The GTH-CSTR Entries for the Speech Synthesis Albayzin 2010 Evaluation: HMM-based Speech Synthesis Systems considering morphosyntactic features and Speaker Adaptation Techniques

R. Barra-Chicote[1], *J. Yamagishi*[2], *J. M. Montero*[1], *O. Watts*[2], *S. King*[2], *J. Macias-Guarasa*[3]

[1]Speech Technology Group, Universidad Politecnica de Madrid
[2]Center for Speech Technology Research, University of Edinburgh
[3]Geintra Group, University of Alcala

barra@die.upm.es, jyamagis@inf.ed.ac.uk, macias@depeca.uah.es

Abstract

This paper describes the GTH-CSTR systems developed for the *Albayzin 2010 Speech Synthesis Evaluation*. We have developed three different HMM-based systems to build synthetic voices in Spanish, using two hours of speech of a male speaker. We have improved our baseline system (GTHCSTR-2008) by using morphosyntactic features, iterative segmentation algorithms, enhanced feature analysis and speaker adaptation techniques.

Index Terms: text to speech synthesis, statistical parametric speech synthesis, morphosyntactic features, speaker adaptation, speech synthesis evaluation

1. Introduction

The quality of HMM-based speech synthesisers has been improving in the recent years, also showing good intelligibility rates. However, the over-smoothing tendency, typical of these synthesisers, causes that most of the sentences are spoken in a very closely form. We have incorporated morphosyntactic features to the system, looking to improve the prosody generation of our text-to-speech system (TTS) and to enrich the way it reads complex sentences.

One of the features of HMM-based synthesis is their flexibility as compared to unit selection synthesis. Since we have an explicit speech model, its parameters can be modify more easily modified to obtain new voices. The application of model adaptation techniques to an average voice [1], allows the possibility of building a target speaker voice using only a few minutes of speech. We have incorporated those techniques to our baseline system and present an additional entry to the evaluation.

2. Albayzin 2010 Speech Synthesis Evaluation

The *Albayzin 2010 Speech Synthesis Evaluation* is an event, similar to the *Blizzard Challenge*, promoted in order to compare different techniques for building corpus-based speech synthesisers applied to Spanish. The challenge consists of building a voice from a released data set and synthesising a predefined set of test sentences, which are perceptually evaluated through listening tests by volunteers and speech experts.

Each voice is evaluated in terms of:

- Similarity with the target speaker
- Naturalness
- Intelligibility

3. Corpora

The organisation has released the UVIGO_ESDA corpus as the target speaker synthetic voice for this challenge. In addition we have also used our Spanish Expressive Voices (SEV) corpus as part of the training data in one of our three submitted systems.

3.1. UVIGO_ESDA Corpus

The UVIGO_ESDA Database contains speech recordings from an amateur male speaker that read prompted texts in "neutral" style. The database approximately contains two hours of speech and 1217 phonetically balanced sentences, automatically extracted from journalistic texts by means of a greedy algorithm. Data collection was performed at a recording studio. Audio files with the original sampling frequency (44100 KHz) were also provided for training tasks (test audio files should be in 16 KHz).

3.2. SEV Corpus

The *Spanish Expressive Voices* (SEV) corpus [2] comprises speech and video recordings of an actor and an actress speaking in a neutral style and simulating six basic emotions: *happiness*, *sadness*, *anger*, *surprise*, *fear* and *disgust*.

The SEV corpus covers speech data in several genres such as isolated word pronunciations, short and long sentences selected from the SES corpus [3], narrative texts chosen from a novel "Don Quijote de la Mancha", a political speech, short and long interviews, question answering situations and short dialogues. The texts of all utterances are emotionally neutral.

More than 100 minutes of speech duration per emotion have been recorded, allowing for comprehensive studies in emotional speech synthesis, prosodic modelling and speech conversion. The amount of data per emotion and speaker is close to one hour of speech.

For this challenge, we have used the neutral voice of the SEV male speaker as the *Base Voice* in the training process of the GTHCSTR-2010 Adaptation-based system described in the following section. However, for further research we are also interested in using positive emotional voices as the *Base Voice*.

