Experiments on automatic detection of filled pauses using prosodic features

H. Medeiros[1,2], H. Moniz[1,3], F. Batista[1,2], I. Trancoso[1,4], H. Meinedo[1]

[1]Spoken Language Systems Lab - INESC-ID, Lisbon, Portugal
[2]ISCTE - Instituto Universitário de Lisboa, Portugal
[3]FLUL/CLUL, Universidade de Lisboa, Portugal
[4]IST, Lisboa, Portugal

Abstract. This paper reports our first experiments on automatic detecting *filled pauses* from a corpus of university lectures using segmentation and prosodic features. This domain is informal, mostly spontaneous speech, and therefore difficult to process. *Filled pauses* correspond 1.8% of all the words in the corpus, and to 22.9% of all disfluency types, being the most frequent type. The detection of *filled pauses* has a great impact on the speech recognition task, since their detection has impact on the recognition of adjacent context and on the improvement of language models.
We have performed a set of machine learning experiments using automatic prosodic features, extracted from the speech signal, and audio segmentation provided by the ASR output. For now, only force aligned transcripts were used, since the ASR system is not well adapted to this domain. The best results were achieved using J48, corresponding to about 61% F-measure. Results reported in the literature are scarce for this specific domain, and are significantly better for other languages. The proposed approach was compared with the approach currently in use by the in-house speech recognition system and promising results are achieved. This study is a step towards automatic detection of *filled pauses* for European Portuguese using prosodic features. Future work will extend this study for fully automatic transcripts, and will also tackle other domains, also exploring extended sets of linguistic features.

Keywords: Filled pauses, University lectures, Statistical methods, Prosody

1 Introduction

Disfluencies are linguistic mechanisms used for online editing a message. These phenomena have been studied in several areas such as Linguistics, Psycholinguistics, Text-to-speech and in Automatic Speech Recognition (ASR). The ASR system is often a pipeline with several modules where each one feeds the subsequent ones with more levels of information. Disfluencies are known to have impact in the ASR modules, since they are

frequently misrecognized and may also lead to the erroneous classification of adjacent words. This has impact in several natural language processing tasks such as part-of-speech tagging, audio segmentation, capitalization, punctuation, summarization, speech translation, etc. It is also known that if disfluencies are recognized, then more reliable transcripts are provided, allowing, for example, to disambiguate between sentence-like units and also between those units and disfluency boundaries, and for speaker characterization.

Although the ASR system has to account for all the disfluent categories (*filled pauses*, *prolongations*, *repetitions*, *deletions*, *substitutions*, *fragments*, *editing expressions*, *insertions,* and *complex sequences*), the focus of the present study is on the detection of *filled pauses*. *Filled pauses* are amongst the most frequent disfluent types produced. In European Portuguese they are mostly characterized by: i) an elongated central vowel only (orthographically transcribed as "aa"; (ii) a nasal murmur only ("mm"); and (iii) a central vowel followed by a nasal murmur ("aam"). *Filled pauses* display several communicative functions, *e.g.*, evidence for planning effort, mechanisms to take or hold the floor, to introduce a new topic in a dialogue. Figure 1 shows an example of a disfluent sequence containing the filled pause *aa*/"uh".

vou <a Lisboa **aa**> ao Porto
I'm going <to Lisbon **uh**> to Oporto

Fig. 1. Example of a disfluent sequence.

In this study, *filled pauses* are predicted in a corpus of university lectures in European Portuguese (EP) based on a feature-set from both the speech signal itself and also from automatically extracted prosodic information. The specific domain is very challenging, mainly because we are dealing with quite informal lectures, contrasting with other corpus already collected of more formal seminars [4].

Our in-house recognizer [8] performs automatic identification of *filled pauses* with the aim of filtering and including rich transcripts for broadcast news. The filtering process was achieved by identifying speech regions with plateau pitch contours and energy values. The inclusion process was exclusively based on the integration of *filled pauses* in the lexicon with alternative pronunciations. The experiments reported are a step forward in the prediction of *filled pauses* by means of encompassing broader set

of acoustic features and also by using distinct classification methods to evaluate the best performance achieved.

This paper is organized as follows: Section 2 overviews the literature concerning *filled pauses* and statistical methods used to predict such events. Section 3 describes the corpus analyzed in our experiments. Section 4 discriminates the acoustic feature set applied. Section 5 presents the evaluation metrics and the results for the identification of *filled pauses*. Section 6 presents concluding remarks and the future work.

