A CONFIDENCE-BASED APPROACH FOR IMPROVING KEYWORD HYPOTHESIS SCORES

M.S. Seigel, P.C. Woodland and M.J.F. Gales

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK
{mss46,pcw,mjfg}@eng.cam.ac.uk

ABSTRACT

The task in keyword spotting (KWS) is to hypothesise times at which any of a set of key terms occurs in audio. An important aspect of such systems are the scores assigned to these hypotheses, the accuracy of which have a significant impact on performance. Estimating these scores may be formulated as a confidence estimation problem, where a measure of confidence is assigned to each key term hypothesis. In this work, a set of discriminative features is defined, and combined using a conditional random field (CRF) model for improved confidence estimation. An extension to this model to directly address the problem of score normalisation across key terms is also introduced. The implicit score normalisation which results from applying this approach to separate systems in a hybrid configuration yields further benefits. Results are presented which show notable improvements in KWS performance using the techniques presented in this work.

Index Terms— keyword spotting, confidence estimation, conditional random fields, spoken term detection

1. INTRODUCTION

The keyword spotting task addresses the problem of reliably detecting the occurrence of specific single or multi-word key terms within audio data. A two-stage approach to KWS is commonly employed to achieve this. Firstly, an automatic speech recognition system is used to recognise the audio and generate an "index" of words and times for the audio from the recognition lattices. Thereafter, a search is performed within the indexed data for the key terms. The result of this process is a list of key term occurrences, each with a corresponding time at which it was hypothesised to exist in the audio.

A score is associated with every key term hypothesis. This score is the basis for the decision of whether to accept the hypothesis as a "hit" or reject it as a false alarm, and is therefore crucial to KWS performance. These scores are effectively confidence measures, and the task of computing them is therefore cast as a confidence estimation problem in this work. Typically, lattice-based keyword posteriors are employed as confidence measures [1, 2, 3, 4], however other measures such as those based on the coherence of the surrounding word context have been proposed in the literature [5, 6, 7, 8].

Many confidence measures are computed in a manner which is not dependent on the actual identity of the key term. Such "term independent" measures are flawed in that they imply the assumption in decision making that hypotheses for different key terms having the same term-independent score can be treated equally. This is however not the case, as differences in characteristics of the key terms such as length, frequency of occurrence, and language model scores mean that the key term scores naturally tend to fall in different ranges. Rank-based and term-weighted normalisation [1, 9, 10] have proven to be effective solutions to this problem. Discriminative score mapping [11, 12] is another approach which aims to achieve this normalisation through direct modelling.

In this work we propose the use of linear-chain CRF models as a unifying framework to perform confidence estimation for key term hypotheses, and account for score normalisation. In contrast to approaches based on support vector machines and multi-layer perceptrons [11, 12], this approach operates on sequences of observations for words within key terms to estimate a confidence score. The CRF model is also extended through the formulation of novel feature functions, which enable it to capture differences between the nature of key terms directly. Features extracted from the recognition lattices and language model are defined, with contextual lattice posteriors being introduced in this work as highly discriminative features which are useful for confidence estimation. Improvements are achieved when applying the approach separately in a word and sub-word level KWS system, with further performance gains being achieved in the combined (hybrid) configuration.

The first section of this paper describes the hybrid KWS system on which this work is based, and defines the features used for confidence estimation. Thereafter, the application of the CRF model to confidence estimation for key term hypotheses is detailed. Finally, the experimental setup and results of applying this technique in KWS are presented.

2. HYBRID KEYWORD SPOTTING

In the KWS approach taken in this work, the audio data is decoded once by the ASR system to generate lattices. These lattices can be represented as an index of words along with the times at which they occur in the audio. Key term search is carried out by looking up the entries for words of a key term. This is referred to as a pre-indexed approach, as the mapping of words to times is static and is pre-computed for given audio data. Techniques in which the ASR decoding phase uses knowledge of the key term list to adapt the search [10] may yield improved performance over the pre-indexed approach. However, adapting to the key term list requires that the audio be re-decoded by the ASR system whenever it is modified. This is not the case with the pre-indexing approach used in this work, which is for this reason more scalable.

