English Spoken Term Detection in Multilingual Recordings

Petr Motlicek, Fabio Valente, Philip N. Garner

Idiap Research Institute, Martigny, Switzerland

petr.motlicek@idiap.ch, fabio.valente@idiap.ch, pgarner@idiap.ch

Abstract

This paper investigates the automatic detection of English spoken terms in a multi-language scenario over real lecture recordings. Spoken Term Detection (STD) is based on an LVCSR where the output is represented in the form of word lattices. The lattices are then used to search the required terms. Processed lectures are mainly composed of English, French and Italian recordings where the language can also change within one recording. Therefore, the English STD system uses an Out-Of-Language (OOL) detection module to filter out non-English input segments. OOL detection is evaluated w.r.t. various confidence measures estimated from word lattices. Experimental studies of OOL detection followed by English STD are performed on several hours of multilingual recordings. Significant improvement of OOL+STD over a stand-alone STD system is achieved (relatively more than 50% in EER). Finally, an additional modality (text slides in the form of PowerPoint presentations) is exploited to improve STD.

Index Terms: Spoken Term Detection (STD), LVCSR, Confidence Measure (CM), Out-Of-Language (OOL) detection

1. Introduction

A large increase in the amount of spoken recordings requires automatic indexation and search in this data. A potential solution is a Spoken Term Detection (STD) system[1] which would be able to quickly detect a word or phrase in large archives of unconstrained speech recordings (e.g. lecture recordings, telephone conversations, ...). A common approach is to split the task into two stages. Firstly, a Large Vocabulary Continuous Speech Recognition (LVCSR) system is used to generate a word or phone lattice. Secondly, lattice search is performed to determine likely occurrences of the search terms. STD systems based on word lattices provide significantly better performance than those based on phoneme lattices (e.g., [1]). Word lattices can be associated with a Confidence Measure (CM) for each word. Traditionally, forward-backward re-estimation is used to represent a confidence using word posterior probability conditioned on the entire utterance. Although such an STD system does not deal with Out-Of-Vocabulary (OOV) words, the problem can be solved by taking into account phone recognition lattices usually generated by the same LVCSR system.

In this paper, we present experimental results with an LVCSR-STD system performing automatic indexation of real lecture recordings provided by Klewel[2] to be eventually implemented into a conference webcasting system. Most of the Klewel lecture talks are recorded in west Switzerland. Speech recordings are mostly uttered in English (usually by non-native speakers), however, some recordings are partially (e.g. at the beginning of the talk), or fully uttered in French or Italian. Blindly applying an English STD system for automatically indexing English pronunciations in such multilingual recordings would lead to a significant decrease of overall STD performance since the system would be employed on "inappropriate" speech input (i.e., speech pronounced in different (alien) languages whose words do not appear in the LVCSR dictionary). The amount of detected False Alarms (FAs) of searched terms would significantly increase. These FAs could potentially be reduced by modifying an operating point of the STD system, but this would lead (directly) to an increase of missed terms.

A straightforward solution is to employ a language identification module. However, such a system would have to exploit the knowledge of other (non-target) languages. In order to keep the entire STD system relatively simple and independent of any non-target language, an OOL detection module is an ideal solution. Such a module exploits only the information stored in the same LVCSR word lattices used for search of the spoken terms. Similar approach can possibly be applied to reduce false detections due to dialect variations of the target language which usually have a severe impact on the performance of speech systems [2].

The paper is organized as follows: Sect. 2 and Sect. 3 describe respectively STD task and the STD system used in our studies. Experiments carried out to improve the STD system and achieved results are given in Sect. 4. Sect. 5 concludes the paper.

2. STD task

2.1. Test data

The study is carried out on the 16 kHz audio lecture recordings (supplemented with video and text) provided by Klewel[2]. In total, 9 hours of recordings pronounced in English, French and Italian languages were used. This data was first transcribed according to the input language and then used for evaluation of the OOL detection module. Then, over one hour of English data (from 9 hours of multilingual speech) was selected for STD evaluations and carefully manually annotated. In order to jointly evaluate STD and OOL modules, an additional two hours of French and Italian recordings were used together with one hour of English data. All audio recordings were automatically segmented using a state-of-the-art Multi-Layer Perceptron (MLP) based speech/non-speech detector [3].

In addition, to evaluate a stand-alone STD English system on a standard database, 3 hours of a two channel 8 kHz CTS English development corpus distributed by NIST for the 2006 STD task was used[1].

