ISSN 2007-9737

Application of Different Statistical Tests for Validation of Synthesized Speech Parameterized by Cepstral Coefficients and LSP

Carlos Franco-Galván[1], Abel Herrera-Camacho[2], Boris Escalante-Ramírez[3]

[1] Universidad Nacional Autónoma de Mexicana, Facultad de Artes BUAP,
Laboratorio de Tecnologías del Lenguaje, Puebla,
Mexico

[2] Laboratorio de Tecnologías del Lenguaje,
Mexico

[3] Universidad Nacional Autónoma de México, Facultad de Ingeniería,
Mexico

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

Abstract. The following document tries out different statistical norms to validate the quality of synthesized voices applied to an HTS-based Spanish synthesizer, which uses LSP and Cepstral Coefficients parameterizations. Standard MOS tests were carried out. Nevertheless, other types of quality tests were performed to reinforce the MOS results. Such as: MUSHRA, ABX and CCR. The subjective test PESQ was also applied. To validate intelligibility a SUS test was used.

Keywords. Speech synthesis, voice parameterization, line spectral pair.

1 Introduction

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*. Among other things, the present work used on the first place, a speech parameterization based on Mel Frequency Cepstral Coefficients. After carrying out a series of tests with different users [3], it was considered 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 as well [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 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 describes the tests related to naturalness, section 3 concerns intelligibility related tests and section 4 discusses the results of all of them.

2 Naturalness Tests

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.

doi: 10.13053/CyS-23-2-2977

Table 1. MOS Results

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

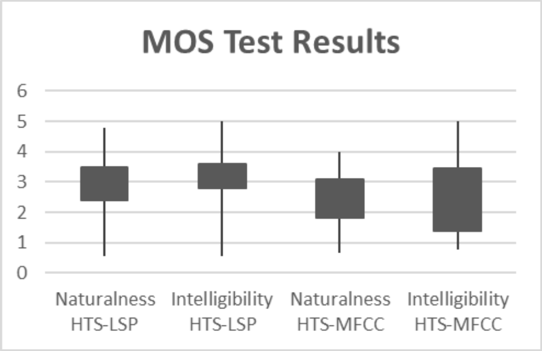

Fig. 1. MOS Results

Four tests were chosen to validate naturalness: MUSHRA, ABX, CCR and PESQ. Intelligibility was validated using SUS. Details and summaries of each naturalness test are shown below.

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:

We can 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. To be certain of the results shown above, another series of tests were carried out. 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.

2.2 MUSHRA Test

Multiple Stimuli with Hidden Reference and Anchor **MUSHRA** [8] 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.

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,

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

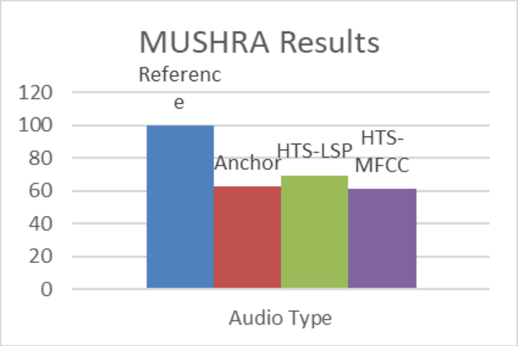

Fig. 2. MUSHRA Results

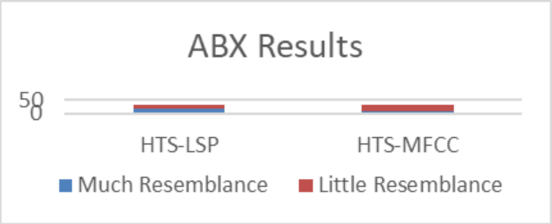

Fig. 3. ABX Results

with HTS-LSP scoring higher. The mean values have a confidence interval of ±95%.

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.

2.3 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.

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.

2.4 CCR Test

In a situation where only quality differences between the two systems are measured, the Comparison Category Rating CCR test [10] can be applied. Only two samples must be listened to and a 7-point scale is used to validate them, where -3 represents "very bad" and 3 represents "very good". The results are then averaged to a comparison mean opinion score CMOS for each sound sample.

A population of 21 answered the CCR test. In this case HTS-LSP overcame HTS-MFCC by over 0.5 points since both scores were 1.04 and 0.47. If -3 to 3 is considered a 7-point scale (0 to 7) HTS-LSP has a 71.42% which again is consistent with the values obtained with MUSHRA and MOS were the scores were 69% close to the maximum. Figure 4 shows these results in a chart.

2.5 PESQ Test

To conclude the statistical validation of naturalness, an objective test was applied which corresponds to the norm ITU-p.862 [11], it is known as Perceptual Evaluation Speech Quality PESQ.

The test was designed to evaluate the quality of a voice signal transmitted through Internet Protocol IP. Its algorithm simulates human sound perception through a comparison between signal delays, details can be found in [12].

X(t) is a voice signal before passing through a communications channel, Y(t) is the voice signal after being transmitted through the channel. Obviously, Y(t) will be degraded, what PESQ aims to is to determinate how bad that degradation is without a human listener's opinion.

