SPA: Web-based Platform for easy Access to Speech Processing Modules

Fernando Batista[1,2], Pedro Curto[1], Isabel Trancoso[1,3], Alberto Abad[1,3], Jaime Ferreira[1], Eugénio Ribeiro[1,3], Helena Moniz[1,4], David M. de Matos[1,3], Ricardo Ribeiro[1,2]

[1] L[2]F – Spoken Language Systems Laboratory, INESC-ID Lisboa
[2] ISCTE-IUL - Instituto Universitário de Lisboa, Lisboa, Portugal
[3] Instituto Superior Técnico, Universidade de Lisboa, Portugal
[4] FLUL/CLUL, Universidade de Lisboa, Portugal
fernando.batista@inesc-id.pt, pkurto@gmail.com, isabel.trancoso@inesc-id.pt

Abstract

This paper presents SPA, a web-based Speech Analytics platform that integrates several speech processing modules and that makes it possible to use them through the web. It was developed with the aim of facilitating the usage of the modules, without the need to know about software dependencies and specific configurations. Apart from being accessed by a web-browser, the platform also provides a REST API for easy integration with other applications. The platform is flexible, scalable, provides authentication for access restrictions, and was developed taking into consideration the time and effort of providing new services. The platform is still being improved, but it already integrates a considerable number of audio and text processing modules, including: Automatic transcription, speech disfluency classification, emotion detection, dialog act recognition, age and gender classification, non-nativeness detection, hyper-articulation detection, dialog act recognition, and two external modules for feature extraction and DTMF detection. This paper describes the SPA architecture, presents the already integrated modules, and provides a detailed description for the ones most recently integrated.
Keywords: Web interface, Web services, Speech modules, Speech Analytics, REST API

1. Introduction

World Wide Web is becoming a development platform (Cervinschi et al., 2010) and web-based applications are emerging everyday. Web-based applications are fairly easy to develop, use, and maintain, thus being a way of providing access to existing and new resources for the whole community. This paper presents a web-based Speech Analytics platform – SPA – that makes it possible to use existing speech processing modules through the web.

The original motivation for the SPA platform was to answer the increasing number of requests received for transcribing audio/video files in European Portuguese. It was firstly created to provide a simple interface, easy to use by non-expert users, but requests have multiplied and diversified, namely in terms of languages (Spanish, English) and varieties covered (European, Brazilian), domains (broadcast news, interviews), etc.. On the other hand, although the majority of the SPA users are only interested in the automatic transcripts, other technology partners showed interest in obtaining information about characteristics of the speakers (e.g. male/female) and all the other metadata that could be retrieved from the speech signal (e.g. language, recording conditions, etc.). In fact, a vast number of other audio and text processing modules have been developed at the Spoken Language Lab (L[2]F), including: speech disfluency classification, emotion detection, dialog act recognition, age and gender classification, amongst others. Thus, a major goal was to develop a platform that could integrate most the existing modules as services, ease the integration of new services, and provide these modules to the user community, without the need of knowing about software dependencies, configurations and other restrictions. Having this in mind we sought to build a generic web-based platform in a way that most of the design and implementation phases are done almost automatically, maintaining

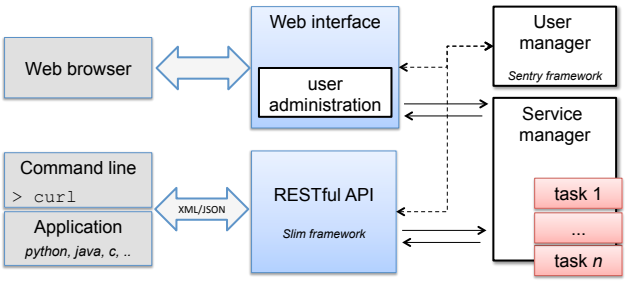

Figure 1: The SPA platform Architecture.

fundamental characteristics such as consistence, flexibility, scalability and security. The current version of the SPA platform already integrates several modules and is available at https://www.l2f.inesc-id.pt/spa. SPA was recently used in the context of the European FP7 project SpeDial (https://sites.google.com/site/spedialproject/) for sharing local resources with the project partners, following our first experience with a previous version with partners in the EUTV project.

This paper describes the SPA architecture and presents the already integrated modules, with an emphasis on the ones that were recently integrated in the context of the SpeDial project. For the most recently integrated modules, a detailed description of each module is provided, together with the corresponding evaluation for different training and test corpora. The paper is organized as follows: Section 2. presents the overall architecture of the system. Section 3. describes the speech analytics modules that can now be used through the platform. Finally, Section 4. draws the major conclusions and presents some plans for the future.

