A Text-to-Speech System for the Brazilian Portuguese Based on Syllabic Units

Solimar de S. Silva, Fernando G. V. Resende Jr., and Sergio L. Netto

Abstract

This paper describes a new text-to-speech (TTS) system, the so-called TalkActive, based on syllabic units for the Brazilian Portuguese. The system employs a concatenation approach, using the TD-PSOLA algorithm. Preliminary intelligibility tests are included comparing the current version of the TalkActive system to a TTS system based on diphones.

1 Introduction

Concatenation text-to-speech (TTS) synthesizers are systems capable of converting text to speech through the concatenation of the appropriate units of speech. The choice of units is a controversial matter: some systems use units of constant size (e.g., diphones), other systems use non-uniform units, with the longer units modeling the situations of heavy coarticulation [1]. Due to phonological differences between languages, to make a good selection of units is a language dependent problem.

In practice, there is also need to use techniques to modify the units of the database, due to the great variability of the speech, since the database cannot have all possible unit realizations for all possible contexts. The modification is done by the concatenation algorithm, that tends to smooth the parameters on the boundaries of the units, imposing a reasonable prosody (intonation and rhythm) on the utterance [1]. The choice of the concatenation algorithm can be described also as language independent. Most algorithms have their pros and cons, but the variations of the PSOLA scheme (e.g., TD-PSOLA or MBR-PSOLA) are very popular concatenation algorithms in TTS systems. For the Brazilian Portuguese, there are some commercial TTS systems that exhibits good segmental quality. For instance, informal listening tests for the Elan Informatique TTS indicate that the synthesized speech is almost natural, being the generation of natural prosody the main problem of this system. The restricted availability of annotated speech databases for the Brazilian Portuguese inhibits the appearance of very high quality TTS systems for this language.

2 TalkActive System

In general, diphones are the most popular units used in concatenation TTS systems. For the TalkActive system, however, we have chosen the syllables as basic units based on the assumption that larger units imply less concatenation points, providing better segmental quality. In addition, the Brazilian Portuguese is a language considered to be highly syllabic, thus indicating that the decision of considering syllabic units should result in a TTS system with good overall characteristics. This decision tends to increase the size of the database required to implement the TTS system. However, hard-disk space in current computer systems are more than suitable for storing a syllabic database.

The Figure 1 shows the development of the units database and the other modules of the TTS system.

Referring to Figure 1, we now proceed to describe the function and the current degree of development of each block in the TalkActive system:

- Word listing - A list containing the words to be recorded and the respective syllable of interest to be extracted from each word. This list was meant to have all the syllables in the Portuguese language, and was designed by Rosana C. de Oliveira by means of a direct search on a dictionary;
- Recording - All the words in the previous list

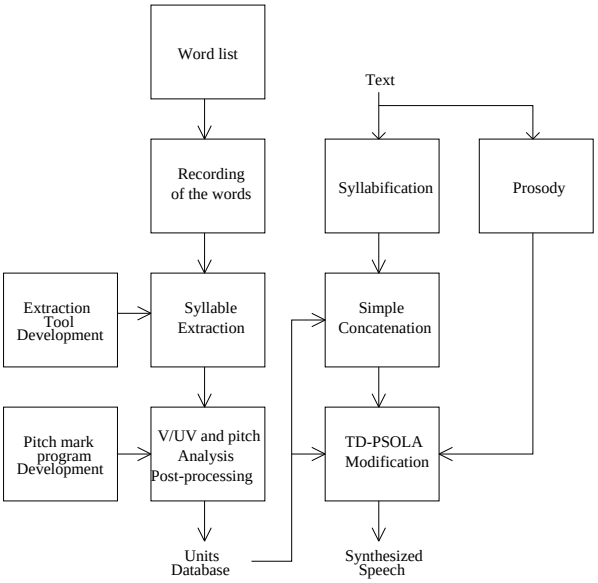

Figure 1: TalkActive system block diagram.