The algorithm compares time intervals between X(t) and Y(t) and point out where a significantly different delay is. The time intervals are analyzed based on psychoacoustical simulations of the human ear based on loudness and frequency. The differences found are then computed and a quality factor is obtained. Such factor is scaled from 0 to 5 to correlate with a regular MOS test.

A single voice signal is used in two versions: The voice signal lossless recorded X(t) and the voice signal filtered with the typical frequency cuts of an IP transmission system Y(t). To distinguish results, MOS-LQO (MOS Listening Quality Objective) is the name given to the PESQ results and MOS-LQS (MOS Listening Quality Subjective) to standard MOS results. [12] mention in their documentation that the norm can be used to measure artificial speech quality, but the subject is not mentioned in depth.

Cenark [13] applied the PESQ test to synthesized speech using single words, in that case the original speaker was used as the undegraded voice and the synthesized speech as the voiced transmitted through IP. The authors recreated this approach using entire sentences instead of single words. The sentences chosen were the same as those used on the MOS test.

The ANSI C implementation software provided by the PESQ authors was used to carry out the test. The synthesized voice was HTS-LSP since it is our latest implementation. The values given by the test are shown in table 3.

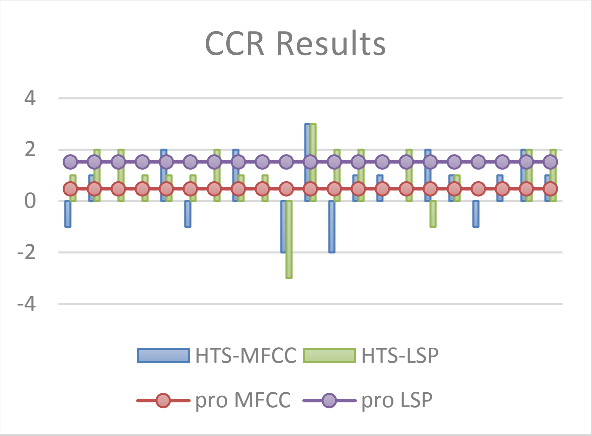

Fig. 4. CCR Results

Table 3. PESQ Results

Value	MOS PESQ
Mean (CI 95%)	1.158
Standard Deviation	0.75
Maximum	1.215
Minimum	1.08
Pearson Coefficient	0.093

The MOS obtained by PESQ is 1.58 which is 2.3 below compared to 3.47 found by the standard MOS. Its root mean square error RMSE is 2.225 and is consistent with the 2.3 natural difference of scores.

Cenark claims that MOS and PESQ results have a linear correlation, he proves this using a Pearson Coefficient Correlation test denoted by:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 \left(\sum x_i\right)^2} - \sqrt{n\sum y_i^2 \left(\sum y_i\right)^2}}.$$

The conditions are the following:

- If r = 1, a perfect correlation exists. Both variables are totally dependent, when one increases, the other one increases proportionally as well.

– If 0 < r < 1, positive correlation exists.
– If r = 0, no relation at all.

Cenark reports values close to 1, in our case Pearson´s coefficient is close to 0 as shown in table 3. Figure 5 compares MOS results with PESQ results, certain linear correspondence can be found in the graphics but is not a perfect linearity. We attribute this to the fact that we used entire phrases instead of single words unlike Cenark.

3 Evaluation of Intelligibility

As it was mentioned earlier, a SUS *Semantically Unpredictable Sentences* [14] test was carried out to validate intelligibility. 30 people took dictation of five synthesized sentences in Spanish using HTS-LSP. The subjects were college students of an average age of 23. The sentences were semantically irregular, without logical meaning. The nonsense contents are on porpose, to avoid the subject unconsciously correct possible mistakes. People usually attribute meaning to words according to the semantical context they are in and not each word individually.

The sentences were:

1. El perro amarillo voló detrás de la almohada. (The yellow dog flew behind the pillow)
2. Me gusta bailar de cabeza sobre el mar. (I like to dance heads down above the sea)
3. Cielos de mermelada sobre lagos de fierro. (Marmalade skies over iron lakes)
4. El club de viento se saturó de pinturas abstractas. (The wind club was filled in of abstract paintings)
5. La hermosa detective se cansó de tanta azúcar. (The beautiful detective got tired of so much sugar)

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 loud speaker.

The dictations were reviewed and graded two points to each sentence written correctly. The mean group score was 6 points. In average, two of five sentences were not clear for the subjects. Table 4 shows the group mistakes in the sentences.

