# Modelling Filled Pauses Prosody to Synthesise Disfluent Speech

Jordi Adell, Antonio Bonafonte
TALP Research Center
Universitat Politècnica de Catalunya
www.talp.cat

David Escudero-Mancebo
ECA-SIMM Laboratory
Universidad de Valladolid
www.infor.uva.es

## Abstract

In the present paper we present a new approach to the synthesis of filled pauses since they are as frequent as most frequent words in conversational speech. The problem is tackled from the point of view of disfluent speech synthesis. Based on the synthetic disfluent speech model, we analyse the features that describe filled pauses and propose a model to predict them. The model was implemented and perceptually evaluated with successful results.

**Index Terms**: speech synthesis, disfluent speech, filled pauses

## 1. Introduction

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. Therefore, we claim it is necessary to move from *reading* to *talking* speech synthesisers. 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 [1]. Despite the lack of propositional content, they may give cues about what is being said to the listener [2]. Disfluencies are very frequent in every day speech [3] so that it is possible to hypothesise the need to include these prosodic events to approximate to talking speech synthesis.

The study of disfluencies has been approach from several disciplines, mainly phonetics [4, 3], psycholinguistics [5, 6] and speech recognition [7, 8]. 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 others approaches used in TTS such as [9] or [10], considers the potential fluent sentences associated with the disfluent sentence and the local modifications produced when the editing term is inserted. These local modifications can affect speech prosody and the quality of the original delivery. We showed the relevance of this local modifications by studying the impact of disfluencies on the duration of the syllables surrounding the editing term of disfluent sentences [11].

In this paper we first show the relevance of filled pauses (FP) by analysing its frequency of appearance in conversational corpora. Second, we review the disfluent speech generation model based on the prosody of constituent fluent sentences plus local modifications around the editing term[11]. Therefore, in section 4, we analyse analyse which are the local modifications in the case of FP and propose a model to predict them. Finally, we present the results of a perceptual evaluation of the models and some conclusions.

## 2. The relevance of filled pauses in conversational speech

We have analysed the presence of filled pauses in two different corpora, none of them is read speech. The LC-STAR and TC-STAR corpora differ in style but they both have conversational style speech, and a significant number of disfluencies as a consequence.

The LC-STAR corpus was developed under the LC-STAR European project. This corpus consists of 55 hours of speech, 580,000 words and ∼100 speakers embracing Spanish and Catalan. The corpus was recorded in a laboratory, and it comprises dialogues of two people who are requested to accomplish a task by phone. The tasks belong to four scenarios, namely *i) a hotel, ii) a travel agency, and iii) a tourism office and iv) railway/airline company*. Communication was semi-duplex so that the database is recorded in turns. Speakers utter disfluencies naturally and frequently because they need to plan their turns at the time they perform the tasks [12].

FPs represent 1.7% (i.e., 5,723 utterances) of the words in the Spanish corpus and 1.3% (i.e., 3,292 utts) in Catalan. The relevance of these numbers can be illustrated by comparing them with the frequency of the most common word in the Spanish corpus. It is the preposition *"de"*[1] and represents 3.4% of the words in the corpus (i.e., 11,038 utts), and the FP is the 7th most frequent *word*. Additionally, the word *"de"* is the most frequent in Catalan and represents 2.6% of the words in the corpus, and the FP is the 10th most frequent. FPs are as frequent as the preposition *"a"*[2], which appears 4,842 times (1.5%) in the Spanish corpus and is equally frequent as the verb *"és"*[3] in Catalan (1.5%). These findings clearly illustrate the importance of disfluencies in these corpora, showing its spontaneous nature. FPs are probably the most studied type of disfluency, and they are as frequent in spontaneous contexts as some common non-content words.

The TC-STAR corpus consists of parliamentary speeches. It contains about 9 hours and 70,000 words in Spanish and English. Speakers are either members of the parliament or interpreters. The Spanish corpus was recorded from the Spanish parliament emission and the English, from the European parliament emission. It contains 1.5% FPs.

Despite the fact that parliamentary speech may be better planned and more carefully released, FPs are present in a similar amount to those in more conversational-like speeches (for Spanish, it is 1.5% in this corpus, and 1.7% in the LC-STAR corpus). This result supports the hypothesis that some disfluencies should not be considered mistakes but solutions to in-

[1] **de**: *prep.* from; *prep.* of

[2] **a**: *prep.* to

[3] **és**: *v.i.* is (to be)

time discourse planning [5], even when no conversational-like speech is produced.

