Direct Posterior Confidence for Out-of-Vocabulary Spoken Term Detection

Dong Wang
Eurecom
BP 193, F-06904
Sophia Antipolis, France
Dong.Wang@eurecom.fr

Simon King
CSTR, University of Edinburgh
10 Crichton Street, EH8 9AB
Edinburgh, UK
Simon.King@ed.ac.uk

Nicholas Evans
Eurecom
BP 193, F-06904
Sophia Antipolis, France
evans@eurecom.fr

Joe Frankel
CSTR, University of Edinburgh
10 Crichton Street, EH8 9AB
Edinburgh, UK
joe@cstr.ed.ac.uk

Raphaël Troncy
Eurecom
BP 193, F-06904
Sophia Antipolis, France
Raphael.Troncy@eurecom.fr

ABSTRACT

Spoken term detection (STD) is a fundamental task in spoken information retrieval. Compared to conventional speech transcription and keyword spotting, STD is an open-vocabulary task and is necessarily required to address out-of-vocabulary (OOV) terms. Approaches based on subword units, e.g. phonemes, are widely used to solve the OOV issue; however, performance on OOV terms is still significantly inferior to that for in-vocabulary (INV) terms.

The performance degradation on OOV terms can be attributed to a multitude of factors. A particular factor we address in this paper is that the acoustic and language models used for speech transcribing are highly vulnerable to OOV terms, which leads to unreliable confidence measures and error-prone detections.

A direct posterior confidence measure that is derived from discriminative models has been proposed for STD. In this paper, we utilize this technique to tackle the weakness of OOV terms in confidence estimation. Neither acoustic models nor language models being included in the computation, the new confidence avoids the weak modeling problem with OOV terms. Our experiments, set up on multi-party meeting speech which is highly spontaneous and conversational, demonstrate that the proposed technique improves STD performance on OOV terms significantly; when combined with conventional lattice-based confidence, a significant improvement in performance is obtained on both INV and OOVs. Furthermore, the new confidence measure technique can be combined together with other advanced techniques for OOV treatment, such as stochastic pronunciation modeling and term-dependent confidence discrimination, which leads to an integrated solution for OOV STD with greatly improved performance.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SSCS'10, October 29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-4503-0162-6/10/10 ...$10.00.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms
Algorithms, Experimentation

Keywords
spoken term detection, speech document search, spontaneous conversational speech, speech recognition

1. INTRODUCTION

The ever-increasing volume of speech material online creates the need for spoken information retrieval techniques. Spoken term detection (STD), as defined by NIST in 2006 [8], is a fundamental task associated with information retrieval from spoken documents. According to NIST, STD aims to provide for the searching of large quantities of audio without the need for reprocessing the audio signal every time a query is performed. The evaluation series organized by NIST have attracted broad interest, including [1, 2, 7, 9, 10].

A typical STD system consists of two components. First, an automatic speech recognition (ASR) component is used to transcribe speech signals into intermediate representations, usually word or subword lattices, and then a detection component searches for occurrences of search terms within the generated lattices. A key task of the search component is to accept reliable detections and reject unreliable ones, which requires an acceptable compromise between hits and false alarms (FAs). This is achieved according to confidence measures, among which the most popular is the lattice-based confidence derived from lattice posterior probabilities, denoted as \( c_{lat} \) in our work and given as follows:

\[
 c_{lat} = \frac{\sum p(O, K^s_{\alpha}, \pi_\alpha, K^s_{\beta}, \pi_\beta)P(\pi_\alpha, K^s_{\alpha})}{\sum \xi p(O|\xi)P(\xi)} \tag{1}
\]

where \( K^s_{\alpha} \) denotes the event that search term \( K \) appears in the speech segment from time \( t_s \) to time \( t_e \) in the audio stream \( O \). \( \pi_\alpha \) and \( \pi_\beta \) denote paths through the lattice...
before and after $K$, with $\pi_a$ starting from the beginning of the audio stream and $\pi_b$ running to the end. The summation in the numerator operates over all valid paths involving the search term $K$, whereas the denominator includes any valid path through the lattice, denoted by $\xi$. We note that $p(0|\pi_a, K^a_s, \pi_a)$ and $P(\pi_a, K^a_s, \pi_a)$ are acoustic model and language model scores respectively, and both have been retained in the lattice following the speech transcription.

