Stochastic Pronunciation Modelling for Spoken Term Detection

Dong Wang, Simon King, Joe Frankel

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

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. Current approaches to STD do not acknowledge the particular properties of OOV terms, such as pronunciation uncertainty. In this paper, we use a stochastic pronunciation model to deal with the uncertain pronunciations of OOV terms. By considering all possible term pronunciations, predicted by a joint-multigram model, we observe a significant performance improvement.

Index Terms: joint-multigram, pronunciation model, 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. Because of its fundamental importance for multimedia information retrieval and the evaluation series run 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. Some state-of-the-art STD systems include those reported in [2]–[7].

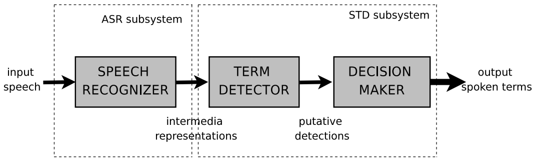

Figure 1: *The standard STD architecture: a speech recogniser converts speech signals to an intermediate representation (e.g., phoneme lattices); a term detector searches this representation for putative occurrences of the search terms; a decision maker ascertains whether each putative detection is reliable.*

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 problems for STD systems because their pronunciations are unknown and they are poorly represented by the acoustic and language models of the system; ironically, these are likely to be common search terms. 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., [3],[4],[9],[10]. 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. Other types of subword units under investigation include word-fragments [11], particles [12], graphones [5],[13], multigrams [6], syllables [14] and graphemes [15].

Both in-vocabulary (INV) terms and OOV terms can be retrieved in the same way by a subword-based system, based on their pronunciations. However, OOV terms have different properties to INV terms. In particular, they may have more unpredictable pronunciations. We present the use of a stochastic model to deal with this uncertainty which we believe comes from: ASR errors, pronunciation variation, and acoustic variation. We employ a joint-multigram model to estimate the distribution of possible pronunciations for OOV terms, then use this distribution when searching the lattice.

2. Joint-multigram model-based pronunciation prediction

In the phoneme-based STD approach, the first step in detecting an OOV term is to predict its pronunciation from its written form using a LTS model.The joint-multigram model, proposed by [16], is a promising approach to LTS, e.g., [17],[18]. This model is motivated by the idea that both spelling and pronunciation depend on some underlying hidden process of human language. To infer the pronunciation, given a spelling, implies inferring the underlying process. This is quite different from approaches that assume pronunciation depends on spelling. We follow Bisani and Ney [19] and call the units of the underlying process *graphones*. A graphone is denoted as $u = (\tilde{g}, \tilde{q})$ where $\tilde{g}$ and $\tilde{q}$ are the grapheme and phoneme component of u respectively. Both $\tilde{g}$ and $\tilde{q}$ contain a sequence of symbols of length of N_{min} to N_{max}, which are defined when constructing the model. With graphones defined, the statistical property of spelling G and pronunciation Q can be written in graphones U as:

$$
\begin{aligned}
p(G,Q) &= \sum_{U;G(U)=G,Q(U)=Q} p(U) & (1)\\
&= \sum_{U;G(U)=G,Q(U)=Q} p(u_1,u_2,...,u_K) & (2)
\end{aligned}
$$

where $G(U)$ and $Q(U)$ denote the grapheme and phoneme component of U, respectively. The task of pronunciation prediction is then formulated as follows:

$$
\begin{aligned}
\hat{Q}(G) &= \arg\max_{Q} p(G,Q) & (3)\\
&= \arg\max_{Q} \sum_{U;G(U)=G,Q(U)=Q} p(U) & (4)\\
&\approx Q(\arg\max_{U;G(U)=G} p(U)) & (5)
\end{aligned}
$$

where Equation 5 shows an approximated decoding approach. Using the approach from [17],[18], we factor $p(U)$ into graphone n-grams:

$$p(U) = \prod_{j=1}^{|U|} p(u_j|h_j) \qquad (6)$$

where h_j is the graphone history of u_j.

We trained and tested a joint-multigram model on the dictionary used by the AMI RT05s LVCSR system [20]. The dictionary was randomly divided into three subsets: 36575 words for training, 4064 words for parameter tuning and 8000 words for evaluation.

