A Unified Confidence Measure Framework Using Auxiliary Normalization Graph

Zhehuai Chen, Yanmin Qian, Kai Yu[⋆]

Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering
Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, China
{chenzhehuai, yanminqian, kai.yu}@sjtu.edu.cn

Abstract. Due to the distinct search space and efficiency demands in different ASR applications, the state-of-the-art confidence measures and their decoding frameworks are heterogeneous among keyword spotting, domain-specific recognition and LVCSR. Inspired by the success in applying a phone level language model to replace the word lattice in discriminative training, the *auxiliary normalization graph* is proposed in this work, and it is constructed to model the observation probability in hypothesis posterior based confidence measure. In this way, confidence measure normalizing term modelling can be independent from the original search space and the confidence measure can be grouped into an unified framework. Experiments on three typical ASR applications show that the proposed method using a unified confidence measure framework achieves comparable performance to the separately optimized system on each task.

Keywords: Confidence measure, Auxiliary normalization graph, Connectionist Temporal Classification, Phone synchronous decoding

1 Introduction

In automatic speech recognition (ASR), *confidence measure* (CM) is used to evaluate the reliability of recognition results. CM is taken as a further verification stage to different ASR applications, e.g., *keyword spotting* (KWS) [1] to guarantee low false acceptance rates, grammar [2] or class [3] language model based domain-specific recognition to verify in-domain recognition result [4][5], and *large vocabulary continuous speech recognition* (LVCSR) [2] to support semantic processing [6]. Due to distinct search space and efficiency demands in

⋆ This work was supported by the Shanghai Sailing Program No. 16YF1405300, the China NSFC projects (No. 61573241 and No. 61603252) and the Interdisciplinary Program (14JCZ03) of Shanghai Jiao Tong University in China. Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

above applications, confidence measures are usually heterogeneous [7][8][9][10], which will be reviewed in section 2.

It is challenging to propose an unified framework across all above ASR applications with high performance in both CM and accurary. The key challenge of the unified framework includes two sides: i) how to normalize the best ASR result with proper overall evaluation, i.e., the normalizing term in CM. ii) how to keep the computational efficiency in resource limited scenarios, e.g., spoken term detection in some personal digital assistant. *Keyword-filler* based method [9] and *utterance verification* [11] can be viewed as such kind of trials. For instance, in KWS, a context independent (CI) linguistic unit, called `filler` is proposed to model all the non-keyword elements, which is imperfect. In domain-specific ASR, the framework even suffers from weakness in context dependency (CD) modelling, which results in worse `filler` recognition ability. In LVCSR, a theoretically better method, i.e., hypothesis posterior based CM [10], is proposed. ASR is formulated as *maximum a posterior* (MAP) in the framework. The posterior probability of ASR output given the whole utterance can be served as CM. The observation probability is modelled by the summation of probability of all the hypothesis from the ASR search space. Because ASR search space is always tremendous, the word lattice recorded in the decoding process is used to constrain the hypothesis. In LVCSR, hypothesis posterior based CM is significantly better than the unified framework, although other applications can't benefit from it, e.g. KWS.

Modeling the search space with an elaborately optimized phone level language model to replace the word lattice, recently shows competitive performance in discriminative training [12][13]. In the paper, similar idea is adopted to the normalizing term modelling, i.e., *auxiliary normalization graph*. Because of the phone level acoustic modelling, such method is theoretically sound in all above applications. Therefore, the CM normalizing term modelling can be independent with both the original search space and the acoustic modelling. To reduce the computational cost from the search space modelling, CTC-based *phone synchronous decoding* (PSD) [14] is further adopted, which shows great efficiency in phone level decoding.

In the paper, an unified and efficient confidence measure framework using auxiliary normalization graph and phone synchronous decoding is proposed to provide consistent performance among all types of ASR applications within a single framework, for the first time. The whole paper is arranged as follow. In section 2, the state-of-the-art confidence measures and decoding frameworks in different ASR search space are compared and summarized. In section 3, the auxiliary normalization graph is proposed to form an unified and efficient framework for varieties of search space. Section 4 describes experiments and analysis, followed by the conclusion in section 5.

