# TTS evaluation campaign with a common Spanish database

**Iñaki Sainz[a], Eva Navas[a], Inma Hernáez[a], Antonio Bonafonte[b], Francisco Campillo[c]**
[a]Aholab Signal Processing Laboratory, University of the Basque Country, Spain
[b]TALP Research Center, Universitat Politècnica de Catalunya, Spain
[c]Signal & Communication Department, University of Vigo, Spain
E-mail: inaki@aholab.ehu, eva@aholab.ehu.es, inma@aholab.ehu.es, antonio.bonafonte@upc.edu, campillo@gts.tsc.uvigo.es

**Abstract**

This paper describes the first TTS evaluation campaign designed for Spanish. Seven research institutions took part in the evaluation campaign and developed a voice from a common speech database provided by the organisation. Each participating team had a period of seven weeks to generate a voice. Next, a set of sentences were released and each team had to synthesise them within a week period. Finally, some of the synthesised test audio files were subjectively evaluated via an online test according to the following criteria: similarity to the original voice, naturalness and intelligibility. Box-plots, Wilcoxon tests and WER have been generated in order to analyse the results. Two main conclusions can be drawn: On the one hand, there is considerable margin for improvement to reach the quality level of the natural voice. On the other hand, two systems get significantly better results than the rest: one is based on statistical parametric synthesis and the other one is a concatenative system that makes use of a sinusoidal model to modify both prosody and smooth spectral joints. Therefore, it seems that some kind of spectral control is needed when building voices with a medium size database for unrestricted domains.

## 1. Introduction

Subjective tests are essential for the quality assessment of text-to-speech (TTS) synthesisers. Objective measures can be employed to evaluate the quality of text processing or prosody prediction modules, but fail to address the whole human hearing process. That is because of its complex and multidimensional nature. Therefore, designing large and time-consuming perceptual evaluation campaigns is still necessary in order to assess the performance of new techniques.

The main purpose of the Albayzin TTS evaluation campaign is to compare the various techniques and implementations used by different systems that were developed with a common speech database, and to favour the collaboration between the research teams involved. Its design is based on the Blizzard Challenge (Black & Tokuda, 2005) international evaluation. While the later evaluates TTS systems for both English and Mandarin languages, Albayzin is focused only on Spanish.

Each participating team had a period of seven weeks to generate a voice from the development material provided, which included a 105 minutes long speech database recorded by a female voice talent (Bonafonte & Moreno, 2008) and the corresponding phone segmentation labels (almost half of which were hand-corrected). Next, a set of sentences were released and each team had to synthesise them and sent the audio files back within a week period. Finally, a subjective web-based evaluation campaign was deployed. Synthesised audio files were subjectively evaluated under the following criteria: Similarity to the original voice, Naturalness and Intelligibility.

This paper is organised as follows. Section 2 presents the participants that took part in Albayzin TTS evaluation and gives a brief description of the main characteristics of each system. Section 3 describes how the test was designed and deployed. A summary of test results and analysis is presented in Section 4. And finally, some conclusions and suggestions for next evaluations are drawn in Section 5.

## 2. Participants

Seven different institutions took part in the Albayzin TTS evaluation campaign. They are listed in Table 1. One of the participants submitted two different systems, so there was a total of eight participating systems. As all of them had native Spanish speakers in their voice development team, everybody competed in completely equal terms.

| |
|---|
| Barcelona Media Center & Cereproc Research |
| Group on Multimodal Processing (Univ. Ramon Llull) |
| Text to speech conversion group of Telefónica R & D |
| Multimedia Technologies Group (Univ. Vigo) |
| BDSM Madrid (UPM, Univ. Edinburgh, Alcalá Univ.) |
| Aholab (Univ. Basque Country) |
| TALP Research Center (UPC) |

Table 1: Institutions that participated in the evaluation.

