# A PHONETICALLY BALANCED MODIFIED RHYME TEST FOR EVALUATING CATALAN SPEECH INTELLIGIBILITY

*Francesc Alías and Manuel Pablo Triviño*

GPMM - Grup de Recerca en Processament Multimodal
Enginyeria i Arquitectura La Salle. Universitat Ramon Llull
Quatre Camins, 2, 08022 - Barcelona, Spain
falias@salle.url.edu

## ABSTRACT

This work introduces a phonetically balanced modified rhyme test (MRT) for evaluating Catalan speech intelligibility. The proposal complies with the standard MRT restrictions, besides yielding phonetic balanced word ensembles so as to avoid biasing the test to scarcely representative phonemes. Hence, it allows testing the intelligibility of any communication system delivering Catalan speech by means of a unique phonetic meaningful comparison framework.

**Keywords:** Rhyme test, speech intelligibility, Catalan language, text-to-speech synthesis.

## 1. Introduction

Since 2005 the speech synthesis community is tackling a large scale multi-site evaluation of text-to-speech (TTS) synthesis systems using common data, named *The Blizzard Challenge* [3]. The main goal of this worldwide challenge is defining a framework for fair TTS systems performance comparison in terms of synthetic speech *intelligibility* and *naturalness*.

There exists a myriad of speech intelligibility evaluation tests posed in the literature, some of them are focused on (*i*) *segmental level* analysis, such as the Rhyme Test [8], the Modified Rhyme Test (MRT) [10], the Diagnostic Rhyme Test [21], or the Minimal Pairs Intelligibility Test (MPIT) [20], while others are devoted to (*ii*) *supra-segmental level* analysis, such as the Harvard Psychoacoustic Sentences (HPS) [6], the Haskins syntactic sentences [13] or the Semantically Unpredictable Sentences (SUS) [4]. Among them, the *challenge* is making use of the MRT (as firstly suggested in [14]) and the SUS test for evaluating TTS systems intelligibility. Both are used as open response tests where the listener is asked to type the perceived words into a text box [3]. Thus, strictly speaking, they are making use of the open version of the MRT [12]. On the other hand, the synthetic speech naturalness is evaluated by means of the well-known 5 points scale Mean Opinion Score (MOS) test [5] (5= excellent, 4= good, 3= fair, 2= poor, 1= bad).

In this context, this work is focused on developing a standardized test to evaluate the intelligibility of Catalan speech. As far as we know, the intelligibility of TTS systems delivering Catalan speech has only been evaluated in [9], where a MPIT approach was followed. Allowing for the lack of standards in this area, this paper introduces a phonetically balanced test as a basis for evaluating Catalan speech intelligibility, which conforms to the standard restrictions of MRT [10], besides yielding a phonetically balanced test (i.e. considering the phonetic distribution of Catalan language when developing the test to avoid biasing the test to scarcely representative phonemes). Although the starting point of this work was to allow the evaluation of the segmental intelligibility for Catalan TTS systems, the proposal is also useful for evaluating, e.g. (*i*) the intelligibility of coded or distorted speech over any communication system [11] or (*ii*) the performance of different acoustic distances in order to find optimal joins between concatenated units [18].

This paper is organized as follows. Section 2 briefly reviews the main tests for evaluating speech intelligibility. Then, sections 3 and 4 describe the proposed rhyme test for evaluating Catalan speech intelligibility –both the design process and its phonetic distribution. Finally, section 5 presents the conclusions of this work.

## 2. Testing speech intelligibility

Rhyme tests are available for many languages, being designed for each language individually (i.e. considering each language orthographic and phonetic particular characteristics) to test speech intelligibility in that language. The US English MRT [10] (inspired in the seminal idea posed by Fairbanks [8]) uses 50 six-word lists of rhyming or similar-sounding meaningful monosyllabic words, consisting of a set of 6 words, each with 25 items varying in the initial consonant (e.g. "*book*, *took*, *shook*, *cook*, *hook*, *look*") plus 25 items varying in the final consonant (e.g. "*bat*, *bad*, *back*, *bass*, *ban*, *bath*"). Each word is con-

structed from a consonant-vowel-consonant (CVC) sound sequence. Listeners are shown a six-word list and then are asked to identify which one has been spoken by the speaker from a list of possible choices, i.e. it is a closed-response test which allows the intervention of untrained listeners in the evaluation process [10]. The MRT results evaluate listeners errors in discriminating both initial and final consonant sounds by means of (*i*) the frequency of particular confusions of consonant sounds, computed by means of a confusion matrix, or (*ii*) the number of words perceived incorrectly (i.e. the word error rate), used, for instance, for evaluating TTS systems [3].