2 Confidence Measure and Search Space

Regarding to the difference in search space and decoding frameworks, ASR applications can mainly be divided into three types, and the state-of-the-art confidence measures are heterogeneous among these various ASR search space.

2.1 Keyword Spotting

KWS is to accurately and efficiently detect words or phrases of interest, i.e., keywords, in continuous speech. Therefore, the search space of KWS is all the keyword sequences [1]. False acceptance reveals to falsely recognize speech spans with interested keywords, which is undesirable. i.e., false alarm segments are not in the original search space. To treat this problem, a branch of methods [8] include a post-processing algorithm to provide a word level CM with specific threshold to potentially model the non-keyword elements. Another branch of methods [9] add non-keyword units into acoustic modelling. Due to the difficulty in non-keyword modelling, the previous branch shows better performance especially in restricted KWS [8].

2.2 Domain-specific Recognition

Recent focus on assistant products has increased the need for users to make voice commands referencing their own personal data, such as favorite songs, names and contacts. In the scenarios, grammar [2] or class [3] based language model with *slots* to dynamically indexing personal words or phrases, is the most suitable search space of the domain-specific recognition [4]. The false recognition in the task includes: i) falsely recognizing out-domain utterance to the in-domain result; ii) correctly recognizing the domain but with false slot values. To verify the recognition result, a sentence level CM should be effective in both cases. For in-domain utterances, the language patterns and *slots* can be viewed as the complete search space. Therefore, the CM discussed in section 2.3 is proper. However for out-domain utterances, the out-domain elements need to be modelled specifically as the methods in section 2.1. To our knowledge, there hasn't been careful research conducted in this aspect.

2.3 Large vocabulary continuous speech recognition

In *large vocabulary continuous speech recognition* (LVCSR) [2], the search space is modelled by a n-gram language model. To support semantic post-processing [6], CM is proposed to model the reliability of spontaneous speech recognition results. Hypothesis posterior based CM [10], is commonly used in LVCSR. In this framework, ASR is formulated as the *maximum a posterior* (MAP) decision process. The posterior probability of ASR output given the whole feature sequence can be served as a word level or sentence level CM. The normalizing term of

[1] The LVCSR based KWS is not included in the discussion because it's mostly a problem to enhance the acoustic model performance and keyword indexing algorithm. Besides, the computational burden is not suitable for resource-limited scenarios.

MAP, i.e., the observation probability, is modelled by the summation of probability of all the hypothesis from the ASR search space. Because the ASR search space is always tremendous, the word lattice recorded through the decoding process is used to constrain the hypothesis.

3 Auxiliary Normalization Graph based Confidence Measure

In this paper, an unified framework among all above ASR applications with good performance in both CM and accurary is proposed using *auxiliary normalization graph* and CTC-based *phone synchronous decoding*.

3.1 Unified Confidence Measure Framework

The posterior probability of ASR output given the whole utterance is served as CM in MAP framework,

$$CM = P(\mathbf{w}|\mathbf{x}) = \frac{P(\mathbf{x}|\mathbf{w}) \cdot P(\mathbf{w})}{P(\mathbf{x})} \tag{1}$$

Here, $P(\mathbf{w})$ is the language model probability and $P(\mathbf{x}|\mathbf{w})$ is the acoustic part. $P(\mathbf{x})$ is the probability of observing $\mathbf{x}$ and can be modelled as below,

$$P(\mathbf{x}) = \sum_{H} P(\mathbf{x}, H) = \sum_{H} P(H) \cdot P(\mathbf{x}|H) \tag{2}$$

Here, H denotes all the alternative competing hypotheses of the recognition results. H is distinct between different ASR applications and always infinite.Accordingly, the modelling on H is the bottleneck of the performance.

