Keyword Spotting Based on Hypothesis Boundary Realignment and State-Level Confidence Weighting

Hong Liu
Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University
China
hongliu@pku.edu.cn

Yuezhao Chen*
Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University
China
yzchen@sz.pku.edu.cn

Runwei Ding
Shenzhen Touching AI Technologies Co., Ltd
China
dingrunwei@pkusz.edu.cn

Cheng Pang
Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University
China
chengpang@sz.pku.edu.cn

ABSTRACT

Keyword[1] spotting (KWS) deals with the identification of keywords in speech utterances. A two-stage approach is often used for the flexibility and high efficiency. The two stages are keyword hypotheses detection stage and hit or false-alarm verification stage in sequence. How to reduce the false-alarms is a key and difficult problem in the verification stage, which is formatted as the confidence measure (CM) problem. In this paper, a novel keyword-filler hidden Markov model (HMM) based method is proposed based on two improved approaches. On one hand, for more effective confidence measure, a hypothesis boundary realignment method is used to gain more precise hypothesized segments for possible keyword. Then an overlap ratio criterion is defined to evaluate this process. On the other hand, a state-level confidence weighting method is proposed to improve the posterior probability based CM. Experiments show that either improvement is effective, and the proposed method based on the two processes gives the best performance.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; *Specialized information retrieval*; Multimedia and multimodal retrieval; Speech/audio search; • **Computing methodologies** → **Artificial intelligence**; *Natural language processing*; Speech recognition

KEYWORDS

Keyword spotting, keyword verification, confidence measure, hypothesis boundary realignment, state-level confidence weighting

1 INTRODUCTION

Keyword spotting [1, 2] is to detect the occurrences of predefined keywords in continuous speech stream. There are kinds of applications using this technology, such as speech surveillance [3], spoken document retrieval [4], voice interaction [5, 6], etc.

In the past decades, many methods have been proposed for KWS, which can be clustered to two groups. One is the large vocabulary continuous speech recognition (LVCSR) based KWS, which commonly assumes offline processing of audio stream to generate word lattices [7-10] or sub-word lattices [9, 10]. It is usually used to search audio documents which are indexed based on the generated lattices. This kind of systems often require a large amount of training data, which is only available for resource-rich languages, but a new trend in KWS is to build an efficient system for resource-limited languages [11]. Besides, due to the restricted vocabulary, handling out-of-vocabulary (OOV) words is a difficult problem [12]. On the other hand, the keyword-filler HMM based KWS [1, 2] requires little or no training data, and has the flexibility of selecting keywords set. This kind of systems use a vocabulary which contains the keywords to be spotted and the fillers to absorb the non-keyword speech [13, 14]. For example, phone-loop filler is often a simple and efficient filler model.

However, phone-loop based keyword-filler KWS has a major problem of high false-alarm rate. In fact, due to the use of the same phone models in the keyword part and filler part, the filler model can potentially model a phoneme sequence corresponding to any word. After generating enough keyword hypotheses in the initial detection stage, a verification stage is necessary to

CSAE '18, October 22–24, 2018, Hohhot, China
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6512-3/18/10...$15.00
https://doi.org/10.1145/3207677.3277918

reduce the false-alarms. For each keyword hypothesis, a confidence score is given to estimate the probability of occurrence of the corresponding keyword. Confidence measure is the key problem in the verification stage. Traditional likelihood ratio (LR) [15], which means the ratio of hypothesized segment's likelihood for the keyword to likelihood for the non-keyword, is often used. Though many methods have been proposed for the non-keyword modelling, such as online dynamic filler model [16] and anti-subword model [17], the non-keyword modelling is still a difficult problem. However, non-keyword model is not necessary for the later frame-level posterior probability based confidence measure [18], which even gives a better performance than LR-based methods.

