Automatic Out-of-Language Detection based on Confidence Measures derived from LVCSR Word and Phone Lattices

Petr Motlicek

Idiap Research Institute, Martigny, Switzerland
motlicek@idiap.ch

Abstract

Confidence Measures (CMs) estimated from Large Vocabulary Continuous Speech Recognition (LVCSR) outputs are commonly used metrics to detect incorrectly recognized words. In this paper, we propose to exploit CMs derived from frame-based word and phone posteriors to detect speech segments containing pronunciations from non-target (alien) languages. The LVCSR system used is built for English, which is the target language, with medium-size recognition vocabulary (5k words). The efficiency of detection is tested on a set comprising speech from three different languages (English, German, Czech). Results achieved indicate that employment of specific temporal context (integrated in the word or phone level) significantly increases the detection accuracies. Furthermore, we show that combination of several CMs can also improve the efficiency of detection.

Index Terms: LVCSR, Confidence Measure, Out-Of-Language (OOL) detection

1. Introduction

Confidence Measures (CMs) are commonly used to detect recognition errors obtained by Large Vocabulary Continuous Speech Recognition (LVCSR) systems. In preliminary work, C_{max} measure estimated from LVCSR word lattices has been shown to be the best performing confidence measure for recognition error detection [1]. Frame-based posterior CMs have been used to improve speech recognition performaces in hybrid HMM/ANN systems [2, 3]. As reported in [4], CMs estimated from LVCSR word and phone lattices were shown to be good estimates, not only for detection of errors, but also for detection of Out-Of-Vocabulary (OOV) words [5].

In this paper we employ techniques similar to those used in [4] for OOV detection, but we focus on a different task: to detect time segments in input recordings containing speech pronounced in a non-target (alien) language. Such a task is called Out-Of-Language (OOL) detection. By alien language, we mean a language for which the LVCSR is not built. Interchangeable use of different languages in short time periods by the same speaker can often be registered in spontaneous speech recordings (e.g., meeting data). This introduces many difficulties for LVCSR systems which are built to recognize spontaneous speech pronounced in one language. However, if the input speech is pronounced in a different language, LVCSR will not be able to detect it. Instead it will output the most probable sequence of words generated according to the dictionary used.

OOL detection can partially be seen as a Language Identification (LID) task. Unlike LID, in OOL detection we do not identify the speaker's language. From a practical point of view, OOL detection system does not exploit the knowledge of other (non-target) languages, which is the case in LID.

The goal of this paper is to explore confidence measures in OOL detection. Specifically, we investigate the use of frame-based word and phone posterior probabilities (referred as posteriors in the paper). Then, we propose to exploit information extracted from temporal context, which is accomplished by the time-domain filtering of previously obtained OOL detection thresholds. This yields a significant improvement in the OOL detection. Finally, we show that combination of post-processed OOL thresholds, obtained from individual word and phone LVCSR posteriors using a Maximum Entropy (MaxEnt) approach, increases detection performance.

2. Estimation of posteriors from LVCSR lattices

In this work, posterior probabilities are estimated from LVCSR outputs represented as recognition lattices. The posterior probabilities are estimated for each 10 ms speech segment (frame). Word and phone is used as a basic unit of the posteriors, denoted as $p(u \mid t)$. u is the respective unit and t denotes time in frames.

Arcs in the word lattice represent hypothesized words W_i^j, where W_i is the word identity (selected from a dictionary) and j is the occurrence of word W_i in the lattice. Word posterior probabilities $p(W_i^j)$ are computed from the associated Acoustic Model (AM) and Language Model (LM) scores using the standard forward-backward algorithm. $p(W_i^j)$ are estimated for every speech frame. Phone posteriors are estimated in the same manner to word posteriors. Phone recognition lattices are generated from LVCSR outputs and processed by the forward-backward algorithm to obtain phone posteriors $p(P_i^j)$, where P_i represents a hypothesized phone.