2. System Architecture

The architecture of SPA is depicted in Figure 1. The web interface provides the easiest way of accessing services through a web browser, but requests may also be performed through a REST API, which makes it possible to integrate the existing services in other applications. Both ways require authentication, handled by the User Manager. The Service Manager validates the request, creates the corresponding task with the service arguments provided, and manages its execution.

The platform has been developed taking into account the time and effort required to integrate and provide service interfaces to new modules. For that reason, we have created a framework that contains all functions that can be used to create new service pages and structured templates for new services. Also, to make this process more transparent, we have built scripts that automatic create the files needed to add new services, based on template files.

The authentication and authorization tasks are accomplished using the Sentry framework[1], developed by Cartalyst and licensed under BSD-3. This is a framework agnostic authentication and authorization system that provides a set of classes for simplifying the task of adding standard authentication and access control flows to a PHP application, without compromising security.

2.1. Access through the Web Interface

The web interface is the most easy way of accessing the integrated modules. The access through the web interface is facilitated by the existing check boxes and combo boxes that contain all possible available options. Most of the integrated services are speech analytics modules that take a single input file and produce an output, taking into account a set of parameters. The output produced becomes available as a link that can be used for downloading the data from the browser to the client. Depending on the service, the output can also be presented as text in the browser for an easy and immediate access. The Automatic Transcription service, being one of the most complex services provided, produces an automatic transcript from audio or video and also provides means for manual editing the produced output. Figure 2 shows a screenshot of the editing interface for a submitted video.

Most of the times a user is interested in using individual services, but when performing speech analytics tasks, there is sometimes the need of executing several different services for the same input file. Multiple Services in SPA allow users to obtain combined results for a single input file, either in a parallel or pipeline fashion, but for the sake of simplicity they must be manually pre-configured in the system backend. Figure 3 shows an example of a service that combines three modules and executes them in parallel. The parameters common to all the services, such as the input filename, are provided only once. Specific parameters are also available for each individual module and can be set independently.

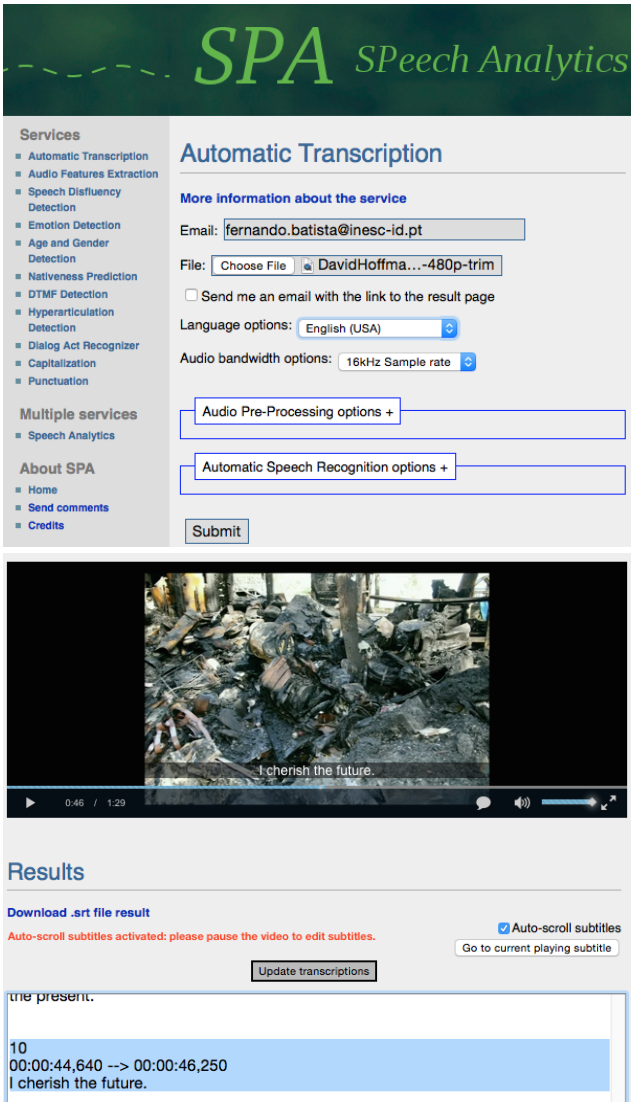

Figure 2: SPA web interface.

