ISSN 2007-9737

Measuring the Quality of Low-Resourced Statistical Parametric Speech Synthesis Trained with Noise-Degraded Data Supported by the University of Costa Rica

Marvin Coto-Jiménez

University of Costa Rica,
Costa Rica

marvin.coto@ucr.ac.cr

Abstract. After the successful implementation of speech synthesis in several languages, the study of robustness became an important topic so as to increase the possibility of building voices from non-standard sources, e.g. historical recordings, children's speech, and data freely available on the Internet. In this work, a measure of the influence of noise in the source speech of the statistical parametric speech synthesis system based on HMM is performed, for a case of a low-resourced database. For this purpose, three types of additive noise were considered at five signal-to-noise ratio levels to affect the source speech data. Using objective measures to assess the perceptual quality of the results and the propagation of the noise through all the processes of building speech synthesis, the results show a severe drop in the quality of artificial speech, even for the cases of lower levels of noise. Such degradation seems to be independent of the noise type, and is at lower proportion to the noise level. This results are of importance for any practical implementation of speech synthesis from degraded data in similar conditions, and shows that applying denoising processes became mandatory in order to keep the possibility of building intelligible voices.

Keywords. Noise, robustness, speech synthesis.

1 Introduction

The purpose of speech synthesis can be established as the production of artificial speech from a given text input using computers. The resulting speech should be perceived with intelligibility and naturalness, in order to apply the results in the desired application. This process of speech synthesis (also referred to as text-to-speech) has a long history, from early mechanic systems to our days, where complex techniques and the release of dedicated software have extended the speech synthesis possibilities to many languages and applications.

The evolution of modern techniques can be traced back to the early 1970s [1], where the waveform generation was made using low-dimensional information, such as formants. And it has evolved to perform direct manipulations of waveforms (e.g. concatenative and unit selection approaches) or high dimensional parameters and deep learning-based models.

The statistical models of speech synthesis, mainly based on Hidden Markov Models (HMM), were popularized among researchers of the field after the first publications of the technique [2, 3], particularly after the release of the HTS software [4]. HMMs were previously successfully applied to speech recognition, and many of the ideas and parameters applied for that task were translated to the speech synthesis field.

With the HTS software, many papers were published on the implementation of statistical parametric speech synthesis in several languages around the world. The case of Spanish was also reported by a reduced number of researchers [5, 6, 7].

The advantages of statistical parametric speech synthesis based on HMM were reported in terms of its flexibility and capacity for producing intelligible

voices with low-training data [8]. The main disadvantages were the buzzy, muffled sound often reported.

With the increased performance and success of deep learning in several fields during the last decade, speech synthesis also benefits from the possibilities of the complex modeling and effective training algorithms of deep neural networks. The first ideas on the implementation of deep learning in speech synthesis were published in [9].

In previous years, many proposals have been made to apply different types of neural networks, such as Restrictive Boltzmann Machines, Deep Belief Networks, Bidirectional Long Short-term Memory Neural Networks, and Convolutional Neural Networks [10]. In some recent reports, the combination of both statistical parametric modeling combined with deep learning was also published [11, 12].

Typically, the deep learning-based approaches report a higher quality of results but require a large amount of training data. There are many situations where the availability of such resources is not possible to achieve. For example, in building speech from historical recordings, children's speech and low resourced languages [13, 14].

For these cases, HMM-based statistical parametric speech synthesis remains the main possibility to produce intelligible artificial voices. In many of such cases, the quality of the recordings was also a shortcoming for the quality of the results.

The usual framework in the building of synthetic voices was considered in the vast majority of cases: the recording of datasets in highly-controlled environments, which has typically done in professional studios with high-quality equipment. According to to [15], given the advances in speech synthesis techniques, the research community can consider building quality voices from data collected in less controlled environments. These new conditions represent several challenges for the process, for example, non-consistent recording conditions, unbalanced phonetic material, and noisy data. It is still not clear how robust speech systems are under such unfavorable conditions [16].

