

Statistical analysis of filled pauses' rhythm for disfluent speech synthesis

Jordi Adell¹, Antonio Bonafonte¹, David Escudero²

¹Dpt. of Signal Theory and Communications, Universitat Politècnica de Catalunya, Spain.

²Dpt. Computer Science, Universidad de Valladolid, Spain.

Abstract

Given that state of the art speech synthesis systems have already reached a high naturalness level, it is time to move to *talking* speech from the actual *read* speech framework. For this purpose it is thus necessary to investigate how disfluencies can be included in speech synthesis and even increase its naturalness. This paper builds on a previously presented work and focuses on finding a local model of filled pauses rhythm. A statistical study of rhythm effects around filled pauses is presented and based on the correlation between rhythm variables, a regression model is proposed to predict filled pauses duration and prepausal lengthening.

1. Introduction

Speech synthesis has already reached high naturalness, mainly due to the use of effective techniques such as unit selection-based systems [1] or other new arising technologies [2] based on the analysis of huge speech corpora. The main application of speech synthesis has been focused by now on reading style speech as it is plausible to assess that reading style is the most generalist style to be extrapolated to any other situation. But nowadays new applications of text-to-speech (TTS) systems like film dubbing, robotics, dialogue systems, speech translation or multilingual broadcasting demand different styles as the users expect the interface to do more than just reading information.

If synthetic voices want to be integrated in future technology, they must *speak* the way people talk instead the way people read. This objective has been already tackled in several manners such as emotional speech synthesis [3], voice quality modelling [4] or even pronunciation variants [5]. In our opinion, style is more important; it is desirable synthetic speech to be more conversational-like rather than reading-like speech. We call this *talking speech*, in contrast to *read speech*.

Talking speech differs significantly from reading speech due to the inclusions of a set 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* [6]. Disfluencies are very frequent in normal speech [7] and they in fact contain information [8] and help human communication [9, 10]. Then, it is plausible to hypothesise the need to include this prosodic event in order to move towards to talking speech synthesis. In the present work we focus in one kind of disfluency: *filled pauses*.

There already exist published works on disfluent speech synthesis like the one done in [11], where they presented an algorithm for insertion of filled pauses and breathing into a text. Also in [12], where they present a study about prosodic cues of hesitations for speech synthesis.

We have also presented experiences on synthesising disfluencies (i.e. filled pauses and repetitions) in TTS systems in previous works [13], and here we present further work on the same direction focusing on filled pauses' rhythm. In our previous work, we claimed that filled pauses' pitch is lower than its segmental context. However, we were not able to find any simple model to predict the filled pause (FP) duration and a constant value was proposed. Although the synthesis of filled pauses reached higher degree of quality than repetition synthesis in informal tests, we have detected two main drawbacks: *coarticulation* and *rhythm*.

Since our work is based on a unit-selection approach, coarticulation problems come from the lack of FP units in the inventory and from the fact that filled pauses can be strongly coarticulated, some times it is hard even to differentiate, in human speech, filled pauses from strong vowel lengthening. The second drawback was that the sentence rhythm was not affected by the presence of the filled pause at all in the synthetic speech. It was inserted into a fully fluent utterance in terms of rhythm and it sounded unnatural.

In this paper, first of all the database used is described. Then in Section 3, the use of silent pauses to avoid coarticulation is discussed. The study on the rhythm of sentence with filled pauses is presented in Section 4. Afterwards, due to the similar naturalness between filled pauses and silent pauses, the findings of the study will be analysed in the case of silent pauses in Section 5. Finally conclusions are summarised in Section 6.

2. Database and Synthesis

A database has been recorded specifically for unit-selection speech synthesis of conversational speech. Two large databases of about 10h of speech each one where recorded in order to build a couple of high quality voices for our TTS system, a male and a female voice. In addition, some extra sentences where recorded to study the synthesis of disfluent speech synthesis.

Prompts to record these sentences where extracted from real utterances from the European Parliament. They consisted on 65 sentences, which contained filled pauses, repetitions, restarts and breathing. These 65 sentences have been recorded by the male as well as by the female speaker. The prompt given to both speakers contained indications of where to do filled pauses, repetitions and others disfluent events.

In the case of filled pauses, prompts signalled when a filled pause had to be uttered but no acoustic specifications was given to the speaker. Therefore, the database contains a variety of realisations: *ehh*, *ahh*, *mmm*, *emm*. However, in the present paper all filled pauses have been considered equally.