2.2. Access through the REST API

The REST API functionalities are accomplished with the help of the Slim framework[2], developed by Cartalyst and licensed under BSD-3. This is a micro-framework designed for development of web applications and APIs, and comes with a sophisticated URL router and support for page templates, flash messages, encrypted cookies and middleware. Slim works by defining router callbacks for HTTP methods and endpoints simply by calling the corresponding method – *get()* for GET requests, *post()* for POST requests, and so on – and passing the URL route to be matched as the first argument of the method. The final argument to the method is a function which specifies the actions to take when the route is matched to an incoming request. The functionalities implemented in the REST API are: i) Authentication/Authorization; ii) One GET method (to retrieve the result of a previous task processed); iii) A POST method for each of the available services.

The developed Rest API accepts both XML and JSON as input. The file output from the service call will have the

[1] https://cartalyst.com/manual/sentry/2.1

[2] http://www.slimframework.com/

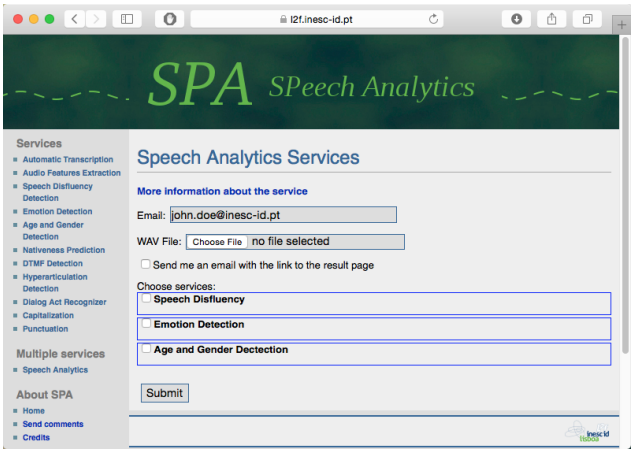

Figure 3: Running multiple modules in parallel.

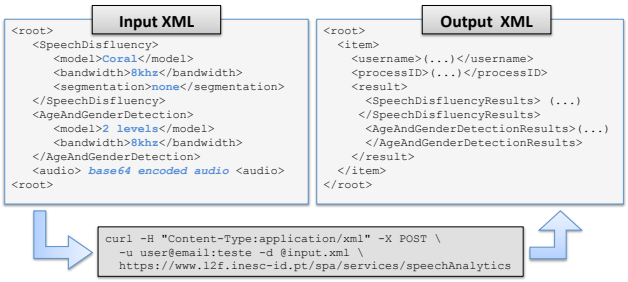

Figure 4: Using the REST API for accessing a SPA service.

same type as the input given, i.e. if the input is an XML file, the output will also be in XML. Audio files can be sent in the same request as the rest of the data, encoded in Base64, or can be simply provided as a valid URL. Figure 4 shows one example for calling a service through the REST API. The example calls a combined service named Speech Analytics that consists of multiple parallel service calls (in this case, speech disfluency, and age and gender classification).

3. Speech Analytics Services

This section describes the audio processing modules that are currently integrated in SPA, starting by the Automatic Transcription service, the oldest available module, already introduced in a first version of the platform. The remaining modules, introduced mainly in the context of the Spedial project, are then described and, whenever possible, results achieved with these modules are presented for the corpora available in the SpeDial project in English (Let's Go corpora), and Greek. When this is not possible (for instance Greek was not covered by our recognition modules), the results will be reported only for specific datasets.

The Automatic Transcription (Meinedo and Neto, 2003; Abad et al., 2008; Meinedo et al., 2010) module was introduced in the first version of the platform and can be used to create automatic transcripts from a video/audio file. As shown in Figure 2, the web interface allows to manual editing the automatic subtitles. The default models have been optimized for broadcast news captioning in several languages (Portuguese, English, Spanish). However, recognition models trained with telephone speech were also recently made available in the context of SpeDial project. This module can be combined with several others, namely with the Punctuation and Capitalization module (Batista et al., 2009; Batista and Mamede, 2011; Batista et al., 2012), in order to provide enriched transcripts.