The problem of producing artificial speech has been addressed by some authors with particular interests in techniques that take advantage of a large corpus of clean data, such as speaker-adaptation in HMM-based speech synthesis. Using such corpus new voices can be built by incorporating information from the corpus in the smaller datasets.

For example, in [17], the authors proved that naturalness is not significantly affected by the presence of noise in the smaller dataset. The unfavorable conditions can be presented in found data, i.e. data freely available on the web. Such data has significant variation in terms of speaking style and channel characteristics [18].

In this paper, an experimental study on the influence of noisy recordings in the results of statistical parametric speech synthesis is performed, for the case of a small database in Spanish. The purpose of the study is to numerically report and compare the influence of several types and levels of noise in the speech data required to produce artificial speech.

The influence of the noise provide information to anticipate the quality of artificial speech that can be produced from recordings with unfavorable conditions. Such information is relevant for the evaluation of low-quality sources of speech resources in building speech synthesis.

The rest of this paper is organized as follows: Section 2 presents the theoretical background of speech synthesis and the effects of noise. Section 3 presents the experimental setup of the proposal. Section 4 presents the results. Finally, the Conclusions are presented in Section 5.

2 Statistical Parametric Speech Synthesis

The Statistical Parametric Speech Synthesis based on HMMs models the speech production process using the source-filter theory of voice production [1]. This model comprises the voicing information using fundamental frequency (or the logarithm of this measure) and the spectral envelope, commonly represented by mel-frequency cepstral coefficients (MFCC). The

speech waveforms are reconstructed from sequences of such parameters, and additional information about dynamic features (e.g. rate of change in form of delta and delta delta features [19]).

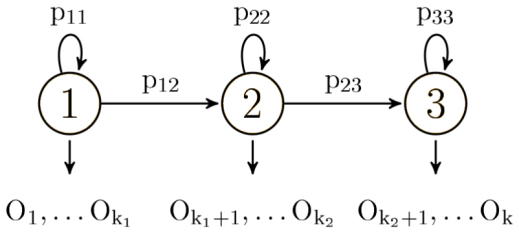

Fig. 1. Left-to-right Hidden Markov Model with three states

First, the HMMs are trained, using a similar approach to that utilized in speech recognition: adjusting the parameters of the HMM model (Figure 1) using information extracted from a speech database. Each HMM can be expressed as:

$$\lambda = (\pi, \mathbf{a}, \mathbf{b}), \tag{1}$$

where π is the probability of initial-state, $\mathbf{a}$ and $\mathbf{b}$ the state-transition and output probability distributions, assumed as multivariate Gaussian distributions (with a mixture of continuous and 0-dimensional distributions).

In statistical parametric speech synthesis based on HMM, the set of models depends not only on the number of phonemes of the particular language, but on the context-dependency of the phonemes (phonetic and prosody contexts) as well. For this reason, a large number of models are trained to represent the temporal, spectral and pitch characteristics of every sound and its context. For example, a model for the $< a >$ phoneme at the beginning of a phrase, followed by consonant, and a model for the $< a >$ phoneme at the beginning of a phrase, followed by a vowel, etc.

The training of each HMM can be expressed as:

$$\lambda_{max} = \arg\max_{\lambda} p(\mathbf{O}|\lambda, W), \tag{2}$$

where $\mathbf{O}$ is the set of speech parameters and W the phoneme labels. A detailed description of the HMM and the procedures involved in the speech synthesis can be found in [1, 20].

For this work, it is of particular importance to state that the quality of the speech synthesis relies on the proper adjustment of the parameters of λ_{max} in Equation 2. And this adjustment depends on the quality of the features $\mathbf{O}$ extracted from the dataset, and its consistency according to the phoneme labels (linguistic specification) W.

