IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 6, Ser. I (Nov. – Dec. 2020), PP 14-18
www.iosrjournals.org

MOS validation on Synthesized Speech parameterized by Cepstral Coefficients and LSP

Carlos Franco-Galván[1], Abel Herrera-Camacho[2]
[1] *Laboratorio de Tecnologías del Lenguaje, UNAM, Mexico City, Facultad de Artes BUAP,*
[2] *Laboratorio de Tecnologías del Lenguaje, UNAM, Mexico City, Mexico,*

Abstract: *After the appliance of different statistical norms to validate the quality of synthesized voices applied to an HTS-based spanish synthesizer, which uses LSP and Cepstral Coefficients parameterizations. The authors firmly concluded both things: LSP parameterization can be as good as the standard Mel-Cepstral parameterization. Nevertheless, both parameterizations are still insufficient to qualify as natural sounding speech synthesis.*

Keywords: *Speech Synthesis, Voice Parameterization, Line Spectral Pair, Mel-Cepstral Parameterization.*

Date of Submission: 10-11-2020 Date of Acceptance: 25-11-2020

I. Introduction

Mel Frequency Cepstral Coefficients were standard parameterization in a HMM-based Text to Speech synthesizer HTS [1] adapted to spanish has been used for over four years [2] in *Laboratorio de Tecnologías del Lenguaje UNAM*. After carrying out a series of tests with different users [3], it was considered insufficient to qualify as high quality synthesized voice. Therefore, it was decided by the authors to employ an alternative voiced parameterization based on Line Spectral Pair LSP [4]. Such parameterization was also implemented in the Spanish HTS synthesizer and statistically validated [5].

The first validation was carried out with MOS tests only [6]. Besides knowing the user's opinion in terms of naturalness and intelligibility, it was necessary to learn in which position LSP parameterization was in relation to Cepstral parameterization. Since both types were programmed in HTS, they were named HTS-LSP and HTS-MFCC respectively. The subjects who validated them qualified HTS-LSP slightly above HTS-MFCC [5].

Given that Mel-Cepstral parameterization is the standard in synthesis and recognition, the authors judged necessary to apply further tests to sustain or even reject the MOS results. This document aims to summarize each test and its results. It is divided as follows: section 2 overviews the tests, section 3 concerns intelligibility related tests and section 4 discusses the results of all of them.

II. MOS Test Overview

Whenever artificial speech is tested, two aspects are considered: naturalness and Intelligibility. Resemblance to a person´s voice is sought for in the first aspect. The second aspect explores how clear the words are articulated.

MOS historically originates from subjective measurements where listeners would sit in a "quiet room" and score a telephone call quality as they perceived it. This kind of test methodology had been in use in the telephony industry for decades and was standardized in ITU-T recommendation P.800. It specifies that "the talker should be seated in a quiet room with volume between 30 and 120 dB and a reverberation time less than 500 ms (preferably in the range 200–300 ms). The room noise level must be below 30 dBA with no dominant peaks in the spectrum."

The MOS is expressed as a single rational number, typically in the range 1–5, where 1 is lowest perceived quality, and 5 is the highest perceived quality. Other MOS ranges are also possible, depending on the rating scale that has been used in the underlying test. The Absolute Category Rating scale is very commonly used, which maps ratings between Bad and Excellent to numbers between 1 and 5, the qualitative measures are: 5 Excelent, 4 Good, 3 Fair, 2 Poor and 1Bad.

Other standardized quality rating scales exist in ITU-T [16] recommendations (such as P.800 or P.910). For example, one could use a continuous scale ranging between 1–100. Which scale is used depends on the purpose of the test. In certain contexts there are no statistically significant differences between ratings for the same stimuli when they are obtained using different scales.[2]

The MOS is calculated as the arithmetic mean over single ratings performed by human subjects for a given stimulus in a subjective quality evaluation test. Thus:

$$MOS = \frac{\sum_1^N R}{N} \quad (1)$$

Where R are the individual ratings for a given stimulus by N subjects.

ITU-T Recommendation P.800.2 prescribes how MOS values should be reported. Specifically, P.800.2 says:

it is not meaningful to directly compare MOS values produced from separate experiments, unless those experiments were explicitly designed to be compared, and even then the data should be statistically analysed to ensure that such a comparison is valid. That is why for this experiment, a MUSHRA test was conducted to prove consistency within the obtained results.

