Less errors with TTS?
A dictation experiment with foreign language learners

Thomas Pellegrini[1], *Ângela Costa*[1,2], *Isabel Trancoso*[1,3]

[1]Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento, Lisbon, Portugal
[2]Universidade Nova de Lisboa, Lisbon, Portugal
[3]Instituto Superior Técnico, Lisbon, Portugal
thomas.pellegrini@inesc-id.pt

Abstract

This article reports a contrastive experiment about the use of Text-To-Speech (TTS) synthesis instead of pre-recorded utterances in a dictation exercise submitted to students of European Portuguese as a second language (PSL). Forty sentences were extracted from a PSL student book. Twenty of them were synthesized and the other twenty were directly taken from the pre-recorded audio documents of the book. It appeared that the synthetic utterances were easier to transcribe than the pre-recorded ones, with word error rates of 26.6% and 33.9% respectively. This result was somehow surprising since the synthetic voice was not built or tuned for learning purposes. Nevertheless, the lower speech rate of the TTS voice (15% slower) may explain this outcome. A manual error categorization showed that less word substitutions and less errors on function words were made on the TTS utterances.

Index Terms: Computer-Assisted Language Learning, speech synthesis, dictation, European Portuguese

1. Introduction

Text-To-Speech (TTS) synthesis allows to generate high quality speech from text input. Nevertheless, its use in Computer-Assisted Language Learning (CALL) systems is still debatable. Many CALL works integrate TTS but do not assess its appropriateness. In [1] for example, listening comprehension and dictation items are automatically generated with texts retrieved from the Web. TTS is used to generate audio speech corresponding to the test items. Nevertheless, no field test over the usability issues was performed.

On the other hand, some studies report in-depth evaluations of TTS for CALL applications. In [2], the author asked herself the question: "Is TTS synthesis ready for use in CALL?" The author assessed the synthesis of four commercial products with three criteria: comprehensibility, acceptability, and appropriateness. She concluded that the best systems were ready for use in various CALL applications, but expressiveness was judged insufficient.

Recently, we developed CALL tools for learners of European Portuguese (EP) as a second language (PSL) [3]. One fundamental aspect of all our tools remains in the automatic generation of curriculum materials for each type of exercises [4]. This is very valuable for teachers, saving them time in search for motivating materials of appropriate quality, level and topic. TTS was already introduced in some of our tools to use the computer as a reading machine [5]. In order to assess the use of TTS in listening comprehension activities, we conducted a first experiment that is reported in this article, where synthesized and pre-recorded utterances were used in a dictation exercise carried out in a real classroom context. The main idea was to test whether the students would perform differently according to the type of speech. Our TTS engine was used "out of the box", with no specific tuning for learning purposes.

Dictation is a traditional language learning writing activity in which the teacher reads aloud a text or plays a pre-recorded text which the student is asked to transcribe it. It provides students with a holistic learning experience that incorporates the essential advantage of developing a focus on both the meaning and form (grammar and spelling) in oral texts. To our knowledge, very few studies directly compared human and synthesized speech in dictations. In [6] for instance, no impact of using TTS was found on the student performance nor on the error types.

2. Method

2.1. Dictation test set

For the dictation exercise, 40 sentences were extracted from all the chapters of the students' book Português XXI level A1. From these 40 sentences, 20 were synthesized and for the other 20 sentences we used the pre-recorded audio from the CD that comes with the students book. The sentence selection was manual, guided by the need of very similar lexical content and equal difficulty level in both sets. The pre-recorded and synthesized sets had 138 and 143 words respectively. On average, each sen-

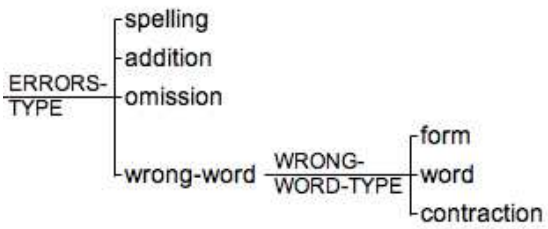

Figure 1: *Taxonomy.*

tence had 7 words. We should also mention that the ratios of function words and content words were identical in both sets: 55.4% function words and 44.5% content words. The vocabulary present in both subsets was mostly constituted of words of every day use connected with, for instance, daily routine, professions, introducing oneself. Syntactically speaking, the majority of the utterances were simple sentences, but some compound and complex sentences were also selected. Most of the sentences were declaratives, but there were also some interrogatives and exclamative sentences.

