

# On the generation of synthetic disfluent speech: Local prosodic modifications caused by the insertion of editing terms

Jordi Adell<sup>1</sup>, Antonio Bonafonte<sup>1</sup> and David Escudero-Mancebo<sup>2</sup>

<sup>1</sup>TALP Research Center, Universitat Politècnica de Catalunya, Barcelona (Spain)

<sup>2</sup>ECA-SIMM Laboratory, Universidad de Valladolid, Valladolid (Spain)

{jadell, antonio}@gps.tsc.upc.es, descuder@infor.uva.es

## Abstract

Disfluent speech synthesis is necessary in some applications such as automatic film dubbing or spoken translation. This paper presents a model for the generation of synthetic disfluent speech based on inserting each element of a disfluency in a context where they can be considered fluent. Prosody obtained by the application of standard techniques on these new sentences is used for the synthesis of the disfluent sentence. In addition, local modifications are applied to segmental units adjacent to disfluency elements. Experiments evidence that duration follows this behavior, what supports the feasibility of the model.

**Index Terms:** speech synthesis, disfluent speech, prosody, disfluencies.

## 1. Introduction

Speech synthesis has already reached a high standard of naturalness [1], mainly due to the use of effective techniques such as unit selection-based systems or other new rising technologies [2] based on the analysis of huge speech corpora. By now, the main application of speech synthesis has been focused on read style speech since it can be considered that read style is the most generalist style to be extrapolated to any other situation. But nowadays, and even more in the future, applications of text to speech (TTS) systems (e.g. automatic film dubbing, robotics, dialogue systems, or multilingual broadcasting) demand for a variety of styles since the users expect the interface to do more than just reading information.

If synthetic voices want to be integrated in future technology, they must simulate the way people talk instead the way people read. Synthetic speech must become conversational-like rather than reading-like speech. Therefore, we claim it is necessary to move from *reading* to *talking* speech synthesizers. Both styles differ significantly from each other due to the inclusion of a variety of prosodic resources affecting the rhythm of the utterances. Disfluencies are one of these resources defined as phenomena that interrupt the flow of speech and do not add propositional content to an utterance [3]. Despite the lack of propositional content, they are cues about what is being said [4]. Disfluencies are very frequent in every day speech [5] so that we hypothesize the need to include these prosodic events to get closer to talking speech synthesis.

The study of disfluencies has been approached from several disciplines, mainly phonetics [6, 5], psycholinguistics [7, 8] and speech recognition [9, 10]. Different approaches model disfluencies according to their specific interest. The use of disfluencies in TTS systems brings additional considerations leading us to introduce an alternative model. This model, in contrast with other approaches used in TTS such as [11] or [12], considers

the potential fluent sentences associated with the disfluent utterance in conjunction with the local modifications produced by the insertion of the editing term. These local modifications can affect speech prosody and the quality of the original delivery. We show the relevance of these local modifications by studying the impact of disfluencies on the duration of the syllables that surround the editing terms.

First we introduce the disfluent speech generation model. Second experimental procedure for the application of this model is presented reflecting the impact on the duration of the syllables that surround the editing terms. Third, we discuss the future work to be done in this ongoing research and the paper ends with conclusions.

## 2. Synthetic disfluent speech model

The synthesis of disfluent speech in the framework of unit-selection speech synthesis presents a series of drawbacks. First of all, most of the existing unit selection systems have a closed inventory which does not contain disfluencies at all. Therefore, the machine learning techniques that are usually applied to analyze prosody in TTS are not able to automatically model these phenomena from data. Secondly, not only the prosodic models but also the text analysis models (e.g. POS tagging), expect sentences to have a rigid structure based on the concatenation of syntactic, accent and intonation groups. When the fluency of an utterance is broken, its structure is also broken. This makes difficult for standard models to predict prosodic parameters accurately. In addition, disfluent speech synthesis requires the use of new segmental units, that are not defined in standard phone-sets (e.g. fillers or interrupted phones). In this section, we present a model which, in one hand, tries to take advantage of standard prosodic models trained from fluent speech and on the other hand, takes into account local modifications at the point where fluency is broken.

