Albayzín 2010: a Spanish text to speech evaluation

Francisco Campillo[1], *Francisco Méndez*[1], *Montserrat Arza*[2], *Laura Docío*[1], *Antonio Bonafonte*[3], *Eva Navas*[4], *Iñaki Sainz*[4]

[1]Group on Multimedia Technologies, Universidade de Vigo, Spain
[2]Instituto Ramón Piñeiro, Santiago de Compostela, Spain
[3]TALP Research Center, Universitat Politècnica de Catalunya, Spain
[4]Aholab Signal Processing Laboratory, University of the Basque Country, Spain

{campillo,fmendez,marza,ldocio}@gts.uvigo.es,
antonio.bonafonte@upc.edu,{eva,inaki}@aholab.ehu.es

Abstract

Albayzín 2010 Text-to-Speech Evaluation Campaign was the second biannual Albayzín Campaign. A Spanish corpus was provided by the Group of Multimedia Technologies of the University of Vigo, and six teams developed a total of ten systems for the evaluation. A set of test sentences was released to be synthesized, and an on-line evaluation was conducted, focusing on naturalness, similarity to the original voice, and intelligibility. In this paper the evaluation details and results are described.

Index Terms: speech synthesis, evaluation, listening test

1. Introduction

The main goal of the Albayzín Text-to-Speech evaluations is to compare the different techniques employed by the research teams in building their TTS systems with Spanish voices. With this purpose in mind, a contest has been proposed among speech synthesis systems, similar to the *Blizzard Challenge* [1] for English and Mandarin. The participating teams had up to seven weeks to build their system from the development material that was supplied. Next, a set of test sentences was released and the teams had five days to synthesize them and send the audio files back. An evaluation was carried on-line including three different listening tests: similarity with the original voice, naturalness and intelligibility. While Albayzín 2008 [2] was organized by Aholab at the University of the Basque Country, Albayzín 2010 was organized by the Group of Multimedia Technologies of the University of Vigo, following the main principles of Albayzín 2008 and previous Blizzard Challenges. The detailed results and analysis, as well as the detailed description of the systems, can be found via the website of Fala2010 [3], the conference in which the evaluation was held. The outline of this paper is as follows: Section 2 describes the different groups that participated in Albayzín 2010, as well as the main characteristics of the submitted systems. Section 3 includes the details of the evaluation: description of the listening tests and listener groups. Sections 4 and 5 are dedicated to the corpus provided for building the systems, and the test sentences. Section 6 describes the analysis methodology. Finally, Sections 7 and 8 are the results and discussion.

2. Participants

The Albayzín 2008 Text-to-Speech Evaluation had 7 participant groups, with a total of 8 systems submitted. In this edition there were 6 groups, who submitted a total of 10 different systems to the evaluation:

- University of the Basque Country (AhoLab)
- University of Vigo, Group on Multimedia Technologies (GTM)
- Technical University of Madrid - University of Edinburgh (GTH-CSTR)
- Microsoft Language Development Center
- UPC-Barcelona Tech
- La Salle – Ramon Llull University, Group on Multimedia Technologies Research (GTM)

Table 1 summarizes the main features of the submitted systems. The system names were anonymised assigning letters from A to K, A representing natural. Note that system J was exactly the same system that participated in Albayzín 2008, obtaining the best results. Although both evaluations were carried out with different corpus and a direct comparison might be misleading, it can serve as a baseline.

3. Evaluation

3.1. Listening test design

The evaluation was carried out on-line, following the design developed for the Blizzard Challenge 2007 [4]. The listening test was open for approximately three weeks. The type of listeners that participated is summarized in table 2. Each participant was expected to provide fifteen speech experts and Spanish native speakers as listeners of the evaluation.

The test was divided into three different sections which could be completed in any order, and in several sessions if desired. Around 30 minutes were needed to complete the whole test. Table 3 summarizes the number of evaluators in each test.