Several factors can affect the outcomes of the process: The amount of information in the database (few information implies less $\mathbf{O}$ to estimate the parameters of the HMMs) and the quality of this information. If the information is corrupted by noise, or the recordings have large variations among phonemes (typically, this can occur in very expressive or emotional speech), the ability of the HMMs to reproduce the parameters of the speech for a natural sounding voice with high intelligibility may be affected. The nature of such noise and its level can also be a relevant factor for the results. In this work, an experimental validation of such assumptions is proposed and measured.

3 Experimental Setup

3.1 Database

For this work, we selected the set of words and sentences of [21], developed at the Center for Language and Speech Technologies and Applications of the Polytechnic University of Catalonia. The 184 utterances were recorded by a professional native Spanish speaker actor in a professional studio, where the recording conditions were controlled completely. The database includes affirmative and interrogative sentences, fifteen paragraphs, digits and isolated words.

3.2 Experiments

To determine how noise affects the building of synthetic voices with such small database, several voices were produced using the HTS system, each one after affecting the speech source with noise. The complete database was degraded with additive noise of three types: two artificial-generated noise (White, Noise) and

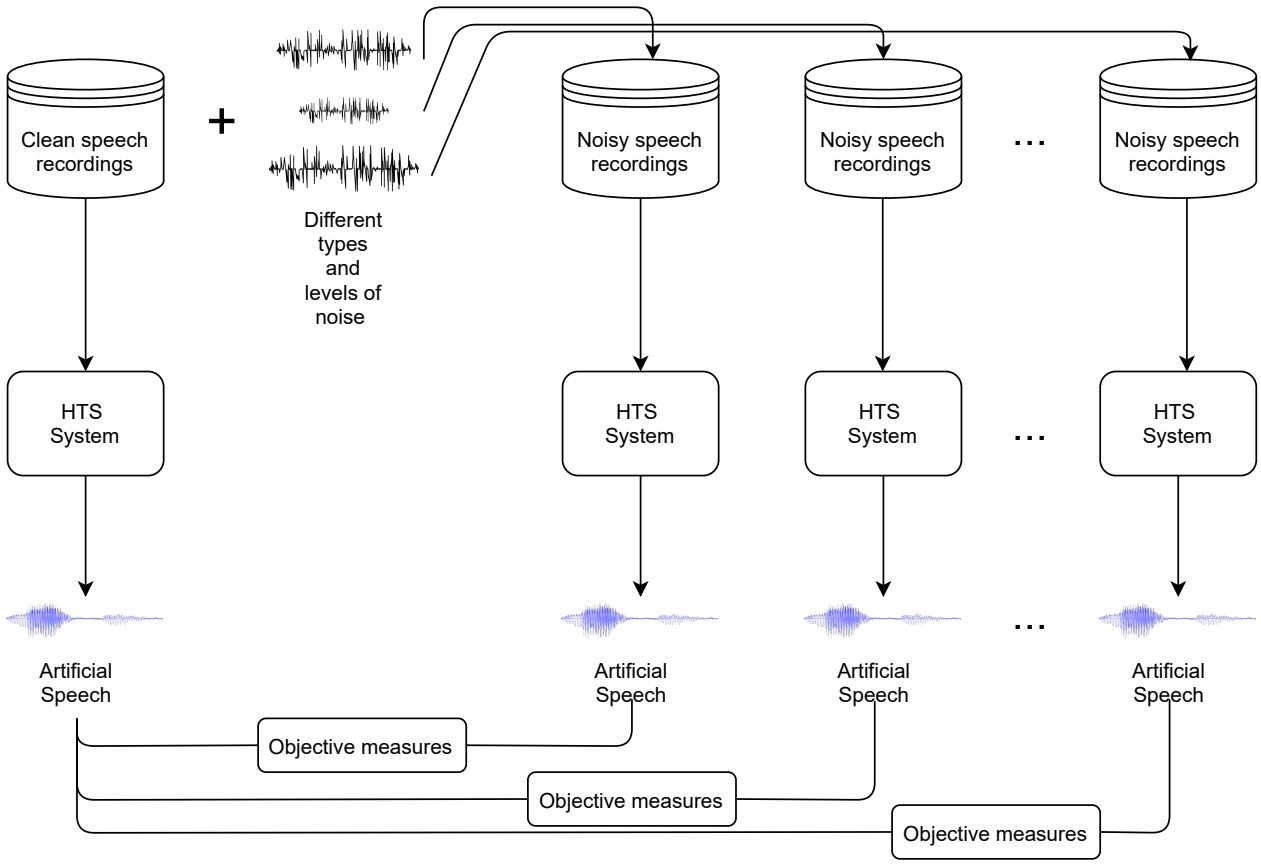

Fig. 2. Diagram of the experimental procedure

one natural noise (Babble). Five levels of Signal-to-noise (SNR) ratio were considered, to cover a range of conditions and comparatively assess the effect on the results.

