# Acoustic Vowel Analysis in a Mexican Spanish HMM-based Speech Synthesis

Marvin Coto-Jiménez[1,2], Fabiola M. Martínez-Licona[2], and John Goddard-Close[2]

[1]Universidad de Costa Rica, Electrical Engineering School,
San José, Costa Rica

[2]Universidad Autónoma Metropolitana, Electrical Engineering Department,
Mexico City, México

marvin.coto@ucr.ac.cr{fmml,jgc}@xanum.uam.mx

**Abstract.** The synthetic voice produced from an HMM-based system is often reported as sounding muffled when it is compared to natural speech. There are several reasons for this effect: some precise and fine characteristics of the natural speech are removed, minimized or hidden in the modeling phase of the HMM system; the resulting speech-parameter trajectories become oversmoothed versions of the speech waveforms. In order to obtain more natural synthetic voices, different training conditions must be tried in the construction of the HMMs. One of the most important issues related to the obtained synthetic voice is that of quality assessment. There are several ways to address this, from subjective to objective approaches, applied to different parameters. This paper presents a comparative analysis of certain acoustic features derived from synthesized speech which has been obtained using different training configurations. Pitch, jitter and shimmer were extracted from the synthesized versions of three training sets of vowels of a Mexican Spanish speech database: the normal training set and sets with alterations in the context and fundamental frequency F0. The results show that these objective features can be part of an adequate quality assessment of synthetic speech.

**Keywords:** HMM, speech synthesis, jitter, shimmer, pitch.

## 1 Introduction

The speech production process may be described using the source filter theory of voice production [5], as shown in Fig. 1. This model is called the source-filter model.

Speech synthesis can be realized using this model, e.g. the Klatt synthesizer: voiced and unvoiced speech sounds are produced by applying a source, defined by a pulse train or white noise, to an LTI filter. The LTI filter serves as the vocal tract.

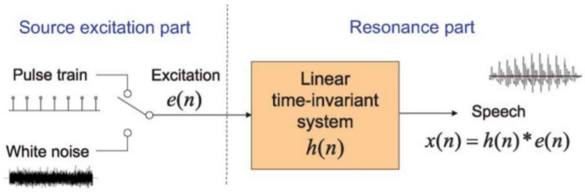


Fig. 1: Source-filter model of speech production [15].

Other methods have been used for speech synthesis, perhaps the most successful to date being that of concatenative synthesis. Recently, HMM-based synthesis, also known as Statistical Parametric Synthesis, has been introduced.

In HMM-based speech synthesis, HMMs are used to generate the speech waveform by modeling pitch, duration, and spectral information, usually in the form of fundamental frequency, f0, parameters, as well as mel frequency cepstal coefficients (MFCC).

HMMs are trained using data obtained from real speakers, and a maximum likelihood criterion is employed to generate these speech parameters [16].

Synthesized speech produced using this technique has been reported as sounding muffled when compared to natural speech, because the generated speech-parameter trajectories are often oversmoothed [15] [1]. This means that detailed characteristics of the speech parameters are removed in the modeling stage, and the model's output is unable to reproduce them.

In order to identify which synthetic voices have better quality, there are several possible approaches; these range from applying either subjective or objective measures. The most popular method for evaluating the quality of synthesized speech is that of a subjective listening test [8].

Several proposals have addressed the issue of relating objective measures to subjective measures such as [3,18,9,14,13,12,17]. In this paper we present a comparative analysis of three acoustic measures: Jitter, Shimmer and Pitch. These acoustic parameters have been used to analyze stress in human speech [11], as well as various pathologies [4] [19], including vowel analysis [2] to mention a few.

We propose extending these studies of pitch, jitter and shimmer, to that of comparing natural and synthesized speech with the objective of using these parameters to aid in assessing the quality of the synthesized voice. More precisely, we use a Spanish speech database consisting of two professional speakers, one male and one female, and under different training conditions employ HTS, an HMM-based speech synthesis system, to build the corresponding synthetic voices. Information on the three parameters is extracted for each voice and statistical tests are conducted and compared to an independent subjective evaluation to assess the possible correlations between them. These results are presented in the paper.