### 2.1 System Description

All the systems were based on the unit selection concatenative approach (Hunt & Black, 1996) except one that opted for statistical parametric speech based on HTS framework (Zen et al., 2006). In order to preserve the anonymity, each participating system was assigned a letter that identified it (A to H). Letter I referred to natural speech signals. In Table 2 a brief description of each system is provided, including the following fields: type of TTS, basic unit, manual revision of labels, pitch modelling and the type of signal modification applied (if any).

| .System | Type | Basic unit | Manual Revision | Pitch Modelling | Signal Modifications |
|---|---|---|---|---|---|
| A | Concatenative | Diphone | No revision | No pitch model | No modifications |
| B | Concatenative | Semiphone | 70 man-hour | Unit selection | Pitch and duration |
| C | Statistical parametric | Pentaphone | 0.5 man-hour | HMM | Vocoder |
| D | Concatenative | Diphone | No revision | Generic pitch model | No modifications |
| E | Concatenative | Diphone | 60 man-hour | Unit selection | Sinusoidal Model |
| F | Concatenative | Semiphone | No revision | CART | Pitch and duration |
| G | Concatenative | Diphone | 60 man-hour | Unit selection | Pitch and duration |
| H | Concatenative | Semiphone | No revision | Unit selection | Pitch and duration |

Table 2: Short description of each system.

## 3. Evaluation

In order to evaluate the quality of each of the participating systems, a subjective evaluation was conducted. Listeners were recruited by the participating research groups. Each team had to provide a minimum of five evaluators. All the listeners were volunteers.

### 3.1 Test Sentences

Once the voice development period had finished, each participant received a set of 350 text sentences. The set was extracted from two domains: novels and news. A certain degree of phonetic balance was achieved during the texts' selection by means of a greedy algorithm (Sesma & Moreno, 2000).

To evaluate the intelligibility of the synthetic signals, 25 SUS (Semantically Unpredictable Sentences) were also included. Due to lack of a robust POS (Part Of Speech) generator, they were manually generated using five structures proposed in (Grice, 1989) and shown in Table 3.

| |
|---|
| DET + NOUN + $VERB_{Intrans}$ + PREP + DET + NOUN |
| DET + ADJ + NOUN +$VERB_{Trans}$ + DET + NOUN |
| ADV + $VERB_{Trans}$ + DET + NOUN + CONJ + DET + NOUN |
| Q-ADV + DET + NOUN + $VERB_{Trans}$ + DET + ADJ + NOUN |
| DET + NOUN + $VERB_{Trans}$ + DET + REL PRON + $VERB_{Intras}$ |

Table 3: SUS structures employed.

### 3.2 Test Design

The test consisted of three sections and although it was designed to be completed in a unique session (about 30 minutes), it was allowed to interrupt the session at the end of each section.

#### 3.2.1. 1st Section: Similarity to the Original Voice

Each evaluator had to listen to three natural voice recordings to become familiar with the original voice. After doing so, the evaluators had to listen to a total of nine audio files (one for each participant system and one natural recording) and give each of them a value on an MOS (Mean Opinion Score) scale. The score ranged from 1 (*completely different voices*) to 5 (*exactly the same voice*).

#### 3.2.2. 2nd Section: Naturalness

Up to 42 signals (5 for each participant system + 2 original recordings) were evaluated with respect to their naturalness. A scale that ranged from 1 (*the voice is completely unnatural*) to 5 (*the voice is completely natural*) was employed. Naturalness was asked to be rated globally, without special consideration of (for example) 'naturalness of the intonation'. As such, discontinuities at concatenation points or other usual noises in synthetic signals have probably contributed towards unnaturalness.

#### 3.2.3. 3rd Section: Intelligibility

Each evaluator listened to 16 signals (2 for each system) and was asked to type what he/she had understood. They were warned that the sentences might not make any sense at all and were requested to restrict the number of listenings per sentence to two. As a measure of the intelligibility, the WER (Word Error Rate) was computed. It must be clarified that no natural speech was included in this section because there were no SUS sentences available for the original voice.

### 3.3 Listener Groups

During the test each subject evaluated a total of 67 signals. In order to minimize possible ordering effects during the presentation of the signals, a Latin Square strategy (Penfold & Street, 1987) was adopted. To do so, groups of listeners were set for each section: as many groups as participants (plus natural voice if necessary). Thus, all the evaluators listened to the same sentences and in the same order, but synthesized with different systems (or natural voice).

### 3.4 Listener Characteristics

The test included a short questionnaire designed in order to identify the circumstances and characteristics of the evaluators. The information concerning the 103 subjects that completed the evaluation test is summarized in Table 4 (figures in number of listeners).

