Speech Synthesis in Mexican Spanish using LSP as Voice Parameterization

Carlos FRANCO
Facultad de Artes, BUAP
Puebla, Mexico, francocarlos@gmail.com

Abel HERRERA
Laboratorio de Tecnologias del Lenguaje, UNAM
Mexico City, Mexico, abelherrerac@hotmail.com

Boris ESCALANTE
Facultad de Ingeniería, UNAM
Mexico City, Mexico, boris@unam.mx

ABSTRACT

A voice parameterization using Linear Spectral was implemented to a Mexican Spanish HMM-based Speech Synthesis System. Five phrases were synthesized and statistically validated by applying a MOS test to 30 listeners who analyzed the original voices compared to a synthetic voice. Results were compared to previous work where MFCC was used as voice parameterization, the comparison shows that LSP parameterization is above the mean score and pointed better than MFCC.

Keywords: Speech Synthesis, Linear Spectral Pair, HTS

1 INTRODUCTION

The search for a Speech Synthesis system which can be indistinguishable from human speech (a simplification of the Turing test) is one of the ongoing goals at Laboratorio de Tecnologías del Lenguaje of the National Autonomous University of Mexico UNAM. Herrera and Del Río [1] developed a spanish speech synthesis system based on the works of Tokuda and his colleagues: [2] Hidden Markov Models as Text to Speech Synthesis (HTS), a system were Hidden Markov Models HMM are used as an alternative means to phoneme selection.

As it was shown in [3] this type of system can be used with two voice parameterization schemes: Mel Frequency Cepstral Coefficients (MFCC) and Line Spectral Pair (LSP). Both schemes have been proved and validated. Nakatani and colleagues [4] hypothesized that LSP is more efficient than other parameterizations but did not carry out proof tests, the authors decided to continue that line of research with its respective experiments. After adjusting and statistically validating the system, the authors conclude that it efficiently produces speech synthesis in spanish language using LSP. Its naturalness and intelligibility were qualified above the mean and above previously validated MFCC based synthesis.

2 RELATED WORK

LSP parameterization of a speech signal has been in the interest of several lines of research for the last three decades. Nakatani [4] and colleagues evaluated LSP parameterized phrases, but their study was exclusively focused on analyzing isolated phonemes in japanese and not entire words or phrases. Arakawa and colleagues [5] applied LSP to improve certain features of STRAIGHT synthesis system, but the principles of such system differ from those in the system the authors experimented with. Bäckstörm in his doctoral project [6] makes a complete mathematical analysis of LSP but his work is theoretical and did not experiment with speech signals. Tokuda and his team [7] left the door open to experiment with Either LSP or MFCC but they focused on the HTS (Hidden Markov Models as Text to Speech Sythesis) system from a global perspective and do not report results on neither speech parameterization effectiveness.

3 SPEECH SYNTHESIS USING HTS

HTS (Hidden Markov Models as Text to Speech Sythesis) is a proposal from the 2000´s. This type of synthesis decomposes a voice signal in three vectors which include its three main features: Mel General Cepstral coefficients MGC [3], F0 and duration. In practice, these vectors are obtained with a software named Signal Processing Tool Kit SPTK [4].

The vectors are accessed non-linearly to obtain the correct phoneme sequence in a spoken phrase. Therefore, the stochastic selection algorithm of Hidden Markov Models HMM is used in contrast with other synthesis systems, such as Festival [8] were phonemes are selected using a linear method named CART [9].

To compute the probability of the HMMs, the creators of HTS took advantage of a free distributed system developed by the university of Cambridge. The program is known as Hidden Markov Model Toolkit HTK [10].

HTK was originally designed for speech recognition.

Figure 1 shows a general scheme of HTS. More details can be found on the references [11] and the HTS website [7].

Before being able to synthesize a phrase, HTS need to be trained using the desired language specifications. Other characteristics are as well defined in the training stage (e.g. parameterization, number of coefficients, sampling frequency, etc.)

