SYNTHESIS OF FILLED PAUSES BASED ON A DISFLUENT SPEECH MODEL

Jordi Adell, Antonio Bonafonte *

TALP Research Center
Universitat Politcnica de Catalunya (UPC)
Barcelona (Spain)

David Escudero †

ECA-SIMM Laboratory
Universidad de Valladolid
Valladolid (Spain)

ABSTRACT

In the present paper we present a new approach to the synthesis of filled pauses. 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

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. The main application of speech synthesis has been focused by now 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 have 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 may give cues about what is being said to the listener [4]. Disfluencies are very frequent in every day speech [5] so that it is possible to hypothesize the need to include these prosodic events to approximate 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]. 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 [10] or [11], 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 [12].

In this paper we first review the disfluent speech generation model. Second, we analyse which are these local modifications in the case of filled pauses and propose a model to predict them. Finally, we present the results of a perceptual evaluation of the models and some 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 existing unit selection systems have a closed inventory which do no contain disfluencies at all. Therefore, it is not possible that state of the art machine learning techniques, that are usually applied to model prosody, can learn these phenomena from data. Secondly, not only prosodic models but also text analysis models (e.g. POS tagging), expect sentences to have a rigid structure based on concatenation of syntactic, accent and intonation groups. When fluency is broken this structure is also broken and makes it more difficult for standard models to predict prosodic parameters at this point. 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 on fluent speech and on the other hand, takes into account local modifications at the point where fluency is broken [13].

Three different elements are taken into account for the generation of any given disfluent sentence (DS). First, the original sentence (OS) expected to be uttered before the disfluency is required. 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 [5] 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:

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

†and by the Ministerio de Ciencia e Innovacion, Spanish Government Glissando project FFI2008-04982-C003-02

$$DS = PrevRM|RM|ET|RR|PostRR$$
$$OS = PrevRM|RM|PostRM$$
$$TS = PrevRM|RR|PostRR$$

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 $PostRM$ is likely to be *right*.

A unit selection TTS system must be able to generate the OS and TS from the DS. A set of ETs 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 the RM is empty) [12]

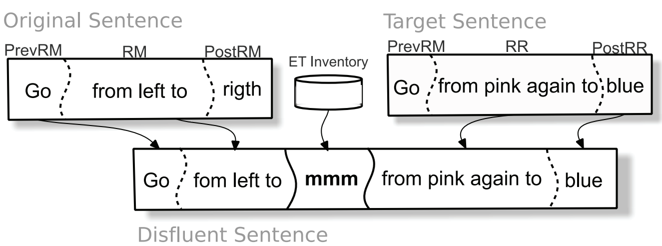

Fig. 1: Synthetic disfluent speech generation process applied to a sample sentence

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 that might appear at joins between elements described in this section ($PrevRM$, RM, ...).

In a previous work we showed evidences that fluent and disfluent sentences only differ from these local variations [12]. 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.

3. 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, modelling of these parameters will lead to the synthesis of disfluent speech.

3.1. Data

The corpus used here is a selection of sentences from the corpus developed under the LCSTAR European project[1]. 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 [15]. 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. One hundred 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.

3.2. Prosodic description of filled pauses

Filled pauses (FP) have been considered as pauses containing non-verbal sounds instead of silences [8]. First, we compared their duration with the one of silent pauses (SP). If the FPs duration followed the same distribution than silent pauses at the middle of sentences, a standard model trained on fluent speech could be used to predict their duration.

However, FPs are placed at points where silent pauses 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. 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 ($\pm$18ms) and for silent pauses is 222ms ($\pm$13ms). The difference between both means is statistically significant ($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.

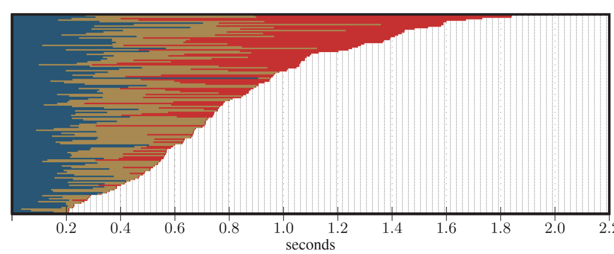

Fig. 2: Length of previous syllable (left), filler (center) and the silence (right). The total length of the bar shows the total FP duration.

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_f 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 $\approx$ 400ms. This picture suggests that syllables cannot be longer than $\approx$ 650ms, 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 $\sim$ 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

[1] http://www.lc-star.org

linear function:

$$D_{syl} = K \quad (1)$$

$$D_{fil} = \begin{cases} D_{FP} - D_{syl} & if\ D_{FP} < D_{fil}^{max} + K \\ D_{fil}^{max} & if\ D_{FP} \geq D_{fil}^{max} + K \end{cases} \quad (2)$$

$$D_{sil} = \begin{cases} 0 & if\ D_{FP} \leq D_{fil}^{max} + K \\ D_{FP} - K - D_{fil}^{max} & if\ D_{FP} > D_{fil}^{max} + K \end{cases} \quad (3)$$

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.

3.2.1. 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 [16].

