APPLICATION OF OUT-OF-LANGUAGE DETECTION TO SPOKEN TERM DETECTION

Petr Motlicek, Fabio Valente

Idiap Research Institute, Martigny, Switzerland
{motlicek,valente}@idiap.ch

ABSTRACT

This paper investigates the detection of English spoken terms in a conversational multi-language scenario. The speech is processed using a large vocabulary continuous speech recognition system. The recognition output is represented in the form of word recognition lattices which are then used to search required terms. Due to the potential multi-lingual speech segments at the input, the spoken term detection system is combined with a module performing out-of-language detection to adjust its confidence scores. First, experimental results of spoken term detection are provided on the conversational telephone speech database distributed by NIST in 2006. Then, the system is evaluated on a multi-lingual database with and without employment of the out-of-language detection module, where we are only interested in detecting English terms (stored in the index database). Several strategies to combine these two systems in an efficient way are proposed and evaluated. Around 7% relative improvement over a stand-alone STD is achieved.

Index Terms— Spoken Term Detection (STD), Large Vocabulary Continuous Speech Recognition (LVCSR), Confidence Measure (CM), Out-Of-Language (OOL) detection

1. INTRODUCTION

Spoken Term Detection (STD) [1] aims at detecting a word or phrase in unconstrained speech and is typically used in searching large archives of recorded speech in many applications (e.g., meeting data, telephone speech, unconstrained conversations). Traditional STD systems perform two steps denoted as indexing and searching. First, the input speech is processed (decoded) and the outputs obtained are stored in the index. i.e., the speech is tagged using the sequence of recognized words or phonemes. Then, the index is searched in order to return the location of the determined term.

Two different approaches are currently used in STD which differ in the basic unit used for indexing. In the first case, the index is represented by a word lattice obtained by a Large Vocabulary Continuous Speech Recognition (LVCSR) system. In the second case, the index is based on a phoneme lattice obtained by phoneme recognition.

STD systems based on word lattices provide significantly better performance than the ones based on phoneme lattices (e.g., [2]). The word recognition lattices, which represent a compact way for storing the most probable hypotheses generated by LVCSR, can be associated with a confidence measure for each word. Typically, word posterior probability conditioned on the entire utterance is estimated from the word lattice by forward-backward re-estimation [3]. However, STD based on word lattice is highly sensitive to the dictionary. For instance those systems are unable to detect Out-Of-Vocabulary (OOV) words.

Many spontaneous speech recordings (e.g., teleconferencing, telephone call recordings provided by call centers or security offices) contain short sentences uttered in different languages. STD performance dramatically decreases when the system is employed on "inappropriate" speech input, such as speech pronounced in a different (alien) language whose words do not appear in the LVCSR dictionary used. This introduces many difficulties for LVCSR, which is designed to recognize spontaneous speech pronounced in one language, including higher number of False Alarms (FAs).

One solution consists in modifying the detection threshold (represented by the operating point given by the application) in order to reduce FAs introduced by "inappropriate" input speech segments. However, this will have a direct effect as an increase of missed spoken terms.

This paper describes an Out-Of-Language (OOL) detection module [4] that, based on a confidence measure, is able to detect speech segments that are not uttered in the same language for which the LVCSR system was designed. By exploiting an OOL detection module in a word lattice based STD system, we can detect "inappropriate" speech segments and thus significantly reduce number of FAs. The OOL detection module exploited in our experiments is based on processing word and phone lattices obtained by LVCSR. It can also be used as an OOV detector, i.e., to detect words (such as names of persons, places, etc.) that do not appear in the dictionary.

The study is carried on a database that contains discussions in English, Czech and German languages [10] uttered by native Czech and German speakers. For development purposes, we use OGI multilanguage corpus [11] to estimate the parameters of the OOL detection module. The investigation is carried out with a STD system based on word lattices designed for English. The work investigates the use of the OOL detection for improving the STD scores i.e., reducing the errors related to the presence of unknown languages in the audio.