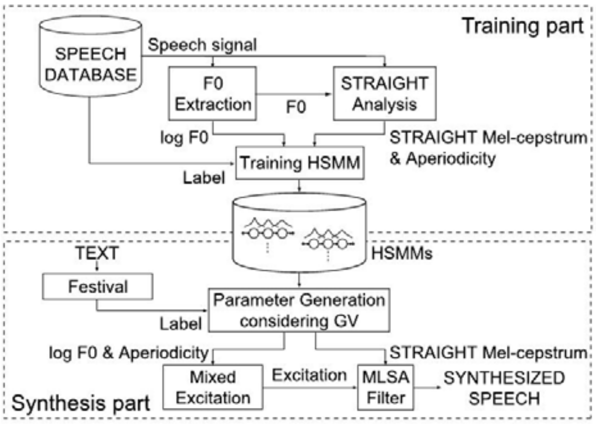

Figure 1: *Blocks diagram GTHCSTR-2008 system.*

4. Systems Description

4.1. GTHCSTR-2008: Baseline System

Our HMM-based voices have been built using a method similar to the Nitech-HTS 2005 system [4] which is publicly available from the HTS toolkit website [5].

The HMM-based speech synthesis system comprises three components: speech analysis, HMM training, and speech generation. In the speech analysis part, three kinds of parameters for the STRAIGHT [6] mel-cepstral vocoder with mixed excitation (the mel-cepstrum, $\log F0$ and a set of aperiodicity measures) are extracted as feature vectors for modelling by the HMMs. These are as described in [4], except that the F0 values we used were more robustly estimated using a vote amongst several F0 extraction algorithms [7]. In the HMM training part, context-dependent multi-stream left-to-right MSD-HSMMs [8] are trained using the maximum likelihood criterion. In the speech generation part, acoustic feature parameters are generated from the MSD-HSMMs using the GV parameter generation algorithm [9]. Finally an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add (PSOLA) [10]. This signal is used to excite a mel-logarithmic spectrum approximation (MLSA) filter corresponding to the STRAIGHT mel-cepstral coefficients, generating the speech waveform. Figure 1 plots the blocks diagram of the GTHCSTR-2008 system.

The GTHCSTR-2008 system exhibited very good performance in the previous *Albayzin 2008 Speech Synthesis Evaluation* [11]. Emotional synthetic voices have been recently developed with this system [12].

4.2. GTHCSTR-SA-2010: Adaptation-based System

The GTHCSTR-2008 system has been improved with the inclusion of adaptation techniques. We have incorporated CSMAPLR an MAP adaptation algorithms in the voice training processes, fully described in [13, 14].

An average voice [1] is usually used in speaker independent TTS systems [14] as the *Base Voice* that is adapted to the target speaker. However, for the challenge we used the neutral male voice (built with GTHCSTR-2008 system) from the SEV corpus as the *Base Voice*. This voice was adapted to the target speaker of UVIGO_ESDA corpus only using the first 50 sentences (5% of training data) from the training set (approximately 5 minutes of speech). Figure 2 plots the block diagram of the GTHCSTR-SA-2010 system.

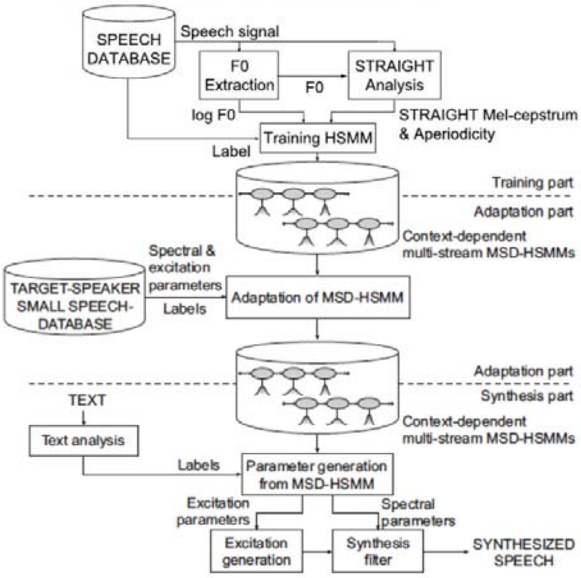

Figure 2: *Block diagram of GTHCSTR-SA-2010 adaptation-based system.*

The objective of this GTHCSTR-2008 system is comparing the results between the speaker dependent voices of the target speaker (GTHCSTR-SD-2010 system described below) and an adapted voice trained using only a very small amount of training data.

4.3. GTHCSTR-SD-2010: Speaker Dependent System

4.3.1. Acoustic processing improvements

We have modified our speaker dependent system by using some acoustic improvements, as compared to the GTHCSTR-2008 implementation:

- A bigger spectral bandwidth using the original sampling frequency (44100 KHz) in the training process. In the synthesis stage, the speech signals were down-sampled to 16000 KHz.
- A higher number coefficients in the analysis of the spectral component.
- An iterative segmentation process based on building partial voices used for relabelling.