The hypothesis space for KWS using this approach is restricted by the size of the ASR system lattices and the vocabulary used. This problem is addressed by generating lattices and carrying out key term search at both the word and sub-word levels, as was investi-

This work was in part supported by DARPA under the RATS program via a subcontract to BBN Technologies, as well as by the Nuance Foundation. The paper does not necessarily reflect the position or the policy of the US Government or the Nuance Foundation, and no official endorsement should be inferred by either party.

gated in [13, 14, 15]. In KWS systems, a trade-off between the number of false positives and false negatives generated must typically be made. The word and sub-word systems represent different operating points within this trade-off, with the combination of the two yielding a hybrid solution with the benefits of both.

2.1. Word-Level Key Term Search

All occurrences of words within the word-level system lattices which constitute key terms are considered as partial key term hypotheses. A start time and a set of features is associated with each such hypothesis. Given the set of partial hypotheses, individual word hypotheses are grouped into key term hypotheses where possible, provided the relevant partial hypotheses follow one another. This approach yields a fairly low number of key term hypotheses, as the lattices naturally represent a constrained hypothesis space. This results in the system having a relatively short range of possible operating points. Furthermore, as an exact match is sought, only key terms which are covered by the ASR vocabulary can be found.

2.2. Sub-Word-Level Key Term Search

In order to perform key term search at the sub-word level, the word-level lattices are marked up with sub-word timing and acoustic score information. Key term search is performed by considering all occurrences of sub-words which constitute any of the key terms as partial hypotheses. Individual sub-word partial hypotheses are clustered together to form longer key term hypotheses. This is permitted provided these partial hypotheses occur in the correct order, and satisfy constraints on the length of time allowed between sub-words. In this process, sub-words which form part of different word-level hypotheses can be grouped together as part of new key term hypotheses that are not present in the lattice. This results in a large number of hypotheses being generated (which can include out-of-vocabulary key terms), therefore covering a large range of KWS operating points.

3. FEATURES

During key term search, features which are indicative of the quality of the hypotheses are extracted from the lattices. One such feature is the word-level Lattice Arc Posterior Ratio (LAPR) [16]. Arc posterior probabilities $p(a|\mathbf{O})$ are calculated for each of the word-level arcs a in a lattice given the acoustic observation vectors $\mathbf{O}$ (as described in [17]). For a word k, the set of intersecting arcs $\mathcal{I}$ is defined as all arcs which overlap with the median time of the word arc for k. The LAPR for a word/keyword k is then calculated by summing the posterior probabilities $p(a|\mathbf{O})$ for all word-level arcs in the intersecting set $\mathcal{I}$ with the same word identity w_a as the word k. This sum is normalised by the sum of the posterior probabilities for all word arcs in $\mathcal{I}$ yielding:

$$\text{LAPR}(k) = \frac{\sum_{a\in\mathcal{I}} \delta(w_a, k) p(a|\mathbf{O})}{\sum_{a\in\mathcal{I}} p(a|\mathbf{O})} \tag{1}$$

where δ is the Kronecker delta function which returns 1 when its arguments match and 0 otherwise. Similarly, a Lattice Sub-Arc Acoustic Ratio (LSAAR) is computed from a sub-word lattice for a key term k by averaging over the N sub-words it contains as follows:

$$\text{LSAAR}(k) = \frac{1}{N} \sum_{g=1}^{N} \frac{\sum_{s\in\mathcal{I}} \delta(w_s, k_g) p(s|\mathbf{O})}{\sum_{s\in\mathcal{I}} p(s|\mathbf{O})}$$

where k_g is the identity of the sub-word at index g, the sub-word arc s has the identity w_s, and $\mathcal{I}$ is the set of intersecting sub-word arcs over which the sub-word-unit acoustic model scores $p(s|\mathbf{O})$ are summed. This is different from the sub-word posteriors described in [18, 19] which incorporate the word-level language model scores.

3.1. Contextual Posterior Features

Features which represent the confidence of the ASR system in the words immediately preceding and following the keyword itself (the context) are of interest. Such features serve to indicate whether the keyword hypothesis occurs within a sequence of likely words. For a lattice arc corresponding to a keyword hypothesis, the most likely preceding arc k' and following arc k'' within the lattice are found. These arcs represent the localised context for the word. Applying Equation 1 to these word arcs yields the contextual features LAPR(k') and LAPR(k''). These features are particularly relevant in the word-level system. However, a similar set of contextual posteriors can be computed in the sub-word level system. In this case, the preceding context posterior corresponds to that of the parent arc of the first sub-word, and the following context posterior is that of the parent arc for the last sub-word.

