FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

Building an HMM-based Spanish TTS system for Albayzin 2010 challenge

Daniela Braga[1,2], *Pedro Silva*[1], *João Freitas*[1], *Difda Monterde*[1], *Miguel Sales Dias*[1,2]

[1] Microsoft Language Development Center, Portugal
[2]ADETTI – ISCTE, IUL, Lisbon, Portugal
{i-dbraga, i-pedros, i-joaof, v-dimont, Miguel.Dias}@microsoft.com

Abstract

In this paper we describe the process of building a new Text-to-Speech (TTS) system in Spanish from the materials provided for the Albayzin 2010 challenge using the deployed Microsoft TTS technology in Spanish. The architecture of the system is described as well as the process of compiling a new voice font and producing a new synthetic voice. The main issues found during this process are presented and discussed.

Index Terms: Text-to-Speech, HTS, Spanish, voice quality, evaluation

1. Introduction

When talking about the Spanish language, one has to keep in mind that it is in the top 5 of the most spoken languages in the world [1]. Spanish is the official language in 21 countries, and it is spoken in nearly 30 countries around the world [2]. This amounts to almost 400 million of Spanish speakers, of which 350 million speak Spanish as the native language [3]. These numbers can be considered a huge motivation to work on Speech Technologies such as Text-to-Speech (TTS) systems.

For European Spanish, we find some university research groups developing text-to-speech systems based on concatenative synthesis, such as *Cotovía*, a TTS for Spanish and Galician from the University of Vigo [4]; the *Ogmios* UPCTTS for Spanish and Catalan from the Talp (Tecnologies i Aplicacions del Llenguatge i la Parla) group of the Politechnical University of Catalunya [5]; and the ViVoLab (Voice input Voice output Lab) from the University of Zaragoza [6]. The Politechnical University of Madrid tested both concatenative-based and HSMM-based synthesis systems for TTS and got better results using the last methodology mentioned [7].

Due to the large size of the Spanish speaking population, some private initiatives were interested in the market of Spanish TTS. Some of them focused on Latin American Spanish, such as *Marta* and *Miguel* from Cepstral [8], *Rosa* and *Alberto* from AT&T Labs and *Violeta* from Neospeech [9]. Some TTS from other companies offer only European Spanish or both varieties, among them we can find *Jorge, Carmen* and *Leonor* from Loquendo [10], *Amaya, Carlos* and *Laura* from Verbio [11], *Isabel, Monica, Diego, Paulina* and *Javier* from Nuance [12], and *Antonio* and *Maria* from Acapela [13]. Microsoft released a Spanish synthetic voice in Exchange 2010 deployment, named Helena, available for download both on server and client side[1], together with other languages for mobile and desktop interfaces.

2. System description

The front-end of the system is dictionary-based, being composed by a lexicon with 599520 words, phonetically annotated with phonetic transcriptions, stress marks and syllable boundaries, and with Part-of-Speech (POS) information. The front-end is also composed by the text analysis, which involves the sentence separator and word breaker modules and includes a couple of other files, such as phone set, features and the POS tags set. It also includes a rule-based Text Normalization module and stochastic-based LTS (Letter-to-Sound) converter to predict phonetic transcriptions for out-of-vocabulary words. The prosody model was trained with 2000 utterances prosodically tagged in terms of breaks, boundaries and intensity.

The front-end outputs phonetic transcriptions that are subsequently input of the TTS runtime engine or back-end, which then outputs synthetic voice. Figure 1 illustrates the system workflow.

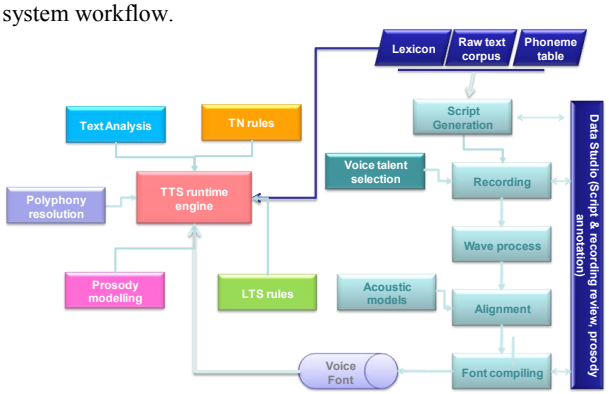

Figure 1: Microsoft TTS system pipeline.

