DEEP NEURAL NETWORK BASED WAKE-UP-WORD SPEECH RECOGNITION WITH TWO-STAGE DETECTION

Fengpei Ge, Yonghong Yan

The Key Laboratory of Speech Acoustic and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

ABSTRACT

This paper presents a novel far-field voice trigger algorithm utilizing DNN with the objective function of state-level minimum Bayes risk for training, customizing the decoding network to absorb the ambient noise and background speech. We adopt a two-stage classification strategy to integrate the phonetic knowledge and model-based classification into detecting wake-up words. Experimental results of the online test show that it can provide a higher than 90% accuracy and meanwhile false alarms are less than once per nine hours in the noisy home environments where the sound pressure level is about 80dB.

Index Terms—Wake-up-word Speech Recognition, Deep Neural Network, Far-field Speech Recognition

1. INTRODUCTION

Recently the technique of automatic speech recognition (ASR) has gained significant improvements as an important part of artificial intelligence[1-3], but it still has a big challenge to let a computer recognize all speech at anywhere and anytime. As one of the most prominent applications, automatic wake-up-word (WUW) speech recognition (WUWSR) becomes more essential. We usually first call her/his name to attract someone's attention before a conversation. The technique of AWUWSR, also called as automatic voice trigger, is to play such a role in human-computer interaction. It can be used as a switch to an ASR system instead of the conventional push-to-talk mode and improve the multitasking lifestyle.

In order to facilitate our life successfully, a key point is to maintain a stable performance in noisy environments. In the far-field pick-up pattern, where speakers are a few meters away from the receiver, the speech received has been polluted seriously by the noise and reverberation. Especially in the conditions of low signal-noise rate (SNR), most ASR systems can't work because of speech distortion. The far-field pick-up and a low SNR bring the AWUWSR system many false alarms and result in unacceptable performance.

AWUWSR algorithms are classified into two categories. One is the method of pattern matching based on templates, and the other is based on ASR framework. Anhao Xing[4] adopted a simple template matching algorithm with distance-based scores to accomplish a compact AWUWSR system on embedded platforms. Jwu-Sheng Hu[5] utilized the consistency of spatial eigenspaces formed by the speech source at different frequencies and the resonant curve similarity of WUW as the features. The method of pattern matching has a low computational complexity and is applicable on the small-resource embedded platforms. In the scenarios where WUW is coming from a legitimate user, such algorithms are economical and practical. However, limited by templates, it does not have the good generalization. To compensate for that defect, A.Zehetner[6] used the Euclidean distance and cross-correlation between MFCCs of the current audio signal and the keyword template in addition to DTW. The algorithms based on the ASR framework usually train a refined acoustic model and then do Viterbi decoding. Hyeopwoo Lee[7] implemented a voice trigger system using the keyword-dependent speaker recognition and compared the template-based method and hidden Markov model (HMM) based method. Namgook Cho's research[8] was also based on the ASR framework and developed an enhanced-voiced activity detection (E-VAD) to improve the efficiency in a continuous listening environment. Because of its better robustness and generalization, the method based on the ASR framework receives more attention. Researchers[9-14] did lots of experiments with English data and investigated many features, including MFCC, LPC and ENH-MFCC. Chih-Ti Shih[15] studied prosodic features and devoted them to this area. In previous work, they still adopt a GMM-HMM structure to model the acoustic space where GMM denotes the output probability distribution. Moreover, there are no experimental results in the condition of low SNR.

DNN-based acoustic models have provided a significant performance improvement in ASR[16]. It is mainly due to that DNN can learn the complex nonlinear relationship between the input and targets. This paper presents a novel DNN based AWUWSR algorithm, which is quite suitable for the continuous listening noisy environment in a far-field pick-up pattern. In this algorithm, a customized decoding network and a two-stage classification strategy are proposed. They can help to achieve a high accuracy and control the emergence of false alarms effectively.

This work was done during the first author's visiting stay at Georgia Institute of Technology in 2016-2017.