2 Related work

Literature on *filled pauses* points out to several features used for predicting such events. [13] shows that *filled pauses* have low pitch and plateau or falling tones. [14,15] evidence that *filled pauses* can be fairly detected using prosodic features related to duration, silent pauses, and pitch. The work of [2] describe experiments on 100 utterances extracted from a Japanese spoken language corpus [3]. Based solely on the analysis of the speech signal, focusing on small and constant pitch transitions and small spectral envelope deformations, this study achieves 91.5% precision and 84.9% recall on filled pause detection. [17] explore the interplay between *filled pauses* and discourse structure based on Dutch spontaneous monologues from a corpus of 45 minutes of speech containing 310 *filled pauses*. This work reports that stronger breaks in the discourse are more likely to co-occur with *filled pauses* than do weaker ones, that *filled pauses* at stronger breaks also tend to be segmentally and prosodically different from the other ones and they have more often silent pauses preceding and following them.

[20] perform experiences on a corpus of Greek university lectures of approximately 7 hours containing 1124 occurrences of *filled pauses*. They report the utility of video information for improving precision and recall on the task of detecting *filled pauses*, achieving a precision rate of 99.6% and recall rate of 84.7%. This represents a considerable improvement over the 98.5% precision and 80.6 recall achieved using solely the audio stream.

For European Portuguese (EP), studies on the characterization of disfluent events are mostly related to *filled pauses* and prolongations [11,21]. [11] accounts for the characterization of *filled pauses* and prolongations based on oral presentations in school context, showing that such events are segmentally different from regular words, since they do not seem do undergo the same sandhi processes or contextual variation. [10] perform an overall characterization of disfluent phenomena for EP. The specific

speech domain addressed in this paper is studied in [19], showing a very high percentage of errors that can be attributed to misrecognized *filled pauses*.

3 Data

Table 1. Properties of the Lectra training subset.

Corpus subset →	train+dev	test
Time (h)	28:00	3:24
Number of sentences	8291	861
Number of disfluencies	8390	950
Number of words (including *filled pauses)*	216435	24516
Number of *filled pauses*	3998	388
Proportion of *filled pauses*	1.85%	1.58%

The data used in this work corresponds to the LECTRA corpus, a corpus of university lectures in European Portuguese originally created for producing multimedia content and to aid hearing impaired students [18]. The lectures vary from 30 to 90 minutes long, and are mainly characterized by both prepared non-scripted and spontaneous speech. From this corpus, a subset of 28 hours was used for training and tuning our models while the remaining portion was used for testing. Table 1 presents overall statistics about the data, including statistics concerning *filled pauses* in each one of the subsets.

The material available for this corpus comprises: i) the raw ASR transcription, ii) an orthographically enriched version of the ASR transcription with information such as punctuation marks, disfluencies, inspirations, etc, iii) forced-aligned transcripts automatically produced by the ASR Audimus [9]. All the information is stored in self contained XML files. The XML files contain all the information coming from the ASR, such as confidence scores, duration associated to different units of analysis such as words, syllables and phones, and also information automatically acquired from the signal, in particular pitch and energy.

The ASR was trained for a different domain and text materials in European Portuguese to train language models for this domain are scarce. For that reason, we have decided to use the ASR in a forced alignment mode, reducing the number of misrecognized words and unbiasing this study by using an out of domain recognizer.

4 Feature Set

An XML parser was specially created with the purpose of extracting and calculating features from the XML files described in the previous section. The following features were extracted either for the current word (cw) or for the following word (fw): $conf_{cw}$, $conf_{fw}$ (ASR confidence scores), dur_{cw}, dur_{fw} (word durations), $phones_{cw}$, $phones_{fw}$ (number of phones), syl_{cw}, syl_{fw} (number of syllables), $pslope_{cw}$, $pslope_{fw}$ (pitch slopes), $eslope_{cw}$, $eslope_{fw}$ (energy slopes), [$pmax_{cw}$, $pmin_{cw}$, $pmed_{cw}$, (pitch maximum, minimum, and median; energy median)], $emax_{cw}$, $emin_{cw}$ (energy maximum and minimum), $bsil_{cw}$, $bsil_{fw}$ (silences before the word). The following features involving two consecutive words were calculated: $equals_{pw,cw}$, $equals_{cw,fw}$ (binary features indicating whether words are equal), $sil.cmp_{cw,fw}$ (silence comparison), $dur.cmp_{cw,fw}$ (duration comparison), $pslopes_{cw,fw}$ (shape of the pitch slopes), $eslopes_{cw,fw}$ (shape of the energy slopes), $pdif_{pw,cw}$, $pdif_{cw,fw}$, $edif_{pw,cw}$, $edif_{cw,fw}$ (pitch and energy differences), $dur.ratio_{cw,fw}$ (words duration ratio), $bsil.ratio_{cw,fw}$ (ratio of silence before each word), $pmed.ratio_{cw,fw}$, $emed.ratio_{cw,fw}$ (ratios of pitch and energy medians). Features expressed in brackets were used only in preliminary tests, but their contribution was not substantial and therefore, for simplification, they were not used in subsequent experiments. It is important to notice that some of the information contained in the features that were not used in subsequent experiments is already encoded by the remaining features, such as slopes, shapes, and differences.

