EVALUATION OF A SPANISH TEXT-TO-SPEECH SYSTEM

Lourdes Aguilar, Josep M. Fernández, Juan M. Garrido, Joaquim Llisterri
Departament de Filologia Espanyola, Edifici B, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain. Fax: (34.3) 581.16.86. E-mail: Joaquim.Llisterri@cc.uab.es

Alejandro Macarrón, Luis Monzón, Miguel Ángel Rodríguez
División de Servicios de Tratamiento del Habla, Telefónica I+D, Emilio Vargas 6, 28043 Madrid, Spain. Fax: (34.1) 337.42.02. E-mail: miguel@craso.tid.es

Abstract.- *A battery of tests for output assessment of Spanish TTS systems is presented, as well as its application to the system developed by Telefónica I+D. This battery is an adaptation to Spanish of tests already used in other languages. At phonemic level, consonants, consonant clusters and vowel combinations are evaluated. At word level, identification of words in meaningful sentences as well as in semantically unpredictable sentences is considered. Evaluation of comprehension is accomplished by means of a listening test, and the global quality and user's acceptance is assessed using a paired adjectives test. The results show that the Telefónica I+D TTS system can be favorably compared with existing English products and offers insights for further improvements.*

1.- INTRODUCTION

Due to the availability of different Spanish text-to-speech (TTS) systems, the need to develop a language-specific battery of tests for evaluating and assessing their performance has arisen. To respond to this need, a battery of tests for the evaluation of segmental and word intelligibility, comprehension and global quality of synthesized speech in Spanish has been developed. One possible approach would have been to produce a Spanish version of SAM tests (SAM, 1992). However, it was decided to adapt more classical tests in order to be able to compare the results with previous global evaluations of other systems (Pisoni, 1987). Of course, this does not preclude future adaptations of SAM protocols.

This battery has been applied to the evaluation of the TTS system developed by *Telefónica I+D* (Rodríguez *et al.*, 1993). The main aims when evaluating this particular system have been: (a) the comparison of the intelligibility and comprehension of the synthesized output with respect to the natural voice; (b) the identification of the aspects of the system that require an improvement; (c) the collection of data allowing the comparison of the system with other systems for Spanish and for other languages; and (d) the evaluation of the user's acceptance in applications such as text transmission over the telephone line.

2.- TESTS

2.1.- Segmental intelligibility tests

The Modified Rhyme Test (MRT, House *et al.*, 1965) has been adapted to Spanish for the evaluation of speech intelligibility at the phonemic level. The basic format of the MRT has been used to prepare three different tests focusing on initial and final consonants, consonant clusters and vowel combinations.

In the first test, consonant intelligibility in initial and final position in monosyllabic CVC words is assessed. The strong tendency to bisyllabic word structure in Spanish prevents from finding the necessary number of words to fulfill the numeric requirements of the original test. Thus, the six choices in the English MRT are reduced to four, and the original fifty groups to forty. Segments considered in initial position are [p, b, m, f, θ, t, d, s, n, l, r, ʧ, k, g]; in final position [p, β, m, f, θ, t, ð, s, n, l, ɾ, j, k, ɣ] are taken into account, according to the phonotactic constraints of Spanish.

The full inventory of syllable onset consonant clusters is evaluated in the second test: [pr, tr, br, dr, gr, fr, pl, kl, bl, gl, fl]. The restriction over the meaning of the words imposed on the MRT has not been maintained due to the lack of lexical items presenting these combinations, and meaningless but phonologically possible words have been used.

The intelligibility of the following vowel combinations and accentual contrasts is evaluated in the last test in this series: ['ee, e'e, 'io, i'o, 'ie, i'e, 'io, i'o, 'ia, 'oe, o'e, 'oo, o'o, 'je]. In this case, all items are actual Spanish words.