The rest of this paper is organized as follows. Section II describes the Spanish speech database and introduces the methods of analysis of pitch, jitter and shimmer in vowels. Section III describes the results, which are discussed in Section IV. Concluding remarks and future work are presented in Section V.

## 2 Methods

Two Mexican speakers, a professional actress and actor, recorded three sets of 184 Spanish speech utterances each. The 184 utterances included isolated words as well as sentences which could be in affirmative or interrogative forms. The distribution is shown in Table 1.

Table 1: Spanish Corpus Contents.

| Identifier | Corpus contents |
|---|---|
| 1-100 | Affirmative |
| 101-134 | Interrogative |
| 135-150 | Paragraphs |
| 151-160 | Digits |
| 161-184 | Isolated words |

The selection of the words, sentences and paragraphs were the same as that of [10], an emotional speech database originally developed by the Center for Language and Speech Technologies and Applications of the Polytechnic University of Catalonia for the purpose of emotional speech research. The Mexican Spanish recordings were carried out in a professional studio where the recording conditions were completely controlled.

Acoustic features were extracted from the speech signals using Praat [6]; the features selected were pitch, jitter and shimmer, which we shall briefly describe in the following subsections.

### 2.1 Pitch

Each utterance was segmented into the corresponding phonemes, and the maximum pitch of each of the five Spanish vowels found was extracted using the autocorrelation method. The results were separated according to the vowel and the training conditions.

### 2.2 Jitter

Jitter is a measure of period-to-period fluctuations in the fundamental frequency. In general it is calculated between consecutive periods of voiced speech as follows:

$$J_t = \frac{|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{1}$$

where $T_i$, $T_{i+1}$ are the present and posterior periods of speech and $N$ the total number of intervals. The jitter reported is the local jitter, which is used as a voice quality feature, and is defined as the rate between the computed jitter and the mean value of the periods of voiced signal.

### 2.3 Shimmer

Shimmer is a measure of the period-to-period variability of the amplitude value and is defined as follows:

$$Shm = \frac{|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{2}$$

where $A_i$, $A_{i+1}$ are the present and posterior periods' amplitude of speech, and $N$ the total number of voiced periods. The shimmer reported is the local shimmer, which is defined as the average absolute difference between the amplitudes of consecutive periods divided by the average amplitude.

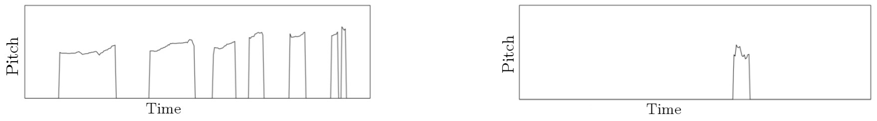
### 2.4 Training conditions

In order to measure the effects of different training conditions on pitch, jitter and shimmer of a synthetic voice, we used the procedures described below, for both speakers, to produce different voices constructed with the HTS system [20].

**Normal training** The male and female voices were trained using the full speech corpus with complete context information (24 features) adapted from HTS-CMU-US-ARCTIC demo in [7]. An analysis of the pitch range was made with Praat, so that a correct $f0$ range could be defined in both cases, and the 24 contextual factors were used, considering prosodic features.

**Context information reduction** In HMM-based speech synthesis, contextual factors are used to capture both segmental and prosodic features [**?**]. In this case, the prosodic contextual factores were removed, and only the phoneme definitions remained. The $f0$ range was the same as in normal training above.

**Distorted $f0$** From our experiences with HMM synthesis, we found that the definition of the $f0$ range in training has a decisive influence on the results. A poorly defined range produces often intelligible but very unnatural voices. The reason for this can be seen in Figure 2, where an $f0$ contour is compared for a phrase pronouncing the hour, for both normal training and also with a poor $f0$ range. The lack of some pitch regions in the latter case has a considerable effect on its naturalness.



(a) Normal training (b) Bad $f0$ range in training

Fig. 2: Pitch contours of the utterance "Son las 8:45" (It's 8:45).