were recorded in a professional recording studio at a 44.1 kHz rate. The speaker was 32-year old Brazilian male that included prosody during the recording stage;

- Syllable extraction - The acoustic realizations of all syllables were identified within the recorded words using an extraction tool designed by the first author of this paper. The extraction process took about 4 months and was performed by Verochile da Silva Jr. generating about 1500 syllable units;

- V/UV and pitch analysis, and post-processing - To implement the concatenation stage using a PSOLA algorithm, the database must have the information about the pitch marks for each realization of the syllables, and this requires a V/UV (*voiced/unvoiced*) and pitch analysis. There is a post-processing stage: the units must also be modified to ensure correct phase alignment during concatenation. The complete processing implemented in this stage will be fully explained in the Section 3;

- Syllabification - This is the TTS module that converts the input text in a sequence of units to be concatenated. This module was developed at the USP - São Carlos University under the supervision of Prof. da Graça, and was kindly made available to the current project;

- Simple concatenation - This block represents the direct concatenation of the units by juxtaposition prior to their modification;

- TD-PSOLA modification - The pitch and duration of the units must be modified to smooth the boundaries of the units and impose natural intonation and rhythm to the synthesized utterance. This block will be fully explained in the Section 4;

- Prosody - This is the TTS module responsible for the assignment of the pitch and duration of the units. This block was not developed yet, and should be the subject of a research during the first author's master program at COPPE/UFRJ.

The Figure 2 shows the TTS system interface, designed by Rodrigo C. Torres. As in its current version, the TalkActive system allows a text input from file (ASCII format) or from the text window made available to the user. Output information can be recorded as a WAV file or send directly to a sound card.

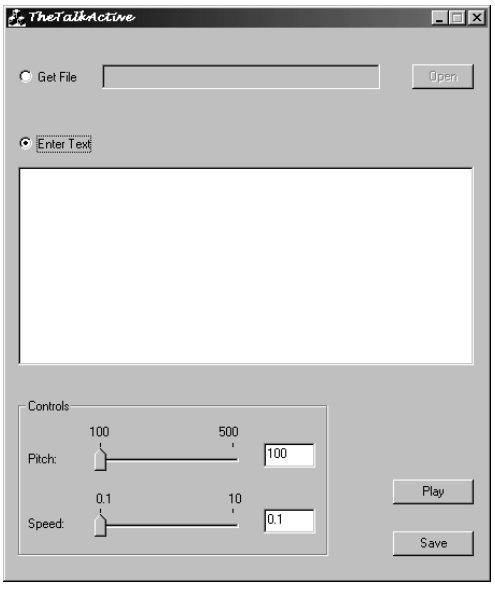

Figure 2: TalkActive system interface.

3 V/UV and Pitch Analysis, and Post-Processing

This block employs a simple autocorrelation pitch determination algorithm, with center-clipping, to estimate the pitch contour of the units. The pitch marks were placed using this pitch contour information. In the voiced regions, they were placed near the highest peak within a period, because it was realized this is one of the most reliable positions to place the pitch marks [2]. In the unvoiced regions, we have placed the pitch marks in such a manner that the spacing between them is equal to the period of the adjacent voiced region.

The original units usually have some samples before the first pitch mark and after the last pitch mark. Since, in the simple concatenation, the last pitch mark of a unit must fall in the position of the first pitch mark of the following unit, to ensure correct phase alignment, the samples of the unit ends must be removed. This is done in the post-processing.

4 TD-PSOLA Modification

This block uses the TD-PSOLA algorithm [3]. The algorithm has the following steps:

- Analysis - The signal $s(n)$ from the "simple concatenation" block is converted to an intermediate representation, multiplying it by a sequence of windows $w(n)$ centered around the pitch marks to obtain a sequence of short-term signals $s_i(n)$:

$$s_i(n) = s(n)w(n - iT_{0i}) \quad (1)$$

  where T_{0i} is the original local pitch period;

- Modification - The original positions of the pitch marks are mapped to new positions to produce the desired pitch contour. To modify the duration, some of the short-term signals are deleted or duplicated in the intermediate representation;

- Synthesis - The synthesized signal $\tilde{s}(n)$ is obtained by an overlap-add (OLA) procedure. We have used a simplified OLA scheme:

$$\tilde{s}(n) = \sum_{i=-\infty}^{\infty} s_i(n - i(T_i - T_{0i})) \quad (2)$$

  where T_i is the synthesis local pitch period.