Filled Pauses are as frequent as the most frequent words in the LC-STAR corpus as well as in the TC-STAR corpus. Theses findings indicate that it is not possible to ignore FPs when conversational speech is the goal.

## 3. Synthesis of disfluent speech

In [11] we proposed to generate any disfluent sentence taking into account three different elements. These elements are taken into account for the generation of any given disfluent sentence (DS)[13]. First, the original sentence (OS) is the one that was originally planned. Second, the target sentence (TS) is what would be uttered if no disfluency was present; and third, the Editing Term (ET). According to the terminology described in [14] ET is the cue mark of the disfluency (e.g. filled pauses). Let us consider the example sentence: *Go from left to mmm from pink again to blue* from [3] whose disfluency elements can be identified as follows:

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

being $RM$ (*Reparandum*), $RR$ (*Repair*), $ET$ (*Editing Term*) and $IP$ (*Interruption Point*) the disfluency elements defined in [14]. This sentence is somehow related to sentences: *Go from left to right* and *Go from pink again to blue*, the OS and TS respectively; and the ET is mmm. We can decompose these sentences as follows:

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

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

A standard TTS system is supposed to be able to generate the OS and TS from the DS. 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 that the insertion of the Editing Term implies local modifications of the acoustic features of $RM$, $PostRM$ and $PrevRM$ (especially if RM is empty) [10]

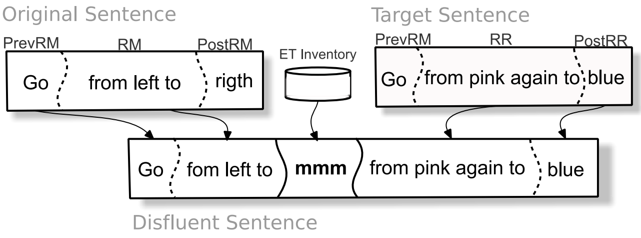

Figure 1: Synthetic disfluent speech generation process applied to a sample sentence

Our disfluent speech generation proposal operates in three stages (Figure 1). First, it uses OS to obtain 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 might appear at joins between elements described in this section ($PrevRM$, $RM$, ...).

In a previous work we showed evidences that fluent and disfluent sentences differ in these local variations [11]. Therefore, local models can be used together with standard models trained on fluent speech. In the following sections we will define the features that model this local variations and also propose a model to predict them.

## 4. Analysis and modelling of filled pauses

In this section first we present the data used in the present study. Then, we analyse filled pauses to build a list of parameters that can completely describe local prosodic variations. Afterwards, proper modelling of these parameters will lead to the synthesis of disfluent speech.

### 4.1. Data

The corpus used here is a selection of sentences from the corpus developed under the LC-STAR European project (Section 2). 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.

### 4.2. Prosodic description of filled pauses

FPs have been considered as pauses containing non-verbal sounds instead of silences [6]. First, we compared their duration with the one of silent pauses (SP). If the FPs duration followed the same distribution than SPs at the middle of sentences, a standard model trained on fluent speech could be used to predict their duration.

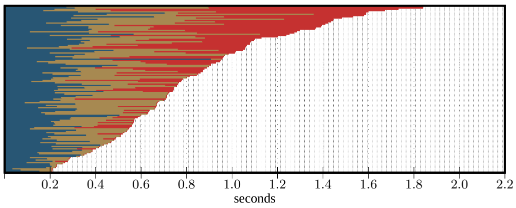

Figure 2: Length of previous syllable (left), filler (center) and the silence (right). The total length of the bar shows the total FP duration.

However, FPs are placed at points where SPs are usually not placed, and so their features might not be consistent; thus, both approaches must be compared. Additionally, because of their value as *pauses*, we expect a pre-pausal syllable lengthening affecting syllables before FPs, as has already been described in the literature [8]. This lengthening must, of course, be one of the features describing filled pauses.

The pre-pausal syllable duration mean for FPs in our corpus ($D_{syl}^{-1}$) is 278ms (±18ms) and for SPs is 222ms (±13ms). The difference between both means is statistically significant ($p = 0.020$). In addition, the difference between FP duration $D_{FP}$ and SP duration $D_{SP}$ distribution means is also statistically significant ($p \sim 10^{-7}$). Therefore, it is not possible to consider SPs equal to FPs, and a specific model for FPs must be proposed.

