IDIAP RESEARCH REPORT

AUTOMATIC SPEECH INDEXING SYSTEM OF BILINGUAL VIDEO PARLIAMENT INTERVENTIONS

Gyorgy Szaszak Milos Cernak Philip N. Garner
Petr Motlicek Alexandre Nanchen Flavio Tarsetti

Idiap-RR-25-2013

JULY 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Automatic Speech Indexing System of Bilingual Video Parliament Interventions

July 3, 2013

Abstract

This paper presents the development and evaluation of an automatic audio indexing system designed for a special task: work in a bilingual environment in the Parliament of the Canton of Valais in Switzerland, with two official languages, German and French. As several speakers are bilingual, language changes may occur within speaker or even within utterance. Two audio indexing approaches are presented and compared: in the first, speech indexing is based on bilingual automatic speech recognition; in the second, language identification is used after speaker diarization in order to select the corresponding monolingual speech recognizer for decoding. The approaches are later combined. Speaker adaptive training is also addressed and evaluated. Accuracy of language identification and speech recognition for the monolingual and bilingual cases are presented and compared, in parallel with a brief description of the system and the user interface. Finally, the audio indexing system is also evaluated from an information retrieval point of view.

Index Terms: spoken document indexing, spoken document retrieval, bilingual speech recognition, language identification, spoken term detection

1 Introduction

Multimedia data containing audio and video plays an important role in communication and has become a valuable information source, which creates a need for the development of technologies capable of searching spoken terms and/or indexing speech. Conventional Automatic Speech Recognizers (ASR) have been largely used to implement such indexing systems based on the audio [1], [2] to obtain transcripts or word lattices for further text based term detection. The video is also often used, especially in speaker diarization and topic segmentation tasks, prior to speech indexing, where cues coming from gestures like gaze, facial and hand movements can be exploited [3], [4]. Several speech indexing systems were developed during the past decades, for instance [5], [6], [7].

This paper presents the design, development and evaluation of an automatic speech indexing system designed for bilingual environment. Bilingual (or multilingual) speech recognition is an active research area [8], [9], but as far as the

authors know, this is the first attempt to implement a bilingual speech indexing system.

In the Canton of Valais, in Switzerland, both French and German are official and spoken. Whilst French is relatively close to the standard French spoken in France, German has a local dialect, which is often hard to understand even for German speaking people coming from another canton or state. In formal interactions, usually standard German (Hochdeutsch) is spoken, but even this is highly influenced by local accent. The automatic speech indexing system is developed for the Cantonal Parliament of Valais, where both languages are in active use. A part of speakers is either bilingual or non-native speaker, and it is not rare that a bilingual speaker changes language from one utterance to the other or uses lexical items or terms borrowed from the other language. Therefore a pure speaker diarization is not sufficient to perform implicit language identification, as speakers can be bilingual or non-native speakers of the other language.

To implement a system capable of handling two different languages, three possibilities are evident [10]: in the first approach, a real bilingual ASR is used, with merged phoneme sets, dictionaries and language models for the two languages. The second approach is to use two separate monolingual ASR, one for each language, and insert a Language IDentification (LID) module into the system to choose which ASR should be used for the given utterance. A third alternative is to combine both approaches: use a LID, but allow for decisions for French, German or 'unknown', this latter in case of weak confidence. Then, in the first case the French, in the second the German, and in the third the bilingual ASR module is loaded.

Another key issue when working in a bi- or multilingual environment, especially if the language environment is rich in dialects and accents like in Valais, is speaker normalization and or/speaker adaptation, as speech data shows very high variability. This paper does not focus on enhancing considerably existing normalization and adaptation methods, but rather evaluating them within the current speech indexing system in order to find the best performing ones for the given bilingual tasks, aiming inclusive adaptation for speakers, accent/dialect and acoustic environment, with a special focus on speaker adaptive training [12].

Beside providing transcription for high amounts of audio/video material, an important task of the speech indexing system is the support for information retrieval from audio archives which are never transcribed. The audio indexing system is therefore also analysed from an information retrieval point-of-view, that is, to perform Spoken Term Detection (STD). In this aspect, the system can be regarded as a Large Vocabulary Continuous Speech Recognition (LVCSR) based keyword spotting application.

