DISFLUENCY DETECTION ACROSS DOMAINS

Helena Moniz[1,2], Jaime Ferreira[1], Fernando Batista[1,3], Isabel Trancoso[1,4]

[1]Spoken Language Systems Lab - INESC-ID, Lisbon, Portugal

[2]FLUL/CLUL, Universidade de Lisboa, Portugal

[3]ISCTE - Instituto Universitário de Lisboa, Portugal

[4]Instituto Superior Técnico, Universidade de Lisboa, Portugal

helenam;jaime.ferreira;fmmb;isabel.trancoso@l2f.inesc-id.pt

ABSTRACT

This paper focuses on disfluency detection across distinct domains using a large set of openSMILE features, derived from the Interspeech 2013 Paralinguistic challenge. Amongst different machine learning methods being applied, SVMs achieved the best performance. Feature selection experiments revealed that the dimensionality of the larger set of features can be further reduced at the cost of a small degradation. Different models trained with one corpus were tested on the other corpus, revealing that models can be quite robust across corpora for this task, despite their distinct nature. We have conducted additional experiments aiming at disfluency prediction in the context of IVR systems, and results reveal that there is no substantial degradation on the performance, encouraging the use of the models in IVR domains.

Keywords: Disfluency detection, acoustic-prosodic features, cross-domain analysis, European Portuguese.

1. INTRODUCTION

Disfluencies are on-line editing strategies with several (para)linguistic functions. They account for a representative portion of our spoken interactions. Everyday we are analists of our own speech and of others, monitoring distinct linguistic and paralinguistic factors in our communications, using disfluencies to make speech a more error-free system, a more edited message, and a more structured system with coherent and cohesive mechanisms.

Disfluencies are an important research topic in several areas of knowledge, namely, Psycholinguistics, Linguistics, Automatic Speech Recognition, and more recently in Text-to-Speech conversion and even in Speech-to-Speech translation. Yet, whereas for several languages one can find much literature on disfluencies, for others, such as European Portuguese, the literature is quite scarce.

Detecting and filtering disfluencies is one of the hardest problems in rich transcription of spontaneous speech. Enriching speech transcripts with structural metadata [15] is of crucial importance for many speech and language processing tasks, and comprises several metadata extraction/annotation tasks besides dealing with disfluencies such as: speaker diarization (i.e. assigning the different parts of the speech to the corresponding speakers); sentence segmentation (also known as sentence boundary detection); punctuation and capitalization recovery; topic and story segmentation. Such metadata extraction/annotation technologies are recently receiving increasing attention [7, 6, 15], and demand multi-layered linguistic information to perform such tasks. A simple segmentation method, for instance, may rely only on information about pauses. More complex methods, however, may involve, e.g., lexical cues, dialog acts cues. In fact, the term *structural segmentation* encompasses all algorithms based on linguistic information that delimit *spoken sentences* (units that may not be isomorphic to written sentences), topics and stories.

Acoustic-prosodic features have been pervasive in disfluency prediction tasks, mostly due to the fact that ASR systems are still not mature enough to account for spontaneous speech in all its complexity. Therefore, the lexical output of a recognizer may not be reliable to train models to account for highly spontaneous data, as the datasets used in this work. Generally, the acoustic-prosodic set of features described for disfluency detection encompass: slower speech rate, lengthier adjacent silent pauses, higher values of spectral tilt, jitter, shimmer, pitch and energy differences relatively to the adjacent contexts (*e.g.*, [13, 18, 19]).

Previous experiments conducted for European Portuguese [12, 1, 9, 8, 11] reported that different structural regions of a disfluency can be predicted based exclusively on discriminative acoustic-prosodic features of disfluencies. Building on that, we now aim at a cross-domain disfluency prediction. Therefore, the aim of this work is threefold: to characterize the acoustic-prosodic features of disfluencies; to evaluate the impact of acoustic features within and across domains. In order to understand the realms of potential transversality in disfluency prediction tasks in dialogues in European Portuguese, domain-specific models will be created for each one of the different scenarios, and then those models will be reused across domains. The literature on cross-domain analysis of disfluencies in dialogues is, in general, quite scarce. It is, therefore, not clear, how one may transpose findings from one domain into another in human-human dialogues. Ultimately, this work may be seen as a contribution to the scarce studies on disfluencies in human-human dialogues, on the most salient features, and on the common vs. distinct properties of such phenomena.

2. DATASETS

This work focuses on two domains to predict disfluencies, namely: university lectures and dialogues. The choice of the corpora was influenced by the availability of large amounts of (highly spontaneous) transcribed data in European Portuguese for these two domains. Both corpora are available through ELRA.