This work was partially supported by the Swiss National Center of Competence in Research (NCCR) on "Interactive Multi-modal Information Management (IM)2"; and by the European commission $7th$ Framework Programme (FP7) ICT project TA2.

[1] NIST STD06 Eval, http://www.itl.nist.gov/iad/mig//tests/std

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

2.2. Evaluation metric

Since STD is a detection task, performance can be characterized by Detection Error Tradeoff (DET) curves of miss (P_{miss}) versus false alarm (P_{fa}) probabilities [4]. In addition, we also present Equal Error Rates (EERs), a one number metric often used to optimize the system performance. Besides DET and EERs, we use the evaluation measure defined by NIST 2006 STD: Maximum Term-Weighted Value (MTWV) [5].

3. STD system

To perform the search of selected spoken terms in lecture audio recordings, the recordings are first pre-processed using the LVCSR system that produces word recognition lattices. The word lattices are then converted into a candidate term index accompanied with times and detection scores. The detection scores are represented by the word posterior probabilities, estimated from the lattices using the forward-backward re-estimation algorithm [6], and defined as:

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

where W_i is the hypothesized word identity spanning the time interval $t \in (t_s, t_e)$. t_s and t_e denote the start and end time interval, respectively. j denotes the occurrence of word W_i in the lattice. $x_{t_s}^{t_e}$ denotes the corresponding partition of the input speech (the observation feature sequence). Q represents a set of all word hypotheses sequences in the lattice that contain the hypothesized word W_i in $t \in (t_s, t_e)$.

3.1. LVCSR system

To achieve robust hypotheses outputs, a 3-pass LVCSR system is employed, based on various acoustic models trained on different audio data (no Klewel recordings used for training). The system achieving the best recognition performance is then selected to be used in the subsequent STD experiments. More specifically, an LVCSR based on the 8 kHz Conversational Telephone Speech (CTS) system derived from AMI[DA][3] LVCSR [7] is used. In the first pass, PLP features are exploited and HMMs are trained using a Minimum Phone Error (MPE) procedure. In the second pass, Vocal Tract Length Normalization takes place together with HLDA, MPE and Speaker Adaptive Training (SAT). In the third pass, posterior-based speech features estimated using a neural network system replace PLPs. For the decoding, a 50k dictionary is used together with a 3-gram Language Model (LM).

In the second potential system, acoustic models trained on 16 kHz Individual Headset Microphone (IHM) recordings from several meeting corpora (ICSI, NIST, AMI) are employed, replacing CTS models. In the third case, Multiple Distant Microphone (MDM) instead of IHM recordings are used to train acoustic models. In both (IHM, MDM) cases, discriminative training in 3-pass system, similar to the previous AMI CTS system, is employed.

To select the most suitable LVCSR setting in the following STD studies, we evaluate the three systems on 1 hour of manually annotated Klewel English lectures. Overall, the best ASR performance measured in terms of Word Error Rates (WERs) is achieved for the LVCSR system trained on 16 kHz IHM meeting recordings (WER = 28.9%). LVCSR systems trained on 16 kHz MDM and 8 kHz CTS acoustic models perform around 4% and 6% worse, respectively. Therefore, 16 kHz IHM LVCSR is selected for subsequent STD studies.

3.2. Evaluation of stand-alone STD system

First, the LVCSR STD system is evaluated on 3 hours of 8 kHz CTS English development database. The automatically segmented speech recordings are processed by the AMIDA LVCSR system employing CTS acoustic models with a 50k dictionary. The generated bigram lattices are subsequently expanded with a trigram language model. For evaluation, 550 English search terms are randomly selected from the STD06 dry-run list. The achieved STD performance is compared to the baseline system described in [8]. The EER of the baseline system is about 10.1%. The presented STD built on 3-pass LVCSR gives about 20% relative improvement.

For automatic indexing of Klewel lecture recordings, an STD system based on word lattices generated using 16 kHz IHM acoustic models is chosen, since the best ASR performance is achieved with such a system. Word recognition lattices are generated in the third pass using HTK (HDecode) with bi-gram language model. The list of English spoken terms consists of 312 items. The terms are selected manually from the available annotations (in a random fashion over all recordings based only on a potential interest of Klewel end-users). The list of terms is then transformed into a format following NIST 2006 STD evaluations. The EER achieved on 3 hours of Klewel multilingual recordings is about 8.1%, as shown in Tab. 2.

