



CRF-based Stochastic Pronunciation Modeling for Out-of-Vocabulary Spoken Term Detection

Dong Wang^{1,2}, Simon King¹, Nicholas Evans², Raphaël Troncy²

¹Centre for Speech Technology Research, University of Edinburgh

²Multimedia Communication Department, EURECOM, France

Dong.Wang@ed.ac.uk, Simon.King@ed.ac.uk

evans@eurecom.fr, Raphael.Troncy@eurecom.fr

Abstract

Out-of-vocabulary (OOV) terms present a significant challenge to spoken term detection (STD). This challenge, to a large extent, lies in the high degree of uncertainty in pronunciations of OOV terms. In previous work, we presented a stochastic pronunciation modeling (SPM) approach to compensate for this uncertainty. A shortcoming of our original work, however, is that the SPM was based on a joint-multigram model (JMM), which is suboptimal. In this paper, we propose to use conditional random fields (CRFs) for letter-to-sound conversion, which significantly improves quality of the predicted pronunciations. When applied to OOV STD, we achieve considerable performance improvement with both a 1-best system and an SPM-based system.

Index Terms: speech recognition, spoken term detection, conditional random field, joint multigram model

1. Introduction

As defined by NIST in 2006 [1], spoken term detection (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. Partly due to the series of evaluations organized by NIST, STD has attracted significant interest, including [2, 3, 4, 5, 6].

Unlike conventional keyword spotting, STD is an open-vocabulary task and must therefore cope with queries containing out-of-vocabulary (OOV) terms. For example, in the searching of broadcast news or educational material, a task that we are addressing through the ACAV project¹, queries may contain OOV entity names or technical terms, which can present a significant challenge. The usual approach to detecting OOV terms employs subword units which searches for subword representations of the search terms that are obtained from letter-to-sound (LTS) conversion. A potential problem with the subword approach, however, is that no special acoustic and linguistic properties of OOV terms are taken into account, which leads to much worse detection performance for OOV terms than for in-vocabulary (INV) terms. It is a reasonable hypothesis that the detection performance for OOV terms can be improved by compensating for the OOV special properties.

One such property is the high degree of pronunciation uncertainty. Different from INV terms, pronunciations of OOV terms are unknown, which leads to reduced familiarity or standardization in their pronunciation, resulting in additional uncertainty in pronunciations which we refer to as *lexical deviation*.

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.

It is distinctly different from acoustic variation and therefore cannot be fully compensated for by commonly employed soft matching techniques.

In previous work we presented a stochastic pronunciation modeling (SPM) approach to deal with lexical deviation [7]. This approach involves the searching of all possible pronunciations of OOV terms generated according to a stochastic pronunciation model; this amounts to treating the pronunciation as a hidden variable, and integrating it out. A potential problem in the original work lies in the use of a joint-multigram model (JMM) to implement SPM. As we will see in the next section, the JMM-based LTS conversion is sub-optimal for pronunciation prediction, and leads to sub-optimal STD performance. This paper reports our efforts to use a conditional random field (CRF) both for LTS conversion and for SPM. In contrast to JMMs, the CRF is a conditional model and performs global inference thus it is better suited to LTS conversion. Of greater importance is the ability of CRFs to generate higher quality n-best predictions than JMMs, which provides a better pronunciation model to implement SPM.

The remainder of this paper is organized as follows. In Section 2 we present the CRF-based approach to LTS conversion, and then show its application to implement SPM in Section 3. Experimental work is reported in Section 4 and some conclusions and future work are discussed in Section 5.

2. Pronunciation prediction using CRFs

LTS conversion has been studied for many years, mostly in the context of text-to-speech (TTS) for OOV word synthesis. Most state-of-the-art LTS conversion systems resort to a model-based approach, which learns phonological rules from representative exemplars and represents them through statistical models, such as artificial neural networks [8], hidden Markov models (HMMs) [9], classification and regression trees (CARTs) [10] and joint-multigram models (JMMs) [11].

In this work, we investigated the use of CRFs to tackle the LTS conversion task. A CRF is a sequence modeling framework that models the conditional probability distribution of a label sequence given an observation sequence. As a discriminative and conditional model, CRFs are a powerful tool for labeling and segmenting sequential data, and have received much interest in a wide range of research fields, e.g. text processing [12], bioinformatics [13] and speech recognition [14].

Compared to existing approaches, CRFs have several characteristics that make them more suitable for the LTS conversion task. First, the CRF is a conditional model and thus relaxes the conditional independence assumption that is required by generative models such as HMMs to make inference tractable; second, CRFs infer entire label sequences, which is different from the piece-wise inference implemented by other conditional

models such as decision trees and artificial neural networks; third, the CRF is a discriminative model and thus does not need to model the joint probability distribution of observations (graphemes) as we have to with JMMs. Finally, the CRF loss function is convex, which guarantees convergence to the global optimum [15].

