Hidden Markov Models · PESQ · Reverberation · Speech synthesis

1 Introduction

Text-to-speech (TTS) synthesis is the technique created for the generation of artificial, intelligible speech from any given text [15], usually from computers or high technology devices. There are many implementations of TTS in commercial applications and many potential areas where it can be applied. For example,

Supported by the University of Costa Rica.

J. L. Crespo-Mariño and E. Meneses-Rojas (Eds.): CARLA 2019, CCIS 1087, pp. 369–382, 2020.
https://doi.org/10.1007/978-3-030-41005-6_25

any circumstance that requires the transfer of information between people and machines is a potential application. One of the main advantages of applying TTS for this purpose is the fact that speech is the most widely used communication method between humans. Additionally, verbal communication is natural and requires no special training [4].

TTS systems are divided into two main components [7]: A "front end", where the text is processed to produce a linguistic specification, so the units of speech (such as phonemes or syllables) can be described in terms of their surrounding components, and a "back end", that take the linguistic specification as input and generates a waveform.

The development of TTS has evolved from the creation of isolated words or phrases to general purpose voices in different languages, with different styles and emotions [1,3]. There is a significant effort in research to obtain improvements in the multiple challenges that TTS systems have today, as its extensive use in applications depends on obtaining more natural and close-to-human voices.

The most recent techniques to generate TTS have emerged from the idea of machine learning algorithms applied to store and reproduce parameters of the speech [19–21]. The first model that successfully applied those techniques was the Hidden Markov Models (HMM), learning parameters such as fundamental frequency (f_0) and Mel-Frequency Cepstral Coefficients (MFCC). This set of parameters and models were known as Statistical Parametric Speech Synthesis. More recently, Deep Learning-based algorithms have been applied to voice generation from text [9,12], or as post-filter to the results obtained with HMM [2,11].

Previous references have reported a significant quality drop in artificial speech when the training parameters of the speech data are noisy. This condition requires the compensation of the voice signals with several techniques [6,17,18]. For example, speech enhancement algorithms can be used to clean the available noisy data.

This problem has been addressed in several references, but only some of them have objectively measured the impact of specific conditions, particularly noise [10]. The interest in predicting the effects of different degrees of reverberation in the results of statistical parametric speech synthesis relies on the prior evaluation of usability for future experiences with speech synthesis.

For this purpose, in this work we want to address the impact of reverberation on objective quality measures in speech synthesis, in comparison to those produced with clean speech.

To answer this question, we made several experiments with different conditions of reverberation, and measure the impact between clean and reverberated speech, and between the artificial speech generated with both.

The rest of this paper is organized as follows: Sect. 2 gives the background and context of the problem. Section 3 describes the experimental setup, Sect. 4 presents the results with a discussion, and finally, in Sect. 5, we present the conclusions.

2 Background

2.1 Hidden Markov Models

Hidden Markov Models (HMM) can be described from a Markov process, in which state transitions are given by probability. There is a second process described with probability, which models the emission of symbols when it comes to each state, according to probabilistic rules. There are several kinds of HMM, applied to model many important areas.

In Fig. 1, a representation of a particular HMM, known as a left-to-right, is shown. This is the most common type of HMM applied in speech technologies. Here, the first state is at the left, from which transitions can occur. These transitions lead to the same state or to the next on the right, according to some probability p_{ij}. Transitions cannot occur in the reverse direction.

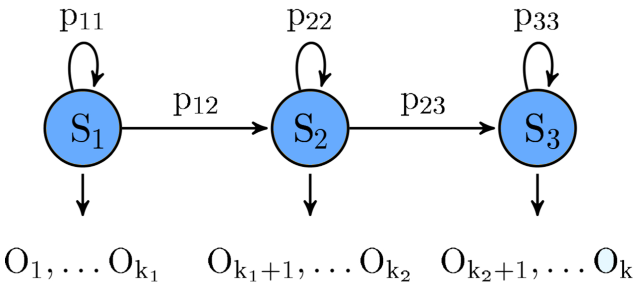


Fig. 1. Left to right example of an HMM with three states. O_k represents the observation emitted in state k.

An HMM can mathematically be described by a tuple:

$$\lambda = (S, \pi_i, a, b) \tag{1}$$

where S is the set of states, π a probability vector that establishes the probability of i to be the initial state. a is the transition probability matrix between states, and b the probabilistic rule of observations of specific symbols in each state.