A particular feature that discriminates STD from other tasks such as speech transcription and keyword spotting is that it is an open-vocabulary task, which means that queries may contain any words that are not limited to the system vocabulary. Typical examples include entity names or technical terms. STD systems must thus cope with queries which contain out-of-vocabulary (OOV) words. For example, in a query task addressed by the ACAV project\(^1\), queries may contain OOV entity names or technical terms, which can present a significant challenge. Search terms involving OOV words are known as OOV terms in STD; correspondingly, terms involving words only within the system vocabulary are in-vocabulary (INV) terms. As OOV terms usually convey important information, a good solution to OOV term detection is highly desirable for spoken document indexing and retrieval. The usual approach to detecting OOV terms is to use language model scores respectively, and both have been retained in the lattice following the speech transcription.

2. DIRECT POSTERIOR CONFIDENCE

2.1 Acoustic posterior confidence

It is well known that a standard 3-layer MLP network with softmax output activation can be used to estimate class posterior probabilities for a classification task. MLPs have been widely used in this fashion for speech recognition, by estimating the posterior probabilities for phone classes, given acoustic features as inputs \([6]\). The direct posterior probability technique proposed in \([13]\) uses an MLP to estimate the frame-wise posterior probability $P(Q_t|O)$, where $Q_t$ is the phone class of the search term $K$ corresponding to frame $t$. For example, if the term search obtains a partial path representing the term $K$, and its phone sequence indicates that frames $t_a$ to $t_b$ belong to a phone $Q$, $Q_{t_a}$ will be $Q$ for all $t$ from $t_a$ to $t_b$. Note that $Q_t$ is easily specified given the lattice.

A detection $d$ is denoted as follows:

$$d = (K, s = (t_s, t_e), v_a, v_t, ...) \tag{2}$$

where $K$ is the search term, $s$ defines the speech interval $t_s$ to $t_e$, during which the detection resides, and where $v_a$ and $v_t$ are the acoustic and language model scores respectively. Other informative factors that might be included are denoted by ‘...’. Using the frame-wise posterior probability $P(Q_t|O)$, the confidence in detection $d$, denoted $c_{MLP}(d)$, is calculated simply by averaging individual frame confidences as follows:

$$c_{MLP}(d) = \frac{1}{t_e - t_s} \prod_{t = t_s}^{t_e} P(Q_t|O) \tag{4}$$

where $W$ is the half-window length of the MLP input.

Instead of resorting to the Bayesian rule as in the lattice-based confidence estimation (Eq. 1), this gives a ‘direct’ posterior confidence since it is based on frame-wise posterior probabilities that are calculated directly from a discriminative model (MLP here). Note that the direct posterior confidence is derived purely from acoustic features at the phone level, and therefore we refer to it as the acoustic posterior confidence. Without considering any linguistic context, the acoustic posterior confidence is less impacted by the OOV issue than the lattice-based confidence. Finally, this new confidence estimate need not be necessarily based on an MLP, but on any discriminative model that evaluates the posterior probability locally.

2.2 LM posterior confidence

Obviously, an implicit assumption behind the derivation of the acoustic posterior confidence is that phone classes of any two frames are independent conditioned on acoustic features. This leads to a simple local confidence measure which, as we will see in Section 2.4, effectively removes the negative impact of linguistic contexts in the case of OOV term detection; however, it also means that some information from linguistic constraints is ignored. This information is potentially beneficial and it is thus of interest to assess its use in both INV and OOV term detection.

\(^1\)“Collaborative Annotation for Video Accessibility” (ACAV) is a project supported by the French Ministry of Industry (Innovative Web call) that aims to develop a collaborative annotation tool for the manual correction of automatically derived transcriptions and for the enriching of content with semantic metadata.
In order to get the information involved in LMs back and use it in a safe way, we consider the evidence that a ‘linguistic lattice’ provides to a putative detection. Similar to the lattice-based confidence estimation, we examine the posterior probability of the phoneme string of the search term given the lattice, but no acoustic scores are considered. This posterior probability represents the confidence we have for a detection when we observe the search term appearing within the phonemic context. Eq. 6-8 formulate this idea, where \( L \) denotes the entire phoneme lattice, \( K \) denotes the phoneme form of search term \( K \), and \( C_{K'} \) is the context of \( K' \).

\[
c_{lm}(d) = \frac{P(K' | L)}{P(L)} \tag{6}
\]

\[
c_{lm}(d) = \frac{P(K' | L)}{P(L)} = \frac{\sum_{C_{K'}} P(K', C_{K'})}{P(L)} \tag{7}
\]

where \( c_{lm} \) is denoted as the LM posterior confidence, given that \( P(K' | L) \) concerns linguistic constraints only.

