Combining Lattice-Based Language Dependent and Independent Approaches for Out-of-Language Detection in LVCSR

Yuxiang Shan, Yan Deng, Jia Liu

Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, China

Abstract

In this paper, Out-Of-Language (OOL) detection problem is handled by both language dependent (LD) and language independent (LI) approaches. In the LD approach, a novel speech content and language joint recognition algorithm is proposed, which integrates a phone lattice-based vector space modeling language recognition (LRE) backend into the conventional speech decoding procedure. In the LI approach, lattice derived confidence measures are used. Since these two approaches reflect two different dimensions of uncertainties encoded in lattices, combining them improves both the LRE and OOL detection performance. Experiments also show that for LD approach the detection accuracies can be significantly increased by applying heuristic phone lattice reconstruction. Evaluated on a Mandarin/English mixed conversational telephone speech corpus with a Mandarin speech recognizer, the proposed method achieves an EER of 12.68% in OOL detection, and reduces the recognition error by 33.06%.

Index Terms: large vocabulary continuous speech recognition (LVCSR), language recognition, confidence measure, Out-Of-Language (OOL) detection

1. Introduction

The temptation to add language recognition ability to a large vocabulary continuous speech recognizer arises from two aspects: firstly, for applications such as speech-to-speech translation systems and dialogue systems which often work in multilingual environment, it is necessary to identify language types user speaks before the conversation started. Secondly, for applications designed for one specific target language, since interchangeable use of different languages in short time periods by the same speaker is very common in overseas telephone calls, meeting and other scenes, there is also a need to filter out speech segments of non-target languages to avoid meaningless recognition results. In this paper, we focus on the latter case, which is also called Out-Of-Language (OOL) detection [1].

The OOL detection problem can be seen as either a language verification task or a confidence measure (CM) estimation task which is similar to Out-Of-Vocabulary (OOV) detection, since we do not need to identify what exactly the speaker’s language is. Based on such consideration, [1] and [2] exploited several language independent (LI) CMs derived from frame-based word and phone posterior for OOL detection and remarkable progress has been made. However, as will be shown later, there are two dimensions of uncertainties encoded in speech recognition results (e.g. word/phone lattices). The LI approach only makes use of uncertainties in the dimension of recognition hypotheses (recognized word or phones) at a fixed time point (frame-based CMs) or in a short period of time (phone/word-based CMs). In this paper, we propose to use a language dependent (LD) approach to make use of uncertainties along the time dimension, namely the context (n-gram statistics) of recognition hypotheses, which is a helpful supplement to the LI approach.

Widely used lattice-based language recognition (LRE) techniques [3][4][5] which model high level features such as words or phones and their context are very suitable candidates for implementing the LD approach. However, it is not appropriate to use them directly, since OOL detection has its own characteristics. For example, the OOL segments are usually very short, in many cases only a few words or seconds long, and embedded without clear boundaries in a continuous spontaneous speech stream. So it is very difficult to precisely locate and segment them out using a voice activity detector (VAD) and then recognize them using a common LRE system. Moreover, in typical lattice-based LRE systems, phone recognizers with simple phone loop and null grammar are used, which guarantees that each phone has equal probability to be recognized. When working with a LVCSR system in which the decoding network is strictly constrained by the target language’s pronunciation dictionary and language model, the word/phone connections in the generated lattices will also keep these restrictions, and thus set a heavy bias towards the target language and cause great performance degradation.

To overcome the above two difficulties, we propose to use a novel speech content and language joint recognition algorithm to implement the LD approach. By tightly integrating a phone lattice based vector space modeling (VSM) LRE backend into the conventional speech decoding procedure, this algorithm can make full use of phonetic hypotheses and their time alignment information collected during decoding, and detect OOL segments simultaneously. The pronunciation and grammar restrictions of the target language are eliminated by a heuristic phone lattice reconstruction algorithm which significantly increases the recognition accuracy. There are, of course, other alternatives for the LRE backend, such as the language model (LM) backend. We choose to use VSM because of its superior performance in many previous works [3][6][7].

The paper is organized as follows. Section 2 and Section 3 describe the LD approach and the LI approach respectively. Section 4 compares and combines these two approaches. Experimental setup, results and analysis are given in Section 5 and 6. Section 7 concludes the paper.