2.2 Statistical Parametric Speech Synthesis

Statistical parametric speech synthesis based on HMM follows a procedure with a training part and a synthesis part. The training part requires recordings of speech and their corresponding text transcriptions. This data is presented to a set of HMM (or other machine learning algorithms) that learns the parameters corresponding to a certain sound of the speech.

In the synthesis part, any text can be applied to the models, which output the corresponding parameters to the specific sounds of the utterance, and then a filter produces the waveform. This scheme has been applied since the creation of the HMM-Based Speech Synthesis (HTS) System [16,24] for several languages, and allows specific definition for phonetic units, customizing training parameters according to needs and the amount of available data.

For applications of speech recognition and synthesis, the probabilistic rule at the output of each state of a HMM, named b in Eq. 1 is assumed as a multivariate Gaussian distribution defined as:

$$b_i(\boldsymbol{o}_t) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_i|}} \exp\left\{\frac{-1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{o}_t) - \boldsymbol{\mu}_i\right\} \tag{2}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are mean vector and covariance matrix, respectively. d is the dimension of vector of acoustic parameters, and $\boldsymbol{o}_t t$ is an observation vector of parameters at frame t.

The training process of a HMMs for a speech synthesis application can be described as finding the best parameters of λ given observed parameters of the speech (O). This process can be written as:

$$\lambda_{max} = \arg\max_{\lambda} p(O|\lambda, W), \tag{3}$$

where p is probability and W a specific word or sound.

In the synthesis part, the problem of getting the best parameters related to a given W which need to be synthesized can be stated as:

$$o_{max} = \arg\max_{o} p(o|\lambda_{max}, w) \tag{4}$$

In the following sections, we describe the application of these models to produce artificial speech and study the influence of reverberating conditions in training.

3 Experiments

In order to test the effects of reverberated speech to Statistical Parametric Speech Synthesis based on HMM, the experimental setup can be summarized in the following steps:

3.1 Database

For the experimentation, we used the SLT voice of the CMU Arctic databases, developed at the Language Technologies Institute at Carnegie Mellon University [8]. This database was specifically designed for research in speech synthesis. It consists of a number 1150 utterances selected from out-of-copyright texts from Project Gutenberg.

For degrade this data with reverberation, we use five impulse responses from the MARDY database [22] and the Center for Digital Music (C4DM) at Queen Mary, University of London [14].

The following nomenclature will be used for each condition:

- MARDY, from the corresponding database.
- GH (Great Hall), from the C4DM database.

- OC (Octagon), from the C4DM database.
- CR1 y CR2 (Classroom 1 y 2), from the C4DM database.

The speech files of the CMU database were convolved with the impulse responses of each condition. The output is the speech signal with the reverberation of the space where the impulse response was recorded.

3.2 Synthesis of Reverberated Speech

With the clean version of the SLT/CMU voice, an artificial voice where build using the HTS system [23]. To compare the influence of the different reverberating cases, the HMM-based synthetic voices were produced with each of the five conditions after the convolution: MARDY, GH, OC, CR1, CR2.

A set of comparisons between clear speech, artificial speech produced with the clear speech, artificial speech produced with reverberated speech and the reverberated speech were performed. This comparison was made to measure the effect of reverberation before and after the process to produce artificial speech.

Figure 2 illustrates the general procedure for each of the conditions of reverberation.

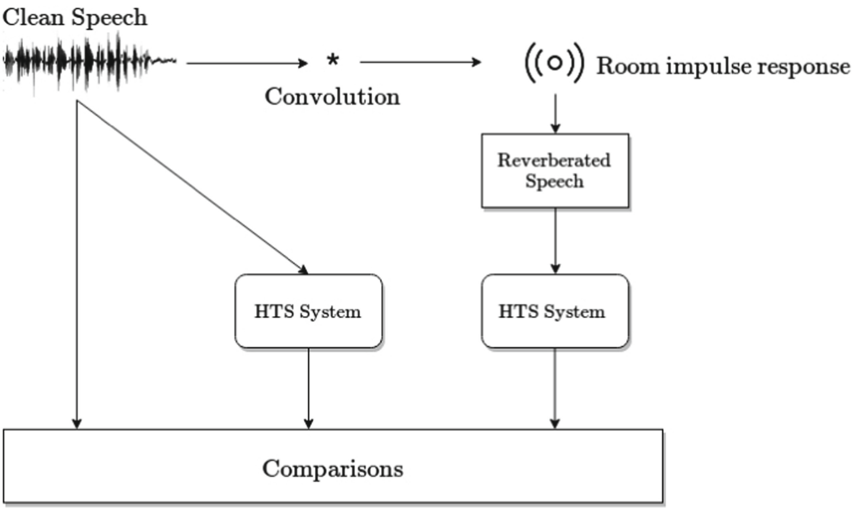


Fig. 2. Procedure to obtain and compare reverberated and artificial speech.