3. Posterior based CMs used for OOL detection

Here, we explore several kinds of Confidence Measures (CMs) estimated from frame-based word and phone posteriors. For simplicity, only word CMs will be described here. Phone CMs are estimated in a similar manner to the word CMs. In particular, five different CMs are considered in our experiments:

- C_{max}**:** As a base-line, C_{max} confidence measure is used defined as a maximum posterior probability of hypothe-

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. We thank to Lukas Burget and Martin Karafiat (Brno University of Technology) for providing the LVCSR setup.

sized word W_i for time t spanning interval $t \in (t_s, t_e)$, where t_s and t_e denote start and end time of the interval, respectively

$$C_{max} = \max_{t \in (t_s, t_e)} p(W_i \mid t). \quad (1)$$

- **Mean word entropy:** First, we define frame-based word entropy $H(W \mid t_n)$ which is a measure of the amount of uncertainty associated with W for the given time instance $t = t_n$

$$H(W \mid t_n) = \sum_i \frac{1}{p(W_i \mid t_n)} log_2\big(p(W_i \mid t_n)\big). \quad (2)$$

In order to obtain word-based CM, we apply the formula from [6]

$$H\big(W \mid t \in (t_s, t_e)\big) = \frac{\sum_{t \in t_s, t_e} H(W \mid t)}{1 + \alpha(t_e - t_s - 1)}. \quad (3)$$

Here, the denominator is used to "smoothly" normalize the frame-based CM (e.g., word entropy). For $\alpha = 0$, no normalization takes place, whereas for $\alpha = 1$, $H(W \mid t)$ is fully normalized by the length of the current word (in frames). Due to this normalization, we can control the significance of the word-based CM which can vary a lot from short to long words.

- **Mean word posterior C_{mean} (later denoted as fWER):** It is defined as a mean posterior probability of a hypothesized word W_i spanning time interval $t \in (t_s, t_e)$

$$C_{mean} = \frac{\sum_{t \in (t_s, t_e)} p(W_i \mid t)}{1 + \alpha(t_e - t_s - 1)}. \quad (4)$$

C_{mean} is related to a minimum time Frame Word Error (fWER) defined in [6]. C_{mean} is normalized by the same denominator as used in Equation 3.

- **Word recognition lattice width:** This CM is obtained by counting number of active arcs at the given time instance $t = t_n$ from the recognition lattices. Similar to $H(W \mid t_n)$, it is a measure estimating the amount of uncertainty in the LVCSR system at the given time frame. Frame-based lattice width counts are smoothed in similar manner as in Equation 3.
- **Number of different active words:** In contrary to the lattice width, number of words (active and unique at the given time frame) is counted, smoothed using Equation 3, and also used as a word-based CM.

4. Post-processing of word and phone CMs and their combination

Frame-based CMs estimated from the frame-based word and phone recognition lattices are converted into word-based CMs, as described above. To increase the influence of CMs in OOL detection, we incorporate temporal dependencies (context) by applying temporal filtering of previously estimated CMs. The importance of temporal context largely depends on the recognition task (kind of input speech recordings). In general, this can be intuitively accepted especially in the case of spontaneous speech recognition. Usually, the language used by a speaker in spontaneous speech recordings does not change too quickly. It can rather be seen as a slowly varying process (e.g., with time response of an average sentence length). On the other hand, assuming too long a time constant would cause a decrease in OOL detection accuracy due to slow response of the system. In our experiments, a relatively simple median filter is employed to represent temporal context. In the experiments, the optimal length of the temporal window is analyzed.

Furthermore, word-based CMs, generated by the individual techniques described above, are combined to obtain a global CM. The combination is provided by a Maximum Entropy (MaxEnt) criterion [7]. MaxEnt uses conditional maximum entropy models which have been shown to provide good performance in speech and language processing (language modeling, parsing).