The voice font building is also a very complex and demanding process that requires the following steps: script selection (using different text genders, phonetically balanced, with a broad prosody coverage (in terms of types of sentences – declaratives, interrogatives, exclamatory sentences), in a total of 11 500 prompts and nearly 13 hours of speech), recording process at 44 kHz, 16 bits of sampling rate), edition of the prompts, recording quality control, re-recording and edition of the prompts which failed in the quality control, wave process, automatic alignment and quality validation, font compiling and conversion of the original recorded waves to 8khz, 8 bits sample rate. Figure 2 depicts the voice font building process.

[1] Microsoft Helena can be downloaded here: http://www.microsoft.com/downloads/en/details.aspx?FamilyID=f704cd64-1dbf-47a7-ba49-27c5843a12d5&displaylang=en

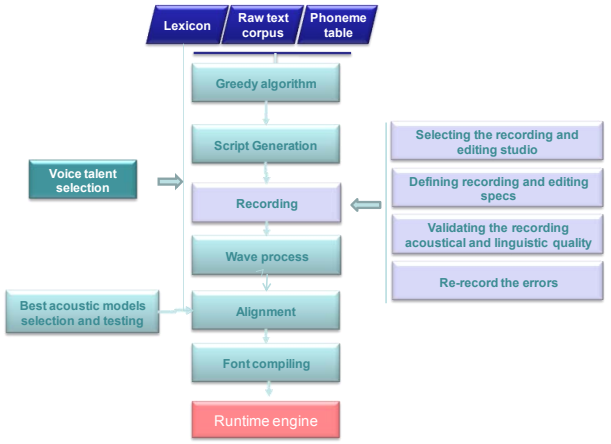

Figure 2: Voice font building process.

3. Building a new Spanish voice for Albayzin 2010

The Spanish Albayzin voice was developed using the same process and tools as the currently shipped Spanish (Spain) voice from Microsoft, which is presently available in several Microsoft commercial products and services. Additionally, the Microsoft Spanish voice is also available through a public Speech Platform SDK [14], which can be used for the development of speech-enabled applications.

As such, the current Microsoft Spanish text analysis front-end module was leveraged for preparing the input data, namely the phone set, lexicon and related components.

The provided Albayzin database consisted of 1217 phonetically balanced sentences (text sentences + wave files) recorded from an amateur male speaker in "neutral" style, comprising ~2 hours of speech.

To build the Albayzin voice the following steps were taken:

1. Prepare input wave files;
2. Prepare input script file;
3. Extract phone segmentation / alignment results;
4. Train HMM-based speech synthesis (HTS) voice font.

For step 1, all the 1217 wave files were first filtered and normalized before the training step.

In step 2, an input script file was prepared with the 1217 text sentences, including phonetic transcription + syllable + stress marks (all obtained from proprietary Spanish lexicon) as well as POS tags, per word. Before doing this, the input script file was first cleaned of some errors that were found originally, such as empty lines and more than one sentence per line.

Step 3 consists of an automatic process for extraction of phone segmentation labels, based on the script file and the wave files.

Finally in step 4, the HTS voice font was trained using the script, wave files and phone alignment results.

Once finished, the public Speech Platform [14] was used to synthesize the output test wave files (the original Spanish voice font was replaced with Albayzin trained voice font, keeping the runtime TTS engine and the frontend text analysis module).

4. System Evaluation

After training our system with the Albayzin database (1217 prompts and its correspondent wave files), a set of 430 sentences were provided in order to be synthesized and uploaded in an online platform designed for this purpose. Then Albayzin organization created an online evaluation tool where all synthesized utterances from the 10 systems participating in the challenge were randomly displayed, always having the original voice as baseline. Three tests were presented, each one testing naturalness (proximity to the original and MOS) and intelligibility (through a SUS test). Microsoft contributed to this evaluation with 13 participants, 11 of each from outside Microsoft. Participants' requirements to perform this evaluation were: being native or nearly native speaker and expert in language technology. The results of this evaluation are described in [17].

