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

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

Dong Wang
Eurecom
BP 193, F-06904
Sophia Antiplis, 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 Antiplis, 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 Antiplis, 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 INVs 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.

SSCS'10, October 29, 2010, Firenze, Italy.

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:

$$c_{lat} = \frac{\sum_{\pi_\alpha,\pi_\beta} p(O|\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta) P(\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta)}{\sum_\xi p(O|\xi)P(\xi)} \quad (1)$$

where $K_{t_s}^{t_e}$ 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. π_α and π_β denote paths through the lattice

before and after K, with π_α starting from the beginning of the audio stream and π_β 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 ξ. We note that $p(O|\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta)$ and $P(\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta)$ 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 retrieving. The usual approach to detecting OOV terms is based on subword units, which searches subword lattices for subword representations of search terms that are obtained from letter-to-sound (LTS) conversion. Among various subword units, phonemes are the most simple and widely used.

Whilst the subword-based approach enables OOV term detection, the detection performance is always inferior to that of INV terms. One of the principle reasons, we hypothesize, is that the acoustic models (AMs) and language models (LMs) tend to represent OOV terms not very well due to the absence of OOV terms in training materials, which in turn leads to unreliable confidence estimation with Eq. 1.

A direct posterior confidence measure has been proposed recently [13] for STD. Instead of being derived from acoustic and language models, the new confidence measure reflects the posterior probability $P(K|O)$ 'directly' from a discriminative model, e.g. a multi-layer perceptron (MLP). In this paper, we propose to use this confidence measure to provide more reliable confidence estimation for OOV term detection. Being fully derived from acoustic properties at the phone level, the new confidence measure ameliorates the modelling weaknesses of OOV terms, thereby enhancing OOV STD performance. This approach to confidence estimation was first presented in [13]. The novelty of this paper is that, first we improve the LM fusion strategy; second we highlight the distinct behaviors of INV and OOV terms with the new confidence; third we combine this technique with stochastic pronunciation modeling (SPM) and the term-dependent confidence discrimination technique. These contributions lead to significant improvements in OOV term detection.

In the next section, we first briefly introduce the direct posterior confidence and then report comparative experiments with INV and OOV terms. The integrated solution and corresponding results are reported in Section 3. Finally we present our conclusions in Section 4.

[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.

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 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:

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

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_l 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:

$$\begin{aligned} c_{mlp}(d) &= \frac{P(K_{t_s}^{t_e}|O)}{t_e - t_s} && (3)\\ &= \frac{1}{t_e - t_s} \prod_{t=t_s}^{t_e} P(Q_t|O) && (4)\\ &= \frac{1}{t_e - t_s} \prod_{t=t_s}^{t_e} P(Q_t|o_{t-W}, ..., o_t, ...o_{t+W}) && (5) \end{aligned}$$

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.

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^l denotes the phoneme form of search term K, and C_{K^l} is the context of K^l.

$$c_{lm}(d) = P(K^l|L) \quad (6)$$

$$= \frac{P(K^l, L)}{P(L)} \quad (7)$$

$$= \frac{\sum_{C_{K^l}} P(K^l, C_{K^l})}{P(L)} \quad (8)$$

where c_{lm} is denoted as the *LM posterior confidence*, given that $P(K^l|L)$ 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:

$$c_{mlp+lm} = 1 - (1 - c_{mlp})^{\alpha}(1 - c_{lm}) \quad (9)$$

where α is a scale factor, and c_{mlp+lm} is the confidence, which integrates the acoustic posterior confidence (c_{mlp}) and LM posterior confidence (c_{lm}), 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}), given by Eq. 5 and Eq. 1 respectively. This gives rise to:

$$c_{mlp+lat} = 1 - (1 - c_{mlp})^{\alpha}(1 - c_{lat}) \quad (10)$$

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 AMI[2] RT05s 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 phoneme decoding; the SRI LM toolkit was used to train phoneme n-gram models. Term detection was implemented with the *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]. STD performance is reported in terms of average term-weighted value (ATWV) [8]; detection error trade-off (DET) curves are also used to show behavior at different hit/FA ratios. The best ATWV that can be obtained with an optimal threshold is denoted as *max-ATWV* [8]. All results reported here are those obtained on the evaluation set, with parameters (e.g. the threshold to make decisions) being tuned to optimize performance on the development set.

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 ($p < 1e^{-5}$) 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-

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

Confidence	ATWV	max-ATWV
c_{lat}	0.4743	0.5058
c_{mlp}	0.4902	0.4994
c_{mlp+lm}	0.4963	0.5022
$c_{mlp+lat}$	**0.5344**	0.5363

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.

