

STOCHASTIC PRONUNCIATION MODELLING AND SOFT MATCH FOR OUT-OF-VOCABULARY SPOKEN TERM DETECTION

Dong Wang, Simon King, Joe Frankel, Peter Bell

The Centre for Speech Technology Research,
University of Edinburgh, UK

ABSTRACT

A major challenge faced by a spoken term detection (STD) system is the detection of out-of-vocabulary (OOV) terms. Although a subword-based STD system is able to detect OOV terms, performance reduction is always observed compared to in-vocabulary terms. One challenge that OOV terms bring to STD is the pronunciation uncertainty. A commonly used approach to address this problem is a soft matching procedure, and the other is the stochastic pronunciation modelling (SPM) proposed by the authors. In this paper we compare these two approaches, and combine them using a discriminative decision strategy. Experimental results demonstrated that SPM and soft match are highly complementary, and their combination gives significant performance improvement to OOV term detection.

Index Terms— stochastic pronunciation modelling, soft match, confidence estimation, spoken term detection, speech recognition

1. INTRODUCTION

Spoken term detection (STD), as defined by NIST [1], involves the search of large, heterogeneous audio archives for occurrences of spoken terms. Partly due to the evaluation series run by NIST, STD is receiving much interest. A typical STD system comprises an ASR subsystem for lattice generation and a STD subsystem for term detection, as illustrated in Figure 1. State-of-the-art STD systems include those reported in [2, 3, 4, 5, 6, 7].

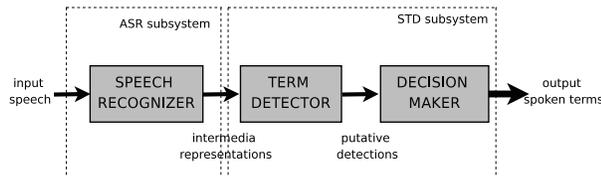


Fig. 1. The standard STD architecture: a speech recogniser converts speech signals to an intermediate representation (e.g., phoneme lattices); a term detector searches this representation for putative occurrences of the search terms; a decision maker ascertains whether each putative detection is reliable.

STD systems have difficulty in detecting out-of-vocabulary (OOV) terms. It is estimated that 20,000 new English words are coined each year: 50 per day [8]. These novel words and terms cause OOV challenges for STD systems in three

aspects: uncertainty in pronunciations, high diversity in properties, and weakness in acoustic and language modelling; ironically, these ‘challenging terms’ are likely search terms in practice. If a STD system is unable to handle OOV terms well, it will be less useful to end users, no matter how well it works on in-vocabulary terms.

Typically, a phoneme-based system is used to handle OOV terms, e.g., [9, 4]. In this approach, search terms are converted to pronunciations by letter-to-sound (LTS) models, and the pronunciations are searched for in a phoneme lattice generated by a speech recogniser. We take this approach in the work reported here.

In a previous study [10], we proposed a stochastic pronunciation model (SPM) which makes use of multiple pronunciations of OOV terms predicted by a joint-multigram model [11], thus compensating for the pronunciation uncertainty when unfamiliar words are spoken. Another approach to treating the pronunciation uncertainty is using a “soft match”, which allows some mismatch in lattice search. Both SPM and soft match have shown significant performance improvement for OOV term detection; however their respective properties and relative advantages have not been extensively studied.

Another issue with the uncertainty treatment approaches, both SPM and soft match, is that they usually suffer from a flood of false alarms (FA), which tends to hurt the overall performance, thus limiting the application of these approaches. In our previous work [12], we presented a term-dependent discriminative decision strategy which employs discriminative models to integrate various decision factors, especially term-dependent factors, into a discriminative confidence measure, leading to a decision strategy that causes minimum decision errors. In this paper, we borrow the discriminative power of this decision strategy to control the overwhelming false alarms caused by SPM and soft match, and combine these two approaches to gain further improvement.