5. Experimental setup

5.1. LVCSR system

The system used in the experiments is based on the Conversational Telephone Speech (CTS) system, partially described in [4], derived from AMI[DA] LVCSR [8]. 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 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 from 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 Word Error Rate (WER) of 2.9% on Wall Street Journal (WSJ1) Hub2 test set composed from November 92 (2.5 hours, with 5k dictionary and a trigram LM).

5.2. Evaluation data

For the evaluation, we used the audio-visual data recorded at CUT [9]. The evaluation set consists of 50 audio (16 kHz) recordings (∼30 minutes). 4 different speakers appear in the evaluation set. The data was recorded to study rare audio-visual events. More particularly, in each recording, a subject asks a question in the native language (Czech, German). Then, the subject is asked to repeat the question in English. The vocabulary used in the LVCSR system contains the 5000 most frequent words from the LM training corpora. In order to eliminate possible OOV words, all English (target language) words appearing in the test recordings are included in the dictionary.

The evaluation data was manually annotated for the OOL detection task. Therefore, each speech recording contains information about the time segments of the target and alien languages.

5.3. Evaluation

The goal of OOL detection is to identify segments in the input recordings which do not contain speech pronounced in the target (English) language. The evaluation procedure is then simi-

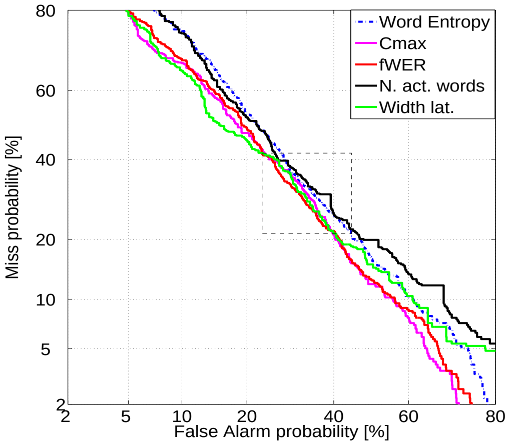

Figure 1: *DET plot - OOL detection using different word CMs without application of temporal context. The box highlights EER - operating point.*

lar to OOV detection – mis-recognized words overlapping with OOL speech segments are detected. More particularly, thresholds provided by CMs are associated with individual words (and their time alignment information) obtained by LVCSR one-best output.

False alarm probabilities and miss probabilities in the OOL detection task are evaluated on the evaluation set. Performance is shown using a standard Detection Error Trade-off (DET) curves. Since one-number metrics such as Equal Error Rate (EER) or Cross-over Error Rate (CER) are dependent on the ratio of correct targets to overall number of tokens, as pointed out in [4], we use them only to optimize the system performance. The operating point of the OOL detection system is therefore open by illustrating the whole DET curves.

6. Experimental results

6.1. Word CMs

In Figure 1, we show the set of DET curves representing the OOL detection using (word-based) word CMs. The best performance is achieved by CMs based on lattice width and fWER (depending on the selected operating point). The smallest EER is achieved by fWER.

6.2. Temporal context

As a result of our experiments, the incorporation of temporal context provided by median filtering of word CMs significantly improves OOL detection performance. An example of OOL detection without and with the application of temporal context is given in Figure 2 for a word entropy based CM. More specifically, by adding 1.5 sec temporal context, the EER decreased from 32.68% to 21.17%. This time interval approximately corresponds to 5 words (the average length of recognized words in the evaluation set is 0.3 sec). To analyze the optimal length of the temporal window (characterizing the median filter), the window length varied from 0.3 to 12 sec. The results are graphically shown in Figure 4 for different word CMs. The smallest EER is achieved by the fWER based CM for the length equal to 1.5 sec. One can see that very long temporal context does not improve OOL detection performance. Figures 3 and 5 graphically compare DET curves for all previously defined word CMs without and with the application of a 1.5 sec long median filter, respectively.

