Term-Dependent Confidence for Out-of-Vocabulary Term Detection

Dong Wang, Simon King, Joe Frankel, Peter Bell

The Centre for Speech Technology Research,
University of Edinburgh, UK
dwang2@inf.ed.ac.uk, Simon.King@ed.ac.uk, joe@cstr.ed.ac.uk, Peter.Bell@ed.ac.uk

Abstract

Within a spoken term detection (STD) system, the decision maker plays an important role in retrieving reliable detections. Most of the state-of-the-art STD systems make decisions based on a confidence measure that is term-independent, which poses a serious problem for out-of-vocabulary (OOV) term detection. In this paper, we study a term-dependent confidence measure based on confidence normalisation and discriminative modelling, particularly focusing on its remarkable effectiveness for detecting OOV terms. Experimental results indicate that the term-dependent confidence provides much more significant improvement for OOV terms than terms in-vocabulary.

Index Terms: confidence estimation, spoken term detection, speech recognition

1. Introduction

Spoken term detection (STD) is the task of automatically retrieving specified terms from speech data. To coordinate the research and boost the technology, NIST has run an evaluation series since 2006 [1], fostering a multitude of leading work and practical systems, including [2]–[10]. A typical STD system comprises two subsystems: an ASR subsystem for lattice generation and a STD subsystem for term detection, as illustrated in Figure 1.

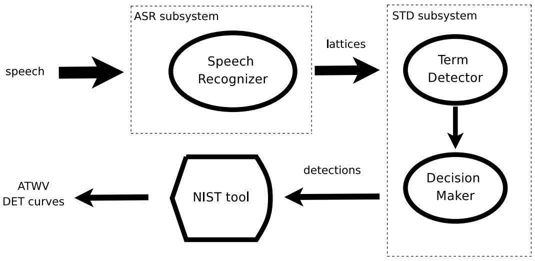

Figure 1: *The standard STD architecture: a speech recogniser converts speech signals to lattices; a term detector searches these lattices for putative occurrences of the search terms; a decision maker ascertains whether each putative detection is reliable. The NIST tool is applied to evaluation system performance, in terms of ATWV and DET curves.*

In a STD system, the *decision maker* plays an important role. It determines if a putative detection from the term detector is reliable enough to be a hit or should be discarded as a false alarm (FA), based on certain confidence measures. The lattice-based confidence derived from detection posterior probabilities is the most popular confidence measure and has been adopted by researchers from BBN [3], IBM [8], BUT [2], SRI&OGI [6], Tsinghua & Microsoft Research Asia [10]. Other confidence measures such as MED [5] and Local Likelihood Ratio [4] have also been studied.

A major problem of the above confidence measures is that they are all term-independent, that is to say, if two detections are assigned the same confidence, they are regarded the same for decision making purposes, no matter how much different properties of the detected terms hold in reality. This term-independence poses a serious problem for out-of-vocabulary (OOV) terms which usually exhibit high diversity in ASR error pattern, pronunciation variation, occurrence rate, confidence distribution, etc. Considering the fundamental importance of OOV terms for a practical STD system, herein we study the term-dependent confidence, analyse how it works, and investigate how much it contributes to OOV term detection.

The first part of our method is a term-dependent normalisation technique based on the ATWV-oriented decision proposed by BBN [3] and adopted by some researchers (e.g., [6],[9]), though we implement the technique from an alternative perspective. The second part of the method relies on discriminative modelling, initially presented for utterance verification [11],[12], which not only leads to a minimum decision cost, but also amends the flaws of the normalisation technique with other confidence. We notice that SRI&OGI presented a similar approach, whereas we focus more on OOV terms, and utilise alternative discriminative models (MLP and SVM) to test the hypothesis. In addition, we design occurrence-derived features to represent term-dependence for discriminative confidence.

In the rest of the paper, we first describe the confidence normalisation and analyse its discriminative power for OOV terms, and then we present the discriminative confidence. Experiments will be presented in Section 4, and some conclusions in Section 5.