In our model, three different elements are taken into account for the generation of any given disfluent sentence (DS). First, the original sentence (OS) that is the sentence expected to be uttered before the disfluency is entered. Second, the target sentence (TS) that is the one expected to be uttered in the case the disfluency is not present; and third, the Editing Term (ET). According to the terminology described in [13] ET is the cue mark of the disfluency (e.g. filled pauses). Let us consider the following example from [5]: *Go from left to mmm from pink again to blue* whose disfluency elements can be identified as follows:

$$Go\ RM\{from\ left\ to\} \overset{IP}{\downarrow} ET\{mmm\},\ RR\{from\ pink\ again\ to\} blue$$

being *RM* (*Reparandum*), *RR* (*Repair*), *ET* (*Editing Term*)

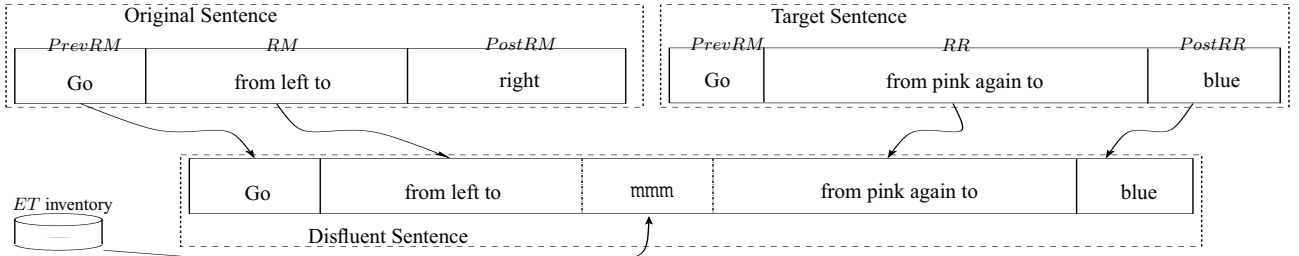


Figure 1: Synthetic disfluent speech generation process applied to a sample sentence

and *IP* (Interruption Point) according to [13]. This sentence is somehow related to the sentences: *Go from left to right* and *Go from pink again to blue* that are the OS and TS sentences respectively; and the ET is *mmm*. These relations are modelled as follows:

$$\begin{aligned}
 DS &= PrevRM|RM|ET|RR|PostRR \\
 OS &= PrevRM|RM|PostRM \\
 TS &= PrevRM|RR|PostRR
 \end{aligned}$$

where *PrevRM* is the part of the sentence preceding the *RM*; *PostRM* is the part that follows the *RM* and *PostRR* is the part following the *RR*. Note that *PostRM* exists only in the OS, since this part of the sentence is not actually uttered given that the *RR* is uttered instead. In the example presented above *PostRM* could be *right*.

Given a DS, a unit selection TTS system is supposed to be able to generate correctly the corresponding OS and the TS. A set of *ET*s can be also selected if the TTS unit inventory has been built from a database which also includes a set of disfluent sentences. There exist evidences of the local modifications of the acoustic features of *RM*, *PostRM* and *PrevRM* (especially if *RM* is empty) [14, 12] caused by the insertion of the Editing Term.

Our disfluent speech generation proposal operates in three stages (Figure 1). First, it uses OS to obtain the prosodic parameters related to *PrevRM* and *RM*, and TS to obtain the ones related to *RR* and *PostRR*. These prosodic parameters are the ones used to guide the unit search in the inventory. In the second step, it obtains the *ET* from the inventory. Finally, it applies local modifications to the syllables adjacent to the *ET*. These modifications correspond to the local deviations from fluent prosody that might appear at joins between the elements described in this section (*PrevRM*, *RM*, ...).

Rhythm is one of the prosodic variables that might deviate from the values predicted by a no disfluent prosodic model. In the next section we will show how the rhythm of the disfluent sentences keep an overall rhythm that equals to fluent sentences and that deviations from this rhythm occur at the segmental units adjacent to the *ET*.