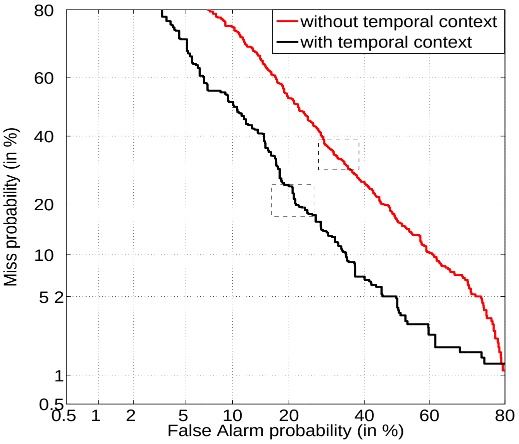

Figure 2: *DET plot - OOL detection using word entropy based CM without (red curve) and with (black curve) application of temporal context (provided by the optimal length of median filter). The boxes highlight EER - operating points.*

6.3. Combination of CMs

In the following experiments, the MaxEnt approach is used to combine individual word CMs. Previously mentioned Figures 3 and 5 also show DET curves for MaxEnt combination. MaxEnt combination does not improve OOL detection performance if temporal context is not included in individual CMs (Figure 3). However, with application of temporal context, the detection due to MaxEnt combination improved, especially for low false alarm operating points. This can partially be seen in Figure 5, although here we also used phone CM for combination using MaxEnt approach.

Due to the low amount of annotated data for OOL detection, the MaxEnt classifier was trained on part of WSJ0 development set (3.6 hours) annotated for OOV detection.

6.4. Phone CMs

We also experimented with phone CMs, i.e., CMs are estimated for every speech frame from phone recognition lattices (generated using the word LM). We performed experiments with the phone entropy based CMs, which are estimated in a similar way to the word entropy (Equation 2) but from the phone LVCSR lattices. Frame-based phone CMs are then converted into word-based phone CMs using Equation 3.

Achieved EER using the phone entropy based CM is equal to 34.6% (compared to 31.02% - the best word CM). However, by adding temporal context (due to employment of median filter), the phone entropy achieves the best performance amongst all the individual techniques used for OOL detection. In Figure 4, EERs obtained using phone CM are compared to word CMs for different lengths of median filter. The optimal window length for the phone entropy based CM (3 sec) is approximately two times longer than for the word entropy based CM (1.5 sec).

Subsequent MaxEnt combination of all hitherto presented word and phone CMs gives the best performance for DET operating points around EER, as seen in Figure 5. However, for systems required to operate with a low number of false alarms,

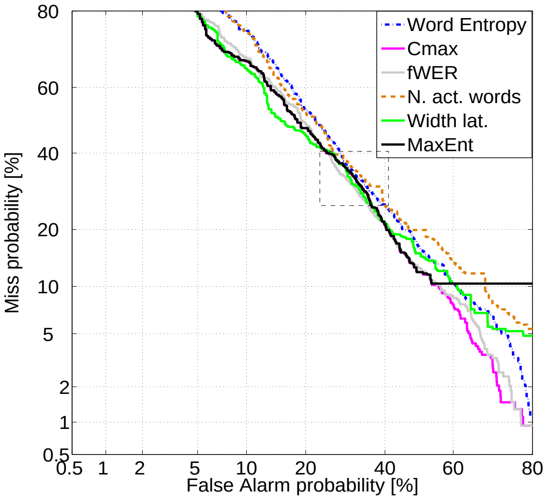

Figure 3: *DET plot - OOL detection using different word CMs and their MaxEnt combination without application of temporal context. The box highlights EER - operating point.*

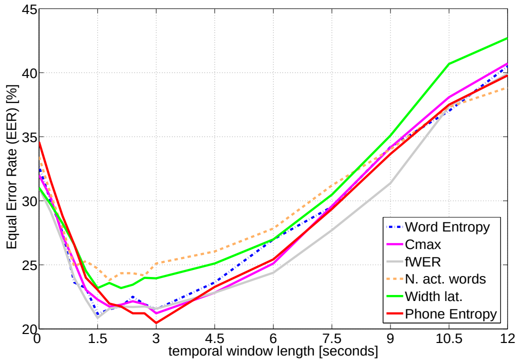

Figure 4: *OOL detection evaluated using EER for various lengths of temporal window applied on word and phone CMs.*

a simple phone entropy based CM outperforms MaxEnt combination.