3.3 Evaluation

To evaluate the results given from our experiments, we use the PESQ (Perceptual Evaluation of Speech Quality), defined in the ITU-T recommendation

P.862.ITU. Results are given in interval $[0.5, 4.5]$, where 4.5 corresponds to a perfect reconstruction of the signal. PESQ is computed as [13]:

$$\text{PESQ} = a_0 + a_1 D_{ind} + a_2 A_{ind} \tag{5}$$

where the D_{ind} is the average disturbance and A_{ind} the asymmetrical disturbance. The a_k are chosen to optimize PESQ in measuring speech distortion, noise distortion, and overall quality.

We also use the MOS-LQO (Mean Opinion Score - Listening Quality Objective) measure, performing a mapping function from the PESQ, by the relation

$$\text{MOS-LQO} = 0.9999 + \frac{4.999 - 0.999}{1 + e^{-1.4945 \cdot \text{PESQ} + 4.6607}}, \tag{6}$$

according to the ITU-T P.862.1 [5].

We are interested in measuring the effects of reverberation in the speech signals before and after the process of generating artificial speech with the HTS System. To perform these measures, we applied the following comparisons between groups of utterances:

- Natural speech and HTS voice produced with natural speech.
- Natural speech and reverberated speech.
- Natural speech and HTS voice produced with reverberated speech.
- HTS voice produced with natural speech and HTS voice produced with reverberated speech.
- Reverberated speech and HTS voice produced with reverberated speech.

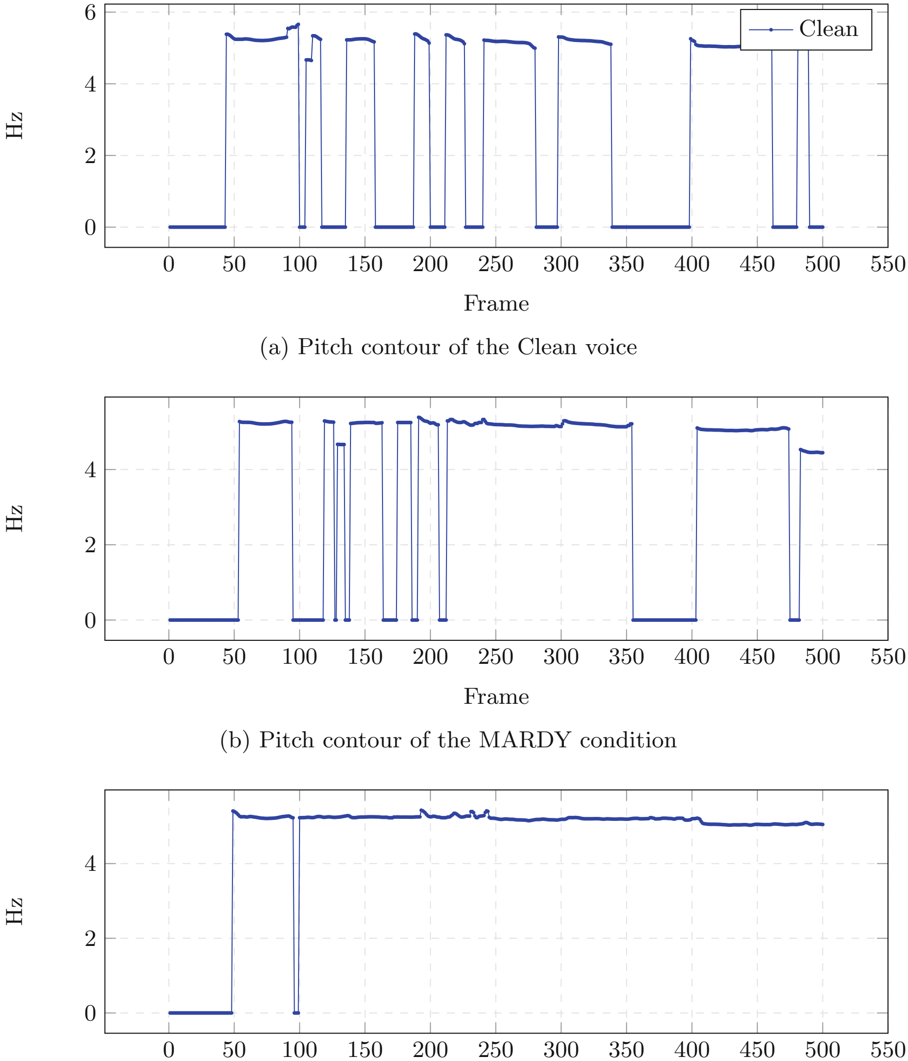
Besides those five comparisons, there are other possible combinations that do not give information about the effects on artificial voice generation. For each of the five cases of reverberation, we compare the PESQ measure. Additionally, we report spectrograms and pitch contours for direct visualization of the results.