In Figure 2, FPs duration is depicted and sorted by length (shortest at the bottom and longest at the top). For each FP and from left to right, $D_{syl}^{-1}$, $D_{fil}$ and $D_{sil}$ are depicted, which are the durations of the previous syllable, the filler and the silence, respectively.

We can observe that nor the silence not the filler are present for short filled pauses. Fillers appear only for filled pauses longer than ≈ 250ms. This picture suggests that syllables cannot be longer than ≈ 400ms, and, when the speaker needs a longer filled pause (i.e. needs more time for planning), a filler is added at the end. This is due to the human limitation to make sounds that last for a long time. In this corpus, the limit for syllables is ∼ 500ms. Finally, if this is not enough, silence is inserted between $Syl_{-1}$ and the filler. This behaviour suggests that speakers have an estimation of the time they need to re-plan speech, since the silence is placed before the filler, and the second has a limit in duration.

Then, the behaviour of the elements that describe filled pauses with respect to their total duration can be modelled by a piece-wise linear function:

$$D_{syl} = K \quad if\ D_{FP} < K \tag{1}$$

$$D_{fil} = \begin{cases} D_{FP} - D_{syl} & if\ D_{FP} > K \\ D_{fil}^{max} & if\ D_{FP} < K + D_{fil}^{max} \end{cases} \tag{2}$$

$$D_{sil} = \begin{cases} 0 & if\ D_{FP} < D_{fil}^{max} + K \\ D_{FP} - K - D_{fil}^{max} & if\ D_{FP} > D_{fil}^{max} + K \end{cases} \tag{3}$$

where $D_{FP}$ is the filled pause duration. $D_{fil}^{max}$ is the maximum possible length for a filler, and $D_{syl}$ the syllable duration and $D_{sil}$ the duration of the silence.

#### 4.2.1. *Pitch contour*

An important feature to take into account when modelling pitch in FPs is slope. For example, syllables at the end of a sentence are mainly pronounced with a descending pitch slope, but an interrogative sentence ends with a rising pitch. Therefore, it is reasonable to investigate whether there is a standard pitch slope for FP or whether, on the other hand, it depends on other aspects, such as semantics or syntax [15].

We measured the slope of the pitch contour as the difference between the pitch evaluated as $F0D = F0E - F0B$, where $F0E$ and $F0B$ are $F0$ at the end and the beginning of the segmental unit. The mean $F0D$ for fluent syllables is $1.83 \pm 0.63$, while for fillers it is $-11 \pm 5.5$ at a 95% confidence level. The difference in means is significant with $p << 0.01$. This clearly shows that the pitch contour of fillers tends to be decreasing compared to that of fluent syllables.

If the speech synthesis system used to generated FP can apply a continuous contour to each segmental unit, then this decreasing slope has to be taken into account. In contrast, many TTS systems do not modify the segments selected by the unit selection algorithm. But the prosody description of the target filler (duration and f0 slope) can be used to select fillers with the appropriated prosodic features. Thus, the unit slope remains the same. For these systems, it is not worthwhile to deal with this issue. Nevertheless, these findings can be used to detect undesired filler units in the inventory (i.e., the ones that do not contain a decreasing pitch contour), so that only fillers with decreasing pitch are used.

We can also consider a different approach. Let us define $\bar{F0}$ as a baseline prediction for the $F0$ mean of the filler. This prediction is a linear interpolation between the mean pitch values of previous and following syllables:

$$\bar{F0} = \frac{F0^{-1} + F0^{1}}{2} \tag{4}$$

If we calculate the difference between $\bar{F0}$ and the mean pitch value of the filler $F0_{fil}$, we can see whether there is a systematic relation between both values.

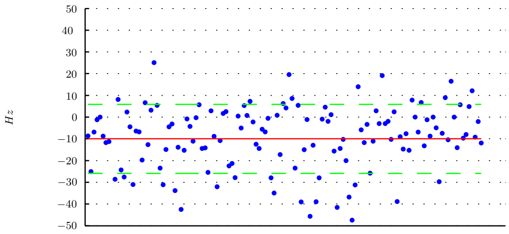

Figure 3: $F0_{fil} - \bar{F0}$ for all fillers in the database. Plain line identifies the mean and dashed lines ± one standard deviation from the mean.