The key challenge of the unified framework includes two sides: i) how to model the task-specific set H in an unified method. ii) how to keep the computational efficiency in modelling the theoretically infinite set H.

3.2 Auxiliary Normalization Graph

To solve this problem, a *auxiliary normalization graph* is proposed to be integrated with the original ASR search space. The architecture of the proposed method is compared with traditional methods in Figure 1.

In lattice based method, $P(\mathbf{x})$ is obtained from the lattice recorded from a subset of decoding graph. In `filler` based method, $P(\mathbf{x})$ is modelled by phone-loop graph. In the proposed method, $P(\mathbf{x})$ can be obtained from the modified search space as,

$$P(\mathbf{x}) \approx \max_{H} P(H) \cdot P(\mathbf{x}|H) \tag{3}$$

Here, three types of auxiliary normalization graphs are proposed for comparison.

- Phone loop graph (`AX1`). All the phone-loop can be combined together as the auxiliary normalization graph. Similar to the traditional keyword-filler CM [9], an all-phone recognition can be conducted to normalize the word based decoding result.

- Lexicon free graph (`AX2`). Innovated by recent progress in lattice-free discriminative training [13], the auxiliary normalization graph can be constructed from a phone level language model to approximate the search space.
- Lexicon based graph (`AX3`). In some language, e.g., Mandarin, the mapping between acoustics to characters is always many-to-one [2]. Therefore, given the limited number of characters, the expected pronunciations are also limited. All the possible pronunciations can be combined together as the auxiliary normalization graph.

Besides, the proposed method can be served as a word level CM. In this case, Equation(1) and (3) can be transformed to Equation(4) and (5),

$$CM = P(w|\mathbf{x}) = \frac{P(\mathbf{x}|w) \cdot P(w)}{P(\mathbf{x}^w)} \tag{4}$$

$$P(\mathbf{x}^w) \approx \max_{H^w} P(H^w) \cdot P(\mathbf{x}^w|H^w) \tag{5}$$

Here, $P(\mathbf{x}|w)$ is the acoustic model probability within the time span of the word w and $P(w)$ is the language model probability of w. $\mathbf{x}^w$ is the feature sequence within the time span of w and H^w is the alternative hypothesis of w. $P(\mathbf{x}^w)$ can be specifically obtained from the auxiliary normalization graph in decoding stage to form the word level CM.

Decoding results in the auxiliary normalization graph is a good approximation of the observing probability to form the CM in MAP framework. Because the acoustic model unit is always phone level, such method of search space modelling is theoretically sound in all above ASR applications. Compared to the traditional lattice based method, the proposed method is more stable among all types of ASR search space, which will be revealed in the experiment part. Besides, the method doesn't need to further include a series of non-keyword model units as `filler` based or utterance verification methods. Therefore, the CM normalizing term modelling can be independent with both the original search space and the acoustic modelling.

3.3 Efficient CTC-based Phone Synchronous Decoding

To reduce the computational cost from the search space modelling, CTC-based *phone synchronous decoding* (PSD) [14] can be adopted, which shows the great efficiency in the phone level decoding. Due to the removal of non-phonemic frames, compared with the traditional *frame synchronous decoding* (FSD), less search errors and phone boundary disambiguity are made [15]. This phenomenon results in less hypothesis H in Equation (2). It has been proved in [14][16] that the decoding process becomes a small part of the overall computation. In the experiment part, the efficiency is also taken into account.

[2] while in language like English, the mapping is many-to-many.

