Disfluent Speech Analysis and Synthesis: a preliminary approach.

Jordi Adell, Antonio Bonafonte and David Escudero

TALP Research Center
Dpt. Signal Theory and Comunication
Universitat Politècnica de Catalunya
Barcelona,Spain
www.talp.upc.edu

Dpt. of Informatics
Universidad de Valladolid
Valladolid, Spain
www.infor.uva.es

Abstract

Despite of the existence of high quality unit selection speech synthesizers, they are based on a reading style approach. However, new applications such as Speech-to-Speech Translation or Speech User Interfaces demand a talking style which is more natural in these contexts. Disfluencies are a major characteristic of talking style so that it is convenient to be able to generate disfluent speech. In the present paper a preliminary analysis of pitch and segmental duration in repetitions and filled pauses is presented. Simple rules to predict these prosodic features are derived from the previous analysis and used for synthesis. Evaluation shows an increase in naturalness while overall quality is decreased.

1. Introduction

Speech synthesis systems have already reached a high quality performance on reading text. Unit selection systems reach high levels of naturalness and intelligibility close to natural speech [1, 2]. Moreover, they are usually designed to read text and evaluated in such task. In fact, we have been building *reading* machines. However, some applications, such as Speech User Interfaces, Speech-to-Speech translation, or even film dubbing, request for a *talking* machine instead.

It is well known that humans do not write as they speak, reading and talking styles are produced under different rules. While speaking is more spontaneous, writing is more planned and do not benefit from other communicative tools such as prosody or gesture in addition to words themselves.

Therefore, there is a need to turn from *reading* to *talking* speech synthesis. A major factor that discriminates talking from reading is the use of disfluencies. *"Most spoken disfluencies are not problems in speaking, but the solutions to problems in speaking"* [3]. For this reason, in the present paper we will focus on pitch and segmental duration of disfluencies in order to work toward their synthesis.

The present work has been performed under the TC-STAR project. It is therefore, focused on a Speech-to-Speech Translation (S2ST) task in the parliamentary domain. This task has some particular conditions that makes from disfluencies, in our opinion, an interesting research field. There are two major characteristics to take into account when working in speech-to-speech translation in the parliamentary domain. In one hand, input and output are in a spoken style rather than written. On the other hand, recognition and translation systems may introduce errors to the text. Therefore, the Text-to-Speech (TTS) system has to take these errors into account.

This two elements affect both naturalness and intelligibility. If we intend to generate exactly the same meaning the speaker wanted to transmit, it is necessary to generate all the elements, also the ones that are not contained in words. Pitch, for example, but also disfluencies, they would also contribute to a more natural speech. Furthermore, intelligibility is affected by recognition and translation errors.

Text coming from the output of an statistical machine translation is usually hard to understand due to errors and due to unstructured sentences. Sometimes it is even hard for a human to read aloud text coming from the output of such systems. Humans can finally understand it if all the necessary words are present but it takes longer to do so.

Therefore, it would be helpful to be able to mark whether what the system is saying is considered to be hard to understand or not (for instance, if the translation confidence is very low). Disfluencies might be used to mark it or to give more time to the listener to decode what is being said as it has been already pointed out by different authors [4, 5, 10].

In summary, speech synthesis of disfluent speech may help to make voice more natural and to understand unstructured texts. For these reasons here we present an approach to synthesise two different kind of disfluencies: *filled pauses* and *repetitions*. First, an analysis is presented in Section 2, setting the overall framework for the present work. Then, the proposed synthesis approach is presented in Section 3 and evaluated in 4. Finally all the work is discussed in Section 5.

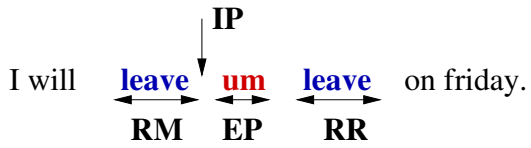
2. Analysis of Speech Disfluencies

Disfluencies in speech has already been analysed from different points of view, either linguistics, cognitive or engineering [7]. We must take profit from all these research fields in order to build a model that allows us to synthesise disfluent speech in a useful way for the listener. For example, and as has been mentioned in previous section, it could help to understand unstructured sentences as in S2ST systems.