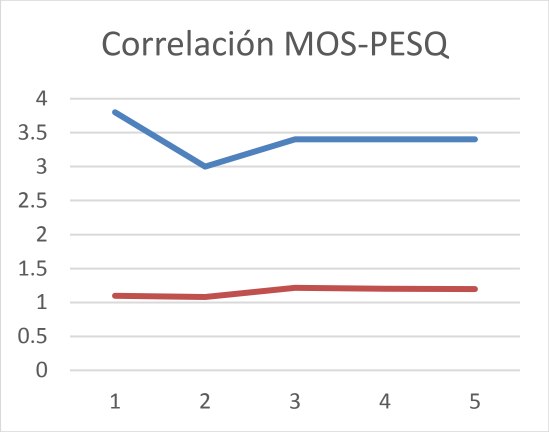

Fig. 5. Correlation between MOS-LQS and MOS-LQO

Table 4. Mistakes in the SUS Test

Phrase	Text	Mistakes
1	El perro amarillo voló detrás de la almohada.	16
2	Me gusta bailar de cabeza sobre el mar	2
3	Cielos de mermelada sobre lagos de fierro.	23
4	El club del viento se llenó de pinturas abstractas.	10
5	La hermosa detective se cansó de tanta azúcar.	3

Sentence number 3 was the hardest to identify, followed by number two. These two phrases are the most irregular semantically. Most of the mistakes in phrase three was a misunderstanding of the word *fierro* which was heard as *hierro*. In phrase one, the most difficult Word to understand was *almohada* were the subjects wrote *alborada* instead. In either of the phrases mentioned the words can be interchanged without losing meaning. As it was said before, the human brain tends to unconsciously give a logical sense where there is not. The other three sentences are perhaps easier to understand because their meaning in not as nonsensical.

4 Conclusions

ABX and CCR tests were used to HTS-LSP and HTS-MFCC to have a relative qualification between both. The authors considered necessary to show which position HTS-LSP was in compared to the MFCC parameterization standard.

The four subjective tests show that HTS-LSP parameterization sounds more natural to the average listener. It is fair mentioning that LSP is not above MFCC by a large score. The ultimate choice of parameterization depends on the system it used on.

The results of all the subjective tests show that naturalness is 30% far from the ideal. Those imperfections reflected on the mean score of 70% deserve a deeper study to check what can be modified to increase their score.

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.

The subjective validation given by PESQ applies only partially since norm p.862 is highly susceptible to time aligning variations between the comparing voice signals. A synthesized voice, given its concatenative phoneme nature, has a considerable number of time irregularities compared to its real voice counterpart. There is a positive aspect of using PESQ given its correlation with the standard MOS. PESQ could be used on a new synthesized voiced preliminary to standard MOS. According to the given results the developer can roughly predict the MOS results with actual listeners.

Acknowledgements

This work has been partially funded by a CONACYT scholarship and PAPIIT IT 102314 of DGAPA-UNAM.

References

1. **Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., & Oura, K. (2013).** Speech Synthesis Based on Hidden Markov Models. *Proc. IEEE*, Vol. 101, No. 5, pp. 1234–1252. DOI: 10.1109/JPROC.2013.2251852.
2. **Herrera-Camacho, A. & Del Rio-Avila, F. (2013).** Development of a Mexican Spanish Synthetic Voice Using Synthesizer Modules of Festival Speech and HTSStraight. *Int. J. Comput. Electr. Eng.*, Vol. 5, No. 1, pp. 36–39. DOI: 10.7763/IJCEE.2013.V5.657.
3. **Franco, C., Del Rio-Avila, F., & Herrera, A. (2016).** *ATINER Conference Paper Series Speech Synthesis of Central Mexico Spanish using Hidden Markov Models*. pp. 1–12.
4. **Nakatani, N., Yamamoto, K., & Matsumoto, H. (2006).** Mel-LSP Parameterization for HMM-based Speech Synthesis. *Eurasip Proc. (SPECOM´06)*, pp. 261–264.
5. **Franco, C., Herrera, A., & Escalante, B. (2017).** Speech Synthesis in Mexican Spanish using LSP as voice parameterization. *IIISCI. ORG*, Vol. 15, No. 4, pp. 72–75.
6. **TU-T (2016).** *Recommendation ITU-T P.800.1 : Mean opinion score (MOS) terminology*.
7. **King, S. & Karaiskos, V. (2016).** The Blizzard Challenge 2016. *Blizzard Challenge workshop*.
8. **Itu-BS.1534 (2015).** *Method for the subjective assessment of intermediate quality level of audio systems Policy on Intellectual Property Right (IPR) Series of ITU-R Recommendations*. pp. 1–34.
9. **Munson, W.A. & Gardner, M.B. (1950).** Standardizing Auditory Tests. *J. Acoust. Soc. Am.*, Vol. 22, No. 5, pp. 675–675.
10. **ITU-T (1996).** *T-REC-P.800-1996*. Vol. 800.
11. **ITU-T (2001).** *ITU-T Recommendation P.862 - PESQ measure*.
12. **Beerends, J.G., Hekstra, A.P., Rix, A.W., & Hollier, M.P. (2002).** 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.

13. **Cernak, M. & Rusko, M. (2005).** An evaluation of synthetic speech using the PESQ measure. *Proc. of European Congress on Acoustics*, pp. 2725– 2728.

14. **Benoit, C., Grice, M., & Hazan, V. (1996).** 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.

Article received on 29/06/2018; accepted on 21/12/2018. Corresponding author is Carlos Franco-Galván.