Speech Technology Expert	Yes	64
	No	83
Spanish Native	Yes	134
	No	13
Listening equipment	Headphones	119
	Loudspeakers	28
Gender	Male	98
	Female	49

Table 2: *Information about registered listeners.*

System	Type	Basic unit	f0 modeling	Signal modification	Remarks
B	Statistical parametric	Quinphone	HSMM context dependent model	Vocoder postfiltering	Re-alignment iterations
C	Statistical parametric	Quinphone	HSMM context dependent model	Vocoder	Uses adaptation
D	Concatenative hybrid	Demiphone	Unit selection + MSD-HSMMs	TD-PSOLA	30' manual revision
E	Statistical parametric	Quinphone	Context dependent MSD-HSMMs	Straight	
F	Statistical parametric	Context dependent HMMs	Context dependent MSD-HSMMs	HNM-based vocoder	30' manual revision
G	HTS	N/A	N/A	N/A	Buggy system
H	Unit selection	Demiphone	Stylized contours of stress groups	f0 and duration	Extra data for prosodic model
I	Concatenative	Demiphone	Unit selection	When needed	Joined acoustic/intonation selection
J	Statistical parametric	Quinphone	HSMM context dependent model	Vocoder	2008's best system
K	Unit selection	Diphone	CBR	PSOLA	Perceptual weight-tuning

Table 1: *System descriptions.*

	Sect. 1	Sect. 2	Sect. 3	Total
Completed	137	135	132	132
Partially completed	3	1	0	8
No response at all	7	11	15	7
Total registered	147			

Table 3: *Number of listeners.*

The next sections describe the different tests conducted in the evaluation.

3.1.1. Section 1 - Similarity to the original voice

In each part listeners could play four fixed reference samples of the original voice talent and one synthetic sample. Then, they had to score the synthetic sample taking only into account how similar to the original voice it was, on a scale ranging from 1 *[Sounds like a totally different person]* to 5 *[Sounds like exactly the same person]*. In this section each evaluator had to listen to 11 audio files, one from each system and another one from the original recording.

3.1.2. Section 2 - Naturalness MOS (Mean Opinion Scores)

In each part evaluators listened to one sample and chose a score which represented how natural or unnatural the sentence looked like on a scale between 1 *[Completely Unnatural]* and 5 *[Completely Natural]*. This was the main section of the listening test, and each evaluator had to listen to a total of 44 audio files, 4 for each participant plus 4 natural ones.

3.1.3. Section 3 - Semantically Unpredictable Sentences (SUS)

Semantically unpredictable sentences were designed to test the intelligibility of synthetic speech. Evaluators listened to one utterance in each part and typed in what they heard. In this part each listener had to transcribe two sentences from each system, 20 in total (no original voice in this section, since there were no recordings of these SUS sentences in the Uvigo_esda database).

3.2. Listener groups

Following the Blizzard 2007 design [4], listeners were grouped in 11 (10 submitted systems plus original voice) groups using a Latin square strategy. For each listener group, each section of the test had a different system ordering: all evaluators listened the same 75 sentences in the same order, but synthesized by different systems.

4. Corpus description

The Uvigo_esda Database was provided by the Group on Multimedia Technologies of the University of Vigo. It contains speech recordings from an amateur male speaker that read prompted texts in "neutral" style, to an extent of approximately 2 hours of speech. Collection was performed at the recording studio of the Signal Theory Group of the University of Vigo. It consists of 1217 phonetically balanced sentences, automatically extracted from journalistic texts by means of a greedy algorithm.

The following information was provided for each utterance:

- Audio data: sampled at 16 kHz, with 16 bits resolution. The original recordings, at 44.1 kHz, were also provided.
- Phone segmentation: phone labels were automatically extracted using the segmentation tools of HTK. SAMPA code is used for phones.
- Pitch marks files, extracted with Praat.
- Prompt text. Silences are always marked with commas or periods, and aligned with the segmentation files.
- Prompt text with information about intonation boundaries: intonation group boundaries ($\#R - E\#$) and intonation boundaries related to commas that the speaker did not realize as silence in the recording ($\#R - C\#$).
- Lexicon table derived from the texts, including all the words in the corpus and the different pronunciations.