4.3.2. Morphosyntatic Features

In order to improve the basic HMM-based system, we have also included new features coming from a morphosyntactic analysis of the input sentences. As the natural language processing (NLP) of the speech synthesis sentences should be very robust (in order to deal with whatever grammatical structures the author of the target texts could use), shallow techniques seem to be a good choice. The first module in our NLP chain is a Spanish Part-Of-Speech tagger (Montero 2003), based on ESPRIT-860's EAGLES-like 10-byte labels (more than 250 possible tags), using a set of dictionaries such as RAE's 159898-word dictionary, richly-tagged ESPRIT-860's 9883-word dictionary, Onomastica's 58129-proper-noun dictionary, GTH's 1960-multiword

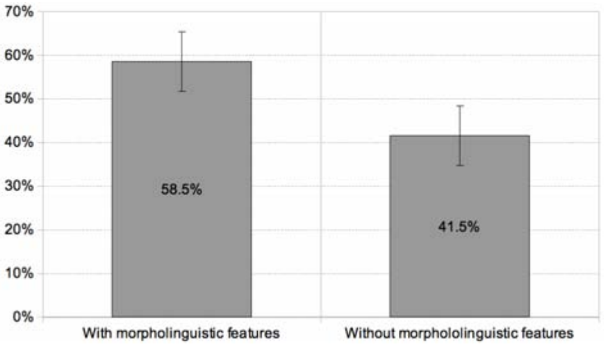

Figure 3: *Comparison between Speech Quality (SQ) scores obtained by GTHCSTR-SD-2010 speaker dependent system considering morphosyntatic features or not.*

expression dictionary, and GTH's verb conjugation analyser (including 102 irregular paradigms and 4325 infinitives).

After assigning all possible tags to the words, several sets of hand-written rules are used for cascade-filtering impossible tag sequences: GTH's 77 high-recall rules, CRATER's 148 rules and GTH's 162 low-recall high-precision rules. On the 38172-word test-set of the ESPRIT-860 Spanish corpus, the recall is as high as 0.9987 when averaging 1.6984 tags per word.

Finally, the TnT stochastic tagger (Brants) is used for disambiguation. This tagger uses an interpolated language model based on trigrams, bigrams and unigrams, resulting in a 98.99% accuracy for a 1-tag-per-word basis, or 99.45% if 1.0238 tags are assigned per word on average.

After tagging the sentence, 2 features are available to be used in the speech synthesis training and testing process:

- A gross 10-category feature (based on 860 tag).
- A 3-byte set of tags in 860 coding scheme, (including a first byte for the 10 main tags and 2 additional bytes for a more detailful sub-tagging).

The final NLP processing module is a shallow parser based on a CYK botton-up algorithm and a set of 2179 hand-written general-purpose CYK parsing rules. As these rules are very ambiguous, many possible parser trees are assigned to each sentence. In order to control the exponential growth of this analysis, a small set of splitting rules were developed (trying to reduce the length of the text to be analysed) and a final filtering process was used, selecting only one tree using a Minimum Description Length approach. In a subset of the test set, for a total 5703 shallow syntactic phrases, there were 0.35% cutting errors, 0.55% tagging-recall errors, 1.10% tagging-precision errors and 1.49% syntactic-analysis errors. These shallow syntactic phrases are the third feature to be used in the synthesis process.

reader to understand the figure Figure 3 shows the results of an internal perceptual test to validate the improvements when adding the morphosyntatic features. Based on these results (a 41% statistically significant relative improvement in Speech Quality) our final system considered the morphosyntatic features described.

5. Results and Discussion

Figure 4 shows the similarity scores for all systems submitted to the evaluation (for all listeners). GTHCSTR-2008 is plotted using a red box, GTHCSTR-SD-2010 is plotted using a green box, and GTHCSTR-SA-2010 is plotted using a blue box. GTHCSTR-SD-2010 has good similarity scores and significantly improves our baseline GTHCSTR-2008 system.