In the context of the Spedial project, the most relevant modules are the ones that may provide potential cues for detecting hot spots in spoken dialogue systems. The following modules were developed mostly in the scope of the SpeDial project and were integrated in the platform: Emotion detection, Disfluency detection, Non-nativeness classifier, Hyper-articulation detector, gender and age classification, and Dialog Act recognition. All these modules can provide cues for detecting hot spots and are described in more detail in the following sections.

3.1. Emotion Detection

A module that may warn the system about user insatisfaction is of course emotion detection. The deployed model was trained using Let's Go 2006 data (Schmitt et al., 2012), extracted from the Let's Go Interactive Voice Response system (IVR) that provides information about bus schedules in the city of Pittsburg, through spoken telephonic interaction with a dialog system(Raux et al., 2006; Eskenazi et al., 2008).

The data, comprising approximately of 2 hours of audio, was manually annotated for the emotional state by one annotator and the annotation scheme was composed of five distinct classes, indicating various degrees of positive or negative state (friendly, neutral, slightly angry,angry, and very angry). The dataset is highly unbalanced for the neutral class, which makes it difficult to create multiclass models with such a small dataset.

The deployed module performs a binary classification: Neutral vs. Angry, which is still useful for detecting hot spots. It uses a 8Khz model that was trained using approximately 75% of the data, achieving about 80.4% accuracy on the remaining 25%.

Additional experiments have been performed using Vera am Mittag German emotional speech database (VAM) (Grimm et al., 2008), an audio-visual corpus annotated using a three dimensional emotion space concept. The FANN Toolkit was used to build an artificial network, and the achieved performance was similar to the approach described in (Grimm et al., 2007), but the corresponding model is yet to be included in the SPA platform.

3.2. Disfluency Detection

It is well known that the performance of speech recognition systems may severely degrade in the presence of disfluencies. It is also known that disfluencies have multiple (para)linguistic functions, explored in several domains, areas, applications. These were the core motivations for integrating a disfluency detection module that processes an audio segment and signals the presence of disfluencies in that segment.

Different models were created based on an automatic segmentation, provided by the in-house ASR system (Meinedo et al., 2008). We have used the large set of openSMILE features from the Interspeech 2013 Paralinguistic challenge (Eyben et al., 2010). OpenSMILE is a publicly available

tool 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 (Schuller et al., 2013). Different classification methods from the Weka toolkit (Hall et al., 2009) have been applied, including: Naïve Bayes, Logistic Regression, Decision trees, Classification and Regression trees, and Support Vector Machines (SVM). However, the best performance was achieved with SVM, which has been setup to use Sequential Minimal Optimization with a Linear kernel as the training algorithm.

The original disfluency detection models trained with 16kHz full bandwidth were downsample to 8kHz, with the use of the telephone simulator FaNT (Hirsch and Finster, 2005), since SpeDial Project targets IVR systems. The results of the experiments with the telephone bandwidth reveal no substantial degradation on the performance and encourage the models' use in IVR domains. Currently, four different models are being provided, either for 8Khz or 16Khz audio files. Two models were created using data from CORAL (ISLRN: 499-311-025-331-2), a corpus of map-task dialogues (Trancoso et al., 1998), and the other two models were created using LECTRA (ISLRN 298-379-572-530-5), a corpus of university lectures. The existing models achieve 80-83% accuracy in 10-fold cross validation scenarios. (Moniz et al., 2015) provides extended details about the deployed system and related experiments.

3.3. Non-nativeness Detection

Knowing the degree of nativeness of a user is relevant for a number of applications. For instance, such information can be used by an ASR system to swap or adapt its language models, minimizing recognition errors in the presence of non-native speech. Furthermore, it can also be used to identify causes of hot spots in the dialogue, which was also relevant for SpeDial. This module, developed during our participation in the INTERSPEECH 2015 ComPaRe challenge (Ribeiro et al., 2015a), receives an audio file and identifies the degree of nativeness of the corresponding speaker, in a continuous scale from 1 to 5. Since the challenge data was labeled in a continuous scale, we have tackled the problem using regression. Multiple approaches were applied, such as phonotactic models, i-vectors, and goodness of pronunciation (GOP), covering both segmental and suprasegmental features. The resultant information was combined in multiple ways to feed a SVM regressor. The performance achieved on the challenge test data was around 0.58 Spearman score (Ribeiro et al., 2015a).