LECTRA (ELRA-S0366) is a university lectures corpus, collected within the national project LECTRA (LECture TRAnscriptions in European Portuguese) [20], aiming at producing multimedia contents for e-learning applications, and also at enabling hearing-impaired students to have access to recorded lectures. It includes seven 1-semester courses, six of them recorded in the presence of students, and only one recorded in a quiet environment. Most classes are 60-90 minutes long. The initial set of 21 hours orthographically transcribed was recently extended to 32 hours [16]. The corpus was divided into train+development (89%) and test (11%). The sets include portions of each one of the courses and follow a temporal criterion, meaning the first classes of each course were included in the training and development sets, whereas the final ones were integrated into the test sets.

CORAL (ELRA-S0367) is a corpus of map-task dialogues [21]. One of the participants (giver) has a map with some landmarks and a route drawn between them; the other (follower) has also landmarks, but no route and consequently must reconstruct it. In order to elicit conversation, there are small differences between the two maps: one of the landmarks is duplicated in one map and single in the other; some landmarks are only present in one of the maps; and some are synonyms. The 32 speakers were divided into 8 quartets and in each quartet organized to take part in 8 dialogues, totaling 64 dialogues, which corresponds to 9h (46k words). The manual annotation of the last two quartets was finished very recently and could not be used in the scope of this work.

The corpora were multilayered annotated, as described in more detail in [11]. The disfluency annotation followed mostly [18, 2]. For the purposes of the present work, sentence-like units were transformed into "chunks" and sequences of distinct types of disfluencies were merged into "disfluent sequences". A binary classification will allow, at a first instance, to discriminate the most important classes. One of the goals is to provide a first evaluation of cross-domain characteristics in order to apply to an IVR system. The datasets comprise manual transcripts and force aligned transcripts, produced by the in-house ASR Audimus [14].

Overall statistics of the corpora are presented in Table 1. Chunks correspond to distinct types of sentence-like units (full stop, question mark, comma) and disfluent sequences encompass all the different types of disfluencies (e.g., filled pauses, complex sequences of disfluencies, repetitions, inter alia). The percentage of disfluent sequences in the university lectures corpus is of 28.5% in the train set and of 29.3% in the test set. The percentage of disfluent sequences in the dialogue corpus is relatively smaller, corresponding to 21%. These results, at a first inspection, point out to dialogues having fewer disfluencies than lectures. However, when considering more refined measures, e.g., the number of sentence-like units (either fluent or with disfluencies) uttered per minute, there are more sentences of both types per minute in dialogues than in lectures, clearly supported on the fact that dialogues have fewer words in both SUs, motivated by temporal constraints in the interaction between interlocutors. Further details on the characetrization of disfluencies in both corpora can be found in [11].

Table 1: Overall characteristics of the datasets.

	Lectra		Coral
	Train	Test	
Chunks	16569	4194	7968
Disfluent sequences	6619	1737	2120
% disfluent sequences	28.5%	29.3%	21.0%

Table 2: Performance for cross-validation and test set.

	C-value	Cross-validation	Test set	
		Acc.	Acc.	Kappa
OS-F	0.01	86.84%	84.60%	0.646
AS-F	1.0	85.51%	83.98%	0.629

3. EXPERIMENTS

The goal of this section is to find a set of features that allows us to discriminate disfluent segments from non-disfluent ones. The reported experiments are based on an automatic segmentation, provided by the in-house ASR system [10]. We have used the large set of openSMILE features from the Interspeech 2013 Paralinguistic challenge. The openSMILE toolkit [3] is capable of extracting a very wide range of speech features and has been applied with success in a number of paralinguistic classification tasks and for disfluency prediction **[17]**. It has been used in the scope of this study to extract a feature vector containing 6125 speech features (henceforth denoted as OS-F) by applying segment-level statistics (means, moments, distances) over a set of energy, spectral and voicing related frame-level features.

Different classification methods from the Weka toolkit [4] have been applied, including: Naïve Bayes, Logistic Regression, Decision trees, Classification and Regression trees, and Support Vector Machines (SVM). However, results reported in this section refer to SVM only, which have almost consistently achieved the best performances. The SVM has been setup to use Sequential Minimal Optimization with a Linear kernel as the training algorithm.

3.1. University lectures

Table 2 presents a summarized view of the results presented in this section. The cross-validation column shows the accuracy performance achieved using only the training data, with 10 folds, and offers useful insights about the model performance. The best experimented complexity of the model, expressed by the parameter C, was also

calculated based on training data only, and the resulting model was applied to the test set. The Kappa statistic is a chance-corrected measure of agreement between the classifications and the true classes. A value close to zero indicates that results could be achieved almost by chance, whereas a value close to 1.0 means an almost complete agreement.