4. Improving STD by detecting OOL segments

Although the English STD system performs reasonably well, while having at the input (unrestricted) multilingual recordings, other improvements can be obtained by detecting OOL segments. The OOL module can be thought of as a probabilistic model that assigns a probability of each processed input segment given the language used in the segment.

4.1. OOL module

The OOL detection used extracts a confidence score of the processed input speech using several Confidence Measures (CMs) [9]. These CMs are derived from word LVCSR lattices. More specifically, we studied these CMs:

- C_{mean} – Probabilities of all hypotheses for the word W_i recognized in the 1-best output, spanning time interval $t \in (t_s, t_e)$, are summed and normalized [10]:

  $$C_{mean} = \frac{\sum_{t=t_s}^{t_e} P(W_i \mid t)}{1 + \alpha(t_e - t_s - 1)}. \tag{2}$$

  α is a constant between 0 and 1.

  C_{max} – The best case probability for a hypothesized word W_i (also found in the 1-best output) occurring in a certain period of time $t \in (t_s, t_e)$ is returned [10]:

  $$C_{max} = \max_{t \in (t_s, t_e)} P(W_i \mid t). \tag{3}$$

- $H(W \mid t_{t_s}^{t_e})$ – Amount of uncertainty of recognized words measured using Entropy information criteria for the given time interval $t \in (t_s, t_e)$:

  $$H(W \mid t_{t_s}^{t_e}) = \frac{\sum_{t=t_s}^{t_e} \sum_i \frac{1}{P(W_i|t)} log_2\big(P(W_i \mid t)\big)}{1 + \alpha(t_e - t_s - 1)}. \tag{4}$$

[3] http://www.amiproject.org

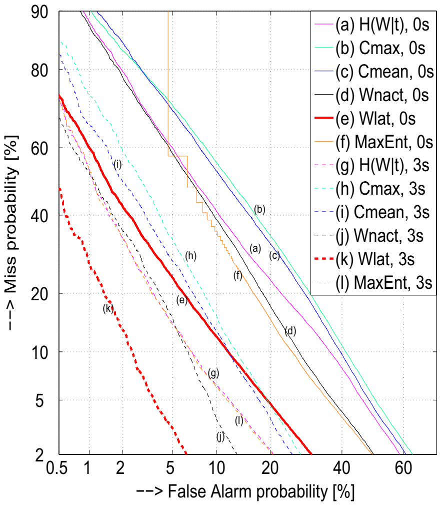

Figure 1: *DET plot – OOL detection using different CMs for temporal context equal to* 0 *sec. and* 3 *sec.*

OOL - EER					
Len [s]	C_{mean}	C_{max}	$H(W \mid t_{t_s}^{t_e})$	W_{lat}	W_{nact}
0	24.9%	25.6%	21.4%	**10.9%**	19.0%
3	11.2%	11.8	8.0%	**4.1%**	7.4%
120	3.6%	3.9%	2.6%	**1.4%**	2.6%

Table 1: *OOL – EER [%] performances achieved on Klewel lecture recordings for different CMs and various temporal context.*

- W_{lat} – Word lattice width - a simple measure provided by counting the number of active arcs from the recognition lattice determines the amount of uncertainty in the LVCSR system at the given time instance $t = t_n$.
- W_{nact} – Number of active and unique words at the given time instance $t = t_n$ is counted and also used as an OOL confidence score.

OOL detection is tested directly on the target Klewel evaluation data. In particular, 9 hours of recordings (3 hours from each of English, Italian and French language) are used. The derived OOL CMs, described in Sec. 4.1, are further post-processed to incorporate a temporal context. This has been shown to significantly improve the detection performance. In case of unconstrained length of processed speech segments, the optimal length of the temporal filter was found to be about 3 sec. [9]. We also experimented with higher lengths (up to 120 sec.) of the filter, since the language usually does not change often in the processed recordings. However, this may cause significant degradation of OOL detection when such a temporal filter were applied on short speech segments, as shown in [9].

Achieved detection performance is shown in the form of DET curves and EERs in Fig. 1 and Tab. 1, respectively. W_{lat} as a confidence score significantly outperforms other CMs used for OOL detection. Additional experiments to fuse all individual CMs using a Maximum Entropy (MaxEnt) technique do not bring any improvements (see Fig. 1). This is probably caused by employing very different data to train the MaxEnt classifier.