This paper is organized as follows: in Section 2, the block scheme and design of the bilingual speech indexing system is presented and explained. Section 3 gives information about training and testing data used for ASR modelling, presented and evaluated in detail in Section 4, including acoustic and language model training and speaker adaptive training. Language identification is presented and evaluated in Section 5. The user interface for accessing indexed

audio and video is described in Section 6. Section 7 describes spoken term detection and presents analysis from the information retrieval point of view. Finally, conclusions are drawn.

2 The MediaParl Video Server and Speech Indexing System

The speech (or audio) indexing system is part of a complex system allowing for browsing in a video database, presented in Fig.1. The video database is an inventory of material recorded at the cantonal parliament of Valais, located in Sion, Switzerland. The political debates at the parliament are recorded regularly. Currently, records from year 2009 are available, but the system is to be launched mid 2013 for continuous, automatic operation, that is, to record video, index its content based on audio, store it and provide access to it via a video streaming server. The role of the audio indexing system is to generate indexes for all video material based on audio, relying on models trained on MediaParl and SwissParl databases (presented later). The potential user can interact via the Internet network with the video server, browse referenced content or search fast and effectively spoken terms in the inventory.

The block scheme of the speech indexing system is shown in Fig.2. From the video inventory of years 2006 and 2009, a bilingual audio database is created, used for system training, development and evaluation. This database (MediaParl DB) is presented in the next section (Section 3). The audio indexing system incorporates a LID module. As already mentioned in the introduction, speaker diarization itself does not solve language detection due to bilingualism, hence, either a LID module is necessary to choose between French and German input language, or a bilingual ASR should be used, which is however expected to be more complex and less effective than the monolingual ones. The ASR system (Section 4) performs large vocabulary speech recognition for both languages, generates transcripts which are used for indexing of audio. Corpus for vocabulary and language models come from the audio transcripts of the MediaParl database, but also from a text database called SwissParl, both of them presented in Section 3. SwissParl contains more transcripts from several cantonal parliaments in Switzerland, it is used to ensure sufficient training data and robust language modelling. Audio indexes are stored in the Indexes Database.

3 Databases

3.1 The MediaParl Database

The MediaParl database is used to train ASR acoustic models. It consists of political debates recorded at the cantonal parliament of Valais. Debates take place always in the same room, they are recorded with distant talker microphones. The recordings mostly contain prepared speeches in both languages. Compared

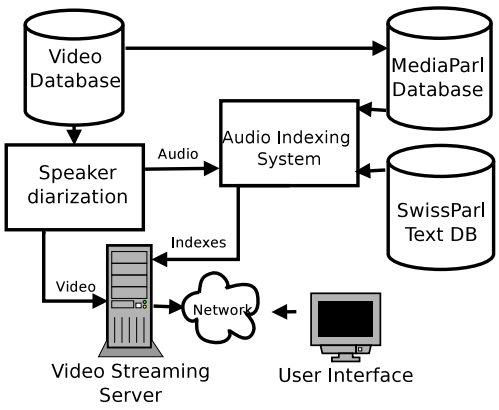

Figure 1: *Complex Audio-Video Server architecture with the Speech Indexing System.*

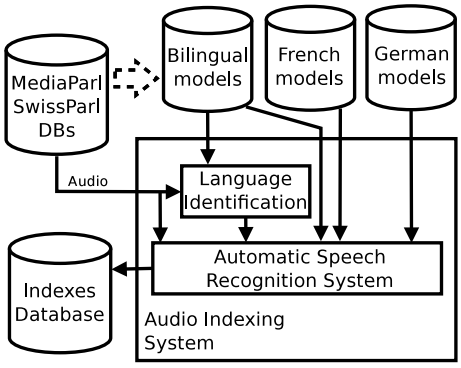

Figure 2: *The Speech Indexing System.*

to similar multi- or bilingual databases, MediaParl stands out because of its size as it contains 20 hours of speech in both German and French [10].

A detailed description about the MediaParl database is given in [10], here only some basic characteristics of data are presented: audio recordings were formatted as MPEG ADTS, layer III, v1, 128 kbps, 44.1 kHz, mono, 16 bits, but then converted to WAVE audio, PCM, 16 bit, mono, 16 kHz. The database is split into train, development and test sets. The test set contains all bilingual speakers who actively used both languages (7 speakers, 2,446 utterances) to allow for bilingual evaluation. The remaining speakers are split into training set (180 speakers, 11,425 utts) and development set (17 speakers, 1,525 utts). In case of adaptation experiments, test set was further split into adaptation set (148 French, 156 German utts) and final test set (925 French, 1,521 German utts, approx. 31k ans 33k words respectively).

