



Tunable Keyword-Aware Language Modeling and Context Dependent Fillers for LVCSR-based Spoken Keyword Search

Tze Siong Lau¹, I-Fan Chen², and Chin-Hui Lee²

¹ Temasek Lab@NTU, Nanyang Technological University, Singapore

² School of Electrical and Computer Engineering, Georgia Institute of Technology

noiseztual@gmail.com, ichen8@gatech.edu, chl@ece.gatech.edu

Abstract

We explore the potential of using keyword-aware language modeling to extend the ability of trading higher false alarm rates in exchange for lower miss detection rates in LVCSR-based keyword search (KWS). A context-dependent keyword language modeling method is also proposed to further enhance the keyword-aware language modeling framework by reducing the number of false alarms often sacrificed in order to achieve the desirable low miss detection rates. We demonstrate that by using keyword-aware language modeling, a KWS system is able to achieve different operating points (misses vs. false alarms) by tuning a parameter in language modeling. We observe a relative gain of 20% in actual term weighted value (ATWV) performance with the keyword-aware KWS systems over the conventional LVCSR-based KWS systems when testing on the English Switchboard data. Moreover the proposed context-dependent keyword language modeling could further achieve a 9% relative ATWV improvement over the original keyword-aware KWS systems for single-word keywords which cause the most false alarms.

Index Terms: keyword spotting, keyword search, spoken term detection, grammar network, context-dependent

1. Introduction

Keyword spotting (KWS) [1, 2] is a task of detecting a set of preselected keywords in continuous speech. The technology has been widely used in various applications, such as spoken term detection [3-6], spoken document indexing and retrieval [7], speech surveillance [8], spoken message understanding [9, 10], etc. Each of these applications has a different tolerance to missed detections and false alarms. Therefore, it is important for KWS systems to be able to trade higher false alarm rates for lower miss detection rates to achieve optimal performance for different applications. In general, most existing KWS systems fall into one of two categories: classic keyword-filler based [1, 2] and large vocabulary continuous speech recognition (LVCSR) based KWS [3-6] systems.

In the classic keyword-filler based KWS systems, speech inputs are decoded into sequences of keywords and non-keywords (often referred to as fillers) [1, 2] with a simple keyword-filler loop grammar (Figure 1 (a)). Though the grammar sometimes tends to generate a great amount of false alarms due to the over simplification of the speech sentence structure, it provides a great freedom for the KWS systems to adjust their operating points. By setting the prior probabilities of the keywords and fillers in the grammar, the tradeoff between missed detections and false alarms of the keywords can be controlled directly. This allows the keyword-filler based KWS systems easily trading higher false alarm rates in

exchange for lower miss detection rates to achieve the desired performance for the target applications.

The LVCSR-based KWS systems, however, do not have such a flexible mechanism for operating points tuning. Instead of assuming the input speech is a sequence of keywords and non-keywords (fillers), LVCSR-based KWS systems convert input speech into text documents using speech-to-text (STT) techniques with n -gram language models first. The systems then search the text documents and return a scored list of putative hits for each target keyword. With better speech syntactic information captured in the n -gram language model grammars (Figure 1 (b)), the LVCSR-based KWS systems usually have better performance with much less false alarms than the classic keyword-filler based KWS systems [11]. And users are able to use a threshold to trade higher miss detection rates for lower false alarm rates. However, it is impossible for the users to lower the miss detection rate since it is bounded by the STT generated text documents. In other words, once the keywords are missed in the STT generated text documents, the KWS systems can no longer recover the miss. This makes LVCSR-based KWS systems more difficult to customize towards applications which require a lower miss detection rate and are tolerant to higher false alarm rates.

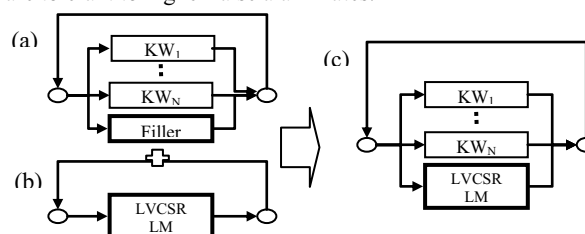


Figure 1. (a) Grammar for classic keyword-filler based KWS, (b) grammar for LVCSR-based KWS, and (c) the keyword-aware grammar, which combines the grammars used in the two KWS frameworks