Pronunciation prediction results are shown in Table 1 in terms of word error rate (WER) on the 8000 word evaluation set. For comparison, the performance using a CART model implemented in the Festival system [21] is also reported. The joint-multigram model-based approach generally outperforms the CART.

A significant advantage of the joint-multigram approach is the ability to predict multiple pronunciations – the distribution of pronunciations of a term. Whereas a CART obtains multiple pronunciations by concatenating alternative pronunciations for each grapheme, the joint-multigram model estimates the confidence of the whole pronunciation. This confidence can be written as a posterior probability given the spelling, and can be estimated easily from the pronunciation lattice that is constructed by the joint-multigram during prediction. The confidence of a prediction in graphone form is given by Equation 7:

$$c(U) = \frac{p(U)}{\sum_{U' \subseteq \Re(\phi)} p(U')} \qquad (7)$$

where $\Re(\phi)$ stands for the decoding lattice for word ϕ, and $p(U)$ denotes the probability of any graphone path U in $\Re(\phi)$. The results of the n-best pronunciation prediction are shown in Figure 2. We observe a significant error reduction, especially with the first few pronunciations.

3. Stochastic pronunciation modelling

3.1. Motivation

Most LTS models suffer from a significant prediction error rate, typically getting the pronunciation of around 30% of words wrong. For INV terms, we assume that the pronunciations obtained from dictionaries are *canonical* and that both pronunciation variation and recognition errors can both be regarded as deviations from the canonical forms which can be handled during the lattice search either by finding alternative paths in the lattice or allowing non-exact matches (e.g., [22]). For OOV terms, however, we can never assume the predicted pronunciations are canonical, otherwise any terms whose pronunciations are incorrectly predicted by LTS will not be detected. Rather, we must take properly account for LTS prediction errors – this suggests the use of a stochastic pronunciation model.

Pronunciation prediction	
Model	WER%
CART (stop=1)	35.2
joint multi-gram	34.4
+Kneser-Ney discounting & interpolation	33.2
+ insertion compensation	32.7
+ reverse decoding	31.3
+ pronunciation unification	30.3

Table 1: *Experimental results of joint-multigram pronunciation prediction. Insertion compensation uses a factor to compensate pronunciations with more phonemes. Reverse decoding means the model uses right context instead of left context. Pronunciation unification employs the more accurate decoding scheme shown in Equation 4 instead of the approximation in Equation 5, following the idea presented in [18].*

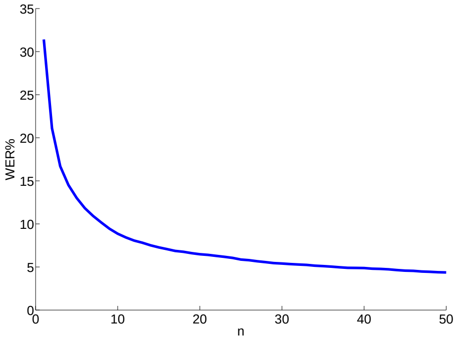

Figure 2: *As we consider more pronunciation variants (n), the chance of the correct pronunciation being the list increases, so the prediction error drops. The reduction in error is particularly rapid for low n – the joint-multigram frequently produces the correct pronunciation near the top of the n-best list.*

3.2. Multiple pronunciation prediction

Our first step in representing the distribution of pronunciations is an n-best list of pronunciations. As Figure 2 shows, the joint-multigram model is good at predicting and ranking multiple pronunciations. Of course, although n-best prediction increases the possibility that the correct pronunciation is produced, it leads to more false alarms so we set a confidence pruning threshold η on the predictions.

3.3. Stochastic pronunciation model

An obvious shortcoming of the n-best approach is that n and η are term-independent, although they should not be. Another problem is that the confidence of the predicted pronunciations is not utilised – all items in the n-best list are assumed to have the same uniform probability.

We therefore wish to integrate the confidence of the predicted pronunciation with the confidence of a term detection so that a jointly-optimal *postponed decision* can be formulated based on a *compound confidence*. So, rather than using n and η to control the quality of the predicted pronunciations, we take into account all possible pronunciations when searching for terms in the lattice. To make description clear, we first define a detection d as a tuple

$$d = \big(K, Q, s = (t_1, t_2), v_a, v_l, \ldots\big) \qquad (8)$$

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 "...".