2. Confidence normalisation

We denote a detection d of a search term K as a tuple:

$$d = (K, s = (t_1, t_2), c_a, c_l, ...) \qquad (1)$$

where s represents the speech segment from t_1 to t_2 within which the detection resides, and c_a and c_l are acoustic likelihood and language model score respectively. Any other informative factors can be included, as denoted by '. . .'.

Denote the confidence of d as $c(d)$, the hard-decision (e.g., [13]) can be formally written as an assertion function as Equation 2:

$$assert(d) = \begin{Bmatrix} 1 & if\ c(d) >= \theta \\ 0 & if\ c(d) < \ \ \theta \end{Bmatrix} \qquad (2)$$

where θ is a pragmatic threshold obtained by parameter tuning on the development set.

An obvious shortcoming of this decision strategy is that θ is term-independent, which raises problems for OOV term detection. OOV terms hold much diversity in the distribution of the confidence, so that the same confidence value might represent quite different confidence for different terms. Furthermore, this decision does not consider the evaluation metric, say, ATWV.

To improve the decision quality, we resort to the definition of ATWV defined by NIST [1]:

$$ATWV = \sum_{K} \frac{N_{hit}^{K}}{N_{true}^{K}} - \beta \frac{N_{FA}^{K}}{T - N_{true}^{K}} \quad (3)$$

where N_{hit}^{K} and N_{FA}^{K} are the number of hits and false alarms of search term K respectively, and N_{true}^{K} is the number of real occurrences of K. T denotes the audio length and β is a weight factor. This definition indicates that if a putative detection is a hit, it will provide benefit $\frac{1}{N_{true}^{K}}$, and if it is a false alarm, it will introduce a cost $\frac{\beta}{T-N_{true}^{K}}$, therefore the expected benefit of the putative detection d is

$$\zeta(d) = \frac{c(d)}{N_{true}^{K}} - \beta \frac{1 - c(d)}{T - N_{true}^{K}} \quad (4)$$

Considering that any putative detection with positive expected benefit will increase the final ATWV, we get the ATWV-oriented decision strategy:

$$assert(d) = \left\{ \begin{array}{ll} 1 & if\ \zeta(d) >= 0 \\ 0 & if\ \zeta(d) < \ \ 0 \end{array} \right\} \quad (5)$$

Note that N_{true}^{K} is unknown when performing the evaluation, and therefore must be estimated from the effective occurrence of K:

$$N_{true}^{K} \approx \sum_{i} c(d_{i}^{K}) \quad (6)$$

where d_{i}^{K} is the i-th detection of K.

Equation 4 can be regarded as a normalisation on confidence $c(d)$, denoted as ζ_K:

$$\zeta_K(c(d)) = \zeta(d) \quad (7)$$

Obviously, ζ_K is term-dependent, and the decision strategy of Equation 5 is correspondingly a term-dependent decision.

Now we apply the normalisation to the widely used lattice-based confidence, which, denoted as c_f, is formulated as the posterior probability that a search term K occurs in the speech segment from t_1 to t_2 as $K_{t_1}^{t_2}$:

$$\begin{aligned} c_f(d) &= p(K_{t_1}^{t_2}|O) & (8) \\ &= \frac{\sum_{C_K} p(O|C_K, K_{t_1}^{t_2}) p(C_K, K_{t_1}^{t_2})}{\sum_{\xi} p(O|\xi)p(\xi)} & (9) \end{aligned}$$

where $K_{t_1}^{t_2}$ denotes the event that K occurs between t_1 and t_2 of the input speech O, C_K is the context of K, and ξ is any path in the lattice. The normalised lattice-based confidence is given by Equation 10.

$$\zeta_K(c_f(d)) = \frac{c_f(d) \times \alpha + \gamma}{\sum_i c_f(d_i^K)} - \beta \frac{1 - c_f(d) \times \alpha - \gamma}{T - \sum_i c_f(d_i^K)} \quad (10)$$

where we have introduced a linear transform of $c_f(d)$ with two adaptable parameters α and γ, to compensate for any bias.