## 4. Results

As mentioned above, the purpose of this evaluation was to assess the similarity with the original voice, the naturalness and the intelligibility of the synthesised voices involved in the campaign. Since the naturalness section had the greatest importance in the evaluation (42 signals out of 67), the figures and tables with the results have been ordered according to the mean score of each system for that section.

| | | |
|---|---|---|
| **Equipment** | Headphones | 81 |
| | Speakers | 22 |
| **Evaluator Information** | Speech Technology Expert | 54 |
| | Non expert | 49 |
| | Native speaker | 94 |
| | Non native speaker | 9 |
| | Male | 66 |
| | Female | 37 |

Table 4: Information about the listeners.

### 4.1 Measurements

As the MOS likert-type scale (Likert, 1932) does not guarantee the interval between scores to be constant (e.g. an improvement from 1 to and 2 is not necessarily proportional to the one found from 3 to 4) it is not statistically significant to compare means among systems (Marcus-Roberts & Roberts, 1987). Therefore, it is recommended to compare medians. To make the tables and plots more readable, we maintain the ordering of the systems according to their means. But it must not be interpreted as an actual ranking.

#### 4.1.1. Box-Plots

For each of the MOS sections a box-plot like the ones shown in Figures 1 and 2 have been generated. The rectangle represents the range between first and third quartile and the whiskers extend to 1.5 times the inter-quartile range. The median (or second quartile) is represented by a horizontal red stripe. Out of range values (outliers) are represented by an "x" and the rest of the values are grouped with a solid line. Along with the box-plot a table like Tables 5 and 7 is attached, containing information on: median, mean, standard deviation, lower confidence limit for the median (LCM) and the upper confidence limit for the median (UCM).

#### 4.1.2. Wilcoxon Test

To determine whether there were statistically significant differences between the MOS of each system, pair-wise Wilcoxon signed rank tests (Wilcoxon, 1945) have been conducted with a level of significance of 0.05 and Bonferroni correction (Abdi, 2007). The results are shown in a table similar to Tables 6 and 8. It is a symmetrical matrix where statistically significant differences between two systems are represented with a 1 and a 0 indicates that there were no significant differences at all.

#### 4.1.3. Word Error Rate

In the intelligibility section the Word Error Rate (WER) is measured as indicated in Equation 1.

$$WER = (S+D+I) / N \quad (1)$$

Where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N the number of words in the reference. Written accent marks were ignored during the WER computation, since quite a few subjects did not employ them in the whole test.

### 4.2 Analysis of the Results

The results obtained in each of the three sections that formed the test are presented and discussed in the following subsections.

#### 4.2.1. Naturalness

Figure 1, Table 5 and Table 6 illustrate the MOS information for naturalness taking all listeners into account. Looking at the median values three main groups can be distinguished: natural speech (MOS of 5), systems C, E, B, F, D, G (MOS of 3) and system A (MOS of 2). There is a big gap in perceived naturalness between the synthetic voices and the natural one. Besides, system A scores considerably lower than the rest. The MOS for the rest of the systems varies between 3.34 (system C) to 2.52 (system G). The Wilcoxon test in Table 6 shows whether these differences are statistically significant or not.

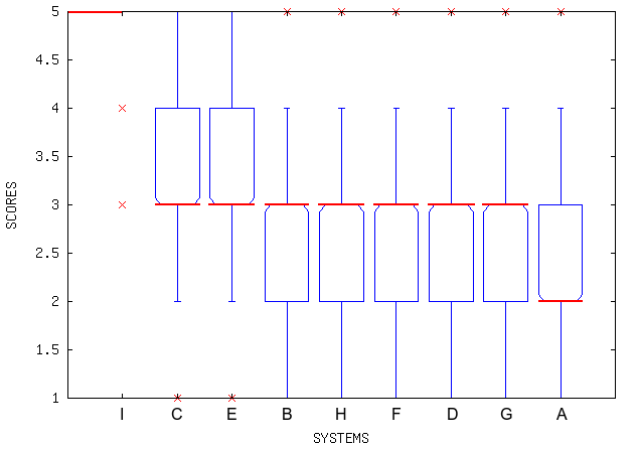


Figure 1: Box-plot for Naturalness, all listeners

Grouping the systems that have no significant differences and sorting them from higher to lower mean scores, the classification goes as follows: Systems C and E, systems B,H and F, system D, system G and system A. The two systems evaluated as the most natural sounding ones are the only ones that make some kind of signal modification of the spectrum: System E does it directly (with a sinusoidal model), and C implicitly as it generates the waveform with a vocoder. These techniques that alter the spectrum always add some kind of quality distortion, but they also generate a smoother voice. Concatenative systems tend to maintain the quality of the natural voice at segmental level, but their overall quality is not consistent (e.g. just one “bad join” can spoil the whole sentence). Apparently, the listeners have preferred the smoother and