3.2 Swiss Parliament Data

Swiss parliament data, organized into the SwissParl database is used for language modelling. Swiss parliaments are bounded by law to transcribe the content of all political debates and make them publicly available. This data is an excellent source for in-domain language modelling. The used text corpora come from two different sources: (i) text normalized transcripts form the MediaParl database with their associated pronunciation lexicons; (ii) another part of text data comes from the following cantonal parliaments: Valais, Fribourg, Vaud, Geneva, Neuchâtel, Basel, Bern, Zürich and Solothurn. SwissParl contains in total 275,159 French and 260,408 German sentences. The French lexicon contains 47,676, the German one 118,186 pronunciations. All sentences have been normalized in a semi-automatic way with human proofing. The lexicons have been constructed using two purchased lexicons, BDLex and Phonolex. Missing pronunciations have been automatically generated using Phonetisaurus software[1] and checked by a native speaker.

4 ASR Baselines for Speech Indexing

Automatic Speech Recognizers used for automatic audio transcription are HMM/GMM based, independent systems both for the two monolingual and the bilingual case. Feature extraction yields 39 Mel-Frequency Perceptual Linear Prediction (PLP) features (C0-C12+Δ+$\Delta\Delta$). Cepstral Mean Normalization (CMN) was applied in all cases.

4.1 Acoustic Modelling

Models are HMM tied-state triphones for all setups, with up to 3000 tied states, trained on the MediaParl database. State tying is based on the MDL criterion [11]. State emissions are modelled by 16 component GMMs. Bilingual models are estimated on the merged train set of the monolingual French and German systems (like in [10]). For the bilingual case, the SAMPA phoneme set is also merged for French and German. Merging the phoneme set was based on following considerations: for phonemes in the two languages with identical SAMPA symbol a merged model is trained (mixed German and French data); for phonemes specific to one of the languages, only data from the corresponding language could be used. The phoneme set for French monolingual case was composed of 37 phonemes, for German 56 phonemes, and for the bilingual model set 62 phonemes.

4.2 Language Models

Language models are tri-gram ARPA format models, trained on mixed MediaParl and SwissParl data. The bilingual language model is trained on mixed

[1] Phonetisaurus is developed by Josef Novak, currently available at: `https://code.google.com/p/phonetisaurus/`

French and German data. Lexicons are also merged for bilingual ASR. Perplexities and OOV rates for monolingual French and German, and for the bilingual language models are shown in Table 1.

Table 1: *Perplexities and OOV rates for the monolingual and bilingual language models, computed on the test set.*

Language	Perplexity	OOV rate [%]
French (FR)	134	7.09
German (GR)	220	2.91
FR+GR	195	4.37

4.3 Speaker Adaptive Training

Speech spoken in Valaisian environment shows very high variability, as the ratio of non-native speakers is high. Whilst real bilingual speakers speak both languages on mother tongue level and are usually closer to standard pronunciation in both languages, non-native speakers speak the second language with some more or less characteristic accent. In addition to this, German has a characteristic local dialect in Valais (Walliser Deutsch), influencing mostly German speakers. Therefore speaker normalization or adaptation methods were exhaustively analysed. Speaker Adaptive Training (SAT) was found more efficient than either MLLR-like (mean and variance MLLR, CMLLR, SMAPLR or CSMAPLR) or MAP based speaker adaptation [14]. The quantity of data available from each individual speaker was indeed insufficient for the slowly converging MAP adaptation.

SAT was performed in selective way, that is after applying parameter clustering of the baseline models. For the clustering, a dynamic, regression tree based approach was used [15]: regression tree was constructed based on state tying statistics until 32 final leaves. However, if for a given speaker, the available data for adaptation was not sufficient, trees were pooled (clusters merged to ensure sufficient training utterances for each remaining cluster). SAT was based on CMLLR transforms (transforming both mixture means and variances), and applied in the feature space. 5 iterative re-estimation cycles were carried out on baseline models using the speaker transforms. The contribution of SAT to relative WER reduction is shown in Table 2 in parallel with baseline ASR performance (word accuracy) for monolingual and bilingual cases. Applying further adaptation after SAT (using SAT transforms as parent transform for test speakers and do another adaptation) did not lead to significant performance improvement [14]. The reason for the lower WER reduction in case of the French system is most probably due to insufficient adaptation data for 2 speakers (regression trees used for adaptation had to be pooled to 6 and 13 leaves, respectively).