To investigate the contribution of the normalisation to discriminative power, we conducted the STD experiment on a development set, and plotted the confidence distribution of hits and false alarms before and after normalisation. The result, shown in Figure 2, confirms that the normalisation substantially improves the discrimination between correct and false detections. Detailed experimental settings will be presented in Section 4.

Note that the basic idea was proposed by [3]; we reformulate it here as a confidence transform, which gives greater flexibility in designing variant confidence, for example, by the introduction of the linear transform in Equation 10.

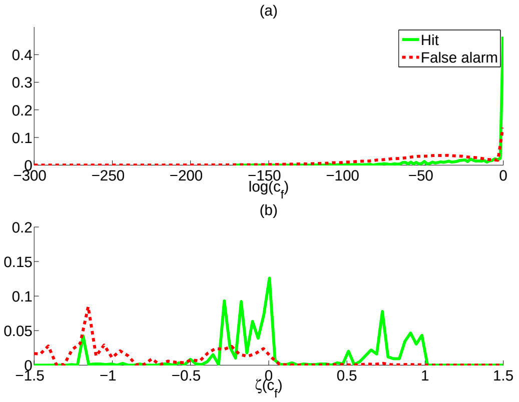

Figure 2: *The class conditional distribution of lattice-based confidence before (a) and after (b) normalisation.*

3. Discriminative confidence

A potential problem of the confidence normalisation ζ is that the confidence $c(d)$ might be not suitable for use in Equations 4 and 6. According to the decision theory, the only valid choice of $c(d)$ is the posterior probability of hit/FA classification. Letting C_{hit} denote the event that a detection is a hit, we define the classification posterior-based confidence $c_p(d)$ as

$$c_p(d) = p(C_{hit}|d) \quad (11)$$

Obviously $c_p(d)$ leads to a minimum decision cost for the hit/FA decision, and therefore is a *discriminative* confidence measure.

We derive $c_p(d)$ by constructing a mapping of the lattice-based confidence $c_f(d)$. We construct either a short mapping g:

$$g : c_f(d) \longrightarrow c_p(d) \quad (12)$$

or incorporate more informative factors using a long mapping f:

$$f : (c_f(d), c_0, c_1, ...) \longrightarrow c_p(d) \quad (13)$$

where c_0, c_1,... denote informative factors.

Although any informative factor can be taken in the long mapping, term-dependent factors are more preferable. We designed two occurrence-derived attributes to represent the term-dependence: effective occurrence rate, $R_0(K)$, and effective false alarm rate, $R_1(K)$, defined as the following:

$$R_0(K) = \frac{\sum_i c_f(d_i^K)}{T} \quad (14)$$

and

$$R_1(K) = \frac{\sum_i (1 - c_f(d_i^K))}{T} \quad (15)$$

To investigate how R_0 and R_1 improve the discrimination, we conducted the STD experiment on the development set with both in-vocabulary (INV) terms and OOV terms, and represent each detection as a dot in the coordinate $c_f \times R_0$ and $c_f \times R_1$, as shown in Figure 3. The interesting observation is that for OOV terms, both R_0 and R_1 improved discrimination, while for INV terms, their contribution is rather marginal.

We employed two alternative discriminative methods to construct g and f: a multiple layer perceptron (MLP) [11] and a support vector machine (SVM) [12]. Both of them can estimate $p(C_{hit}|d)$ with unlimited accuracy given sufficient training data. Figure 4 illustrates the distribution of the lattice-based confidence and the MLP-based discriminative confidence, before and after normalisation. We see clearly that the normalised discriminative confidence has a greater discriminative power.