2.2.- Word identification tests

Intelligibility evaluation at word level is carried out by means a Spanish adaptation of the Harvard Psychoacoustic Sentences (HPS, Egan, 1948) and of the Haskins Semantically Anomalous Sentences (HSAS, Nye-Gaitenby, 1974). The version of the HPS consists of 100 syntactically and semantically well-formed sentences covering a wide range syntactic structures. All sentences are affirmative and they are phonetically balanced in groups of 10. The adaptation of the HSAS consists of 50 sentences with a fixed syntactic structure: <article> + [<adjective>/ <noun>/ <adverb>] + <noun>+ <verb>+ <article>+ <noun>.

2.3.- Comprehension test

The comprehension test developed for Spanish (based in Pisoni, 1987) is composed of three texts with different styles -news, scientific and literary- , each one followed by 6 multiple-choice questions with four possible answers.

2.4.- Global quality test

The global quality test is an enlarged version of the one proposed by Robert *et al.* (1989). The test for Spanish includes 37 pairs of bipolar adjectives or expressions taking into account the following aspects: voice quality, reading style, geographic variety, segmental and suprasegmental quality, global quality, appropriateness to possible applications and user's acceptance. A 1-to-5 point scale has been used in the answers, 1 being close to the negative member of the pair and 5 to the positive one. The test is preceded by a 127 words synthesized text.

3.- EVALUATION OF THE *TELEFÓNICA I+D* TTS SYSTEM

3.1.- Methodology

A specific methodology was devised to meet the aims of the evaluation of the *Telefónica I+D* TTS system: (a) each test had to be presented in four different conditions: male natural, male synthesized, female natural and female synthesized voice, except the global quality test which was only presented for synthesized voice; (b) the number of subjects to be tested had to be chosen to meet statistical reliability criteria, so that a total of 4000 answers (25 subjects) for each condition were necessary for the consonant intelligibility test, 1000 (25 subjects) for the consonant cluster and vowel combination tests, 8000 (20 subjects) for the HPS test, 4000 (20 subjects) for HSAS test, 360 (20 subjects) for the comprehension test and 1140 (30 subjects) for the global quality test; (c) each subject had to be tested with natural speech and with synthesized speech in this order; (d) each subject had to listen to both male and female voices; (e) the experience had to reproduce telephone line conditions.

To fulfill these requirements, a total of 280 subjects -93 men and 187 women, aged between 18 and 30 with a high level of education - organized in 56 groups participated in the evaluation. Stimuli were band-pass filtered at 300-3400Hz and recorded on magnetic tapes by *Telefónica I+D*. A Tascam 112 tape recorder and Sennheiser HD-25-1 binaural headphones were used in

the experiment. Subjects listened to the stimuli at a comfortable intensity level in a sound-treated room at the *Universitat Autònoma de Barcelona.*

3.2.- Results

3.2.1.- Segmental intelligibility tests

Results were organized in several confusion matrices in order to determine intelligibility scores and consonant confusions. Table II presents the overall percentage of correct identifications obtained in the segmental intelligibility tests for each of the four conditions.

	Natural male voice	Natural female voice	Synthesized male voice	Synthesized female voice
Initial consonants	98%	98%	93.2%	90.6%
Final consonants	94.6%	96.8%	79.7%	79.3%
Consonant clusters	95%	93.5%	84%	59.6%
Vowel combinations	97%	94.2%	79.4%	79.2%

Table II. Percentage of correct identifications obtained in the segmental intelligibility tests.

An overall error rate of 13.55% in male synthesized voice and of 15.05% in female synthesized voice for initial and final consonant is obtained. These scores are higher than those found for some English TTS systems such as DECtalk , Prose 2000 or MITalk, with error rates in the 3% to 7% range, but lower than those found for other systems such as Votrax, Smoothtalker and Echo, in which error rates are higher than 25% (Logan *et al.*, 1989).

3.2.2.- Word intelligibility tests

The correct identification percentages obtained in key words in meaningful (HPS) and meaningless but syntactically correct sentences (HSAS) for each of the conditions are presented in table III.