Table 2: *Baseline and SAT-normalized word accuracies and relative WER reduction in ASR used for audio indexing.*

System	Language	Acc [%]	Rel. WER red. [%]
Baseline	French	74.8	-
+ SAT	French	78.8	15.8
Baseline	German	74.4	-
+ SAT	German	79.1	21.4
Baseline	Bilingual	71.0	-
+ SAT	Bilingual	77.7	23.3

5 Language Identification

Initial experiments with automatic LID for MediaParl database were described in [10], where a hierarchical multilayer perceptron based approach was investigated. This approach was ASR-independent, and the reported average accuracy was 98.5%.

For the two languages used, we hypothesized that an ASR-dependent approach could perform better. The bilingual ASR system (see Section 4) was used for LID based on the obtained transcripts (text). We investigated two criteria for this text-based LID: (i) word count and (ii) a naive Bayes classifier. The later is proposed to have a system less sensitive to the OOV words.

The word count criterion works with French and German lists of words (generated from MediaParl and SwissParl databases), and counted the number of French and German words appearing in the transcript. The higher count classifies the language. The second approach is based on the Naive Bayes classifier that was trained using the Natural Language Toolkit[2]. We used Europarl[3] data for training the classifier. For classification, the following features were used: (i) words, based on dictionaries (like for the word count criterion), (ii) first three letters of words, (iii) last three letters of words and (iv) character counts of words.

We evaluated the performance of both approaches on the test set (2446 utterances, 64k words). In both cases, LID accuracy was significantly improved, from 98.5% to 99.9%. The Naive Bayes classifier approach performed better than the word count criterion, however, on the test set the improvement was not significant.

[2]NLTK is a leading platform for building Python programs to work with human language data, available at `http://nltk.org/`.

[3]European Parliament Proceedings Parallel Corpus 1996-2011, available at `http://www.statmt.org/europarl/`.

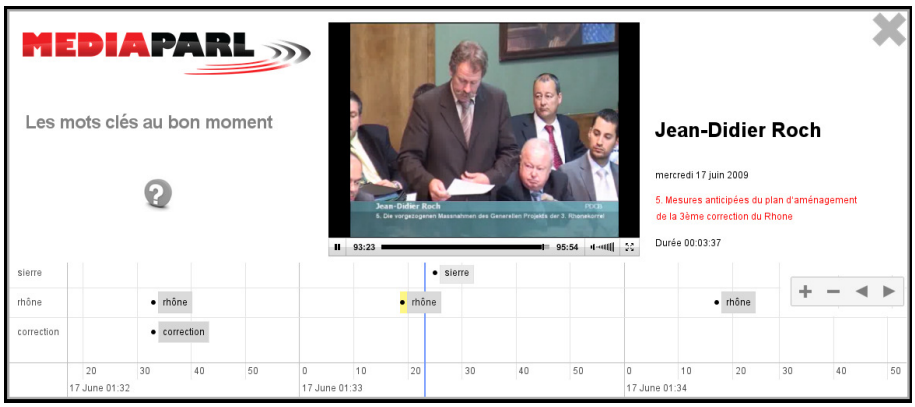

Figure 3: *The second layer of the user interface.*

6 User Interface

The MediaParl website[4], available in French and German versions, is built on the concept of two layers: the first layer presents a query interface, displaying a list of links related to the searched information. Further filtering and refining of the search is possible based (i) on specific deputy/state councilor's name, (ii) on the subject of the debate (this is done via the TV information displayed and using Idiap's optical character recognition system to extract this information) and (iii) on the transcription/keywords occurring during the relevant debate. Filters for legislature and for political groups can also be used. The second layer (see Fig. 3) opens up when clicking on a proposed item from the first layer. It presents the requested information in parallel with time labelling in the video for the selected debate. A navigable time-line is displayed at the bottom of the video, enabling the user to view the region of interest in the video. The user can easily click on the time-line to move to a specific location in the video. This layer can even present the searched keywords in the region of interest: by clicking on these indexed keywords, the user gets to the region of the video where the word is pronounced.