Participants were asked to build a synthetic voice from this database. There was freedom to choose the number of sentences to be included, and to use other techniques as voice adaptation. No manual intervention was allowed during synthesis (neither prompt sculpting, nor using different subsets of the database for different test sentences or sentence types unless this is a fully automatic part of the system).

5. Test sentences

There were two different sets of test sentences:

- 350 held-out phonetically balanced sentences from the Spanish corpus Uvigo_esda, automatically extracted and belonging to four different broad types: declarative, interrogative, exclamatory and suspensive.
- 82 semantically unpredictable sentences, manually designed for the intelligibility test. These sentences are seven words long, all with the same morphosyntactic structure: DETERMINER + NOUN + ADJECTIVE + VERB + DETERMINER + NOUN + ADJECTIVE.

Participants had five days to synthesize all 432 sentences and send back the synthetic audio files.

The sentences actually used in the on-line evaluation were a subset of the ones described before. From the original 350 sentences, 55 sentences were randomly chosen to meet the following criteria:

- Interrogative: 2 sentences in section 1 and 6 in section 2.

- *Short* sentences (5-7 words): 2 sentences in section 1 and 6 in section 2.
- *Long* sentences (20-24 words): 2 sentences in section 1 and 6 in section 2.
- *Normal* sentences (10-15 words): 5 sentences in section 1 and 26 in section 2.
- Neither exclamatory nor suspensive sentences were selected.
- Sentences with foreign words or words which had to be text normalized were excluded.

SUS sentences for the listening test were selected manually.

6. Analysis methodology

A complete statistical analysis was made using the programs and scripts available from the Blizzard Challenge [5]. For each section in the listener test, a summary table with descriptive statistics is presented: median, median absolute deviation (mad), mean, standard deviation (sd), number of samples (n) and number of missing samples (na).

In these tables, systems are sorted in descending order of the mean scores in sections 1 and 2. This order cannot be interpreted as a ranking. It is intended only for readability, as it is more appropriate to compare medians than means in this MOS Likert-type scale (see [6]).

For word error rates, it makes sense to compare means, so ordering is in ascending order of the means.

To determine whether there are significant statistical differences between the MOS scores of systems a series of Bonferroni-corrected pairwise Wilcoxon signed rank significance tests with $\alpha = 0.01$ are used. It is represented with a symmetrical matrix in which significant differences between two systems are represented with a "•", while blank spaces denote no significant statistical differences.

For section 3, word error rates (WER) are calculated automatically, using the same methodology and scripts as in the Blizzard Challenge [4][5]. Capitalization and written accent marks were ignored, and certain common orthographic errors, such as confusion between *b* and *v* or erroneous misuse of *h*, were allowed. Also allowance was made for certain spelling variations in listener responses. Compounding or splitting words (e.g. *chico leo* instead of the correct word *chicoleo*) are also handled.

7. Results

In this section the results obtained in each part of the listening test are presented. Unless notated, figures and tables display the results for all listener types combined, including those listeners who completed only partially any section of the test.

7.1. Section 1: Similarity to the original speaker

Results for section one are presented in tables 4 and 5. As expected, natural speech obtained the highest score, with a median of 5. Two systems, D and I, perform significantly better than the average, scoring a median of 4. Although comparing to the results obtained in the last Albayzín 2008 TTS Evaluation [2] (best system with a median of 3) is not completely fair since a very different speech database was used, these two systems show a great progress, although there is a lot of room for improvement yet. Then, a group of 6 systems (B, H, K, E, J, F) with a median of 3, shows no significant statistical differences.