To apply CRFs to LTS conversion, we treat word spellings (grapheme sequences) as observations and pronunciations (phoneme sequences) as labels. The task of LTS conversion thus amounts to assigning an optimal label sequence given the entire observation, a problem to which CRFs are ideally suited. According to the definition given by Lafferty et al. [15] and when applied to such a task, the CRF can be formally written as follows:

$$P(Q|G) = \frac{1}{Z(G)} \exp\left\{\sum_{k=1}^K \lambda_k F_k(Q, G)\right\}, \quad (1)$$

where G is the spelling (grapheme sequence) of the word whose pronunciation we seek, Q is a candidate pronunciation, F_k is the k -th aggregated feature, and λ_k is a factor to scale its contribution to the global probability. $Z(G)$ is a normalization quantity given by:

$$Z(G) = \sum_Q \exp\left\{\sum_{k=1}^K \lambda_k F_k(Q, G)\right\}. \quad (2)$$

Considering the Markov assumption, the undirected graph of the CRF is separated into cliques, each of which contains two consecutive phonemes and the entire grapheme sequence. Therefore, the aggregated feature $F_k(Q, G)$ can be factored into feature functions of cliques, given by:

$$F_k(Q, G) = \sum_{j=1}^{n-1} \{f_k(Q_j, Q_{j-1}, G, j)\}, \quad (3)$$

where $f_k(Q_j, Q_{j-1}, G, j)$ is the k -th feature function of the j -th clique, and n is the length of the grapheme sequence.

A commonly used family of features are binary functions that return binary values by examining the graphemes and phonemes at various positions in the clique. For example, the following feature function returns a non-zero value if and only if the current and the previous graphemes are H and I respectively, and the current phoneme is $/i/$.

$$f(Q_j, Q_{j-1}, G, j) = \begin{cases} 1 & G_{j-1} = H \ G_j = I \ Q_j = /i/ \\ 0 & \text{otherwise} \end{cases}. \quad (4)$$

A practical problem when building the CRF-based LTS conversion system is that the phoneme and grapheme sequences of a word are often of different length, which is inconsistent with the CRF structure. To solve this problem, an *empty* symbol can be inserted into the original sequence so that 1-to-1 alignment is achieved. This can be conducted manually; here we chose an automatic approach that aligns the phoneme and grapheme sequences based on a joint-multigram model whose grapheme and phoneme components consist of 1 symbol at most.

To examine performance of the CRF-based LTS conversion, we conducted our experiments on the dictionary used by the AMI RT05s LVCSR system [16], with 36575 words randomly selected for training, 4064 words for parameter tuning and 8000 words for evaluation. The CRF++ v0.52 toolkit [17] implemented by Taku Kudo at NTT Communication Science Laboratories in Japan was used to train the CRF model and perform the test.

To evaluate the proposed CRF-based approach, we compare it to a JMM-based baseline system which was reported to give the best performance among other conventional models [18]. For the CRF-based system, a range of context configurations were examined. For example, $(-2, +2)$ indicates that the feature function (Equation 3) covers 2 graphemes before and after the current position. Results are presented in terms of word error rate (WER) and are shown in Table 1. We observe that prediction accuracy of the CRF-based system increases rapidly as the context increases (67.6% to 25.4% WER). The best performance was achieved with the 4-grapheme context, i.e. $(-4, +4)$; broader contexts were prohibited by memory limitations. Moreover, we see that the $(-4,+4)$ CRF-based approach achieves much better performance than the JMM-based approach (25.4% cf. 31.3% WER). A pair-wise t -test shows that the performance improvement with the CRF over the JMM is highly significant ($p < 10^{-14}$). This result supports our hypothesis that a global, conditional model such as the CRF is well suited to the task of LTS conversion.

3. CRF-based stochastic pronunciation modeling

We assume that degradations in STD performance caused by OOV terms are partly due to the inherent high variability in their pronunciations, to a large extent arising from lexical deviation. In [7] we proposed an SPM approach to compensate for the lexical deviation. This approach considers all possible pronunciations of OOV terms, and assigns to each putative detection a composite confidence according to:

$$c(d) = \gamma c_{lat}(d) + (1 - \gamma) c_{pron}(d), \quad (5)$$

where d denotes a putative detection and $c(d)$ denotes the associated confidence. $c_{lat}(d)$ is the lattice-based detection confidence, $c_{pron}(d)$ is the pronunciation confidence given by a stochastic pronunciation model, and γ is a factor for linear interpolation.

