Albayzín 2010 Evaluations: TTS

F. Méndez, Montserrat Arza, Laura Docío, F. Campillo

November 10, 2010

Albayzín 2010 TTS: Overview

- Objectives
- Timeline
- Participants
- Speech Database
- Test Sentences
- Listening Test
- Results
- Conclusions

Albayzín 2010 TTS: Objectives

- Compare the different techniques employed by the research teams in building their TTS systems with a common Spanish database

Albayzín 2010 TTS: Timeline

May 18th Registration form available in the FALA2010 web.

June 14th Development material released

July 15th Registration deadline

September 6th Test sentences released

September 10th Deadline for submitting synthesized speech

September 20th – October 10th Evaluation campaign

October 27th Notification of the evaluation results to each of the participants

November 10th Presentation of results in the FALA 2010 Workshop

Albayzín 2010 TTS: Participants

AhoHTS University of the Basque Country (AhoLab)

AhoTTS University of the Basque Country (AhoLab)

Cotovía University of Vigo, Group on Multimedia Technologies (GTM)

Cotovía–hts University of Vigo, Group on Multimedia Technologies (GTM)

GTHCSTR–2008 Technical University of Madrid - University of Edinburgh (GTH-CSTR)

GTHCSTR–2010 Technical University of Madrid - University of Edinburgh (GTH-CSTR)

GTHCSTR–2010–adaptation Technical University of Madrid - University of Edinburgh (GTH-CSTR)

MS–HTS–TTS Microsoft Language Development Center

Ogmios UPC–Barcelona Tech

SalleTTS La Salle – Ramon Llull University, Group on Multimedia Technologies Research (GTM)

Albayzín 2010 TTS: Speech Database

- Amateur male speaker, neutral style, Spanish language. Recorded at University of Vigo.
- Two hours, 1217 phonetically balanced sentences. Journalistic texts.
- 44 and 16 KHz waveforms
- Phone segmentation
- Pitch marks
- Intonation boundaries
- Lexicon

Albayzín 2010 TTS: Test sentences

- 350 held-out phonetically balanced sentences from the database.
- 82 semantically unpredictable sentences. 7 word long, with structure DET + NOUN + ADJ + VERB + DET + NOUN + ADJ
- Participants had up to 5 days to synthesize the 432 test sentences.
- 75 sentences were randomly selected for the listening test.

Albayzín 2010 TTS: Listening test

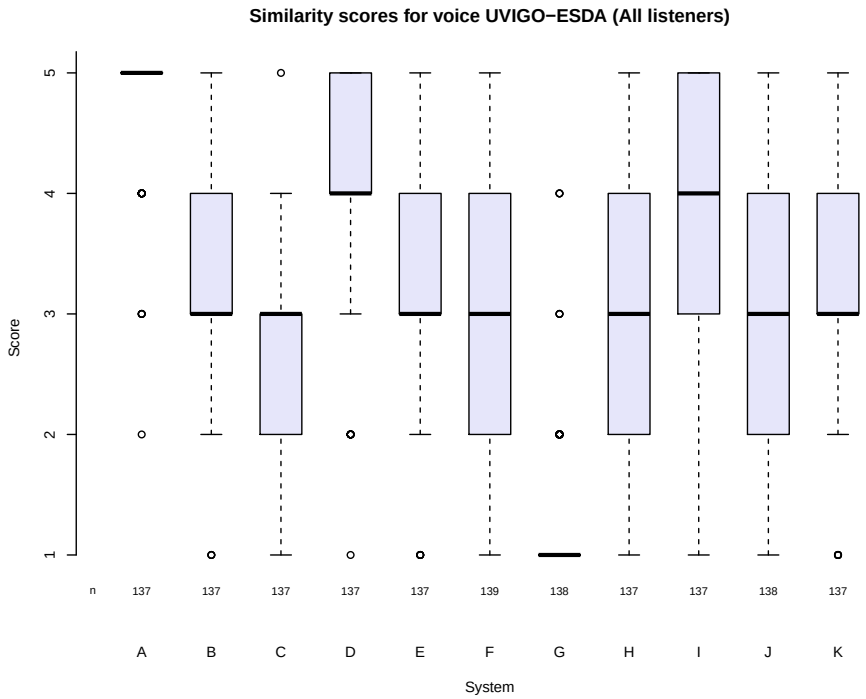
- Online Evaluation Campaign: opened 3 weeks
- Three sections:
  - Similarity with the original voice: 11 sentences
  - Naturalness Mean Opinion Scores: 44 sentences
  - Intelligibility: 20 Semantically Unpredictable Sentences
- About 30 minutes long