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 ± 0.63, while for fillers it is -11 ± 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 generate FP can apply a continuous contour to each segmental unit, then this decreasing slope has to be taken into account. In contrast, some TTS systems only apply a factor to each unit, but do not modify the contour inside units. 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 $F0'$ 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:

$$F0' = \frac{F0^{-1} + F0^{1}}{2} \quad (4)$$

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

In Figure 3, we can observe that the difference between both values is $\sim -10Hz$, and for 80% 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 generate a correct 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.

3.3. 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.

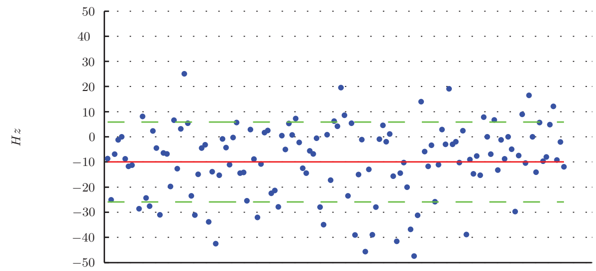

Fig. 3: $F0_{fil} - F0'$ for all fillers in the database. Plain line identifies the mean and dashed lines ± one standard deviation from the mean.

3.3.1. Duration

The necessary parameters 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 the regression lines that can be used to predict the local parameters that describe filled pauses from their global duration.

$$D_{syl} = 0.277 \quad (5)$$

$$D_{f} = 0.184 * D_{FP} + 0.130 \quad (6)$$

$$D_{sil} = 0.717 * D_{FP} - 0.320 \quad (7)$$

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.

3.3.2. Fundamental frequency 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 $F0' = \frac{F0^{-1}+F0^{1}}{2}$ can be used to predict $F0_{fil} = a * F0' + c$. 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 $F0'$ gives the following result:

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

4. EVALUATION

The proposed model has been implemented in our TTS system [17]. Our goal was to create a system able to generate filled pauses without degrading the quality of the existing system. If we achieved it we could say that our system can generated filled pauses thanks to the synthetic disfluent speech model. To evaluated whether we achieved the goal 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.

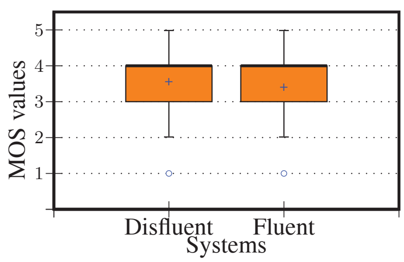

Fig. 4: MOS values of the system with and without filled pause.

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

5. CONCLUSIONS

In the present paper we overviewed the synthetic disfluent model. This model assumes that the prosody of a dislfuent 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 a real unit-selection system and perceptually evaluated. Results showed that the model can successfully be used in generating disfluent sentences with filled pauses without degrading the quality of the original system.

6. REFERENCES

[1] Andy Aaron, Ellen M. Eide, and John F. Pitrelli, "Conversational computers," *Scientific American*, vol. 292, no. 6, pp. 64–69, June 2005.

[2] Simon King and Vasilis Karaiskos, "The blizzard challenge 2009," in *Proceedings of the Blizzard Challenge 2009 Workshop*, Edinburg,UK, 2009.

[3] Jea 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.

[4] Michiko Watanabe, Keikichi Hirose, Yasuharu Den, and Nobuaki Minematsu, "Filled pauses as cues to the complexity of following phrases," in *Proceedings 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 *Proceedings of International Congress of Phonetic Science (ICPhS)*, San Francisco, CA, USA, 1999, Symposium on The Phonetics of Spontaneous Speech (S. Greenberg and P. Keating, organisers), vol. 1, pp. 619–622.

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

[8] Daniel C. O'Connell and Sabine 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.

[9] Elizabeth Shriberg, Rebecca Bates, and Andreas Stolke, "A prosody-only decision-tree model for disfluency detection," in *Proceedings of Eurospeech*, Rhodes, Greece, September 1997.

[10] Shiva Sundaram and Shrikanth Narayanan, "An empirical text transformation method for spontaneous speech synthesizers," in *Proceedings of Eurospeech*, Geneva, Switzerland, September 2003, pp. 1221–1224.

[11] Rolf Carlson, Kjell Gustafson, and Eva Strangert, "Cues for hesitation in speech synthesis," in *Proceedings of Interspeech*, Pittsburgh, PA, USA, September 2006.

[12] Jordi Adell, Antonio Bonafonte, and David 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.

[13] Jordi Adell, *Analysis and Modelling of Conversational Elements for Speech Synthesis*, Ph.D. thesis, Dpt. of Signal Theory and Communications, Universitat Politcnica de Catalunya (UPC), Barcelona, Spain, July 2009.

[14] Elizabeth Ellen Shriberg, *Preliminaries to a Theory of Speech Disfluencies*, Ph.D. thesis, Berkeley's University of California, 1994.

[15] David Conejero, Jess Gimnez, Victoria Arranz, Antonio Bonafonte, Neus Pascual, Nria Castell, and Asuncin Moreno, "Lexica and corpora for speech-to-speech translation: A trilingual approach.," in *Proceedings of Eurospeech*, Geneva, Switzerland, September 2003.

[16] Elizabeth Shriberg, "Disfluencies in switchboard," in *Proceedings of International Conference on Speech and Language Processing (ICSLP)*, Pittsburg, PA, USA, 1996.

[17] Antonio Bonafonte, Pablo D. Agüero, Jordi Adell, Javier Pérez, and Asunción 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.