2.1. General Framework

Disfluencies are *"phenomena that interrupt the flow of speech and do not add propositional content to an utterance"* [4]. There already exist several studies on how classifying and modelling them [8]. Here we will use the standardisation presented on [8] based on the *Reparandum*(RM), *Interruption Point*(IP), *Editing Phase*(EP) and *Repair*(RR) structure shown below:

This work has been partially sponsored by the European Union under grant FP6-506738 (TC-STAR project, http://www.tc-star.org) and by the Spanish Goverment (MCYT) under contract TIC2003-08382-C05-03



The RM is the speech to be repaired, the IP corresponds to the time at which the speaker realises something is going wrong. The EP may or may not exist, and it might be used as an editing function such as *"I mean"* or it might just contain a filled or unfilled pause used for re-planing. Finally, RR has a relation to the RM and is the onset of the fluent speech.

Therefore, it is possible to list some disfluencies and relate them to this structure as shown below:

- *Substitution:* Repair is different to reparandum.
- *Insertion:* Repair repeats reparandum adding new element.
- *Deletion:* Part of reparandum or all of it is removed in the repair.
- *Repetition:* Repair is equal to reparandum.
- *Filled Pause:* Editing phase is a filler.

Here in the present paper we will focus on filled pauses and repetitions. We will acoustically analyse them from real speech in order to extract a small set of rules that will allow us to generate them. The work presented explores the segment duration and pitch modelling of these disfluent events. There already exists literature about acoustic analysis of disfluencies [7]. However, here we have tackled the analysis problem as well as the synthesis.

2.2. Acoustic analysis

An acoustic analysis of both duration and pitch characteristics of disfluent speech has been made. The purpose for such study is to investigate on how to model prosodic features of disfluent speech. We do not intend to build an exhaustive model, but a simple one instead, that would allow to generate disfluencies in a general way. For this purpose a database of 37,000 words from the European Parliament in Spanish has been recorded. It is a framework with rich speeches. However, due to its formal style, it is not *too much* spontaneous, what would have made it too difficult to study. These recordings contain 67 different speakers. The number of filled pauses found are 300 (≈1%) and the number of repetitions of one single word is 56, mostly function words.

First, the acoustic analysis for filled pauses is presented and afterwards the one for repetitions.

2.2.1. *Filled Pauses*

Filled pauses are considered disfluencies where the EP is a filler (such as *uh*, *um*, *ehh*, etc.). RM and RR may or may not contain anything. Therefore, filled pauses (FP) may be a disfluency on their own or they may be part of another kind of disfluency such as deletions for example. Furthermore, in the present work filled pauses are studied with respect to their context words.

The pitch contour and the duration have been extracted for each disfluent event, Figure 1 shows four examples. Mean and standard deviation of the contour have been calculated. Then, three mean values ($\bar{f0}_{w-1}, \bar{f0}_{fp}, \bar{f0}_{w+1}$) and standard deviations ($\sigma_{w-1}, \sigma_{fp}\sigma_{w+1}$) for each contour are obtained, one for previous word, one for the filled pause and the third one for the next word after the filled pause.

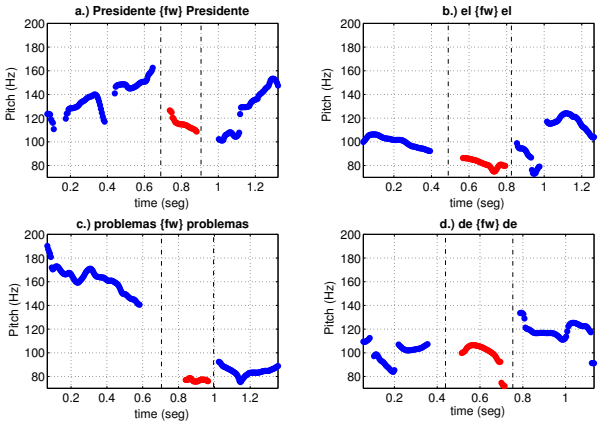


Figure 1: Pitch contour examples for Filled Pauses.

We are interested on the prosody of the FP with respect to their adjacent words. It has been stated previously that pitch of filled pauses is systematically lower than their context [7]. This can also be observed in examples in Figure 1. To analyse this hypothesis, correlation between mean values has been obtained. $\bar{f0}_{fp}$ has a 0.51 correlation with respect to $\bar{f0}_{w-1}$ and 0.52 with respect to $\bar{f0}_{w+1}$. Therefore, there is a clear relation between pitch of FPs and their context.