The whole set of voices to compare can be listed as:

— HTS Clean: The produced with the clean database, without any noise added.

— White Noise added at five SNR levels: SNR 5, SNR 7.5, SNR 10, SNR 12.5, SNR 15.

— Pink Noise added at five SNR levels: SNR 5, SNR 7.5, SNR 10, SNR 12.5, SNR 15.

— Babble Noise added at five SNR levels: SNR 5, SNR 7.5, SNR 10, SNR 12.5, SNR 15.

The evaluation metrics proposed in the following section were used to compare the level of degradation on the artificial voice in comparison with the base system (HTS clean). A diagram of the complete process is presented in Figure 2.

3.3 Evaluation

To determine the quality of each case of synthetic voice, two objective measures were applyed. These measures have been reported in speech synthesis reports as reliable in measuring the quality of synthesized voices:

— Segmental SNR (SegSNR): This measure calculate the average of SNR at frame level,

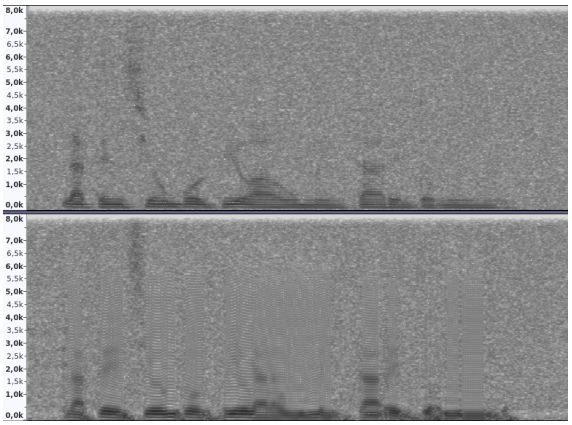

Fig. 3. Spectrograms of an utterance with White noise at SNR5 (above) and the same utterance synthesized from a database degraded with the same type and level of noise (below)

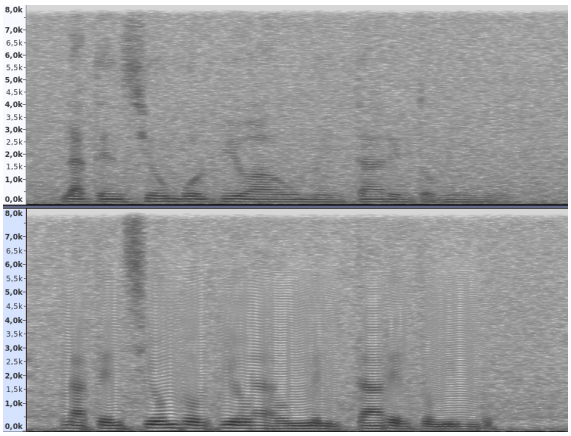

Fig. 4. Spectrograms of an utterance with Pink noise at SNR10 (above) and the same utterance synthesized from a database degraded with the same type and level of noise (below)

according to the equation:

$$\text{SegSNR} = \frac{10}{N}\sum_{i=1}^{N}\log\left[\frac{\sum_{j=0}^{L-1} s^2(i,j)}{\sum_{j=0}^{L-1}\left(s(i,j)-x(i,j)\right)^2}\right], \tag{3}$$

where $x(i)$ is the original sample and s_i the i^{th} synthetic speech sample. N is the total number of samples of the utterance and L is the frame length.