In the references [12-14], we proposed keyword-aware language modeling approaches to alleviate the miss detection problem for the LVCSR-based KWS framework. By integrating the keyword-filler loop grammar into the n -gram LM grammar, we constructed the keyword-aware grammar (Figure 1 (c)) which allows the keyword probabilities in the LVCSR-based KWS systems to be boosted to reduce the miss detection rate. Experimental results on OpenKWS Vietnamese [12-14] and Tamil [13] tasks show the proposed language modeling approach could significantly improve KWS performances. In this study, we demonstrate that by using the keyword-aware language modeling, the tradeoff between missed detections and false alarms in LVCSR-based KWS systems can be easily controlled to achieve optimal

10.21437/Interspeech.2015-725

operating points for different applications. A context-dependent keyword language model (CD-KWLM), which is a revision of our previously proposed context-simulated keyword language model (CS-KWLM), is also presented to further reduce the false alarm rate of the keyword-aware LM based KWS systems across all operating points.

2. Keyword-Aware Language Modeling for LVCSR-based KWS Systems

In this section, we first review the keyword-aware language modeling technique. The CS-KWLM interpolation approach [12] is specifically focused here since it is very effective and easy to be realized to all the LVCSR-based KWS systems. The revision of the method is then presented.

2.1. Context-Simulated Keyword Language Model

In keyword-aware language modeling, boosting keyword prior probabilities to alleviate the miss detection problem of LVCSR-based KWS systems is of primary concern. In the CS-KWLM interpolation approach, the boosting effect is realized by interpolating the original n -gram LM with a keyword LM.

To build the keyword LM, in the CS-KWLM training text we put context terms before and after each keyword to simulate the situation that keywords appear in real sentences. Figure 2 illustrates such a training text setup. The context terms can be selected as bigrams or trigrams with high probabilities in the original language model of the LVCSR-based KWS system. Once the CS-KWLM is trained, a linear interpolation between the original language model and the CS-KWLM is performed to create the final language model for the KWS system:

$$P_{INT_LM}(w|h) = \alpha \cdot P_{CS-KWLM}(w|h) + (1-\alpha)P_{LM}(w|h) \quad (1)$$

In Eq. (1), α is an interpolation weight for the CS-KWLM and is ranged between 0 and 1. $P_{INT_LM}(w|h)$ is the interpolated probability of the CS-KWLM and the original language model for the n -gram (h,w) . Note that for the system keywords, $P_{LM}(w|h)$ is usually small compared to $P_{CS-KWLM}(w|h)$ due to the large vocabulary of the training corpus relative to the number of keywords. As α increases to 1, $P_{INT_LM}(w|h)$ increases from $P_{LM}(w|h)$ to $P_{CS-KWLM}(w|h)$ for any n -gram inside the keyword. This increase of keyword prior probabilities results in a reduction in the missed detections and a growth of false alarms of the KWS systems.

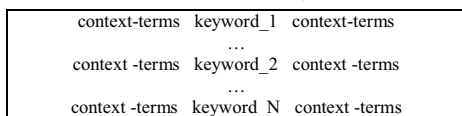


Figure 2. Illustration of the training text for context-simulated keyword language model (CS-KWLM)

2.2. Context-Dependent Keyword Language Model

Though the CS-KWLM method provides a great flexibility for tuning the system operating points, in our previous research [14] we found the method sometimes tends to generate a lot of false alarms for single-word keywords. The problem is especially serious for the keywords seen in the training data. Since their prior probabilities in the original LM are already high enough to achieve low miss detection rates, boosting the probabilities of those seen single-word keywords reduce only a marginal miss detection rate with a dramatically increase on the number of false alarms. As a result, the KWS performance

for the single-word keywords can barely be improved by the CS-KWLM approach due to the large number of false alarms.

In this research a context-dependent keyword language modeling (CD-KWLM) method is proposed to alleviate this false alarm problem for the single-word keywords seen in the training text. Similar to CS-KWLM, the keyword LM is trained with the training text built up from the same process described in Section 2.1. However, for the single-word keywords appearing in the training text, instead of using n -grams with high probabilities in the original LM as context terms, we use only n -grams before and after the keywords in the training data as their context terms for the keyword language model training text building. With these new context terms (or context-dependent fillers) for the seen single-word keywords, the CD-KWLM trained with this new training text would therefore have less probabilities for those single-word keywords seen in the training data, and thus their false alarm rates could be reduced. Once the context-dependent keyword language model is trained, we can use Eq. (1) again to get the interpolated language model for the KWS system use.