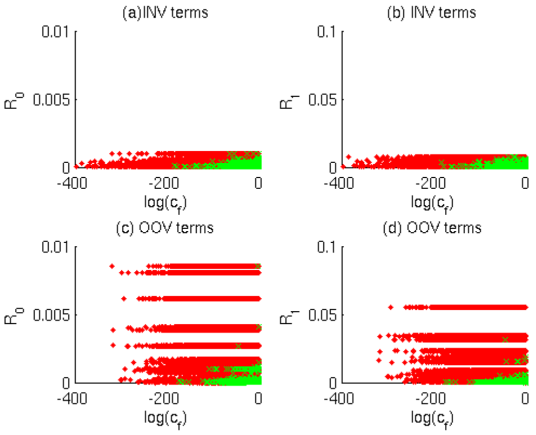

Figure 3: *The discriminative power of R_0 and R_1. A red point represents a false alarm, and a green cross represents a hit. The two plots (a)(b) show the INV terms,and the other two show the OOV terms.*

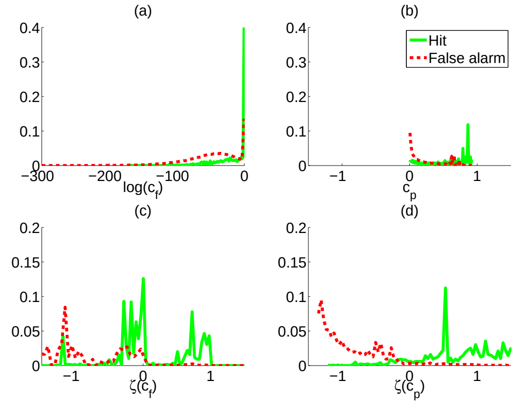

Figure 4: *The class-conditional distribution of lattice-based confidence (left) and MLP-based discriminative confidence (right). The top two plots (a)(b) present the distributions before normalisation, and the bottom plots (c)(d) present the distributions after normalisation.*

4. Experiments

We conducted experiments on speech from the meeting domain, recorded using individual headset microphones (IHM). The same training speech and text data used for building the AMI RT05s LVCSR system [14] were used to train the acoustic models (AM) and language models (LM). The NIST RT04s dev set was used for parameter tuning, and the evaluation corpora comprised three sub-sets: the NIST RT04s and RT05s eval set, and a new meeting corpus recorded at the University of Edinburgh as part of the AMIDA project.

We first selected 256 terms from the AMI dictionary as INV terms, which are all content terms and have 2329 occurrences in the evaluation data. Then we compared the AMI dictionary (in active use and assumed to represent current usage) and the COMLEX Syntax dictionary v3.1 (published by LDC in 1996 and therefore historical from a STD perspective), and selected 412 terms as OOV terms from the AMI dictionary that do not occur in the COMLEX dictionary. These terms simulate the evolution of novel terms over time. Additionally, we selected 70 *artificial OOV terms* that have more occurrences and plausible as search terms. In total we have 482 OOV terms and 2736 occurrences in the evaluation data. To ensure the OOV terms in the experiment represent truly novel terms, we purged all of them from the training speech and text.

		ATWV	
System	Conf. Norm.	INV terms	OOV terms
Word	NO	0.5661	-
Word	YES	0.5678	-
Phoneme	NO	0.4173	0.0273
Phoneme	YES	0.4743	0.2761

Table 1: *STD performance of the word and phoneme based systems with and without confidence normalisation. Results are reported for both INV terms and OOV terms, in terms of ATWV.*

We built a word-based system and a phoneme-based system. Both systems shared the same state-clustered triphone models, using 39-dimensional MFCC features. A 3-gram word LM was used for the word-based system and a 6-gram phoneme LM was used for the phoneme-based system. Cambridge University's HTK was used to train acoustic models and perform lattice generation, and the SRI LM toolkit was used to train LMs. An enhanced Joint-Multigram model [15] trained with the AMI dictionary was applied to predict pronunciations for the OOV terms.