As far as we know, the original US English MRT has been adapted for different languages, such as Chinese [11] and Czech [19] –with 40 six-word lists– or Spanish with 40 four-word lists [1]. However, for the meantime, there has been no MRT proposal for Catalan –even though a Catalan MPIT was introduced in [9] (page 127).

## 3. Designing phase

According to the US English MRT definition [10], each test row of the Catalan MRT proposal must be composed of meaningful monosyllabic CVC words (allowing #CV or CV# forms, where # stands for silence hereafter), with constant orthographic representation of vowel nucleus. Notice that these restrictions make no attempt to achieve a phonetic balanced test. Hence, the stimulus items used in the MRT may not entirely be representative of the distribution of phonemes found in English language [13]. In order to avoid this drawback, the degree of phonetic balance can also be considered as a restriction when designing the MRT components [1], e.g. as in the HPS test [6].

Following a counterpart scheme for adapting the MRT to Castilian Spanish [1], the present work defines a Catalan MRT proposal so as to comply with the MRT standard [10] besides adapting its contents to the Catalan language phonetic distribution according to [16] (see the 6th column of table 1). In this manner, the test will better reflect the language characteristics, and hence, will allow performing more realistic speech intelligibility evaluations.

### 3.1. Catalan general phonetic particularities

Catalan is a Romance language spoken by about 6 million people in Eastern Spain mainly. Like other languages, Catalan has several dialects: Central and Western Catalan, Valencian, and Majorcan, which are spoken in Catalonia, the Valencian Region and Majorca (and the rest of Balearic Islands), respectively. All four dialects share the same vowel system composed of 8 vowels: /a/, /e/, /E/, /i/, /o/, /O/, /u/ and /@/ (schwa) (hereafter, all phones are represented according to SAMPA notation [17]).

It is important to note that obtaining monosyllabic words with CVC structure in Catalan is a challenging task, as the proportion of monosyllables is only the 35% of Catalan words in contrast to the 80% in English [15]. Thus, it was necessary to resort to unfamiliar words during the collection phase. Moreover, Catalan is a language with (*i*) a rather strong assimilation property, occurring at final word position, e.g. /b/, /d/ and /g/ turn to /p/, /t/ and /k/, respectively (as in Czech language [19]), or 't' and 'r' are unpronounced (e.g. "*pont*" (bridge) and "*por*" (fear) are transcribed as /pon/ and /po#/, respectively), and (*ii*) consonantal palatalizations, e.g. 'c' and 'g' before 'e' and 'i' turn to /s/ and /Z/, respectively (nor /k/ neither /g/). As a consequence, the phonetic transcription of the collected words had to be controlled to avoid having words with identical spoken form within the same ensemble. Finally, focusing on the CVC words structure, it is to note that phonemes /r/ and /N/ are not allowed to be the word initial consonant, while /b/, /d/, /g/, /Z/, /r/ and /z/ can not be the final consonant –yielding the null rows of table 1.

### 3.2. Collecting CVC words

In order to fulfill the aforementioned test restrictions, the first step of the design process was focused on collecting as many meaningful Catalan CVC words as possible. To that effect, we looked up different bibliographic sources, such as the Catalan-Valencian-Majorcan Dictionary [2] and the *Diccionari General de la Llengua Catalana* [7], searching for monosyllabic nominal lexical entries.

After conducting an exhaustive search, 486 monosyllabic words were obtained, with 30 combinations of 6 stimuli with fixed initial consonant (e.g. "*cas*, *cal*, *call*, *cant*, *cap*, *car*") and 39 with fixed last consonant (e.g. "*sic*, *xic*, *dic*, *nyic*, *pic*, *ric*"). The next step was devoted to find the *optimal* set of monosyllables which maximized the trade-off between the number of CVC forms per ensemble and their phonetic balancing (i.e. looking for a large enough language balanced intelligibility test).