In this paper, a novel method based on hypothesis boundary realignment and state-level confidence weighting is proposed to improve the performance of confidence measure. On one hand, the hypothesis boundary realignment network is proposed to obtain more precise keyword boundary. For performance evaluation purpose, the overlap ratio between cross duration and the total duration of a hypothesis and its reference is defined as the matching criterion. On the other hand, as an improved posterior probability based CM, the state-level confidence weighting approach is proposed to estimate the reliability of each keyword hypothesis. Combination of the two approaches is natural and effective, which can achieve better KWS performance. The framework of the proposed KWS system is shown in Fig. 1, which mainly includes detection and verification stages.

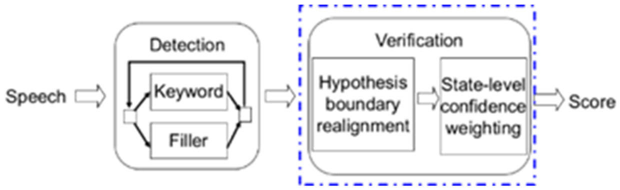
Speech
Detection
Keyword
Filler
Verification
Hypothesis boundary realignment
State-level confidence weighting
Score

Figure 1: The framework of the proposed KWS system.

Overview: The KWS result data is produced by the keyword-filler HMM based detector. From the decoding sequence, multiple hypotheses for each keyword are generated. Then each hypothesis is processed by the proposed boundary realignment method. Based on the hypothesis with more precise boundaries, a confidence score is then calculated by the state-level weighting based CM method. At last, based on the list of keyword hypotheses with corresponding confidence scores, overall performance of the KWS system is evaluated on the MTWV metric. The TIMIT corpus [19] is used for evaluation, which contains 6300 English sentences. The HTK toolkit [20] is used in the extraction of acoustic features, training of HMM models and Viterbi decoding.

2 BASELINE KWS SYSTEM

2.1 Keyword-Filler HMM based Keyword Detection

The keyword-filler HMM based detector is adopted in the KWS system. The filler model is the phone-loop HMM. The posterior probability-based CM is the baseline in the verification stage.

The framework of keyword-filler HMM [2] is shown in Fig. 2. The keyword model is constructed as the connected phones of the keyword pronunciation. The filler model is used to absorb all non-keyword speech or silence segments. There are several choices, such as phone-loop network (i.e. parallel connection of phones), LVCSR network without the current keyword. However, the LVCSR-based approach requires a higher computational cost and a larger memory requirement. In this paper, the phone-loop based filler model is used.

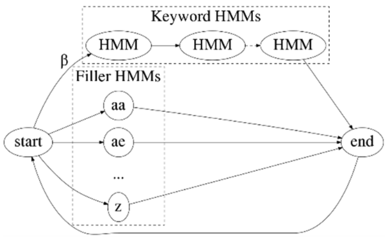

Figure 2: The framework of keyword-filler HMM for keyword detection.

During the detection stage, Viterbi decoding is used to select the best path, which gives the keyword and filler model sequence. By backtracking the decoding path, hypothesized keyword segments are produced. A keyword hypothesis is a positive sample, or hit, if the start and end times of hypothesis lie either side of the mid-point of an identical label in the reference; otherwise it is a negative sample, or false-alarm. The trade-off between detection rate and false-alarm rate is achieved by adjusting the parameter β, which is the filler-to-keyword transition probability. To achieve a high detection rate, a large enough value of β is necessary.

2.2 Posterior Probability based Confidence Measure

For each keyword hypothesis, a posterior probability based confidence score is calculated. The procedure is conducted at three levels, namely frame level, phone level and word level. At the frame level, the posterior probability of state s given the observation o_t at frame t, is calculated by

$$p(s \mid o_t) = \frac{p(o_t|S)p(s)}{\sum_{i=1}^{N_s} p(o_t|S_i)p(s_i)} \qquad (1)$$

where $p(o_t|s_i)$ is the likelihood of observation o_t at state s_i, $p(s_i)$ is the prior probability of state s_i, which is assumed to be equal with each other, and N_s is the number of all states.