2. THE FRAMEWORK OF AWUWSR

The framework is shown in Figure 1. Firstly, a continuous audio stream is processed by the speech enhancement algorithm [17]. Then VAD listens continuously to detect the presence and absence of human speech and acoustic events, with the aim of filtering out acoustic events, silence and noise, and yielding speech segments for the back-end detection. The perceptual linear predictive (PLP) feature of speech segments is extracted and input to the decoder. The decoder cooperates with the acoustic model and completes the soft alignment for phoneme boundaries. With the decoder output, variety of confidence measures are calculated for each pronunciation unit from multiple perspectives. Finally, a classification module is used to determine whether the voice fragment is WUW or not. In VAD we use a self-learning parameter modification strategy, which can automatically optimize parameters according to the SNR change. Thus non-speech can be discarded as much as possible and more speech can be preserved. If the speech segment obtained is far shorter or longer than the normal speech length of WUW, it is forcibly judged as non-WUW and thrown away in order to remove unnecessary back-end processing. The details of other modules are demonstrated in the following sections.

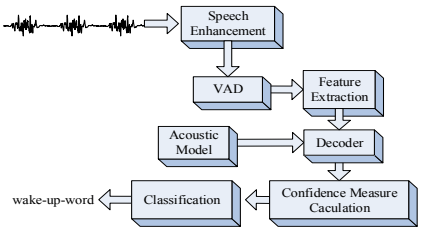

Figure 1. The diagram of the AWUWSR system

3. ACOUSTIC MODELING WITH DNNS

In a DNN-HMM hybrid system, DNNs are trained to provide posterior probability estimates for the HMM states. Specifically, for an observation O_{ut} corresponding to time t in utterance u, the output $y_{ut}(s)$ of the DNN for the HMM state s is obtained using the softmax activation function:

$$y_{ut}(s) \cong P(s|O_{ut}) = \frac{\exp\{a_{ut}(s)\}}{\sum_{s'}\exp\{a_{ut}(s')\}} \quad (1)$$

where $a_{ut}(s)$ is the activation at the output layer corresponding to state s. The recognizer uses a pseudo log-likelihood of state s given observation O_{ut},

$$\log p(O_{ut}|s) = \log y_{ut}(s) - \log P(s) \quad (2)$$

where $P(s)$ is the prior probability of state s calculated from the training data.

The networks are trained to optimize a given training objective function using the standard *error back propagation* procedure[18]. Typically, cross-entropy is used as the objective and the optimization is done through stochastic gradient descent (SGD). For any given objective, the important quantity to calculate is its gradient with respect to the activations at the output layer. The gradients for all the parameters of the network can be derived from this one quantity based on the back propagation procedure.

3.1. Cross-entropy (CE)

For multi-class classification, it is common to use the negative log posterior as the objective:

$$F_{CE} = -\sum_{u=1}^{U}\sum_{t=1}^{T_u}\log y_{ut}(s_{ut}) \quad (3)$$

where s_{ut} is the reference state label at time t for utterance u. This is also the expected cross-entropy between the distribution represented by the reference labels and the predicted distribution $y(s)$. The necessary gradient is:

$$\frac{\partial F_{CE}}{\partial a_{ut}(s)} = -\frac{\partial \log y_{ut}(s_{ut})}{\partial a_{ut}(s)} = y_{ut}(s) - t\delta_{s;s_u} \quad (4)$$

where $\delta_{s;s_u}$ is the Kronecker delta function. Minimizing the cross-entropy is the same as maximizing the mutual information between $y(s)$ and $\delta_{s;s_u}$ computed at the frame-level.

3.2. State-level minimum Bayes risk (sMBR)

While minimizing F_{CE} minimizes expected frame-error, the MBR family of objectives are explicitly designed to minimize the expected error corresponding to different granularity of labels [19]:

$$F_{MBR} = \sum_u \frac{\sum_W p(O_u|S)^{\alpha} P(W) A(W,W_u)}{\sum_{W'} p(O_u|S)^{\alpha} P(W')} \quad (5)$$

where $\mathrm{A}(\mathrm{W},\mathrm{W_u})$ is the raw accuracy, that is, the number of correct state labels corresponding to the word sequence W with respect to that corresponding to the reference W_u. Differentiating (5) w.r.t. $\log p(O_{ut}|r)$, we get:

$$\frac{\partial F_{MBR}}{\partial \log p(O_{ut}|r)} = \alpha\gamma_{ut}^{DEN}(r)\{\bar{A}_u(s_t = r) - \bar{A}_u\} = \alpha\gamma_{ut}^{MBR}(r) \quad (6)$$

where $\bar{A}_u(s_t = r)$ is the average accuracy of all paths in the lattice for utterance u that pass through state r at time t; $\bar{A}_u$ is the average accuracy of all paths in the lattice; and $\gamma_{ut}^{MBR}(r)$ is the MBR 'posterior' as defined in [20]. Like before, we get:

$$\frac{\partial F_{MBR}}{\partial a_{ut}(s)} == \alpha\gamma_{ut}^{MBR}(s) \quad (7)$$

4. DECODING NETWORK

In a far-field pick-up pattern, the target speech is often contaminated by variety of acoustic events, especially when there are some other people talking. In order to filter out interferences, we design a special network to implement Viterbi decoding, which is shown in Figure 2. The network centers in the phoneme sequence of WUW and there is a recyclable sub-network of filler phonemes in both ends respectively. The WUW path can be skipped during decoding. For example, if '你好电视' is the WUW, its central line is the phoneme sequence 'n-i3-h-ao3-d-ian4-sh-i4'. The recyclable sub-network is constructed with several parallel filler phonemes and a loopback path. Thus it can absorb both speech interference and noise perfectly and accomplish a soft alignment for five kinds of audio segments, which are only

WUW speech, WUW speech with front interferences and/or noise, WUW speech with rear interferences and/or noise, WUW speech with both front and rear interferences and/or noise, and pure interferences and/or noise. Those five types cover all the potential segments to be detected. Therefore, the phoneme boundary of WUW can be aligned more accurately and segments of pure interferences and/or noise are removed.

Filler phonemes are obtained through statistics and analysis of the pronunciation variations with plenty of speech data [21]. According to the correlation between them, all common phonemes are clustered into several classes, and each class is a filler. Additionally, there are some other filler phonemes representing different types of environmental noise.

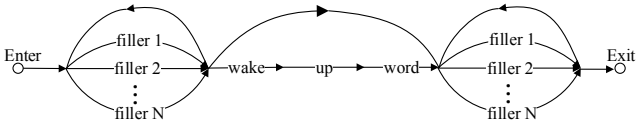

Figure 2. The architecture of decoding network

5. THE TWO-STAGE CLASSIFICATION STRATEGY

5.1. Confidence measures

In this paper we present six effective confidence measures from different perspectives, which are the normalized phoneme duration, the normalized phoneme log-likelihood, the phone log-posterior probability (PLPP)[22], the state number of one frame, the shortest syllable duration and the total duration of WUW.

The normalized phoneme duration is formulated as

$$dur_{NOR}(p_i) = \frac{dur(p_i)}{\sum_{i=0}^{S} dur(p_i)} \quad (8)$$

where p_i is the ith phoneme of WUW, $dur(p_i)$ is the duration of phoneme p_i, and S is the total phoneme number of WUW. The normalized phoneme log-likelihood can be expressed as

$$LL_{NOR}(p_i) = \frac{ln(P(O|p_i))}{dur(p_i)} \quad (9)$$

where $P(O|p_i)$ is the likelihood of phoneme p_i, and $ln(P(O|p_i))$ can be obtained from the decoder output.

PLPP is calculated by Eq. (10)

$$P_{PLPP}(p_i|O) = \frac{ln\, P(p_i|O)}{dur(p_i)} = \frac{ln\left(\frac{P(O|p_i)P(p_i)}{\sum_{q\in Q} P(O|q)P(q)}\right)}{dur(p_i)} \approx \frac{ln\left(\frac{P(O|p_i)}{\sum_{q\in Q} P(O|q)}\right)}{dur(p_i)} \quad (10)$$

where $\sum_{q\in Q} P(O|q)$ is the sum of all phoneme likelihood in the similar phoneme set Q of p_i.