— PESQ: This is a measure intended to predict the subjective perception of speech, in ITU-T recommendation P.862.ITU. The results are reported in the interval $[0.5, 4.5]$. A PESQ value of 4.5 means an exact reconstruction of the speech. PESQ is computed following the equation:

$$\text{PESQ} = a_0 + a_1 D_{ind} + a_2 A_{ind}. \tag{4}$$

The coefficients a_k are chosen to optimize PESQ measure in signal distortions and overall quality.

Additionally, we propose the visualization of spectrograms as a mean to represent the noise and its effect on the spectrum of the speech signals.

4 Results

This section presents the evaluation metrics on the different experiments and its analysis in terms of how the presence of noise affects the building of synthetic voices. For example, in the spectrograms of Figure 3, the silence segments at the beginning and the end of the noisy speech (with SNR 5), and the synthesized version of the same utterance preserves similar patterns of the noise. On the other hand, in the speech segments, the spectrogram presents noticeably blurred bands of frequencies.

A similar observation can be made for the case of Pink noise at SNR 10, as presented in Figure 4. The particular pattern in the form of bands of frequencies can be explained for the process of adjusting the trajectories of parameters in the HMMs. The noisy information became part of the information adjusted in the models, and in the process of generating parameters, the characteristics of flat trajectories also affected the noise.

Unfortunately, such characteristic during the speech segments in the spectrograms represent considerable decrease in the objective measures of the synthesized voice. For example, Figure 5 shows how the noisy condition of the data severely affects the perceptual quality of the

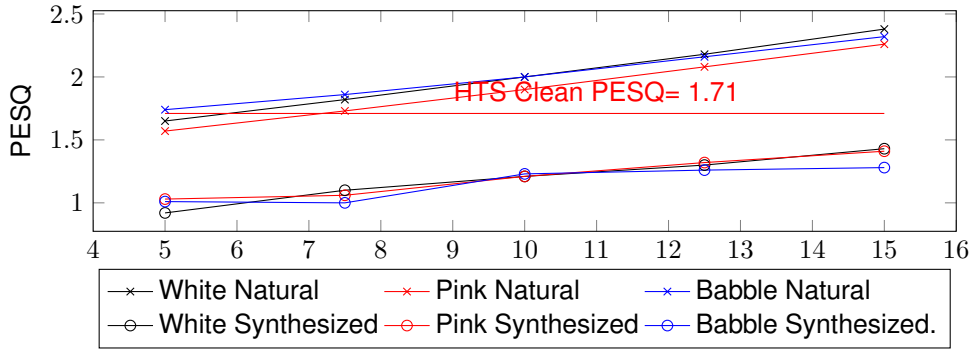

Fig. 5. PESQ results for the noise-degraded speech and the artificial version produced from the same speech