For the phone-level and word-level confidence measures, the confidence score is estimated as some kind of average value of the confidence scores in previous level respectively. The confidence score of the phone *ph* with aligned observations from frame *a* to frame *b* is calculated by

$$CM(ph) = \frac{1}{b-a+1}\sum_{t=a}^{b} log\big(p(s_t \mid o_t)\big) \qquad (2)$$

where s_t denotes the aligned state at frame *t*.

Though several phone-level confidence weighting methods [21-23] are proposed to calculate the word-level confidence, they mainly depend on both the spotted keywords and the speech utterances database. The arithmetic average of phone confidences is adopted in the experiments. The confidence score of word w based on the hypothesized speech segment aligned with phone sequence $\{ph_1, ph_2, \ldots, ph_{N_{ph}^w}\}$ is

$$CM(w) = \frac{1}{N_{ph}^w}\sum_{i=1}^{N_{ph}^w} CM(ph_i) \qquad (3)$$

where N_{ph}^w is the phone number of pronunciation for word *w*.

3 HYPOTHESIS BOUNDARY REALIGNMENT

Due to the high transition probability *β*, many negative hypotheses, i.e. false alarms, are usually made in the detection stage. It should be noted that the positive hypotheses, i.e. hits, usually have longer durations than that of the transcription reference. It is the non-keyword deletion problem from the non-keyword's view. The problem was caused by the similar reason as the phone deletion error in phone recognition, which is the insertion penalty inter models. However, in the keyword-filler HMM network, these non-keyword segments happen to be absorbed by the keyword HMM, causing longer hypothesis duration than the ground truth. With the biased hypothesis boundaries, confidence measure cannot be reliably achieved. To solve the problem, the hypothesis boundary realignment is proposed to extract the true or more precise keyword boundaries. The network of hypothesis boundary realignment is shown in Fig. 3.

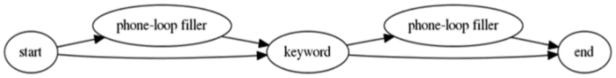

Figure 3: The network of hypothesis boundary realignment.

Two alternative phone-loop filler components are connected with the keyword HMM in the beginning and ending positions. In the Viterbi decoding stage, the hypothesized speech segment is passed through the realignment network, producing a realigned keyword hypothesis with more compact boundaries. For example, the processing result of a positive sample for keyword 'always' is shown in Fig. 4.

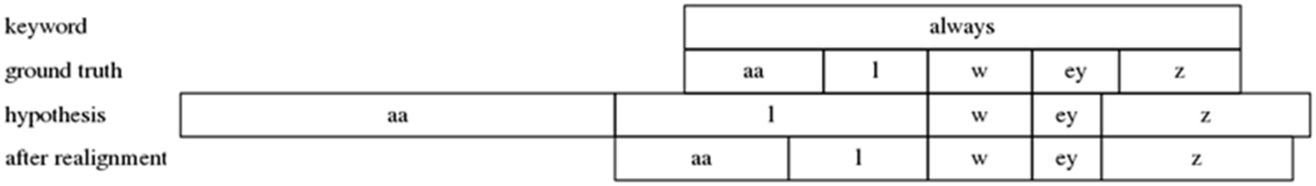

Figure 4: Example of boundary realignment. The phonetic transcriptions with time spans are schematically depicted in the figure. The width of the unit is proportional to the duration.

To validate the effectiveness of the proposed method, the matching degree between the hypothesis and the reference is defined as the overlap ratio:

$$OverlapRatio(ref, hyp) = \frac{CrossDuration(ref,hyp)}{TotalDuration(ref,hyp)} \qquad (4)$$

with

$$CrossDuration(ref, hyp) = min(E(ref), E(hyp)) - max(S(ref), S(hyp)) + 1 \qquad (5)$$

$$TotalDuration(ref, hyp) = max(E(ref), E(hyp)) - min(S(ref), S(hyp)) + 1 \qquad (6)$$

where S(•) and E(•) mean the indexes of the beginning frame and ending frame respectively. The higher overlap ratio for a hypothesis means the more precise boundaries. Through realigning the hypothesis, the boundaries will become more precise, as shown in the experiment section. With the more precise hypothesis boundary, the CM in the proposed method works more reliably.