2.1 MOS Test

MOS Test is by far the most widely applied test to measure audio quality in Telecommunications [6]. It is the standard used in the academical workshop known as the Blizzard Challenge [7] whose aim is to statistically validate artificial voices, therefore it was the obvious choice to validate the HTS-LSP parameterization. A population of 31 listeners was selected. Five phrases were played to each listener in three different versions: The voice of the speaker used to create the synthesizer, the synthesized voice HTS-MFCC and the synthesized voice HTS-LSP. Naturalness and Intelligibility were validated using a scale from 0 to 5. The average results are shown below:

Table 1. MOS Results taken form [5]

Variable	Naturalness HTS-LSP	Intelligibility HTS-LSP	Naturalness HTS-MFCC	Intelligibility HTS-MFCC
Mean Score (CI 95%)	3.47	3.6	3.07	3.44
St. Dev.	0.56	0.57	0.65	0.76
Max.	4.8	5	4	5
Min.	2.4	2.8	1.8	1.4

We can at first glance learn from the results that HTS-LSP gained better acceptance from the listeners. The mean scores have a confidence Interval CI o 95%. Both parameterizations are above the medium of 2.5 which means the parameterizations are around 60% of the highest score. The HTS-LSP parametrization, being the most recent modification to the Synthesizer was favored by the author [5]. Figure 1 presents the results in a chart. After inspecting the chart, we can see that statistically speaking the standard deviation of HTS-MFCC and HTS-LSP are 0.65 and 0.57 respectively they are fairly close. Therefore, another set of tests were carried out to look for significative changes in the average score.

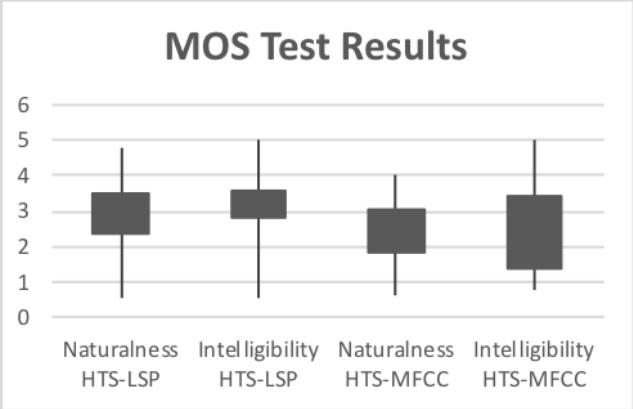

Fig. 1. MOS Results taken from [15]

2.2 Second Battery of MOS Tests

After increasing our population of listeners from 31 to 100 the results shown in Table 2 were obtained.

Table 2. Second MOS Results

Variable	Naturalness HTS-LSP	Intelligibility HTS-LSP	Naturalness HTS-MFCC	Intelligibility HTS-MFCC
Mean Score (CI 95%)	3.5	3.6	3.0	3.4
St. Dev.	0.68	0.6	0.69	0.74
Max.	4.8	5	4	5
Min.	2.4	2.8	1.8	1.4

The mean scores were basically the same. The standard deviation averages 0.7 in all cases for what we can infer that no relevant changes occurred in listeners opinions.

In recent tests [15] Multiple Stimuli with Hidden Reference and Anchor **MUSHRA** [8] together with ABX and three other tests were carried out. The results from such tests are considered at this point for the purpose of distinguishing MOS only. The above-mentioned results can be consulted in the cited paper by the same authors.

2.3 MUSHRA Test

Multiple Stimuli with Hidden Reference and Anchor **MUSHRA** [1] is a norm recommended by the *International Telecommunications Union* ITU. Specially designed to validate the quality of audio codecs. It is organized as follows: A subject listens to the same audio content codified in different ways. The reference is the original audio included in a lossless file and that same audio is also shown low pass filtered with a frequency cut of 3500 Hz as an anchor. This anchor prevents the listener to unconsciously correct his o herself with the reference. The rest of the audio are codifications of the original (e.g. mp3 or wma).

A population of 11 listeners took the test. The norm requests that the subjects need some expertise in audio engineering, 5 of the listeners were professionals in audio and the rest were music technology students. The files each subject listened to were four: The original recording, the anchor, synthesized voice HTS-LSP and HTS-MFCC. The subject sat in front of the computer and the files were randomly played through headphones with SNR of 93 dB.