5.2. The first classification stage

With confidence measures, we firstly utilize a pre-decision algorithm, named as the first stage classification. It filters out a large number of acoustic events which is similar to WUW. The pre-decision process is demonstrated as Figure 3. Each threshold in the chart can be previously obtained through empirical knowledge and statistics analysis. If not sentenced to non-WUW in this process, it will be sent into the next stage.

5.3. The second classification stage

Segments which are not filtered out in the first classification stage, usually have higher correlation with the correct WUW pronunciation. Therefore, we take a more sophisticated classification method for them in the second stage.

In this paper we adopt support vector machine (SVM) classifier. All the confidence measures for each phoneme and the whole segment are concatenated in order and become a confidence vector as the input of SVM classifier. For instance, if the WUW is "你好电视", it contains four syllables, which are equal to eight phonemes. According to Section 5.1, the confidence vector has 8*3+3 elements. The classifier output is the final detection result of whether it is WUW or not.

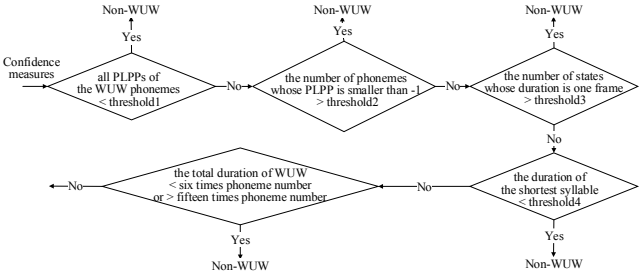

Figure 3. The flow chart of the first classification stage

6. EXPERIMENTAL RESULTS

6.1. Evaluation metrics

Similar to keyword spotting, the performance of AWUWSR contains the accuracy and false alarm rate (FAR). To monitor a continuous audio stream, there are thousands of segments per hour. So FAR is no longer appropriate to indicate false alarms. In this paper, we use the number of false alarms per hour and call it as false alarm frequency (FAF).

6.2. Datasets and experiment settings

We use about 200 hour's speech as the train set. It is recorded ourselves by an array with two microphones in a room and the room size is about 5*6*3 meters. The distance between speaker's mouth and microphone array is about 3 meters, and it covers 500 persons, including 250 males and 250 females. Every speaker needs to speak a word/phrase/sentence three times with the speed of fast, medium and slow respectively. Environmental noise comes from random TV shows and room reverberation, and the sound pressure level (SPL) ranges from 0db to 80db. In fact, SPL with 80db represents a quite noisy condition and it is far higher than that of our general home environment. The data is processed by the speech enhancement module mentioned in Section 2 before training acoustic model. The test data contains positive and negative sample set, which are both obtained in home environment with smart TV as the receiving device and accompanying with various noise. The positive sample set is used for the accuracy and negative sample set is for FAF. The former includes three subsets (named as subset1, subset2 and subset3) whose SPLs are 10db, 70db and 80db respectively. Each subset contains 260 samples. The latter has 12391

samples amounting to about 25 hours. Additionally, we take the online test in which it runs on a playing smart TV and there are other neighbor TVs playing programs. The SPL reaches about 80db. There are 20 testers, 10 males and 10 females. Each tester speaks WUW 20 times. Moverover, We make the proposed system running continuously for 48 hours and observe the occurrence number of false alarms.

With the HMM framework, the acoustic model includes 65 Chinese phonemes without tones, a 'sil' for silence, a 'sp' for short pause, a garbage phoneme for background noise and 15 fillers. There are 3893 tied-states labeled as senones. 13 PLPs and their first and second derivatives, with a context window of 7 frames (central frame +3 context frames), are used as parametric representation of the speech signal. Furthermore, the classic cepstral mean subtraction is applied. Thus, the speech vector has 273 components (39×7) as the input to DNN model. The DNN is constructed with five hidden layers. Each hidden layer has 512 neurons and the output softmax layer has 3893 output units (according to 3893 senones). The following scheme is used for training DNN. It is initialized randomly. All the utterances and frames are randomized before being fed in the DNN. The mini-batch size is set to 256 and the initial learning rate is set to 0.008. Then the network is discriminatively trained with the objective function of sMBR using back propagation. After each training epoch, we validate the frame accuracy on the development set, if the improvements is less than 0.5%, we shrink the learning rate by the factor of 0.5. The training process is stopped after the frame accuracy improvement is less than 0.1%. General purpose graphics processing units (GPUs) are utilized to accelerate the training process. For performance comparison, we also train the traditional GMMs [9-15] and DNNs based on the objective function of CE. Each GMM contains eight mixtures and the DNN based on CE is constructed in the same way as that based on sMBR.