The platform has been developed in *Django* using the *Python* programming language. The interface makes calls to the Postgres database using *ajax* and uses usual web programming languages: *html*, *javascript*, *jquery*.

7 System Evaluation

Browsing the video inventory is possible based on several search criteria presented so far (speaker, subject, keywords). An important application of these is multimedia retrieval based on user specified keywords. The evaluation of the ASR system in terms of word accuracies was presented in Section 4. However,

[4]The MediaParl website is currently available at: `www.idiap.ch/webapps/webgrandconseil`

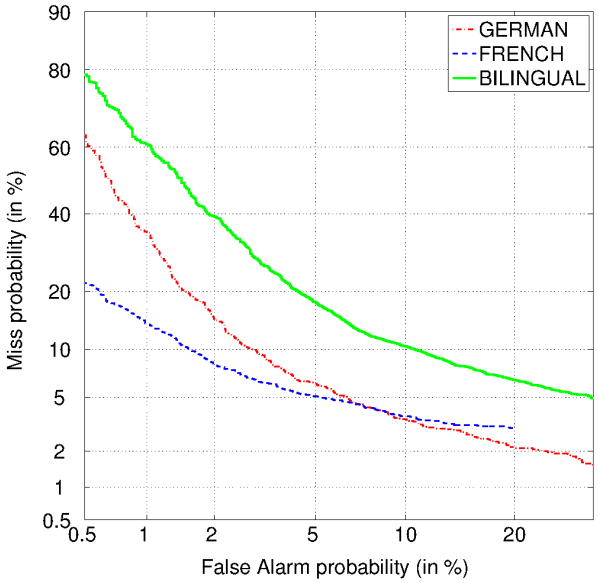

Figure 4: *DET curves for French, German and bilingual STD.*

a very informative measure on system usability can be drawn from an information retrieval point of view evaluation. Therefore the audio indexing system was analysed as a STD system, working on word recognition lattices. Performance is assessed using Detection Trade-off (DET) plots [16] and Equal Error Rates (EER).

A list of 200-200, keywords were selected for both French and German, and passed on for STD analysis. For the bilingual case, lists were merged. Confidence scores were estimated based on word posterior probabilities as described in [18].

The lattice contains word posterior probabilities, used as confidence scores, for each candidate computed as in [18]:

$$P(W_i; t_s, t_e) = \sum_{Q} P(W_i^j, t_s, t_e | x_{t_s}^{t_e}), \tag{1}$$

where W_i is a word identified by index i, starting at t_s and ending at t_e; j denotes the occurrence of word W_i in the lattice. $x_{t_s}^{t_e}$ corresponds to the observation sequence between $t_s - t_e$. Q represents a set of all hypothesis sequences in the lattice that contain the hypothesized word W_i in a time interval $t \in (ts, te)$.

STD performance is shown in Fig. 4 for each ASR setup in terms of DET curves. EERs are provided in Table 3, with precision and recall for the operation points defined by EER.

Whilst in ASR transcription accuracy, bilingual system performance was quite close to monolingual system performance, this is not the case for STD

Table 3: *EER rates, precision and recall for monolingual French and German, and for the bilingual language models.*

ASR	EER [%]	Precision	Recall
FR monolingual	5.07	82.6	94.9
GE monolingual	5.68	82.4	94.3
FR+GR bilingual	10.29	87.5	89.7

based on lattices, where it is more important to use the LID module to favor monolingual systems if possible.

8 Conclusions

This paper focused on audio indexing in order to allow for browsing video content. State-of-the-art technologies were evaluated and used to develop a complex system including language identification, mono- and bilingual modelling for ASR, and a browsing interface, supporting spoken term detection. The system and many of its main components were presented and evaluated from different points of view with a special emphasis on bilingual modelling. Evaluation results have shown the importance of speaker adaptive training, yielding up to 23.3% relative WER reduction for speech transcription. Language identification worked by 99.5% accuracy, therefore a system incorporating LID was implemented, using still a bilingual ASR beside the monolingual ones, as LID is allowed to classify the language as 'unknown' in case of uncertainty. This approach is believed to be optimal to handle sudden code switches, which might occur even within an utterance. Evaluating the system from a STD point of view, the role of LID is even more important, as STD evaluated for the bilingual setup had significantly higher EER (10.29%) than the monolingual ones (5.07% and 5.68% for French and German, respectively).