The first row of Table 2 show results achieved with the initial set of features. The OS-F feature set comprises a very high number of features, most of them possibly not useful for our specific task. In this context, the need to search for a smaller subset, without lowering the classification performance, is crucial. Having a subset of these features, selected by their usefulness to distinguish the disfluent and non-disfluent classes is useful only if it allows to have simpler models that are equally discriminative. To this end, we automatically selected features using the WEKA's implementation on correlation-based feature subset selection [4], and used a best first search strategy. This approach evaluates the relevance of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Therefore, subsets of features that are highly correlated with the class, while having low intercorrelation, are preferred. Selecting over the training data gives us a total of 222 automatically selected features (AS-F) that best discriminate between the two classes of segments. Results achieved with this subset of features are presented in the second row of Table 2. Apart from a small degradation in accuracy performance, this subset allowed us to use a higher complexity model and significantly reduced the time required to train and test the model.

3.2. Cross-domain experiments

To test the robustness of our features in a cross domain scenario we created and applied models across corpora. We have started by creating an additional model using the CORAL corpus for training. These models were then applied to the LECTRA test set, and Table 3 shows the corresponding results. The achieved results are about 9% lower than the corresponding results achieved using the model created with the LECTRA training data (*vide* Table 2). Two main factors contribute to interpret these results, as shown in [11] on an analysis of speaking style effects in the production of disfluencies, considering the same corpora studied in the present work. Firstly, the distributional patterns of disfluencies evidenced that the selection of disfluency types is corpus dependent. Excluding filled pauses, the remaining disfluency categories have different distributional patterns. In dialogues, speakers produce more often repetitions and fragments than in lectures. In lectures, teachers prefer complex sequences of disfluencies (mostly repetitions and substitutions used for lexical search). Those strategies were associated with teachers having more time to edit their speech, displaying strategies associated with more careful word choice and speech planning, whereas dialogue participants had stricter time constraints. Secondly, the prosodic parameters analyzed showed that, although there is a cross corpora prosodic contrast marking between disfluency/fluency repair, there are significant differences in the degrees of contrast made in both corpora. Lectures exhibit the highest pitch maxima in all units of analysis, whereas dialogues exhibit the highest energy maxima. As shown in Tables 3 and 4, there is, in fact, an impact of speaking styles in automatic disfluency detection, however the features that best characterize the behavior of the disfluent sequences in the dialogue corpus fairly predict the disfluencies in the university lectures corpus.

Table 3: CORAL models applied to LECTRA test set.

	C	CORAL	LECTRA test set	
		Cross-val. Acc.	Acc.	Kappa
OS-F	0.01	82.00%	75.74%	0.495
AS-F	1.0	80.71%	74.85%	0.484

Table 4: LECTRA models applied to CORAL.

	Upperbound	CORAL	
	Acc.	Acc.	Kappa
OS-F	82.00%	74.25%	0.393
AS-F	80.71%	71.36%	0.353

The same parameters used in previous experiments were applied to the models used in Table 3. As we did not adjust the parameters, the performance reported for the cross-validation can be assumed to be an upper-bound for the CORAL data. Finally, we have applied our initial models, trained using the LECTRA training data, to the CORAL corpus. Table 4 shows the corresponding results, where the cross-validation performance achieved previously serve as an upperbound.

In the scope of an European Project and aiming at disfluency prediction in the context of IVR systems, the original disfluency detection models trained with 16kHz full bandwidth were downsample to 8kHz, with the use of the telephone simulator FaNT [5]. The main goal is to compare the performance between full and telephone bandwidth, replicating the same conditions of the previous experiments, using the openSMILE features, with the C parameter at 0.01, using the test set of LECTRA and the k-fold cross-validation (k=10) for CORAL. The results of the experiments with the telephone bandwith are displayed in Table 5, revealing no substantial degradation on the performance and encouraging the models' use in IVR domains.

The most informative features include the following: *mfcc_sma[6]_quartile3*, *mfcc_sma[14]_amean*, *mfcc_sma[12]_skewness*, *audspec_lengthL1norm_sma_lpc3*, and *pcm_Mag_harmonicity_sma_de_iqr1-3*. This selec-

Table 5: Telephone bandwith models.

	Acc.	Kappa
LECTRA	85.95%	0.64
CORAL	86.75%	0.54
LECTRA to CORAL	80.87%	0.45
CORAL to LECTRA	80.05%	0.55

tion points out to the importance of the MFCCs features and for the audiospectral differences. It is known that MFCCs highly contributes to distinct types of tasks, being quite transversal in a plethora of speech prediction tasks, what is specific of disfluencies are the audiospectral differences, in line with [19].