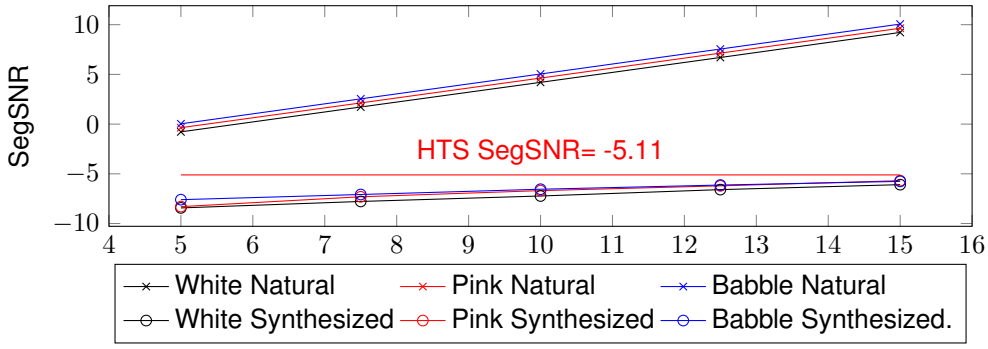

Fig. 6. SegSNR results for the noise-degraded speech and the artificial version produced from the same speech

synthesized speech at all SNR levels. At SNR5 of White Noise, the resulting synthesized speech is closed to the lowest value of PESQ. All artificial voices produced under noisy conditions have considerably lower PESQ values than the base system: the HTS voice.

There are no significant differences between the three types of noise analyzed in this work. The Babble noise seems to affect the results more than the artificial voices, which is expected due to the speech nature of such noise (consisting of a crowd talking in the background).

Considering SNR levels below SNR 5 is a common practice in the study of robust speech recognition. But with these results, it seems that below this level, the synthesized speech for a low resource database cannot be considered for any practical application.

The results of the measure SegSNR are presented in Figure 6. Like the previous measure, there is a significant drop in the quality of synthetic voices at all SNR levels, and very similar among the noise types. All the cases present values below the base system (HTS Clean voice, with SegSNR=-5.11) as expected, but there is a decrease in the slope of the lines in the synthesized speech that can be considered an unexpected result of this study. Such behavior in the SegSNR trends at all SNR levels can be explained by the averaging process performed during the training of the HMMs.

All the results presented have similar trends in the dropping of the quality of synthetic voices in the presence of noise; thereby, preserving the slope of the degraded speech for the case of PESQ. It is important to remark that the results were obtained from a Spanish speech database that can be considered low-resourced. The robustness of the HTS system under such conditions can be considered very low in contrast to the experiences

reported in the references that took advantage of adaption systems or the complement of clean speech from other speakers during the process of generating the artificial speech.

5 Conclusions

In this work, an experimental study on the quality of synthetic speech built from a Spanish noisy database was performed. The amount of data available for the experiments can be considered low-resourced in contrast to larger speech databases available in other languages.

The obtained results show how the presence of noise in the recordings severely affects the synthetic voices produced, regardless of the type of noise and the SNR. In particular, the perceptual quality measured using PESQ shows how the resulting voices have lower quality than the voices produced from clean speech. The type of noise seems to make no difference in the quality of the synthetic speech.

The results are relevant to the building of synthetic voices where data cannot be collected in controlled environments, from historic recordings, data freely available on the Internet, or recordings performed during videoconferencing.

In addition, the results help to establish the importance of building a clean larger speech corpus for endangered languages, children's speech, and many other potential applications of speech synthesis in new languages or languages where such resources have not been produced.

For future work, several relevant questions can be addressed for experimental validation, in terms of the robustness of speech synthesis systems under partially noise-corrupted data, and a broader range of noise types and levels. Applying denoising algorithms before the building of the voices is an important opportunity to preserve the possibility of generating synthetic voices from noise-degraded data.