Following this observations a factor that relates the pitch of the filler with its context pitch is calculated and named Filler Pitch factor (FP_f) as shown in Equation 1, its histogram is shown in Figure 2.

$$FP_f = \frac{2 \cdot \bar{f0}_{fp}}{\bar{f0}_{w-1} + \bar{f0}_{w+1}} \quad (1)$$

We can observe in the histogram how most of the values (aprox. 80%) are under 1. Around 45% of the values are between 0.8 and 1. This supports the hypothesis that pitch of filled pauses are generally lower than their contexts.

From duration values of filled pauses and their previous and posterior words (d_{fp}, d_{w-1}, d_{w+1}). Correlation values across them are too low to be able to predict duration of the filled pause from the duration of the context words. However, Figure 2 shows that most of filled pause durations go from 100ms to 200ms. Then, a representative value of filled pauses could be their median, for this database it is 180ms.

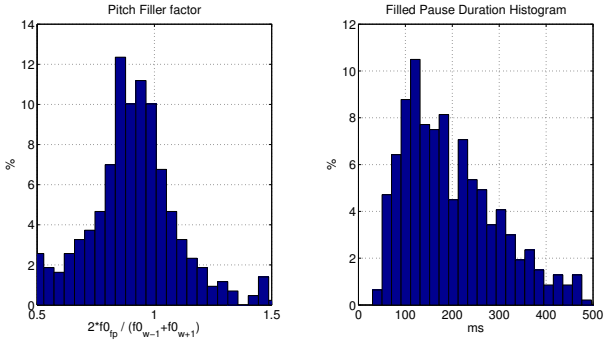


Figure 2: Histograms for pitch factor in Equation 1 and filled pause durations

2.2.2. *Repetitions*

Repetitions are disfluencies where RM and RR are the same. They can be uttered by a speaker because he needs more time for planning, to emphasise the word repeated, or to restart again a sentence with a previous new intention. EP can be empty or not in repetitions. In the present work we will study only the RM and RR parts of repetitions. EP should be studied aside. For example, if EP is a filled pause it has been treated in previous section.

Pitch contours and durations from RM and RR have also been extracted here ($f0_{RM}(t)$ and $f0_{RR}(t)$) and temporally normalised from 0 to 1. Some examples are shown in Figure 3. Correlation and RMSE have been calculated across both repetitions for each example. They are presented as an histogram in Figure 4. It can be observed how ≈40% of the examples analysed gave a correlation bigger than 0.75 and RMSE lower than < 15Hz. This indicates that both pitch contours are quite similar to each other in most cases. This fact can also be observed in example of Figure 3.

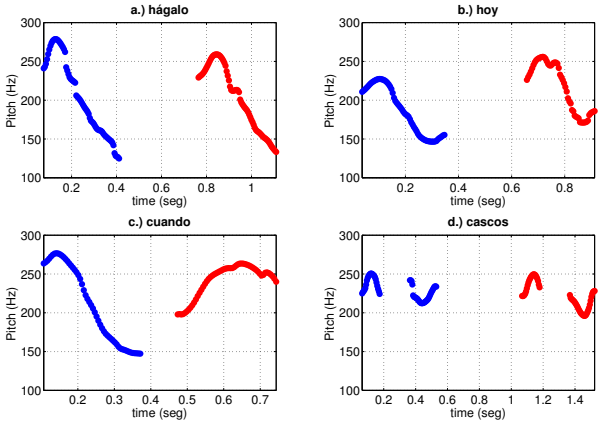


Figure 3: Four pitch contour examples of repetitions.

Durations have also been analysed. In Figure 4 the histogram of RR duration with respect to RP ($d_{RR} \div d_{RM}$) has been plotted. It can be observed that in most times the RM is longer than the repair. This must be taken into account when modelling repetitions.

Another important thing is to know whether the reparandum is longer than usual or the repair is shorter. In order to discuss this, repetitions of the word *la* and the word *de* have been studied. The mean duration of word *la* in fluent speech is 135ms, while for repetition when it is in the RM is 422ms and 163ms when it is in the repair. For word *de* values are 147ms for fluent speech, and 284ms and 136ms for RM and RR respectively. It can clearly be observed how RR duration is closer to fluent speech and RM is lengthened with respect to it.