Word lattice based STD system, even when combined with the OOL detection module, can not recognize words pronounced in a different language (or words not appearing in the dictionary, in the case of the OOV module). However, such detected speech segments can then be processed by another STD system, such as a phoneme lattice based STD. This approach is capable of providing the best ratio between low number of incorrectly detected terms (due to the OOL detection) and low number of missed terms (due to the use of word lattice based STD).

The paper is organized as follows: Section 2 describes and evaluates the STD module used. In Section 3, the OOL module is described. Results on combination of STD and OOL techniques are given in Section 4, followed by discussions and conclusions.

This work was partially supported by the Swiss National Center of Competence in Research (NCCR) on "Interactive Multi-modal Information Management (IM)2"; by the European commission 6^{th} Framework Programme (FP6) ICT project AMIDA; and by the $7th$ Framework Programme (FP7) ICT project TA2.

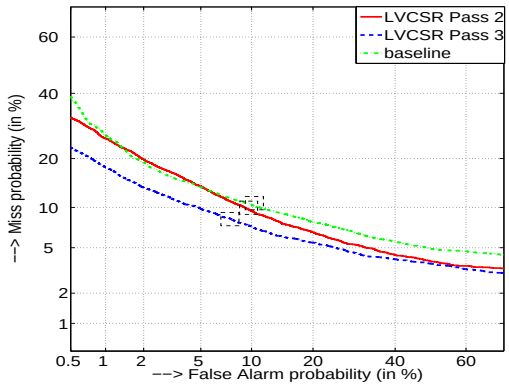

Fig. 1. DET plot - Performance of STD system for different LVCSR pass. The boxes highlight EER - operating points.

2. STD BASED ON WORD LATTICES GENERATED BY LVCSR

The STD system used in the experiments is based on a Large Vocabulary Continuous Speech Recognition (LVCSR) system. Multiple hypotheses are obtained and compactly represented in the form of a word lattice. A word lattice is a directed, acyclic, and weighted graph, where each node represents a time instance and each edge represents the word hypothesis along with its acoustic model likelihood and the language model probability.

The word posterior probability is defined as:

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

where W_i is the hypothesized word identity spanning time interval $t \in (t_s, t_e)$. t_s and t_e denote 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 hypothesis sequences in the lattice that contain the hypothesized word W_i spanning time interval $t \in (t_s, t_e)$.

$P(W_i; t_s, t_e)$ can be estimated from the lattice using the forward-backward re-estimation algorithm [3]. Given a spoken term, the hypothesized word with the maximum confidence score $P(W_i; t_s, t_e)$ is selected from the cluster of overlapping word hypotheses.

2.1. LVCSR system

The LVCSR used in the experiments is based on the Conversational Telephone Speech (CTS) system, derived from AMI[DA] LVCSR [5]. 250 hours of Switchboard data is used for training Hidden Markov Models (HMMs). The decoding is done in three passes, always with a simple bigram Katz backoff Language Model (LM). In the first pass, PLP features (accompanied with delta coefficients) are used and processed by Heteroscedastic Linear Discriminant Analysis (HLDA) to perform a robust data-driven dimension reduction. HMMs are trained using a Minimum Phone Error (MPE) procedure. In the second pass, Vocal Tract Length Normalization (VTLN) is employed on similar features to pass 1. In addition to HLDA, MPE and Speaker Adaptive Training (SAT) are applied. Finally, the third pass is similar to the second pass, except input PLP features are replaced by posterior-based features estimated using a Neural Network (NN) system. The NN processes 300 ms long temporal trajectories of Mel filter-bank energies. The NN is represented by a Multi-Layer Perceptron (MLP) with 1 hidden layer (500 neurons). The LVCSR system reaches a Word Error Rate (WER) of 2.9% on the Wall Street Journal (WSJ1) Hub2 test set from November 1992 (2.5 hours, with 5K dictionary and a trigram LM).