4 Results and Discussion

In this section, we show the influence of the different reverberations on clean and artificial speech. The reverberation in speech signals greatly affects the estimation of the pitch, which is one of the most important parameters for speech recognition and generation.

For example, in Fig. 3 it is noticeable how the reverberation produces more voiced frames (those with positive values for pitch) in the MARDY condition. The GH, with a bigger degree of reverberation, almost produces only voiced frames, introducing great distortion and affecting the quality of the speech.

The spectrograms also show different levels of distortion when compared to the Clean voice and the correspondent artificial voice 4. For example, Fig. 5 show some recognizable characteristics of the spectrum in the MARDY condition, which seems to produce some light distortions in the artificial voice constructed from this data.



(a) Pitch contour of the Clean voice

(b) Pitch contour of the MARDY condition

(c) Pitch contour of the GH condition

Fig. 3. Comparison of pitch contours for clean voice and two reverberating conditions in the utterance: "Author of the danger trail, Philip Steels, etc."

On the other hand, Fig. 6 shows evident degradation of the signal with the OC condition and almost unrecognizable spectrum in the artificial speech. From this spectrograms, it is remarkable how different levels of reverberation can affect the quality of artificial speech.

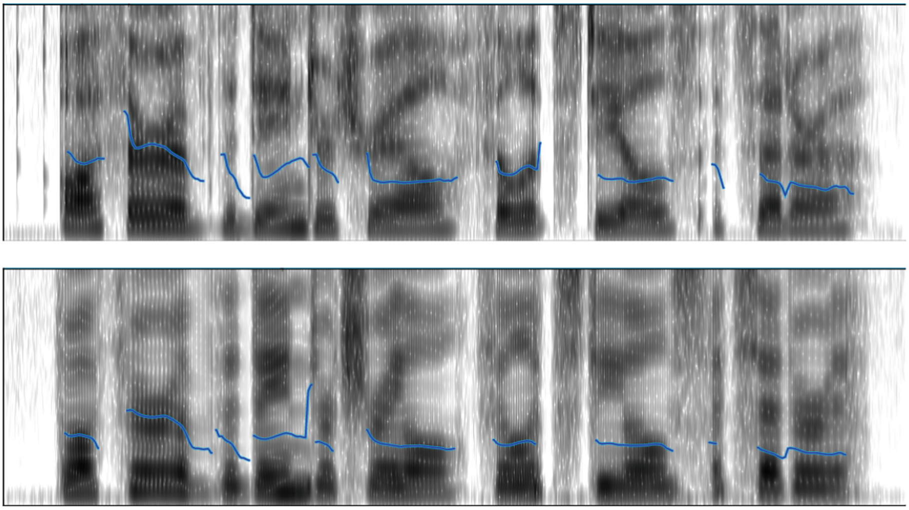


Fig. 4. Spectrograms of the utterance "Not at this particular case, Tom, apologized Wittmore“, with the Clean Voice (at the top) and artificial voice (at the bottom). Pitch contour is also highlighted.