In the rest of the paper, we first describe the configurations of our experiments, and then compare the SPM and soft match. In Section 4, we apply the discriminative decision to SPM and soft match, and combine them using discriminative confidence measures.

2. EXPERIMENTAL CONFIGURATIONS

We conducted experiments on meeting speech in the condition of individual headset microphones (IHM), and focused on OOV terms in English, using phoneme-based ASR and STD systems.

To ensure the OOV terms in the experiment represent truly novel terms, we defined OOV terms strictly as those containing no words existing in the dictionaries of the ASR system and the term detector and not appearing in training material

for acoustic or language models. In order to simulate real cases of newly-coined terms, 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). We selected 412 terms from the AMI dictionary that do not occur in the COMLEX dictionary. We also chose another 70 *artificial* OOV terms that occur more frequently and are plausible search terms. This results in 482 search terms having a total of 2736 occurrences in the evaluation data. We purged these terms from the system dictionary and all training speech and text data.

We trained acoustic models (AM) and language models (LM) on the corpora used by the AMI RT05s system [13]. After OOV term purging, there were 80.2 hours of speech for AM training and 521M words of text for LM training. The development set was the RT04s dev set, which contains 67 OOV terms for system development; the evaluation set consisted of the RT04s and RT05s eval sets and a new meeting corpus recorded recently at the University of Edinburgh in the AMIDA project, amounting to 11 hours of speech, containing all the OOV terms.

39-dim MFCC features were used with cepstral mean and variance normalisation (CMN + CVN); 3-state triphone HMMs and 6-gram phoneme LMs were employed. HTK was used to train acoustic models and conduct phoneme decoding; the SRI LM toolkit was used to train grapheme and phoneme n-gram models. The term detector was implemented with *Lattice2Multigram* provided by the Speech Processing Group, FIT, Brno University of Technology. Word-dependent thresholds were applied to improve decision quality [3, 14]. STD performance is reported in terms of ATWV [1]. Detection Error Tradeoff (DET) curves are used to show behaviour at different hit/FA ratios.

3. STOCHASTIC PRONUNCIATION MODELLING AND SOFT MATCH

3.1. Stochastic pronunciation modelling (SPM)

The basic idea of SPM [10] is to employ multiple pronunciations predicted by a joint-multigram model in OOV term detection. For a clear description, we first define a detection d as a tuple

$$d = (K, Q, s = (t_1, t_2), v_a, v_l, \dots) \quad (1)$$

where K denotes the search term, Q denotes its pronunciation, and s represents the speech segment from t_1 to t_2 within which the detection resides. v_a and v_l are the acoustic likelihood and language model score respectively. Other informative factors could be included in d , as denoted by "...".

Then the confidence score of a detection d is written as Equation 2

$$c_{fp}(d) = (1 - \gamma)c_f(d) + \gamma c_p(d) \quad (2)$$

where γ is an interpolation factor optimised with the development set, $c_p(d)$ is a *pronunciation confidence*, calculated as a posterior probability of the pronunciation when predicted by the joint-multigram model,

$$c_p(d) = p(Q|K) \quad (3)$$

and $c_f(d)$ is a lattice-based confidence given by Equation 4,

$$\begin{aligned} c_f(d) &= p(K_{t_1}^{t_2}, Q(d)|O) \\ &= \frac{\sum_{\zeta_K} p(O|\zeta_K, K_{t_1}^{t_2}, Q(d))p(\zeta_K, K_{t_1}^{t_2}, Q(d))}{\sum_{\zeta} p(O|\zeta)p(\zeta)} \end{aligned} \quad (4)$$

where $K_{t_1}^{t_2}$ denotes the event that K occurs between frame t_1 and t_2 of speech O , ζ is any complete path in the lattice, and ζ_K is any complete path passing through the partial path representing the detection of K .

In the implementation, all the possible pronunciations predicted by the joint-multigram model are searched within the lattices, and all putative detections are collected with confidence scores measured according to Equation 2.