The existing models were created based on the following corpora resources: i) datasets provided for the INTERSPEECH 2015 ComPaRe challenge (Schuller et al., 2015), including parts of the AUWL, ISLE, and C-AuDiT corpus; ii) the euTV corpus, consisting of data used to develop the euTV (Bertini et al., 2013) system for media monitoring and publishing. One of its services is able to identify the 12 most spoken languages across the European Union - English, Spanish, Polish, Greek, Portuguese, Hungarian, Czech, German, Italian, French, Dutch, and Swedish. Data was obtained from previously existing corpora used for automatic speech recognition, from the podcasts and archives made available online by the respective national radios and TV stations, and also from the podcasts and archives of the SBS (http://www.sbs.com.au/) multi-language radio site; iii) the LRE2011 corpus, consisting of data used by INESC-ID's Spoken Language Systems Laboratory to develop the language recognition systems (Abad, 2011) submitted to the 2011 NIST Language Recognition Evaluation. It comprises data from 24 different languages obtained from different sources, including the data provided for the challenge; previous LRE campaigns; and several available Linguistic Data Consortium (LDC) sets.

3.4. Hyper-articulation

Hyper-articulation is a speech adaptation phenomenon that consists of adopting a clearer form of speech, in an attempt to improve recognition levels. However, although it may work in child directed speech or when talking to people with hearing impairment, it typically has the opposite result when talking to an ASR system, decreasing its performance (Soltau and Waibel, 1998; Litman et al., 2000). This happens because ASR systems are not trained with hyper-articulated speech and, thus, are unable to perform well in its presence. Furthermore, these situations typically occur in attempts to correct previous recognition errors on unmarked speech. This means that the supposed correction will also be misrecognized, leading to further hyper-articulation and recognition errors, completely disrupting the dialogue flow (Oviatt et al., 1998). Automatic detection of hyper-articulated speech may be relevant for dialog systems, where the user can be guided towards the use of unmarked speech or alternate ASR models may be used. Automatic hyper-articulation detection can also be used to find possible causes of hot spots in a dialog, reducing the need for manual annotations.

We have tackled the problem using data from three years (2009, 2012, and 2014) of the Let's Go corpus. The data from the three years was annotated for hyper-articulation in a joint effort by KTH and INESC-ID (Lopes et al., 2016). The ComParE 2013 feature set (Schuller et al., 2013) was used to obtain a baseline. Furthermore, we used multiple acoustic-prosodic features and combined them in multiple sets, to assess their contributions for the task. Also, since many of the features are speaker dependent, we also calculated the differences between the values of the features in the turn being classified and in the first turn of the dialog. In the end, we concluded that this task benefits from the use of large feature sets, out of which the most important features can be selected using automatic methods.

In terms of classification approaches, we used both SVMs and Random Forests in our experiments. However, the latter systematically outperformed the first, achieving about 81.6% accuracy and surpassing the ones obtained by (Fandrianto and Eskenazi, 2012) on similar data, which are the only ones we are aware of for this task. However, since these results were obtained on balanced datasets, they do not reflect real situations, where hyper-articulation is rare. Thus, we also performed experiments on highly unbalanced versions of the datasets, using the same classifiers. By adjusting confidence thresholds for the classification of a turn as hyper-articulated, we were able to obtain accuracy re-

sults around 98% on every unbalanced dataset, while maintaining a high precision.

The model deployed on SPA does not take features that require information from the first turn into account, since that information is not always available. However, we intend to add the option to also upload a file corresponding to the first turn, so that the classifier can take advantage of it when available. Furthermore, since the accuracy gains were reduced in comparison to the time required to extract our acoustic-prosodic features, the deployed model relies solely on a selected subset of the ComParE 2013 feature set. In terms of the confidence threshold, we also intend to allow user customization, so that the classifier can adapt to specific situations, into which the user has more insight.

3.5. Gender/Age Classification

Our first experiments with gender classification were done in the framework of our broadcast news (BN) automatic captioning system, with the goal of reducing the load on subsequent clustering, providing more flexibility in clustering settings (for example female speakers may have different optimal parameter settings to male speakers), and supplying more side information about the speakers in the final output. The initial classification of male vs. female was later modified to include a third class - child - with the original goal of automatically detecting child pornography on the web (Meinedo and Trancoso, 2011). Although we have achieved top results in related paralinguistic challenges with methods based on fusion of several sub-systems trained with short and long term acoustic and prosodic features, and different classification paradigms (GMM-UBM, MLP and SVM), the age/gender modules currently integrated in the SPA platform were optimized for BN captioning with very low latency. The gender module is based on an MLP with 9 input context frames of 26 coefficients (12th order PLP coefficients plus deltas), two hidden layers with 350 sigmoidal units each and the appropriate number of softmax output units (one per target class) which can be viewed as giving a probabilistic estimate of the input frame belonging to each class. The age module consists of a first segment-level front-end extraction stage followed by a 3-class SVM classification. Feature vectors are formed by 450 features extracted with OpenSMILE and corresponds to the official reduced-set of the InterSpeech 2010 Paralinguistic Challenge (Schuller et al., 2010).