more consistent voices.
All the system but E&C share some characteristics (i.e. concatenative systems with no spectral modifications). But there also are some structural differences among them that could explain the perceived differences. The group formed by B, H & F uses the semiphone as their basic unit, while the rest use the diphone. Although a smaller unit can potentially lead to an increase in the number of concatenations, it offers more flexibility to form diphones from non-consecutive semiphones. The system A is the only one that neither has a pitch model nor makes any signal modification. Its design seems to be more oriented to a restricted domain application and that could explain why it obtains the poorest score in the evaluation.

| Syst | Med | Mean | SD | LCM | UCM | Num |
|---|---|---|---|---|---|---|
| I | 5 | 4.82 | 0.41 | 5 | 5 | 212 |
| C | 3 | 3.34 | 0.92 | 3 | 3.06 | 524 |
| E | 3 | 3.2 | 0.89 | 3 | 3.06 | 524 |
| B | 3 | 2.91 | 0.93 | 2.93 | 3 | 524 |
| H | 3 | 2.86 | 0.96 | 2.93 | 3 | 524 |
| F | 3 | 2.81 | 0.91 | 2.93 | 3 | 524 |
| D | 3 | 2.6 | 0.94 | 2.93 | 3 | 524 |
| G | 3 | 2.56 | 0.91 | 2.93 | 3 | 524 |
| A | 2 | 2.28 | 0.94 | 2 | 2.06 | 524 |

Table 5: Naturalness statistics for all listeners.

| | I | C | E | B | H | F | D | G | A |
|---|---|---|---|---|---|---|---|---|---|
| **I** | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **C** | 1 | | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| **E** | 1 | 0 | | 1 | 1 | 1 | 1 | 1 | 1 |
| **B** | 1 | 1 | 1 | | 0 | 0 | 1 | 1 | 1 |
| **H** | 1 | 1 | 1 | 0 | | 0 | 1 | 1 | 1 |
| **F** | 1 | 1 | 1 | 0 | 0 | | 1 | 1 | 1 |
| **D** | 1 | 1 | 1 | 1 | 1 | 1 | | 1 | 1 |
| **G** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | | 1 |
| **A** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |

Table 6: Wilcoxon Test for Naturalness.

The results for the naturalness section were also calculated for the different groups of listeners (made according to their characteristics and the listening environment as shown in Table 4). The results were pretty similar and therefore are not presented here. We compared the correlation between mean scores of each system for opposite groupings and the most different results were obtained between subjects that used headphones and the ones that used loudspeakers. Even in this worst case the correlation values are pretty high (ρ=0.98 for naturalness) and the system ordering is preserved.

#### 4.2.2. Similarity to the original voice

The MOS for the similarity to the original voice is shown in Figure 2 and Tables 7 and 8. Once again, there is a clear difference between the natural voice (system I) that has a median value of 5 and the synthetic voices with a median value of 3. The Wilcoxon test shown in Table 8 proves that too, but more information can also be extracted from it. Systems B & E are significantly more similar to the original voice than system A; and System B is significantly better than system G. The pair-wise comparison among any other combination of systems shows that there are not statistically significant differences. The worst system in naturalness section maintains that position in this section too. Besides, the spectral modifications made by system C & E have apparently not degraded the similarity to the original voice.