### 3.3. Selecting the CVC candidates

The initial collection of CVC words was composed of $35.46\%$ of plosives (basically /b/,/p/,/k/ and /t/), $24.56\%$ of fricatives (mainly /s/ and /f/), $20.15\%$ of liquids (/l/, /L/ and /rr/), and $19.83\%$ of nasals (basically /m/ and /n/), which is quite close to their distribution in language: $37.22\%$, $20.22\%$, $23.15\%$ and $19.41\%$, respectively (percentages computed from the relative consonant frequencies reported in [16]). However, after analyzing the collection at phoneme

**Table 1:** Phonetic content of the Catalan MRT proposal. Consonants relative frequencies vs. language distribution.

| Phoneme | Initial | Final | Total | MRT (%) | Rafel [16](%) | Difference (%) |
|---|---|---|---|---|---|---|
| /p/ | 11 | 6 | 17 | 2.97 | 2.76 | +0.21 |
| /t/ | 15 | 17 | 32 | 5.59 | 5.24 | +0.35 |
| /k/ | 10 | 17 | 27 | 4.72 | 4.41 | +0.31 |
| /b/ | 18 | 0 | 18 | 3.14 | 2.96 | +0.18 |
| /d/ | 25 | 0 | 25 | 4.37 | 4.48 | −0.11 |
| /g/ | 5 | 0 | 5 | 0.87 | 0.95 | −0.08 |
| /s/ | 21 | 29 | 50 | 8.73 | 8.57 | +0.16 |
| /S/ | 1 | 1 | 2 | 0.35 | 0.35 | +0.00 |
| /f/ | 7 | 1 | 8 | 1.40 | 1.18 | +0.22 |
| /z/ | 2 | 0 | 2 | 0.35 | 0.76 | −0.41 |
| /Z/ | 3 | 0 | 3 | 0.52 | 0.44 | +0.08 |
| /l/ | 7 | 27 | 34 | 5.94 | 6.00 | −0.06 |
| /L/ | 3 | 3 | 6 | 1.05 | 0.85 | +0.20 |
| /m/ | 11 | 12 | 23 | 4.02 | 3.83 | +0.19 |
| /n/ | 12 | 23 | 35 | 6.11 | 6.38 | −0.27 |
| /J/ | 1 | 1 | 2 | 0.35 | 0.28 | +0.07 |
| /N/ | 0 | 3 | 3 | 0.52 | 0.36 | +0.16 |
| /rr/ | 8 | 9 | 17 | 2.97 | 2.32 | +0.65 |
| /r/ | 0 | 0 | 0 | 0.00 | 3.77 | −3.77 |
| Total | 160 | 149 | 309 | 53.97 | 55.89 | $-0.10\% \pm 0.91$ |

level, a large RMSE of $3.19\%$ was obtained when computing the deviation, consonant per consonant, between the collection contents and the language consonantal distribution.

Therefore, the initial set of CVC words candidates was sequentially searched until a *good* phonetic balance was achieved. Specifically, the iterative process was repeated until no consonant presented an absolute deviation higher than 1% w.r.t the language distribution [1] (see last column of table 1).

## 4. Final proposal

As a result of the design process, the number of combinations per row was adjusted to be four, as in the Spanish MRT proposal [1]. That is, the final Catalan MRT is composed of $40 \times 4$ CVC words, as shown in table 2. Columns A-D represent the four test forms while their rows represent response ensembles, i.e. the first 20 rows correspond to the CVC words varying in the initial consonant, while the last 20 rows contain the CVC words varying in the final consonant –including their corresponding phonetic transcription in columns 5th to 8th. As it can be observed, there is no #VC combination in the table, whereas there are 8 CV# forms (i.e. the silence only represents the $1.7\%$ of the total number of phonemes, thanks to the optimization process). Moreover, notice that the liquid consonant /r/ is missing due to the linguistic reasons described in section 3.2. (i.e. this phoneme only appears in VCV forms, which do not conform to the CVC structure of MRT components). Hence, this phoneme is not present in the final proposal (see last row of table 1).

### 4.1. Phonetic distribution

As previously mentioned, the initial CVC words collection presented a high deviation in terms of language consonantal distribution (RMSE=$3.19\%$). Thanks to the expert pruning of the collection contents, an RMSE relative reduction of $71.8\%$ is achieved, yielding a RMSE = $0.899\%$. It is important to note that, on one hand, this improvement is statistically significant in terms of ANOVA ($F(1, 38) = 4.14$, $p = 0.493$), and, on the other hand, besides reducing the error, the correlation between test and language consonant distribution is also improved from $\rho = 0.666$ to $\rho = 0.933$. Moreover, if the /r/ phoneme is not considered in the computation, the results are even better, i.e. the RMSE is reduced from $2.79\%$ to $0.58\%$, being the improvement more significant in terms of ANOVA ($F(1, 36) = 14.14, p = 0.0006$), and the correlation is increased from $\rho = 0.762$ to $\rho = 0.995$.