The DIXI TTS engine was used to generate the synthesized utterances [7]. This engine uses the standard concatenative unit selection synthesis. All the synthetic sentences used the same male voice ("Vicente"), whereas the pre-recorded utterances featured several male and female voices.

2.2. Experimental setup

The group of students that have participated in our experimental exercise were ERASMUS students that were living and studying in Portugal for 5 months. They all had classes in Portuguese at the New University of Lisbon and they attended a Portuguese Language and Culture class of level A2 according to the Common European framework of reference for languages. When the experiment took place, the students already had 56 hours of formal teaching of Portuguese. From the 11 students that took part in the experiment 6 were Italians, 4 were Spanish and 1 was Swedish. The students were familiar with dictation exercises, as it was practiced several times during the classes. We should also mention that the dictation is one of the parts of the final Portuguese exam. This experiment was carried out in the classroom during a regular lesson of Portuguese. The students were asked to write down the dictated sentences. Each sentence was played 3 times, with a small pause between the reproductions. Two loudspeakers were used.

2.3. Error taxonomy

Several classificatory systems have been used in Error Analysis. For instance, the Linguistic Category Classification [8] based in the linguistic categories of the errors (such as morphology, lexis, and grammar) and the Surface Structure Taxonomy [8] that focuses on the way surface structures have been altered by learners (e.g. omission, addition, misformation, and misordering). For our purposes, we have decided to use a simplified taxonomy inspired on the latter. The taxonomy, illustrated in figure 1, is divided into four main categories: Spelling, Addition, Omission, and Wrong-word. Spelling problems correspond to a substitution of one or more letters, resulting in a word that does not exist. Under addition we have considered words that were not present on the audio stimulus and that were added to the sentences by the students. Omission is exactly the opposite. Contrarily to the spelling mistakes, the wrong-word category considers substitutions with words that do exist in Portuguese. In this case, the spoken word was substituted by a different one. Inside this type of error, we distinguished three other subtypes: Form, Word, and Contraction. In the first subtype the word is correct but not in its correct form (masculine/feminine, singular/plural). The second type corresponds to words with no specific relation to the correct ones. Finally, contraction problems correspond to compulsory contractions that were not respected, such as the contraction of prepositions with articles.

3. Results

The global word error rate (WER) was 34.1%. Without considering orthographic accents – for example *sabado* (Saturday) would be correct even if *sábado* is the correct spelling – the WER decreases to 30.2%. In all the following, orthographic accents will not be considered since this study focuses more on the listening perception than the spelling capabilities of the subjects.

Table 1 shows the detail of the average word error rates (WER) for the synthesized ("S") and the prerecorded utterances ("H"). Figure 2 further illustrates the percentages of errors by showing the WER median values according to the speech type. It appeared that many more errors were made in H than in S, with average WERs of 33.9% (sd=15.1) and 26.6% (sd=12.5) respectively. Correspondent median values reached 40.6% and 30.1% as shown in the figure. The difference between the *S* and *H* WER means per student was found to be statistically significant by running a paired t-test that gave a p-value of 0.0019 for a 95% confidence threshold. More errors were made in function words in H (53.2%), whereas content words were more subject to mistakes in S (53.1%). Nevertheless, this difference was not found to be statistically significant (p-value of 0.907). Since the lexical content and difficulty level were very similar in both sets, the differences in WER were expected to be due to the speech characteristics of the human and synthetic recordings, i.e. pronunciation aspects. The following subsections propose an analysis of the potentially involved fac-

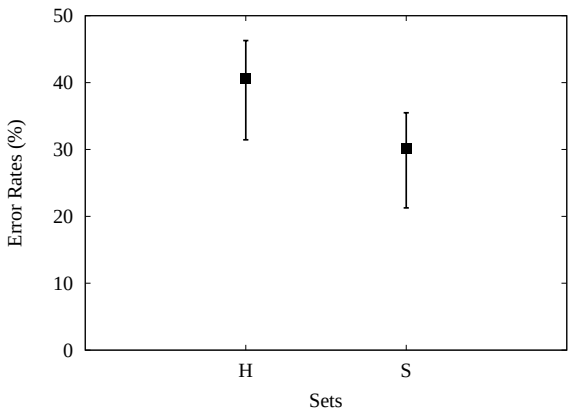

Figure 2: *Word error rates (%) as a function of both types of speech. Median values and standard deviations are represented.*

tors.

