Towards a Gendered Mexican Spanish Emotive Speech Synthetic Voice

Abel Herrera-Camacho and Fernando Del R ó-Ávila
Laboratorio de Procesamiento de Voz, Facultad de Ingenier á, UNAM
Email: {abelhc, hitosan}@hotmail.com

***Abstract*—A new Mexican Spanish voice was created from a set of emotive recordings (neutral, happy, sad and angry) taken from two speakers (male and female). All recordings were used to generate a single database, from this database we extracted the emotional information of each phrase and added new tags to the phonetic transcription to select the correct gender and emotion during training and synthesis time.**

***Index Terms*—emotive speech synthesis, HTS synthesis technique, hidden Markov models, Mel frequency cepstral coefficients**

I. Introduction

At the beginning of this century, a synthesis system was created using Hidden Markov Models (HMM's). This system selects subphonemes from centroid subphonems of VQ clustering. It was created by Dr. Tokuda and his group, it was called 'HMM-based Text to Speech' (HTS) [1]-[3].

Also the FESTIVAL group created a synthesis system based on HMM's, it was called CLUSTERGEN [4], with similar voice naturalness to the Tokuda System [5], [6]; The most important difference between HTS and CLUSTERGEN is the last one does not use an impulse filter, it takes the subphonemes directly from the corpus.

At HTS, the subphonemes are represented by MFCC's, F0 and time duration. Other core change at this century, was the use of other parametrization instead of MFCC's. The most famous is STRAIGHT [7], [8]. It uses a set of parameters from the spectral envelope of the subphoneme, this one is in database.

A natural voice was created for Mexican Spanish, at our laboratory, using HTS [5], [6], obtaining good results. Once this baseline has been generated, a new set of recordings were now generated, to obtain a new synthetic voice with emotive components (neutral, happy, sad and angry). Two sets of emotive recordings, one for male and other for female were used, as well as one purely neutral set.

Three database sets were created, one for each recording (male and female) and an additional one using all voices. In the combined database, gender and mood tags were used to select the appropriate set of parameters during synthesis time.

Manuscript received November 1, 2015; revised May 4, 2016.

During synthesis, a modified Festival [9], [10] build is used, that allows the use of additional tags during the text analysis phase to add mood and gender information to the phone transcription.

II. Database Recording and Labelling

For the database recording a set of random phrases were taken from different sources (including fairy tales, novels and news sources) to generate phrases of different styles and length. These phrases were scanned for phonetic balance, replacing them as needed to obtain better coverage of uncommon phones.

Two hundred fifty phrases were used for each mood (approximately 30 minutes of audio for neutral, happy, sad and angry speech). We generated 2 sets of emotive speech (one male and one female) and one purely neutral set (male).

Labelling was done using the festival EHMM labeler [11], using 27 phones (17 consonants, and 10 vowels (stressed and unstressed vowels were treated as different phones to facilitate processing). This process extracts the most likely starting and ending points of each phone from the transcription of the recordings.

Labels were stored in the festival utterance format, which stores a tree containing the context of each phone.

Context information contains phone, syllable and word position within the larger groups, phone name and classification (consonant, vowel, articulation and voicing type, etc.).

III. Database Training

Database training was done using the HTK toolkit [12] with HTS. The voice recordings were first normalized at -3dB. It was found that at higher values the synthesis filter became unstable, while at lower volumes the parameter extraction was not always successful.

Once the recordings were normalized, the Mel-cepstral parameters as well as the f0 contours are extracted and the context information for each of the recorded phones was generated from the previously generated utterance files.

For the context information, additional parameters were inserted into the label files to account for different moods and voice providers. In addition, a Spanish dictionary was created to indicate the part of speech of each word in the recording. Manual labelling was

doi: 10.18178/ijsps.4.6.519-522

required to account for homonyms, as linguistic analysis is not yet complete.

The dictionary is stored as a SQLite database, which is read during runtime to extract the part of speech of each word, as well as a mood weight for the most common words of the language.