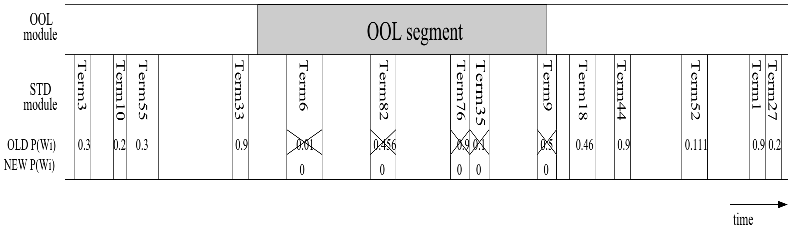

Figure 2: *Combination of OOL and STD modules: STD detection scores are set to zero if detected in speech segments marked as OOL.*

STD						
	OOL - W_{lat}		OOL - no		OOL - manual	
Len [s]	EER	MTWV	EER	MTWV	EER	MTWV
0	5.9%	0.70	8.1%	0.64	3.5%	0.82
3	4.0%	0.78				
120	3.6%	0.81				

Table 2: *STD – EER [%] performances achieved on Klewel lecture recordings w.r.t. OOL detection module. Len denotes length of the temporal filter of the OOL detection module. OOL-W_{lat}, OOL-manual and OOL-no denote OOL detection based on W_{lat} CM, OOL detection taken from manual annotations and the STD system without OOL detection module, respectively.*

4.2. Exploiting OOL in STD system

The OOL detection module is applied in the STD system to automatically remove input speech segments pronounced in non-target languages. Therefore, false alarm terms caused by processing non-English speech segments will potentially be removed in an optimal way (i.e., without any effect on correctly detected terms in English segments).

More specifically, the confidence scores of those terms (already detected by STD system) which correspond to speech segments classified to be OOL segments are set to zero, as graphically shown in Fig 2. In order to "hard threshold" STD detection scores using the OOL detection module, an OOL detection threshold needs to be introduced. In our studies, an optimal threshold is found on development data. A development set comprising of 30 min. of audio recordings uttered in Czech and German languages (i.e., different to French and Italian) as well as in English is used to tune the operating point of OOL detection module [9]. The threshold corresponding to EER is selected as the operating point of the OOL detection module.

Experimental results of the English STD system, in terms of EERs and MTWVs, achieved on 3 hours of multilingual Klewel lecture recordings are given in Tab. 2. Graphical representation in terms of DET curves is shown in Fig 3. Since the best automatic OOL detection performace is achieved with W_{lat} CM, that system is exploited in STD experiments. As seen in Tab. 2, the temporal filter of the OOL detection module with a length of 3 sec. gives performance close to the STD system with manually classified OOL speech segments.

4.3. Exploiting prior information from other modality

Many Klewel lecture audio recordings are supplemented with corresponding slide (PowerPoint) presentations. Therefore, we attempted to exploit this modality in our STD experiments.

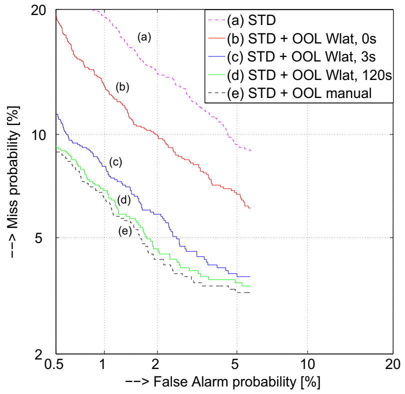

Figure 3: *DET plot – STD on Klewel multilingual recordings.*

STD			
slide	OOL	EER	MTWV
no	no	5.3%	0.74
yes	no	4.5%	0.76
yes	Wlat, 3s	2.0%	0.80
yes	manual	1.6%	0.83

Table 3: *STD – EER [%] performances achieved on a subset of Klewel lecture recordings when additional modality is exploited. c was chosen to be equal to* 50.

More specifically, word posterior probabilities $P(W_i; t_s, t_e)$ of searched terms are modified using a prior which represents a relevance of a term to the topic (given by corresponding text slides). The prior is introduced by a multiplicative constant c:

$$\begin{array}{ll} P_{new} = cP_{old}, & if \quad c <= 1/P_{old}, \\ P_{new} = 1, & otherwise. \end{array} \tag{5}$$

The experiments are run on a subset ($\sim 1/3$) of the multilingual lecture recordings (those supplemented with text slides). First, for each lecture recording, a new list of terms (which is a subset of original 312 searched terms) is automatically generated based on the occurrence of searched terms in the text of corresponding PowerPoint slides. Since no time allocation of the individual slides and their precise alignment with the audio segments of a lecture is available (only the general lecture number assignation), no precise temporal information is employed. Then, posterior probabilities P_{old} (initially estimated from the LVCSR lattices) associated with search terms occurring in the new list of a given lecture are updated according to Eq. 5.