Table 1: *Word error rates (WER) for the human (H) and synthesized (S) speech.*

	WER(%)	WER (%) Function Words	WER (%) Content Words
H	**33.9**	**53.2**	46.8
S	26.6	46.9	**53.1**

3.1. Error categorization

The errors were manually categorized according to the taxonomy presented in section 2.3. Spelling problems showed that the "new" words usually demonstrate the influence of the students' mother tongue. For instance several Italian students wrote *journal* instead of *jornal* (newspaper). As we can see, the spellings are very similar in Italian and Portuguese. It is also worth mentioning the influence of sound over orthography. Words with an orthography that has no straightforward congruency with its pronunciation may be difficult for the students. An example of this is the spelling *mai* instead of *mãe* (mother, phonetic transcription: m6~i~). In this case, the pronunciation of the word can be misleading.

Concerning additions, an example was found in the sentence *Sou a amiga da *tua* mãe* (I am your mother's friend). The possessive pronoun *tua* was added to the sentence, when the original sentence was *Sou a amiga da mãe*. In this particular case, the sentence is grammatically and semantically correct even with the additional word, but in general additions lead to non-sense sentences.

Like addition, omission does not always lead to grammatically incorrect sentences. For instance in the sentence *Subiram a pé* (they climbed by foot), the complete sentence is *Subiram a Serra a pé*. The object is missing, but the sentence it is not wrong. It is interesting to verify that the higher number of omissions occurred for very short function words (*a*, *o*, *de*, *que*, *se*), which are strongly coarticulated with the neighboring words.

An example of a form problem is present in the phrase *reportagem desse tipo* (an article of this type) where the correct word would be the same noun but in its plural *reportagens*.

Concerning the wrong-word category of errors, an example would be the sentence *casa na praia* (house on the beach) instead of *casa de praia*. The preposition was substituted by another one. Other errors on prepositions are due to contractions. The preposition *em* (in) for instance, was not contracted with the article *no* (*em* + *o*). Another example: *de este* (from this) should be written *deste*.

In general, it appeared that the numbers of error categories in both sets were very similar. For example, percentages of additions were 5.1% and 6.0%, and of omissions 39.6% and 40.6% in H and S respectively. The largest difference was observed in the wrong-word category. There were less errors in S (37.3%) than in H (40.9%). On the counterpart, there were more spelling mistakes in S (16.1%) than in H (14.4%). This lead us to think that the words spoken by the TTS voice were globally more understandable than the ones of the prerecorded voices.

3.2. Speech rate

To estimate the speech rate, words per second (w/s) and phonemes per second (p/s) were computed for both sets of utterances. The synthetic voice presented slightly smaller rates: 2.32 w/s and 10.02 p/s versus 2.74 w/s and 11.23 p/s for the human speech.

In [9], it was shown that fast rates lower the comprehension of messages presented in TTS synthesis for both native and non-native listeners. The better results of the S set confirmed this observation. The lower speech rate of the TTS voice is one of the factors that could explain the better results of the students on the S set.

3.3. Pronunciation reduction

The reference sentences were phonetically transcribed by a grapheme-to-phoneme tool that provided a "*canonical*" broad phonetic transcriptions, where all the phonemes would be pronounced as if there was no reduction nor co-articulation. A second narrow phonetic transcription was created manually by listening to the utterances. Only 8.5% of the 281 reference words were manually transcribed with a reduced phonetic form (7.7% for the S set and 9.4% for the H set). These words are expected to be more difficult to understand than words with a non-reduced pronunciation, and in fact these words showed a slightly larger error rate than the other ones, 33.3% and

28.4% respectively.

Beside reductions, inter-word phenomena such as co-articulation is expected to make the understanding of an utterance more difficult. To illustrate this point, table 2 shows one of the utterances that caused more problems to the students: "aquele ali ao lado da pastelaria" (*That one next to the bakery*). The second and third rows of the table give the "canonical" transcription (words in isolation) and the manual phonetic transcriptions respectively. The last row shows the WERs for each word.

No student transcribed correctly the first word, and the second word also presented a high error rate (63.6%). The schwa of the first word was not pronounced due to co-articulation with the second word. The use of deictics for this level may also contribute to increased difficulty. One can notice that the last word "pastelaria" was also error-prone. In this case, this may be due to the fact that it is a relatively unfamiliar 3-syllable word (for their knowledge level), with an [S] in coda position, and a strongly reduced schwa in the middle.