Pitch slopes were calculated based on semitones rather than raw frequency values. Slopes in general were calculated using linear regression. Silence and duration comparisons assume 3 possible values, expanding to 3 binary features: $>$ (greater than), $=$ (equal), or $<$ (less than). The pitch and energy shapes expand to 9 binary features, assuming one of the following values {RR, R–, RF, –R, – –, –F, FR, F–, FF}, where $F = Fall$, $- = stationary$, and $R = Rise$. The ratios assume values between 0 and 1, indicating whether the second value is greater than the first. All the above features are based on audio segmentation and prosodic features, except for the feature that compares two consecutive words at the lexical level. In future experiments, we plan to replace it by an acoustic-based feature that compares two segments of speech on the acoustic level.

5 Experiments and Results

This section describes binary classification experiments aimed at automatically identifying *filled pauses*, using a set of acoustic features. The

described results were achieved using various state of the art machine learning algorithms, which are widely used in the literature for this kind of task, namely: Logistic Regression, Multilayer Perceptron, CARTs (Classification and Regression Trees) and J48. All experiments were conducted using Weka[1], a collection of open source machine learning algorithms and a collection of tools for data pre-processing and visualization.

In order to assess the performance, the following standard performance metrics are considered: Precision, Recall, F-measure and SER (Slot Error Rate) [5]. Only elements that we aim at identifying, in this case *filled pauses*, are considered as slots and used by these metrics. Hence, for example, the SER is computed by dividing the number of *filled pause* errors (misses and false alarms) by the number of *filled pauses* in the reference. The F-measure simultaneously accounts for precision and recall in a way that the overall error is not considered. For that reason, the SER may be a more meaningful metric for our task, which will be grater than 100% whenever the number of errors exceed the number of existing slots. Receiver Operating Characteristic (ROC) is also used and consists of plotting the false alarm rate in the horizontal axis, while the correct detection rate is plotted on vertical [1]. This performance metric provides a notion of the relation between the amount of risk taken and the amount of correct classifications. The results related to this metric account for the area bellow the curve.

Table 2. Performance analysis on predicting *filled pauses*.

	time(sec.)		overall perf.		detailled slot performance							
	train	test	acc.	kappa	cor	ins	del	prec	rec	F	SER	ROC
ZeroR	0.1	2.1	98.42	0.00	0	0	388					0.50
Logistic Regression	33	2.1	98.74	0.47	139	59	249	**70.2**	35.8	47.4	79.4	0.98
Multilayer Perceptron	3516	7.9	98.71	0.55	201	129	187	60.9	51.8	56.0	81.4	0.97
J48	1800	1.9	98.87	0.60	217	107	171	67.0	**55.9**	**61.0**	**71.6**	
Simple Cart	1257	1.7	98.82	0.55	179	80	209	69.1	46.1	55.3	74.5	

Table 2 summarizes the results achieved using all the different methods. The first line of the table corresponds to a baseline obtained by ignoring all predictors (features described in Section 4) and orienting all classifications towards the most common element, regular words. The methods

[1] Weka version 3-6-8. http://www.cs.waikato.ac.nz/ml/weka

comprise two probabilistic classifiers: Logistic Regression (LR) and Multilayer Perceptron (MLP), and two methods based on trees: Simple Cart (CART) and J48. The first two columns of the table report on the time (in seconds) taken for training and testing the models, revealing that LR is considerably faster when compared with the other methods, being 38x faster than any other method. The percentage of correctly classified instances (acc.) and kappa represent overall performance values, which take into account all classification outcomes and not only slots. The relatively high accuracy values reveal that data is highly unbalanced, since regular words correspond to 98.42% of the total events, which difficults the classification task. The values presented in the remaining columns consider only slots and correspond to more meaningful performance metrics for the current task. The number correct (corr), inserted (ins), and deleted (del) slots provide the basis for calculating the precision (prec), recall (rec), F-measure (F) and Slot Error Rate (SER) presented in the subsequent columns. Methods based on trees do not provide probabilities over the classes and for that reason the corresponding ROC area cannot be fairly computed.