3. Synthesis of Speech Disfluencies.

This section focus on synthesis of disfluencies. This may be useful for different purposes such us making unstructured speech more clear, or simply to achieve a more expressive or natural speech. Disfluencies cannot only show problems on generating speech but also change the actual meaning of a sentence [9].

The purpose of the present work is to find simple relations between disfluency units and their context in order to be able

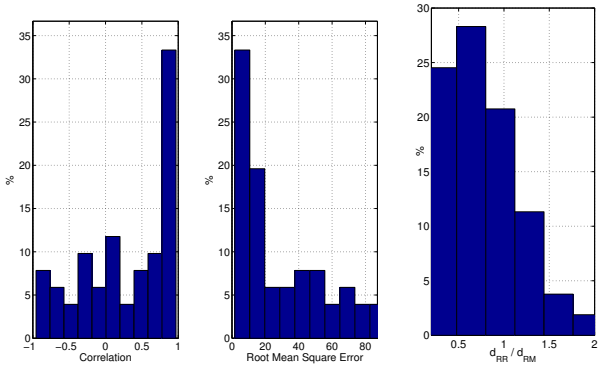


Figure 4: Correlation and RMSE histograms for Repetition.

to synthesise the disfluencies from fluent realisation adding the modifications described by the rules. We present results for two different types of disfluencies: *filled pauses* and *repetitions*.

Prediction of disfluencies is out of the scope of the present paper [10], we aim instead to generate them properly. The approach presented here is based on a Speech Synthesis Markup Language that indicates the type and elements of the disfluency. Other works can be found in the literature that tackle the prediction problem [10].

It is clear from Section 2.2 that there exists a relationship between pitch and duration of disfluencies and their context. We propose thus here an approach to synthesise disfluencies based on these findings and on the definition of disfluency.

Disfluencies are generally due to a need for planning what will be said, or entered by a need to correct what has already been said. They are usually co-articulated with what had to be said in the original plan. Therefore, we propose, based on a unit selection framework, to have a prosodic model for fluent speech and use it to predict the prosodic features of the structured text. Afterwards, a disfluency model, based in presented findings and by means of simple rules, could add new units and their desired prosody before the selection and concatenation is performed.

In the case of repetitions, these rules consist on considering the units generated by the synthesiser as the RR and add new units for the RM. New units have now same pitch contour as the RR but their length is $r_{duration}$ times the duration of the RR. $r_{duration}$ is the median of $\frac{d_{RM}}{d_{RR}}$ distribution in Figure 4. Rules are summarised in Table1.

	Repetition
F0 Contour	$F0Contour_{RM} = F0Contour_{RR}$
D	$d_{RM} = r_{duration} \cdot d_{RR}$
	Filled Pause
F0 Contour	$F0Contour_{FP} = \frac{r_{pitch}}{2} \cdot (f0_{w-1} + f0_{w+1})$
D	d_{RM} =*constant value*

Table 1: Disfluency generation rules.

Filled pauses consist on one single unit modelled by one single pitch value obtained by multiplying the mean pitch value of adjacent words by a factor r_{pitch} corresponding to the median of distribution in Figure 2. Filled pauses duration is a constant value obtained corresponding to the median value of duration distributions in Figure 2.

Input text contains XML marks with information about the elements of the disfluency (DF) and each of its elements (DFE),

as in the example shown bellow:

```
I will
<DF TYPE="repetition">
   <DFE TYPE="RM"> leave</DFE>
   <DFE TYPE="EP">  um  </DFE>
   <DFE TYPE="RP"> leave</DFE>
</DF>
on friday
```

The TTS system generates the desired prosodic features for a fluent sentence. Afterwards new units are added and new features are predicted from the context in the fluent sentence. Finally, selection and concatenation is performed as usual.

4. Evaluation

It is now necessary to evaluate results from previous sections. The analysis presented in this paper is a first approach to the problem of disfluent speech synthesis. We attempted to find simple rules for the prosodic modelling of repetitions and filled pauses.