4 STATE-LEVEL CONFIDENCE WEIGHTING

Though the phone-level confidence measure is directly derived from frame-level local posterior probability, Eq. (2) can be rewritten as duration weighted state-level confidences.

$$CM^{'}(ph) = \frac{\sum_{i=1}^{N_s^{ph}} CM(s_i)T(s_i)}{\sum_{i=1}^{N_s^{ph}} T(s_i)} \qquad (7)$$

where N_s^{ph} is the number of states in HMM model of phone *ph*, which is set to *3* for all phones in the experiments, $T(s_i)$ is the frame number for state s_i, and $CM(s_i)$ denotes the confidence score for state s_i, aligned from frame a_i to frame b_i:

$$CM(s_i) = \frac{1}{b_i - a_i + 1} \sum_{t=a_i}^{b_i} \log\big(p(s_i \mid o_t)\big) = \frac{1}{T(s_i)} \sum_{t=a_i}^{b_i} \log\big(p(s_i \mid o_t)\big) \quad (8)$$

The Eq. (7) shows that the baseline method only takes the state duration as weight in the derivation of the phone-level confidence. Assuming that each state contributes to the phone confidence in different degrees, apart from the duration, a weight coefficient is added to each state:

$$ICM(ph) = \frac{\sum_{i=1}^{N_s^{ph}} CM(s_i) T(s_i) w_i}{\sum_{i=1}^{N_s^{ph}} T(s_i)} \quad (9)$$

where w_i is the contribution weight of state s_i. If each weight w_i is assigned to 1, the proposed method becomes the baseline form as shown in Eq. (2).

The advantage of this kind of weighting is that it works at state level, for each phone, so no further adjustment is needed for the new keywords to be spotted. This method works well for both phone verification and keyword verification, which are validated in the experiment section, respectively.

To obtain the weights, Viterbi algorithm is used for phone-loop decoding on the training set. Then, for each phone *ph*, the positive hypotheses set H_{ph}^{+} and negative hypotheses set H_{ph}^{-} are collected by checking the decoding sequence and the phone-level transcription. Based on that, the optimization target is to maximize the area under curve (AUC) of receiver operating characteristics (ROC) curve. The AUC for a phone *ph* can be approximated as Wilcoxon-Mann-Whitney statistic [24]

$$\hat{A}\big(ph\big) = \frac{\theta_{max}\big(ph\big)}{|H_{ph}^{+}||H_{ph}^{-}|} \Sigma_{u \in H_{ph}^{+}} \Sigma_{v \in H_{ph}^{-}} S\Big(ICM\big(u\big) - ICM\big(v\big)\Big) \quad (10)$$

where | • | means the cardinality of a set, $\theta_{max}(ph)$ is the callback rate for the phone *ph*, *ICM(u)* denotes the confidence for the phone given hypothesis *u*, $S(\bullet)$is the sigmoid function. The objective function is optimized by the generalized probabilistic descent algorithm. The vector w^{*}corresponding to the optimal target gives the learned weights for the current phone.

5 EXPERIMENTAL RESULTS

5.1 Experimental Setup

TIMIT contains 6300 English sentences. Since this database only includes training and test sets, it is redivided into training, development and test sets. A small part of the original training set is used as the development set (400 utterances, 0.34 hours). The left part of the original training set is the training set (3296 utterances, 2.79 hours). The original test set is taken as the test set (1344 utterances, 1.14 hours). The dialect utterances (the SA sentences) are not used.