Table 1: Disfluency elements definition for each of the studied disfluency type.

type	RM	ET	RR
hesitation	$\emptyset$	$\emptyset$	$\emptyset$
filled pause	$\emptyset$	filler (e.g. mmm)	$\emptyset$
repetition	1 <sup>st</sup> utt.	$\emptyset$	2 <sup>nd</sup> utt.

Three types of disfluencies have been studied: hesitations, repetitions and filled pauses. Model elements can be defined for each type as shown in Table 1.

### 3. Local modifications of the rhythm in disfluent sentences

Usually two main language categories are considered in the literature for describing language rhythm: *Accent-timed* and *syllable-timed*. Both categories relate to the principle of isochrony by which phonological units tend to be equally spaced in time [15]. Spanish, as well as all other Romance languages, such as Catalan or French; is considered to be syllable-timed [16, 17]. Since in this work we will only deal with Spanish, it is appropriate to measure the rhythm of a supra-segmental unit (e.g. a sentence) as the mean syllable duration in this unit. As a reference see Table 2 for duration measures calculated over all data.

Table 2: Syllable duration means and 99% confidence interval lower and upper bounds.

	mean	lower bound	upper bound
non-Accented	105ms	102ms	108ms
Accented	136ms	132ms	140ms
pre-Pausal	222ms	210ms	236ms
All syllables	123ms	120ms	125ms

The corpus used here is a selection of sentences from the corpus developed under the LCSTAR European project. It was recorded in a laboratory and it collects dialogs of two people that are requested to accomplish a task by phone. Communication was semi-duplex so that the database is recorded in turns [18]. Although it has been recorded in a laboratory speech is spontaneous because speakers were not guided. Speakers utter disfluencies naturally and frequently because they need to plan their turns at the time they perform the tasks. 100 sentences were selected from four different speakers (3 male, 1 female) to contain as much disfluencies as possible: 133 filled pauses, 71 repetitions and 65 hesitations. Phonetic segmentation was performed automatically and manually corrected.

A set of measurements has been calculated for each element described in section 2 (Table 3). Note that  $Syl_{p1}$  is the last syllable of *PrevRM* and  $Syl_{n1}$  the first one of *PostRR*. We expected to find that rhythm keeps constant across the whole sentence and that the changes are produced at ET boundaries according to the proposed model.

The study is based on a set of box-plot drawings [19] which indicate how the rhythm remains constant across the sentence and that it is only modified in syllables adjacent to the *ET*. In order to make interpretation easier, variable are listed in the same order they appear in the sentence.

The Least Significant Difference (LSD) test also known as multiple t-test has been used to compare mean of distributions in order to establish which differences are significant [20]. All significances presented in this paper are at 99% confidence level.

Table 3: List of variables used to study rhythm variations due to the presence of a disfluency.

Variable	Definition
$R_s$	Mean syllable duration of a whole sentence.
$Rw_{pN}$	Mean syl. duration of the $N$ th word previous to $RM$ .
$Syl_{pN}$	Duration of the $N$ th syllable previous to the $RM$ .
$R_{RM}$	Mean syllable duration in the $RM$
$R_{ET}$	Mean syllable duration in the $ET$
$R_{RR}$	Mean syllable duration in the $RR$
$Syl_{nN}$	Duration of the $N$ th syllable next to the $RR$ .
$Rw_{nN}$	Mean syllable duration of the $N$ th word next to the $RR$ .

### 3.1. Hesitations

They are composed of a syllable lengthening. In Figure 2 it can be observed how there is no significant difference between rhythm of words before the hesitation ( $Rw_{p1}$  and  $Rw_{p2}$ ) and  $R_s$ . Neither there is for following words ( $Rw_{n1}$  and  $Rw_{n2}$ ). However, the lengthened syllable (i.e.  $Syl_{p1}$ ) has a wider range and larger mean.