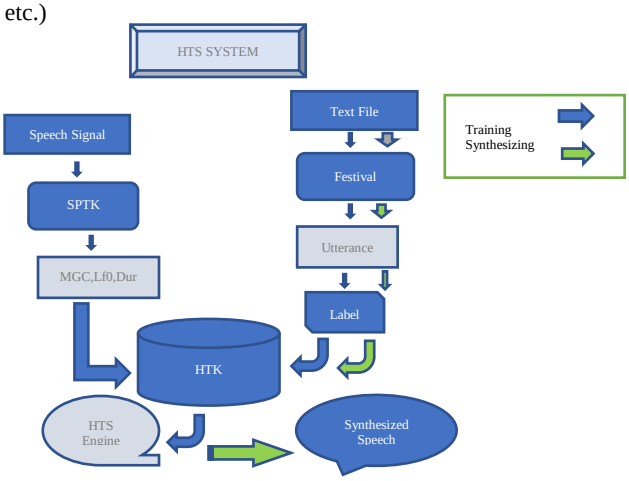

Figure 1. “HTS General Scheme”

The system is trained by inputting 300 audio files containing phonetically balanced phrases and their respective text transcription. The highest probabilities of occurrence of a phoneme sequence will be calculated within the HMMs to obtain the better combination. Text to phoneme conversion is done through Festival [8]. Since Festival was originally designed for english language synthesis, when a different language is used, the system must be adapted to process the grammatical features of such a language. All these grammatical features are coded in a software called lexicon. A lexicon in spanish indicates Festival the use of stressed vocals, letter "ñ", differences between phonemes like /c/ -next to vowels /e/ and /i/- or /z/ among others. The current system uses a lexicon created originally for Andalusian spanish named Junta de Andalucía. It was chosen because iberic spanish is grammatically identical to mexican Spanish, no further modifications were needed. Except for substituting "c" and "z" letters for an "s" when the desired synthesized phrase is being written. Once the phrase is written, text to phoneme conversion occurs in the following order: Sentence to phrase, phrase to word, word to syllable and syllable to phoneme [12].

When the conversion is finished, Festival delivers a utterance (.utt) file. The actual synthesis process takes place in a software named HTS Engine, utterance files must be reorganized to be compatible with it. For that purpose, they are changed into label (.lab) files.

Input data to the system were used in a previous experiment involving MFCC parameterization training [13]. Such data consists of 300 phrases recorded as wave files in an anechoic chamber by a male professional radio speaker. The wave files were coded into RAW files which contain the same information of the wave file except for a header.

The other input data simultaneously processed are the label (.lab) file. These are text files which indicate HTS Engine the desired phoneme sequence (e.g. sentence, phrase, word, syllable) of the phrase to be synthesized.

The RAW files are decomposed in three vectors: One vector contains Mel General Cepstral Coefficients; the second vector contains the phrase LogF0 and the third one the phrase duration. These three elements are stored in three-streamed HMM which is in practice a Gaussian matrix. Their delta and double delta Coefficients are also considered to smooth out the wave transitions within each other, a common practice in speech processing. This model is named hmm0. The calculations are done based on a previously given phoneme probability master label file MLF [10].

The model hmm0 should be divided into smaller models to separate the different phoneme values. For that matter, the mean of hmm0 is calculated generating a new three-streamed model named hmm1. The probabilities stated in the MLF are then condensed in a Master Macro File MMF. Based on this file probabilities, the process is repeated iteratively until several HMM models are formed. The number of HMM models is previously defined by the user. After the HMM models are completed, their probabilities are computed following a Viterbi algorithm and grouped into single phoneme gaussians. Thus, for example, all the /a/ phonemes are together in a same group. Consequently, the selection process will be linear.

The synthesis takes place in a piece of software named HTS Engine [11] which is a vocoder filter driven by two sound sources: Sinusoidal for voiced sounds and white noise for unvoiced sounds. The formers emulate those voice sounds produced by the vocal cord vibrations and the others are phonemes produced by air currents passing from the lungs to the mouth. The filter frequencies correspond to those the phonemes in the phrase to be produced.

4 LSP

Line Spectral Pair (LSP) is a voice parameterization based on a theory proposed by Itakura [14]. It has been used in different voice processing system for synthesis and recognition [15],[16].