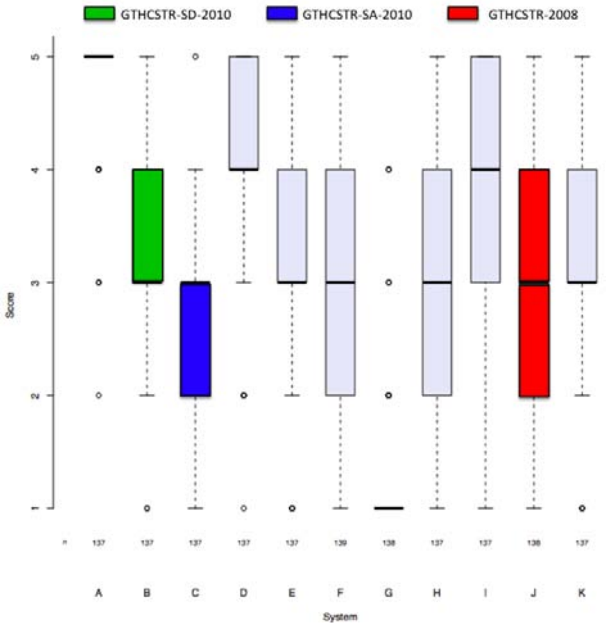

Figure 4: *Similarity scores for voice UVIGO-ESDA (All listeners).*

Our speaker adapted system, GTHCSTR-SA-2010, obtains excellent similarity results (median value of 3, equivalent to the speaker dependent systems) considering that we have only used 50 sentences (5% of the whole training data) of the target speaker to build its voice. This result strongly supports the goodness and the high potential of the speaker adaptation algorithm.

Figure 5 shows the MOS scores (considering all listeners). Again, GTHCSTR-SD-2010 significantly improves GTHCSTR-2008. In this case, as expected, the MOS scores obtained by GTHCSTR-SA-2010 are lower than the speaker dependent system. These results are reasonable since the *Base Voice* was built only with 50 minutes of speech (in comparison with the 2 hours of UVIGO_ESDA) and we only used 5 minutes of adaptation data.

Figure 6 shows the MOS scores only considering the listeners that did not used headphones. In this case, our GTHCSTR-SD-2010 system did not show significant diferences with the two systems which obtained better MOS scores. We presume that the over-smoothing introduced by our synthesis technique is filtered by the channel when synthetic speech is heard using speakers instead of headphones.

Figure 7 shows the Word Error Rate in the intelligibility tests (WER considering all listeners). GTHCSTR-SD-2010 has lower WER than GTHCSTR-2008. Also, there are no significant differences between the best system (system E) and GTHCSTR-SD-2010 and GTHCSTR-2008.

6. Conclusions

This paper described the GTH-CSTR systems submitted to the *Albayzin 2010 Speech Synthesis Evaluation*. All of them are

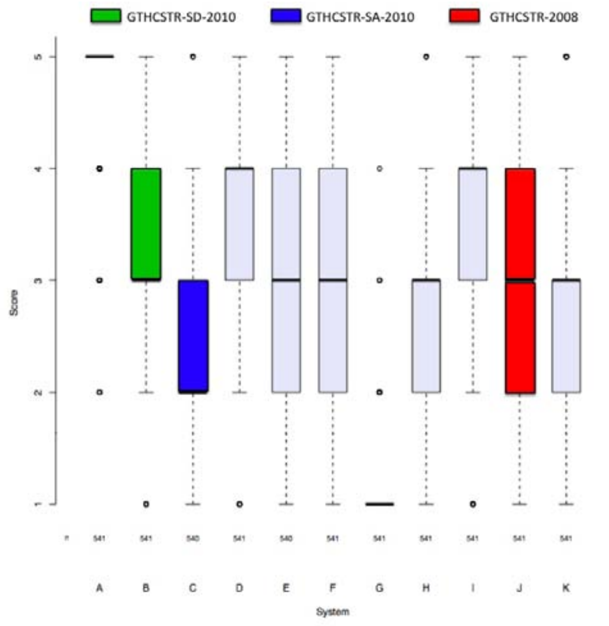

Figure 5: *Mean opinion scores for voice UVIGO-ESDA (All listeners).*

based on HMM-based synthesis to build synthetic voices in Spanish.

Three systems have been presented:

- A baseline system submitted to the 2008 Albayzin Evaluation (GTHCSTR-2008)
- A system based on speaker adaptation algorithms (GTHCSTR-SA-2010)
- An improved speaker dependent system (GTHCSTR-SD-2010).

The synthetic voice built with GTHCSTR-SA-2010 system is reasonably perceived as the target speaker, in spite of having used only the 5% of the training data.

An internal evaluation validated the goodness of adding morphosyntatic features in order to improve the quality of our GTHCSTR-SD-2010 synthetic voices.

7. Acknowledgements

The authors which to express their gratitude to the members of GTH and CSTR for their collaboration in the preparation of this work.