Furthermore, despite lying beyond the initial scope when defining the MRT proposal, the vocalic distribution of the final proposal (160 instances) is also analyzed. As a result, it can be observed that the proposal implicitly attains a good correlation with respect to language vocalic distribution, after grouping /@/+/a/ in the same category (i.e. $\rho = 0.935$). Notice that including the /@/ vowel (schwa) in the MRT was particularly complicated, due to the predominant tonicity of CVC words in Catalan. In order to allow the presence of the schwa vowel (the most frequent vowel in Catalan [16]), a set of pseudo-CVC forms was explicitly included in the proposal (see the 4th row of table 2).

**Table 2:** Test word lists arranged to A-D forms. Each row represents the test ensemble with its SAMPA [17] phonetic transcription in italics. The first 20 ensembles correspond to the fixed initial consonant forms, while the latter 20 ensembles accomplish the reverse situation.

| A | B | C | D | /A/ | /B/ | /C/ | /D/ |
|---|---|---|---|---|---|---|---|
| bat | baf | vas | ban | *bat* | *baf* | *bas* | *ban* |
| bis | bit | vint | vim | *bis* | *bit* | *bin* | *bim* |
| cot | cós | com | con | *kot* | *kos* | *kom* | *kon* |
| se'l | se'm | ses | se'n | *s@l* | *s@m* | *s@s* | *s@n* |
| sin | cinc | sis | si | *sin* | *siN* | *sis* | *si#* |
| cent | sés | cec | sé | *sen* | *ses* | *sek* | *se#* |
| duc | dull | dur | dus | *duk* | *duL* | *du#* | *dus* |
| des | deix | dent | de | *des* | *deS* | *den* | *de#* |
| dot | do | dol | don | *dOt* | *dO#* | *dOl* | *dOn* |
| gal | gall | gat | gas | *gal* | *gaL* | *gat* | *gas* |
| ra | ras | ram | ran | *rra#* | *rras* | *rram* | *rran* |
| la | lar | las | lat | *la#* | *larr* | *las* | *lat* |
| ret | rep | rent | res | *rret* | *rrep* | *rren* | *rres* |
| ben | bes | vet | bé | *ben* | *bes* | *bet* | *be#* |
| das | dalt | dà | dany | *das* | *dal* | *da#* | *daJ* |
| nus | nuc | nul | nu | *nus* | *nuk* | *nul* | *nu#* |
| nan | nap | nas | nat | *nan* | *nap* | *nas* | *nat* |
| pus | pul | punt | puny | *pus* | *pul* | *pun* | *puJ* |
| tall | tan | tanc | tac | *taL* | *tan* | *taN* | *tak* |
| tun | tul | tuc | tu | *tun* | *tul* | *tuk* | *tu#* |
| puc | cuc | suc | duc | *puk* | *kuk* | *suk* | *duk* |
| sic | fic | dic | tic | *sik* | *fik* | *dik* | *tik* |
| bac | sac | tac | mac | *bak* | *sak* | *tak* | *mak* |
| vas | cas | zas | fas | *bas* | *kas* | *zas* | *fas* |
| dus | pus | lluç | mus | *dus* | *pus* | *Lus* | *mus* |
| les | pes | ves | tes | *lEs* | *pEs* | *bEs* | *tEs* |
| dalt | val | mal | pal | *dal* | *bal* | *mal* | *pal* |
| zel | mel | del | cel | *zEl* | *mEl* | *dEl* | *sEl* |
| gol | mol | dol | col | *gOl* | *mOl* | *dOl* | *kOl* |
| mul | nul | tul | pul | *mul* | *nul* | *tul* | *pul* |
| vil | quil | fil | Gil | *bil* | *kil* | *fil* | *Zil* |
| fat | lat | mat | nat | *fat* | *lat* | *mat* | *nat* |
| vim | quim | cim | llim | *bim* | *kim* | *sim* | *Lim* |
| com | som | dom | mom | *kom* | *som* | *dom* | *mom* |
| dent | ment | fent | sen | *den* | *men* | *fen* | *sen* |
| sin | gin | nin | tint | *sin* | *Zin* | *nin* | *tin* |
| xap | nap | nyap | pap | *Sap* | *nap* | *Jap* | *pap* |
| far | lar | tar | mar | *farr* | *larr* | *tarr* | *marr* |
| llur | mur | pur | tur | *Lurr* | *murr* | *purr* | *turr* |
| sit | dit | fit | git | *sit* | *dit* | *fit* | *Zit* |