4. CONCLUSIONS AND FUTURE WORK

We have performed a disfluency detection task using openSMILE features, extensively used in Interspeech paralinguistic challenges. Amongst different machine learning methods being applied, Support Vector Machines achieved the best performance. Feature selection experiments revealed that the dimensionality of a large set of features can be further reduced at the cost of a small degradation (about 1% absolute). The robustness of the features across domains was investigated using university lectures and dialogues. Different models trained with one corpus were tested on the other, revealing that models can be quite robust across corpora for this task, despite their distinct nature. Models trained using university lectures achieved about 74% accuracy on the dialog corpus, and about 76% accuracy in the opposite direction. This later results is possibly due to the fact that CORAL contains more contrastive tempo characteristics, shares with LECTRA most of the pitch and energy patterns on disfluent sequences and therefore a model created with such data generalizes better. In the scope of an European Project, we have conducted additional experiments aiming at disfluency prediction in the context of IVR systems. Results show no substantial degradation on the performance, encouraging the use of the models in IVR domains.

Future work will encompass merging features now being extracted with openSMILE with knowledge-based features in order to their contribution in cross-domain disfluency prediction. Our work will also tackle other spontaneous domains and cross-language studies in human-machine interfaces.

5. ACKNOWLEDGEMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, under Post-doc grant SFRH/PBD/95849/2013, and by EU-IST FP7 project SpeDial, under contract 611396.

6. REFERENCES

[1] Batista, F., Moniz, H., Trancoso, I., Mamede, N., Mata, A. I. 2012. Extending automatic transcripts in a unified data representation towards a prosodic-based metadata annotation and evaluation. *Journal of Speech Sciences* (3), 115–138.

[2] Eklund, R. 2004. *Disfluency in Swedish Human-Human and Human-Machine Travel Booking Dialogues*. PhD thesis University of Linköpink.

[3] Eyben, F., Wollmer, M., Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. ACM, , (ed), *Proceedings of the international conference on Multimedia, MM '10* New York, NY, USA. 1459–1462.

[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. 2009. The weka data mining software: an update. *SIGKDD Explororations* 11(1), 10–18.

[5] Hirsch, H., Finster, H. 2005. The simulation of realistic acoustic input scenarios for speech recognition systems. *Proceedings of Interspeech 2005* 2697–2700.

[6] Jurafsky, D., Martin, J. H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR second edition.

[7] Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Harper, M. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transaction on Audio, Speech and Language Processing* 14(5), 1526–1540.

[8] Medeiros, H., Batista, F., Moniz, H., Trancoso, I., Nunes, L. 2013. Comparing different machine learning approaches for disfluency structure detection in a corpus of university lectures. ACM, , (ed), *SLATE 2013* Oporto, Portugal. 259–269.

[9] Medeiros, H., Moniz, H., Batista, F., Nunes, L., Trancoso, I. 2013. Disfluency detection based on prosodic features for university lectures. *Interspeech 2013* Lyon, France. 2629–2633.

[10] Meinedo, H., Viveiros, M., Neto, J. 2008. Evaluation of a live broadcast news subtitling system for Portuguese. *Interspeech* Brisbane, Australia.

[11] Moniz, H., Batista, F., Mata, A. I., Trancoso, I. December 2014. Speaking style effects in the production of disfluencies. *Speech communication* 65(6), 20–35.

[12] Moniz, H., Batista, F., Trancoso, I., Mata, A. I. 2012. Prosodic context-based analysis of disfluencies. *Interspeech 2012* Portland, Oregon.

[13] Nakatani, C., Hirschberg, J. 1994. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America (JASA)* (95), 1603–1616.

[14] Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D. 2008. Broadcast news subtitling system in Portuguese. *ICASSP 2008* 1561–1564.

[15] Ostendorf, M., Favre, B., Grishman, R., Hakkani-Tür, D., Harper, M., Hillard, D., Hirschberg, J., Ji, H., Kahn, J. G., Liu, Y., Maskey, S., Matusov, E., Ney, H., Rosenberg, A., Shriberg, E., Wang, W., Wooters, C. 2008. Speech segmentation and spoken document processing. *IEEE Signal Processing Magazine* 25(3), 59–69.

[16] Pellegrini, T., Moniz, H., Batista, F., Trancoso, I., Astudillo, R. 2012. Extension of the lectra corpus: classroom lecture transcriptions in european portuguese. *GSCP 2012* Brazil.

[17] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S. 2013. Paralinguistics in speech and language - state-of-the-art and the challange. *Computer Speech and Language* (27(1)), 4–139.

[18] Shriberg, E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis University of California.

[19] Shriberg, E. 2001. To "errrr" is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 31, 153–169.

[20] Trancoso, I., Martins, R., Moniz, H., Mata, A. I., Viana, M. C. May 2008. The Lectra corpus - classroom lecture transcriptions in European Portuguese. *LREC 2008* Marrakesh, Morocco.

[21] Trancoso, I., Viana, M., Duarte, I., Matos, G. 1998. Corpus de dialogo CORAL. *PROPOR'98* Porto Alegre, Brasil.