Overview of speech keyword recognition technology

Sixuan Li1, Guanyu Li1*, Jiao Han1, Tiankai Zhi1

1Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China

*Corresponding author’s e-mail: xxlgy@xbmu.edu.cn

Abstract: Since the 1990s, speech keyword recognition technology has gradually separated from speech recognition technology and officially become an important branch of speech recognition technology. In this paper, several speech keyword recognition technologies are studied and reviewed, including sample recognition methods, filler model methods, basic speech recognition systems based methods, neural network classifiers and end-to-end methods. Finally, the development prospect and advanced technology of speech keyword recognition will be discussed.

1. Introduction

Keyword detection was first studied by Bridle[1] in 1973, when keywords were called "given words", and the term of Keyword was proposed by Christinansen in 1977. Christinansen used LPC (Linear Predictive Coding) method to detect keywords from continuous speech streams and proposed the concept of "keywords"[2]. So far, it has a history of nearly fifty years. With the continuous development of speech keywords, in recent years, it has gradually replaced continuous speech recognition and been applied to many fields such as robot interaction, device voice command control, smart home equipment.

At present, there are three kinds of traditional Keyword Spotting methods. In the first category, speech Keyword Spotting based on the sample[3,4,5,6] mainly uses the Dynamic Time Warping (DTW)[7] algorithm to calculate the similarity between the input speech sequence and the template speech sequence in the database, and the success of Keyword Spotting is determined when a certain threshold is reached[8,9]. The second one is the speech Keyword Spotting based on the whitening model, which processes the language model and makes all words independent into sentences, reduces the size of the speech model, so that all keywords in the decoding graph have independent paths, instead of all keywords sharing a path, that is, garbage edge. The third category is speech keyword recognition based on LVCSR. The trained LVCSR system is used to establish a time factor converter for candidate word cases and generate an inverted index[10]. In recent years, with the development of speech recognition technology, neural network and end-to-end have become the mainstream of speech Keyword Spotting technology. These two methods are an upgrade and extension of technology based on traditional methods.

2. Speech keyword spotting technology based on sample

Based on the keyword spotting of the sample, the problem is considered as a matching problem. Consider the audio sample of the keyword, and several test audio, and calculate their similarity respectively. If the similarity between the test audio and the keyword exceeds a certain threshold, it is
considered to be the detected keyword. In this way, users can record their own audio and define it as keywords, making it more personalized.

The keyword spotting based on the sample can be divided into two categories. One is based on Dynamic Time Warping (DTW) algorithm, which uses DTW algorithm to calculate the similarity between two audio feature sequences. The other is based on embedded learning, which encodes two audio frequencies as vectors and calculates the distance between the two vectors directly. The method based on DTW has been applied since the 1970s, but it has a high computational complexity when matching two sequences, and is mainly used in unsupervised situations at present. Based on the method of embedded learning\textsuperscript{[11,12]}, it is simpler to match, and becomes popular after deep learning becomes popular.

In the time series, the length of the two time series that need to compare the similarity may not be the same\textsuperscript{[13]}, which is shown in the field of speech recognition that different people speak at different speeds. And different phonemes within the same word can be pronounced at different speeds. For example, some people will make the "A" sound very long, or the "I" very short. In addition, different time series may only have displacement on the time axis, that is, in the case of reduction displacement, the two time series are consistent. In these complex cases, the distance (or similarity) between two time series cannot be effectively determined using the traditional Euclidean distance. DTW calculates the similarity between the two time series by extending and shortening the time series\textsuperscript{[14,15]}:

Order to compute the similarity of two time series for $X$ and $Y$, the length of the $|X|$ and $|Y|$ respectively.

The form of the rectified path is $W=w_1, w_2, ..., w_k$, among them

$$\text{Max}(|X|, |Y|) \leq K \leq |X| + |Y|$$

(1)

The form of $w_k$ is $(i, j)$, where $i$ represents the $i$ coordinate in $X$ and $j$ represents the $j$ coordinate in $Y$.

Attribute path $W$ must begin from $w_1 = (1,1)$, to $w_k = (|X|, |Y|)$ at the end, to ensure that each of the $X$ and $Y$ coordinates are in the $W$.

In addition, the $i$ and $j$ of $w(i, j)$ in $W$ must be monotonically increasing to ensure that the dashed lines do not intersect. The so-called monotonically increasing means:

$$w_k = (i, j), w_{k+1} = (i, j) \quad i \leq i \leq i + 1, j \leq j \leq j + 1$$

(2)

The final aggregation path is the one with the shortest distance:

$$D(i, j) = \text{Dist}(i, j) + \min\{D(i - 1, j), D(i, j - 1), D(i - 1, j - 1)\}$$

(3)

Obtained under control path distance of $D(|X|, |Y|)$, the use of dynamic programming for solving.