STD	EER [%]	MTWV
baseline	10.13	0.358
LVCSR pass 2	9.66	0.478
LVCSR pass 3	8.04	0.565

Table 1. Equal Error Rates (EERs) and Maximum Term Weighted Values (MTWVs) computed for NIST STD 2006.

2.2. Evaluation of stand-alone STD system

The LVCSR word lattice based STD system is evaluated on three hours of two channel CTS English development database distributed by NIST for the 2006 Spoken Term Detection (STD06) task [1]. The speech recordings are first segmented into shorter sub-segments using a speech-silence segmentation algorithm which removed around 50% of the data. Then, word lattices are generated using the previously described LVCSR system with a dictionary containing 50K words. The generated bigram lattices are subsequently expanded with a trigram language model.

One-half of the 1107 English search terms are randomly selected from the list defined for the dry-run set distributed for the STD06 evaluation. False alarm probabilities and miss probabilities in the STD task are evaluated. Performance is shown using a standard Detection Error Trade-off (DET) curves [6]. In addition, we also present Equal Error Rates (EERs), a one-number metric, mainly used to optimize the system performance. Figure 1 and Table 1 shows the performance of the STD built on a 3-pass LVCSR system. One can see that word lattices generated in the third pass provide significantly better performance than those in the second pass. STD performance is also compared to the baseline system described in [7]. The baseline system achieves EER of about 10.1%. The STD built on 3-pass LVCSR gives EER about 8% (20% relative improvement).

Besides EER and DET, we use a Term-Weighted Value (TWV) evaluation measure defined by NIST STD06 [1], which is also a one-number metric. TWV is estimated first by computing the miss and false alarm probabilities for each term separately, then using these and a pre-determined prior probability to compute term-specific values, and finally averaging these term-specific values over all terms to produce an overall system value $TWV(\theta)$ [8]. In particular, we use Maximum Term-Weighted Value (MTWV) computed over the range of all possible values (θ). MTWV ranges from 0 to $+1$. The achieved results together with EERs are shown in Table 1.

3. OOL DETECTION MODULE

The goal of an OOL detection is to identify segments in the input recordings which do not contain speech pronounced in the target language. The STD system used in the experiments is developed for detection of English spoken terms.

The OOL detection module exploits several individual frame-based Confidence Measures (CMs), which are later combined into a global confidence score. Individual CMs are derived from word as well as phoneme recognition lattices generated for each speech

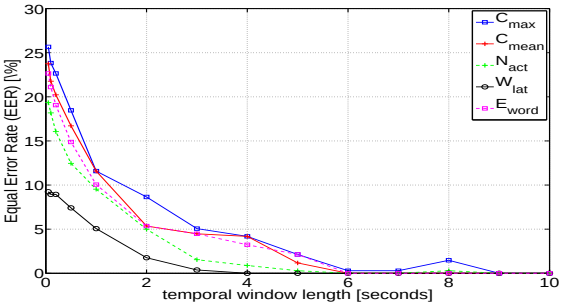

Fig. 2. Equal Error Rates (EERs) of particular OOL CMs for varying length of the temporal window.

CMs - OOL detection	EER [%]
E_{word} (individual CM)	32.68
Filtered E_{word} (individual CM)	21.17
MaxEnt combination (global CM)	18.62

Table 2. Equal Error Rates (EERs) for OOL detection provided by three different CMs.

frame by the previously described LVCSR system. We exploit several frame-based CMs derived using various approaches. More particularly, CMs used in the OOL detection are based on maximum and mean word posterior probability estimates (C_{max} and C_{mean}, respectively), frame-based entropy of word and phoneme posterior probabilities (E_{word} and E_{phone}, respectively), width of word recognition lattices (W_{lat}), and number of different active words in word lattices (N_{act}) [4].

To increase the influence of CMs in OOL detection, we incorporate temporal dependencies (context) by performing temporal filtering of previously estimated CMs. A relatively simple median filter is employed to incorporate temporal context. Optimal length of the temporal window is analyzed on a development set.