7. Discussions and conclusions

In this paper, we proposed a system for OOL detection, i.e., a system to automatically detect segments containing speech pronounced in non-target languages. The detection is based on confidence measures estimated by processing word and phone recognition lattices obtained by LVCSR. LVCSR was trained on telephone speech of English pronunciations. The proposed OOL detection system was tested on the data comprising speech pronounced in three languages (English, German, Czech).

The base-line system employing CMs estimated from LVCSR lattices gives EER performance about 32%. Incorporating temporal context by employing the median filter, OOL detection performance significantly improved (over 30% relative improvement). Subsequent combination of word and phone CMs using a MaxEnt approach yields the best performance for operating points around EER (EER $\sim$ 18.7%). However, for a low number of false alarms, a simple phone entropy based CM performs the best.

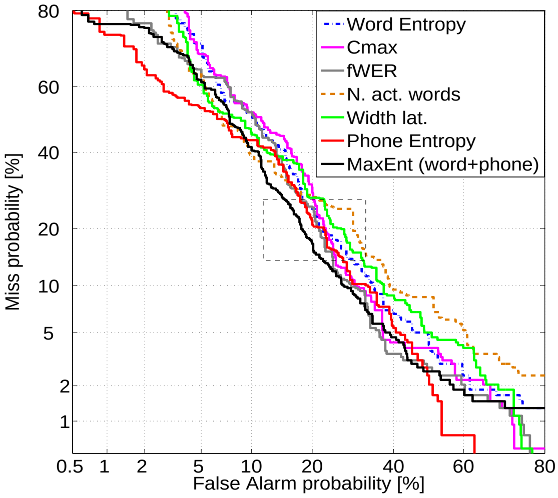

Figure 5: *DET plot - OOL detection using different word and phone CMs and their MaxEnt combination with application of temporal context. The plot clearly shows that the best performance for EER (highlighted by the box) is provided by MaxEnt combination. The best performance for low alarms is provided by the phone entropy based CM.*

8. References

[1] 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. 288298, 2001.

[2] G. Bernardis and H. Bourlard, "Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems", *in Proc. ICSLP*, pp. 775-778, Sydney, Australia, 1998.

[3] G. Williams and S. Renals, "Confidence Measures for Hybrid HMM/ANN speech recognition", *in Proc. of Eurospeech*, pp. 1955-1958, Rhodes, Greece, 1997.

[4] L. Burget, et al. "Combination of Strongly and Weakly Constrained Recognizers for Reliable Detection of OOVs", *in Proc. of ICASSP*, pp. 4081 - 4084, Las Vegas, USA, 2008.

[5] H. Ketabdar, M. Hannemann and H. Hermansky, "Detection of Out-of-Vocabulary Words in Posterior Based ASR", *in Proc. of Interspeech*, pp. 1757-1760, Antwerp, Belgium, 2007.

[6] F. Wessel, R. Schluter, H. Ney, "Explicit Word Error Minimization using Word Hypothesis Posterior Probabilities", *in Proc. of ICASSP*, pp. 33-36, Salt Lake City, USA, 2001.

[7] 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.

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

[9] Czech University of Technology - data website: <http://cmp.felk.cvut.cz/projects/dirac/data/Dirac-CMPdata-16.html>, and mirror:
<http://sirius.physik.uni-oldenburg.de/members/bach/>.