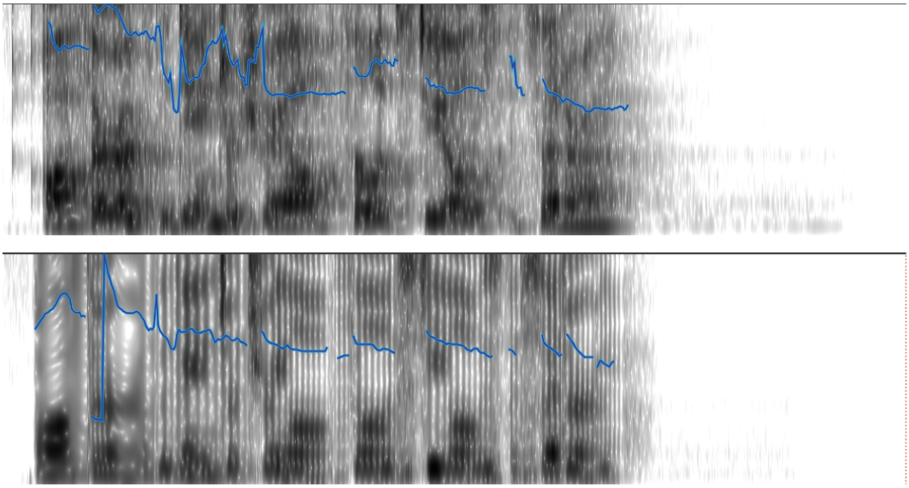


Fig. 5. Spectrograms of the utterance "Not at this particular case, Tom, apologized Wittmore", with reverberated voice with MARDY condition (at the top) and artificial voice produced with this reverberation (at the bottom). Pitch contour is also highlighted.

The results and comparisons for the PESQ measure are presented in form or radar plots. The radar plots allow the comparison between all the measures indicated in Sect. 3.3. The more contracted the radar plot, the lower perceptual quality in the reverberated and artificial voice. All the plots have the same scale.

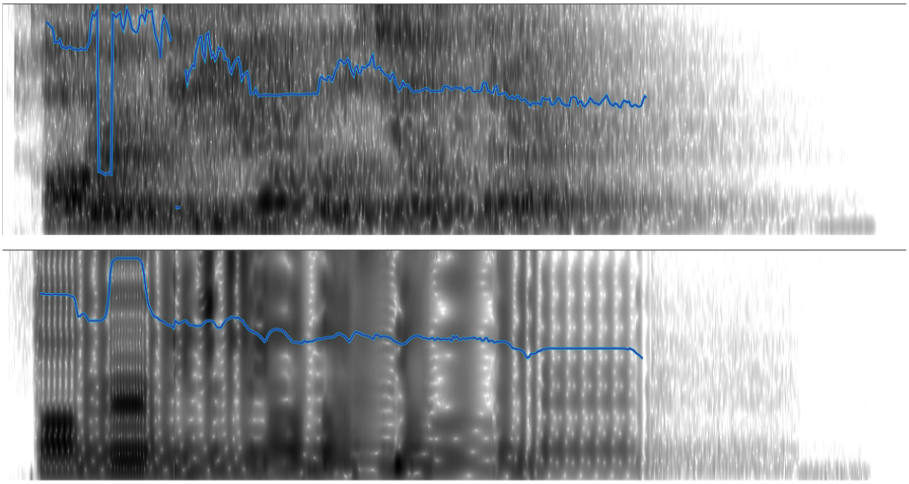


Fig. 6. Spectrograms of the utterance “Not at this particular case, Tom, apologized Wittmore”, with reverberated voice with OC condition (at the top) and artificial voice produced with this reverberation (at the bottom). Pitch contour is also highlighted.

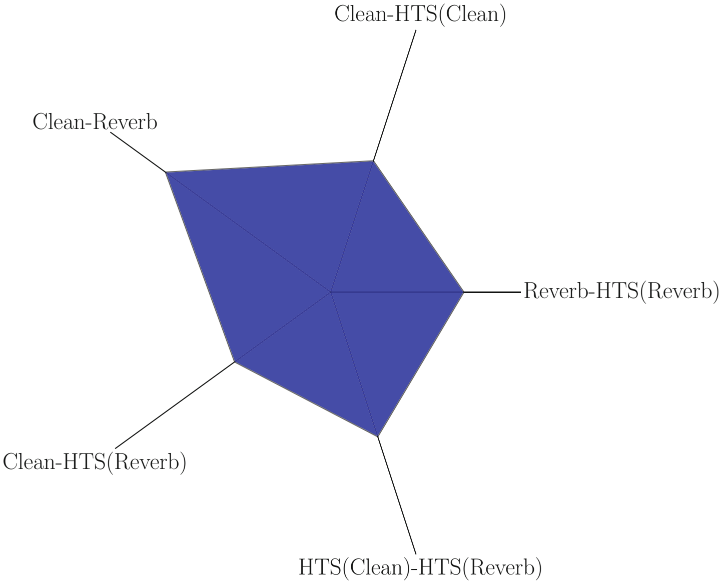


Fig. 7. Radar plot of Mean PSQ Values for MARDY Condition

Figure 7 shows the radar plot for the MARDY reverberation condition. As shown previously, this is the case where the reverberation produces lower distortion on the signal. When compared to the rest of the radar plots, this is the less contracted plot.