We have recently worked towards improving the gender/age classification module, taking into account the SpeDial core requirements: telephone speech and multilinguality. We have used additional corpora for adapting the respective modules. A multilingual telephone speech corpus composed of subsets of SpeechDat corpora of the Portuguese, English, Spanish, French, German, and Italian languages was used for training, while the Let's Go 2014 corpus and the Greek Movie Ticketing system corpus (Lopes et al., 2016) were used for assessment of the classifiers.

Gender classification has been applied separately to each speaker turn of the Let's Go and Movie Ticketing datasets (Lopes et al., 2016) in order to obtain automatic turn gender classification. Also, the complete speaker sides have been processed to obtain per dialogue results. Different approaches have been tested, an improved MLP-based system similar to the baseline one including MLP retraining and low-energy frame dropping, but also other methods based on segment-level features in combination with neural network modeling and i-vector based classifiers. Here, only results with the improved frame-level MLP classifier are reported. In the case of turn-level classification, gender accuracies obtained are 79.6% and 89.8% in the Let's Go and Movie Ticketing datasets respectively, when considering all the speaker turns. Notice that in both datasets not only most of the turns are extremely short, but there is also a significant number of turns without speech content. In particular, around 12% of the Let's Go turns do not contain useful speech, which affects negatively the performance of the classifiers. When considering only the turns annotated as containing speech, the performance increases up to 84.9% in the Let's Go corpus (speech content annotation is not available in the Movie corpus). Regarding dialogue level evaluation, gender accuracies obtained are 91.7% and 98.0% in the English Let's Go and Greek Movie Ticketing SpeDial datasets, respectively. Overall, the module performs consistently in both datasets, independently of the language (notice that Greek data was not included in the training set). We consider these results quite satisfactory, particularly considering the reduced amount of actual speech in most of the speaker turns.

Regarding age classification, we tested a classification approach similar to the one described for gender. The frame-level MLP-based age classifier has been applied both to each speaker turn and to the complete dialogues of the Let's Go dataset in order to obtain automatic age classification, achieving around 67% classification accuracy. The Movie Ticketing dataset was not considered in these experiments, given that age annotations are not available for this corpus. The results are not as satisfactory as the ones obtained in the multi-lingual telephone speech corpus. The new gender/age classification models are now in the process of being integrated in the SPA framework.

3.6. Dialog Act Recognition

Identifying the dialog acts is important for spoken dialogue systems, since they reveal the intention of the speaker. Thus, an inconsistent sequence of dialog acts is a cue for the presence of communication problems in the dialog.

In an effort to gather information that may be relevant to detect and find the root causes for hot spots, we have integrated a text-based module for dialog act recognition based on previous experiments on the influence of context on that task (Ribeiro et al., 2015b). The deployed system uses a model trained on data from the Switchboard corpus (Godfrey et al., 1992), consisting of about 2400 telephone conversations among 543 American English speakers (302 male and 241 female). Each pair of speakers was automatically attributed a topic for discussion, from 70 different ones. Furthermore, speaker pairing and topic attribution were constrained so that no two speakers would be paired with each other more than once and no one spoke more than once on a given topic. However, only a subset of 1155 manual transcriptions (annotated with disfluency, abandonment, and interruption information), contain-

ing 223606 utterances, was annotated for dialog acts, using the SWBD-DAMSL tag set (Jurafsky et al., 1997). The deployed system uses the 42-label variant of this tag set.
In terms of features, it uses unigrams, bigrams, wh-words, and punctuation as base features extracted from the utterance being classified. Furthermore, it uses context information extracted from the three previous utterances in the form of the dialog act label predicted by the SVM classifier itself. The number of previous utterances was fixed in three since beyond that the accuracy improvements are negligible. Using this approach we were able to obtain 79.6% accuracy in a 10-fold cross-validation evaluation. Furthermore, we were able to surpass the results obtained by (Gambäck et al., 2011) under the same evaluation conditions, making our approach the state-of-the-art on the dialog act recognition task.
Finally, in terms of input and output, the dialog act recognition module receives the sequence of utterances in textual form, one per line, and outputs the corresponding dialog act labels, one per line, in the same order.