In Figure 3, we can observe that the difference between both values is $\sim -10Hz$, and, for 90% of cases, this value is negative. This implies that the filler pitch is systematically lower than the sentence pitch contour. Therefore, we will be able to model filler pitch by lowering the pitch from the pitch of the contour defined by adjacent syllables. In both approaches, we are taking the decreasing slope into account.

### 4.3. Prosodic model for Filled Pauses

In previous section we claimed that it is possible to describe FPs by setting their duration ($D_{FP}$). It consists of the sum of the previous syllable($D_{syl}$), the silence ($D_{sil}$) and the filler ($D_{fil}$) durations. Additionally, a pitch decreasing factor must be applied. The distribution of $D_{FP}$ can be modelled by a piecewise function, and the pitch factor can be modelled by linear regression.

#### 4.3.1. *Constituents and duration*

The necessary parameter to model filled pauses are $D_{sil}$, $D_{syl}$, $D_{fil}^{max}$ and $D_{fil}$. This model has to predict three duration values. Given that the duration of the syllable previous to the FPs can be considered a constant, the maximum length of FP without silence is limited by the maximum length of the filler ($D_{fil}^{max}$). Silences appear only if $D_{FP}$ is larger than a certain value. This value corresponds to the maximum $D_{FP}$ when $D_{sil}$= 0. Since this maximum value corresponds to a constant plus $D_{fil}$, we can obtain $D_{fil}^{max}$ from it.

Equations 5 6 and 7 correspond to regression the lines that can be used to predict the local parameters that described filled pauses from their global duration.

$$D_{syl} = 0.277 \tag{5}$$

$$D_{fil} = 0.184 * D_{FP} + 0.130 \tag{6}$$

$$D_{sil} = 0.717 * D_{FP} - 0.320 \tag{7}$$

From these equations, let $D_{sil}$= 0, since $D_{FP}^{max}|_{D_{sil}=0} = \frac{0.320}{0.717} = 0.446$, then $D_{FP}^{max}|_{D_{sil}=0} = D_{syl} + D_{fil}^{max}$, and: $D_{fil}^{max} = 0.446 - 0.277 = 0.169$.

Therefore, the proposed model will distribute the duration of a FP, giving 277ms to the syllable previous to the filler and the rest up to 169ms will be the duration of the filler itself. Finally, if there is still duration to be covered, the rest will be used to insert a silence in between the syllable and the filler.

### 4.3.2. Pitch contour

As stated in previous sections, the pitch of a FP is systematically lower than its context (i.e., the previous and following syllables). Therefore, the proposed approach consists of calculating the $F0$ value using the $F0$ values of the preceding ($F0^{-1}$) and following ($F0^{1}$) syllables with a regression model: $a * F0^{-1} + b * F0^{1} + c$. However, taking into account that the pitch of filled pauses is systematically lower than the interpolation line between $F0^{-1}$ and $F0^{1}$, a new feature $\bar{F}0 = \frac{F0^{-1}+F0^{1}}{2}$ can be used to predict $F0_{fil} = a * \bar{F}0 + c$. The use of this feature will make coefficients $a$ and $c$ meaningful. Since $a$ is a decreasing factor, that gives an idea of how low the pitch of the FP is with respect to the fluent context, and $c$ is an offset that models a systematic decrease in pitch that does not depend on the context. The linear regression of $F0_{fil}$ with respect to $\bar{F}0$ gives the following result:

$$F0_{fil} = 0.99 * \bar{F}0 - 7.72 = 0.99 * \frac{F0^{1} + F0^{-1}}{2} - 7.72 \quad (8)$$

Again, this is a speaker independent model, but speaker dependent models with values estimated for every single speaker would generate more accurate predictions.

## 5. Evaluation

The proposed model has been implemented in our TTS system [16]. The system selects units from a database of 10h speech recorded by a professional speaker. In addition, we recorded 57 filler utterances as the segmental units selected as fillers.

Our goal was to create a system able to generate filled pauses without degrading the quality of the existing system. If we achieved this we could say that our system can generated filled pauses thanks to the synthetic disfluent speech model. To evaluated whether we achieved it we carried out a perceptual evaluation.

The evaluation consisted of a Mean Opinion Score (MOS) test. We used 5 sentences and 28 listeners participated in the test. Two versions of each sentence were presented to the participants. First, the fluent version (without filled pauses) and second, the disfluent version, which contained one filled pause in it. Participants were asked to rated the naturalness of the sentences they listened to in a 1-5 scale.

