The GTM-UVigo Systems for Albayzín 2010 Text–to–Speech Evaluation

Francisco Campillo Díaz[1], *Francisco Méndez Pazó*[1],
Montserrat Arza Rodríguez[2], *Eduardo Rodríguez Banga*[1]

[1]Group on Multimedia Technologies, Universidade de Vigo, Spain
[2]Centro Ramón Piñeiro para a Investigación en Humanidades. Xunta de Galicia.
{campillo, fmendez, marza, erbanga}@gts.uvigo.es

Abstract

This paper describes the synthesizers Cotovía and Cotovía-hts developed by the Group on Multimedia Technologies of the University of Vigo, in cooperation with "Centro Ramón Piñeiro para a Investigación en Humanidades". Cotovía is a state–of–the–art concatenative system based on a combined selection of acoustic and intonation units, while Cotovía–hts is a preliminary version of a synthesiser based on HMM technology.

Index Terms: speech synthesis, unit selection, statistical parametric, HMMs

1. Introduction

This paper describes the current state of the synthesizers Cotovía and Cotovía–hts at the moment of the Albayzín 2010 Text–to–Speech Evaluation. Cotovía is a state–of–the–art corpus based text to speech synthesizer [1]. In this kind of systems, synthetic speech is generated by concatenation of natural segments selected from a large database recorded from the same speaker. The underlying assumption is that synthetic speech will be indistinguishable from natural if segments are used in similar contexts to those from which they were originally extracted. In practice, this technology alternates chunks of completely close to natural speech, with sporadic and hard to predict concatenation artifacts that degrade severely the quality of the synthetic speech.

Regarding Cotovía–hts, is the first immersion of GTM on HMM based speech synthesis [2], and the result of three months of work, so it can only be considered as a preliminary version. Synthetic speech generated by this kind of systems is smooth and very stable, without the frequent artifacts of unit selection, but with a general artificial impression.

The article is outlined as follows: Section 2 describes the process of building the voice for the two systems, both regarding the information that is used and the system requirements; Section 3 summarises the main characteristics of the synthesizers: system structure, features that are taken into account and prosodic modelling; Section 4 shows the results obtained in the evaluation, and, finally, Section 5 is dedicated to the overall conclusions and future lines of research.

2. Building the voice

The Spanish corpus uvigo_esda was released by the FALA2010 organizers for comparison of the different systems. It consists of around 2 hours of speech (mono, 16 kHz sampling frequency and 16 bits/sample) of isolated sentences read in a neutral style by an amateur speaker. The corpus is phonetically balanced according to the frequency of appearance of phones in the Spanish language, and contains sentences of different lengths and belonging to four broad types: declarative, interrogative, exclamatory and suspended.

The organizers provided the wave files, the text files including information of intonation boundaries, the phonetic segmentation files (not manually revised), and the pitchmark files for voiced segments, as obtained directly from Praat [3].

In the process of building the voice for this evaluation, phonetic segmentation files were not manually corrected, besides some gross errors that were detected while testing the systems, specially regarding major phrase boundaries. Pitchmarks files were postprocessed with several tools developed by the research team, both to fill up unvoiced segments with equally spaced pitch marks, and to ensure that pitchmarks were always positioned at the same point of the local period, in order to avoid concatenation artifacts related to phase mismatch. Also, pitch–synchronous MFCC vectors (12 coefficients) were computed with Festival [4] to model spectral envelope continuity in joints, for the unit selection system. As for the intonation model, stylized intonation contours were extracted from the wave files using Praat, and postprocessed later to smooth out wrong values mainly related to unvoiced segments.

The automated process of building the voice for the unit selection synthesizer took around one hour in an Intel© Xeon™ server, 2.50 GHz processor and 8 GB RAM.

With regards to Cotovía–hts, we used Cotovía, Straight [5] and HTS [6] for the training process of the voice. This process took around 48 hours in an Intel© Xeon™ server, 2.40G GHz processor and 18 GB RAM. Please refer to Section 3.3 for a detailed description of the system.