The radar plot for the Octagon condition (Fig. 8) shows a smaller value of PESQ for the reverberated voice. This lower quality also influences the lower perceptual quality for synthetic speech in relation to natural and artificial speech produced without reverberation.

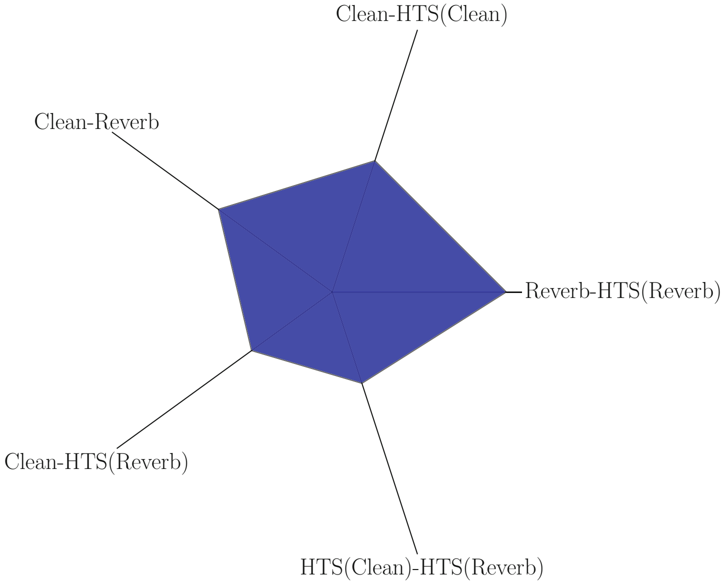


Fig. 8. Radar plot of Mean PSQ Values for Octagon Condition

The GH reverberation produces a degradation of the signal which heavily affects all the process, from the reverberated speech to the synthetic speech. As shown in Fig. 9, this is the most contracted plot in terms of all categories of speech without reverberation. According to these plots, this seems to be the condition that affects more the speech signal and the correspondent artificial voice.

Finally, the two CR conditions (Figs. 10 and 11) show similar degrees of reverberation and similar degradation on the perceptual quality of artificial speech. In comparison with GH, OC presents lower PESQ when compared the reverberated signal with the clean speech, and a better measure in the comparison of the reverberated signal and the artificial speech.

The results of the MOS-LQ measure, obtained from Eq. 6 are presented in Table 1. The greater effect on this measure before the generation of synthetic speech tend to produce bigger negative effects on the results. But the relationship does not seem to be linear.