3. Filler model

The complement model, sometimes referred to as the spam model, considers the Keyword Spotting problem as a frame-by-frame sequential tagging problem. The keywords are assigned to different annotations, and an additional "whitening" annotation is used to match all non-keywords.

HMM was first used in Keyword Spotting\textsuperscript{[16]}. It builds a hidden Markov model for each keyword and an additional hidden Markov model for non-keywords, and the observation probability is modeled by mixed Gaussian or neural network. Directly for keyword modeling in data sparsity problems. At present, the popular hidden markov model adopts subword units, such as phonemes, for modeling. In this case, it is very similar to the acoustic model used in speech recognition based on the HMM hybrid model, except that the decoding diagram is a hand-designed grammar rather than a statistical language model generated. The Keyword Spotting system used by Amazon Alexa voice assistant is based on this kind of method\textsuperscript{[17]}. There are mainly two kinds of search networks based on the whitening model:
Generally, the first search network does not limit the number of keywords, and the second search network limits the number of keywords to 1. After the keyword model and the whitening model are obtained by training, the search network is formed. The Viterbi algorithm\cite{18} is used to match the detection and model, and the best matching results are output.

4. Keyword spotting based on LVCSR

Given a pre-trained ASR system and key phrases to be recognized. The recognition Vocabulary Of the ASR system may not include keywords in advance, but the basic phonemes must be able to synthesizing keywords, so that keywords can be added to the Vocabulary and even the language model. Therefore, the ASR can effectively solve the problem Of words Out Of Vocabulary (OOV). LVCSR part: decodes the keyword set of the query and generates the corresponding lattice. Key words index part:

1) Preprocessing: Each input automaton is preprocessed to obtain a posterior Lattice, in which the non-overlapping arc sets are marked respectively;
2) Construct a time factor converter: for each processed input automata, construct an intermediate factor converter that can accurately identify the input factor subset;
3) Factor selection: these intermediate factor converters are converted into deterministic converters by extending each factor with disambiguation symbols, and then applying weighted automata optimization;
4) Search in time factor converters: these deterministic converters are combined and further optimized To obtain the deterministic inversion index of the entire data set.

In this method, the strategy to solve OOV is to use the proxy-word method, find a series of words in the vocabulary that have similar pronunciation with the keyword, and use these words as the proxy-words of the keyword. Proxy keyword fuzzy search method is used to solve the OOV keyword problem and improve the IV(in-vocabulary) keyword problem. The generation formula is as follows\cite{19}: 
\[ K' = \text{Project}\left(\text{ShortestPath}\left(\text{prune}\left(\text{prune}\left(K \cdot L_2 \cdot E^\prime\right) \cdot L_1^{-1}\right)\right)\right) \] (4)

Where \( K' \) is the original keyword, \( L_2 \) is a dictionary containing \( K \) pronunciation, and \( E^\prime \) is the edit distance converter, including the degree of pronunciation confusion calculated by the training set. \( L_1 \) is the original dictionary. \( K' \) is a WFST containing a series of words with similar sounds to \( K \), and then these similar words are used as keywords for searching. In the above formula, you need prune, especially if you have a large vocabulary.

5. immediate development

5.1 The neural network

In the past few years, keyword spotting has taken off, thanks to the development of neural networks. In addition to the basic whitening model based on hidden horse, a whitening model directly classified by DNN appeared in the later stage\(^{[20]}\). Continuous speech streams are fed segment by segment into the neural network for classification. The category is all command words, and an additional Filler category, such as the 10 command words, has 11 classes. After the classification is completed, a smoothing post-processing is performed because the output probability may appear "burr", and then if the probability of a certain category exceeds a threshold, a certain command word is considered to have been detected. This method occupies little memory, does not need decoding search, and has high accuracy. However, because a large number of corpus containing command words need to be prepared, if the command words are changed, another batch of corpus needs to be collected, so it is difficult to use in practice. In contrast, the Keyword Spotting based on HMM is more commonly used because it is for the modeling of subword units and the corpora can be used in general.