Albayzín 2010 TTS: Listeners

	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			

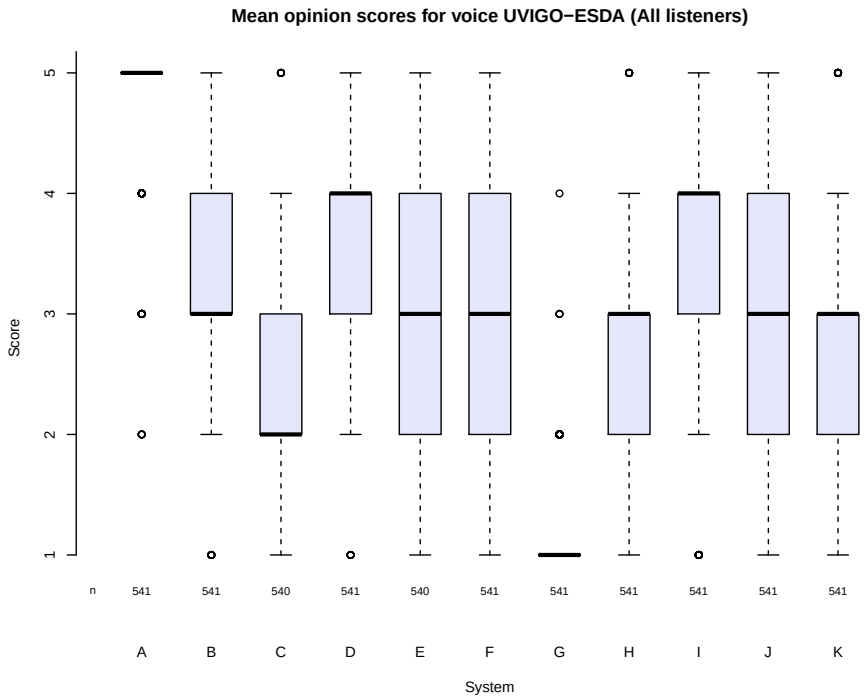
Speech Technology Expert	Yes	64
	No	83
Spanish Native	Yes	134
	No	13
Hearing equipment	Headphones	119
	Loudspeakers	28
Gender	Male	98
	Female	49

Albayzín 2010 TTS: Results - Similarity



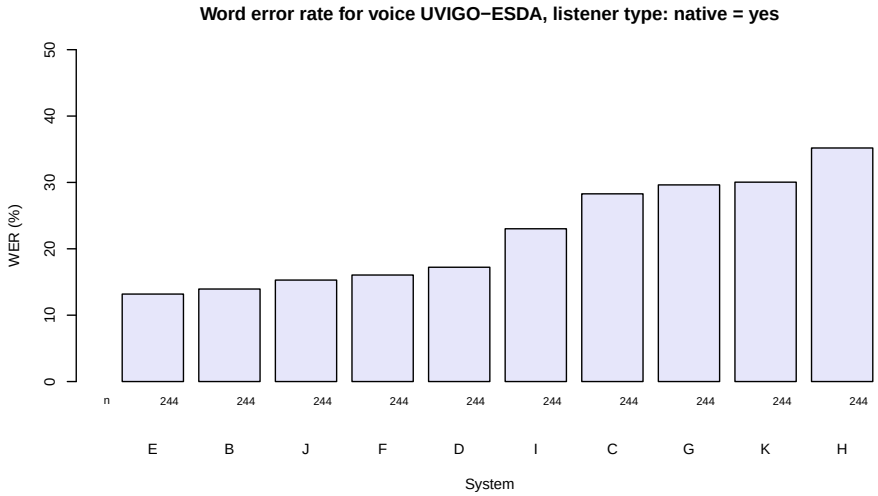
	median	MAD	mean	sd	n	na
A	5	0.00	4.83	0.49	137	10
D	4	1.48	4.07	0.94	137	10
I	4	1.48	4.02	0.94	137	10
B	3	1.48	3.34	0.94	137	10
H	3	1.48	3.23	1.14	137	10
K	3	1.48	3.20	0.99	137	10
E	3	1.48	3.15	0.97	137	10
J	3	1.48	3.13	1.11	138	9
F	3	1.48	2.91	1.12	139	8
C	3	1.48	2.54	0.96	137	10
G	1	0.00	1.25	0.60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

Albayzín 2010 TTS: Results - Naturalness



	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

Albayzín 2010 TTS: Results - Intelligibility



	median	MAD	mean	sd	n	na
E	0.00	0.00	0.13	0.17	244	24
B	0.00	0.00	0.14	0.20	244	24
J	0.14	0.21	0.15	0.21	244	24
F	0.14	0.21	0.16	0.19	244	24
D	0.14	0.21	0.17	0.20	244	24
I	0.14	0.21	0.23	0.22	244	24
C	0.29	0.42	0.28	0.25	244	24
G	0.29	0.42	0.30	0.26	244	24
K	0.29	0.21	0.30	0.26	244	24
H	0.29	0.42	0.35	0.27	244	24

Albayzín 2010 TTS: Conclusions

- Trend change in TTS system's technologies:
  - 3 Concatenative
  - 6 HMM-based
  - 1 hybrid
- Improvements in similarity with original and naturalness, achieving medians of 4
- Worse results in intelligibility due to the difficulty of the 2010 SUS test
- In general, concatenative systems got better scores in naturalness and similarity with the original voice, HMM-based systems tend to get better results in intelligibility

Albayzín 2010 TTS: Conclusion

Thank you very much to all participants and volunteer evaluators!