5 Comparison Results

To evaluate the quality of our system, some preliminary quality-assessing tests were performed [4]:

- A comparison test: This test evaluates the personal preferences between some TTS systems and our system, in the task of synthesizing a set of four phonetically balanced phrases. The TTS systems compared to our system, in this test, were:
  - Elan Informatique TTS: A commercial TTS system for the Brazilian Portuguese designed by Elan Informatique;
  - Delta Talk: A commercial TTS system designed by MicroPower Software;
  - DOSVOX: A system designed to aid visually impaired people. This system was designed by a group from NCE/UFRJ;
  - A diphone based TTS: A system designed at DEL/UFRJ, using LPC parameters of diphone units with a uniform pitch contour [4].
  - The TalkActive without TD-PSOLA: At this stage of development, the TD-PSOLA modification only imposes a uniform pitch contour on the utterances. Therefore, the absence of TD-PSOLA concatenation stage allows the pitch included during the recording stage to interfere with the final intelligibility characteristics of the TTS system.

- An intelligibility test: This test evaluates solely the intelligibility (segmental quality) of the TTS, using a set of 20 words and 20 short phrases, phonetically balanced, using the TD-PSOLA modification to impose a uniform pitch contour. These words and phrases were played for 40 listeners, and each listener wrote down the word (s)he understood. The intelligibility is computed for the words and phrases using the following criteria, as done in [4]:
  - Word intelligibility: The score is incremented by 1 for each word correctly understood;

Table 1: Results of the intelligibility tests.

Synthesizer	Words	Phrases
Diphones	67%	92%
Syllables	(67±10)%	(72±12)%

– Phrase intelligibility: The score is incremented by 1 for each phrase completely understood and by 0.5 if more than 70% of the phrase is understood.

The Figures 3 and 4 show the histograms for the distribution of listener according to the intelligibility score. The Figures 5 and 6 show the score for each word and phrase used in the test. The Table 1 gives a summary of the intelligibility tests for our TTS system and the system based on diphones.

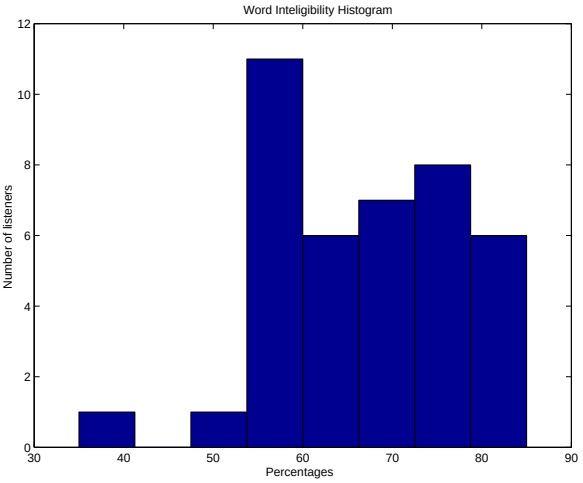

Figure 3: Word intelligibility histogram.

There is a drastic increase of intelligibility in the diphone based system when using phrases that is not observed in our TTS system. This can be explained by the fact that anticipatory coarticulation acts as a redundancy, giving a hint of what is the next phone in the speech stream. But context information were not taken into account when extracting the syllables, eliminating these coarticulation benefits. In the case of diphones, the definition of the units naturally forces the TTS to keep some context information.