We define the detection confidence $c_f(d)$ of a detection d as the posterior confidence of the event that the search term K occurs between t_1 and t_2 with pronunciation Q. This posterior probability can be estimated from the lattice, as in our previous study [15], using Equation 9.

Model	ATWV	max-ATWV	P(FA)	P(Miss)
CART	0.2126	0.2607	0.00002	0.766
joint-multigram	**0.2761**	0.2770	0.00006	0.667

Table 2: *STD performance using CART or joint-multigram. max-ATWV is the maximum ATWV value along the DET curve; P(FA) and P(Miss) are false alarm rate and missing rate respectively, as defined in the NIST STD06 evaluation plan [1].*

$$\begin{aligned} c_f(d) &= p(K_{t_1}^{t_2}, Q(d)|O) && (9) \\ &= \frac{\sum_{C_K} p(O|C_K, K_{t_1}^{t_2}, Q(d))p(C_K, K_{t_1}^{t_2}, Q(d))}{\sum_{\zeta} p(O|\zeta)p(\zeta)} && (10) \end{aligned}$$

where $K_{t_1}^{t_2}$ denotes the event that K occurs between frame t_1 and t_2 of speech O, C_K is the context of K, and ζ is any path in the lattice.

Now we define the pronunciation confidence $c_p(d)$ of the detection d as the posterior probability of pronunciation Q given the term K:

$$c_p(d) = p(Q|K). \qquad (11)$$

We will denote the model describing $p(Q|K)$ a "stochastic pronunciation model" (SPM). The joint-multigram model is a suitable SPM and can estimate $p(Q|K)$ according to Equation 7, where G will be the spelling of K.

From the detection confidence and the pronunciation confidence we can estimate the compound confidence of a pronunciation-bearing detection d as a combination of $c_f(d)$ and $c_p(d)$. We tried several methods, and found that linear interpolation of c_f and c_p gave best performance (Equation 12).

$$c(d) = (1-\gamma)c_f(d) + \gamma c_p(d) \qquad (12)$$

where γ is a weight. Note that c_f is related to the AM and LM scores, while c_p only relates to pronunciation.

4. Experiments

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 have more occurrences 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 [20]. 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; 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. The evaluation corpora comprise 11 hours of speech.

System	# Pron.	ATWV	max-ATWV	P(FA)	P(Miss)
1-best	484	0.2761	0.2770	0.00006	0.667
4-best	854	0.3013	0.3025	0.00006	0.636
SPM	20877	**0.3153**	0.3303	0.00008	0.604

Table 3: *STD performance for a joint-multigram with 1-best, 4-best and stochastic prediction. The second column reports the number of different pronunciations predicted for the set of search terms (for the SPM, a maximum of 50 pronunciations are allowed per term, to limit computational cost).*

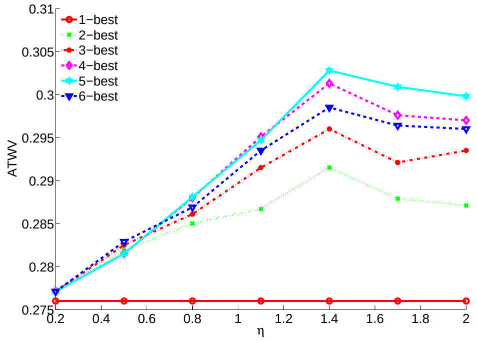

Figure 3: *STD performance with n-best prediction by joint-multigram. Each curve represents the performance with a particular value of n as confidence pruning threshold η is varied.*

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 graphone 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 [2],[13]. STD performance is reported in terms of ATWV [1]; detection (DET) curves are used to show behaviour at different hit/FA ratios.

4.1. STD performance with single pronunciation

We first examined STD performance with single 1-best pronunciations predicted by the joint-multigram; for comparison, the same experiment was conducted with the CART LTS model implemented in Festival. Results are shown in Table 2 and DET curves in Figure 5. A pairwise t-test shows the joint-multigram significantly outperforms the CART ($p < 0.001$).