Fig. 4 graphically shows a dependence of EER on varying c for two STD systems (without and with application of the OOL detection module). c varied from 10^{-4} to 10^{3}. Corresponding MTWV values are given in Tab. 3. Although a very simple model is used, which takes into account neither time allocation of searched terms nor quantity of their occurrence in the corresponding slides of each lecture, a relative EER improvement of about 15% is achieved (in both cases with and without the OOL detection module).

5. Discussions and conclusions

This paper summarizes experimental results achieved with an English STD system on Klewel lecture recordings. Due to the unconstrained multilingual input, the system is augmented with an OOL detection module which assigns to each input segment (e.g. frame) a probability given the language used in the segment. Such a module performs as a binary classifier (target-*English* / non-target-*any* language). An OOL detection module can use different lengths of temporal context, which has a significant effect on performance of the subsequent STD system.

STD performance is measured using several criteria (DET curves, EER, MTWV values) on 3 hours of multilingual recordings. Incorporation of the OOL detection module (with 3 sec. long temporal filter) into the STD system increases EER performance relatively by more than 50%.

We also experimented with an additional source of information available from associated text slides on a subset of Klewel recordings. Posterior probabilities (initially estimated from the LVCSR lattices) of those spoken terms which are detected in the corresponding slides of a given lecture recording are modified by a multiplicative constant. A relative improvement of $\sim 15\%$ in STD EER was found.

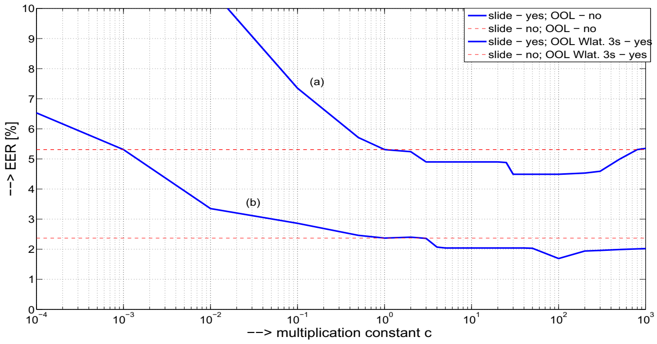

Figure 4: *Overall EERs of STD on the subset of Klewel multilingual recordings when additional prior information is exploited: (a) STD system without OOL module, (b) STD system with OOL module.*

6. References

[1] D. Vergyri et al., "The SRI/OGI 2006 Spoken Term Detection System", *in Proc. of Interspeech*, pp. 2393-2396, Belgium, 2007.

[2] M. Mehrabani, H. Boril and J. Hansen, "Dialect Distance Assessment Method based on Comparison of Pitch Pattern Statistical Models", *in Proc. of ICASSP*, pp. 5158-5161, Dallas, USA, 2010.

[3] J. Dines, J. Vepa, and T. Hain, "The segmentation of multichannel meeting recordings for automatic speech recognition", *in Proc. of the ICSLP*, pages 1213-1216, Pittsburgh, USA, 2006.

[4] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, "The DET curve in assessment of detection task performance", *in Proc. of Eurospeech*, pp. 1895-1898, Greece, 1997.

[5] NIST Spoken Term Detection (STD) 2006 Evaluation Plan, <http://www.itl.nist.gov/iad/mig//tests/std/2006/docs/std06-evalplan-v10.pdf>

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

[7] T. Hain, et al, "The AMI System for the Transcription of Speech in Meetings", *in Proc. of ICASSP*, pp. 357-360, Hawaii, USA, 2007.

[8] I. Szoke et al., "BUT System for NIST Spoken Term Detection 2006 - English", *in Proc. of NIST Spoken Term Detection Workshop (STD 2006)*, pp. 15, Washington D.C., USA, 2006.

[9] P. Motlicek, "Automatic Out-of-Language Detection based on Confidence Measures derived from LVCSR Word and Phone Lattices", *in Proc. of Interspeech*, Brighton, England, 2009.

[10] F. Wessel, R. Schluter, K. Macherey and H. Ney, "Confidence measures for large vocabulary continuous speech recognition", *in IEEE Trans. Speech and Audio Processing*, vol. 9, no. 3, pp. 288-298, 2001.