Furthermore, the sentences have been manually segmented at phone level. These sentences have been added to the unit-selection inventory. The present work is focused on filled pauses and the database contains 138 of them.

These units, i.e. *ehh* or *mmm*, have been turned into phone-like units that can be used in the selection and concatenation process which need prosodic values to choose the most appropriate unit. Therefore, as well as for phones, prosodic models are requested and this is the main motivation of this work.

3. Silent Pauses Insertion

In order to avoid coarticulation problems, the insertion of silent pauses at both sides of filled pauses is proposed here. Experimental observations have motivated such proposal. In Figure 1 it can be observed how both silent pauses are present at both sides of the filled pause (*ehh*).

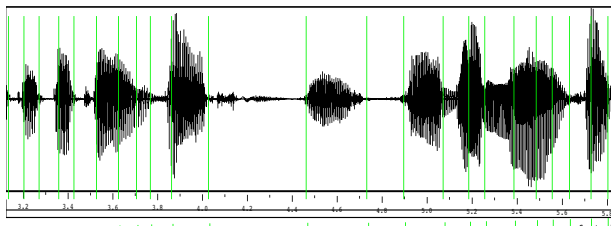


Figure 1: Audio example. It can be observed the pre-pausal syllable lengthening (*/fa/*) and also the short silences before and after the filled pause (*/ehh/*).

Table 1 shows how many times in the database silent pauses appear next to a filled pause. It can be observed that if we consider both speakers together 49% of filled pauses contain at least one of those silent pauses. This fact supports the insertion of these silent pauses. On one hand, the female speaker do not include both of them never in the database and only one forth of the filled pauses contain at least one SP. On the other hand, the male speaker uses this silent pauses more often since two thirds of the FP contains at least one SP.

Combinations	Both Spk.		Male Spk.		Female Spk.	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
FP	70	51%	25	32%	46	74%
SP·FP·SP	16	12%	16	20%	0	0%
SP·FP	12	9%	8	10%	5	8%
FP·SP	39	28%	28	36%	11	17%

Table 1: Number (*n*) and frequency of occurrences of silent pauses together with filled pauses. Silent pauses at both sides (SP·FP·SP), at one side (SP·FP or FP·SP), and no silent pauses (FP) have been taken into account.

Therefore, the use of both silent pauses is not the most frequent structure used by the two speakers we analyse here.

However, it is a possible structure and thus we are allowed to use it in order to avoid coarticulation problems in the insertion of FP. Therefore, the silent pauses' length is a variable that has to be predicted by the prosodic model.

4. Rhythm Study

In this section we will discuss rhythm implications in filled pauses. For this purpose, a set of rhythm-related variables will be defined. Afterwards, some summary statistics are presented in order to identify the general behaviour of such variables. Then, correlation between variables is explored and a regression model is proposed for prosody modelling of filled pauses.

4.1. Feature-set definition

We define “*rhythm*” as the mean syllable length of an utterance. This can be done in Spanish since syllable is the basic segmental unit for timing [14]. Since we are interested to discuss whether the filled pause produces a rhythm change or not, three rhythm variables are define: the total rhythm of the sentence (i.e. mean syllable length across the whole sentence), the rhythm previous to the filled pause, from the beginning of the sentence; and the rhythm after the FP. We will refer to these variables as: *totrh*, *prerh* and *posrh*. The filled pause duration has been excluded and is evaluated separately (i.e. *fpdur*). Through experimental observations we realised that prepausal syllables were larger than the mean syllable length. This phenomena can be observed in Figure 1. In this audio example, the prepausal syllable (*/fa/*) is significantly larger than the rest of syllables. Therefore, this value has also been excluded from rhythm calculus and variable *syl₋₁* will represent syllable length of syllable previous to filled pause. Moreover, since silent pauses before and after the FP will be part of the model, two more variables are included in the study, they represent both silent pauses' length: *paupre* and *paupos*. In addition, in order to examine whether only the prepausal length is lengthened or not, *syl₋₂* was added; and given the importance of the syllable nucleus (i.e. the vowel) in the syllable length also its duration has been included: *nuc₋₁* and *nuc₋₂*. In Summary, the set of features extracted from the database for each filled pause are: *totrh*, *prerh*, *posrh*, *syl₋₂*, *nuc₋₂*, *syl₋₁*, *nuc₋₁*, *paupre*, *fpdur* and *paupos*.