2. LD approach for OOL detection

2.1. Speech content and language joint recognition

Given an observation sequence \( \mathbf{O} \), the conventional speech recognition algorithm determines the most probable word sequence \( \mathbf{W}^* \) using the Bayes’ rule:

\[
\mathbf{W}^* = \arg \max \limits_{\mathbf{W}} p(\mathbf{W} | \mathbf{O}) .
\] (1)
In order to perform joint recognition, we add language dependent items to Eq. (1):

\[
(W^*, L^*) = \arg\max_{W, L} p(W, L | O) = \arg\max_{W, L} \left[ p(L | W, O) \right] \cdot p(W | O)] , \tag{2}
\]

where \(L^*\) represents the most probable language sequence. When a phone lattice based VSM backend is used for language recognition, only speech recognition results (e.g., phone lattices) are needed as the input to the classifier, so Eq. (2) can be further written as:

\[
(W^*, L^*) = \arg\max_{W, L} \left[ p(L | W, O) \right] \cdot p(W | O)] . \tag{3}
\]

In Eq. (3), \(p(W | O)\) is the speech recognition item, which is normally solved using the Viterbi decoding. \(p(L | P)\) is the language recognition item, which is solved by VSM method in this paper. \(p(P | W)\) represents a mapping relation from a word sequence \(W\) to a phone sequence \(P\), which, in most cases, is a one-to-one mapping due to the existence of pronunciation dictionary. We will modify this mapping item using a heuristic phone lattice reconstruction algorithm (to be described in Section 2.2) to eliminate the restrictions and get better LRE performance. Since the directly optimization of Eq. (3) is very difficult, and it is not necessary to perform language recognition every frame, we simplify the decoding procedure as follows:

**Step 1** Start decoding the observation sequence \(O\) using the same Viterbi decoding procedure as the conventional LVCSR. Collect phone hypotheses at each frame using the method described in Section 2.2.

**Step 2** At the \(n\)-th possible language turning point \(t_n\) reconstruct phone lattice spanning time interval \([t_n, t_{n+1}]\) using the previously collected phone hypotheses.

**Step 3** Feed the reconstructed lattice into VSM LRE backend and record the recognition results.

**Step 4** Go to Step 1, continue decoding if the observation sequence is not finished, else trace back and generate \(W^*\) and \(L^*\).

The language turning point \(t_n\) can be selected in many ways. In our implementation, we choose it dynamically from word boundaries according to the time alignment information in the 1-best LVCSR result up to that frame. This solves the “location” problem in OOL segment detection. In order to get a robust and reasonable LRE performance, a minimum interval length \(T_{min}\) is set as a threshold that \(t_n \geq t_{n-1} + T_{min}\). LRE performances under different segment lengths are analyzed in experiments to select appropriate \(T_{min}\).

### 2.2. Phone lattice reconstruction

For the convenience of description, we use the term original phone lattice to denote phone lattice generated by the conventional LVCSR. A lattice \(L = (N, A, n_{start}, n_{end})\) is a directed acyclic graph (DAG) which is specified by a finite set of nodes \(N\), a finite set of arcs \(A\), a start node \(n_{start} \in N\) and an end node \(n_{end} \in N\). Given an arc \(a \in A\), \(S[a]\) denotes its start node, \(E[a]\) its end node, \(l[a]\) its phone/word identity, \(ac[a]\) its acoustic likelihood and \(lm[a]\) its language model likelihood.

The purposes of the lattice reconstruction are 1) to encode richer decoding alternatives and contexts 2) to break pronunciation and grammar restrictions in the original lattice. These are achieved in two phases:

- **Phone hypothesis collecting phase**
  
  In this phase, phone hypotheses generated at each frame during the Viterbi decoding (Step 1 in Section 2.1) are pushed into a phone hypothesis pool. If a phone hypothesis of the same identity and start/end frame number has already existed in the pool, the one with larger acoustic likelihood is kept. Once a phone hypothesis is pushed, it will never be taken out, even the path that it belongs to is pruned. In this way, we can get a huge number of phone hypotheses that never appears in the original phone lattices. To avoid data sparseness in the VSM backend, left and right context of triphones are squashed out and only monophones are registered in the pool.

- **Lattice reconstruction phase**
  
  In this phase, we reconstruct phone lattice by concatenating the previously collected phone hypotheses frame by frame, regardless of whether they come from the same word or path. This procedure has 3 steps:

  **Step 1** The time interval \([t_n, t_{n+1}]\) initialize phone lattice \(L = (N, A, n_{start}, n_{end})\) with \(N = \{t : t = t_n, t_{n+1}, \ldots, t_{n'}\}, A = \{\}, n_{start} = t_n\) and \(n_{end} = t_{n'}\).

  **Step 2** For \(t = t_n\) to \(t_{n'}\) get the N-best \((N=10\) in our experiments\) phone hypotheses with end frame number \(t\) from the phone hypotheses pool, and store them into a list \(P\). The N-best order is decided by duration normalized log likelihood. For a phone hypotheses \(p\), this normalized score is defined as

  \[
  score_{norm}(p) = \log ac(p) \cdot \frac{E[p] - S[p]}{I[p]} . \tag{4}
  \]

  For each phone hypotheses \(p \in P\), create a new arc \(a\) with \(S(a) = S[p], E[a] = t, I[a] = I[p], ac[a] = ac[p], lm[a] = 0\). These notations have the same meaning for \(p\) and \(a\). Finally, add \(a\) to \(A\).