Confidence	ATWV	max-ATWV
c_{lat}	0.2761	0.2770
c_{mlp}	0.2971	0.2986
c_{mlp+lm}	0.2941	0.2980
$c_{mlp+lat}$	**0.2973**	0.3011

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

gest that context information, which is captured in context-dependent models in acoustic modeling and n-gram models in language modeling, is not suited to OOV detection; OOV terms are detected more reliably with local confidence with less context interference, i.e. as with the acoustic posterior confidence.

Finally, when the lattice-based confidence and the direct posterior confidence are combined, significant improvements are obtained for both INV terms and OOV terms. For OOV terms, the improvement is marginal, supporting our conjecture that the lattice-based confidence is unreliable for OOV terms. For INV terms, the improvement is significant ($p < 0.01$), suggesting that the two confidence measures are both valuable and complementary.

The DET curves, shown in Figures 1 and 2 illustrate the differences in detection performance for INV and OOV terms respectively as the detection thresholds are varied. We first observe that the INV curves extend to a much lower miss probability (lower right side of the DET plot) than the OOV curves, indicating that much higher precision is obtained on INV terms than on OOV terms. Secondly, we see that the INV curves are almost linear while the OOV curves are concave. This means for OOV terms, it is rather difficult to get more hits by just allowing more false alarms, suggesting that performance of OOV STD is limited by inaccurate speech transcription.

Concentrating on the curves for INV terms (Figure 1), we see that the acoustic posterior confidence does not show better performance than lattice-based confidence, either with or without the LM posterior confidence. This shows that the lattice-based confidence is good enough for INV term detection and that the new confidence measure does not give any benefit. For OOV terms (Figure 2), however, we find that the acoustic posterior confidence performs significantly better than the lattice-based confidence, particularly in the area of low false alarms. When integrated with the LM posterior confidence, further gains are obtained, particularly in the low FA area. This is somewhat inconsistent with the ATWV results in Table 2, where the LM posterior confidence contributes very little. This might be due to the fact that the FA suppression is predominantly important in this operating area, so that the linguistic constraint, although noisy, is still beneficial. Nevertheless, the conclusions drawn from the DET profiles and the ATWV results are largely consistent: the direct posterior confidence is much more effective than the lattice-based confidence for OOV term detection, and the combination of the two confidences delivers even better performance.

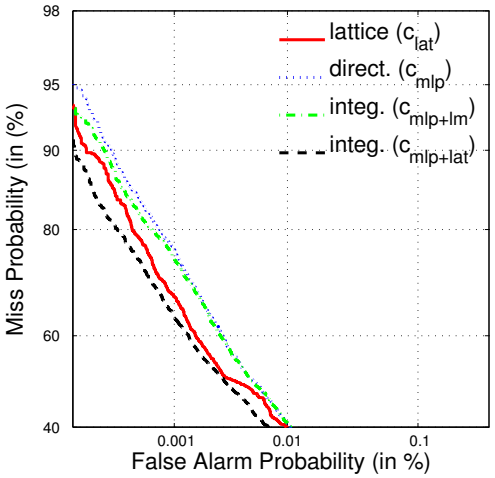

Figure 1: DET curves for STD system performance on INV terms using various confidence measures.

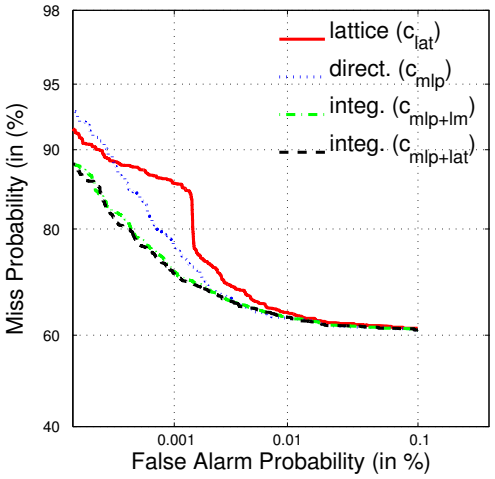

Figure 2: DET curves for STD system performance on OOV terms using various confidence measures.

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:

$$d = (K, Q, s, v_a, v_l, ...). \tag{11}$$

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}:

$$c_{pron}(d) = P(Q_d|K_d) \tag{12}$$

where K_d is the search term and Q_d 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}:

$$c_{spm}(d) = f(c_{lat}, c_{pron}) \tag{13}$$

where c_{spm} 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:

$$c_{svm}^{disc}(d) = f_{svm}(c_{lat}, R_0, R_1, ...) \tag{14}$$

where f_{svm} 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

$$R_0(K) = \frac{\sum_i c_{lat}(d_i^K)}{T} \tag{15}$$

and

$$R_1(K) = \frac{\sum_i (1 - c_{lat}(d_i^K))}{T} \tag{16}$$

where T is the length of the audio stream, and d_i^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:

$$c_{svm}^{disc}(d) = f_{svm}(c_{lat}, c_{pron}, c_{mlp}, R_0, R_1, ...). \tag{17}$$

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_{pron} give 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.