## 3 Results

In order to compare the parameters of pitch, jitter and shimmer of the speakers and the HTS synthesized voices, an analysis of these parameters on all the database utterances was performed.

100 utterances were also produced using each of the synthesized voices, and 5 were randomly chosen for each voice. A mean opionion score (MOS) test was applied to these 5 utterances using 20 volunteers. We use these subjective evaluations as a reference to compare the possibly significant differences between the three acoustic parameters of the original voice with the synthesized ones.

Both the male and female voices obtained using the normal training conditions received the best subjective evaluation, while the voices obtained with a reduced $f0$ range scored the lowest.

Figure 3 shows the pitch value boxplots for each vowel obtained from the voices constructed using the different training conditions and the original natural male voice, while Figure 4 presents those of the female voice.

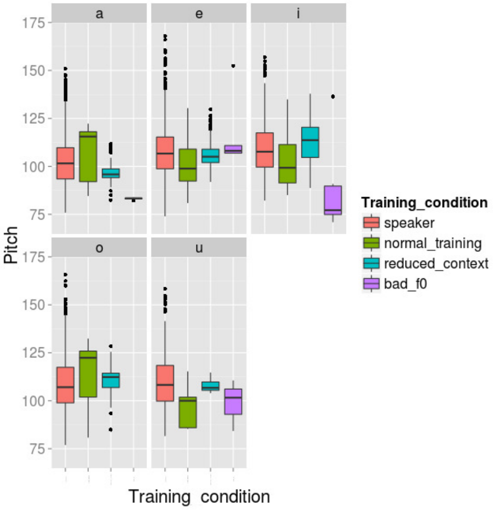


Fig. 3: Boxplots for the pitch values of each vowel using male voice.

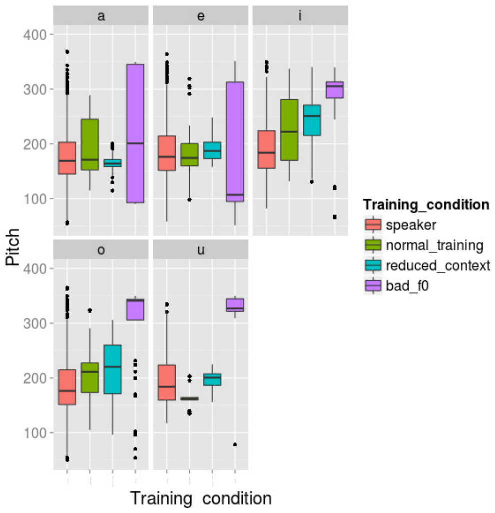


Fig. 4: Boxplots for the pitch values of each vowel using female voice.

Figure 5 shows the jitter value boxplots for the original database and the different training conditions, for the male speaker, and Figure 6 the corresponding shimmer value boxplots for the female speaker.

In order to decide which of the observed differences in the three parameters were statistically significant from the original speaker, we conducted a Friedman test, with a significance level of $\alpha = 0.05$.

Statistically significant differences were detected for several groups of variables, compared to the original speaker, as shown in Table 2. Friedman's test was carried out for all the training conditions, and a Post-hoc test was used to decide which groups are significantly different from each other, with special interest in differences related to the original speaker and with the training conditions.

Table 2: Friedman test Post-Hoc statistically significant differences with original speaker. M: Male voice, F:Female voice

| Training conditions | Statistical significant difference | | |
|---|---|---|---|
| | Pitch | Jitter | Shimmer |
| Normal training (M) | | ✓ | |
| Normal training (F) | ✓ | | ✓ |
| Distorted $f0$ (M) | ✓ | ✓ | ✓ |
| Distorted $f0$ (F) | ✓ | ✓ | ✓ |
| Reduced context (M) | | ✓ | ✓ |
| Reduced context (F) | ✓ | ✓ | ✓ |