3. Systems description

Cotovía and Cotovía–hts share a common linguistic module that extracts the information needed for the next stages, so this Section begins with the description of the main features of this module. After that, the distinctive characteristics of both synthesizers are presented.

3.1. Linguistic module

The linguistic module comprises several stages through which the input sentence is translated into a sequence of acoustic target units characterized by a set of features that are used both for prosodic modeling and waveform generation. In this sense, features are extracted related to phone identity, phonetic context, phone boundaries, accentuation, syllabic structure, type of sentence, position in the phonic group, part–of–speech (POS) labels and syntactic information.

With respect to POS, a hybrid analysis is performed [7]. First, a reduced set of highly reliable linguistic rules is used to eliminate from each word those categories that are not possible according to the context. Second, a statistical tagger makes up the final decision combining a contextual 5–gram model of sequences of categories, and a lexical model that considers the probability of each word having a certain category. Figure 1 is an example of the use of the 5–gram window to consider the ambiguous context around each word.

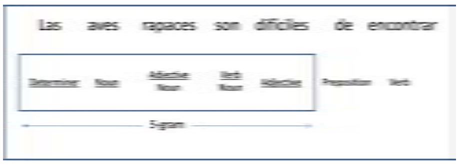

Figure 1: *Example of POS disambiguation*

3.2. Cotovía

3.2.1. *Prosody estimation*

Cotovía includes different modules to estimate duration, energy, intonation and phrasing.

Intonation Similarly to acoustic unit selection, Cotovía integrates a corpus based intonation module [8], with the accent group (defined as a sequence of unaccented words finishing in an accented one) as the basic unit for concatenation. This model is characterized by:

- Accent groups are described by their position within the phonic group and the sentence, the types of boundary surrounding them, the number of syllables, the position of the accent, the type of sentence, the POS of the accented word and the syntagma following it.
- The target cost penalizes the differences from the aforementioned features to the estimated ones. Perhaps the most interesting detail is that syntactic and morphosyntactic information is used to decide both the strength of the accent and the insertion of minor phrase boundaries [9].
- The concatenation cost only takes into account f0 continuity and boundary continuity, since it was found that joining two accent groups with different boundaries degrades severely the quality of the synthetic contour.

Phrasing Although a combined approach would probably yield better results, in Cotovía major and minor phrasing algorithms are implemented as different stages. First, major phrasing is accomplished by means of a decision tree, with factors such as the distance in syllables from the last pause and the distance in syllables to the next pause, and a POS window of three places to the left and right of the current word. And second, minor phrasing is integrated into the intonation module [9], taking major phrasing as an input. This way, minor phrasing is modeled as another subcost in the intonation target cost function, considering the POS and syntactic information as input. For every target accent group, candidate groups that can be followed or not by a minor phrase boundary are considered. Therefore, the best sequence of candidate accent groups resulting from the Viterbi search includes the best prosodic structure for the input sentence. Figure 2, where shaded and unshaded circles represent candidate accent groups with different boundaries, depicts this situation.

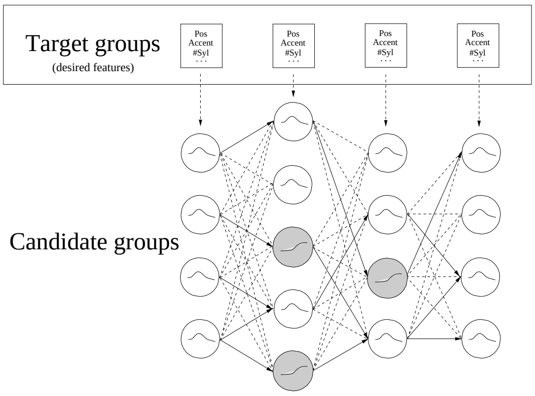

Figure 2: *Combined selection of minor phrasing and intonation contours*

Duration Phones are clustered into ten classes (open vowels, mid vowels, close vowels, voiced plosives, voiceless plosives, fricatives, laterals, nasals, vibrants and silence), and multivariant linear regression models are trained for all of them. The identity of the phone, the phones surrounding it in a window of size five, the position within the word and the phonic group, the type of sentence and the lexical accent are the features used in each model.