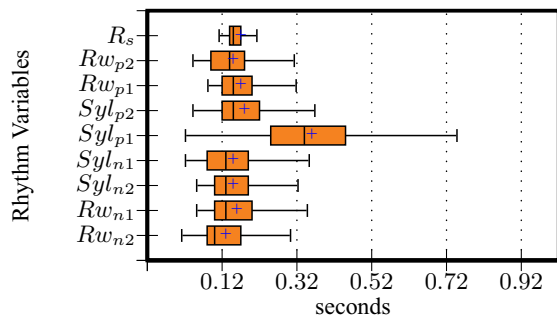


Figure 2: Box-plot of variables for Hesitations

The *LSD* test have shown that the only significant difference is the one between  $Syl_{p1}$  and all other variables. *LSD* intervals indicate that the lengthened syllable is between 141ms and 239ms longer than the mean syllable duration in the sentence. Local modifications in case of hesitations can be summarized as a syllable lengthening while all other syllables would remain the same.

### 3.2. Repetitions

They are composed of two main elements: both repeated utterances, namely  $RM$  and  $RR$ . If we look at Figure 3 it shows that while  $R_{RR}$  is similar to  $R_s$ ,  $R_{RM}$  and  $Syl_{p1}$  are larger than  $R_s$ . These findings agree with previous published studies [14, 21].

*LSD* test shows that there is a significant difference between means of  $R_{RM}$  and  $Syl_{p1}$  and the rest of features. We can conclude that for repetitions  $RM$  is uttered more slowly than the rest of the sentence and that last syllable of  $PrevRM$  ( $Syl_{p1}$ ) is systematically lengthened. *LSD* also indicates that mean of  $Syl_{p1}$  is between 10ms and 88ms longer than  $R_s$ . Also  $R_{RM}$  is between 70ms and 148ms larger than mean syllable duration of the sentence.

We must take into account that the reason for this lengthening might be the existence of a silent pause just before the repetition. However, only 17% of instances contain a silent pause there (i.e. pre-pausal lengthening). If we remove these instances from data, there is still the same significant difference. Then, we can conclude that these rhythm variations are due to the presence of a repetition rather than to the presence of a silent pause.

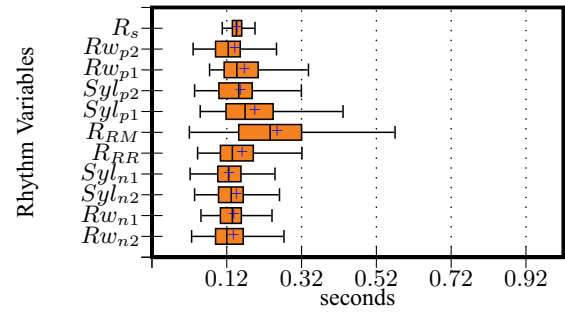


Figure 3: Box-plot of variables for Repetitions

### 3.3. Filled Pauses

Filled Pauses can not be strictly considered syllables. However, filler duration ( $R_{ET}$ ) can be consistently compared with syllable durations. Figure 4 presents the box-plot for filled pauses. It can be observed that all measures except  $R_{ET}$ ,  $Syl_{p1}$  and  $Rw_{p1}$  follow a similar distribution to the  $R_s$ . This is also supported by the *LSD* test which indicates that there is no significant difference between any other variable means.

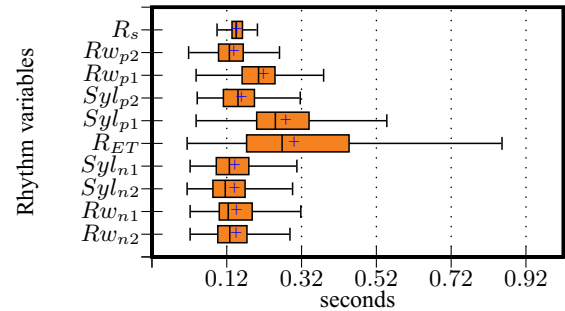


Figure 4: Box-plot of variables for Filled Pauses

Given the *LSD* test, the pre-filler syllable ( $Syl_{p1}$ )—in contrast to pre-pausal—is between 100ms and 161ms longer than  $R_s$ . Filled Pause duration has a mean value of 300ms and 185ms standard deviation. This implies that the length of filled pauses clearly exceeds the length range of pre-pausal syllables. (see Table 2)

Again, for this type of disfluency, the global sentence rhythm is not modified, only local rhythm alterations appear in form of pre-filler syllable lengthening. This supports the use of the proposed synthesis model.