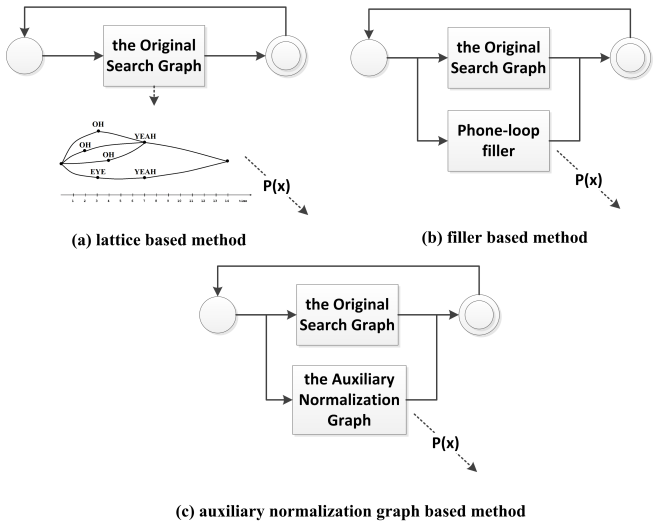

Fig. 1. *Architecture comparison. The original search graph can be from KWS, domain-specific recognition and LVCSR.*

3.4 Empirical Implementation

The search space construction of PSD based system is discussed in [16]. In KWS, the search graph G is a series of linear acceptors of keywords[1]. G can be grammar or class based language model in domain-specific recognition, and n-gram [17][3] language model in LVCSR.

The proposed auxiliary normalization graph is finally unified with the original search space. The output symbol of the auxiliary normalization graph is the symbol $\langle$`fil`$\rangle$ as in `filler` based methods. In the decoding stage, the probability of $\langle$`fil`$\rangle$ symbols can be obtained specifically to form Equation (5).

4 Experiments

To evaluate the unified framework in all ASR search space, experiments are conducted on three typical ASR applications, i.e., KWS, domain-specific recognition [3] and LVCSR. A 5000 hours Mandarin corpus is used to train the CTC model [4] and the training configuration is the same as [14].

In KWS, the word level CM is evaluated by the false alarm and false rejection of the keyword sequences in each utterance. In domain-specific recognition, the sentence level CM is used to filter out two types of recognition errors discussed in section 2.2. *Equal error rate* (EER) is taken as the sentence level metric in KWS and domain-specific recognition, which reflects the average error rate of

[3] Grammar language model based decoding is taken, as the in-domain and out-domain evaluation discussed in section 2.2 are similar between grammar and class based model.

[4] The comparison between CMs in CTC and HMM frameworks has been conducted in previous research [15], all the comparisons below are within the CTC.

the false alarm and false rejection. The lower EER is the better. *Normalised cross entropy* (NCE) [15] is taken as the metric of the word level CM quality in LVCSR. The higher NCE is the better. To ensure that the ASR precision isn't deteriorated by the unified framework, the sentence level recalling rate of the positive examples is provided in KWS and domain-specific recognition, denoted as *snt recall*. In LVCSR, *character error rate* (CER) is used. To evaluate the efficiency of proposed unified framework, the portion of the decoding time except the acoustic model computation, versus all the decoding time is also measured, denoted as *portion of time except acoustic model* (PEA). Because experiments are all conducted with the same acoustic model, the lower PEA shows the less time taken in the other process, e.g., graph searching, lattice generation and post-processing[5]. i.e., the higher PEA indicates the more computational cost from the CM framework.

The baseline CM methods include the predictor feature based CM and the hypothesis posterior based CM, denoted as `AC` and `CN`. They show the best reported result in CTC framework and outperform their HMM competitors [15] . The proposed methods are compared in the experiments, i.e., `AX1`, `AX2` and `AX3` in section 3.2. The auxiliary normalization graph `AX2` is generated from a tri-gram phone language model with *145K* grams. `AX3` is generated from a lexicon graph L mapping phone to syllable, and an all syllable-loop graph G, by $L \circ G$. `filler` based method is not included as it's theoretically similar to `AX1`.

4.1 Keyword Spotting

In the task, 398 home appliance keywords are chosen and tested in 17332 utterances (11789 positive and 5543 negative examples). Table 1 shows the CM quality of different methods.