5.2 End-to-end

In recent years, end-to-end models have emerged and become a hot topic in current research. In speech tasks, end-to-end ends generally refer to the two ends of the neural network, that is, the given input, the output of the neural network is the target result. KWS system based on within DNN - HMM, we can transform the language model, makes all keywords are independent of the path, and all the keywords share a path ("junk" edge), similarly, we can transform the basic output unit for word end-to-end ASR, namely all keywords have independent output node, and all the key words share an output node ("junk" nodes). Fig. 3 shows the neural network structure of end-to-end keyword recognition, which is a standard classifier and can be trained by directly using the Cross Entropy loss function\(^{[21]}\).

![Fig. 3 End-to-end keyword recognition](image-url)
6. Summary and outlook
With the development of science and technology as well as the Internet, and the rise of voice control fields such as smart home, the importance of voice keyword recognition technology is self-evident. The current mainstream approaches are based on the above three traditional methods, which have led to the development of neural network classifiers, end-to-end speech keyword recognition techniques, and, more recently, streaming serial-to-sequence models. In short, speech keyword recognition technology is becoming more and more accurate and efficient.

Acknowledgments
This work was supported by The National Natural Science Fund (NO.61633013).

References
[1] J.S.Bridle An efficient elastic-template method for detecting given words in running speech,British Acoustical Society Meeting ,1973.1-4.
[2] Christiaensen R,Rushforth C.Detecting and locating key words in continuous speech using linear predictive coding[J].IEEE Trans.on Acoust.speech & Signal Processing,1977,25(5):361-367.
[3] Saon G , Chien J T . Large-Vocabulary Continuous Speech Recognition Systems[M]// Fundamentals of speech recognition / . Tsinghua University Press, 2012.
[4] Chelba C , Hazen T J , Saraclar M . Retrieval and browsing of spoken content[J]. Signal Processing Magazine IEEE, 2008, 25(3):39-49.
[5] Higgins A L . Keyword recognition using template concatenation[J]. 1985.
[6] Motlicek P , Valente F , Szoke I . Improving Acoustic based Keyword Spotting using LVCSR Lattices[C]// Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP). IEEE, 2012.
[7] Vintsyuk T K. Speech Discrimination by Dynamic Programming.Cybernetics and Systems Analysis,1968,4(1):81—88.
[8] Saraclar M . Lattice-based Search for Spoken Utterance Retrieval[J]. Proc. of HLT-NAACL, 2004, 2004.
[9] Rohlicek J R , Russell W , Roukos S , et al. Continuous hidden Markov modeling for speaker-independent word spotting[C]// International Conference on Acoustics. IEEE, 1989.
[10] Can D, Saraclar M. Lattice Indexing for Spoken Term Detection[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(8): 2338-2347.
[11] Chen G,Parada C, Sainath T N, et al. Query-by-example keyword spotting using long-short-term memory networks[C]. international conference on acoustics, speech,and signal processing, 2015: 5236-5240.
[12] KamperH, Wang W, Livescu K, et al. Deep convolutional acoustic word embeddings using word-pair side information[J]. international conference on acoustics, speech, and signal processing, 2016: 4950-4954.
[13] Sakoe H,Chiba S. Dynamic programming algorithm optimization for spoken word recognition[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing,1978, 26(1): 159-165.
[14] ItakuraF. Minimum prediction residual principle applied to speech recognition[J]. IEEETransactions on Acoustics, Speech, and Signal Processing, 1975, 23(1): 154-158.
[15] Mantena G V, Achanta S, Prabhallad K, et al. Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2014,22(5): 946-955.
[16] Wilpon JG, Lee C, Rabiner L R, et al. Application of hidden Markov models for recognition of a limited set of words in unconstrained speech[C]. international conference on acoustics, speech, and signal processing, 1989: 254-257.
[17] Sun M,Snyder D, Gao Y, et al. Compressed Time Delay Neural Network for Small-Footprint Keyword Spotting.[C]. conference of the international speechcommunication association, 2017: 3607-3611.
[18] Viterbi AJ. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory. 1967. April, 13 (2): 260–269. doi:10.1109/TIT.1967.1054010.

[19] Chen G, Yilmaz O, Trmal J, et al. Using proxies for OOV keywords in the keyword search task[C]. IEEE automatic speech recognition and understanding workshop, 2013: 416-421.

[20] Chen G, Parada C, Heigold G, et al. Small-footprint keyword spotting using deep neural networks[C]. International Conference on Acoustics, Speech, and Signal Processing, 2014: 4087-4091.

[21] Guoguo Chen, Carolina Parada, and Georg Heigold. “Small-footprint keyword spotting using deep neural networks”. In: IEEE International Conference on Acoustics, Speech and Signal Processing. 2014, pages 4087–4091 (cited on page 115).