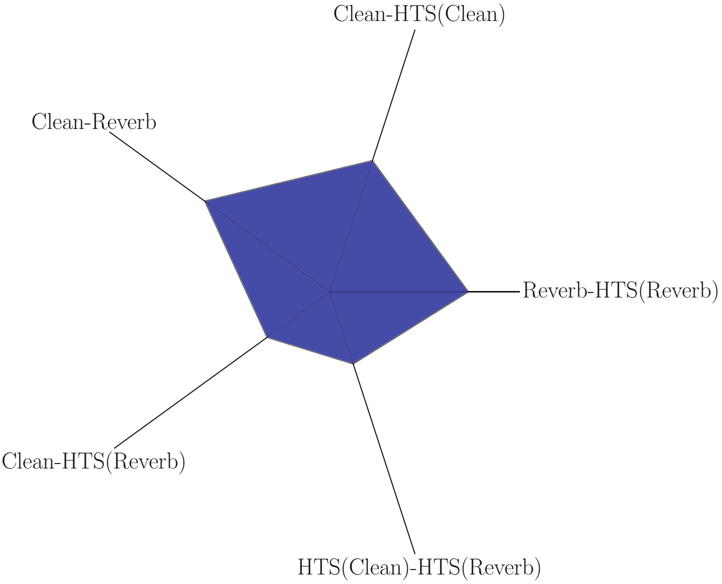


Fig. 9. Radar plot of Mean PSQ Values for GH Condition

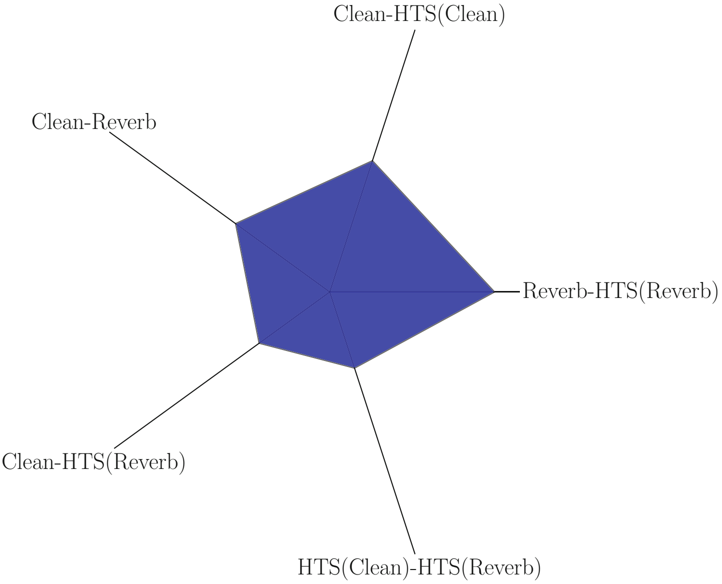


Fig. 10. Radar plot of Mean PSQ Values for CR1 Condition

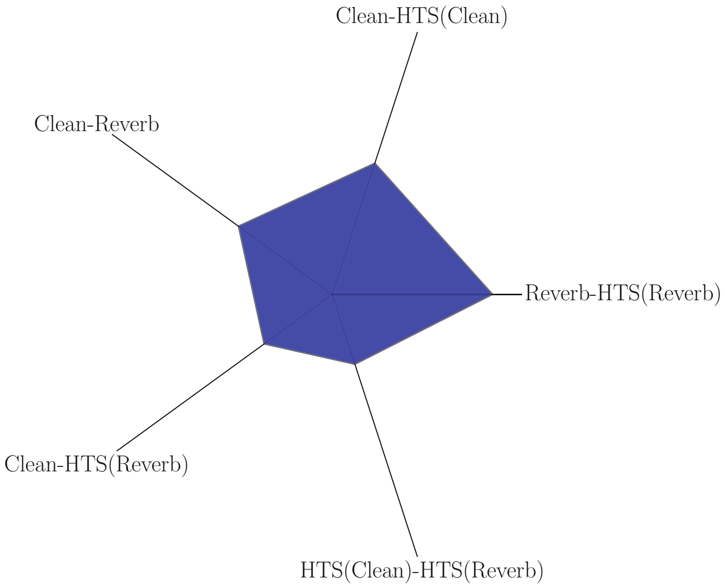


Fig. 11. Radar plot of Mean PSQ Values for CR2 Condition

Table 1. MOS-LQ values from the different cases of reverberation. The results are ordered from worst to best level of reverberation. Clean voice does not have MOS-LQ for being the reference.

Reverberation	MOS-LQ reverberated speech	MOS-LQ HTS
Clean	-	1.30
CR1	1.18	1.15
CR1	1.18	1.12
OC	1.26	1.13
GH	1.30	1.11
MARDY	1.56	1.16

All cases of reverberation produce artificial voice with lower MOS-LQ value than those produced with clean speech. But, different degrees of reverberation produces similar degradation, according to this measure. Being the reverberation a non-additive process, the results show also a complex relationship between the source speech and the result of the statistical parametric speech.

5 Conclusions

In this paper, it was explored the effects of reverberated speech on the creation of artificial voices obtained with statistical parametric techniques, based on Hidden Markov Models. The importance of this research relies on the application of objective measure to the quality of speech before and after the process of generating artificial voices.

For comparison purposes, we proposed the application of radar plots for the multiple visualizations of PESQ measures on all the relevant combinations of clean/artificial speech. These plots show how different levels of reverberation affects the signal before and after the generation of voices with the HTS system.

The results show that reverberation in all analyzed degree is an undesirable condition for the generation of artificial voices with statistical parametric techniques. Particularly for the effects on pitch detection.

This knowledge allows the discrimination of future sources of speech for generating synthetic voices. Having all degrees of reverberation significant negative effects on the quality of synthetic speech, it is critical for the speech synthesis the use of de-reverberation or enhancement procedures before the application of machine learning models.