According to the norm, each file must be qualified from 0 to 100 and at least one of them must be graded 100. Table 2 shows the results and their graph is in figure 2.

Table 2. MUSHRA Results

Statistical Variable	Reference	Anchor	HTS-LSP	HTS-MFCC
Mean Score (CI 95%)	100	62.63	69.54	61.45
St. Dev.	0	15.85	19.77	17.17
Max.	100	86	90	83
Min.	100	30	30	30

The reference was always recognized and given maximum score by the listeners. The anchor was surprisingly poorly valued compared with HTS-LSP and 1.5 above HTS-MFCC. Between these two there is a 7-point difference, with HTS-LSP scoring higher. The mean values have a confidence interval of ±95%.

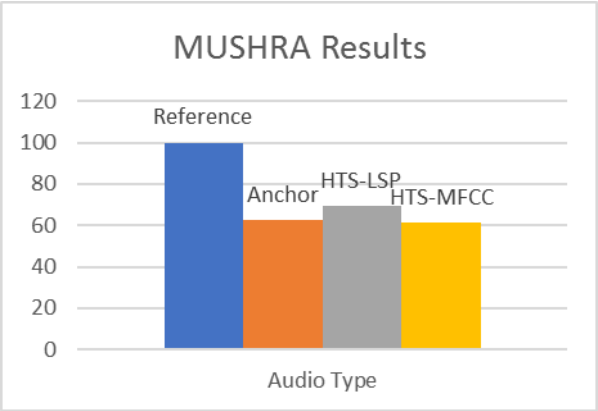

Fig. 2. MUSHRA Results taken from [5]

Compared with the MOS results, HTS-LSP had a mean score of 3.47 which is 69.4% of the maximum score. This result is consistent with MUSHRA where HTS-LSP had a mean score of 69.54.
The population in both tests was entirely different which reinforces consistency in the subjects' opinions.

2.4 ABX Test

ABX validation [9] consists of presenting the listener two sound examples A and B to point out which of those two resembles the reference X, which is a third sound sample. The authors considered the test relevant, because it makes a direct comparison of both parameterizations.

For our study, A was a synthesized sentence using HTS-LSP and B contained the same sentence created form HTS-MFCC. The reference X was the sentence recorded by the speaker whose voice was taken to produce the synthesis.

The test is simple, the listener can play the three audio samples and then answers with "much" or "little" to the following questions: "How close is A to X?" and "How close is B to X?".

30 people participated on the survey, most of them were 23 years old college students. 17 of them thought that HTS-LSP was closer to the reference and 13 said it had little resemblance to the reference. Concerning HTS-MFCC, 10 people judged it closer to the reference whereas 20 said it had little resemblance.

Fig. 3. ABX Results

Once again, as figure 3 shows, the results confirmed HTS-LSP sows better similitude to the original recording than HTS-MFCC. Although ABX is a qualitative test, if the answer "much" was 1 and "little" was 0, given our population of 30, 56.6% of the population (17 people) said HTS-LSP was better which is not far from the 69% obtained in the MOS and MUSHRA tests.

III. Evaluation of Intelligibility

Dictation of single synthesized words was carried out to validate intelligibility. 30 people took dictation of five synthesized words in spanish using HTS-LSP. The subjects were college students of an average age of 23. All listeners had healthy audition, an informal audition test was individually performed to each listener an hour prior to the dictation. The chosen words were phonetically varied.

The dictation took place on a classroom of 10x10 square meters. The sentences were played through a Bose *Soundlink* Speaker connected via Bluetooth to a laptop computer. The audio could be clearly heard on the back of the room 10 meters from the loudspeaker.

The dictations were reviewed and graded two points to each sentence written correctly. The mean group score was 8 points. In average, Two of ten words were not clear for the subjects.

IV. Conclusions

Given these results from the MOS tests and according to those previously obtained in [15] the authors can conclusively affirm that HTS-LSP parameterization can be interchangeable with HTS-MFCC parameterization.

HTS-LSP on the other hand, sounds brighter than HTS-MFCC. Usually brighter sounds are clearer than darker ones, since the human ear is more efficient on high frequencies. The ultimate choice of parameterization depends solely on the system it used on. Compared to human voice, the MOS show that naturalness is 30% far from the ideal. It is fair to say that none of both parameterizations excel at naturalness.

It is important to notice that naturalness is a complex concept and its acceptation depends on multiple factors such as the listener expectations and personal experience. A final validation stems from the situation where the synthesized voice is applied, the results can notably vary when it is used to receive instructions from a GPS map than when an animated character is brought to life.