3.2. Soft match

Soft match is another approach widely used to handle pronunciation uncertainty. The basic idea is to allow some mismatch between the lexical form of the search term and the detected form in the lattice. To extend our discussion to soft match, we write detection d as

$$d = (K, Q, Q_l, s = (t_1, t_2), v_a, v_l, \dots) \quad (6)$$

where Q_l denotes the detected form in the lattice.

Similar to the reasoning behind SPM, the confidence score of a detection is written as a linear interpolation

$$c_{fm}(d) = (1 - \nu)c_f(d) + \nu c_m(d) \quad (7)$$

where ν is an interpolation factor optimised with the development set, and $c_m(d)$ is a *match confidence* of the lexical form and the detected form, derived from an acoustic confusion matrix.

In the implementation, all the putative detections that meet some ‘mismatch constraints’ are collected, with confidence scores measured according to Equation 7.

3.3. Experimental results

Experiments were conducted with SPM and soft match individually. The baseline system used the 1-best pronunciation predicted by the joint-multigram model. In the SPM-based system, the maximum number of predictions were limited to 50 for sake of efficient computation; in the soft match-based system, only one substitution was allowed in term search, as allowing more mismatches produced too many false alarms without additional performance gain.

Table 1 shows the experimental results. We see that both SPM and soft match substantially improved the STD performance in terms of ATWV. A t -test shows that both improvements are statistically significant ($p < 0.01$). Another observation is that the soft match approach achieved better performance in terms of ATWV, although it is not the case in terms of max-ATWV. This looks like a tuning issue, if we notice that the interpolation weight ν is highly biased.

The DET curves shown in Figure 2 reveal more information. We can see the soft match-based system performs the best with a high FA rate, but does not work well when high precision is crucial. In contrast, the SPM-based system provides a consistent performance improvement with a range of precision values. This behaviour can be attributed to the phonetic constraints imposed by the joint-multigram model when predicting alternative pronunciations.

System	γ/ν	ATWV	max-ATWV
Baseline	-/-	0.2761	0.2770
SPM	0.7/-	0.3153	0.3303
Soft match	-/0.9998	0.3275	0.3300

Table 1. The STD results with SPM and soft match. *max-ATWV* is the maximum ATWV with an ideal decision threshold; γ and ν are interpolation weights for SPM and soft match respectively.

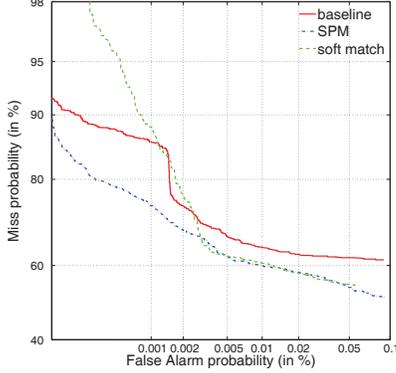


Fig. 2. The DET curves of the baseline system, SPM-based system and soft match-based system.

4. DISCRIMINATIVE DECISION-MAKING FOR SPM AND SOFT MATCH

4.1. Discriminative decision-making

A critical problem of the uncertainty treatment approaches, both SPM and soft match, is that they often generate a large number of false alarms, increasing the difficulty of the decision-making task. To improve the detection performance, we need a more powerful decision strategy that can pick up correct detections from the flood of false alarms.

In a previous study [12], we have proposed a discriminative decision strategy which utilises discriminative models to integrate some term-dependent factors into classification posterior probabilities, thus enhancing the discriminative power of the decision. We can use this technique to control the false alarms introduced by the uncertainty treatment.