Energy Similar to duration, phones are divided into eleven classes (silence, open vowel, mid vowel, close vowel, voiced plosive, voiceless plosive, aproximant, fricative, nasal, vibrant and lateral), and multilayer perceptrons are trained for each of them, with features such as the identity of the phone, the energy of the previous phone, the lexical accent, the position within the sentence and the type of sentence.

3.2.2. *Acoustic unit selection*

As mentioned before, Cotovía is a corpus based synthesizer [1], with the demiphone as the basic unit for concatenation. Perhaps the most interesting difference lies in considering more than a single candidate intonation contour, as most of the other synthesizers do. In natural speech, a sentence can be realized in many different ways just by changing prosody, without affecting the meaning of the message that is conveyed. This way, in Cotovía several coherent candidate intonation contours are extracted from the intonation module, giving another degree of freedom to the acoustic unit search, and improving the quality of synthetic speech [8]. For example, Figure 3 shows two candidate intonation contours for the same sentence.

Summing up, the main characteristics of the acoustic unit selection stage are the following:

- Demiphones are parameterized according to the identity of the surrounding phonemes (in a window of size five), the lexical accent, the position within the phonic group, the type of sentence and the types of boundary. Regarding prosody, duration, energy and fundamental frequency at the beginning and end of the demiphones are considered.
- Target cost comprises two parts. First, similarity of phonemic context is computed using only symbolic features (unlike previous versions of the synthesizer, where MFCC were used with the same goal [10]). And second,

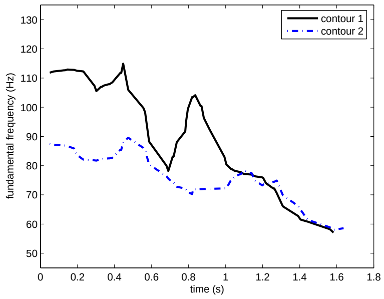

Figure 3: *Two intonation contours with different prosodic structure. Notice the minor phrase break in the solid line contour, around 0.9 seconds*

differences to the target prosody, as given by the modules mentioned before, are included.

- Concatenation cost: continuity of fundamental frequency, energy and spectral envelope are considered.

3.2.3. *Waveform generation*

Synthetic speech is generated by concatenation of the waveforms of the sequence of candidate acoustic units resulting from the Viterbi search. Demiphones close enough to the target duration and fundamental frequency values (40 ms and 5 Hz, respectively) are not prosodically modified, in order to preserve the micropsody and quality of the original recording.

3.3. Cotovía–HTS

In this section we describe Cotovía–HTS, our first statistical parametric speech synthesis system based on hidden Markov models (HMMs). This kind of systems [2] are now very popular, largely due to the release of the HMM–based Speech Synthesis System (HTS) [6][11]. Moreover, the results obtained by such systems in the Blizzard Challenge and also in the last edition of the Albayzín TTS Evaluation show that this is a very interesting and promising field in the speech synthesis research.

In brief, HTS works in two different phases: training and synthesis. In the training part, both spectrum and excitation (and its dynamic features) parameters are extracted from the speech database. This features are modeled by context–dependent HMMs (CD–HMMs), taking account of phonetic, linguistic and prosodic contexts extracted from the labelled speech database.

In the synthesis stage, the input text has to be processed to obtain a context–dependent label sequence, which HTS uses to obtain a sequence of CD–HMMs. Next, excitation and spectral parameters are obtained using the speech parameter algorithm [12], which in turn are used to generate the synthetic speech output.

Figure 4 shows an overview of our system. We use the text processing module from our unit–selection TTS synthesizer, Cotovía, to provide the context–dependent labels used for the CD–HMMs training. The linguistic features used are:

- Phoneme level:
  - Current phoneme identity.
  - Start and end time instants.
  - 2 previous and 2 next phonemes identity.