## 4. Discussion and Future Work

Experimental evidences presented in last section support the model presented in section 2. For each disfluency type a set of local modifications have been observed in real data. These modifications must be applied after joining different elements from the OS and the TS. In summary, the synthesis of hesitations requires a lengthening of last syllable of  $PrevRM$ .  $RR$  in repetitions must be synthesized at same syllable rate than the overall sentence rate but  $RM$  has to be slowed down, also last syllable of  $PrevRM$  has to be longer. When synthesizing filled pauses last syllable of  $PrevRM$  has to be lengthened.

In the present paper we have identified rhythm variations, but for speech synthesis these variations have to be predicted. For this purpose there are several options. A couple of this options have been tested in previous works such as a rule set [22]

or regression models [23].

The proposed model has been implemented in our synthesizer [24] for a few set of disfluencies (filled pauses and repetitions). Informal tests have shown that local modifications of segmental durations allowed the inclusion of disfluencies without a degradation of the standard of naturalness of our system while doing so without taking these local variations into account resulted in less natural speech.

In the future we plan to investigate local variations in pitch contours. Further research will be carried out concerning other prosodic parameters such as energy or voice quality. Our aim is to study variations of a set of prosodic parameters to synthesize disfluencies by taking advantage of fluent models plus local modification of segmental units adjacent to the *ET*.

Furthermore, a different problem that needs to be faced up in disfluent synthesis is the disfluency prediction. This not *how* to synthesize disfluencies but *where* to generate *which* type of disfluency. This is a complex task out of the scope of our ongoing research. Also, the identification of a proper *PostRM* element is an open issue. Nevertheless, the focus of the present work is on applications such as automatic dialogue systems, where this information can be given to the synthesis system; or spoken translation, where cues about the position of disfluencies can be found in the original speech.

## 5. Conclusions

In this paper, it has been presented a model for the prosody of disfluencies based on the definition of three sentences: Disfluent sentence (DS), Original Sentence (OS) and Target Sentence (TS). OS and TS provide a context where Reparandum and Repair respectively are pronounced fluently. The Editing Term (*ET*) is inserted from the speech synthesis inventory and local prosodic modifications are applied at segmental units adjacent to it. We claim that this model can be used for the generation of synthetic disfluent speech taking advantage from already existing prosodic models of fluent speech.

Local modifications of rhythm have been studied for three types of disfluencies: hesitations, repetitions and filled pauses. For these three types it has been probed that rhythm is only significantly altered in the *ET* and the ends of the Reparandum.

These results support the feasibility of the proposed model. Future work has been presented as further research on local modifications of new prosodic parameters such as pitch or voice quality.

## 6. Acknowledgements

This work has been partially funded by the Spanish Government under the AVIVAVOZ project (TEC2006-13694-C03)