In the original work [7], SPM was implemented using a JMM. JMMs are a powerful means of determining n -best predictions and can compute their posterior probabilities easily, which provides the pronunciation confidence $c_{pron}(d)$ requested by SPM, giving

$$c_{pron}(d) = P_{jmm}(Q_d|G_d) \quad (6)$$

where Q_d is the pronunciation based on which d is detected, and G_d is the word spelling of the term that d belongs to.

A potential problem of JMM-based SPM is that the pronunciation confidence (posterior probability $P_{jmm}(Q_d|G_d)$) is derived from joint probabilities applying the Bayesian rule. This calculation is usually based on lattices that are generated by the decoding process, and are thus potentially inaccurate. In contrast, CRFs compute the posterior probability of each candidate from the model, which is comparatively more accurate and straightforward.

Model	WER (%)
JMM	31.3
CRF (-1,+1)	67.6
CRF (-2,+2)	40.9
CRF (-3,+3)	29.7
CRF (-4,+4)	25.4

Table 1: The LTS result of 1-best prediction using the JMM and CRF.

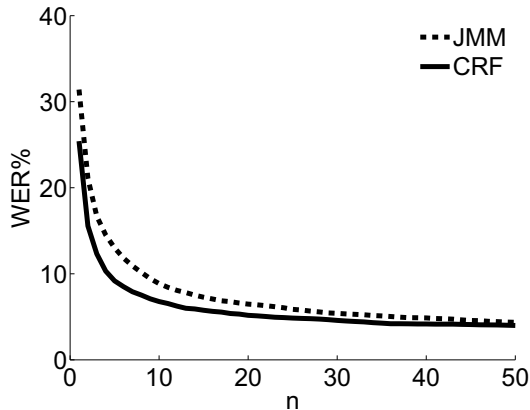


Figure 1: Result of n-best predictions with the JMM and CRF.

To compare the quality of CRF-based and JMM-based SPM, we examine the quality of n-best predictions provided by CRFs and JMMs. The experiment was conducted under the same conditions as in the previous section, except that n-best pronunciations were predicted. The results are evaluated in terms of n-best WER, i.e. the proportion of the words for which none of the n-best predictions is correct. Figure 1 shows the results, where the number of predictions n varies from 1 to 50. We see that for lower values of n , the CRF-based approach provides a higher quality n-best list than does the JMM-based approach.

With CRF-based n-best prediction Equation 6 is replaced by:

$$c_{pron}(d) = P_{crf}(Q_d|G_d), \quad (7)$$

where $P_{crf}(Q_d|G_d)$ is given by the CRF-based LTS conversion according to Equation 1.

4. Experiments

In this section, we apply the CRF-based LTS conversion to OOV spoken term detection, which can be either 1-best STD systems based on 1-best pronunciation prediction, or SPM-based STD systems based on n-best pronunciation predictions.

The experiments were conducted on meeting speech recorded from individual headset microphones (IHM), and focused on OOV terms in English, using phoneme-based ASR and STD systems. 482 search terms were carefully selected as OOV terms and were removed from both ASR and STD dictionaries, in addition to all materials used for acoustic model (AM) and language model (LM) training. After the OOV purge, there remain a total of 2736 occurrences of OOV terms in the evaluation data.

The AMs and LMs were trained on the corpora used by the AMI RT05s system [16]. 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 and 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 [6] provided by the Speech

Processing Group at the Brno University of Technology. Term-dependent normalization was applied to improve decision quality, as described in [19]. STD performance is reported in terms of average term-weighted value (ATWV) [1]; 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*[1]. More information about the experimental system can be found in [20].

4.1. CRF-based LTS conversion

We first applied CRF-based LTS conversion to OOV STD, i.e. employing the CRF model to predict 1-best pronunciations for each OOV term. Results are reported in Table 2 in terms of ATWV and max-ATWV. The first line presents results for the JMM-based system and the second line presents corresponding results for the CRF-based system. The CRF-based system gives marginally better STD performance than the JMM-based system (0.2761 cf. 0.2887), but the improvement is not statistically significant (a *t*-test gives $p \approx 0.2$).

Model	ATWV	max-ATWV
JMM	0.2761	0.2770
CRF (-4,+4)	0.2887	0.2947
JMM+CRF	0.3279	0.3280

Table 2: STD performance with 1-best pronunciation prediction, using JMM and CRF-based LTS conversion, as well as their detection combination.

Considering that JMMs and CRFs model the spelling-to-pronunciation relationship in different ways, they are likely to be complementary. To verify this conjecture, we conducted a third experiment where the detections hypothesized by the JMM and CRF-based systems are combined as proposed in [21]. Results are presented in the third line of Table 2. It can be seen that the combination leads to a considerable improvement in performance. A *t*-test shows that the improvement is highly significant ($p < 0.001$).