4.1. Confidence normalisation

We first examine the performance with the confidence normalisation. The results are reported in Table 1, and the DET curves are shown in Figure 5. We find that the confidence normalisation is much more helpful for the phoneme-based system than the word-based system, which may be because word-based systems have many fewer false alarms so that the term-dependent decision, which aims at FA control, can contribute little. The second observation is that the normalisation is much more effective for OOV terms than INV terms. This supports our hypothesis that OOV terms have more diverse properties, therefore requiring normalisation.

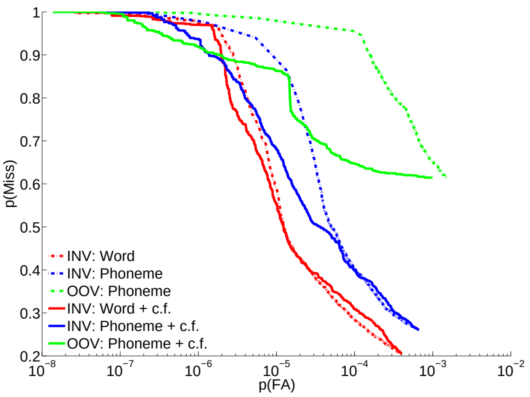

Figure 5: *The DET curves of the word and phoneme -based systems with and without confidence normalisation. Results are reported for both INV terms and OOV terms. $N.B.$ The legend denotes confidence normalisation, and the x-axis has a log scale.*

4.2. Discriminative confidence

We trained a MLP and a SVM to estimate the discriminative confidence. STD experiments were first conducted on the development set, and then detections were collected with hits as false alarms labelled, which were employed to train the MLP and SVM. In the case of short mapping, the input attribute of each training example was just the lattice-based confidence c_f,

		ATWV	
Confidence	Mapping	INV terms	OOV terms
Lattice-based	-	0.4743	0.2761
Discriminative (MLP)	SHORT	0.5453	0.2927
Discriminative (SVM)	SHORT	0.5432	0.2892
Discriminative (MLP)	LONG	**0.5460**	**0.2931**
Discriminative (SVM)	LONG	0.5421	0.2914

Table 2: *STD performance of the phoneme-based STD system with discriminative confidence estimated by the MLP and SVM.*

while in long mapping, R_0 and R_1 were incorporated. To account for the imbalance between positive and negative training examples, we first trained balanced models with equal numbers of hits and false alarm detections, and then adjusted the output using class prior probabilities.

Table 2 shows the experimental results and Figure 6 shows the DET curves. For simplicity, we just report the phoneme-based system (the word-based was impacted very little), and only MLP-based systems are presented in Figure 6 (the SVM-based system exhibited the same behaviour). We observe that the discriminative confidence, whether estimated by the MLP or the SVM, substantially improved system performance, especially for OOV terms. Moreover, significant improvement came from the short mapping, whilst the marginal contribution from the long mapping was relatively small. A pairwise t-test shows that the improvement achieved by the short mapping over the lattice-based confidence is significant ($p < 0.01$), no matter which discriminative model is used; while the additional improvement achieved by the long mapping over the short mapping is weakly significant ($p < 0.05$). Finally, although the best ATWV values were achieved with the MLP-based confidence, SVM-based confidence exhibited higher significance level.

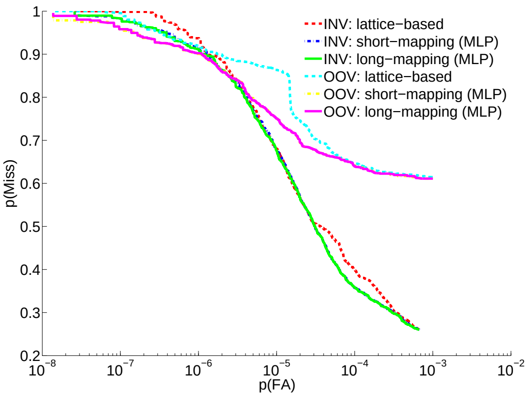

Figure 6: *The DET curves of the phoneme-based systems with MLP-based discriminative confidence. Results are reported on both INV terms and OOV terms, and the x-axis has a log scale.*