## 5. Conclusions

This paper introduces a rhyme test for evaluating Catalan speech intelligibility for any communication system or experimental condition, e.g. for evaluating Catalan TTS systems performance. The proposal complies with the standard restrictions of the Modified Rhyme Test, besides yielding a phonetically balanced distribution of words. As a result, the proposal is composed of 40 four-lists of rhyming meaningful CVC monosyllabic words. In the near future, we will keep working in improving the definition of intelligibility tests so as to better reflect Catalan language particularities, e.g. by defining further test proposals towards improving test phonetic coverage in phonemes such as /r/ and schwa.

## 6. REFERENCES


[1] Aguilar, L. 1991. Propuesta de un Test de Rimas Modificado para el español, Universitat Autònoma de Barcelona.

[2] Alcover, A. M., Moll, F. de B. *Diccionari català-valencià-balear*, ed. Moll.

[3] Bennett, C. and Black, A.W. 2006. *The Blizzard Challenge*, Pittsburgh, USA.

[4] Benoit, C. and Grice, M. 1996. The SUS test: a method for the assessment of text-to-speech intelligibility using Semantically Unpredictable Sentences, *Speech Communication*, vol 18, 381-392.

[5] CCITT. 1984. Absolute category rating method for subjective testing of digital processors, *Red Book*, vol V.

[6] Egan, J. 1948. Articulation testing methods. *Laryngoscope* 58, 955-991.

[7] Gran Diccionari de la Llengua Catalana, ed. Enciclopèdia Catalana `http://www.grec.net/home/cel/dicc.htm` visited 6-March-07.

[8] Fairbanks, G. 1958. Test of phonemic differentiation: The Rhyme Test, *J. Acoust. Soc. Am.*, vol 30, nº 7, 596-600.

[9] Febrer, A. 2001. Síntesi de la parla per concatenació basada en la selecció, *Ph. D Thesis*, Universitat Politècnica de Catalunya,

[10] House, A.S., Williams, C.E., Hecker, M.H.L., and Kryter, K. D. 1965. Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set *J. Acoust. Soc. Am.*, vol 37, nº 1, 158-166.

[11] Li, Z., Tan, E.C, McLoughlin, I. and Teo, T.T. 2000. Proposal of standards for intelligibility tests of Chinese speech, *IEE Proc. on Vision, Image and Signal Processing*, vol. 147, nº 3, 254-260

[12] Logan, J. S., Greene, B. G. and Pisoni, D. B. 1989. Segmental intelligibility of synthetic speech produced by rule. *JASA*, vol 86, nº 2, 566-581.

[13] Nye, P.W. and Gaitenby, J.H. 1974. The Intelligibility of Synthetic Monosyllabic Words in Short, Syntactically Normal Sentences. *Haskins Labs Status Report on Speech Research*, 37/38, 169-190.

[14] Pisoni, D. and Hunnicutt, S. 1980. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system, *Proc. of ICASSP*, vol. 5, 572 - 575.

[15] Prieto, P. 2006. The Relevance of Metrical Information in Early Prosodic Word Acquisition: A Comparison of Catalan and Spanish. *Language and Speech* 49 (2), 233-261.

[16] Rafel, J. 1979. Dades sobre la freqüència de les unitats fonològiques en català, *Estudis Universitaris Catalans XXIII*, vol 2, 473-496.

[17] SAMPA computer readable phonetic alphabet. `http://www.phon.ucl.ac.uk/home/sampa/index.html` visited 7-Feb-07.

[18] Syrdal, A. K. and Conkie A. D. 2005. Perceptually-based data-driven join costs: comparing join types *Proc. of InterSpeech*, Lisbon, 2813-2816.

[19] Tihelka D., Matoušek J. 2004. The Design of Czech Language Formal Listening Tests for the Evaluation of TTS Systems. *Proc. of LREC*, Lisbon, vol VI, 1099-2002.

[20] van Santen, J.P.H. 1993. Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language*, 7, 49-100.

[21] Voiers, W.D. 1977 Diagnostic evaluation of speech intelligibility. *Speech Intelligibility and Speaker Recognition*, vol 2, 374-384.