For future work, new quality measures and more conditions of reverberation can be included. Additionally, statistical validation of results and extended graphical evidence of the degraded signals of natural and artificial speech.

Acknowledgements. This work was supported by the University of Costa Rica (UCR), Project No. 322-B9-105.

References

1. Black, A.W.: Unit selection and emotional speech. In: Eighth European Conference on Speech Communication and Technology (2003)
2. Coto-Jiménez, M.: Improving post-filtering of artificial speech using pre-trained LSTM neural networks. Biomimetics **4**(2), 39 (2019)
3. Coto-Jiménez, M., Goddard-Close, J.: LSTM deep neural networks postfiltering for enhancing synthetic voices. Int. J. Pattern Recognit Artif Intell. **32**(01), 1860008 (2018)
4. Holmes, W.: Speech Synthesis and Recognition. CRC Press, Boca Raton (2001)
5. ITU-T, R.P.: 862.1: Mapping function for transforming P. 862 raw result scores to MOS-LQO. International Telecommunication Union, Geneva, Switzerland, November 2003 (2003)
6. Karhila, R., Remes, U., Kurimo, M.: Noise in HMM-based speech synthesis adaptation: analysis, evaluation methods and experiments. IEEE J. Sel. Top. Signal Process. **8**(2), 285–295 (2013)
7. King, S.: Measuring a decade of progress in text-to-speech. Loquens **1**(1), e006 (2014)
8. Kominek, J., Black, A.W.: The CMU arctic speech databases. In: Fifth ISCA Workshop on Speech Synthesis (2004)
9. Lee, J., Song, K., Noh, K., Park, T.J., Chang, J.H.: DNN based multi-speaker speech synthesis with temporal auxiliary speaker id embedding. In: 2019 International Conference on Electronics, Information, and Communication (ICEIC), pp. 1–4. IEEE (2019)
10. Moreno Pimentel, J., et al.: Effects of noise on a speaker-adaptive statistical speech synthesis system (2014)

11. Öztürk, M.G., Ulusoy, O., Demiroglu, C.: DNN-based speaker-adaptive postfiltering with limited adaptation data for statistical speech synthesis systems. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7030–7034. IEEE (2019)
12. Prenger, R., Valle, R., Catanzaro, B.: WaveGlow: a flow-based generative network for speech synthesis. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3617–3621. IEEE (2019)
13. Rix, A.W., Hollier, M.P., Hekstra, A.P., Beerends, J.G.: Perceptual evaluation of speech quality (PESQ) the new itu standard for end-to-end speech quality assessment Part I-time-delay compensation. J. Audio Eng. Soc. **50**(10), 755–764 (2002)
14. Stewart, R., Sandler, M.: Database of omnidirectional and B-format room impulse responses. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 165–168. IEEE (2010)
15. Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K.: Speech synthesis based on hidden Markov models. Proc. IEEE **101**(5), 1234–1252 (2013)
16. Tokuda, K., Zen, H., Black, A.W.: An HMM-based speech synthesis system applied to English. In: IEEE Speech Synthesis Workshop, pp. 227–230 (2002)
17. Valentini-Botinhao, C., Wang, X., Takaki, S., Yamagishi, J.: Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks. In: Interspeech, pp. 352–356 (2016)
18. Valentini-Botinhao, C., Yamagishi, J.: Speech enhancement of noisy and reverberant speech for text-to-speech. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(8), 1420–1433 (2018)
19. Valin, J.M., Skoglund, J.: LPCNet: improving neural speech synthesis through linear prediction. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5891–5895. IEEE (2019)
20. Wang, X., Lorenzo-Trueba, J., Takaki, S., Juvela, L., Yamagishi, J.: A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4804–4808. IEEE (2018)
21. Wang, X., Takaki, S., Yamagishi, J.: Investigating very deep highway networks for parametric speech synthesis. Speech Commun. **96**, 1–9 (2018)
22. Wen, J.Y., Gaubitch, N.D., Habets, E.A., Myatt, T., Naylor, P.A.: Evaluation of speech dereverberation algorithms using the MARDY database. In: Proceedings of the International Workshop Acoustic Echo Noise Control (IWAENC). Citeseer (2006)
23. Zen, H., et al.: The HMM-based speech synthesis system (HTS) version 2.0. In: SSW, pp. 294–299. Citeseer (2007)
24. Zen, H., et al.: Recent development of the HMM-based speech synthesis system (HTS) (2009)