This kind of synthesis is a variation of the remarkable Linear Predictive Coefficients LPC, which constituted one of the first efforts to reconstruct a voice signal during the second half of the 20th century. LPC parts from a difference between the original signal and its equivalent "deduced" within past samples. The input signal passes through a filter represented by the following equation:

$$A(z) = 1 + \sum_{p=1}^{N} a_k z^{-p} \qquad (1)$$

Synthesized signal corresponds to the sum of different past samples multiplied by a coefficient a_k. The coefficients can be used to map the original signal spectrum. Details on the mathematical procedure can be found in the referring literature. When LSP are to be obtained, instead of computing auto-correlation, two polynomials are proposed as a solution to equation (1).

$$P(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}) \qquad (2)$$

$$Q(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}) \qquad (3)$$

The polynomial roots (poles and zeroes) must be within the unit circle Z, their conjugated pairs correspond to their frequencies called Linear Spectral Frequencies LSF, the sum of both polynomials in equation (4) represent the filter in (1). At the same time, they represent the formants generated in the vocal tract.

$$A(z) = \frac{1}{2}(P(z) + Q(z)) \quad (4)$$

5 USING LSP TO SYNTHESIZE A SPEECH SIGNAL

The authors decided to test this type of parameterization and adapt it to the current HTS Spanish system. The system by default decomposes the speech signal in Mel General Cepstral Coefficients. It is based on a mathematical concept that unifies MFCC and LSP based on the equation (5) proposed in [3].

$$H(z)\begin{cases}(1 + \gamma \sum_{m=0}^{M} C_m x(n-m))^{\frac{1}{\gamma}}, 0 < |\gamma| \leq 1 \\ \exp \sum_{m=0}^{M} C_m x(n-m), \gamma = 0\end{cases} \quad (5)$$

When γ=0, the system becomes a Mel-Cepstral representation, whereas if γ=1 a LPC representation is obtained from which LSP are extracted.
During the training stage of the software, the value of γ was unity to produce the LPC and therefore obtain its LSP. This process

takes place using SPTK. A technical manual with coding details will be published by the authors.
The authors decided to test it for several reasons: First, LSP is based on Linear Predictive Coding (LPC) which parts form seeing the human vocal tract seen as a filter and the formant frequencies as the filter coefficients. The spectra obtained based on vocal tract models tend to resemble natural speech remarkably. Even more, LSP consider more data than LPC which results in a richer quantization of the original speech signal. An LSP voice filter is more stable in nature, the mathematical demonstration can be found in [6]. The size of the audio files is smaller than that of the files using MFCC. Finally, and most important: There are little documented on Speech synthesis using LSP and particularly applied to spanish, where no documentation was found.

6 EVALUATION AND COMPARISON OF BOTH PARAMETERIZTAIONS

With the purpose of verifying the quality of the synthesized voiced, the authors tested both parameterization techniques: LSP and the before tested MFCC [13].
Two statistical test were performed: Mean Opinion Score [17] tests (Subjective) and Statistical distance measure (objective).

6.1 Mean Opinion Score

A population of 30 listeners was surveyed. Each person listened to 5 phrases in three different versions: original speaker, MFCC synthesis, LSP synthesis. The conditions of the conducted experiment were the following: 5 phonetically balanced phrases were used. The subject sat in front of a computer and listened to the phrases through headphones with a SNR of 93 dB. Two aspects on the audio were validated, naturalness and intelligibility on a scale of 0 to 5. Table 1 shows the obtained mean scores.

Table I MOS Results

Type	Naturalness	Intelligibility	Difference
MFCC	3.07	3.44	0.33
LSP	3.4	3.6	0.2

We can infer from the results that LSP was better accepted. Both parameterization schemes are above the 2.5 average score.

6.2 Distance Measure

Since the MOS tests provide a qualification according to a human listener, a statistical distance comparison between *Gaussian Mixture Models* GMM was performed on the synthesized phrases. Usually, this distance is applied to speaker detection and identification. It is part of another line of research in *Laboratorio de Tecnologías del Lenguaje* [18]. It would aid to measure the phrases objectively.

The idea behind this test is to process the synthesized phrases as a human subject to identification. Our reference was the actual speaker whose voice was used to train the system. Therefore, if the synthesis is good enough, the distance between reference and subject to ID should be close to zero.

The GMMs were composed of the MFCCs extracted form 300 phrases of the original voice and another 300 of synthesized voice using LSPs. The distance between the two models (original and synthesized) means was measured using the Mahalanobis distance [19] and the Euclidean distance.