Microsoft's system (system G) results are depicted in Figures 3-5:

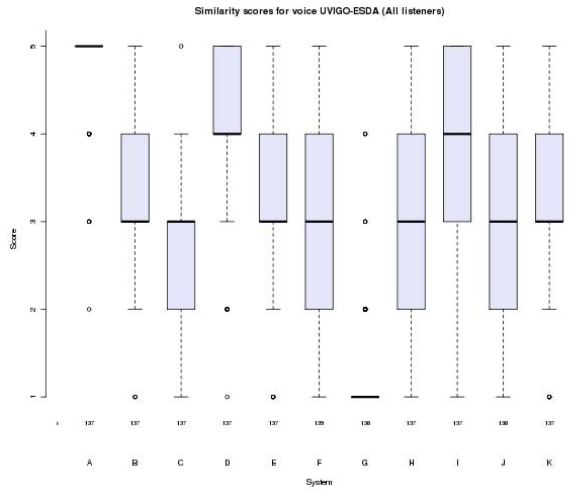

Figure 3: Test 1 results: similarity to the original voice overall scores.

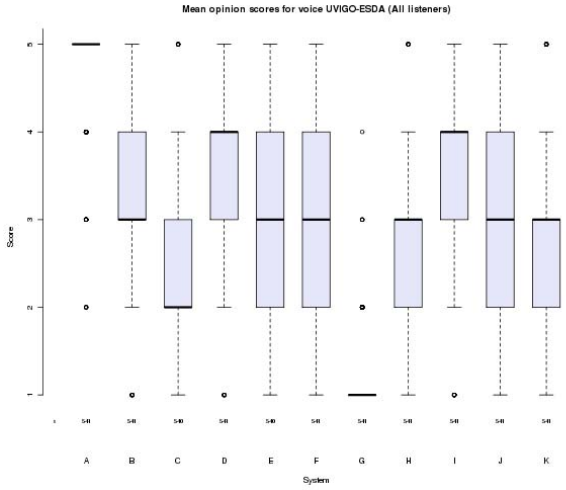

Figure 4: Test 2 results: overall Mean Opinion Scores.

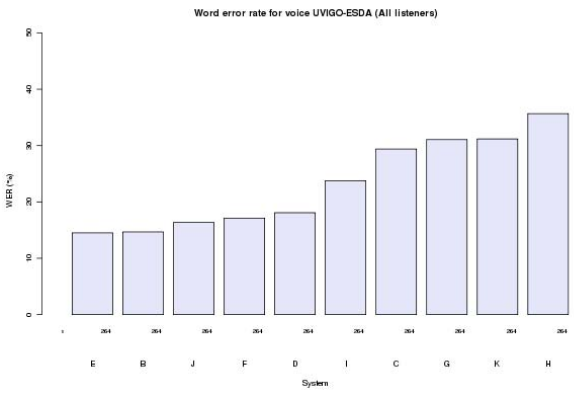

Figure 5: Test 3 results (SUS): overall WER.

The overall results show that our system was rated with 1 in a 5 points scale in Test 1 and Test 2. Regarding the SUS test, Microsoft's system had 30% of WER, which ranked it in the 8th position out of 10 systems presented in the challenge. These results were very surprising to us especially because they don't reflect our internal assessment of the released technology. For confidentiality reasons, we cannot publish those results against the competitors' systems, but we can state that our system passed Microsoft's quality bar regarding intelligibility and voice quality (MOS) after several test rounds with more than 40 native listeners each. Not to mention other performance and back-end tests, where our technology passed with very good scores. The reasons for these bad results may be the following: 1) our system is HMM-based which means it has a good intelligibility rate but not so good naturalness assessment, especially if compared against unit selection based systems; 2) in our system, voice quality improves dramatically when trained with at least 5000 sentences. In this challenge, only 1217 utterances were provided, which explains the bad results: 3) the listeners' profiles are different: we used more listeners, balanced in gender and explicitly with no experience with speech technology, whereas in Albayzin listeners were in less number and preferably experts in speech technology.