	Meaningful sentences	Meaningless sentences
Natural male voice	99.33%	95.13%
Natural female voice	99.33%	96.48%
Synthesized male voice	96.65%	87.3%
Synthesized female voice	94.69%	84.2%

Table III. Percentage of correct identification of key-words in meaningful and in meaningless sentences.

Word recognition results are similar to those shown by other English systems: 95.3% for DECtalk's male voice and 90.5% for the female voice version. In meaningless sentences, results seem to be better than those obtained for English: 86.8% for DECtalk's male voice and 75.1% for the female version (Pols, 1991).

3.2.3.- Comprehension test

Results of the comprehension test are computed as global percentages of correct answers to the multiple-choice questions. The results are shown in Table IV:

Natural male voice	76.11%
Natural female voice	83.88%
Synthesized male voice	73.88%
Synthesized female voice	72.22%

Table IV: Percentage of correct answers to the comprehension test

The results for synthesized speech comprehension are comparable to those obtained in a similar test using MITalk - 70.3% of correct answers (Pisoni, 1987).

3.2.4.- Global quality test

The best scores in the global quality test were obtained for the female voice - 2.4 for male voice and 3.1 for male voice in a 1-to-5 scale-. The system was considered as more adequate, efficient, satisfactory and acceptable in its female voice version, specially when the application to information services and the potential frequency of use of the system were assessed.

4.- CONCLUSIONS

If the intelligibility and comprehension scores are considered, the system performs better in its male version than in its female one. However, in the global quality test a clear preference for the female synthesized voice appeared, justifying the need for further improvement of this version. This large-scale evaluation has identified the main errors in the segment inventory used by the system, and has also shown the need to improve the modeling of vowel combinations as well as the voice quality. The comparison with well-known TTS systems for English has shown a good performance in word intelligibility and text comprehension, although segmental intelligibility is still behind the performance of the best products. Since the tests have also been carried out with natural voice, the results obtained can be used as a reference for future work.

References

EGAN, J.P. (1948) "Articulation testing methods", *Laryngoscope* 58: 955-991.

HOUSE, A.S.- WILLIAMS, C.E.- HECKER, M.H.L.- KRYTER, K.D. (1965) "Articulation Testing Methods: Consonantal Differentiation with a Closed- Response Set", *Journal of the Acoustical Society of America*, 37, 1: 158- 166.

LOGAN, J.S.- GREENE, B.G.- PISONI, D.B. (1989) "Segmental intelligibility of synthesized speech produced by rule", *Journal of the Acoustical Society of America*, 86 (2): 566- 581

NYE, P.W.- GAITENBY, J. (1974) "The Intelligibility of Synthesized Monosyllable Words in Short, Syntactically Normal Sentences", *Haskins Laboratories Status Report on Speech Research* SR-37/38: 169-190

PISONI, D.B. (1987) "Some measures of intelligibility and comprehension" in ALLEN, J. - HUNNICUTT, M. S. - KLATT, D. H. *From Text to Speech. The MITalk System.* Cambridge: Cambridge University Press. pp. 151-171.

POLS, L.C.W. (1991) "Quality assessment of text-to-speech synthesis by rule", in FURUI, S.- SONDHI, M.M. (Eds.) *Advances in Speech Signal Processing.* Marcel Dekker Inc. pp. 387-416

ROBERT, J.M.- CHOINIÈRE, A.- DESCOUT, R. (1989) "Subjective evaluation of the naturalness and acceptability of three TTS systems in French", in TUBACH, J.P.- MARIANI, J.J. (Eds.) *Eurospeech 89. European Conference on Speech Communication and Technology.* Edinburgh: CEP Consultants Ltd. vol. 2. pp. 640-643

RODRÍGUEZ, M.A.- ESCALADA, J.G.- MACARRÓN, A.- MONZÓN, L. (1993) "AMIGO: Un conversor texto-voz para el español", *Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural* 13: 389-400.

SAM (1992) ESPRIT 2589 SAM "User guide to output assessment", Ref. SAM-UCL-G006, University College London, April 1992.