Table 1. *KWS Task*

CM	setup	EER(%)	snt recall(%)	PEA(%)
Phonemic	`AC`	11.65	88.4	10
Hypothesis	`CN`	12.55	88.4	11
	`AX1`	11.60	88.0	10
	`AX2`	10.16	88.2	16
	`AX3`	10.10	88.2	15

Result shows that `AC` outperforms `CN`. The reason is that the overall search space is very limited in KWS, resulting imperfect observing probability modeling from decoding lattice. The imprecision of normalizing term in MAP results in worse CM quality. The proposed auxiliary normalization graph can alleviate the problem and bring about better CM. Concretely, `AX1` can't benefit much. We suspect it is because the end-to-end model CTC intrinsically shares similar

[5] As in [14], the acoustic model is a small size one applied in the embedded application. Therefore, computation time is comparable between all above portions in the three tasks

normalization of all-phone recognition in the model. `AX2` and `AX3` are significantly better than traditional methods, while `AX3` is slightly better. The reason is from better characterization of the linguistic search space in Mandarin, discussed in section 3.2.

Besides, the recalling rate shows that proposed unified confidence measure and its decoding framework slightly affects the model precision. Regarding to the improvement in false alarm revealed by EER, the side effect is tolerable.

In aspect of efficiency, the computation of `CN` is around 10% more than `AC` because of lattice and confusion network generation. Although PEA of proposed method is notably more than both `AC` and `CN`, the total time except the acoustic model computation still takes a very small portion in the task. The reason is from application of PSD. Detail comparison of decoding time versus search space size in PSD can be referred to [14].

4.2 Domain-specific Recognition

A task of larger ASR search space is examined in the section, i.e., grammar based language model ASR decoding. The test-set includes 13186 voice assistant utterances (7923 positive examples and 5263 negative examples), e.g., phone calls, voice commands and etc. The grammar contains several supported speaking styles and different contact information. The negative examples include both in-domain and out-domain situations discussed in section 2.2.

Table 2. *Grammar based ASR Task*

CM	setup	EER(%)	snt recall(%)	PEA(%)
Phonemic	`AC`	19.86	87.4	38
Hypothesis	`CN`	15.78	87.4	43
	`AX1`	19.80	86.0	38
	`AX2`	16.23	87.2	41
	`AX3`	16.12	87.2	40

Table 2 shows the result. In the task of larger ASR search space, `CN` significantly outperforms `AC`, because `AC` doesn't harness the competitive relationship between hypotheses and results in false acceptance. `AX2` and `AX3` are similar to `CN`. The reason is that with larger ASR search space, both the decoding lattice and the auxiliary normalization graph can be a good approximation of observing probability. `AX3` is still slightly better than `AX2`, so we only use `AX3` in the latter experiment.

In the task, `AX1` notably does harm to the recalling rate. It reveals that a series of CI model units is hard to model all the non-keyword and out-domain elements. Besides, the weakness in modelling context dependency also affects the recognition results.There is no such weakness in `AX2` and `AX3`.

Regarding to the efficiency, the proposed framework only slightly affects the computation time compared with the previous task. The reason is compared with original larger search space in grammar based decoding, the search space increment from the proposed auxiliary normalization graph is more ignorable.

4.3 LVCSR

In the section, a mandarin spontaneous conversation test-set (about 25 hours) is taken. A tri-gram language model with *118K* words and *1.9M* grams is used in the decoding stage.

Table 3. *LVCSR Task*

CM	setup	NCE	CER(%)	PEA(%)
Phonemic	`AC`	0.182	10.2	45
Hypothesis	`CN`	0.302	10.2	50
	`AX3`	0.260	10.1	46

As in section 4.2, `CN` outperforms `AC` in LVCSR. `AX3` shows worse but comparable result compared with `CN`. The reason is that in large ASR search space, the auxiliary normalization graph can't provide extra competing information. Meanwhile, in `AX3`, only the best decoding path is taken to simulate the search space, which is worse than `CN` from the decoding lattice.