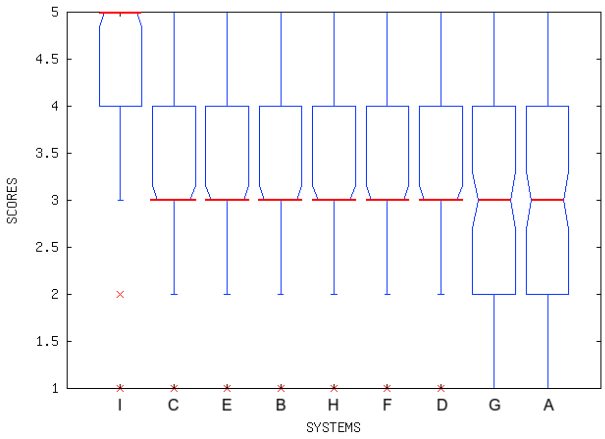


Figure 2: Box-plot for Similarity to the original voice

| Syst | Med | Mean | SD | LCM | UCM | Num |
|---|---|---|---|---|---|---|
| I | 5 | 4.11 | 1.17 | 4.84 | 5 | 107 |
| C | 3 | 3.25 | 0.84 | 3 | 3.15 | 107 |
| E | 3 | 3.35 | 0.94 | 3 | 3.15 | 107 |
| B | 3 | 3.36 | 0.93 | 3 | 3.15 | 107 |
| H | 3 | 3.29 | 0.91 | 3 | 3.15 | 107 |
| F | 3 | 3.23 | 0.81 | 3 | 3.15 | 107 |
| D | 3 | 3.23 | 0.89 | 3 | 3.15 | 107 |
| G | 3 | 3.11 | 0.92 | 2.69 | 3.30 | 107 |
| A | 3 | 2.96 | 0.98 | 2.69 | 3.30 | 107 |

Table 7: Similarity to the original voice: statistics for all listeners.

As far as the comparison among groups of listeners is concerned, the worst correlation was obtained when comparing headphones and loudspeaker groups (ρ=0.91). Therefore, the results for this section are quite independent from listeners' characteristics too.

#### 4.2.3. Intelligibility

Figure 3 shows the data concerning the WER. Only the responses of native listeners (94 out of 103) are considered as they are more reliable. In fact, there is a correlation of only ρ=0.46 between the WER calculated for native and non-native speakers in this section, and the system ordering differs. Table 9 details the types of errors involved in the WER. A two sample t-test has been

conducted for every pair of system with a level of significance of 0.05. Its results are displayed in Table 10. There are no significant differences among the four systems that get the lowest WER (C, E, B & G). For all these systems a certain degree of manual revision of labels was carried out during the voice development process. And it could be part of the cause of their good performance in this section. However, it must be noted that the system with the lowest WER (system C) had very little label revision. Its good intelligibility results might be due to the robustness of the statistical averaging in the modelling process.

| | I | C | E | B | H | F | D | G | A |
|---|---|---|---|---|---|---|---|---|---|
| **I** | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **C** | 1 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **E** | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 1 |
| **B** | 1 | 0 | 0 | | 0 | 0 | 0 | 1 | 1 |
| **H** | 1 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 |
| **F** | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 |
| **D** | 1 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 |
| **G** | 1 | 0 | 0 | 1 | 0 | 0 | 0 | | 0 |
| **A** | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | |

Table 8: Wilcoxon Test for Similarity to the original voice.

| Syst | WER (%) | Samples | Words | S | I | D |
|---|---|---|---|---|---|---|
| C | 3.49 | 188 | 1233 | 29 | 5 | 9 |
| E | 4.78 | 188 | 1234 | 40 | 4 | 15 |
| B | 4.95 | 188 | 1233 | 47 | 13 | 1 |
| H | 6.40 | 188 | 1234 | 68 | 8 | 3 |
| F | 8.19 | 188 | 1233 | 62 | 29 | 10 |
| D | 7.21 | 188 | 1235 | 61 | 20 | 8 |
| G | 3.65 | 188 | 1234 | 30 | 3 | 12 |
| A | 7.78 | 188 | 1234 | 70 | 4 | 22 |

Table 9: WER statistics.

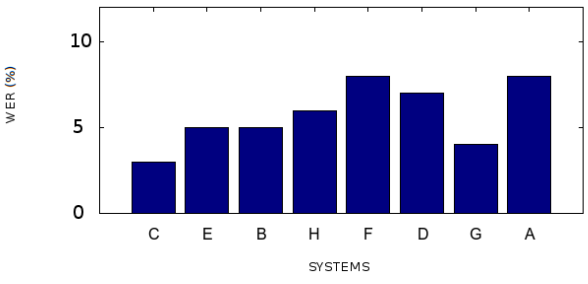


Figure 3: Word Error Rate for Native Listeners

## 5. Conclusions

Analysing the results for the naturalness two main conclusions can be drawn. On the one hand, there is considerable margin for improvement to reach the level of the natural voice (just compare the Median of 5 of the original voice, with the value of 3 for the best synthetic system in Figure 1). On the other hand, two systems (C and E) get significantly better results than the rest (see tables 5 and 6). Interestingly, both systems are the only ones that make some kind of modification of the spectrum: either directly (with a sinusoidal model), or implicitly (using statistical parametric synthesis). Therefore, it seems that for databases of small/medium size and a non limited domain of use, such kinds of systems are more suitable than the ones where signal modifications, if any, are constrained to the prosodic domain (pitch, duration and energy).