Table 1. Performance comparison with different algorithms

algorithm	Accuracy			FAF
	subset1	subset2	subset3	
baseline	97.1%	81.9%	55.7%	0.50
System1	100%	90.5%	66.9%	0.28
System2	100%	100%	80.8%	0.14
System3	100%	100%	86.7%	0.12

6.3. Results

In this paper we use the method proposed by Veton Kepuska [10] as the baseline. The approaches with the framework in Section2 covers three systems. System1 utilizes HMM-GMM based acoustic model. System2 and System3 use HMM-DNN with CE and sMBR respectively. Because of the mutual constraints between accuracy and FAF, we adjust FAF to a lower level firstly and then do the performance comparison. Table 1 shows the results of different algorithms. Comparing the results of baseline and System1, we learn that System1 increases the accuracy with 2.9%, 8.6% and 11.2% in three subsets respectively. The accuracy reaches one hundred percent in subset1 whose SPL is 10db. Meanwhile, FAF is nearly cut in half, that is, one false alarm occurs every four hours. It is mainly attributed to the customized decoding network and the two-stage classification strategy which can control false alarms greatly. When replacing GMMs with DNNs, there are absolute 9.5% and 13.9% improvement in subset2 and subset3 respectively. Meanwhile, false alarms has a large-scale reduction from 0.28 to 0.14 times per hour, equivalent to once every seven hours. When we adopt sMBR as the objective function, the performance is improved further. The accuracy surpasses the *CE*-based system with 7.9% absolutely and reaches 86.7% in Subset3 whose SPL is about 80db. FAF remains stable or even drops slightly, which is less than once per eight hours. It implies that the optimization criterion of sMBR has a better ability of discrimination and can be applied to AWUWSR successfully.

Table 2. Online test results with the proposed system

algorithm	Accuracy				FAF
System3	19/20	18/20	18/20	20/20	0.104
	15/20	19/20	20/20	18/20	
	18/20	20/20	16/20	19/20	
	17/20	20/20	18/20	18/20	
	19/20	15/20	20/20	19/20	
	Average Accuracy = 91.5%				

To verify the performance on the actual products, we conduct an online test. The results are shown in Table 2. The accuracy for all testers and their average value display in the middle. The average accuracy is 91.5%, which is much higher than the result of subset3 in Table 1. Maybe the testers are more user-friendly when using it and speech subconsciously in quieter gap. While the proposed system runs for 48 hours continuously, five false alarms come out. On average, FAF is 0.104. In other words, it provides higher than 90% accuracy and meanwhile the false alarms are less than once per nine hours in the severe noisy environment.

7. CONCLUSION

This paper proposes a novel AWUWSR algorithm for the far-field pick-up pattern and noisy environments. During training the acoustic model, we utilize sMBR as the objective function to make it more discriminative. The customized decoding network can absorb background noise and interference greatly. Moreover, a two-stage classification strategy is adopted, in which the pre-decision makes use of phonetic knowledge to filter out many anomalous fragments that are difficult to be identified by the model-based classifier. Discarding interference audio clips in advance can remove lots of unnecessary calculation. With the experiments, it is learned that the proposed algorithm can improve the performance significantly and meet the needs of smart home applications. In the future, we will further explore how to utilize long-time spectral information of WUW speech, such as making use of more history and future frames with a novel DNN architecture and analyzing its long-rhythm structure.

8. REFERENCES

[1] L. R. Rabiner and B. H. Juang, "Statistical Methods of Speech Recognition", Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005.

[2] Ge F., Pan F., Liu C., "Some acoustic improvements for pronunciation quality assessment for strongly accented mandarin speech", International Conference on Audio, Language and Image Processing. IEEE, 2008:691-696.