The training process generates three sets of parameters (cepstral, f0 and duration) [1]. For cepstral and f0 parameters each phone is divided into states (in this case, 5 states per phone were used). These parameters are clustered using HMMs according to context, selecting the most usual phone variation for each context item.

For each state of every phone, a CART (Classification and Regression Tree) is generated. These are binary decision trees which select a particular copy of a phone segment according to context. For each tree, context information that adds little or no variations to the phones are discarded.

During synthesis time, each tree is traversed independently and the most appropriate segment is selected, interpolation can be used to minimize discontinuities in the selected data.

IV. Results

Six different voices were generated from the recorded data, 3 neutral (both males and female) and 3 emotive (male, female and all voices combined).

In Table I, the number of final nodes in the selection tree is shown. As can be seen the number of nodes for f0 is much larger than the ones for cepstral, as phones change little between different contexts, as most of the intonation and emotive information is carried by pitch and duration changes.

TABLE I. Number of States Used on Different Voice Combinations

Set	Type	Number of final states				
		Cepstral coefficients used				
		1	2	3	4	5
Male 1	Neutral	70	79	78	66	88
	Emot.	131	160	152	122	132
Female	Neutral	73	66	63	68	69
	Emot.	124	131	128	152	142
Male 2	Neutral	95	88	88	84	94
All	Emot.	366	352	327	340	348

Set	Type	f0 values used					dur
		1	2	3	4	5	
Male 1	Neutral	285	511	402	214	315	214
	Emot.	552	769	751	287	512	367
Female	Neutral	329	292	412	391	262	222
	Emot.	417	562	822	563	409	400
Male 2	Neutral	228	336	327	243	217	201
All	Emot.	1349	1277	1362	1013	950	795

In the same table we can see that the number of final nodes for each of the 5 states used for each phone remains roughly similar, so even in the less stable beginning and ending parts of the phone, the variations between different contexts are not as large as expected.

The combined voice has little savings (around 10%), as the parameters are too different between each recording, so few parameters can be combined together and the end result is mostly all parameters copied together into a single data set, but it makes selection between different moods and speaker easier during synthesis time as it can be selected by parameters instead of by loading a new dataset.

From a more detailed analysis on the generated CART trees, we can see (Table II) that the cepstral parameters take into consideration fewer context parameters than in the case of F0, this is consistent with the analysis done for final nodes, as most of the changes due to context appear in F0 and duration, than in variations of the harmonic content of the signal.

TABLE II. Number of Context Questions Used and Filesize on Different Voice Combinations

Set	Type	Context questions used			Filesize (kb.)		
		cep	f0	dur	cep	f0	dur
Male 1	Neutral	129	645	158	321	82	9
	Emotive	185	761	214	415	136	15
Female 1	Neutral	126	679	153	287	80	9
	Emotive	181	796	226	403	131	16
Male 2	Neutral	147	532	122	377	64	8
All	Emotive	269	901	333	1022	280	32

The outputs from the combined voice are not exactly the same as with the individual emotive voices (see Fig. 1), but are very similar, with the original voice owner being easily distinguished.

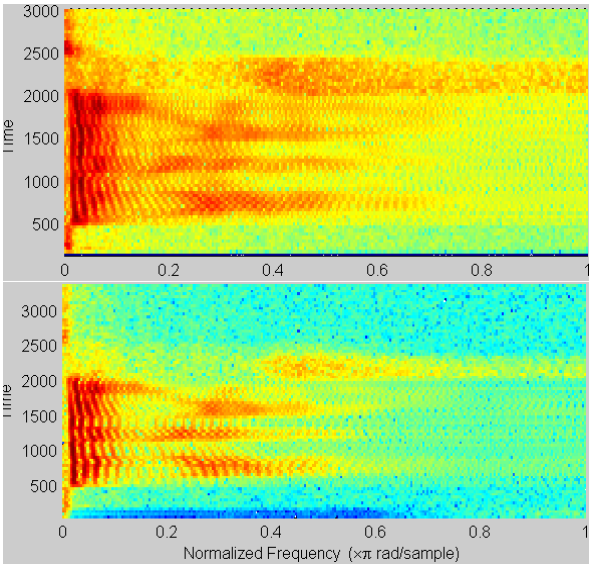

Figure 1. Comparison of the male voice. Top: Individual voice bank. Bottom: Combined voice bank