8. References

[1] J. Yamagishi and T. Kobayashi, “Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training,” *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[2] R. Barra-Chicote, J. Montero, J. Macias-Guarasa, S. Lufti, J. M. Lucas, F. Fernandez, L. D’haro, R. San-Segundo, J. Ferreiros, R. Cordoba, and J. Pardo, “Spanish Expressive Voices: Corpus for emotion research in Spanish,” in *Proceedings of 6th international conference on Language Resources and Evaluation*, 2008.

[3] J. M. Montero, J. M. Gutierrez-Arriola, S. Palazuelos, E. Enriquez, S. Aguilera, and J. M. Pardo, “Emotional speech synthesis: From speech database to TTS,” in *Proc. ICSLP-98*, Dec. 1998, pp. 923–926.

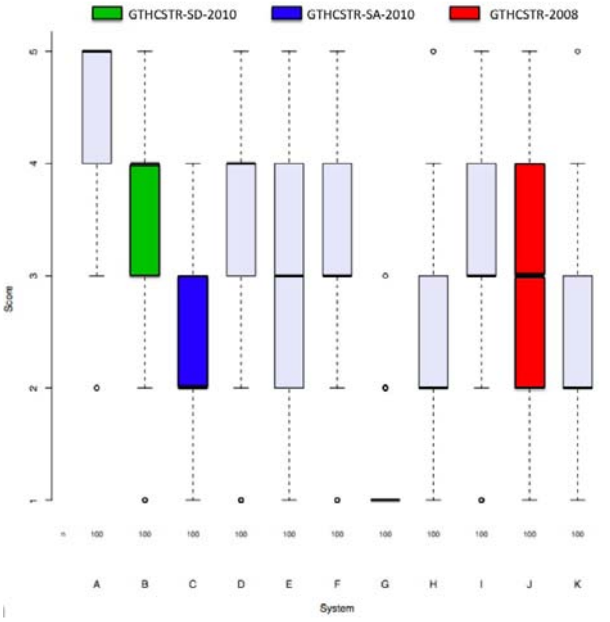

Figure 6: *Mean opinion scores for voice UVIGO-ESDA (listeners headphones=NO).*

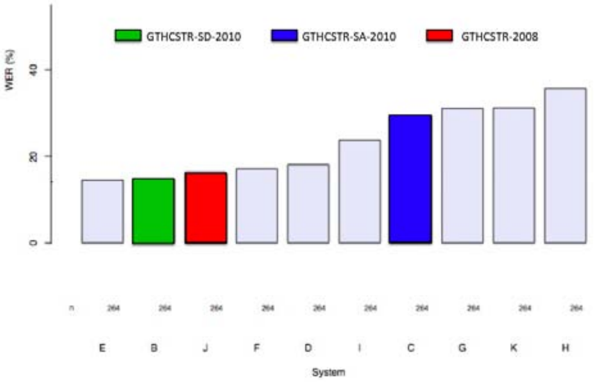

Figure 7: *Word error rate for voice UVIGO-ESDA (All listeners).*

[4] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, “Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005,” *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[5] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, and T. Nose, *The HMM-based speech synthesis system (HTS) Version 2.1*, 2008, http://hts.sp.nitech.ac.jp/.

[6] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, “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,” *Speech Communication*, vol. 27, pp. 187–207, 1999.

[7] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, “A robust speaker-adaptive HMM-based text-to-speech synthesis,” *IEEE Trans. Speech, Audio & Language Process.*, 2009, (in press).

[8] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “A hidden semi-Markov model-based speech synthesis system,” *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[9] T. Toda and K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[10] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech Communication*, vol. 9, no. 5-6, pp. 453–468, 1990.

[11] R. Barra-Chicote, J. Yamagishi, J. Montero, S. King, S. Lutfi, and J. Macias-Guarasa, “Generacion de una voz sintetica en Castellano basada en HSMM para la Evaluacion Albayzin 2008: conversion texto a voz,” in *V Jornadas en Tecnologia del Habla*, Nov. 2008, p. Please add page range.

[12] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, “Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech,” *Speech Communication*, vol. 52, no. 5, pp. 394 – 404, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/B6V1C-4XY4GDS-1/2/7e701c2305a5ff0713d2c2e83af6e760

[13] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm,” *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 1, pp. 66–83, 1 2009.

[14] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, “Speaker-independent HMM-based speech synthesis system — HTS-2007 system for the Blizzard Challenge 2007,” in *Proceedings BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.