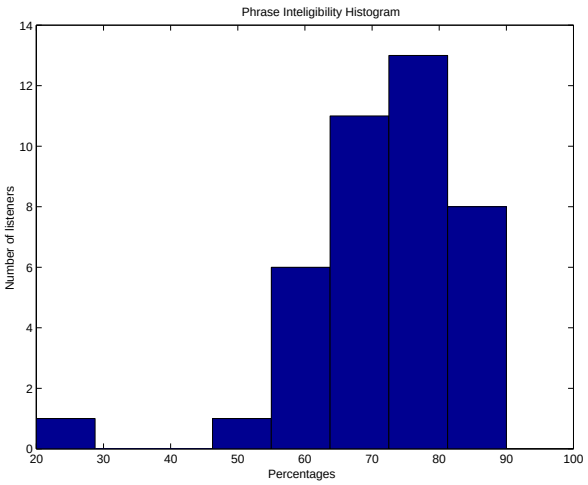

Figure 4: Phrase intelligibility histogram.

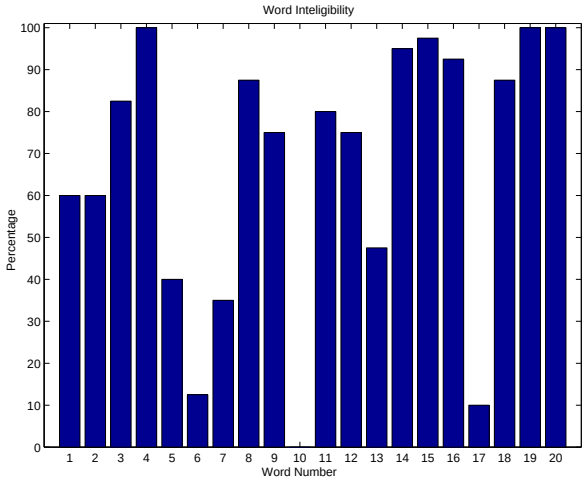

Figure 5: Word intelligibility.

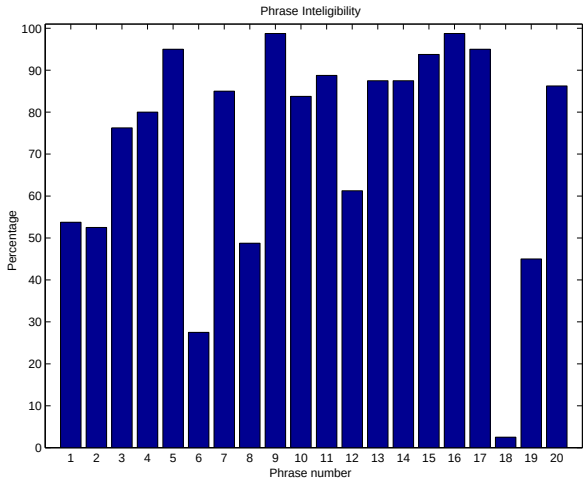

Figure 6: Phrase intelligibility.

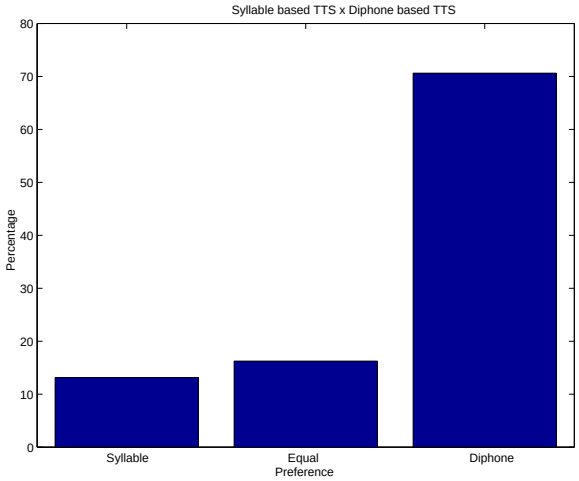

Figure 7: Comparison between the TalkActive system and a diphone based system.