3. Experimental Setup

Experiments were conducted on the English Switchboard-I dataset. The training data consist of the whole 309-hour switchboard data and were used for both acoustic and language model training. We used the standard 3.7-hour NIST 2000 Hub5 evaluation set for the system performance evaluation. Both Switchboard and CallHome recordings were used for evaluation. We built our baseline system following the Kaldi Switchboard recipe [15]. The acoustic features are fMLLR transformed MFCC features. For acoustic models, DNN models with sMBR sequential training [16] were used. And the baseline 3-gram language model was trained with the transcriptions of the 309-hour data. The word-error-rate (WER) of the baseline system on the evaluation data is 20%.

For KWS performance evaluation, one hundred keywords were carefully selected from the Hub5 evaluation data. To study the effect of keyword length to the KWS system performance, we categorized the keywords into three types according to the number of words in the keyword as shown in Table 1 and made sure the number of selected keywords in each type are balanced. Further, for each type of the keywords, we ensured that 60% of the keywords are seen in the language model training text, while the rest 40% of the keywords are unseen to the LM. This setting allows us to study the whether the keywords seen and unseen to the language models affects their KWS performance. Table 2 presents some examples of the keywords used in the experiments.

Table 1. Number of the three types of the keywords seen and unseen to the language model training text.

Type	#keyword	#Seen	#Unseen
1-word	34	20	14
2~3-word	33	20	13
4~5-word	33	20	13

Table 2. Examples of the three types of the keywords.

Type	Example
1-word	anorexia / schools / Libya / shoes / York
2~3-word	torsion wrench / ultra violet light / stone quarry
4~5-word	needle in a haystack / daycare for your child

The performance of keyword search was measured by miss detection rate, false alarm per keyword per hour (FA/KW/Hr), and Actual Term Weighted Value (ATWV), which is computed by

$$ATWV = 1 - \frac{1}{K} \sum_{kw=1}^K \left(\frac{N_{Miss}(kw)}{N_{True}(kw)} + \beta \frac{N_{FA}(kw)}{T - N_{True}(kw)} \right). \quad (2)$$

where K is the number of keywords, $N_{Miss}(kw)$ is the number of true keyword tokens that are not detected, $N_{FA}(kw)$ is the number of false alarms, $N_{True}(kw)$ is the number of keywords in reference, T is the number seconds of the evaluation audio, and β is a constant set as 999.9.

Three KWS systems were compared in this study. While all the systems share the same acoustic model, they are different in the language models. The first system is the baseline which used the original 3-gram LM. The second used the CS-KWLM interpolated LM as the system language model (denoted as "CS-KWLM Int"). And the third system is the KWS system with the CD-KWLM interpolated LM proposed in this research (denoted as "CD-KWLM Int").

4. Experimental Results and Discussion

4.1. System Performance

Table 3 compares the three KWS systems in this research. The interpolation weights, α , for CS-KWLM and CD-KWLM systems were selected for the highest ATWV, and were 0.9 for both systems. In Table 3, the baseline KWS system achieved 0.6645 of ATWV. It is clear that both of the keyword-aware LM based systems significantly outperform the baseline system on the ATWV performance. The CS-KWLM system attained 20% relative improvement of ATWV (from 0.6645 to 0.7992) over the baseline system. While the proposed CD-KWLM system further reached 0.8079 of ATWV, which is the best performance among the three systems. Note that despite the keyword-aware language modeling was originally designed for KWS systems with limited LM training data [12], the English Switchboard results here show that the method could still achieve a significant improvement to the KWS systems with great amounts of data for LM training.

Table 3. Performances for the three KWS systems in this research (α chosen for highest ATWV)

	ATWV	FA/KW/Hr	Miss Rate
Baseline	0.6645	0.1056	0.34
CS-KWLM Int ($\alpha=0.9$)	0.7992	0.3639	0.21
CD-KWLM Int ($\alpha=0.9$)	0.8079	0.3167	0.20

For false alarms, the baseline LVCSR-based KWS system had the lowest false alarm rate as 0.1056 FA/KW/Hr. However, as mentioned in the introduction section, the system suffered from the serious miss-detection problem. The miss detection rate of the baseline system was rather high at 34%, which makes the system not be suitable for applications such as surveillance requiring better sensitivities to the important keyword terms. By tuning the interpolation weights, α , in both CS-KWLM and CD-KWLM systems to 0.9, the systems significantly reduced the miss detection rates to 21% and 20%. In general, reducing the miss detection rates of KWS systems is more critical than diminishing the false alarm rates for many

applications. The false alarms can be easily removed by using a further utterance verification stage [17-20], while it is usually difficult for the KWS systems to recover the missed detections from an incomplete putative list.