[3] Liu C., Pan F., Ge F., Dong B., Zhao Q., Yan Y., "Application of LVCSR to the Detection of Chinese Mandarin Reading Miscues", 2008 Fourth International Conference on Natural Computation-volume (Vol.5, pp.447-451). IEEE Computer Society.

[4] Anhao Xing, Ta Li, Jielin Pan, Yonghong Yan, "Compact Wake-up Word Speech Recognition on Embedded Platforms", Applied Mechanics and Materials ISSN: 1662-7482, Vol. 596, pp 402-405.

[5] Jwu-Sheng Hu, Ming-Tang Lee, Ting-Chao Wang, "Wake-Up-Word Detection for Robots Using Spatial Eigenspace Consistency and Resonant Curve Similarity", International Conference on Robotics and Automation, 2014

[6] A. Zehetner, M. Hagmiiller, F. Pernkopf, "WAKE-UP-WORD SPOTTING FOR MOBILE SYSTEMS", International Conference on EUSIPCO, 2014

[7] Hyeopwoo Lee, Sukmoon Chang, Dongsuk Yook, Yongserk Kim, "A voice trigger system using keyword and speaker recognition for mobile devices", Consumer Electronics, IEEE Transactions on, 2009, 55(4): 2377-2384.

[8] Namgook Cho, Taeyoon Kim, Sangwook Shin, Eun-Kyoung Kim, "Voice Activation System Using Acoustic Event Detection and Keyword/Speaker Recognition", 2011 IEEE International Conference on Consumer Electronics (ICCE), 2011: 21-22.

[9] Kepuska V Z, Eljhani M M, Hight B H., "Wake-Up-Word Feature Extraction on FPGA", World Journal of Engineering and Technology, 2014, 2, 1-12.

[10] Veton Kepuska, "Wake-Up-Word Speech Recognition", Speech Technologies, June 2011, pp.237-263.

[11] Veton Kepuska, "Wake-up-word recognition", 6 October 2010, SPIE Newsroom. DOI: 10.1117/2.1201009.003154.

[12] Kepuska V Z, Klein T B, "A novel wake-up-word speech recognition system, wake-up-word recognition task, technology and evaluation", Nonlinear Analysis: Theory, Methods & Applications, 2009, 71(12): e2772-e2789.

[13] Veton Kepuska, "Wake-up-word speech recognition application for first responder communication enhancement", Jason. Proc.SPIE 6201, Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense V, 62011E (May 10, 2006);

[14] Veton Z. Kepuska, Mohamed M. Eljhani, Brian H. Hight,Int. "Voice Activity Detector of Wake-Up-Word Speech Recognition System Design on FPGA", Journal of Engineering Research and Applications ISSN: 2248-9622, Vol.4, Issue 12(Part 3), December 2014, pp.160-168

[15] Chih-Ti Shih, "Investigation of Prosodic Features for Wake−Up-Word Speech Recognition Task", Florida Institute of technology, 2009，Master's thesis.

[16] Karel Vesel´y, Arnab Ghoshal, Luk´aˇs Burget, Daniel Povey, "Sequence-discriminative training of deep neural networks", Interspeech, 2013

[17] Xiaofei Wang, Yanmeng Guo, Fengpei Ge, Chao Wu, Qiang Fu, Yonghong Yan, "Speech-picking for speech systems with auditory attention ability", SCIENCE CHINA PRESS, 2015, vol.45(10), pp.1310-1327.

[18] D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, pp.533–536, October 1986.

[19] M. Gibson, T. Hain, "Hypothesis spaces for minimum Bayes risk training in large vocabulary speech recognition," in Proc. INTERSPEECH, September 2006, pp. 2406–2409.

[20] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, Cambridge, UK, 2003.

[21] Ge F., Pan F., Liu C., Dong B., Yan Y., "Forward optimal modeling of acoustic confusions in Mandarin CALL system", InINTERSPEECH-2008, 2815-2818.

[22] Ge F., Liu C., Shao J., Pan F., Dong B., Yan Y., "Effective acoustic modeling for pronunciation quality scoring of strongly accented mandarin speech". IEICE Trans.inf. & Syst, 2008,91(10), 2485-2492.