9 Acknowledgements

The authors would like to express their gratitude to the Parliament Service of the State of Valais, Switzerland for their financial support and for providing access to the audio-video recordings.

References

[1] Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen, L., Schwartz, R. and Srivastava, A., "Speech and language technologies for audio indexing and retrieval", In: Proc. IEEE, 88(8): 1338-1353, 2000.

[2] Thambiratnam, K. and Sridharan S., "Rapid Yet Accurate Speech Indexing Using Dynamic Match Lattice Spotting", IEEE Trans. Speech and Audio Proc., 15(1):346-357, 2007.

[3] Friedland, G., Hung, H. and Yeo, C., "Multimodal speaker diarization of real-world meetings using compressed-domain video features", In: Proc. 2009 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 4069-4072, 2009.

[4] Garau, G. and Bourlard, H., "Using audio and visual cues for speaker diarisation initialisation", In: Proc. 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp 4942-4945, 2010.

[5] Alberti, C., Bacchiani, M., Bezman, A., Chelba, C., Drofa, A., Liao, H., Moreno, P., Power, T., Sahuguet, A., Shugrina, M. and Siohan, O., "An audio indexing system for election video material", In: Proc. 2009 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 4873-4876 , 2009.

[6] Van Thong, J.M., Goddeau, D., Litvinova, A., Logan, B., Moreno, P. and Swain, M. "Speechbot: a speech recognition based audio indexing system for the web", In: Proc. RIAO, (Content-Based Multimedia Information Access), Center for the Advanced Study of Information Systems, Paris, pp. 106-115. 2000.

[7] Gauvain, J.L., Lamel, L., and Adda, G., "The LIMSI Broadcast News Transcription System", Speech Communication, 37(1-2):89-108, 2002.

[8] Zhang, Q., Pan, J. and Yan, Y., "Mandarin-English bilingual Speech Recognition for real world music retrieval", In: Proc. 2008 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 4253-4256 , 2008.

[9] Weng, F., Bratt, H., Neumeyer, L., and Stolcke, A., "A Study of Multilingual Speech Recognition", In: Proc. EuroSpeech97, vol. 1, pp. 359-362. 1997.

[10] Imseng, D., Bourlard, H., Caesar, H., Garner, P. N., Lecorvé, G. and Nanchen, A., "MediaParl: Bilingual mixed language accented speech database", In: Proc. of the 2012 IEEE Workshop on Spoken Language Technology, pp. 263-268, 2012.

[11] Shinoda K. and Watanabe, T., "Acoustic Modeling based on the MDL Criterion for Speech Recognition", In: Proc. EuroSpeech97, vol. 1, pp. 99102, 1997.

[12] Anastasakos T., McDonough J. and Makhoul, J., "Speaker adaptive training: a maximum likelihood approach to speaker normalization" In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2:1043-1046, 1997.

[13] Motlicek, P., Valente, F., Szöke, I., "Improving Acoustic Based Keyword Spotting Using LVCSR Lattices", In: Proc. of International Conference on Acoustic Speech and Signal Processing (ICASSP) 2012, Japan, pp. 4413-4416, 2012.

[14] Szaszák, Gy., "Adaptation Experiments on French MediaParl ASR", Research Report: Idiap-RR-10-2013, Idiap, Martigny, Switzerland, 2013.

[15] Gales, M.J.F., "The generation and use of regression class trees for MLLR adaptation", Technical Report CUED/F-INFENG/TR263, Cambridge University, pp.249-264, 1996.

[16] Martin, A., Doddington, G., Kamm, T., Ordowski, M. and Przybocki M., "The DET Curve in Assessment of Detection Task Performance", In: Proc. of Eurospeech97. Rhodes, Greece, pp. 1895-1898, 1997.

[17] Evermann, G. and Woodland, P., "Large Vocabulary Decoding and Confidence Estimation using Word Phoneme Accuracy Posterior Probabilities", In: Proc. of ICASSP, Istanbul, Turkey, pp. 2366-2369, 2000.

[18] Motlicek P. and Valente, F., "Application of out-of-language detection to spoken term detection", In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Dallas, USA, pp. 50985101, 2010.