4.2. System Operating Point Tuning with α

In this section, we demonstrate how the false alarm and miss detection rates and ATWV of a CS-KWLM system are changed with the interpolation weight α . In Figure 3, it is clear that as α increased from 0 to 0.9, the miss detection rate of the system monotonically declined from 34% to 21% (38% relative reduction). In exchange for the decrease of the miss detection rate, the false alarm rate was raised from 0.1056 to 0.3639 FA/KW/Hr. Therefore, by tuning the interpolation weight, α , a CS-KWLM system can easily control the its operating point for the missed detection and false alarm performance. Since the ATWV setting in this study penalized missed detections more than false alarms, the best ATWV performance was achieved when $\alpha=0.9$.

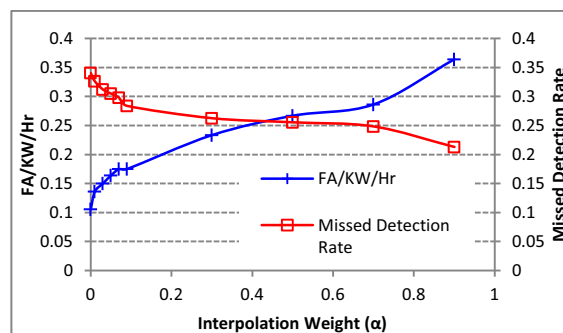


Figure 3. Illustrations of α against FA/KW/Hr and miss detection rate for the CS-KWLM system.

We further investigate the relationship between the interpolation weight, α , and the ATWV performance of each types of keywords. In Figure 4, it is clear that the trends of the three keyword types are very different as the interpolation weight, α , increases. For the 4~5-word keywords, the ATWV performance monotonically increased from 0.6440 to 0.9347. This observation is consistent with our previous discovery in [13, 14] that prior probabilities of multi-word keywords are usually seriously underestimated by conventional n -gram LM and resulting in a high miss rate. Thus this type of keywords benefits from the keyword-aware LM method most. The considerable ATWV improvement of the 4~5-word keywords also made the KWS system having the best overall ATWV when $\alpha = 0.9$. For the 2~3-word keywords, the ATWV remained about the same after α got larger than 0.3. The increased false alarms after $\alpha \geq 0.3$ somehow cancelled the effect of the miss-detection-rate reduction to the ATWV enhancement.

On the other hand, for 1-word keywords, their ATWV achieved the highest value of 0.5589 when $\alpha = 0.3$ and then decreased gradually due to the increased false alarms as the value of α went up. Figure 5 shows that the number of false alarms for 1-word keywords increased rapidly with the value of α . While the number of false alarms for 2~3-word and 4~5-word keywords remained at less than 10 for all α values, the number of false alarms for 1-word keyword escalated from 33 ($\alpha=0$) to 121 ($\alpha=0.9$). From Figure 5, it is obvious that about 90% of the CS-KWLM system's false alarms are from the single-word keywords. Thus one of the keys to further

enhance the KWS system performance is to reduce the false alarms generated by 1-word keywords.

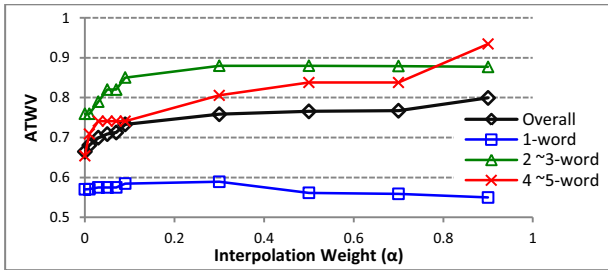


Figure 4. Illustrations of α against the ATWV performances for the CS-KWLM system.