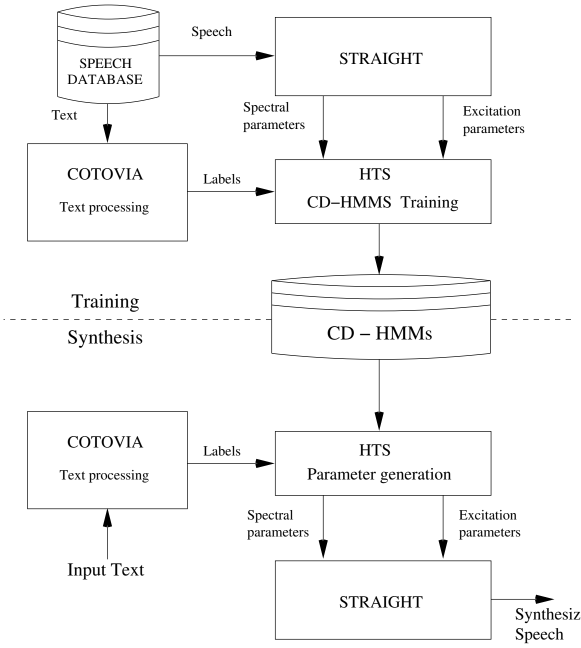

Figure 4: *Overview of Cotovía–HTS*

  - Position of the current phoneme in the current syllable (backward and forward).
- Syllable level:
  - Stress and number of phonemes in the previous, current and next syllable.
  - Position of the current syllable in the current word and phrase (backward and forward).
  - Number of stressed syllables after and before the current syllable in the current phrase.
  - Number of syllables, counting from the previous stressed syllable to the current syllable in this utterance.
  - Number of syllables, counting from the current syllable to the next stressed syllable in this utterance.
  - Vowel in the current syllable.
- Word level:
  - Part of speech (content or function) of the previous, current and next words.
  - Number of syllables of the previous, current and next words.
  - Position of the current word in the current phrase (backward and forward).
  - Number of content words after and before the current word in the current phrase.
  - Number of words, counting from the previous content word, to the current word in this utterance.
  - Number of words, counting from the current word to the next content word in this utterance.
- Phrase level:

  - Number of syllables and words in the previous, current and next phrases.
  - Position of the current phrase in the utterance (backward and forward).
  - Type (declarative, interrogative, exclamatory and suspensive)
- Utterance level:
  - Number of syllables, words and phrases in the current utterance.

Straight [5] was used to obtain the spectral and excitation parameters. In our case 39th order Mel–cepstrum, logf0 and 5 band–aperiodicity coefficients together with their dynamic features (first and second derivatives) were extracted from the speech database. These parameters were used for training, amounting in total 88878 multi–stream —5 stream and 7 states— context–dependent HMMs.

For the synthesis part, again we used Cotovía to extract the context–dependent labels from the test sentences. Then, HTS converts this label sequence into a sequence of CD–HMMs, and the speech parameter generation algorithm provides the spectral and excitation parameters. The final speech waveform is synthesized from these parameters using Straight.

4. Results

The results of the Albayzín 2010 Text–to–Speech evaluation [13] can be considered very positive for both systems (Cotovía was system "I" on the evaluation, while Cotovía–hts was system "E").

Since Cotovía participated too on the Albayzín 2008 evaluation [14] (system "B"), we can compare the results in both cases to have an idea of the improvement. Table 1 shows the results of the evaluations regarding MOS (Mean Opinion Score) and similarity to the original voice, on a scale of 1 to 5, with 1 being the worst and 5 being the best. The results of Cotovía–hts and the best system in the 2010 evaluation are also included. Comparing the output of a system with two different voices might lead to wrong conclusions, but the relative performance of Cotovía regarding the other participants in both evaluations shows an improvement as well: in 2008 Cotovía was third in both MOS and similarity to the original voice, while in 2010 it was second in MOS and similarity, with no statistical differences to the best system in this last test.