Specifically, we want to build a mapping function f which converts some decision factors into a discriminative confidence measure which is in fact the classification posterior probability. For the SPM-based system, the function has the form

$$c_{disc}(d) = f(c_{fp}(d), c_f(d), c_p(d), R_0, R_1) \quad (8)$$

where $c_{disc}(d)$ represents the discriminative confidence of detection d , and c_{fp} has been defined in Equation 2. R_0 and R_1 are two occurrence-derived term-dependent factors defined as follows,

$$R_0(K) = \frac{\sum_i c_f(d_i^K)}{T} \quad (9)$$

and

$$R_1(K) = \frac{\sum_i (1 - c_f(d_i^K))}{T} \quad (10)$$

where d_i^K is the i -th detection of term K , and T is the length in seconds of the audio.

Similarly, the discriminative mapping function for the soft match-based system is defined as follows,

$$c_{disc}(d) = f(c_{fm}(d), c_f(d), c_m(d), R_0, R_1) \quad (11)$$

where c_{fm} is defined as in Equation 7.

4.2. Detection combination

An advantage of the discriminative decision approach is that the scores from different systems are normalised to become classification posterior probabilities, and therefore can be merged directly. Supposing a term is hypothesised as detection d_1 and d_2 by two systems respectively, and the hypothesised detections are overlapped, we then merge them as a single detection d which has the earliest and latest hypothesised start and end times, and a confidence computed as Equation 12,

$$c_{disc}(d) = 1 - (1 - c_{disc}(d_1))(1 - c_{disc}(d_2))^\alpha \quad (12)$$

where α is a tunable scale factor. Note that detections from individual systems are duplicated to the final result directly, along with the confidence scores. A useful property of the discriminative confidence-based combination is that the confidence score of the merged detection is still discriminative.

4.3. Experimental results

We experimented with two alternative discriminative methods to construct the discriminative mapping function f : a multiple layer perceptron (MLP) and a support vector machine (SVM) [15]. To prevent data sparsity, STD experiments were first conducted on the development set with *both* OOV terms and in-vocabulary (INV) terms. Afterwards, each detection was labelled according to whether it was a hit or a false alarm, and these were employed to train the MLP and SVM. In experiments, we found that the SVM model worked better with SPM, and the MLP worked better with soft match. This may be attributed to the bias training with SPM¹, and the robustness of the SVM against this problem.

Table 2 shows the experimental results of the SPM and soft match -based systems with the discriminative decision, as well as their combination. We can see that the discriminative decision strategy improved both the SPM and soft match -based systems, although the improvement is more substantial for the soft match-based system. In addition, when these two systems are combined, further improvement was obtained. A t -test shows that the combined system outperformed both individual systems significantly ($p < 0.01$).

The DET curves are shown in Figure 3. This confirms again that the discriminative decision approach enhances the soft match-based system more than the SPM-based system. Again, this may be due to the biased training with SPM.

¹The discriminative models were largely trained on INV terms for which the SPM tends to model very differently from OOV terms, thus providing biased training data.

System	ATWV	Model	max-ATWV
SPM (disc)	0.3235	SVM	0.3352
Soft match (disc)	0.3379	MLP	0.3409
SPM (disc)+soft match (disc)	0.3593	-	0.3604

Table 2. The STD results of the SPM and soft match -based systems with discriminative decision, as well as their combination.

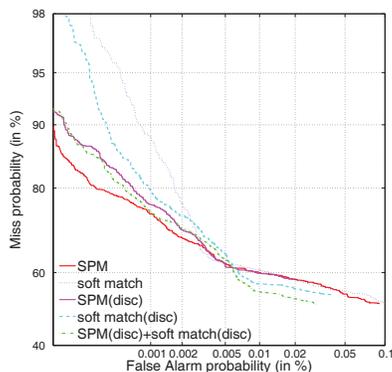


Fig. 3. The DET curves of the SPM-based system and soft match-based system with discriminative decision, as well as their combination.

5. CONCLUSIONS

We compared two approaches, SPM and soft match, for dealing with the pronunciation uncertainty of OOV terms in spoken term detection, and employed a discriminative decision strategy to control the overwhelming false alarms with both approaches. Experimental results demonstrated that the SPM approach works well with a low FA rate while the soft match approach is superior with more false alarms. Applying the discriminative decision enhanced the soft match-based system significantly, and the discriminative confidence-based combination of SPM and soft match gave additional and significant performance improvement.