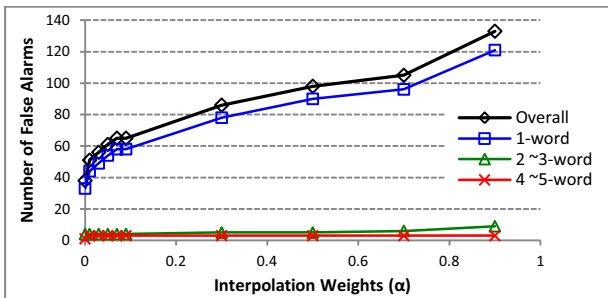


Figure 5. Illustrations of α against the number of false alarms for the CS-KWLM system.

4.3. Performance Analysis for single-word keywords

In this study, we proposed the CD-KWLM approach to address false alarm problems for the single-word keywords in the CS-KWLM systems. With the help of the context-dependent fillers to restrict the probabilities of seen 1-word keywords in the keyword LM, the system false alarm can be reduced to some extent. Table 4 compares the ATWV performance of the 1-word keywords for the three KWS systems.

In Table 4, though the CS-KWLM system had an ATWV improvement on the unseen 1-word keywords, its performance for seen 1-word keywords was worse than the baseline due to the false alarms. On the contrary, the ATWVs of both seen and unseen 1-word keywords in the CD-KWLM system were considerably improved, and achieved 0.6430 in the overall ATWV performance. It is also notable that with lower false alarm rate, the best α for 1-word keywords in the CD-KWLM system was increased from 0.3 to 0.5.

Table 4. ATWVs for 1-word keywords (α chosen for highest ATWV).

1-word keywords	Seen (20)	Unseen (14)	Overall (34)
Baseline	0.5730	0.5701	0.5701
CS-KWLM Int ($\alpha=0.3$)	0.5378	0.6514	0.5889
CD-KWLM Int ($\alpha=0.5$)	0.6188	0.6788	0.6430

To further study the effectiveness of the proposed CD-KWLM approach, Figure 6 shows the receiver operating characteristic (ROC) curves for the CS-KWLM and CD-KWLM systems on the 20 single-word keywords that were seen during original LM training. Note that the ROC curve of the CD-KWLM system always stayed on the left side of the

curve of the CS-KWLM system. In general, the CD-KWLM system could achieve the same detection rate of the CS-KWLM system with only half of the false alarm rate the CS-KWLM system attained.

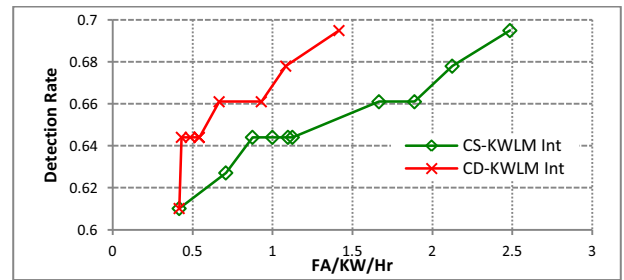


Figure 6. Illustrations of the ROC curves for the 20 single-word keywords seen during baseline LM training.

4.4. OpenKWS13 Vietnamese LimitedLP Task

Similar experiments were also performed on the OpenKWS13 Vietnamese LimitedLP task (with the experiment setup described in [12]) to verify the results discovered in this research. We observed the same behaviors for the desired system operating point tuning with α . Table 5 shows the performance of the 79 single-word keywords in the task on the evaluation part 1 data. It is clear that, like what we found in the Switchboard data, the CD-KWLM system achieved the best performance on the seen 1-word keywords.

Table 5. MTWVs of 1-word keywords in OpenKWS13 Vietnamese LimitedLP task for the three KWS systems.

evalpart1 data	Seen (52)	Unseen (27)	Overall (79)
Baseline	0.0451	0.0103	0.0332
CS-KWLM Int ($\alpha=0.6$)	0.0500	0.0101	0.0364
CD-KWLM Int ($\alpha=0.6$)	0.0650	0.0075	0.0430

5. Conclusion

In this paper, we have explored the flexibility of keyword-aware language modeling based KWS systems to control the tradeoff between the false alarms and missed detections. By increasing the interpolation weight, α , we are able to increase the false alarm rates in exchange for a decrease of the miss detection rates by up to 38% relative (from 34% to 21%). A relative gain of 20% in ATWV performance with the keyword-aware LM based KWS system over the conventional LVCSR-based KWS system is also observed when testing on the English Switchboard data. Moreover, the proposed context-dependent keyword language modeling could further achieve a 9% relative ATWV improvement over the original keyword-aware KWS systems for single-word keywords. Similar MTWV improvement is also observed for single-word keywords in the OpenKWS13 Vietnamese LimitedLP task.