4.2. Summary statistics

Table 2 shows mean, standard deviation, lower and upper quartiles for each feature corresponding to male speaker. Table 3 shows same statistics for the female speaker. It can be observed how rhythm distributions are very similar for the total, the previous and the posterior rhythm in the case of the male as well as for the female speaker. Hypothesis tests have shown that at 95% confidence level rhythm means are equal. This supports our claim that filled pauses do not imply a rhythm change in the sentence. Therefore, the prosody of the corresponding fluent sentence can be modelled and rules to predict *fpdur*, *syl₋₁*, *paupre* and *paupos* could afterwards be applied. This is specially useful in our case, since the biggest part of the synthesis inventory are built by fluently uttered sentences, while only a small part of it contains disfluencies.

Unit.ms Name	Mean	Std Deviation	Lower Quartile	Upper Quartile
totrh.	167	35	143	182
prerh.	173	63	149	181
posrh.	180	42	154	199
syl₋₂	234	125	151	301
nuc₋₂	114	83	74	112
syl₋₁	394	179	280	494
nuc₋₁	228	118	140	288
paupre	348	282	100	465
paupos	242	270	71	404
fpdur	464	223	294	655

Table 2: Summary statistics for the male speaker and for filled pauses.

It can be observed how there is a significant lengthening of the prepausal syllable. Note that mean value of *syl₋₁* is 2.3 times bigger than the mean syllable length of the sentences for

the male speaker and 2.45 in the case of the female speaker. Figure 2 show the Box-and-Whisker graphic of the three rhythm-related and the syllable duration distributions. It can intuitively be observed how the rhythm has the same distribution before and after the filled pause, and how the prepausal syllable distribution is moved through the right in the graphics, what implies a lengthening of the syllable with respect to the sentence rhythm. Same effect appear for both speakers.

<i>Unit:ms</i> <i>Name</i>	<i>Mean</i>	<i>Std</i> <i>Deviation</i>	<i>Lower</i> <i>Quartile</i>	<i>Upper</i> <i>Quartile</i>
totrh.	154	24	146	169
prerh.	150	35	142	167
posrh.	165	28	150	176
syl₋₂	232	131	145	317
nuc₋₂	108	61	71	123
syl₋₁	378	140	304	424
nuc₋₁	222	70	168	277
paupre	434	326	180	595
paupos	268	298	82	360
fpdur	506	184	406	629

Table 3: Summary statistics for the female Speaker and for filled pauses.

These observations support the fact that there exists a prepausal syllable lengthening in filled pauses. A further issue will be to predict this lengthening. It can also be observed that the FP duration is much larger than the sentence rhythm.

The filled pause duration is significantly larger than the mean rhythm, also its standard deviation is bigger. This is related with the fact that filled pauses are used to re-plan what is going to be said. However, experimental synthesis have shown that not any length sounds natural. We believe that a certain relation between the syllable lengthening and the filled pause duration must exist, i.e. the prepausal length and the filled pause duration will be larger for slower speeches and vice versa.

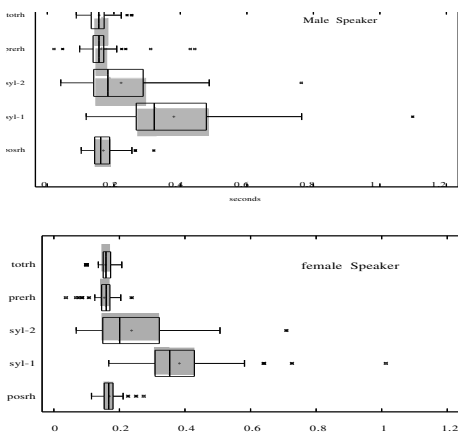


Figure 2: Box-and-Whisker graphics for rhythm variables and syllable lengths for filled pauses.

For these reasons in next sections we will look at the relation between the sentence rhythm and these two variable plus silent pauses' duration.

4.3. Correlation between variables

In order to analyse the relation between the sentence rhythm and the syllable, silent and filled pauses duration; we have calculated the correlation values between all variables. Since the database is small, we have also compute the statistical significance of the correlation values and only significant values ($P < 0.05$) are given. Table 4 summarises all correlation values.

Silent pauses duration do not have significant correlation with any other variable except for the female speaker, in this case they are only correlated with the FP duration. However, since the filled pause duration is an unknown variable, it can not be used to predict the silent pause duration. Therefore, with the approach presented here there is no way to predict the silent pause duration.