V. Output Generation

Once the output voices were generated a small frontend was created (see Fig. 2). This frontend allows

the selection of the speaker and mood that will be used to generate the synthetized voice.

This frontend is linked to a modified festival speech synthetizer. This modified system allows the introduction of the additional context information (speaker, mood and part of speech), as well as preliminary auto mood tagging.

In the case of automatic mood tagging a mood weight is read from the dictionary and a total mood value is calculated for each paragraph. In this case for each paragraph a mood is selected depending on the mood that has the highest weight. In case of a tie a neutral mood is used.

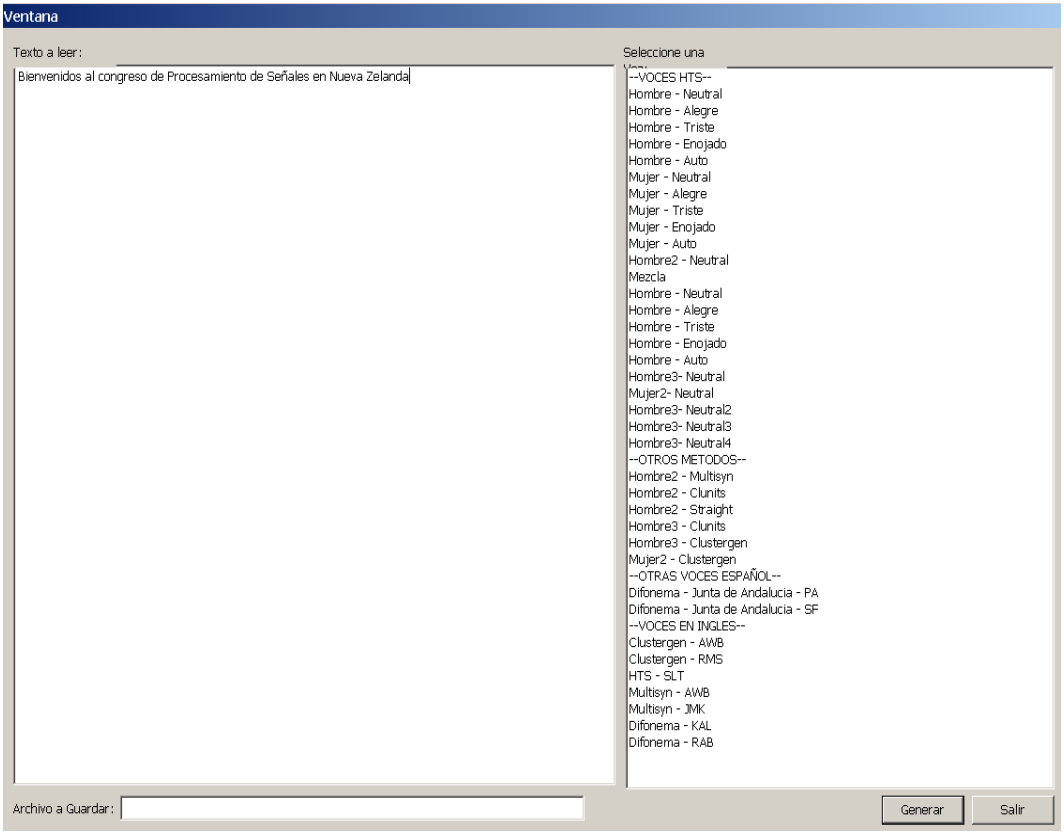

Figure 2. The synthesis frontend

VI. Future Work

This work has created the basic framework for emotional speech synthesis, however much work still need to be done.