  **Step 3** Remove unreachable nodes and arcs from \(L\).

### 2.3. VSM backend

In a typical VSM LRE system, phone lattices are first converted into n-gram statistics. Then super-vectors composed of these statistics are taken as samples for support vector machine (SVM) training and testing to identify languages. The VSM backend used in this paper is exactly the same as [7], so we won’t bother the algorithms here. But the training and testing configurations will be detailed in Section 5.2.

### 3. LI approach for OOL detection

In this section, following the work in [1], two typical CMs derived from word LVCSR lattices are employed for OOL detection. Before extracting these CMs, word lattices are converted into posterior lattices using the standard forward-backward algorithm. Same as [1], we use \(p(W_t | t)\) to denote frame-based posterior of word \(W_t\). The two CMs are:

- **Mean word entropy (Wen):** This CM measures the uncertainty of word hypotheses for given time interval \((t_i, t_f)\) using Entropy information criteria:

  \[
  H(W | (t_i, t_f)) = \sum_{n \in N | t_i \leq n} \sum_{a \in A} \frac{\log p(W | t)}{l(a)} , \tag{5}
  \]

  where \(\alpha\) is a smoothing factor between 0 and 1.

- **Word lattice width (Wlat):** This is a simple but very effective CM by counting the number of lattice arcs at a given time point \(t\). Frame-based Wlat can also be smoothed using the similar manner in Eq. (5). The above CMs are further post-processed to incorporate a temporal context using the same manner as [1], i.e. a simple median filter is employed to represent temporal context. The window size of the temporal filter was experimentally chosen to be 2.5s.
4. Comparison and combination of LD and LI approaches

The language recognition scores generated by LD approach can be viewed as a form of CMs which measures the uncertainty of contexts (more specifically, the n-gram statistics) of the recognition hypotheses. Although temporally filtered, the CMs used in the LI approach still mainly reflect the uncertainty of word hypotheses for a given time interval or time instance. These two dimensions of uncertainties (as illustrated in Figure 1) can both be used for OOL detection.

As is generally agreed, fusion of multiple CMs of different natures improves performance. In this paper, we combine the LD and LI CMs into a global CM. This combination is carried out on word level by logistic regression using the FoCal toolkit [8]. More specifically, the LD score of a time interval is assigned to all the words it overlapped with in the 1-best LVCSR result, and then fused with LI CMs of that word.

5. Experimental setup

5.1. Evaluation data set

For experiments, we evaluate the proposed approaches as a Mandarin/English detector. Mandarin is the target language for which a LVCSR system is built. English is the OOL that we need to filter out. Due to the low amount of annotated data for OOL detection, we manually construct two data sets using the CallHome (CH) databases:

- **Dataset A**: This data set is used for evaluating LRE performance of the LD/LI approach as well as the lattice reconstruction and the optimal $T_{\text{min}}$ selection. This data set contains 17,948 utterances, 5.5 hours of Mandarin and 7.2 hours of English from CH development and test set. Each utterance is segmented from original recordings according to the LDC provided transcriptions. The utterance length distribution is illustrated in Table 1.

- **Dataset B**: This data set is constructed for evaluating OOL detection performance by randomly concatenating utterances in Dataset A under the patterns (M-Mandarin, E-English): M-E-M, E-M-E, M-E, E-M, M, E. To make sure Mandarin dominates this data set and languages won’t switch too often (i.e. less than 1s), only Mandarin utterances longer than 3s and English utterances of length 1–3s from Dataset A are selected for concatenation.

5.2. The baseline LVCSR system

The baseline Mandarin LVCSR system used in this paper is a one-pass lexicon-tree based Viterbi decoder with a 68k-word dictionary. PLP with delta and acceleration coefficients are used as feature. Cepstral mean/variance normalization is applied. A hidden Markov model of cross-word triphones with 6k states and 48 mixtures is trained with about 300 hours speech selected from the training set of HKUST, CallFriend (CF), CH and Chinese 863 corpus. The language model is trained with Gigaword corpus and interpolated with transcriptions of the acoustic model training data. The acoustic model is trained with Minimum Phone Error criterion.