Furthermore, word and phoneme-based CMs, generated by the individual techniques and post-processed by median filter, are combined to obtain a global CM. The combination is provided by a Maximum Entropy (MaxEnt) criterion [9]. MaxEnt uses conditional maximum entropy models, which have been shown to provide good performance in speech and language processing (language modeling, parsing). The MaxEnt classifier was trained on different data (WSJ corpus), as described in [4].

3.1. Evaluation of stand-alone OOL detection module

The OOL detection module is evaluated on a test set comprising 30 min. of audio-visual recordings [10]. More particularly, in each recording, a subject poses a question in the native language (Czech, German). Then, the subject is asked to repeat the question in English (non-native but target language). In addition, the test set also contains speech recordings pronounced by subjects in one language only. In order to eliminate possible OOV words during decoding, all English words appearing in the test recordings are included in the vocabulary. The evaluation data were manually annotated for the OOL detection task. Therefore, each speech recording contains information about the time segments of the target and alien languages.

A length of temporal window (filter) is estimated on 10 min. of development data (OGI multilanguage corpus [11]). More particularly, recordings from 4 different languages (English, Farsi, German, Mandarin Chinese) of the corpus are used. OOL detection is then performed on test data where CMs are processed by such the median filter. Figure 2 plots EERs for various temporal window lengths on OGI development data for various OOL CMs. Since each recording in OGI corpus is pronounced in one language, the perfect OOL detection can be achieved by applying long enough window of the median filter, as shown in Figure 2. However, as mentioned in [4], too long window would cause a significant decrease of the OOL detection accuracy for mixed language (target and alien) scenarios. Withe respect to the analysis results shown in Figure 2, we chose the window length equal to 3 sec. (most of CMs already achieve good performance with such the filter length).

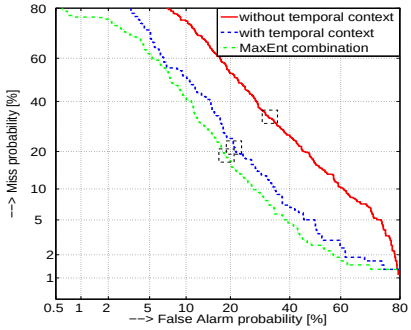

Fig. 3. DET plot - OOL detection using E_{word} without and with application of temporal context, and subsequent MaxEnt combination of all the individual (frame-based) CMs. The boxes highlight EER - operating points.

Similar to the STD evaluation, false alarm probabilities and miss probabilities in OOL detection are represented using DET curves. Figure 3 shows the set of DET curves representing the OOL detection on test data using various CMs: individual word entropy based CM (E_{word}), its subsequent temporal filtering, and the combination of all individual (filtered) CMs using MaxEnt combination. Table 2 shows EER performances of these three CMs on test data. E_{word} (base-line) estimated from word lattices gives EER of about 32%. Incorporating temporal context by employing median filter and subsequent MaxEnt combination significantly improves OOL detection performance (over 40% relative improvement).

4. COMBINATION OF STD AND OOL DETECTION SYSTEMS

Similar test data, used for evaluation of the OOL detection module, is used to get the performance of the STD system combined with the OOL detection module. A list of terms contains 59 English words occurring in the test set. More than 30% of the test data contains speech pronounced in a different language than the target language of the STD system. The speech segments pronounced in English do not contain OOVs.

The baseline STD performance on the test set (without using any OOL information) represented by DET curve and EER is given in Figure 4 and Table 3, respectively. If we remove all false alarms (spoken terms) detected in the speech segments of the alien language, the STD performance (denoted as STD - OOL ground truth) obviously improves, as can also be seen in Figure 4 and Table 3. This is done using ground truth information accessible from the manual annotation.

In other experiments, STD uses confidence scores from the OOL

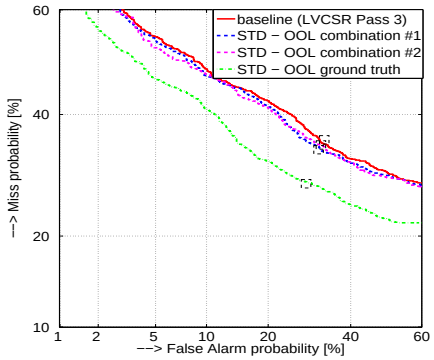

Fig. 4. DET plot - STD performance (baseline) and its combination with OOL detection module. The boxes highlight EER - operating points.