Since we have concluded that the rhythm does not change across the sentence, but that the rhythm across the whole sentence, the one previous to the FP and the posterior follow the same distribution instead, there are only two features left to model: syl_{-1} and $fpdur$.

<i>Variable</i>	Male Speaker		Female Speaker	
	<i>syl₋₁</i>	<i>fpdur</i>	<i>syl₋₁</i>	<i>fpdur</i>
totrh	-0.24	-	-0.45	-
prerh	-	-	-0.47	-
syl₋₂	0.26	-	0.27	-
nuc₋₂	0.32	-	0.56	-
syl₋₁	1	0.29	1	-
nuc₋₁	0.63	0.48	0.39	-
paupre	-	-	-	0.37
fpdur	0.30	1	-	1
paupos	-	-	-	0.39
posrh	-	-	-	0.30

Table 4: Statistically significant correlation between previous syllable length, FP duration a defined variables.

Significant correlation will guide us in order to find independent variables for modelling these features. It can be observed in Table 4 that both features are correlated with utterances that occur in advance in the sentences. For example, syl_{-1} is significantly correlated with the previous syllable and also with the total rhythm of the sentence. In addition, $fpdur$ is correlated with previous rhythm, the total rhythm, and also syl_{-1} in the case of the male speaker. Unexpectedly, the $fpdur$ in the case of the female speaker, is correlated with the silent pauses. However, in Table 1 in Section 3 we have seen that the female speaker do not insert any silent pause in 74% of the utterances.

Since the database was recorded in a studio, we have observed that the female speaker is less systematic in the realisation of such filled pauses, and also less natural. What would explain the lack of significance in the correlation between $fpdur$ and syl_{-1} . However, the filled pause is significantly correlated, in this case, with the posterior rhythm. Moreover, since the rhythm do no change significantly across the sentences, the fact that $fpdur$ is correlated with the posterior rhythm implies that is correlated with the other two rhythm variables in the study (i.e. $totrh$ and $prerh$), but that lack of data makes this correlation not statistically significant.

In next section we will discuss the use of these correlation between features, in order to generate a regression model for synthesis of filled pauses.

4.4. regression models

When trying to synthesise filled pauses within the unit-selection framework, the first issue to take into account is what units to be used. Here we have choose to record a small database containing disfluencies. Filled pauses was one kind of disfluencies recorded. Therefore, filled pauses units are now available in the inventory to its use for disfluent speech synthesis (see Section 2).

After the unit inventory issue is solved, the desired prosody has to be generated. For this purpose our synthesiser already have a pitch, duration and energy model [15]. However, this model is trained on fluent speech. As we have stated in Section 4.2, it is possible to use state of the art prosody modelling to predict the rhythm of the whole sentence as if it was a fluent sentence, and afterwards some local model can be applied to modify this fluent prosody in order to achieve the desired disfluent one.

For this purpose, in the case case of filled pauses, only two variables need to be predicted: syl_{-1} and $fpdur$ (i.e. pre-pausal syllable length, and filled pause duration). Also the silent pauses (i.e. *paupre*, and *paupos*) should be predicted, but we have not found any significant correlation here. Here we propose to use a multiple regression model due to its simplicity and to that these variables are correlated with the rhythm of the sentence.

Given results from Table 4 the prepausal syllable duration can be predicted by means of the total rhythm of the sentence and its previous syllable. In both cases the syllable duration and the syllable nucleus duration are very correlated thus only one of them is used, the one that gives a better fitting are mentioned here. In the case of the male speaker the regression function proposed is:

$$syl_{-1} = 568 + 0.58nuc_{-2} - 1.45totrh \quad (1)$$

and it fits the data with a 106ms of mean absolute error(MAE). For the female speaker the regression function proposed is:

$$syl_{-1} = 692 + 0.96nuc_{-2} - 1.68totrh - 1.07prerh \quad (2)$$

and it fits the date with a MAE of 81ms.

The filled pause duration now can be predicted by means of the sentence rhythm but also depends on the pre-pausal lengthening. Since prepausal lengthening is part of the whole model then a cumulative error effect will be produced, since the error done on prepausal length prediction will be passed to regression function for the filled pause duration. The proposed regression function for the male speaker is as follows:

$$fpdur = 338 + 0.86nuc_{-1} \quad (3)$$

and for the female speaker this functions is proposed:

$$fpdur = 181 + 1.96posrh \quad (4)$$

both functions fit the data with a MAE of 126ms.