With regards to WER (Word Error Rate), there was a surprising decrease in performance, from 4.95% in 2008 to 24% in 2010. The authors consider this to be a result of a more difficult task in the intelligibility test of 2010: while in 2008 the WER ranged between 3.49% and 8.19%, in 2010 it ranged between 15% and 36%.

	MOS		Similarity	
	Mean	Median	Mean	Median
Cotovía–2008	2.91	3	3.36	3
Cotovía–2010	3.50	4	4.02	4
Cotovía–hts (2010)	3.10	3	3.15	3
Best–2010	3.78	4	4.07	4

Table 1: Cotovía: comparison between 2008 and 2010

Cotovía-hts was first in intelligibility (15%), which is clearly remarkable on being compared with other synthesizers with many years of development.

5. Conclusions

This paper describes the current state of the synthesizers Cotovía and Cotovía–hts, as were presented at the Albayzín 2010 TTS evaluation, including both the steps to build a new voice and the process followed to generate synthetic speech. The results of the evaluation were very positive, confirming also the general trend of unit selection systems being more natural and similar to the original voice, and HTS being more intelligible. Cotovía showed a clear improvement when compared with the results of the Albayzín 2008 evaluation (from 2.91 to 3.50 in MOS, and from 3.36 to 4.02 in similarity to the original voice), while Cotovía–hts was first in the intelligibility test and also outperformed in MOS and similarity other synthesizers with many years of development.

6. Acknowledgements

The work reported here was partially funded by MEC under the project TEC2009-14094-C04-04 "Búsqueda de Información en Contenidos Audiovisuales plurilingües" and Xunta de Galicia "Isidro Parga Pondal" research programme.

7. References

[1] A. Hunt, A. Black, "Unit selection in a concatenative speech synthesis using a large speech database", Proceedings of ICASSP, Volume 1, pp. 373-376, Atlanta, 1996

[2] Heiga Zen, Keiichi Tokuda, Alan W. Black, "Statistical parametric speech synthesis", Speech Communication, Volume 51, Issue 11, November 2009, pp. 1039-1064

[3] http://www.praat.com

[4] http://www.cstr.ed.ac.uk/projects/festival

[5] Hideki Kawahara, Ikuyo Masuda-Kasuse, Alain de 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", Speech Communication, 27, pp.187-207, 1999

[6] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, "The HMM-based speech synthesis system version 2.0" , Proceedings of ISCA SSW6, Aug. 2007

[7] F. Méndez, F. Campillo, E.R. Banga, "Análisis morfológico estadístico en lengua gallega", Procesamiento del lenguaje natural, Volume 31, pp 159-166, 2003

[8] F. Campillo, E.R. Banga, "A method for combining intonation modelling and speech-unit selection in corpus-based speech synthesis systems", Speech Communication, Volume 48, Issue 8, pp. 941-956, 2006

[9] F. Campillo, Jan Van Santen, E.R. Banga, "Integrating phrasing and intonation modelling using syntactic and morphosyntactic information", Speech Communication, Volume 51, Issue 5, pp. 452-465, 2009

[10] F. Campillo, E.R. Banga, "On the design of the cost functions for a unit selection speech synthesis", Proceedings of Eurospeech, Volume 1, pp. 289-292, Geneva, 2003

[11] Online: http://hts.sp.nitech.ac.jp/

[12] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech parameter generation algorithms for HMM–based speech synthesis", Proceedings of ICASSP, pp.1315-1318, June 2000.

[13] Francisco Méndez Pazó, Laura Docío-Fernández, Montserrat Arza Rodríguez, Francisco Campillo Díaz, "The Albayzín 2010 Text–to–Speech Evaluation", Proceedings of Fala2010, 2010.

[14] Iñaki Sainz, Eva Navas, Inma Hernáez, Antonio Bonafonte, Francisco Campillo, "TTS evaluation campaign with a common Spanish database", Proceedings of LREC, 2010.