The acoustic features are 12-dimension mel-frequency cepstrum coefficients (MFCCs) plus energy and their first and second time derivatives. The training set is used to train the monophone HMMs. There are three states per phone, and 40 Gaussians per states after the optimization on the development set. The CMU/MIT phone set [25], which contains 39 phonemes, is used for HMM models training, and these phones are used in keyword model and phone-loop based filler model.

5.2 Evaluation Metric

Fifteen words in the TIMIT vocabulary are selected as the keywords, which are shown in Table 1.

Table 1: Keywords Set Used in KWS Experiments

after	always	before	these
without	please	money	also
began	dirty	forces	morning
only	overalls	small	-

KWS performance measure is based on Term-Weighted Value (TWV) [26, 27], which is a linear combination of the probability of missed detections and the probability of false alarms:

$$TWV(\theta) = 1 - [P_{Miss}(\theta) + \beta \cdot P_{FA}(\theta)] \quad (11)$$

where β is a constant set to 999.9, and θ is the threshold to determine a hit or a miss. TWVs for each keyword are averaged to yield actual TWV (ATWV). Maximum TWV (MTWV) is the best TWV after a search over all possible thresholds.

5.3 Hypothesis Boundary Realignment

Fig. 5 shows the evaluation of the realignment procedure by the overlap ratio criterion which is defined in Eq. (4). For the words 'forces' and 'overalls', the overlap ratio becomes lower. By investigating the samples, the main reason is that the last phone 'z' is mistaken as 's' in non-ignorable amount of the hits, which are taken as pronunciation error in this paper (Maybe seen pronunciation variation problem in other research, but not here.). Though lower overlap ratio after the realignment procedure, the realigned hypotheses match better with the standard pronunciation so that a higher confidence can be produced. For the word 'small', the overlap ratio change is negligible, because of the consonants in the starting and ending position, the hypotheses boundaries directly from the detection stage are fairly precise. And the proposed method improves the overlap ratio of the hits for each other keyword. Overall, the proposed boundary realignment method solves the boundary bias problem and obtains better performance.

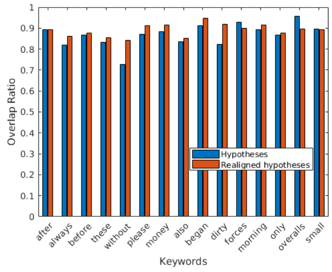

Figure 5: Overlap ratio of the total duration.

5.4 State-Level Confidence Weighting based Phone Verification

Table 2: Performance Comparison for Phone Confidence Measure

Phone	Verification Performance (AUC, %)		
	Baseline CM	State-level confidence weighting based CM	Improvement
aa	75.04	75.43	0.39
ae	73.16	73.35	0.19
ah	64.68	64.75	0.07
aw	80.66	81.60	0.94
ay	78.48	78.89	0.41
b	70.08	71.43	1.35
ch	80.90	82.37	1.47
d	74.38	75.39	1.01
dh	71.69	72.66	0.97
dx	80.95	81.13	0.18
eh	68.20	68.46	0.26
er	74.22	74.50	0.28
ey	75.42	75.51	0.09
f	83.34	83.35	0.01
g	78.69	78.89	0.20
hh	80.79	80.88	0.09
ih	68.22	68.30	0.08
iy	74.05	75.08	1.03
jh	83.75	84.66	0.91
k	79.95	80.59	0.64
l	72.49	72.49	0.00
m	78.32	78.47	0.15
n	76.30	76.69	0.39
ng	80.00	80.54	0.54
ow	76.12	76.32	0.20
oy	88.44	89.13	0.69
p	79.35	79.93	0.58
r	72.12	72.62	0.50
s	72.08	72.93	0.85
sh	85.64	86.41	0.77
t	78.92	80.00	1.08
th	77.46	78.03	0.57
uh	80.36	81.91	1.55
uw	74.82	75.40	0.58
v	77.57	77.87	0.30
w	76.56	76.93	0.37
y	77.09	77.30	0.21
z	77.41	78.50	1.09
Average	76.78	77.33	0.55

Phone recognition is often used to build sub-word based speech database index. The verification for each phone is independent, which is more appropriate to evaluate the state-level confidence weighting method. In this experiment, the speech data of test set is used for phone recognition. After the recognition stage, all the phones are evaluated respectively. For the hypotheses of one phone, each one is given a confidence score. And each hypothesis can be labelled as hit or false-alarm based on the reference.