### 2.3 Confidence integration

The acoustic and LM posterior confidences relate to different aspects of a detection and can thus be combined to improve accuracy. Assuming that the acoustic-based and language-based confidences are given by two independent tests, and if we also assume that at least one test signifies a positive detection, then the AM and LM confidences may be combined, or fused, as follows:

\[
c_{mlp+lat} = 1 - (1 - c_{mlp})^\alpha (1 - c_{lat}) \tag{9}
\]

where \( \alpha \) is a scale factor, and \( c_{mlp+lat} \) is the confidence, which integrates the acoustic posterior confidence \( (c_{mlp}) \) and LM posterior confidence \( (c_{lat}) \), given by Eq. 5 and Eq. 8 respectively. Note that the LM posterior confidence does not provide any information more than the LM does in the lattice-based confidence estimation; it is just a convenient form to fuse the acoustic posterior confidence.

The same approach can be used to combine the acoustic posterior confidence \( (c_{mlp}) \) and the lattice-based confidence \( (c_{lat}) \), by given Eq. 5 and Eq. 1 respectively. This gives rise to:

\[
c_{mlp+lat} = 1 - (1 - c_{mlp})^\alpha (1 - c_{lat}) \tag{10}
\]

where \( c_{mlp+lat} \) is again the combined confidence.

### 2.4 Experiments

Experiments were conducted using English language meeting speech recorded from individual headset microphones (IHMs) and using phoneme-based ASR and STD systems. Meeting speech is highly ‘conversational’ or ‘spontaneous’, which presents a significant challenge to ASR systems; moreover, meetings tend to involve many OOV terms.

To ensure the OOV terms in the experiment have similar properties to genuine novel terms that could be expected in a real application, we defined OOV terms strictly as: those containing no words listed in the dictionaries of the ASR system or of the term detector, and not appearing in the training material for either the acoustic or language models.

To create a list of OOV terms, we compared the AMI dictionary (recently created, in active use and so assumed to represent current usage) and the COMLEX Syntax dictionary v3.1 (published by LDC in 1996 and therefore historical from a STD perspective). We selected 412 terms from the AMI dictionary that do not occur in the COMLEX dictionary. We also added another 70 artificial OOV terms (which occur more frequently) that are plausible search terms. This results in 482 search terms having a total of 2736 occurrences in the evaluation data. These terms were removed from the system dictionaries; furthermore, all utterances and sentences that contain these terms were deleted from the speech and text training corpora. This ensures that they were entirely unseen during system training and tuning. In addition, 256 INV terms which are mostly person and city names were chosen to perform comparative study.

The AMs and LMs were trained on the same corpora used for training in the AMi2 RT04s ASR system [5]. After the OOV purge, there were 80.2 hours of speech for AM training and 521M words of text for LM training. The RT04s development dataset was used for development work. Evaluation work was performed with the RT04s and RT05s evaluation datasets in addition to a new meeting corpus recorded recently at the University of Edinburgh through the AMIDA project. This amounts to 11 hours of speech and there is no overlap between the data used for development and evaluation.

HTK was used to train acoustic models and conduct phone decoding; the SRI LM toolkit was used to train phoneme n-gram models. Term detection was implemented with the \textit{Lattice2Multigram} tool [9] provided by the Speech Processing Group at the Brno University of Technology. Pronunciations of OOV terms were predicted using a letter-to-sound (LTS) approach based on a joint-multigram model (JMM) [4, 11]. Term-dependent normalization was applied to improve decision quality, as described in [12].

### Results

Results are presented in Table 1 and 2 for INV and OOV terms respectively. In each case results are illustrated for the four different systems outlined above. We see that the system based on the acoustic posterior confidence \( c_{mlp} \) outperforms the baseline system that uses the lattice-based confidence \( c_{lat} \), for both INV terms and OOV terms. However the behavior on INV and OOV terms is rather different: for INV terms, a t-test shows that the improvement is not significant \((p = 0.2)\), and the max-ATWV actually decreases. This seems to indicate that the higher ATWV is unreliable. In contrast, for OOV terms, the improvement in ATWV is significant \((p < 0.01)\), and the max-ATWV also increases accordingly. This observation is consistent with our conjecture that the acoustic posterior confidence is more appealing to OOV terms for which the lattice-based confidence tends to be unreliable. Furthermore, when integrated with the LM posterior confidence \( c_{lm} \), there is no improvement with OOV terms, but significant improvement \((p < 10^{-5})\) with INV terms. This suggests that the LM constraint brings no benefit in the case of OOV terms, but that it is rather informative for INV terms. All of these observations sug-