3.2. Unigram Prior Features

Certain key terms are generally more likely to occur than others. A feature indicative of this prior information is therefore of interest. Unigram probabilities for words within a given key term are obtained from the language model of the ASR system and multiplied together. The logarithm of this value is used as the unigram prior (UP) feature for this key term. The aim of this feature is to provide some contrast to the posterior scores, and inform the confidence estimation model when confidence scores should be boosted or discounted to account for the differences between key terms.

4. APPLYING THE CRF MODEL TO KEYWORD CONFIDENCE ESTIMATION

Linear-chain CRF models were first proposed as a discriminative modelling framework for segmenting and labelling sequence data [20]. These models define the distribution of a label sequence $\mathbf{y}$ conditioned on an observation sequence $\mathbf{X}$ as follows:

$$p(\mathbf{y}|\mathbf{X}) \propto \exp\Big(\sum_k \lambda_k t_k(\mathbf{y}) + \sum_l \mu_l g_l(\mathbf{y}, \mathbf{X})\Big)$$

where $t_k(\mathbf{y})$ are the transition feature functions and $g_l(\mathbf{y}, \mathbf{X})$ are the observation feature functions, with parameters λ_k and μ_l respectively. In applying this model to confidence estimation for key terms, the label sequence $\mathbf{y}$ corresponds to a sequence of True Positive (TP) or False Positive (FP) labels for the words in the key term. The observation sequence $\mathbf{X}$ corresponds to a sequence of vectors for each word or key term, which includes the features described in section 3.

During training, the model parameters λ_k and μ_l are estimated by optimising the conditional log-likelihood of the model using a gradient-based approach (Limited-Memory BFGS [21]). During test, the marginal probability of the model assigning the label for a true positive ("TP") to each word/observation in the sequence is calculated. The overall confidence score for the key term is obtained by averaging these individual scores.

4.1. Feature Functions

CRF models are highly flexible as they allow arbitrary feature functions which act on the features to be defined. In this work, we make use of the CRF modelling toolkit developed for confidence estimation in previous work [16]. Within this framework, it is possible for a set of custom observation feature functions (g_l) to be engineered, which are aimed at addressing a specific aspect of the task. These feature functions take a literal feature as an argument, and return the value of a continuous feature, provided the literal value matches that for which the feature function is defined. These "Literal Moment" feature functions are defined as follows:

$$\text{LITM1}_{y,l}(x_i^p, x_i^s) = x_i^p \; \delta(x_i^s, l) \; \delta(y_i, y)$$

where y and l are the label and literal value for which the feature function is defined, x_i^s and x_i^p are the literal (string) and continuous feature values corresponding to the current observation and y_i is the label for the current observation. These feature functions therefore learn separate first-order moment statistics of a continuous feature, for each possible value the literal feature can take. The model also allows for continuous features to be represented using spline feature functions [22], which have been shown to yield good performance for confidence estimation with CRFs in previous work [16].

5. EXPERIMENTS

Key term spotting experiments were carried out on a state-of-the-art recogniser which was built using data from the DARPA Robust Automatic Transcription of Speech (RATS) program for Arabic keyword-spotting [23]. The data consists of Levantine Arabic conversational telephone speech retransmitted over 8 degraded communication channels. This is a highly challenging recognition task, and the recogniser used employs link and speaker adaptive training, as well as front-end CMLLR to compensate for the difficulty of the task. The WER achieved varies between 62% and 81% across communication links, which naturally has an effect on the level of difficulty in performing KWS. The decoding structure of the ASR system consists of multiple passes, the first two of which are the main lattice generation phase, with adaptation being applied in the final phase. A bigram language model (LM) is used during initial decoding of this final phase, before the lattices are rescored using a trigram LM. It was found that using an LM scaling factor value lower than that tuned for optimal ASR performance resulted in improved KWS performance. The system lattices output from the final rescoring phase are used for KWS experiments. These lattices are converted from their romanised form to a normalised UTF-8 Arabic representation before the key term search is carried out entirely in this domain. The recogniser uses graphemes as a sub-word level representation, which is therefore the sub-word unit used in the KWS system.

The RATS program defines a list of 219 key terms for KWS evaluation, 64 of which are single word terms with 155 multi-word terms. The KWS operating point which is of particular interest in this program is that at a false alarm rate of 4%. All experiments in this work are based on the dev-1 dataset, which comprises data held out from the original training set for each channel. This dataset was further split into a training set for the CRF models, and a test set for evaluation. The training subset contains 181 of the total 401 key term occurrences, with the remaining 220 occurrences in the test subset. The results shown in Table 1 for the word and grapheme-based KWS systems are obtained when all key term hypotheses are accepted (i.e. no score threshold is applied below which hypotheses are rejected). Scoring the training data provides supervised training data labels ("TP/FP") for the CRF models, which are augmented with the set of features relevant for each system.