5.3. VSM backend and lattice generation

The VSM training data consists of 9,475 Mandarin utterances and 10,827 English utterances select from CF and CH training set. Each utterance was automatically segmented using VAD to have about 30s speech. For comparison purpose, two VSM backends are built, one for the original lattices and the other for the reconstructed lattices. In the following experiments, the training and testing lattices for each of the two VSMs are generated using the same manner to make training and testing conditions matching. Reconstructed phone lattices as well as original phone lattices and word lattices are generated at the same time using the Step 1 and 2 in Section 2.1. Since the VSM training data are pure in language, in generating the VSM training lattices, we set the minimum interval length $T_{\text{min}} = \infty$, which means phone lattice is reconstructed for the whole utterance. From phone lattices, we count up to trigram statistics which result in super-vectors of 894,048-dim. (There are 96 distinct monophones in our dictionary excluding sil and sp.) SVMTorc [9] is used for SVM training and test.

6. Experimental results

6.1. LRE performance and $T_{\text{min}}$ selection

In this experiment, we examine language detection accuracies of different approaches on Dataset A. Utterance false alarm and miss probabilities are evaluated, and results are presented using Detection Error Trade-off (DET) curves in Figure 2. To get utterance-based CMs, we set $T_{\text{min}} = \infty$ in the LD approach. For the LI approach, we accumulate and normalize frame-based CMs on the whole utterance in the similar way as Eq. (5) using the original phone lattices. As expected, both the LD and LI approaches show their language discriminability. The best performance is achieved by LD approach with phone lattice reconstruction (LD RL). Compared with the original lattices (LD OL), it is interesting that lattice reconstruction has no effects on low false alarm or low missing operating points although the performances around EER are significantly improved. One can also notice that LI CMs estimated from the reconstructed lattices (Wlat RL and Went RL) result in much worse performances. This is because lattice reconstruction adds too much confusion in the recognition hypothesis dimension that distributions of the LI CMs are flattened.

To select the most appropriate $T_{\text{min}}$ value for the following OOL detection experiments, Table 1 gives the LRE performances of different utterance lengths in form of Equal Error Rate (EER). The results suggest that longer utterances have better performances. However, since $T_{\text{min}}$ controls how often the language recognition is triggered, choosing a too long $T_{\text{min}}$ may cause significant performance degradation in detecting short OOL segments. Considering this trade-off, $T_{\text{min}} = 3$s is chosen, about 10–15 characters in Mandarin Chinese.

6.2. LD and LI approaches for OOL detection

In this experiment, we examine the OOL detection accuracies of different approaches on Dataset B. LD approaches with and without phone lattice reconstruction are evaluated using $T_{\text{min}} = 3$s. In LI approaches, word CMs are used. False alarm and miss probabilities are evaluated on word level. This procedure is similar to OOV detection, i.e. mis-recognized words overlapping with OOL segments are detected, and DET curves are shown in Figure 3. The best performance (EER 13.82%) is
also achieved by LD approach with phone lattice reconstruction (LD_RL), which significantly outperforms the others. The EER of LD approach without phone lattice reconstruction (LD_ORL), Wlat and Went are 21.41%, 31.10% and 33.95% respectively.

6.3. Combination of the LD and LI approaches

Based on experiments in Section 6.2, we combine the LD and the LI approaches using logistic regression (LR) as described in Section 4. CMs used in this combination include LD_RL, Wlat and Went. Due to the lack of annotated data for OOL detection, the logistic regression weights were trained using 500 Mandarin utterances and 500 English utterances selected from the VSM training data. DET curve of the combined CM (labeled as “LR fusion”) is plotted in Figure 3. The finally achieved EER is 12.68%, and performances on low false alarm operating points are improved.

6.4. LVCSR with OOL detection

Finally, we improve the baseline Mandarin LVCSR system using the combined LD and LI approaches (logistic regression fusion of LD_RL, Wlat and Went) for OOL detection. This experiment is conducted on Dataset B. The operating point of the OOL detector is simply selected on EER. Table 2 lists the LVCSR performance with and without OOL detection in form of Character Error Rate (CER). By incorporating OOL detection, the substitution and insertion errors (mostly introduced by recognizing English speech) are significantly reduced, but the deletion error is greatly increased due to the OOL detection error. Totally, the CER is reduced by 33.06%.

7. Conclusions

In this paper, we first propose a novel LVCSR decoding procedure which integrates lattice-VSM-based LRE ability. Then we combine this LD approach with lattice-derived LI CMs for OOL detection. Heuristic phone lattice reconstruction is employed to improve both the LRE and OOL detection performance. The LD and LI approaches reflect two different dimensions of uncertainties encoded in phone/word lattices. Evaluated on a Mandarin/English mixed conversational telephone speech corpus with a Mandarin LVCSR system, the logistic regression fusion of the LD and LI approaches achieves an EER of 12.68% in OOL detection, and reduces the LVCSR CER by 33.06%.

8. Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 90920302 and 61005019, by the NSFC and Research Grants Council (RGC) of Hong Kong under Grant No. 60931160443, and in part by the National High Technology Development Program of China (863 Program) under Grant No. 2008AA040201.

9. References