Therefore, they must be perceptually evaluated. We have synthesised a set of ten sentences, chosen from transcriptions of real speeches and which contained disfluencies. Also a paragraph has been selected and synthesised. Speeches came from the European Parliament and were in Spanish. In order to evaluate the performance of the proposed method informal tests have been done and findings are presented in this section.

Sentences have been generated by modifying synthetic speech using praat [11] and its PSOLA functionalities in order to implement rules from Table 1. This was done as a preliminary step to their implementation inside the synthesis system. However, the paragraph was synthesised using a unit selection synthesis system [12], but without any specific pitch nor duration modelling. Although findings were consistent across speakers in the analysis step and listening tests have shown promising results, the desired performance has not been reach.

There are a variety of effects that we consider contributed to these results. On one hand repetitions had concatenation problems, since RM was copied and modified from RR. Some repetitions were perceived more like a synthesis error than disfluent speech. This may be due the fact that the model is too simple. On the other hand filled pauses showed more encouraging results, they have shown to be easier to generate because it is possible to add a small silence before and after it, thus avoiding co-articulation problems.

A closer to talking style have been perceived, however the overall quality of the system has decreased, and therefore intelligibility is reduced. It must be noticed that synthesis of the paragraph showed better results than the sentences.

5. Discussion

The analysis of repetitions and filled pauses have shown a consistency across speakers of some simple rules for the prediction of pitch and duration. Pitch of filled pauses is systematically lower than their context, and reparandum in repetitions is systematically longer than repair and fluent realisations of same words.

An informal test has been performed and it highlighted some problems of the system presented. Filled pauses presented few listening problems and helped the rhythm of the speech. However, repetitions present major problems as they were considered synthesis problems rather than disfluencies. Further studies must be performed for more than one word length repetitions and in order to find more complex models for disfluent synthesis.

Findings are somehow encouraging, despite the decrease of intelligibility when applying the rules proposed, the use of disfluencies clearly lead us to a style closer to talking speech.

In summary, in this paper we presented a preliminary approach to disfluent speech synthesis, simple rules have been derived from disfluent speech analysis and used to generate synthetic speech. Despite consistent rules have been found they have shown not to be enough for synthesis purposes. However, the synthesis of disfluent speech is promising within the framework of speech-to-speech translation technologies, since they break the usual monotony of speech synthesis, and makes it more suitable to be listened for a longer time than reading style speech synthesis.

6. References

[1] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *Proc. of Joint meeting of ASSA*, March 1999, Berlin, Germany.

[2] Alan W. Black, "Perfect Synthesis for All of the People All of the Time," in *Proceedings of IEEE Text-to-Speech Workshop*, 2002, Santa Mónica, CA.

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

[4] Jea E. Fox Tree, "The effects on of false starts and repetitions on the processing of subsequent words in spontaneous speech," *Journal of Memory and Language*, vol. 34, pp. 709–738, 1995.

[5] Jean E. Fox Tree, "Listeners' uses of *um* and *uh* in speech comprehension," *Memory & Cognition*, vol. 29, no. 2, pp. 320–326, 2001.

[6] Martin Corley and Robert J. Hasrtsuiker, "Hesitation in speech can... um... help a listener understand.," in *Proc. of 25th Meeting of Cognitive Science Society*, July 2003, Boston,USA.

[7] 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.

[8] Elizabeth Ellen Shriberg, *Preliminaries to a Theory of Speech Disfluencies*, Ph.D. thesis, Berkeley's University of California, 1990.

[9] E. Eide, A. Aaron, R. Bakis, W. Hanza, M. Picheny, and J. Pirelli, "A corpus-based approach to <AHEM/> expressive speech synthesis," in *Proccedings of 5th ISSW*, June 2004, pp. 79–84, Pittsburgh, USA.

[10] Shiva Sundaram and Shrikanth Narayanan, "Spoken language synthesis: Experiments in synthesis of spontaneous monologues," in *Proc. of IEEE Speech Synthesis Workshop*, September 2002, Santa Monica, CA, USA.

[11] Paul Boersma and David Weenink, "Praat: doing phonetics by computer (version 4.3.04)," March 2005, http://www.praat.org/.

[12] Antonio Bonafonte, Ignasi Esquerra, Albert Febrer, José A. R. Fonollosa, and Francesc Vallverdú, "The UPC text-to-speech system for Spanish and Catalan," in *Proceedings of ICSLP*, November 1998, Sydney, Australia.