The CER of proposed framework is slightly better than the baseline. We believe it's because the auxiliary normalization graph fitters out a portion of false decoding paths, which reduces the insertion and substitution errors. And regarding to the efficiency, it's similar to the previous task.

5 Conclusion

In the paper, the unified confidence measure and efficient decoding framework using auxiliary normalization graph and CTC-based phone synchronous decoding achieves comparable performance to the separately optimized systems on three typical ASR applications. The proposed unified framework can be independent with both the original search space and the acoustic modelling. Future work includes extending the proposed framework to other state-of-the-art acoustic models and language models, e.g., LFMMI and NNLM.

6 Relation to Prior Work

To form an unified confidence measure framework for different ASR applications, prior trial on *keyword-filler* [9] and *utterance verification* [11] are not very successful. Besides, they both need to further include a series of non-keyword units into the acoustic model. The proposed auxiliary normalization graph is inspired by the success in applying an elaborately optimized phone level language model to replace the word lattice in discriminative training [12][13]. In [18], the word level n-gram language model is combined with the keyword search graph to improve its recognition. The work shares similar profits from the auxiliary graph added into the original search space, but it's different in: i) the language model in the auxiliary normalization graph is phone level; ii) the proposed auxiliary normalization graph is combined with varieties of search space; iii) most importantly, the motivation is different, i.e., the proposed graph is to model the

normalization term in CM. Compared with separately optimized confidence measures [7][8][9][10] in different ASR applications, the proposed method achieves consistent performance within a single framework for the first time.

References

1. M. Weintraub, "Keyword-spotting using sri's decipher large-vocabulary speech-recognition system," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1993, pp. 463–466.
2. P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large vocabulary continuous speech recognition using htk," in *1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1994, pp. II–125.
3. W. Ward and S. Issar, "A class based language model for speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 416–418.
4. L. Vasserman, B. Haynor, and P. Aleksic, "Contextual language model adaptation using dynamic classes," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 441–446.
5. J. Cleveland, D. Thakur, P. Dames, C. Phillips, T. Kientz, K. Daniilidis, J. Bergstrom, and V. Kumar, "Automated system for semantic object labeling with soft-object recognition and dynamic programming segmentation," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 2, pp. 820–833, 2017.
6. D. Hakkani-Tür, F. Béchet, G. Riccardi, and G. Tur, "Beyond asr 1-best: Using word confusion networks in spoken language understanding," *Computer Speech & Language*, vol. 20, no. 4, pp. 495–514, 2006.
7. W. Hu, Y. Qian, and F. K. Soong, "A new dnn-based high quality pronunciation evaluation for computer-aided language learning (call)." in *INTERSPEECH*, 2013, pp. 1886–1890.
8. G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4087–4091.
9. S. R. Young, "Detecting misrecognitions and out-of-vocabulary words," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 2. IEEE, 1994, pp. II–21.
10. F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on speech and audio processing*, vol. 9, no. 3, pp. 288–298, 2001.
11. R. C. Rose, B.-H. Juang, and C.-H. Lee, "A training procedure for verifying string hypotheses in continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 281–284.
12. S. F. Chen, B. Kingsbury, L. Mangu, D. Povey, G. Saon, H. Soltau, and G. Zweig, "Advances in speech transcription at ibm under the darpa ears program," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1596–1608, 2006.
13. D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," *Submitted to Interspeech*, 2016.

14. Z. Chen, W. Deng, T. Xu, and K. Yu, "Phone synchronous decoding with ctc lattice," in *Interspeech 2016*, 2016, pp. 1923–1927. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-831
15. Z. Chen, Y. Zhuang, and K. Yu, "Confidence measures for ctc-based phone synchronous decoding," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017.
16. Z. Chen, Y. Zhuang, Y. Qian, and K. Yu, "Phone synchronous speech recognition with ctc lattices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 86–97, Jan 2017.
17. A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit." in *Interspeech 2002*, vol. 2002, 2002, p. 2002.
18. 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 *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, 2014, pp. 192–196.