Other authors claim that naturalness cannot be achieved relying exclusively on parameterization but on the selection system. State of the art synthesizers are based on selection methods better than HTS, the tendency is to work on Deep Neural Networks as the preferred method.

The authors would like to add that even when HTS-LSP is the latest parameterization tested, it is not even close to be the final. There are some ideas to be tried out on speech and audio parameterization. One of them is based in granular synthesis (usually used in music) where the signal is chopped into small time intervals, known as grains. Those grains can be re-arranged linearly or not linearly.

In the case of speech, the linear method would be preferred maintaining the grains corresponding to the minimum required formant frequencies to keep the utterance meaning intact. [16] Reports inspiring results, a possible line of research might start from that point.

STRAIGHT is still today one of the most popular forms of voice parameterization, usually combined with DNN as phoneme selection method.[18] The authors think that this other line of research is a feasible way to go in spanish speech synthesis. For example in Laboratorio de Tecnologías del Lenguaje UNAM a mexican spanish synthesizer has been designed based on RNN and BLSTM [19].

References

[1]. K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech Synthesis Based on Hidden Markov Models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[2]. A. Herrera-Camacho and F. D. R. Ávila, "Development of a Mexican Spanish Synthetic Voice Using Synthesizer Modules of Festival Speech and HTSStraight," *Int. J. Comput. Electr. Eng.*, pp. 36–39, 2013.

[3]. C. Franco, F. Del Rio, and A. Herrera, "ATINER Conference Paper Series Speech Synthesis of Central Mexico Spanish using Hidden Markov Models," pp. 1–12, 2016.

[4]. N. Nakatani, K. Yamamoto, and H. Matsumoto, "Mel-LSP Parameterization for HMM-based Speech Synthesis," *Eurasip Proc. SPECOM 2006*, 2006.

[5]. C. Franco, A. Herrera, and B. Escalante, "Speech Synthesis in Mexican Spanish using LSP as voice parameterization," *iiisci.org*, 2017.

[6]. ITU-T, "Recommendation ITU-T P.800.1 : Mean opinion score (MOS) terminology," 2016.

[7]. S. King and V. Karaiskos, "The Blizzard Challenge 2016," in *Blizzard Challenge workshop*, 2016.

[8]. Itu-BS.1534, "Method for the subjective assessment of intermediate quality level of audio systems Policy on Intellectual Property Right (IPR) Series of ITU-R Recommendations," pp. 1534–3, 2015.

[9]. W. A. Munson and M. B. Gardner, "Standardizing Auditory Tests," *J. Acoust. Soc. Am.*, vol. 22, no. 5, pp. 675–675, Sep. 1950.

[10]. ITU-T, "T-REC-P.800-1996," vol. 800, 1996.

[11]. ITU-T, "ITU-T Recommendation P.862 - PESQ measure," 2001.

[12]. J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment Part II: Psychoacoustic Model," *J. Audio Eng. Soc*, vol. 50, no. 10, pp. 765–778, 2002.

[13]. M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," in *Proc. of European Congress on Acoustics*, 2005, pp. 1–4.

[14]. C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Commun.*, vol. 18, no. 4, pp. 381–392, Jun. 1996.

[15]. Franco-Galván, C., Herrera-Camacho, A., & Escalante-Ramírez, B. (2019). Application of Different Statistical Tests for Validation of Synthesized Speech Parameterized by Cepstral Coefficients and LSP. Computación y Sistemas, 23(2), 461-467.

[16]. Itu-BS.1534, "Method for the subjective assessment of intermediate quality level of audio systems Policy on Intellectual Property Right (IPR) Series of ITU-R Recommendations," pp. 1534–3, 2015.

[17]. S. Fasciani, "Spectral granular synthesis," in *ICMC 2018 - Proceedings of the 2018 International Computer Music Conference*, 2018, pp. 99–103.

[18]. A. Tjandra, S. Sakti, and S. Nakamura, "End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator," in *ICASSP, IEEE International Conference on*

[19]. E. Morales, and A. Herrera. D"NN Synthetizer for Mexican Spanish Language". Proceedings of 26ª Reunión de Otoño de Comunicaciones, Computación, Electrónica y Exposición Industrial, ROC&C'2016. IEEE Mexico Section, to be published.