Part of speech analysis needs to be improved to better extract the type of word, particularly in the case of homophone words. Extraction of the function of each word (subject, object, etc.) could also help in improving the phrasing of the synthetic output. This might also improve the mood weighting, as the weight could be modified depending on the context where the words are used.

Currently, the voice parameters are stored using mel-cepstral. A higher quality parametrization (for example, STRAIGHT) could be used to improve audio quality at the exchange of much higher disk requirements (STRAIGHT would increase disk requirements by approximately 50-100x).

VII. Conclusions

An emotive speech synthesis voice has been created for Mexican Spanish. This system has support for automatically tagging mood information based on weighting from emotional target words.

The main result of the current system is the generation of emotive information without the need of human intervention. The current database is still in its preliminary stages and further training needs to be carried out for improvements in mood selection.

As the recordings used are not from professional speakers and were done on consumer grade equipment, the final audio quality still needs improvement. Even with these recordings, the emotional information is easily distinguished.

Acknowledgment

This research is part of a Project being carried out with support from UNAM-DGAPA-PAPIIT (IT102314), the authors wish to give thanks for the provided support.

References

[1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2000, pp. 1315-1318.

[2] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE SSW*, 2002.
[3] H. Zen and T. Nos. The HMM-based Speech Synthesis System (HTS) Version 2.0. [Online]. Available: htt://mir.cs.nthu.edu.tw/users/heycat/relatedPapers/2_The%20HMMbased%20Speech%20Synthesis%20System%20Version%202.0.pdf
[4] A. W. Black. CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling. [Online]. Available: http://www-2.cs.cmu.edu/~awb/papers/is2006 /IS06139 4.PDF
[5] A. Herrera and F. D. R ó, "Development of a Mexican Spanish HMM-based synthetic voice," in *Proc. 18th Congreso Mexicano de Ac ústica*, 2011.
[6] A. Herrera-Camacho; and F. D. Rio-Ávila, "Development of a Mexican Spanish synthetic voice using synthesizer modules of festival speech and HTS-straight," *International Journal of Computer and Electrical Engineering*, vol. 5, no. 1, pp. 36-39, 2013.
[7] H. Kawahara, I. Masuda-Kastuse, and A. Cheveigne, "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Journal of Speech Communication*, no. 27, pp. 187-207, 1999.
[8] L. Chang and D. Kewley-Port, "STRAIGHT: A new speech synthesizer for vowel formant discrimination," *Acoustic Research Letters Online*, vol. 5, no. 31, 2004.
[9] Festival speech synthesis system. [Online]. Available: http://www.cstr.e d.ac.uk/projects/festival/
[10] A. Black and K. Lenzo. (2000). Building voices in the festival speech synthesis system. [Online]. Available: http://festvox.org/bsv/
[11] HTK toolkit. [Online]. Available: http://htk.eng.cam.ac.uk/
[12] HMM-Based speech synthesis system. [Online]. Available: http//hts.sp.nitech.ac.jp/

Abel Herrera Camacho graduated in 1979 as a mechanical-electrical engineer, acquiring the Master in Electrical Engineering in 1985 and his Ph.D. in 2001, all at the University of Mexico (UNAM), the Ph.D. with help from the University of California-Davis. He carried out a post-doctoral internship at Carnegie Mellon University in 2001, as well as a research internship at the University of Southern California. He is author of over 50 papers in speech codification, recognition and synthesis. He has involved in several projects with the industry and UNAM worth about $50 million dollars. He is Head of the Speech Processing lab, and has taught at the University since 1979.

Fernando del R ío graduated in 2003 as a computer engineer at the University of Mexico (UNAM). He obtained a master degree in electrical engineering at UNAM in 2005. He has a master in engineering manager from McNeese State. He is actually working in a software private company.