	Train			Test		
System	TP	R	P	TP	R	P
Word	103	0.569	0.138	102	0.464	0.128
Grapheme	150	0.829	0.003	174	0.790	0.003

Table 1. Results for the word and grapheme-based KWS systems on training and test subsets of dev-1, with no threshold applied to confidence scores. TP = True Positives, R = Recall and P = Precision.

5.1. Word-Based System Results

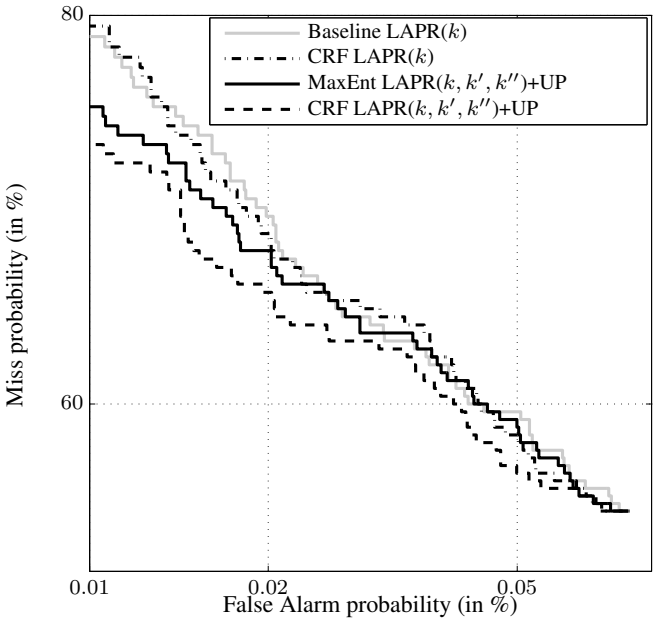

Fig. 1. DET curves showing performance of CRF-based confidence estimation applied in the word-level KWS system.

The DET plot in Figure 1 shows the results of various experiments in applying the CRF model to assign confidence scores in the word-based system. As relatively few key term hypotheses are generated by this system, the range of possible operating points is limited. Spline feature functions were applied to the continuous features in all CRF experiments. Five evenly-spaced knot points are used in the cubic spline approximation, as this was found to yield the best performance. In the baseline system, the confidence score is calculated by taking the product of the posteriors (LAPR(k)) for words making up a key term. It can be seen from the plot that using the LAPR(k) feature in isolation with the CRF model yields performance similar to that of the baseline. This is however not surprising as no additional information is incorporated in the model to improve the confidence scores. Incremental improvements are observed when adding features to the model, but are not shown in 1 for clarity. The best performance is achieved using the combination of LAPR(k), the unigram prior (UP) and the contextual posteriors

(LAPR(k') and LAPR(k'')). The results using this configuration are shown in Figure 1. The sequential CRF approach is shown to outperform an equivalent non-sequential maximum entropy (MaxEnt) model which makes use of the same features. The inclusion of the LITM1 feature functions do not improve performance in this system. This is due to the fact that there are very few training examples for each key term, such that generalisation becomes an issue.

5.2. Grapheme-Based System Results

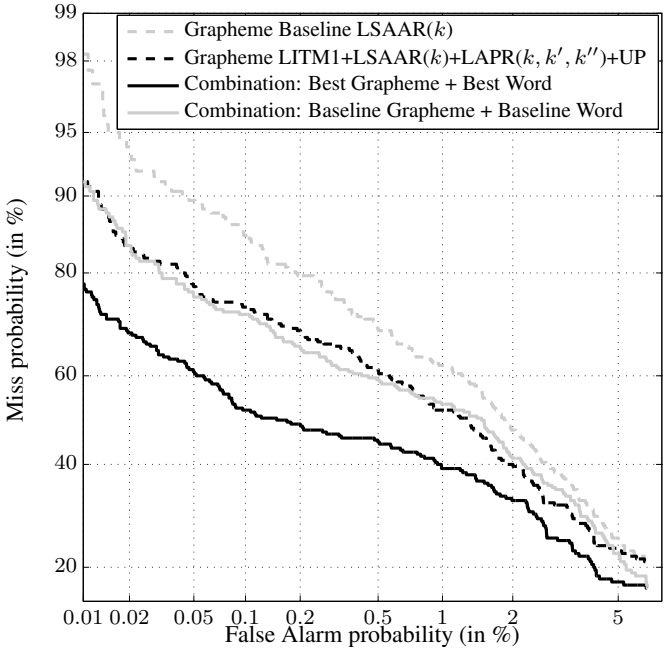

Fig. 2. DET curves showing performance of CRF-based confidence estimation in grapheme-based KWS, as well as in the combined word and grapheme-based KWS systems.