For each phone, the posterior probability and the proposed state-level confidence weighting based CM methods are compared. The results for the test set are shown in Table 2. From the results, it can be seen that the proposed CM method performs better for each phone, and the averaged AUC improvement is 0.55%. Though the averaged improvement is not huge, it should be noted that the phone speech unit is short, which means fewer cues available for reliable confidence estimation and higher confusable possibility with each other. Moreover, a phoneme is the basic unit of a word or phrase. The improvement on the basic unit commonly means more advancement can be accumulated in a higher level unit, as will be shown in the following KWS experiment.

5.5 KWS Performance

In addition to the baseline system, the two proposed methods are also evaluated. In fact, the proposed two approaches above is not conflicting, they can be composed by first hypothesis boundary realignment to reach more precise keyword position and followed state-level confidence weighting based CM to achieve better verification performance.

Table 3: Performance Comparison of The Baseline And The Proposed Kws Systems at Different Optimization Levels

KWS Systems	Overall Performance (MTWV)
Baseline	0.3814
Baseline + (Hypothesis boundary realignment) P1	0.4010
Baseline + (State-level confidence weighting) P2	0.3960
Proposed Method (Baseline + P1 + P2)	0.4066

The comparison results of the three above methods are shown in Table 3. In the baseline system, the posterior probability based CM is used. With the boundary realignment procedure (P1) applied, the MTWV is improved by 0.0196. On the other hand, the state-level confidence weighting method (P2) is effective solely, without P1. Compared to the baseline system, 0.0146 improvement is achieved. The last system uses both the approaches, P1 and P2, and its performance is the best, with 0.0252 improvement on MTWV.

6 CONCLUSIONS

A novel keyword-filler KWS system based on two improvement approaches is proposed. The hypothesis boundary realignment is first used to extract more precise keyword position from an initial hypothesis. The overlap ratio criterion is defined to estimate the matching degree between a keyword hypothesis with the corresponding reference. Besides, the state-level confidence weighting is proposed to improve the confidence measure, which not only produces better performance on the phone verification, but also brings a significant improvement for the KWS system. The keyword-filler HMM based keyword spotting system is promising for its flexibility and effectiveness properties. In the future work, how to make use of more speech context or other prior knowledge in either keyword detection or verification stage is worth exploring.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (NSFC, No. 61673030, U1613209), Natural Science Foundation of Guangdong Province (No. 2015A030311034), Specialized Research Fund for Strategic and Prospective Industrial Development of Shenzhen City (No. ZLZBCXLJZI20160729020003), Scientific Research Project of Shenzhen City (CKCY 2017050810242781).