	median	MAD	mean	sd	n	na
A	5	0.00	4.83	0.49	137	10
D	4	1.48	4.07	0.94	137	10
I	4	1.48	4.02	0.94	137	10
B	3	1.48	3.34	0.94	137	10
H	3	1.48	3.23	1.14	137	10
K	3	1.48	3.20	0.99	137	10
E	3	1.48	3.15	0.97	137	10
J	3	1.48	3.13	1.11	138	9
F	3	1.48	2.91	1.12	139	8
C	3	1.48	2.54	0.96	137	10
G	1	0.00	1.25	0.60	138	9

Table 4: *Similarity scores for all listeners.*

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•		•	•		•		
C	•	•		•	•		•	•	•	•	•
D	•	•	•		•	•	•	•		•	•
E	•		•	•			•		•		
F	•	•		•			•		•		
G	•	•	•	•	•	•		•	•	•	•
H	•		•	•			•		•		
I	•	•	•		•	•	•	•		•	•
J	•		•	•			•		•		
K	•		•	•			•		•		

Table 5: *Wilcoxon test: Similarity scores for all listeners.*

System C also scored a median of 3, but with a somewhat lesser mean (2.54). Note that system J was the best system in 2008, and in 2010 is the seventh one. Finally, G was the worst performing system in this section, with a median of 1 which states as *Sounds like a totally different person*. As a conclusion, this test shows that concatenative systems (D, I, H, K) can deliver speech more similar to the original than the statistical parametric synthesizers.

7.2. Section 2: Naturalness

Tables 6 and 7 show the results for section 2. Again, only original speech achieved a median of 5. Then, two systems, D and I scored a median of 4 (*mostly natural* voice). These two systems again outperformed the best system in the last Albayzín 2008 TTS Evaluation, which in this case ranked sixth with a median of 3. The rest of participants obtained a median of 3, except system C (median of 2) and finally system G was again the least scored system with a median of 1 (*Completely Unnatural* voice). By inspecting the significant statistical differences shown by Wilcoxon test (table 7), two groupings can be established: systems E and F (MOS of 3.15 and 3.10) and another group of three systems, K, H and C, with means around 2.50-2.60. The rest of positions are well defined.

7.3. Section 3: Word error rates

The results for the intelligibility test are displayed in figure 1. They show only the responses of native listeners (122 out of 132 who completed this section) due to the unusual and difficult kind of words that were used in the SUS sentences design. In fact, by inspecting the results obtained by non-native listeners, much higher error rates are observed, ranging from 24 to 49%, showing no significant differences between any system, thus suggesting that their results might not be taken into account. Therefore, system E was the best system here, scoring a

	median	MAD	mean	sd	n	na
A	5	0.00	4.75	0.57	541	47
D	4	1.48	3.78	0.98	541	47
I	4	1.48	3.50	1.02	541	47
B	3	1.48	3.33	0.95	541	47
F	3	1.48	3.15	1.00	541	47
E	3	1.48	3.10	0.96	540	48
J	3	1.48	2.91	1.00	541	47
K	3	1.48	2.62	0.98	541	47
H	3	1.48	2.60	0.97	541	47
C	2	1.48	2.51	0.90	540	48
G	1	0.00	1.10	0.34	541	47

Table 6: *Mean opinion scores for all listeners.*

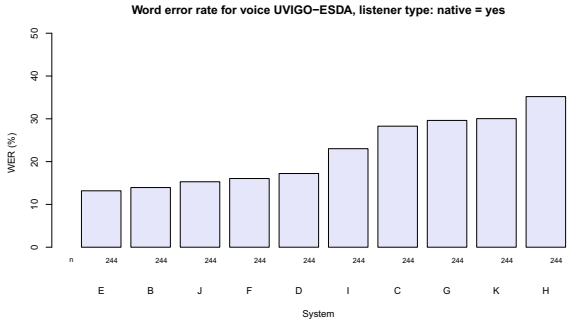

Figure 1: *Word error rates for native listeners.*

WER of 13%, although there are no significant statistical differences with systems B, D, F and J, all of them with word error rates ranging between 14 and 17%. System I achieved a WER of 23% and finally C, G, H and K got around 30%.