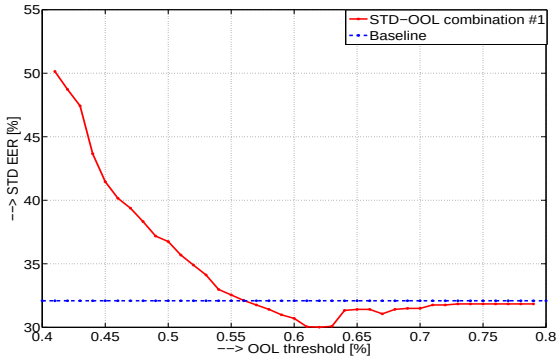

Fig. 5. Equal Error Rates (EERs) and Maximum Term-Weighted Values (MTWVs) of STD for varying OOL detection threshold.

STD	**EER** [%]	**MTWV**
baseline (LVCSR Pass 3)	32.1	0.1904
STD - OOL ground truth	26.0	0.2476
STD - OOL combination (approach #1)	**30.0**	**0.2427**
STD - OOL combination (approach #2)	31.3	0.1872

Table 3. Equal Error Rates (EERs) for STD without and with application of OOL detection module.

detection module. In approach #1, the STD confidence scores are "hard" thresholded, i.e., once the input speech segment is classified to be OOL segment, the scores of corresponding terms detected by STD (appearing in the same segment) are set to zero. Graphically, the resulting performance is given in Figure 5, where EER of STD is plotted for varying OOL threshold. The best achieved EER of the STD system for the optimal OOL detection threshold is equal to 30%, as shown in Table 3. The corresponding STD DET curve is given in Figure 4.

In approach #2, the confidence scores from the OOL detection module are used as weights for STD scores, i.e., STD confidence scores are multiplied by the corresponding OOL detection confidence scores. The best achieved STD EER is then 31.3%.

Audio recordings used for testing the STD – OOL combination are pronounced by non-native English speakers and contain significantly higher level of noise compared to STD06 test data. Performance of the LVCSR system on English parts of test data (in terms of WER) is about 42.5%. Due to that, the overall STD accuracies are worse than those achieved with the same STD system on STD06 test data.

5. DISCUSSIONS AND CONCLUSIONS

This paper proposes a combination of an STD system with an OOL detection module to improve detection accuracies of English spoken terms. First, the current version of the stand-alone STD system is compared to the baseline system on NIST 2006 evaluation data. About 20% relative improvement in EER is achieved. Then, two approaches to combine STD and OOL systems are evaluated on recordings with potential occurrence of speech segments from different languages. Although, performance of the STD system in such the scenario does not reach results when the manual OOL annotation is provided (6% absolute improvement in EER over the stand-alone STD), the final STD – OOL performance improves by more than 2% absolute (7% relative) achieved by approach #1.

6. REFERENCES

[1] NIST Spoken Term Detection Evaluation, <http://www.nist.gov/speech/tests/std>, 2006.

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

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

[4] 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.

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

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

[7] 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.

[8] D. R. H. Miller, M. Kleber, C. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and Accurate Spoken Term Detection", *in Proc. of Interspeech*, Antwerp, Belgium, 2007.

[9] C. White, J. Droppo, A. Acero and J. Odel, "Maximum entropy confidence estimation for speech recognition", *in Proc. of ICASSP*, pp. 809-812, Hawaii, USA, 2007.

[10] Czech University of Technology - data website: <http://cmp.felk.cvut.cz/projects/dirac/data/Dirac-CMPdata-16.html>

[11] R. Cole and Y. Muthusamy, "OGI Multilanguage Corpus", Linguistic Data Consortium, Philadelphia, USA, 1994.