REFERENCES

[1] R. C. Rose and D. B. Paul. 1990. A hidden Markov model based keyword recognition system. In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 1, 129–132.
[2] J. R. Rohlicek, W. Russell, S. Roukos, and H. Gish. 1989. Continuous hidden Markov modeling for speaker-independent word spotting. In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 1, 627–630.
[3] R. L. Warren. 2001. Broadcast speech recognition system for keyword monitoring. U.S. Patent 6332120 B1.
[4] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees. 2000. The TREC spoken document retrieval track: a success story. In *Text Retrieval Conference*, NIST, 26, 1–20.
[5] G. G. Chen, C. Parada, and G. Heigold. 2014. Small-footprint keyword spotting using deep neural networks. In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 4087–4091.
[6] A. H. Michaely, X. D. Zhang, G. Simko, C. Parada, and P. Aleksic. 2017. Keyword spotting for Google assistant using contextual speech recognition. In *Automatic Speech Recognition and Understanding Workshop*, IEEE, 272-278.
[7] D. R. H. Miller, M. Kleber, C. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish. 2007. Rapid and accurate spoken term detection. In *International Conference of the Speech Communication Association*, 314–317.
[8] J. Mamou, B. Ramabhadran, and O. Siohan. 2007. Vocabulary independent spoken term detection. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 615–622.
[9] D. Vergyri, I. Shafran, A. Stolcke, R. R. V. Gadde, M. Akbacak, B. Roark, and W. Wang. 2007. The SRI/OGI 2006 spoken term detection system. In *International Conference of the Speech Communication Association*, 2393–2396.
[10] V. T. Pham, H. H. Xu, X. Xiao, N. F. Chen, E. S. Chng, and H. Z. Li. 2016. Keyword search using query expansion for graph-based rescoring of hypothesized detections. In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 6035–6039.
[11] T. Alume, D. Karakos, W. Hartmann, R. Hsiao, L. Zhang, L. Nguyen, S. Tsakalidis, and R. Schwartz. 2017. The 2016 BBN Georgian telephone speech keyword spotting system. In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 5755–5759.
[12] D. Karakos and R. M. Schwartz. 2015. Combination of search techniques for

improved spotting of OOV keywords. In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 5336–5340.

[13] I. Szöke. 2010. Hybrid word-subword spoken term detection.

[14] K. M. Knill and S. J. Young. 1996. Fast implementation methods for Viterbi-based word-spotting. In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 522–525.

[15] M. Weintraub. 1995. LVCSR log-likelihood ratio scoring for keyword spotting. In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 1, 297–300.

[16] H. Bourlard, B. D'Hoore, and J. M. Boite. 1994. Optimizing recognition and rejection performance in wordspotting systems. In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 1, I/373–I/376.

[17] R. A. Sukkar and C. H. Lee. 1996. Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(6), 420-429.

[18] S. Abdou and M. S. Scordilis. 2004. Beam search pruning in speech recognition using a posterior probability-based confidence measure. *Speech Communication*, 42(3), 409–428.

[19] W. E. Fisher, G. R. Doddington, and K. M. Goudle-Marshall. 1986. The DARPA speech recognition research database: specifications and status. *CMU Arctic Speech Databases for Speech Synthesis Research.*

[20] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2009. The HTK Book (for HTK version 3.4.1), http://htk.eng.cam.ac.uk: Cambridge University.

[21] J. Liang, M. Meng, X. R. Wang, and P. Ding. 2006. An improved Mandarin keyword spotting system using MCE training and context-enhanced verification. In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 1145–1148.

[22] H. Y. Li, J. Q. Han, and T. R. Zheng. 2011. AUC optimization based confidence measure for keyword spotting. In *International Conference of the Speech Communication Association*, 1917–1920.

[23] Y. C. Liu, M. X. Xu, and L. H. Cai. 2014. Improved keyword spotting system by optimizing posterior confidence measure vector using feed-forward neural network. In *International Joint Conference on Neural Networks*, IEEE, 2036–2041.

[24] C. Cortes and M. Mohri. 2004. Confidence intervals for the area under the ROC curve. In *International Conference on Neural Information Processing Systems*, MIT Press, 305–312.

[25] K. F. Lee and H. W. Hon. 1989. Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics Speech and Signal Processing*, 37(11), 1641–1648.

[26] J. G. Fiscus, J. G. Ajot, J. Garofalo, and G. Doddington. 2007. Results of the 2006 spoken term detection evaluation. In *SIGIR Workshop on Searching Spontaneous Conversational Speech*, 51–57.

[27] J. Cui, X. D. Cui, B. Ramabhadran, J. Kim, B. Kingsbury, J. Mamou, L. Mangu, M. Picheny, T. N. Sainath, and A. Sethy. 2013. Developing speech recognition systems for corpus indexing under the IARPA Babel program. In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 6753–6757.