Results clearly show that J48 is the best suited method for this task, achieving simultaneously the highest percentage of overall correct classifications and the best performance when considering slots only. The number of correct instances is significantly above any other method, while keeping the number of insertions and deletions quite low. Logistic Regression achieved the best precision, but the corresponding recall is the lowest, leading to a very low F-measure. Both methods based on trees take roughly half the time taken by the Multilayer Perceptron while achieving the best performance.

5.1 Feature analysis

The resulting tree has 498 leaves, but the top most decisions in the tree lead to the set of the most informative features. All the features in this set are concentrated in the characterization of the unit "current word". Thus, the features sorted by order of relevance are the following: i) confidence score of the current word; ii) current word is composed of a single phone and is lengthier than the following word; and iii) current word has adjacent silent pauses, plateau pitch contours; and iv) current word energy maximum.

The features discriminated above are in line with findings for English, reported by [13,15], a.o., since adjacent silent pauses, plateau pitch contours and constant energy values stand out as the most discriminant

features. The fact that European Portuguese shares with English those properties is a contribution more to cross-language comparisons. What is added by our experiments is the crucial importance of two additional features: the confidence level and the number of phones.

5.2 Comparing with our current filled pause detection system

The lexicon of our speech recognition system was recently upgraded with entries containing possible phonetic sequences for a *filled pause*. Such additional entries made it possible to automatically detect *filled pauses*. This study proposes an alternate way of automatically identifying such events based of both the existing audio segmentation given by the recognizer and additional prosodic features. These experiments aim at assessing whether an approach that uses prosodic features may be useful for extending our current system.

Table 3. Current ASR results.

	Cor	Ins	Del	Prec	Rec	F	SER
Current ASR system	223	146	140	60.4	**61.4**	60.9	78.8
J48	217	107	171	**67.0**	55.9	61.0	**71.6**

The performance of the current ASR system when detecting *filled pauses* has been calculated on the same evaluation subset. Table 3 shows the corresponding results. For simplification, the best results from Table 2 were also included, allowing to easily compare the two approaches. The table reveals that results are quite similar in terms of F-measure, but somewhat different in terms of precision and recall. While the current ASR system tends to perform similarly in terms of precision and recall, using J48 with our features achieves a significantly higher precision.

The results achieved are a parallel way to assess the prediction of *filled pauses* in the ASR system. Although the set of prosodic features proved to be very informative and outperformed the system in the precision values, it is also noteworthy the impact of including *filled pauses* in the lexicon with alternative pronunciations. Results, thus, suggest that combining both approaches may lead to better performances.

6 Conclusion and future work

This paper reports a number of experiments for automatically detecting *filled pauses* in a corpus of university lectures. Four different ma-

chine learning methods were applied and the best results turned out to be achieved by J48, a method based on decision trees. This is in accordance with the literature, which has shown that such approach is suitable for this kind of problem [14,12,15,16]. Our recent work on the detection of disfluencies [7,6] assumed that information about *filled pauses* was previously given by a manual annotation. This work is a step forward for automatically detecting disfluencies, since the performance for automatically calculating such information is now given. On the other hand, we aimed at assessing how well a system relying on prosodic features could complement or outperform the current ASR filled pause detection system. Our approach achieves a better precision while the current ASR achieves a better recall. Despite being quite similar in terms of F-measure, the SER is almost 7% (absolute) better, which might be a more appropriate metric.

The achieved results cannot be directly compared with other related work because different corpora, languages, domains, and evaluation setups are being used. From a linguistic point of view, Portuguese filled pauses are often ambiguous with very frequent functional words. For example, the filled pause "aa" is ambiguous with the preposition a/*to,for,...*; the article a/*the*; and the clitic pronoun a/*it,her*. The filled pause "aam" is also ambiguous with verbal forms such as the ones of the verb amar/*to love,* due mostly to possible vowel reduction or deletion in the word final position, as well as with *sigla*. Finally, the filled pause "mm" may be recognized as the article um/*a* or the cardinal number um/*one*. From a state-of-the-art perspective, this study does not include phone related information, part-of-speech, syntactic and other multimodal information. However, the main outcome of this study concerns the fact that prosodic features by themselves do have a strong impact in this task, comparable to accounting for these phenomena in both language and acoustic models.