## 7. References

- [1] A. Aaron, E. Eide, and J.F. Pitrelli, "Conversational computers," *Scientific American*, vol. 292, no. 6, pp. 64–69, June 2005.
- [2] Mark Fraser and Simon King, "The blizzard challenge 2007," in *Proceedings of the Blizzard Challenge*, 2007, number 1.
- [3] Jea E. Fox Tree, "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech," *Journal of Memory and Language*, vol. 34, no. 6, pp. 709–738, December 1995.
- [4] Michiko Watanabe, Keikichi Hirose, Yasuharu Den, and Nobuaki Minematsu, "Filled pauses as cues to the complexity of following phrases," in *Proc. of Eurospeech*, September 2005, pp. 37–40, Lisbon, Portugal.
- [5] Shu-Chuan Tseng, *Grammar, Prosody and Speech Disfluencies in Spoken Dialogues.*, Ph.D. thesis, Department of Linguistics and Literature, University of Bielefeld, April 1999.
- [6] Elizabeth Shriberg, "Phonetic consequences of speech disfluency," in *Proc. of International Congress of Phonetic Science. Symposium on The Phonetics of Spontaneous Speech* (S. Greenberg and P. Keating, organisers), 1999, vol. 1, pp. 619–622, San Francisco.
- [7] Daniel C. O'Connell and Sabine Kowal, "The history of research on the filled pause as evidence of the written language bias in linguistics (linell, 1982)," *Journal of Psycholinguistic Research*, vol. 33, no. 6, pp. 459–474, November 2004.
- [8] Herbert H. Clark, "Speaking in time," *Speech Communication*, vol. 36, no. 1-2, pp. 5–13, January 2002.
- [9] Elizabeth Shriberg, Rebecca Bates, and Andreas Stolke, "A prosody-only decision-tree model for disfluency detection," in *Proceedings of EUROSPEECH*, 1997.
- [10] Masataka Goto, Katunobu Itou, and Storu Hayamizu, "A real-time filled pauses detection system for spontaneous speech recognition," in *Proc. of EUROSPEECH*, Budapest, Hungary, 1999, pp. 227–230.
- [11] Shiva Sundaram and Shrikanth Narayanan, "An empirical text transformation method for spontaneous speech synthesizers," in *Proc. of EUROSPEECH*, Geneva, Switzerland, September 2003, pp. 1221–1224.
- [12] R. Carlson, K. Gustafson, and E. Strangert, "Cues for hesitation in speech synthesis," in *Proceedings of Interspeech 06*, Pittsburgh, USA, 2006.
- [13] Elizabeth Ellen Shriberg, *Preliminaries to a Theory of Speech Disfluencies*, Ph.D. thesis, Berkeley's University of California, 1994.
- [14] Elizabeth E. Shriberg, "Acoustic properties of disfluent repetitions," in *Proc. of International Conference on Phonetic Sciences (ICPhS)*, 1995, vol. 4, pp. 384–387, Stockholm, Sweden.
- [15] Manuel Almeida, "Organización temporal del español: el principio de isocronía," *Revista de Filología Románica*, vol. 1, no. 14, pp. 29–40, 1997, Madrid.
- [16] Guillermo Andrés Toledo, *El ritmo en el español : estudio fonético con base computacional*, Number 361 in Biblioteca románica hispánica ; II. Estudios y ensayos. Gredos, 1988.
- [17] Mar Carrió and Antonio Ríos, "Compensatory shortening in spanish spontaneous speech," in *Proceedings of ESCA Workshop on Phonetic and Phonology of Speaking Styles.*, September 1991, vol. 16, pp. 1–5, Barcelona, Spain.
- [18] David Conejero, Jesús Giménez, Victoria Arranz, Antonio Bonafonte, Neus Pascual, Núria Castell, and Asunción Moreno, "Lexica and corpora for speech-to-speech translation: A trilingual approach," in *Proc. of Eurospeech*, September 2003.
- [19] Michael Frigge, David C. Hoaglin, and Boris Iglewicz, "Some implementations of the boxplot," *The American Statistician*, vol. 43, no. 1, pp. 50–54, February 1989.
- [20] D.J. Saville, "Multiple comparison procedures: The practical solution," *The American Statisticians*, vol. 44, no. 2, pp. 174–180, May 1990.
- [21] Jordi Adell, Antonio Bonafonte, and David Escudero, "Disfluent speech analysis and synthesis: a preliminary approach," in *Proc. of 3th International Conference on Speech Prosody*, May 2006, Dresden, Germany.
- [22] Jordi Adell, Antonio Bonafonte, and David Escudero, "Filled pauses in speech synthesis: towards conversational speech," *Text, Speech and Dialogue, 10th International Conference, LNAI*, vol. 1, pp. 358–365, September 2007, Springer Verlag.
- [23] Jordi Adell, Antonio Bonafonte, and David Escudero, "Statistical analysis of filled pauses' rhythm for disfluent speech synthesis," in *Proc. of the 6th IWSS*, Bonn, Germany, August 2007.
- [24] Antonio Bonafonte, Pablo Daniel Agüero, Jordi Adell, Javier Pérez, and Asunción Moreno, "Ogmios: The upc text-to-speech synthesis system for spoken translation," in *TC-STAR Workshop on Speech-to-Speech Translation*, June 2006.