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*, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 10 edition, September 2006.
- [2] Jonathan Mamou, Bhuvana Ramabhadran, and Olivier Siohan, "Vocabulary independent spoken term detection," in *Proc. ACM-SIGIR'07*, Amsterdam, July 2007, pp. 615–622.
- [3] David R. H. Miller, Michael Kleber, Chia lin Kao, Owen Kimball, Thomas Colthurst, Stephen A. Lowe, Richard M. Schwartz, and Herbert Gish, "Rapid and accurate spoken term detection," in *Proc. Interspeech'07*, Antwerp, Belgium, August 2007, pp. 314–317.
- [4] Roy Wallace, Robbie Vogt, and Sridha Sridharan, "A phonetic search approach to the 2006 NIST spoken term detection evaluation," in *Proc. Interspeech'07*, Antwerp, Belgium, 2007, pp. 2393–2396.
- [5] Dimitra Vergyri, Izhak Shafran, Andreas Stolcke, Ramana R. Gadde, Murat Akbacak, Brian Roark, and Wen Wang, "The SRI/OGI 2006 spoken term detection system," in *Proc. Interspeech'07*, Antwerp, Belgium, 2007, pp. 2393–2396.
- [6] Igor Szoke, Lukas Burget, Jan Cernocky, and Michal Fapso, "Sub-word modeling of out of vocabulary words in spoken term detection," in *Proc. IEEE Workshop on Spoken Language Technology (SLT'08)*, Goa, India, 2008.
- [7] Timo Mertens and Daniel Schneider, "Efficient subword lattice retrieval for german spoken term detection," in *Proc. ICASSP'09*, 2009, pp. 4885–4888.
- [8] Don Watson, *Death Sentence, The Decay of Public Language*, Knopf, Sydney, 2003.
- [9] Igor Szoke, Michal Fapso, Martin Karafiat, Lukas Burget, Frantisek Grezl, Petr Schwarz, Ondrejlembek, Pavel Matejka, Stanislav Kontar, and Jan Cernocky, "BUT system for NIST STD 2006 - English," in *Proc. NIST Spoken Term Detection Evaluation workshop (STD'06)*, Washington D.C., US, 2006, National Institute of Standards and Technology.
- [10] Dong Wang, Simon King, and Joe Frankel, "Stochastic pronunciation modelling for spoken term detection," in *Proc. Interspeech'09*, Brighton, UK, September 2009.
- [11] Sabine Deligne, Francois Yvon, and Frederic Bimbot, "Variable length sequence matching for phonetic transcription using joint multigrams," in *Proc. Eurospeech'95*, Madrid, 1995, pp. 2243–2246.
- [12] Dong Wang, Simon King, Joe Frankel, and Peter Bell, "Term-dependent confidence for out-of-vocabulary term detection," in *Proc. Interspeech'09*, Brighton, UK, September 2009.
- [13] Thomas Hain, Lukas Burget, John Dines, Giulia Garau, Martin Karafiat, Mike Lincoln, Jithendra Vepa, and Vincent Wan, "The AMI meeting transcription system: Progress and performance," in *Machine Learning for Multimodal Interaction*, vol. 4299/2006, pp. 419–431. Springer Berlin/Heidelberg, 2006.
- [14] Murat Akbacak, Dimitra Vergyri, and Andreas Stolcke, "Open-vocabulary spoken term detection using grapheme-based hybrid recognition systems," in *Proc. ICASSP'08*, Las Vegas, US, April 2008, pp. 5240–5243.
- [15] Rong Zhang and Alexander I. Rudnicky, "Word level confidence annotation using combinations of features," in *Proc. Eurospeech'01*, Aalborg, Denmark, September 2001, pp. 2105–2108.