Table 1: The performance of STD systems on INV terms with direct posterior confidence. \(c_{lat}\) denotes the lattice-based confidence, and \(c_{mlp}\) denotes the direct, acoustic posterior confidence. \(c_{mlp+lm}\) and \(c_{mlp+lat}\) are integrated confidence measures presented in Eq. 9 and Eq. 10 respectively. The best results are shown in bold face.

<table>
<thead>
<tr>
<th>Confidence</th>
<th>ATWV</th>
<th>max-ATWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c_{lat})</td>
<td>0.4743</td>
<td>0.5058</td>
</tr>
<tr>
<td>(c_{mlp})</td>
<td>0.3902</td>
<td>0.4994</td>
</tr>
<tr>
<td>(c_{mlp+lm})</td>
<td>0.4963</td>
<td>0.5022</td>
</tr>
<tr>
<td>(c_{mlp+lat})</td>
<td>0.5344</td>
<td>0.5363</td>
</tr>
</tbody>
</table>

Table 2: The performance of STD systems on OOV terms with direct posterior confidence. The notations are the same as in Table 1.

<table>
<thead>
<tr>
<th>Confidence</th>
<th>ATWV</th>
<th>max-ATWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c_{lat})</td>
<td>0.2764</td>
<td>0.2770</td>
</tr>
<tr>
<td>(c_{mlp})</td>
<td>0.2971</td>
<td>0.2986</td>
</tr>
<tr>
<td>(c_{mlp+lm})</td>
<td>0.2941</td>
<td>0.2950</td>
</tr>
<tr>
<td>(c_{mlp+lat})</td>
<td>0.2973</td>
<td>0.3011</td>
</tr>
</tbody>
</table>

Figure 1: DET curves for STD system performance on INV terms using various confidence measures.

Figure 2: DET curves for STD system performance on OOV terms using various confidence measures.

3. INTEGRATED SOLUTION FOR OOV TERM DETECTION

3.1 SPM and term-dependent confidence discrimination

Stochastic pronunciation modeling (SPM) [11] and term-dependent confidence discrimination [12] are two techniques...
to enhance OOV term detection. The SPM approach addresses the high degree of pronunciation variability which is typical with OOV terms, and term-dependent confidence discrimination copes with the high diversity of OOV terms with respect to linguistic properties. In the following we describe how the two techniques are combined with the direct posterior confidence to give a comprehensive solution for OOV term detection.

With the SPM approach [11], all possible pronunciations of a search term are predicted by a letter-to-sound (LTS) model, using for example a joint multigram model (JMM) as in [11]. Term detection is then applied using the full set of pronunciations. Letting \( Q \) denote one such pronunciation then a detection \( d \), based on this particular pronunciation, may be denoted by:

\[
d = (K, Q, s, v_a, v_1, \ldots),
\]

where all other symbols have the same meaning as in Eq. 2.

If we further define the probability of a pronunciation \( Q \) of term \( K \) as a pronunciation confidence \( c_{pron} \):

\[
c_{pron}(d) = P(Q|d,K)
\]

where \( K \) and \( Q \) is the detected pronunciation represented by the detection \( d \).

The confidence in the detection \( d \) is then determined according to some composite function of \( c_{lat} \) and \( c_{pron} \):

\[
c_{disc}(d) = f_{disc}(c_{lat}, c_{pron})
\]

where \( c_{disc} \) denotes confidence according to the SPM. In the original proposal [11], a linear composition was utilized.

In contrast, the term-dependent confidence discrimination technique [12] is based upon a unified discriminative confidence measure by integrating various informative factors using a certain discriminative model. This can be formally represented as:

\[
c_{disc}(d) = f_{disc}(c_{lat}, R_0, R_1, \ldots)
\]

where \( f_{disc} \) represents the discriminative model, which in our work is always a support vector machine (SVM). \( R_0 \) and \( R_1 \) are two occurrence-derived informative factors introduced in [12] and defined as

\[
R_0(K) = \frac{\sum c_{lat}(d^K)}{T}
\]

and

\[
R_1(K) = \frac{\sum (1 - c_{lat}(d^K))}{T}
\]

where \( T \) is the length of the audio stream, and \( d^K \) denotes the i-th detection of term \( K \).