3.7. Third-party modules

In order to contribute towards hot spot detection, but also to many other applications, we have recently integrated not only our own modules, but other publicly available modules. This was the motivation for including a general purpose Audio Feature Extraction module, based on openSMILE (Eyben et al., 2010). It was also the motivation for including a DTMF Detection module, which may warn that the user resorts to the telephone keys when the spoken dialogue flow does not progress as expected. The algorithm in this module has been adapted from the DTMF Detection Library, written by David Luu[3]. The module was tuned and evaluated on part of the Let's Go 2006 data (Schmitt et al., 2012), achieving about 86.5% Precision and 79.0% Recall.

4. Conclusions

This paper presents a web-based platform that can be used to provide access to existing speech processing modules. It was developed with the aim of facilitating the usage of the modules, without the need to know about software dependencies, and specific configurations. Apart from being accessed by a web-browser, the platform also provides a REST API for easy integration with other applications. The platform is flexible, scalable, provides authentication for access restrictions, and was developed taking into consideration the time and effort of providing new services. It was recently expanded within the Spedial European project to integrate modules that could provide potential cues for hot spot detection in spoken dialogue systems. The platform currently integrates a considerable number of audio and text processing modules, including: speech disfluency classification, emotion detection, dialog act recognition, age and gender classification, non-nativeness detection, hyperarticulation detection, dialog act recognition, and two external modules for feature extraction and DTMF detection.
Since the platform provides convenient means to provide access to the project partners, as well as internal and external users, it is continuously being improved and expanding to integrate new functionalities. Part of the modules previously presented are also being improved and new versions and models will soon be made available in the platform. The platform may constitute an important asset for the ongoing project INSIDE that investigates how to extend multi-agent planning under uncertainty to cooperative scenarios involving human and robot agents. The existing modules for emotion detection and for gender/age classification, may provide relevant features for the involved tasks. The platform may also be relevant for the ongoing project LAW-TRAIN that provides means for law enforcement units to practice the interrogation of suspects in a multicultural and virtual reality context. Emotion detection and disfluency detection may be relevant modules for the involved tasks. We expect these projects to contribute further to the platform with other modules developed in the context of theses projects. In the near future, we plan to promote the platform for the web community and to make most of the modules available for guest access.

[3]available on https://pypi.python.org/pypi/dtmf-detector

Acknowledgments

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, by project INSIDE CMUP-ERI/HCI/0051/2013, and by project LAW-TRAIN H2020-EU.3.7 contract 653587.

5. Bibliographical References

Abad, A., Meinedo, H., and Neto, J. (2008). Automatic classification and transcription of telephone speech in radio broadcast data. In *Computational Processing of the Portuguese Language*, pages 172–181. Springer.

Abad, A. (2011). The L2F language recognition system for NIST LRE 2011. In *The 2011 NIST Language Recognition evaluation (LRE11) Workshop, Atlanta, US*.

Batista, F. and Mamede, N. (2011). *Recovering Capitalization and Punctuation Marks on Speech Transcriptions*. Ph.D. thesis.

Batista, F., Trancoso, I., and Mamede, N. J. (2009). Comparing automatic rich transcription for portuguese, spanish and english broadcast news. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 540–545. IEEE.

Batista, F., Moniz, H., Trancoso, I., and Mamede, N. J. (2012). Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. *IEEE Transactions on Audio, Speech and Language Processing, Special Issue on New Frontiers in Rich Transcription*, 20(2):474–485, feb.

Bertini, M., Bimbo, A. D., Ioannidis, G., Bijk, E., Trancoso, I., and Meinedo, H. (2013). euTV: a system for media monitoring and publishing. In Alejandro Jaimes, et al., editors, *ACM Multimedia*, pages 453–454. ACM.

Cervinschi, C. L., Butucea, D., et al. (2010). Integration of web technologies in software applications. is web 2.0 a solution? *Database Systems Journal*, 1(2):39–44.

Eskenazi, M., Black, A. W., Raux, A., and Langner, B. (2008). Let's go lab: a platform for evaluation of spoken dialog systems with real world users. In *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, page 219.

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

Fandrianto, A. and Eskenazi, M. (2012). Prosodic Entrainment in an Information-Driven Dialog System. In *Proceedings of INTERSPEECH 2012*, pages 342–345.