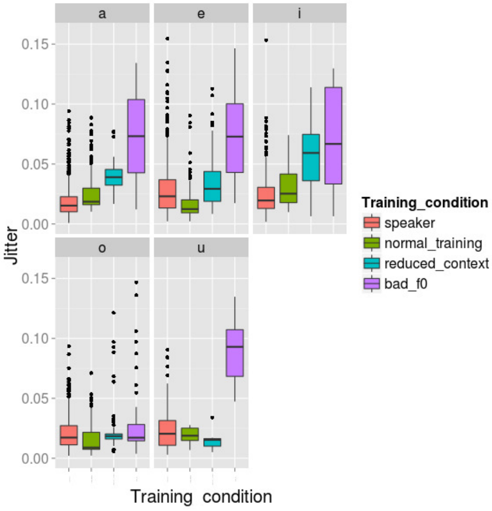


Fig. 5: Boxplots for the jitter values of each vowel using male voice.

Table 3 shows the result of the MOS test, considering the naturalness and intelligibility in pronouncing the time. The scale was from 1 (completely unnatural or completely unintelligible) to 5 (completely natural or completely intelligible).

Table 3: MOS of synthesized voices. M: Male voice, F:Female voice.

| Training conditions | MOS Test | |
|---|---|---|
| | Naturalness | Intelligibility |
| Normal training (M) | 2.96 | 3.61 |
| Normal training (F) | 1.96 | 2.52 |
| Distorted $f0$ (M) | 2.43 | 3.18 |
| Distorted $f0$ (F) | 1.82 | 2.79 |
| Reduced context (M) | 2.74 | 3.32 |
| Reduced context (F) | 2.60 | 3.13 |

## 4 Discussion

We find that the synthesized voices with a lower subjective value of naturalness and intelligibility have significant differences compared to the original speaker in the case of pitch and shimmer.

Table 2 shows that the voices with less natural values (i.e.those obtained with a poor definition of the f0 range in training), have statistically significant

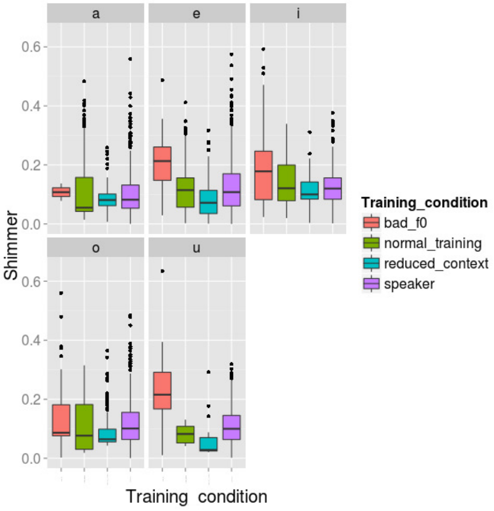


Fig. 6: Boxplots for the shimmer values of each vowel using female voice.

differences with the original speaker in the three acoustic parameters. Similar characteristics are found with the female voice trained with a reduced context, which gets lower subjective values than the voice with normal training. That voice, specially in the case of the male speaker, receives the best subjective evaluation in naturalness and intelligibility.

A combination of the three acoustic parameters can be related to the quality of synthesized voices. For example, Figure 7 shows the scatterplot of the six synthetic voices, where we see that the voices with significant difference compared with the original speaker are differentiated. The voices that have been evaluated with lower scores in the MOS test have these differences with the database.

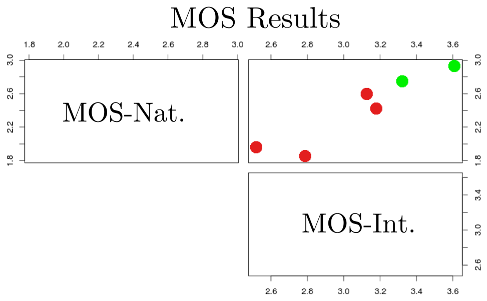


Fig. 7: Scatterplot for MOS test of the six synthesized voices.

## 5 Conclusions

An analysis of the acoustic parameters of pitch, jitter and shimmer for two speakers, and the voices synthesized from them using HMM-based synthesis was presented. Six synthetic voices were produced using different training conditions, and a MOS test applied to obtain different subjective values of naturalness and intelligibility for them.

The aim of the analysis was to establish a relationship between these acoustic parameters and the MOS results.