The comparison tests indicated a higher preference for the commercial TTS systems. Our TTS system outperforms only the DOSVOX TTS. The Figures 7, 8 and 9 show the results of the preference tests for the non commercial systems. As we can see, the system using TD-PSOLA is slightly better than the system without the TD-PSOLA modification.

There are some problematic units in the units database, for instance, units having small durations or extracted from stressed syllables. These units cause the low inteligibility of some words and phrases used in the test. Because of these problems, the comparison between the diphone based system and our TTS system is unfair, and the results of the test cannot be used to verify the viability of the syllable approach. However, we have made experiments that suggest the possibility to increase the segmental quality using syllables extracted from carefully articulated spoken words.

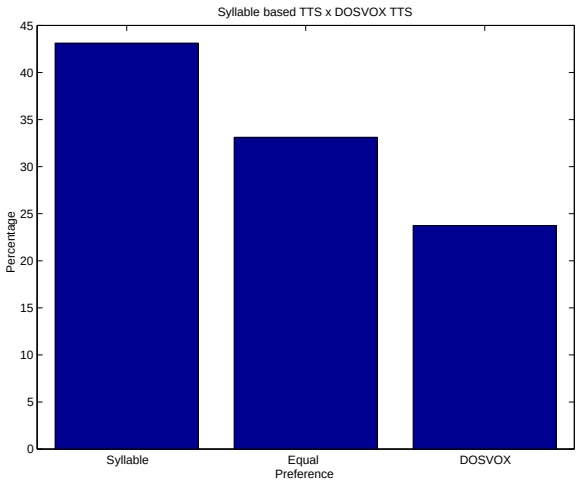

Figure 8: Comparison between the TalkActive system and DOSVOX.

Acknowledgment

The authors would like to thank several people that highly contributed for the current version of the TalkActive system, including: Rosana C. de Oliveira, for determining the necessary syllabic database for the Brazilian Portuguese; Verochile da Silva Jr., for building up the database creating the syllable units as .WAV files; and Rodrigo C. Tor-

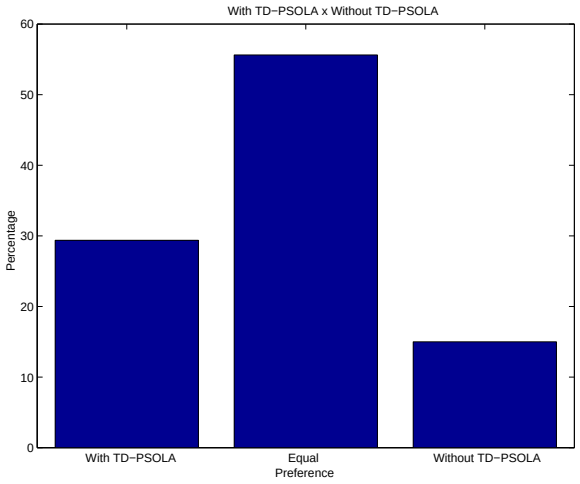

Figure 9: Comparison between the TalkActive system with and without TD-PSOLA.

res for creating the first operational version of the TalkActive system.

6 Conclusion

This paper has described the early stage of a new TTS system for the Brazilian Portuguese based on syllabic units. A complete block diagram of the system development was presented. The current stage uses TD-PSOLA concatenation algorithm although no prosody is being incorporated to the final acoustic signal. Preliminary intelligibility tests showed that the intelligibility does not increase significantly when synthesizing phrases, suggesting the importance of context information in the extraction of units. Comparison tests also showed a slight preference for the utterances synthesized using TD-PSOLA when compared to direct concatenation.

References

[1] T. Dutoit. *An introduction to text-to-speech synthesis*. Kluwer Academic Publishers, 1997.

[2] C. Hamon E. Moulines & F. Charpentier. A diphone synthesis system based on time-domain modifications fo speech. *Proceedings of the ICASSP*, pages 238–241, 1989.

[3] E. Moulines & F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, (9):453–467, 1990.

[4] A. R. Franco. Sintetizador paramétrico para a língua portuguesa. UFRJ, 1999. Projeto Final.