Both systems achieve a median MOS score of 4. Although the mean value was slightly higher for the disfluent system this is not significant. Therefore, it has been possible to include filled pauses in the TTS system without decreasing its naturalness and our goal achieved.

## 6. Conclusions

In the present paper we showed evidences that filled pauses are as frequent as most frequent words in conversational speech. Then, we overviewed the synthetic disfluent model. This model assume that the prosody of a disfluent sentence can be generated by means of a standard fluent model, plus a model for local variations. Then, we have presented an analysis of local variations of segmental duration and pitch contour of filled pauses. Afterwards, a regression model has been proposed to predict the variations. Finally, the proposal was implemented in real unit-selection system and perceptually evaluated. Results showed that the model can successfully be used in generated disfluent sentences with filled pauses without degrading the quality of the original system.

## 7. Acknowledgements

This work has been partially funded by the Spanish Government under the AVIVAVOZ project (TEC2006-13694-C03) and the GLISSANDO Project (FFI2008-04982-C003-02).

## 8. References

[1] J. E. F. 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.

[2] M. Watanabe, K. Hirose, Y. Den, and N. Minematsu, “Filled pauses as cues to the complexity of following phrases,” in *Proceedings of Eurospeech*, September 2005, pp. 37–40, Lisbon, Portugal.

[3] S.-C. Tseng, “Grammar, prosody and speech disfluencies in spoken dialogues.” Ph.D. dissertation, Department of Linguistics and Literature, University of Bielefeld, April 1999.

[4] E. Shriberg, “Phonetic consequences of speech disfluency,” in *Proceedings of International Congress of Phonetic Science (ICPhS)*, vol. 1. San Francisco, CA, USA: Symposium on The Phonetics of Spontaneous Speech (S. Greenberg and P. Keating, organisers), 1999, pp. 619–622.

[5] H. H. Clark, “Speaking in time,” *Speech Communication*, vol. 36, no. 1-2, pp. 5–13, January 2002.

[6] D. C. O’Connell and S. Kowal, “The history of research on the filled pause as evindence of the written language bias in linguistics (linell, 1982),” *Journal of Psycholinguistic Research*, vol. 33, no. 6, pp. 459–474, November 2004.

[7] E. Shriberg, R. Bates, and A. Stolke, “A prosody-only decision-tree model for disfluency detection,” in *Proceedings of Eurospeech*, Rhodes, Greece, September 1997.

[8] M. Goto, K. Itou, and S. Hayamizu, “A real-time filled pauses detection system for spontaneous speech recognition,” in *Proceedings of Eurospeech*, Budapest, Hungary, 1999, pp. 227–230.

[9] S. Sundaram and S. Narayanan, “An empirical text transformation method for spontaneous speech synthesizers,” in *Proceedings of Eurospeech*, Geneva, Switzerland, September 2003, pp. 1221–1224.

[10] R. Carlson, K. Gustafson, and E. Strangert, “Cues for hesitation in speech synthesis,” in *Proceedings of Interspeech*, Pittsburgh, PA, USA, September 2006.

[11] J. Adell, A. Bonafonte, and D. Escudero-Mancebo, “On the generation of synthetic disfluent speech: Local prosodic modifications caused by the insertion of editing terms.” in *Proceedings of Interspeech*, Brisbane, Australia, September 2008, pp. 2278–2281.

[12] D. Conejero, J. Giménez, V. Arranz, A. Bonafonte, N. Pascual, N. Castell, and A. Moreno, “Lexica and corpora for speech-to-speech translation: A trilingual approach.” in *Proceedings of Eurospeech*, Geneva, Switzerland, September 2003.

[13] J. Adell, “Analysis and modelling of conversational elements for speech synthesis,” Ph.D. dissertation, Dpt. of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, July 2009.

[14] E. E. Shriberg, “Preliminaries to a theory of speech disfluencies,” Ph.D. dissertation, Berkeley’s University of California, 1994.

[15] E. Shriberg, “Disfluencies in switchboard,” in *Proceedings of International Conference on Speech and Language Processing (ICSLP)*, Pittsburg, PA, USA, 1996.

[16] A. Bonafonte, P. D. Agüero, J. Adell, J. Pérez, and A. Moreno, “Ogmios: The upc text-to-speech synthesis system for spoken translation,” in *Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation*, June 2006.