One of the future directions is to combine our proposal with the current ASR system for better identifying *filled pauses* in European Portuguese. Additional lexical features will also be explored as a way of improving the filled pause detection task. *Filled pauses* are also an interesting feature to characterize speaking styles and even the speaker. Future experiments will apply the proposed system for such tasks. Finally, this work will be also performed in other domains, such as Broadcast News or map-task dialogues.

7 Acknowledgments

This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under Ph.D grant SFRH/BD/44671/2008 and project PEst-OE/EEI/LA0021/2013, by DIRHA European project FP7-ICT-2011-7-288121, and by ISCTE – IUL.

References

1. Tom Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, 2006.
2. Masataka Goto, Katunobu Itou, and Satoru Hayamizu. A real-time filled pause detection system for spontaneous speech recognition. In *In Proceedings of Eurospeech '99*, pages 227–230, 1999.
3. O. Hasegawa S. Hayamizu K. Tanaka K. Itou, T. Akiba. A japanese spontaneous speech corpus collected using automatic inference wizard of oz system. *The journal of the acoustical society of Japan*, 1999.
4. L. Lamel, G. Adda, E. Bilinski, and J. Gauvain. Transcribing lectures and seminars. In *Interspeech 2005*, Lisbon, Portugal, September 2005.
5. J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. Performance measures for information extraction. In *Proc. of the DARPA Broadcast News Workshop*, pages 249–252, Herndon, VA, Feb. 1999.
6. Henrique Medeiros, Fernando Batista, Helena Moniz, Isabel Trancoso, and Luis Nunes. Comparing Different Methods for Disfluency Structure Detection. In José Paulo Leal, Ricardo Rocha, and Alberto Simões, editors, *2nd Symposium on Languages, Applications and Technologies*, volume 29 of *OpenAccess Series in Informatics (OASIcs)*, pages 259–269, Dagstuhl, Germany, 2013. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
7. Henrique Medeiros, Helena Moniz, Fernando Batista, Isabel Trancoso, and Luis Nunes. Disfluency detection based on prosodic features for university lectures. In *Proc. of Interspeech 2013 (accepted)*, 2013.
8. H. Meinedo. *Audio Pre-processing and Speech Recognition for Broadcast News*. PhD thesis, Technical University of Lisbon, 2008.
9. Hugo Meinedo, Diamantino Caseiro, João Neto, and Isabel Trancoso. Audimus.media: a broadcast news speech recognition system for the european portuguese language. In *Proceedings of the 6th international conference on Computational processing of the Portuguese language*, PROPOR'03, pages 9–17, Berlin, Heidelberg, 2003. Springer-Verlag.
10. Helena Moniz, Fernando Batista, Isabel Trancoso, and Ana Isabel Mata da Silva. Prosodic context-based analysis of disfluencies. In *Interspeech 2012*, 2012.
11. Helena Moniz, Ana Isabel Mata, and Céu Viana. On filled-pauses and prolongations in european portuguese. *INTERSPEECH*, 2007.
12. C. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America (JASA)*, (95):1603–1616, 1994.
13. Douglas O'Shaughnessy. Recognition of hesitations in spontaneous speech. In *Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing - Volume 1*, ICASSP'92, pages 521–524, Washington, DC, USA, 1992. IEEE Computer Society.

14. E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California, 1994.
15. Elizabeth Shriberg, Rebecca Bates, and Andreas Stolcke. A prosody-only decision-tree model for disfluency detection. In *Proc. EUROSPEECH*, pages 2383–2386, 1997.
16. Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Mari Ostendorf, Dilek Hakkani, Madelaine Plauche, Gokhan Tür, and Yu Lu. Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proc. of the ICSLP*, volume 5, Sydney, November 1998.
17. M. Swerts. Filled pauses as markers of discourse structure, 1998.
18. Isabel Trancoso, Rui Martins, Helena Moniz, Ana Isabel Mata, and Céu Viana. The lectra corpus - classroom lecture transcriptions in european portuguese. In *LREC*, 2008.
19. Isabel Trancoso, Ricardo Nunes, and Luís Neves. Recognition of classroom lectures in european portuguese. *Interspeech*, 2006.
20. Vassilis Tsiaras, Costas Panagiotakis, and Yannis Stylianou. Video and audio based detection of filled hesitation pauses in classroom lectures, 2009.
21. A. Veiga, S. Candeias, C. Lopes, and F. Perdigão. Characterization of hesitations using acoustic models. In *International Congress of Phonetic Sciences - ICPhS XVII*, volume -, pages 2054–2057, August 2011.