These figures clearly state that the SUS sentences used in the intelligibility task this year were considerably more difficult than 2008's, when error rates were around 5%. As far as no original recording of the SUS sentences was available, no comparison against the natural voice was done, so it is difficult to discern if these bad results are due to the difficult nature of the test sentences or a consequence of poor readings by the submitted systems.

8. Discussion

Comparing with the previous Albayzín evaluation, in 2008 there were 7 concatenative systems and only 1 HMM-based, while in 2010 there were 6 HMM-based, 3 purely concatenative and a hybrid one, which seems to follow the trend change in technology of the last years, being HMM-based systems a very promising approach. Regarding performance, however, although in 2008 the winner was a statistical parametric system, this year concatenative systems D and I got better results both in MOS and similarity, with reasonable results in intelligibility too. Note that although system D is a hybrid system, HMM's are only used for prosodic and spectral estimation, and waveform generation is purely concatenative.

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•	•	•	•	•	•	•	•
C	•	•		•	•	•	•		•	•	
D	•	•	•		•	•	•	•	•	•	•
E	•	•	•	•			•	•	•	•	•
F	•	•	•	•			•	•	•	•	•
G	•	•	•	•	•	•		•	•	•	•
H	•	•		•	•	•	•		•	•	
I	•	•	•	•	•	•	•	•		•	•
J	•	•	•	•	•	•	•	•	•		•
K	•	•		•	•	•	•		•	•	

Table 7: *Wilcoxon test: Mean opinion scores for all listeners.*

The general main characteristics of current main trends are also reflected in the results. While concatenative systems (D, I) tend to get better scores in naturalness and similarity with the original voice, HMM-based systems (E, B, J) tend to get better results in intelligibility.

The corpus provided for the evaluation included information about intonation boundaries, and although there was not a task explicitly designed to check out the relevance of this information, some groups did their own test and found it very valuable. Perhaps in a future challenge a task related to this topic should be included.

Although comparing performances of systems in different tests, and with different voices, might be misleading, it can give an idea of the evolution of the systems. With respect to similarity with the original voice, results clearly outperform those obtained in Albayzín 2008. Moreover, there are two systems (D and I) that achieve a median of 4. About naturalness, these same two systems improve the results of Albayzín 2008, but both of them are still far from 5, which shows that there is still a lot of room for improvement. With regards to intelligibility, results were definitely worse than 2008, but the authors consider it to be a consequence of the test being much more difficult.

Finally, the bad results of system G were a consequence of a bug during the waveform generation. The developers noticed it when the perceptual test was already being conducted, and the mistake could not be corrected.

9. Acknowledgments

The Albayzín 2010 TTS Evaluation organizing committee would like to thank the organizing committee of the Albayzín TTS 2008 and the Blizzard Challenge, whose work and tools has served as a model for the Albayzín 2010 TTS Evaluation. Finally, thanks to all participants and volunteer listeners who completed the on-line evaluation.

The work reported here was partially supported by the Spanish government and ERDF funds under the project TEC2009-14094-C04-04 "Búsqueda de Información en Contenidos Audiovisuales plurilingües", Xunta de Galicia "Isidro Parga Pondal" research programme and *Centro Ramón Piñeiro para a Investigación en Humanidades*.

10. References

[1] Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2010", in Proc. Blizzard Challenge workshop 2010.

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

[3] "Fala2010 website", http://fala2010.uvigo.es/

[4] Mark Fraser and Simon King, "The Blizzard Challenge 2007" , In Proc. Blizzard Workshop (in Proc. SSW6), 2007

[5] Online http://www.cstr.ed.ac.uk/projects/blizzard/tools.html

[6] R.A.J. Clark, M. Posiadlo, M. Fraser, C. Mayo and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results" , In Proc. Blizzard Workshop (in Proc. SSW6), 2007