5. Comparison with Silent Pauses

As we have said in Section 1 we also want to evaluate whether conclusions concerning filled pauses can also be extended to silent pauses. For this purposes, we have used the whole databases recorded for speech synthesis, which contains about 10h of speech. Same features described in Section 4.1 have been extracted from this database and same statistics have been

<i>Unit:ms</i> <i>Name</i>	<i>Mean</i>	<i>Std</i> <i>Deviation</i>	<i>Lower</i> <i>Quartile</i>	<i>Upper</i> <i>Quartile</i>
totrh.	158	17	149	161
prerh.	162	24	149	164
posrh.	160	26	147	165
syl₋₂	164	53	128	196
nuc₋₂	78	27	63	88
syl₋₁	237	61	200	272
nuc₋₁	110	42	80	136
spdur	340	265	116	480

Table 5: Summary statistics for the male speaker and for silent pauses..

computed. However, now 15,300 silent pauses are available to compute statistics, what means a much larger amount of examples than for filled pauses.

Tables 5 and 6 presents summary statistics for silent pauses. Same analysis is presented for filled pauses in Tables 2 and 3. We can observe how again the rhythm do no change in the silent pause, since rhythm previous to the pause and after it follow same distribution than total sentence rhythm.

<i>Unit:ms</i> <i>Name</i>	<i>Mean</i>	<i>Std</i> <i>Deviation</i>	<i>Lower</i> <i>Quartile</i>	<i>Upper</i> <i>Quartile</i>
totrh.	165	13	158	169
prerh.	167	18	158	172
posrh.	165	19	156	170
syl₋₂	173	50	136	204
nuc₋₂	82	24	68	96
syl₋₁	245	64	195	288
nuc₋₁	118	26	104	132
spdur	313	185	196	364

Table 6: Summary statistics for the female speaker and for silent pauses.

Furthermore, the well-known prepausal lengthening is observed. It can be observed more clearly in Figure 3, which is very similar to the corresponding to filled pauses (see Figure 2). Until now, same conclusion extracted from analysing filled pauses are extracted. This means, that it might be possible to predict sentence rhythm without taking silent pauses into account, and afterwards the silent pause prosody (i.e. pause duration plus prepausal syllable length) can be modelled locally.

Also correlations across features have been computed for silent pauses. Table 7 shows the corresponding values. Note that all correlations are significant since a lot more value are given. It can be observed how syl_{-2} is not correlated at all with silent pause duration (named as *spdur*) neither with pre-pausal length syllable. However, syllable length is strongly correlated with the rhythm and the pause length. Of special interest is the correlation value between pre-pausal syllable and pause duration since it is negative. This implies that the longer the pre-pausal syllable is the shorter the pause. These results are similar than the ones published in [12] claiming that what is perceptually important in hesitations is the sum of the pre-pausal syllable length plus the silent pause. We can conclude from correlations in Table 7 that the faster this speaker talk, the pre-pausal syllable is shorter but there is a longer pause. In contrast, if we talk slowly the pre-pausal syllable is, of course, longer but the silence is shorter.

Finally, note that results are similar than the ones presented

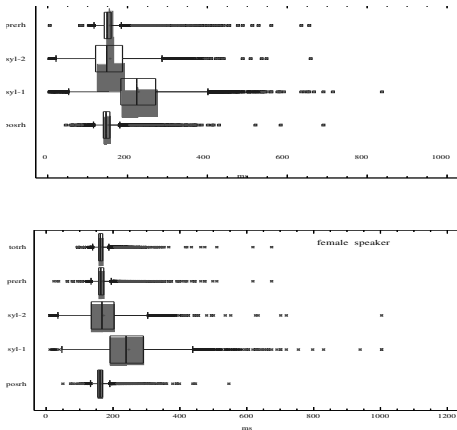


Figure 3: Box-and-Whisker graphics for rhythm variables and syllable lengths for silent pauses.

Variable	Male Speaker		Female Speaker	
	syl ₋₁	spdur	syl ₋₁	spdur
totrh	0.36	-0.22	0.28	-0.15
prerh	0.37	-0.18	0.33	-0.11
syl₋₂	-0.05	-0.06	-0.13	-0.04
nuc₋₂	-0.04	-0.04	-0.10	-0.02
syl₋₁	1	-0.28	1	-0.03
nuc₋₁	0.07	-0.24	-0.10	-0.05
spdur	-0.28	1	-0.03	1
posrh	0.20	-0.18	0.11	-0.13

Table 7: Statistically significant correlation between previous syllable length, SP duration a defined variables.

here, what means that filled pauses behave in a similar way than silent pauses, except for these negative correlations. Therefore, it could be possible to apply same rules and regression proposed here for filled pauses to silent ones.