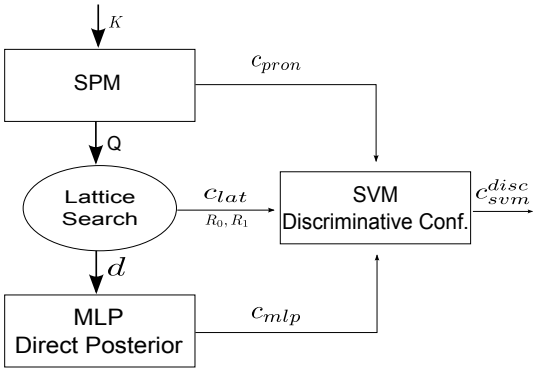

Figure 3: An illustration of OOV term detection with SPM, confidence discrimination and direct posterior confidence estimation.

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 disrimination, 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).

	ATWV	max-ATWV
baseline	0.2761	0.2770
+SPM	0.3153	0.3303
+conf. disc.	0.3235	0.3352
+direct post.	**0.3318**	0.3502

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.

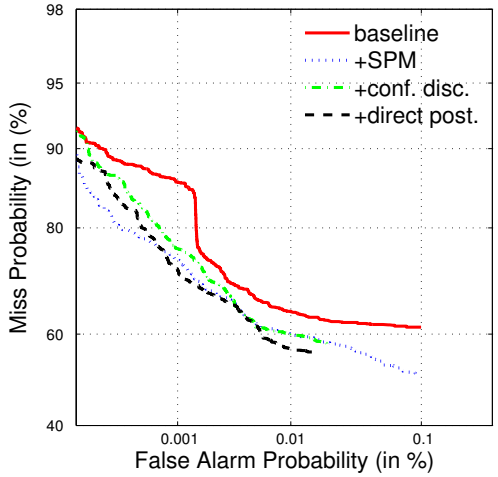

Figure 4: DET curves for STD systems on OOV terms with the integrated solution.

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

[1] M. Akbacak, D. Vergyri, and A. Stolcke. Open-vocabulary spoken term detection using graphone-based hybrid recognition systems. In *Proc. ICASSP'08*, pages 5240–5243, Las Vegas, Nevada, USA, March 2008.

[2] D. Can, E. Cooper, A. Sethy, C. White, B. Ramabhadran, and M. Saraclar. Effect of pronunciations on OOV queries in spoken term detection. In *Proc. ICASSP'09*, pages 3957–3960, Taipei, Taiwan, April 2009.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: A library for support vector machines*, 2001.

[4] S. Deligne, F. Yvon, and F. Bimbot. Variable-length sequence matching for phonetic transcription using joint multigrams. In *Proc. Eurospeech'95*, pages 2243–2246, Madrid, Spain, September 1995.

[5] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan. The AMI meeting transcription system: Progress and performance. In *Machine Learning for Multimodal Interaction*, volume 4299/2006, pages 419–431. Springer Berlin/Heidelberg, 2006.

[6] H. Hermansky, D. P. Ellis, and S. Sharma. Tandem connectionist feature extraction for conventional HMM systems. In *Proc. ICASSP'00*, pages 1635–1638, Istanbul, Turkey, June 2000.

[7] J. Mamou and B. Ramabhadran. Phonetic query expansion for spoken document retrieval. In *Proc. Interspeech'08*, pages 2106–2109, Brisbane, Australia, September 2008.

[8] NIST. *The spoken term detection (STD) 2006 evaluation plan*. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 10 edition, September 2006.

[9] I. Szöke, M. Fapšo, L. Burget, and J. Černocký. Hybrid word-subword decoding for spoken term detection. In *Proc. Speech search workshop at SIGIR (SSCS'08)*, Singapore, 2008. Association for Computing Machinery.

[10] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, and W. Wang. The SRI/OGI 2006 spoken term detection system. In *Proc. Interspeech'07*, pages 2393–2396, Antwerp, Belgium, August 2007.

[11] D. Wang, S. King, and J. Frankel. Stochastic pronunciation modelling for spoken term detection. In *Proc. Interspeech'09*, pages 2135–2138, Brighton, UK, September 2009.

[12] D. Wang, S. King, J. Frankel, and P. Bell. Term-dependent confidence for out-of-vocabulary term detection. In *Proc. Interspeech'09*, pages 2139–2142, Brighton, UK, September 2009.

[13] D. Wang, J. Tejedor, J. Frankel, and S. King. Posterior-based confidence measures for spoken term detection. In *Proc. ICASSP'09*, pages 4889–4892, Taiwan, April 2009.