In the grapheme-based system, a significant number of key term hypotheses are generated. It is therefore possible to apply the literal moment (LITM1) feature functions here, as there is significant training data to reliably estimate the parameters for individual key terms. The continuous feature related to the LITM1 feature functions, for which moment statistics are estimated, is the acoustic score ratio (LSAAR(k)). First-order moment parameters are estimated for all key terms which occur at least 100 times in the training data, of which there are 197. An additional "out-of-shortlist" parameter is estimated to cover the remaining 22 less frequent key terms.

The results for grapheme-based KWS experiments are presented in the DET plot of Figure 2. As this system is able to achieve a wider range of operating points than the word-based system, this DET extends beyond the 4% false alarm rate of interest. The acoustic score ratio features (LSAAR(k)) are used as the confidence measure in the baseline system. Using this feature in isolation with the CRF model results in performance similar to the baseline, which is to be expected. However, incremental performance gains are achieved when adding more features to the model. A configuration which uses the extended model with LITM1 feature functions, and combines LSAAR(k), LAPR(k), contextual posterior features (LAPR(k') and LAPR(k'')) and the unigram prior (UP) yields the best performance. The results using this system configuration are shown in Figure 2. The contextual posteriors proved to be less useful here than in the word-based system. An explanation for this is that the graphemes used to form a key term hypothesis in the sub-word system can form part of any word. The confidence in the context of these parent words therefore has no real bearing on the quality of the key term hypothesis. The performance gains evident in the figure for the best configuration are therefore primarily due to the LITM1 feature functions. This result highlights the importance of score normalisation across key terms, which is particularly useful in this system.

5.3. Combined System Results

Results showing the effect of combining the outputs of the word and grapheme-based systems in the hybrid KWS configuration are also shown in Figure 2. The baseline for these comparisons is formed by combining the individual word and grapheme-based system baselines. These confidence scores are therefore not mapped before combination. After combination, this baseline performance is improved over a grapheme-based system used in isolation. This is a result of the additional true positives gained from the word-based system, at a cost of relatively few false alarms. Combining systems for which confidence scores have been estimated using the approach presented in this work however, yields greatly improved performance. The DET plot for the combined CRF system in Figure 2 corresponds to the combination of the best word and grapheme-based systems. The individual performance of these systems is improved over their respective baselines to begin with. However, when these systems employing CRF-based confidence estimation are combined, further gains are achieved. This is due to the fact that the assigned confidence scores in each system are effectively normalised. The key term hypothesis scores generated by these different systems can thus be treated equally for decision making purposes. The performance of the word-based system is therefore maintained in the combined configuration, with the grapheme-based system contributing many additional key term hypotheses, extending the operating range to desired levels. This results in a consistent relative improvement over the baseline of 26% in miss probability at false alarm rates of both 0.2% and 4%, as well as a relative improvement of 36% in the false alarm rate at a miss probability of 30%.

6. CONCLUSION

A CRF-based approach for estimating confidence scores of key terms in KWS was investigated in this work. This was shown to yield promising results, particularly when applied to systems that are combined in a hybrid KWS configuration. In the approach taken, multiple discriminative features are combined using a CRF model to estimate accurate confidence scores. Amongst these features, the contextual posteriors and unigram priors introduced here proved to be highly informative. The sequential nature of the CRF model used was shown to yield gains over a comparable non-sequential model, thus supporting its use in this task. An elegant approach to the problem of score normalisation across keywords, which makes use of key term-specific CRF feature functions was presented. This was also shown to improve KWS performance in the grapheme-based KWS system. Overall, in both the word and grapheme-based KWS systems, the techniques described in this work resulted in notable performance gains. Furthermore, large improvements were achieved in the combined (hybrid) system. This is because accurate confidence scores are estimated for each system using essentially the same technique, which results in implicit inter-system normalisation of the confidence scores.

7. REFERENCES

[1] 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*. 2007, pp. 314–317, ISCA.