Both formulas for distance fulfill the three basic properties of arithmetic distance: semi-positivity, symmetry and triangular inequality. The one of interest for the study is semi-positivity which states:

$$d(a,b) \geq 0 \forall\, a,b\ \in X, d(a,b) = 0\ \ if\ a = b \quad (6)$$

This means that when elements a and b are equal, the distance between both will be zero. In our case this would be the definition of an *ideal synthesis*: A voice signal which is indistinguishable between human and machine.
4 GMM models were built using 256, 512, 1024 and 2048 elements for the sake of precision. Table 2 shows the results.
As the last two columns show, a near zero value is obtained when 1024 GMMs are used. Independently of the number of the GMMs all distance values are below unity. We have an average Mahalanobis distance of 0.005 and an ecludiean distance of 0.093.
We can learn from the results that human voice and synthetic voice are not too far from each other. Were we testing wth a human subject, the speaker ID would mark positive. These results are consistent with those obtained from the MOS tests which reveal, in opinion of several listeners that good quality synthesis has been produced in terms of naturalness and intelligibility.
As an additional reference, distance between GMM containing the original voice signal only was computed. -A comparison of the human speaker with himself-. 512 GMMs were used and the Mahalanobis distance was 1.62 e-33 and the Euclidean distance was 5.7 e-17. Both are practically zero. Similar values should be expected when a close to ideal synthesis was achieved.
There is still a long way ahead before reaching those synthesis quality levels.

Table II Distance Results

First Group	N0. of GMM	Dist. Mahal	Dist. Euklid
1-150 (synt vs human)	256	0.0063	0.1118
1-150 (synt vs human)	512	0.0072	0.1196
1-150 (synt vs human)	1024	0.000763	0.0391
1-150 (synt vs human)	2048	0.0052	0.1016

7 CONCLUSIONS

As we could learn from the results in the MOS test and the statistical distances - shown in Table I and Table II respectively- , there is an improvement when LSP is chosen as voice parameterization. In terms of size, LSP speech parameterization files are smaller than MFCC parameterization files. This

reduction can be important in terms of data transferring and data storing economization.
The authors consider LSP speech parameterization as a new standard in future works related to speech synthesis in Laboratorio de Tecnologías del Lenguaje FI UNAM.

After conducting the experiments described in this document, to new voices were developed using male and female speakers. Both were parameterized with LSP. They haven´t been statistically validated but early tests showed certain success in intelligibility and naturalness, their validation remains for future work.
Further studies imply making a detailed analysis of the HMM usage within the system. Adjustments in that stage may lead to an improvement in quality independently of the chosen parameterization. Current studies on speech synthesis and recognition propose the use of Deep Neural Networks instead of HMM.

9 **REFERENCES**

[1] A. H. 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.

[2] K. Tokuda, T. Yoshimura, and T. Masuko, "Speech parameter generation algorithms for HMM-based speech synthesis," *, Speech, Signal …*, 2000.

[3] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "MEL-GENERALIZED CEPSTRAL ANALYSIS — A UNIFIED APPROACH TO SPEECH SPECTRAL ESTIMATION," 1994.

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

[5] A. Arakawa, Y. Uchimura, H. Banno, F. Itakura, and H. Kawahara, "High quality voice manipulation method based on the vocal tract area function obtained from sub-band LSP of STRAIGHT spectrum," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. 2, pp. 4834–4837, 2010.

[6] T. Bäckström, "LINEAR PREDICTIVE MODELLING OF SPEECH - CONSTRAINTS AND LINE SPECTRUM PAIR DECOMPOSITION."

[7] 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.

[8] P. Taylor, A. Black, and R. Caley, "The architecture of the Festival speech synthesis system," 1998.

[9] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis.," 1997.

[10] S. Young, "The HTK Book," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2013.

[11] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," *IEEE Speech Synth. Work.*, 2002.

[12] A. Black, "CLUSTERGEN: a statistical parametric synthesizer using trajectory modeling.," *INTERSPEECH*, 2006.

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

[14] F. Itakura and N. Sugamura, "LSP speech synthesizer its principle and implementation," *Trans. Comm. Speech Res.*, 1979.

[15] I. McLoughlin, "Line spectral pairs," *Signal Processing*, 2008.

[16] A. Härmä, M. Karjalainen, L. Savioja, and V. Välimäki, "Frequency-warped signal processing for audio applications," *J. Audio*, 2000.

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

[18] J. Trangol and A. Herrera, "Traditional Method and Multi-Taper to Feature Extraction Using Mel Frequency Cepstral Coefficients," *Int. J. Inf.*, 2015.

[19] P. Mahalanobis, "On the generalized distance in statistics," *Proc. Natl. Inst. Sci. (*, 1936.