| | C | E | B | H | F | D | G | A |
|---|---|---|---|---|---|---|---|---|
| **C** | | 0 | 0 | 1 | 1 | 1 | 0 | 1 |
| **E** | 0 | | 0 | 0 | 1 | 0 | 0 | 1 |
| **B** | 0 | 0 | | 0 | 1 | 0 | 0 | 1 |
| **H** | 1 | 0 | 0 | | 0 | 0 | 1 | 0 |
| **F** | 1 | 1 | 1 | 0 | | 0 | 1 | 0 |
| **D** | 1 | 0 | 0 | 0 | 0 | | 1 | 0 |
| **G** | 0 | 0 | 0 | 1 | 1 | 1 | | 1 |
| **A** | 1 | 1 | 1 | 0 | 0 | 0 | 1 | |

Table 10 Two sample t-test for WER results.

Smoother spectral transitions could be considered to cause a considerable loss in voice quality. But the test proves that both systems (C and E) obtain good scores in the section concerning the resemblance to the original voice. In fact, there is quite a high correlation between the mean scores of the systems in different sections: $\rho$=0.74 between naturalness and similarity and $\rho$=-0.64 between intelligibility and naturalness. So, systems tended to perform in a similar way in all sections.

As far as the intelligibility is concerned, all the systems get quite a low WER. It seems that this section of the test is not as important for Spanish as it still appears to be for other languages like English (Karaiskos, V. et al., 2008). However, no clear conclusion can be set as the intelligibility of the synthetic voices was not tested against the natural voice.

The feedback received from the listeners suggested to include a preference test section and longer paragraphs to evaluate prosody. We believe that in next evaluations, naturalness should be measured with at least three scales: overall naturalness, prosody and segmental quality. That way, more specific conclusions about the failures of each system could be drawn, without requiring too much effort from the evaluators. As the evaluation lasts longer the subjects tend to lose concentration and their answers become less reliable. Therefore, during the test design a balance among those variables must be established.

## 6. Acknowledgements

We want to thank all the research groups and listeners that took part in the evaluation campaign.

The evaluation campaign was funded by the *Red Temática en Tecnologías del Habla* (TEC2009-06876-E) and the work presented in this paper has been partially funded by

the Spanish Government under grants TEC2006-13694-C03-02 (AVIVAVOZ project) and TEC2009-14094-C04-02 (BUCEADOR project).

## 7. References

Abdi, H (2007). Bonferroni and Šidák corrections for multiple comparisons. In: N.J. Salkind (ed.). *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage.

Black, A.W., Tokuda, K. (2005). The Blizzard Challenge-2005: Evaluating corpus-based speech synthesis on common datasets. In: *Ninth European Conference on Speech Communication and Technology*. ISCA, pp. 77--80.

Bonafonte, A., Moreno, A. (2008). Documentation of the upc_esma Spanish database. Barcelona, Spain.

Grice, M. (1989). Syntactic Structures and Lexicon Requirements for Semantically Unpredictable Sentences in a Number of Languages. In: *Speech Input/Output Assessment and Speech Databases*. pp. 19--22.

Hunt, A.J., Black, A.W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In: *ICASSP '96: Proceedings of the Acoustics*. vol 1 , pp. 373--376.

Karaiskos, V., King, S., Clark, R.A.J., Mayo, C. (2008). The blizzard challenge 2008. In *Proc. Blizzard Challenge Workshop*, Brisbane, Australia, September 2008.

Marcus-Roberts, H. M., Roberts, F. S. (1987). Meaningless Statistics. In: *Journal of Educational and Behavioral Statistics*. 12 (4), pp. 383--394.

Penfold Street, A., Street, D. J. (1987). *Combinatorics of Experimental Design*. New York, NY: Oxford University Press.

Sesma, A., Moreno, A. (2000). CorpusCrt 1.0: Diseño de corpus orales equilibrados. UPC Technical Report.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. In: *Biometrics*, vol. 1, pp. 80--83.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., Tokuda, K. (2006): The HMM-based speech synthesis system (HTS) version 2.0. In: *The 6th International Workshop on Speech Synthesis*. vol 6 , pp. 294--299.