The results shows a relationship between the detection of statistically significant differences from a Friedman test and with the lowest quality of the synthetic voices.

These results may lead to establishing the statistical analysis of pitch, jitter and shimmer as a useful reference to determining synthetic speech quality, as related to subjective evaluations.

It is necessary to extend the experiments to other voices, ideally with larger speech databases, to allow a more extensive analysis based on training experiments of individual parameters, their pitch, jitter and shimmer and the corresponding subjective evaluations.

**Acknowledgements.** This work was supported by the SEP and CONACyT under the Program SEP-CONACyT, CB-2012-01, No.182432, in Mexico, as well as the University of Costa Rica in Costa Rica. We also want to thank ELRA for supplying the original Emotional speech synthesis database.

## References

1. Black, A.W., Zen, H., Tokuda, K.: Statistical parametric speech synthesis. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. vol. 4, pp. IV–1229. IEEE (2007)
2. Brockmann, M., Drinnan, M.J., Storck, C., Carding, P.N.: Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. Journal of Voice 25(1), 44–53 (2011)
3. Chu, M., Peng, H.: An objective measure for estimating mos of synthesized speech. In: INTERSPEECH. pp. 2087–2090 (2001)
4. Falcone, M., Yadav, N., Poellabauer, C., Flynn, P.: Using isolated vowel sounds for classification of mild traumatic brain injury. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 7577–7581. IEEE (2013)
5. Fant, G.: Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations, vol. 2. Walter de Gruyter (1971)
6. Goldman, J.P.: Easyalign: An automatic phonetic alignment tool under praat. In: INTERSPEECH. pp. 3233–3236 (2011)
7. HTS-Group: Hts voice demos. `http://hts.sp.nitech.ac.jp/?VoiceDemos` (September 2014)

8. Huang, D.Y.: Prediction of perceived sound quality of synthetic speech. Proc. APSIPA (2011)
9. Huang, D.Y.: Prediction of perceived sound quality of synthetic speech. Proc. APSIPA (2011)
10. Maegaard, B., Choukri, K., Calzolari, N., Odijk, J.: Elra–european language resources association-background, recent developments and future perspectives. Language Resources and Evaluation 39(1), 9–23 (2005)
11. Martínez-Licona, F., Goddard, J., Martínez-Licona, A., Coto-Jiménez, M.: Assessing stress in mexican spanish from emotion speech signals. In: Proc. 8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, MAVEBA. pp. 239–242 (2013)
12. Möller, S., Hinterleitner, F., Falk, T.H., Polzehl, T.: Comparison of approaches for instrumentally predicting the quality of text-to-speech systems. In: INTERSPEECH. pp. 1325–1328 (2010)
13. Norrenbrock, C.R., Hinterleitner, F., Heute, U., Moller, S.: Instrumental assessment of prosodic quality for text-to-speech signals. Signal Processing Letters, IEEE 19(5), 255–258 (2012)
14. Remes, U., Karhila, R., Kurimo, M.: Objective evaluation measures for speaker-adaptive hmm-tts systems. Proc. SSW, to appear (2013)
15. Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K.: Speech synthesis based on hidden markov models. Proceedings of the IEEE 101(5), 1234–1252 (2013)
16. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for hmm-based speech synthesis. In: Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on. vol. 3, pp. 1315–1318. IEEE (2000)
17. Valentini-Botinhao, C., Yamagishi, J., King, S.: Can objective measures predict the intelligibility of modified hmm-based synthetic speech in noise? In: INTERSPEECH. pp. 1837–1840 (2011)
18. Valentini-Botinhao, C., Yamagishi, J., King, S.: Evaluation of objective measures for intelligibility prediction of hmm-based synthetic speech in noise. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. pp. 5112–5115. IEEE (2011)
19. Wertzner, H.F., Schreiber, S., Amaro, L.: Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders. Revista Brasileira de Otorrinolaringologia 71(5), 582–588 (2005)
20. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K.: The hmm-based speech synthesis system (hts) version 2.0. In: SSW. pp. 294–299. Citeseer (2007)