Gambäck, B., Olsson, F., and Täckström, O. (2011). Active Learning for Dialogue Act Classification. In *Proceedings of INTERSPEECH*, pages 1329–1332.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of the IEEE International Conference on Speech, and Signal Processing, ICASSP'92*, volume 1, pages 517–520.

Grimm, M., Kroschel, K., Mower, E., and Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Commun.*, 49(10-11):787–800, October.

Grimm, M., Kroschel, K., and Narayanan, S. S. (2008). The vera am mittag german audio-visual emotional speech database. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 865–868, Hannover, Germany, June.

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

Hirsch, H. and Finster, H. (2005). The simulation of realistic acoustic input scenarios for speech recognition systems. In *Proceedings of Interspeech 2005*, pages 2697–2700.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coder manual. Technical Report Draft 13, University of Colorado, Institute of Cognitive Science.

Litman, D. J., Hirschberg, J., and Swerts, M. (2000). Predicting Automatic Speech Recognition Performance Using Prosodic Cues. In *Proceedings of NAACL*, pages 218–225.

Lopes, J., Chorianopoulou, A., Palogiannidi, E., Moniz, H., Abad, A., Louka, K., Iosif, E., and Potamianos, A. (2016). The spedial datasets: datasets for spoken dialogue systems analytics. In *LREC 2016 - 10th Language Resources and Evaluation Conference*, Portoroz, Slovenia, May.

Meinedo, H. and Neto, J. P. (2003). Automatic speech annotation and transcription in a broadcast news task. In *In MSDR'2003 - ISCA Workshop on Multilingual Spoken Document Retrieval*, Hong Kong, China.

Meinedo, H. and Trancoso, I. (2011). Age and gender detection in the i-dash project. *ACM Trans. Speech Lang. Process.*, 7(4):13:1–13:16, August.

Meinedo, H., Viveiros, M., and Neto, J. (2008). Evaluation of a live broadcast news subtitling system for Portuguese. In *Interspeech*, Brisbane, Australia.

Meinedo, H., Abad, A., and Pellegrini, T. (2010). The l2f broadcast news speech recognition system. *In Fala2010*.

Moniz, H., Ferreira, J., Batista, F., and Trancoso, I. (2015). Disfluency detection across domains. In *Proc. of DISS 2015 - Disfluency in Spontaneous Speech an ICPhS Satellite Meeting*, Edinburgh, Scotland, UK, August.

Oviatt, S., MacEachern, M., and Levow, G.-A. (1998). Predicting Hyperarticulate Speech During Human-Computer Error Resolution. *Speech Communication*, 24(2):87–110.

Raux, A., Bohus, D., Langner, B., Black, A. W., and Eskenazi, M. (2006). Doing Research on a Deployed Spoken Dialogue System: One Year of Lets Go Experience. In *Proceedings of INTERSPEECH 2006*, pages 65–68.

Ribeiro, E., Ferreira, J., Olcoz, J., Abad, A., Moniz, H., Batista, F., and Trancoso, I. (2015a). Combining multiple approaches to predict the degree of nativeness. In *Proc. of Interspeech 2015*, Dresden, Germany.

Ribeiro, E., Ribeiro, R., and de Matos, D. M. (2015b). The influence of context on dialogue act recognition. *CoRR*, abs/1506.00839.

Schmitt, A., Ultes, S., and Minker, W. (2012). A Parameterized and Annotated Spoken Dialog Corpus of the CMU Let's Go Bus Information System. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC'12*, Istanbul, Turkey. European Language Resources Association (ELRA).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. (2010). The INTERSPEECH 2010 Paralinguistic Challenge - Age, Gender, and Affect. In ISCA, editor, *Proceedings of Interspeech*, pages 2794–2797.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K. R., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (2013). The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proceedings of INTERSPEECH 2013*, pages 148–152.

Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J. R., Nöth, E., Zhang, Y., and Weninger, F. (2015). The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition. In *Proceedings INTERSPEECH 2015, ISCA, Dresden, Germany*.

Soltau, H. and Waibel, A. (1998). On the Influence of Hyperarticulated Speech on Recognition Performance. In *Proceedings of ICSLP*.

Trancoso, I., Viana, M., Duarte, I., and Matos, G. (1998). Corpus de dialogo CORAL. In *PROPOR'98*, Porto Alegre, Brasil.