References

1. **Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K. (2013).** Speech synthesis based on hidden Markov models. Proceedings of the IEEE, pp. 1234–1252.
2. **Masuko, T., Tokuda, K., Kobayashi, T., Imai, S. (1996).** Speech synthesis using HMMs with dynamic features. IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Vol. 1.
3. **Tokuda, K., Kobayashi, T., Imai, S. (1995).** Speech parameter generation from HMM using dynamic features.International Conference on Acoustics, Speech, and Signal Processing. Vol. 1.
4. **Zen, H., Nose, T., Yamagishi, J. (2007).** The HMM-based speech synthesis system (HTS) version 2.0. SSW.
5. **Gonzalvo, X., Sanz, I.I., Socoró-Carrié, J.C., Alías F. (2007).** HMM-based Spanish speech synthesis using CBR as F0 estimator. ITRW on NOLISP, pp. 788–793.
6. **Gonzalvo, X., Taylor, P., Monzo, C., Sanz, I.I. (2009).** High quality emotional HMM-based synthesis in Spanish. International Conference on Nonlinear Speech Processing, Springer. DOI:10.1007/978-3-642-11509-7_4.
7. **Franco, C.A., Herrera, A., Escalante B. (2017).** Speech synthesis in Mexican Spanish using voice parameterization. IIISCI, 15(4), pp. 72–75.
8. **Ekpenyong, M., Urua, E.A., Watts, O., King, S., Yamagishi, J. (2014).** Statistical parametric speech synthesis for Ibibio. Speech Communication, Vol. 56, pp. 243–251. DOI: 10.1016/j.specom.2013.02.003.
9. **Ze, H., Senior, A., Schuster, M. (2013).** Statistical parametric speech synthesis using deep neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing. DOI: 10.1109/ICASSP.2013.6639215.

10. **Ning, Y., He, S., Wu, Z., Xing, Ch. (2019).** A review of deep learning based speech synthesis. Applied Sciences, Vol. 9, No. 19, pp. 4050. DOI: 10.3390/app9194050.

11. **Hu, Y.J., Ling, Z.H. (2016).** DBN-based spectral feature representation for statistical parametric speech synthesis. IEEE Signal Processing Letters, Vol. 23, No. 3, pp. 321–325. DOI: 10.1109/LSP.2016.2516032.

12. **Hu, Y.J., Ling, Z.H. (2018).** Extracting spectral features using deep autoencoders with binary distributed hidden units for statistical parametric speech synthesis. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 26, No. 4, pp. 713–724. DOI: 10.1109/TASLP.2018.2791804.

13. **Suraj-Pandurang, P., Laxman-Lahudkar, S. (2019).** Hidden-Markov-model based statistical parametric speech synthesis for Marathi with optimal number of hidden states. International Journal of Speech Technology, Vol. 22, No. 1, pp. 93–98.

14. **Sefara, T.J., Mokgonyane, T.B., Manamela, M.J., Modipa, T.I. (2019).** HMM-based speech synthesis system incorporated with language identification for low-resourced languages. International Conference on Advances in Big Data, Computing and Data Communication Systems (ICABCD). DOI: 10.1109/ICABCD.2019.8851055.

15. **Junichi, Y., Ling, Z., King, S. (2008).** Robustness of HMM-based speech synthesis.

16. **Valentini-Botinhao, C., Wang, X., Takaki, S., Yamagishi, J. (2016).** Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks. Interspeech.

17. **Karhila, R., Remes, U., Kurimo, M. (2013).** Noise in HMM-based speech synthesis adaptation: Analysis, evaluation methods and experiments. IEEE Journal of Selected Topics in Signal Processing, Vol. 8, No. 2, pp. 285–295.

18. **Baljekar, P. (2018).** Speech synthesis from found data. PhD thesis, Carnegie Mellon University.

19. **Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T. (2000).** Speech parameter generation algorithms for HMM-based speech synthesis. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3. DOI: 10.1109/ICASSP.2000.861820.

20. **Toda, T., Tokuda, K. (2007).** A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Transaction on Information and Systems, Vol. 90, No. 5, pp. 816–824. DOI: 10.1093/ietisy/e90-d.5.816.

21. **Maegaard, B., Choukri, K., Calzolari, N., Odijk, J. (2005).** Elra – European language resources association - background, recent developments and future perspectives. Language Resources and Evaluation, Vol. 39, No. 1, pp. 9–23. DOI: 10.1007/s10579-005-2692-5.

Article received on 09/10/2020; accepted on 16/02/2021. Corresponding author is Marvin Coto-Jiménez.