Table 2: *Example of a sentence and its canonical and manual phonetic transcriptions with the SAMPA phonetic alphabet. The last row shows the average error rates per word.*

aquele	ali	ao	lado	da	pastelaria
6k"el@	6l"i	aw	l"adu	d6	p6St@l6r"i6
6k"el	6l"i	O	l"adu	d6	p6St@l6r"i6
100.0	63.6	27.3	18.2	27.3	72.7

4. Summary and Future Work

A contrastive study about the use of Text-To-Speech (TTS) synthesis instead of pre-recorded utterances in a dictation exercise was reported in this article. PSL students transcribed forty sentences extracted from a student book. Twenty of them were synthesized and the other twenty ones directly taken from the pre-recorded audio documents of the book. The synthetic utterances were more correctly transcribed than the human ones, with WERs of 26.6% and 33.9% respectively. This outcome was somehow surprising since the synthetic voice was not built for learning purposes whereas the pre-recorded utterances were. The lower speech rate of the TTS voice (15% slower) may explain this result.

A manual categorization of the errors showed that the same types of errors were made in both sets, except a small difference for the wrong-word category that corresponds to word substitutions. There were less errors in the TTS (37.3%) than in the pre-recorded set (40.9%) for this category. Students identified better the words with the TTS voice. Finally, less errors on function words than on content words were made by the students in the TTS set, with 46.9% and 53.1% WER respectively. Meanwhile for the pre-recorded set, the contrary was observed, with 53.2% and 46.8% respectively.

All these results tend to demonstrate that the TTS samples were more understandable than the pre-recorded ones, even for small-length words like function words. The small number of students available for this study does not allow to generalize this outcome, but the use of TTS had a significant impact on their performance in this particular study. It should also be noticed that no student identified the TTS utterances as synthetic speech, although a specific question was asked about the quality of the sentences at the end of the dictation.

As future work, we plan to repeat the experiment with a larger number of students, with several proficiency levels and with a larger diversity of linguistic background. The easy way in which synthesized sentences can be created for the same voice also allows interesting controlled experiments to assess perception difficulties of PSL students with particular phone sequences. Finally, the TTS voice was used "out of the box". Speech rate and co-articulation reduction effects could be explored since TTS allows control on these parameters.

5. Acknowledgments

The authors would like to thank the students from the New University of Lisbon who participated in the experiment. This work was partially supported by FCT (INESC-ID multi-annual funding) through the PIDDAC Program funds.

6. References

[1] S. Huang, C. Liu, and Z. Gao, "Computer-assisted item generation for listening cloze tests and dictation practice in english," in *Proc. of ICWL*, 2005, pp. 197–208.

[2] Z. Handley, "Is text-to-speech synthesis ready for use in computer-assisted language learning?" *Speech Communication*, vol. 51, no. 10, pp. 906–919, 2009.

[3] T. Pellegrini, W. Ling, A. Silva, I. Trancoso, R. Correia, J. Baptista, and N. Mamede, "Overview of computer-assisted language learning for European Portuguese at L2F," in *To appear In Proc. of the 4th International Conference on Computer Supported Education*, Oporto, 2012.

[4] T. Pellegrini, R. Correia, I. Trancoso, J. Baptista, and N. Mamede, "Automatic generation of listening comprehension learning material in European Portuguese," in *Proc. Interspeech*, Florence, 2011, pp. 1629–1632.

[5] J. Lopes, I. Trancoso, R. Correia, T. Pellegrini, H. Meinedo, N. Mamede, and M. Eskenazi, "Multimedia Learning Materials," in *Proc. IEEE Workshop on Spoken Language Technology SLT*, Berkeley, 2010, pp. 109–114.

[6] C. Santiago-Oriola, "Vocal synthesis in a computerized dictation exercise." in *Proc. of Eurospeech*, 1999.

[7] S. Paulo, L.-C. Oliveira, C. Mendes, L. Figueira, R. Cassaca, C. Viana, and H. Moniz, "DIXI – A Generic Text-to-Speech System for European Portuguese," in *Computational Processing of the Portuguese Language*, ser. LNAI, vol. 5190. Springer-Verlag, 2008, pp. 91–100.

[8] H. Dulay, M. Burt, and S. Krashen, *Language Two*. Newbury House, Rowley, 1982.

[9] C. Jones, L. Berry, and C. Stevens, "Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners," *Comput. Speech Lang.*, vol. 21, no. 4, pp. 641–651, Oct. 2007.