6. Conclusion

In the present paper, we have studied the rhythm of filled pauses. Filled pauses may or not contain silent pauses before and after them. Despite in the database used here it appears in few cases, it is plausible to insert both silences in order to avoid coarticulation problems.

The main issue of the study presented here was to find significant correlations between a set of rhythm features in order to be able to predict filled pauses duration and rhythm related effects.

It has been found that when a filled pause is produced there is not any significant rhythm change in the sentence. However, a prepausal lengthening similar to the one produced before silent pauses is produced. The length of this syllable is correlated with the sentence rhythm. In addition, the filled pauses duration is correlated with the prepausal syllable length as well as with the sentence rhythm.

These both findings plus the evidence that global sentence rhythm is not affected by the filled pause presence, leded us to propose a duration model for speech synthesis. It is linear regression model able to predict prepausal length based on the sentence rhythm, and filled pause duration is predicted using the previously predicted prepausal length and the sentence rhythm

by means of another linear regression model.

Informal tests have shown a noticeable improvement with respect to the previously proposed method in [13].

7. References

- [1] D. Mostefa, M.-N. Garcia, O. Hamon, and N. Moreau, "Deliverable 16: Evaluation report," ELDA, Tech. Rep., Sept. 2006. [Online]. Available: <http://www.tc-star.org>
- [2] C. L. Bennett and A. W. Black, "The blizzard challenge 2006," in *Proceedings of Blizzard Challenge 2006 Workshop*, 2006, Pittsburgh, PA. [Online]. Available: <http://www.festvox.org/blizzard/blizzard2006.html>
- [3] M. Shröder, "Emotional Speech Synthesis: A Review," in *Proceedings of Eurospeech*, vol. 1, Sept. 2001, pp. 561–564, Aalborg, Denmark.
- [4] C. Gobl, E. Bennet, and A. N. Chasaide, "Expressive synthesis: How crucial is voice quality," in *Proceedings of IEEE Workshop on Speech Synthesis*, Sept. 2002, pp. 91–94, Santa Monica, California.
- [5] S. Werner and R. Hoffman, "Pronunciation variant selection for sptaneous speech synthesis - a summary of experimental results," 2006, dresden, Germany.
- [6] J. E. F. 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.
- [7] S.-C. Tseng, "Grammar, prosody and speech disfluencies in spoken dialogues." Ph.D. dissertation, Department of Linguistics and Literature, University of Bielefeld, Apr. 1999.
- [8] H. H. Clark, "Speaking in time," *Speech Communication*, vol. 36, no. 1-2, pp. 5–13, Jan. 2002.
- [9] J. E. F. Tree, "Listeners' uses of *um* and *uh* in speech comprehension," *Memory & Cognition*, vol. 29, no. 2, pp. 320–326, 2001.
- [10] M. Watanabe, K. Hirose, Y. Den, and N. Minematsu, "Filled pauses as cues to the complexity of following phrases," in *Proc. of Eurospeech*, September 2005, pp. 37–40, lisbon, Portugal.
- [11] S. Sundaram and S. Narayanan, "An empirical text transformation method for spontaneous speech synthesizers," in *Proc. of Eurospeech*, Sept. 2003, Geneva, Switzerland.
- [12] R. Carlson, K. Gustafsson, and S. Strangert, "Modelling hesitation for synthesis of spontaneous speech," in *Proceedings of Speech Prosody 2006*, Dresden, may 2006. [Online]. Available: <http://www.speech.kth.se/prod/publications/files/1087.pdf>
- [13] J. Adell, A. Bonafonte, and D. Escudero, "Disfluent speech analysis and synthesis: a preliminary approach," in *Proc. of 3th International Conference on Speech Prosody*, May 2006, dresden, Germany. [Online]. Available: <http://gps-tsc.upc.es/veu/personal/jadell/>
- [14] G. A. Toledo, *El ritmo en el español : estudio fonético con base computacional*, ser. Biblioteca románica hispánica ; II. Estudios y ensayos., Gredos, Ed., 1988, no. 361.
- [15] A. Bonafonte, P. D. Agüero, J. Adell, J. Pérez, and A. Moreno, "Ogmios: The UPC text-to-speech synthesis system for spoken translation," in *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006, pp. 199–204.