### 3.2 Integrated solution

So far, three techniques have been proposed to deal with OOV STD: the direct posterior confidence measure, as presented for the first time here, in addition to SPM and term-dependent discrimination as originally reported in [11, 12]. Each of these techniques tackle the OOV challenge from a unique perspective and addresses different, particular properties of OOV terms. Additional gains in performance might thus be expected by combining these techniques into a integrated solution. The overall system is illustrated in Figure 3 and can be formulated as follows:

\[
c_{disc}(d) = f_{disc}(c_{lat}, c_{pron}, c_{mlp}, R_0, R_1, \ldots)
\]  (17)

Here, according to the SPM, all possible pronunciations are considered in the lattice search, and then each resulting putative detection is assigned a pronunciation confidence \( c_{pron} \) given by the LTS model, a lattice-based detection confidence \( c_{lat} \) given by the lattice search, and a direct posterior confidence \( c_{disc} \) given by the MLP-based phone posterior prediction. These three confidences, in addition to informative factors \( R_0 \) and \( R_1 \) are fed into the SVM-based discriminative confidence estimation function. The resulting discriminative confidence, after normalization, is employed to determine the final hit/FA decision.

### 3.3 Experiments

Experiments were conducted under the same conditions as described in Section 2.4, except that only OOV terms are considered here (SPM works only for OOV terms). A JMM was employed to implement SPM. For confidence discrimination, an SVM was trained with the LIBSVM toolkit from the National Taiwan University [3]. The readers should refer to the original papers [11, 12] for details regarding JMM and SVM training.

Results are illustrated in Table 3. It is clear that each step of the integrated solution contributes a significant improvement in performance \((p < 0.01)\), and that the final result is much better than that of the baseline system (0.33 cf. 0.28 ATWV).

<table>
<thead>
<tr>
<th>Method</th>
<th>ATWV</th>
<th>max-ATWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>0.2764</td>
<td>0.2770</td>
</tr>
<tr>
<td>+SPM</td>
<td>0.3183</td>
<td>0.3303</td>
</tr>
<tr>
<td>+conf. disc.</td>
<td>0.3245</td>
<td>0.3352</td>
</tr>
<tr>
<td>+direct post.</td>
<td>0.3318</td>
<td>0.3392</td>
</tr>
</tbody>
</table>

Table 3: The performance of OOV STD with the integrated solution. The baseline system used lattice-based confidence and single best pronunciation prediction. ‘conf. disc.’ denotes confidence discrimination, and ‘direct post.’ denotes direct posterior confidence estimation.

DET curves for the integrated system are shown in Figure 4. They show that the SPM approach provides the greatest
contribution to performance improvement: the DET curve not only falls in the region of lower FA, but also extends to the area of lower miss probability. This means that SPM not only improves detection accuracy, but also improves system potential, i.e. the maximum occurrences that the system can detect, by considering the variation in pronunciation. Confidence discrimination does not give much improvement, however it provides a way to integrate various informative factors including the direct posterior confidence. The integration of the three techniques results in the best performance across most of the operating region, but poorer performance than the SPM-only approach when the FA rate is low. This might be due to the insufficient amount of data that we used to train the SVM, which leads to unreliable estimation in the area of high precision.

4. CONCLUSIONS

This paper proposes the use of direct posterior confidence estimates that are derived from an MLP-based phone classifier to enhance OOV STD. Compared to the conventional lattice-based confidence estimates, the new confidence is a local measurement and is thus less vulnerable to context sparsity. It is therefore better suited to the detection of OOV terms which are usually inadequately represented by acoustic and language models. Our experiments, which were set up on meeting speech which is highly spontaneous and conversational, demonstrate that the direct posterior confidence is more beneficial for OOV terms than for INV terms, and is complementary to the lattice-based confidence. Moreover, results improve significantly when the new confidence measure is integrated with stochastic pronunciation modeling and confidence discrimination in a comprehensive solution for OOV term detection.

5. ACKNOWLEDGMENTS

This work was carried out while Dong Wang was a Fellow on the EdSST interdisciplinary Marie Curie training programme at CSTR, University of Edinburgh. This work used the Edinburgh Compute and Data Facility which is partially supported by eDIKT, and has been partially supported by the French Ministry of Industry (Innovative Web call) under contract 09.2.93.0966, “Collaborative Annotation for Video Accessibility” (ACAV).

6. REFERENCES