4.2. CRF-based SPM

In the second set of experiments, we seek to assess the utility of CRFs for SPM. Results are presented in Table 3, again in terms of ATWV and max-ATWV. They clearly show that the CRF-based system considerably outperforms the JMM-based system, suggesting that CRFs provide a higher quality pronunciation model for SPM than do JMMs. A *t*-test shows that this improvement is weakly significant ($p < 0.05$). Unfortunately, when we tried to combine the detections from both systems in this case, we did not observe any gain in performance, indicating that the complementarity between JMM and CRF approaches does not play an important role in SPM-based systems. Further analysis shows that the combined system leads to increases in false alarms which degrade overall performance.

Model	ATWV	max-ATWV
JMM SPM	0.3153	0.3303
CRF SPM	0.3352	0.3603
JMM+CRF	0.3253	0.3451

Table 3: The STD performance with SPM, based on the JMM and CRF, as well as their detection combination.

Figure 2 shows the DET curves of the 1-best and SPM-based STD systems using both JMM and CRF models for LTS conversion. The CRF-based approach performs as well as the

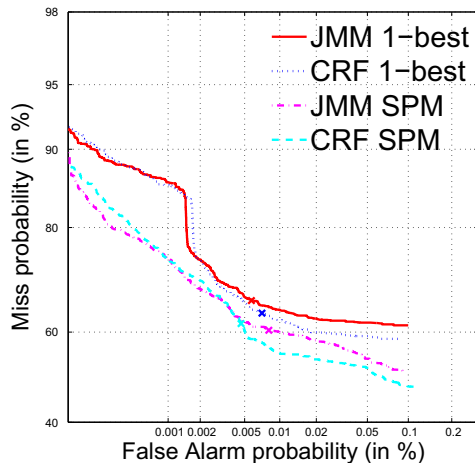


Figure 2: DET curves of 1-best and SPM-based STD systems, for JMM and CRF-based approaches. The decision point on which the reported ATWV resides is indicated with an 'x' on each curve.

JMM-based approach over most of the operating region but leads to superior performance for FA probabilities in excess of 0.005 (bottom right of Figure 2) i.e. where the number of correctly detected OOV terms is the greatest, as is often preferred in information retrieval tasks.

5. Conclusions

The contributions of this paper are two-fold: first we propose a novel CRF-based approach to LTS conversion, second we apply the new LTS model to OOV spoken term detection. We show that CRFs provide significantly better performance than JMMs when applied to LTS conversion, and that they substantially improve performance of OOV term detection when applied to STD, with either the 1-best system or the SPM-based system.

Future work includes the application of CRFs to the detection process directly, so that context information can be integrated and utilized for term search and confidence estimation. Through the ACAV project, we are also working to integrate the approach to help transcribe and index a huge volume of multimedia data hosted and shared on the Dailymotion² platform. Content will be rendered in various accessibility scenarios including those of broadcast news and educational contexts where OOV terms occur frequently.

6. Acknowledgements

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).

7. References

[1] NIST, *The spoken term detection (STD) 2006 evaluation plan*, 10th ed., National Institute of Standards and Technology (NIST),

²<http://www.dailymotion.com/>

Gaithersburg, MD, USA, September 2006. [Online]. Available: <http://www.nist.gov/speech/tests/std>

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

[3] 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*, Taipei, Taiwan, April 2009, pp. 3957–3960.

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

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

[6] 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: Association for Computing Machinery, 2008.

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

[8] T. J. Sejnowski and C. R. Rosenberg, "Parallel networks that learn to pronounce English text," *Complex Systems*, vol. 1, no. 1, pp. 145–168, 1987.

[9] P. Taylor, "Hidden Markov models for grapheme to phoneme conversion," in *Proc. Interspeech'05*, Lisbon, Portugal, September 2005, pp. 1973–1976.

[10] A. W. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in *Proc. 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 77–80.

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

[12] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proc. HLT/NAACL-03*, Edmonton, Canada, 2003.

[13] A. Culotta, D. Kulp, and A. McCallum, "Gene prediction with conditional random fields," University of Massachusetts, Amherst, Tech. Rep., 2005.

[14] Y. Hifny and S. Renals, "Speech recognition using augmented conditional random fields," *ASLPA05*, vol. 17, no. 2, pp. 354–365, February 2009.

[15] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th International Conf. on Machine Learning*, San Francisco, CA, USA, 2001, pp. 282–289.

[16] 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.

[17] T. Kudo, 2009. [Online]. Available: <http://crfpp.sourceforge.net/>

[18] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, May 2008.

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

[20] D. Wang, "Out-of-vocabulary spoken term detection," Ph.D. dissertation, The Center for Speech Technology Research, Edinburgh University, December 2009.

[21] D. Wang, S. King, J. Frankel, and P. Bell, "Stochastic pronunciation modelling and soft match for out-of-vocabulary spoken term detection," in *Proc. ICASSP'10*, Texas, US, March 2010.