6. Acknowledgements

This effort uses the IARPA Babel Program Vietnamese language collection release babel107b-v0.7 with the LimitedLP training set. This work was done while the author, Tze Siang Lau, was a visiting scholar in the School of Electrical and Computer Engineering, Georgia Institute of Technology.

7. References

- [1] J. G. Wilpon, L. R. Rabiner, C.-H. Lee, and E. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, pp. 1870-1878, 1990.
- [2] R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," in *ICASSP*, 1990, pp. 129-132.
- [3] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, *et al.*, "The SRI/OGI 2006 spoken term detection system," in *Interspeech*, 2007, pp. 2393-2396.
- [4] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 615-622.
- [5] D. R. Miller, M. Kleber, C.-I. Kao, O. Kimball, T. Colthurst, S. A. Lowe, *et al.*, "Rapid and accurate spoken term detection," in *Interspeech*, 2007.
- [6] R. Wallace, R. Vogt, and S. Sridharan, "A Phonetic Search Approach to the 2006 NIST Spoken Term Detection Evaluation," in *Interspeech*, 2007.
- [7] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, *et al.*, "Speech and Language Technologies for Audio Indexing and Retrieval," *Proc. IEEE*, vol. 88, pp. 1338-1353, 2000.
- [8] R. L. Warren, "BROADCAST SPEECH RECOGNITION SYSTEM FOR KEYWORD MONITORING," U.S. Patent 6332120 B1, 2001.
- [9] T. Kawahara, C.-H. Lee, and B.-H. Juang, "Key-Phrase Detection and Verification for Flexible Speech Understanding," *IEEE Trans. on Speech and Audio Proc.*, vol. 6, pp. 558-568, Nov. 1998.
- [10] B.-H. Juang and S. Furui, "Automatic Recognition and Understanding of Spoken Language – A First Step Toward Natural Human-Machine Communication," *Proc. IEEE*, vol. 88, pp. 1142-1165, Aug. 2000.
- [11] I. Szoke, P. Schwarz, P. Matejka, L. Burget, M. Karafiat, M. Fapso, *et al.*, "Comparison of Keyword Spotting Approaches for Informal Continuous Speech," in *EuroSpeech*, 2005.
- [12] I.-F. Chen, C. Ni, B. P. Lim, N. F. Chen, and C.-H. Lee, "A Novel Keyword+LVCSR-Filler Based Grammar Network Representation for Spoken Keyword Search," in *ISCSLP*, Singapore, 2014.
- [13] I.-F. Chen, C. Ni, B. P. Lim, N. F. Chen, and C.-H. Lee, "A KEYWORD-AWARE GRAMMAR FRAMEWORK FOR LVCSR-BASED SPOKEN KEYWORD SEARCH," in *ICASSP*, 2015.
- [14] I.-F. Chen, C. Ni, B. P. Lim, N. F. Chen, and C.-H. Lee, "A Keyword-Aware Language Modeling Approach to Spoken Keyword Search," *accepted by Journal of VLSI Signal Processing Systems*, 2015.
- [15] D. Povey, A. Ghoshal, G. Boulianne, L. S. B. S., O. R. Glembek, N. Goel, *et al.*, "The Kaldi Speech Recognition Toolkit," in *ASRU*, Hilton Waikoloa Village, Big Island, Hawaii, US, 2011.
- [16] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Interspeech*, Lyon, France, 2013.
- [17] R. A. Sukkar and C.-H. Lee, "Vocabulary Independent Discriminative Utterance Verification for Non-Keyword Rejection in Subword Based Speech Recognition," *IEEE Trans. on Speech and Audio Proc.*, vol. 4, pp. 420-429, Nov. 1996.
- [18] M.-W. Koo, C.-H. Lee, and B.-H. Juang, "Speech Recognition and Utterance Verification Based on a Generalized Confidence Score," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, pp. 821-832, Nov. 2001.
- [19] J. Ou, K. Chen, X. Want, and Z. Lee, "Utterance Verification of Short Keywords Using Hybrid Neural-Network/HMM Approach," in *ICII*, 2001.
- [20] I.-F. Chen and C.-H. Lee, "A Hybrid HMM/DNN Approach to Keyword Spotting of Short Words," in *Interspeech*, 2013.