[2] Jonathan Mamou, Bhuvana Ramabhadran, and Olivier Siohan, "Vocabulary independent spoken term detection," in *Proc. SIGIR*. 2007, pp. 615–622, ACM.

[3] Dong Wang, Javier Tejedor, Joe Frankel, Simon King, and Jos Cols, "Posterior-based confidence measures for spoken term detection," in *Proc. ICASSP*. 2009, pp. 4889–4892, IEEE.

[4] Javier Tejedor, Doroteo Torre Toledano, Miguel Bautista, Simon King, Dong Wang, and Jos Cols, "Augmented set of features for confidence estimation in spoken term detection," in *Proc. Interspeech*. 2010, pp. 701–704, ISCA.

[5] Carolina Parada, Mark Dredze, Denis Filimonov, and Frederick Jelinek, "Contextual information improves OOV detection in speech," in *Proc. HLT-NAACL*. 2010, pp. 216–224, The Association for Computational Linguistics.

[6] Daniel Schneider, Timo Mertens, Martha Larson, and Joachim Khler, "Contextual verification for open vocabulary spoken term detection," in *Proc. Interspeech*. 2010, pp. 697–700, ISCA.

[7] Taichi Asami, Narichika Nomoto, Satoshi Kobashikawa, Yoshikazu Yamaguchi, Hirokazu Masataki, and Satoshi Takahashi, "Spoken document confidence estimation using contextual coherence," in *Proc. Interspeech*. 2011, pp. 1961–1964, ISCA.

[8] Hung yi Lee, Tsung wei Tu, Chia-Ping Chen, Chao-Yu Huang, and Lin-Shan Lee, "Improved spoken term detection using support vector machines based on lattice context consistency," in *Proc. ICASSP*. 2011, pp. 5648–5651, IEEE.

[9] Dimitra Vergyri, Izhak Shafran, Andreas Stolcke, Venkata Ramana Rao Gadde, Murat Akbacak, Brian Roark, and Wen Wang, "The SRI/OGI 2006 spoken term detection system.," in *Proc. Interspeech*. 2007, pp. 2393–2396, ISCA.

[10] Bing Zhang, Richard Schwartz, Stavros Tsakalidis, Long Nguyen, and Spyros Matsoukas, "White listing and score normalization for keyword spotting of noisy speech," in *Proc. Interspeech*. 2012, ISCA.

[11] Rong Zhang and Alexander I. Rudnicky, "Word level confidence annotation using combinations of features," in *Proc. Interspeech*. 2001, pp. 2105–2108, ISCA.

[12] Dong Wang, Javier Tejedor, Simon King, and Joe Frankel, "Term-dependent confidence normalisation for out-of-vocabulary spoken term detection," *J. Comput. Sci. Technol.*, vol. 27, no. 2, pp. 358–375, 2012.

[13] David A. James, "A system for unrestricted topic retrieval from radio news broadcasts," in *Proc. ICASSP*. 1996, pp. 279–282, IEEE.

[14] Beth Logan, Pedro Moreno, and Om Deshmukh, "Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio," in *Proc. HLT*, 2002, pp. 31–35.

[15] Peng Yu, Kaijiang Chen, Chengyuan Ma, and Frank Seide, "Vocabulary-independent indexing of spontaneous speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5-1, pp. 635–643, 2005.

[16] Matthew S. Seigel and Philip C. Woodland, "Combining information sources for confidence estimation with crf models," in *Proc. Interspeech*. 2011, pp. 905–908, ISCA.

[17] Gunnar Evermann and Philip C. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Proc. ICASSP*. 2000, pp. 1655–1658, IEEE.

[18] Wai Kit Lo, F.K. Soong, and S. Nakamura, "Generalized posterior probability for minimizing verification errors at subword, word and sentence levels," in *Proc. ISCSLP*, 2004, pp. 13–16.

[19] Yi-Cheng Pan, Hung lin Chang, Berlin Chen, and Lin-Shan Lee, "Subword-based position specific posterior lattices (s-pspl) for indexing speech information," in *Proc. Interspeech*. 2007, pp. 318–321, ISCA.

[20] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, 2001, pp. 282–289.

[21] Dong C. Liu and Jorge Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1-3, pp. 503–528, Aug. 1989.

[22] Dong Yu, Li Deng, and Alex Acero, "Using continuous features in the maximum entropy model," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1295–1300, 2009.

[23] Mark J. F. Gales and Federico Flego, "Model-based approaches for degraded channel modelling in robust ASR," in *Proc. Interspeech*. 2012, ISCA.