4.2. STD performance with n-best pronunciations

In this experiment, we applied the n-best pronunciations predicted by the joint-multigram model to STD. Tuning the maximum number of pronunciations n and the confidence threshold η, we achieved the STD performance shown in Figure 3. t-test shows that all the n-best systems ($n > 1$) outperform the 1-best system significantly ($p < 0.01$).

4.3. STD performance with the SPM

The stochastic pronunciation model relies on interpolation factor γ to combine detection confidence c_f and pronunciation confidence c_p. The performance with varying γ is shown in Figure 4. γ was tuned to optimise STD performance on the dev set, with a value of 0.7 giving best performance. Results are shown in Table 3, with results for 1-best and 4-best prediction for comparison. DET curves of these systems are shown in Figure 5. A t-test shows the SPM-based system significantly outperformed the 1-best system ($p < 0.005$), but the improvement over the n-best system is not significant ($p \approx 0.2$).

4.4. Conclusions

We proposed that a stochastic pronunciation model will improve detection of OOV terms by properly representing the distribution of possible pronunciations and showed that this significantly outperformed a single prediction approach.

5. 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.

6. 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] D. Watson, *Death Sentence, The Decay of Public Language*. Knopf, Sydney, 2003.

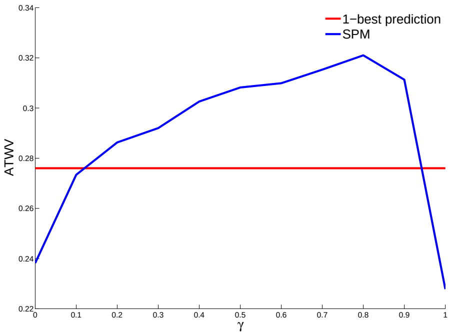

Figure 4: *STD performance using a joint-multigram SPM as the interpolation factor varies from 0 to 1.*

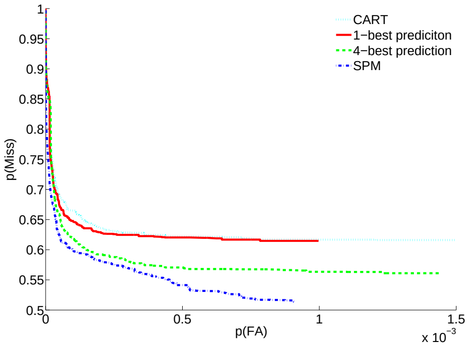

Figure 5: *DET curves for STD systems using CART or joint-multigram pronunciation prediction.*

[9] 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.

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

[11] F. Seide, P. Yu, C. Ma, , and E. Chang, "Vocabulary-independent search in spontaneous speech," in *Proc. ICASSP'04*, vol. 1, Quebec, Canada, May 2004, pp. 253–256.

[12] B. Logan, J. V. Thong, and P. Moreno, "Approaches to reduce the effects of OOV queries on indexed spoken audio," *IEEE Transaction on Multimedia*, vol. 7, no. 5, pp. 899–906, October 2005.

[13] M. Akbacak, D. Vergyri, and A. Stolcke, "Open-vocabulary spoken term detection using graphone-based hybrid recognition systems," in *Proc. ICASSP'08*, Los Angels, US, April 2008, pp. 5240–5243.

[14] S. Meng, P. Yu, F. Seide, and J. Liu, "A study of lattice-based spoken term detection for Chinese spontaneous speech," in *Proc. ASRU'07*, Japan, 2007, pp. 635–640.

[15] D. Wang, J. Frankel, T. Tejedor, and S. King, "A comparison of phone and grapheme-based spoken term detection," in *Proc. ICASSP'08*, April 2008.

[16] 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.

[17] S. F. Chen, "Conditional and joint models for grapheme-to-phoneme conversion," in *Proc. Eurospeech'03*, Geneva, Switzerland, November 2003, pp. 2033–2036.

[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] ——, "Investigations on jointmultigram models for grapheme-to-phoneme conversion," in *Proc. ICSLP'02*, vol. 1, Denver, CO, September 2002, pp. 105 – 108.

[20] 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.

[21] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[22] K. Audhkhasi and A. Verma, "Keyword search using modified minimum edit distance measure," in *Proc. ICASSP'07*, vol. 4, April 2007, pp. 929–932.