5. Conclusions and future work

When leveraging Microsoft Helena for Exchange 2010, we had very good MOS-scale results (which cannot be published for confidentiality reasons), especially regarding intelligibility of synthetic speech. The success rates of our system's intelligibility, when compared with other available systems, may be explained mainly by two reasons: the HTS technology enabled in the back-end, which largely increases intelligibility by making the segmental phone transitions smoother, and the application of several rule-based modules in the front-end, which allows a better accuracy rate in the grapheme-phoneme conversion. Another key aspect that has a significant impact in the synthetic speech voice quality is the choice of the voice talent or the professional speaker. In Speech at Microsoft, we have an accurate process to select the voice talent and to control voice quality with objective and acoustic measurements, which will be published soon. Preliminary work on this topic can be found here: [15], [16]. Albayzin results presented here are not consistent with our internal results and this difference will be investigated. The reasons for the bad results of our system in Albayzin may be related with the fact that most of the systems presented in the challenge are unit selection based, our systems requires more training data and listeners' profiles are different from the ones we used in our internal assessment. More work on naturalness and expressiveness of synthetic speech is ongoing.

6. References

[1] Francisco Moreno Fernández; Jaime Otero Roth. *Demografía de la lengua española.* Instituto Complutense de Estudios Internacionales y Fundación Telefónica. 2006. Spain.

[2] https://www.cia.gov/library/publications/the-world-factbook/fields/2098.html last visited on: 15/10/10

[3] http://es.wikipedia.org/wiki/Anexo:Hablantes_de_espa%C3%B1ol_como_lengua_materna_en_el_2003_(seg%C3%BAn_el_Britannica_Book) last visited on: 15/10/10

[4] Eduardo R. Banga; Francisco Méndez; Francisco Campillo; Gonzalo Iglesias; Laura Docío. *Descripción del sintetizador de voz Cotovía para la evaluación Albayzin 2008.* Grupo de Teoría de la Señal. Universidad de Vigo. 2008. Spain.

[5] Antonio Bonafonte; Ignasi Esquerra; Albert Febrer; José A. R. Fonollosa; Francesc Vallverdú. *The UPC Text-to-Speech System for Spanish and Catalan.* Universitat Politècnica de Catalunya. 1998. Spain.

[6] http://dihana.cps.unizar.es/investigacion/voz/ctv.html last visited on: 15/10/10

[7] R. Barra-Chicote; J. Yamagishi; J. M. Montero; S. King; S. Lufti; J. Macias-Guarasa. *Generación de una voz sintética en castellano basada en HSMM para la evaluación Albayzín 2008: Conversión texto a voz.* Universidad Politécnica de Madrid, University of Edinburgh, Universidad de Alcalá. 2008. Spain.

[8] http://www.cepstral.com/ last visited on: 15/10/10

[9] http://www2.research.att.com/~ttsweb/tts/demo.php last visited on: 15/10/10

[10] http://tts.loquendo.com/ttsdemo/default.asp last visited on: 15/10/10

[11] http://www.verbio.com/webverbio3/html/demos_ttsonline.php last visited on: 15/10/10

[12] http://www.scansoft.com/realspeak/languages/ last visited on: 15/10/10

[13] http://www.acapela-group.com/text-to-speech-interactive-demo.html last visited on: 15/10/10

[14] Microsoft Language Development Center website – Downloads section (http://www.microsoft.com/portugal/mldc/downloads.mspx)

[15] Daniela Braga, Luis Coelho, Fernando Gil Resende Jr., and Miguel Dias. 2007. "Subjective and Objective Assessment of TTS Voice Font Quality", XII International Conference Speech and Computer - SPECOM 2007, October 15-18, 2007, Moscow, Russia. ISBN 6-7452-0110-x

[16] Daniela Braga, Luis Coelho, Fernando Gil Resende Jr., and Miguel Dias. 2007. "Subjective and Objective Evaluation of Brazilian Portuguese TTS Voice Font Quality", Kacic, Zdravko and Markus, Aleksandra (eds) Advances in Speech Technology – International Workshop. June 28-29, 2007. Maribor, Slovenia. ISBN: 978-961-248-105-6. pp. 129-138.

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