5. Conclusions

This paper investigated the effectiveness of term-dependent confidence on OOV terms. Experimental results indicate that with the term-dependent confidence normalisation, model-based discriminative confidence provides significant performance improvement over the lattice-based confidence, no matter which discriminative model is used. This improvement is much more greater when phoneme-based systems are used to detect OOV terms.

6. Acknowledgements

DW is a Fellow of the EdSST Marie Curie training programme. SK is an EPSRC Adv. Res. Fellow. This work used the Edinburgh Compute and Data Facility which is partially supported by eDIKT.

7. References

[1] NIST, *The spoken term detection (STD) 2006 evaluation plan*, 10th ed., National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, September 2006. [Online]. Available: http://www.nist.gov/speech/tests/std

[2] I. Szoke, M. Fapso, M. Karafiat, L. Burget, F. Grezl, P. Schwarz, Ondrejlembek, P. Matejka, S. Kontar, and J. Cernocky, "BUT system for NIST STD 2006 - English," in *Proc. NIST Spoken Term Detection Evaluation workshop (STD'06)*. Washington D.C., US: National Institute of Standards and Technology, 2006.

[3] D. R. H. Miller, M. Kleber, C. lin Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Proc. Interspeech'07*, Antwerp, Belgium, August 2007, pp. 314–317.

[4] K. Iwata, K. Shinoda, and S. Furui, "Robust spoken term detection using combination of phone-based and word-based recognition," in *Proc. Interspeech'08*, Brisbane, Australia, 2008.

[5] R. Wallace, R. Vogt, and S. Sridharan, "A phonetic search approach to the 2006 NIST spoken term detection evaluation," in *Proc. Interspeech'07*, Antwerp, Belgium, 2007, pp. 2393–2396.

[6] 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*, Antwerp, Belgium, 2007, pp. 2393–2396.

[7] I. Szoke, L. Burget, J. Cernocky, and M. Fapso, "Sub-word modeling of out of vocabulary words in spoken term detection," in *Proc. IEEE Workshop on Spoken Language Technology (SLT'08)*, Goa, Inida, 2008.

[8] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proc. ACM-SIGIR'07*, Amsterdam, July 2007, pp. 615–622.

[9] S. Parlak and M. Saraclar, "Spoken term detection for Turkish broadcast news," in *Proc. ICASSP'08*, Los Angels, US, April 2008.

[10] S. Meng, P. Yu, J. Liu, , and F. Seide, "Fusing multiple systems into a compact lattice index for Chinese spoken term detection," in *Proc. ICASSP'08*, Los Angels, US, April 2008.

[11] L. Mathan and L. Miclet, "Rejection of extraneous input in speech recognition applications using multi-layer perceptrons and the trace of HMMS," in *Proc. ICASSP'91*, vol. 1, Toronto, Canada, May 1991, pp. 93–96.

[12] R. Zhang and A. I. Rudnicky, "Word level confidence annotation using combinations of features," in *Proc. Eurospeech'01*, Aalborg, Denmark, September 2001, pp. 2105–2108.

[13] O. Siohan, B. Ramabhadran, and J. Mamou, "The IBM 2006 spoken term detection system," in *Proc. NIST Spoken Term Detection Evaluation workshop (STD'06)*. Washington D.C., US: National Institute of Standards and Technology, 2006.

[14] 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*. Springer Berlin/Heidelberg, 2006, vol. 4299/2006, pp. 419–431.

[15] S. Deligne, F. Yvon, and F. Bimbot, "Variable length sequence matching for phonetic transcription using joint multigrams," in *Proc. Eurospeech'95*, Madrid, 1995, pp. 2243–2246.