End-to-end Keywords Spotting Based on Connectionist Temporal Classification for Mandarin

[1,3]Ye Bai, [1]Jiangyan Yi, [1]Hao Ni, [1]Zhengqi Wen, [1]Bin Liu,[1]Ya Li, [1,2,3]Jianhua Tao

[1]National Laboratory of Pattern Recognition,
[2]CAS Center for Excellence in Brain Science and Intelligence Technology
Institute of Automation, Chinese Academy of Sciences
[3]School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

baiye2016@ia.ac.cn, {jiangyan.yi, hao.ni, zqwen, liubin, yli, jhtao}@nlpr.ia.ac.cn

Highlights

- In this paper, we construct an end-to-end acoustic model based ASR for keywords spotting in Mandarin.
- This model is constructed by LSTM-RNN and trained with objective of connectionist temporal classification.
- The input of the network is feature sequences, and the output the probabilities of the initials and finals of Mandarin syllables.
- Compared with hybrid based ASR systems, the end-to-end system achieves a improvement of 6.32% on ATWV relatively.

Keywords Spotting Based on CTC

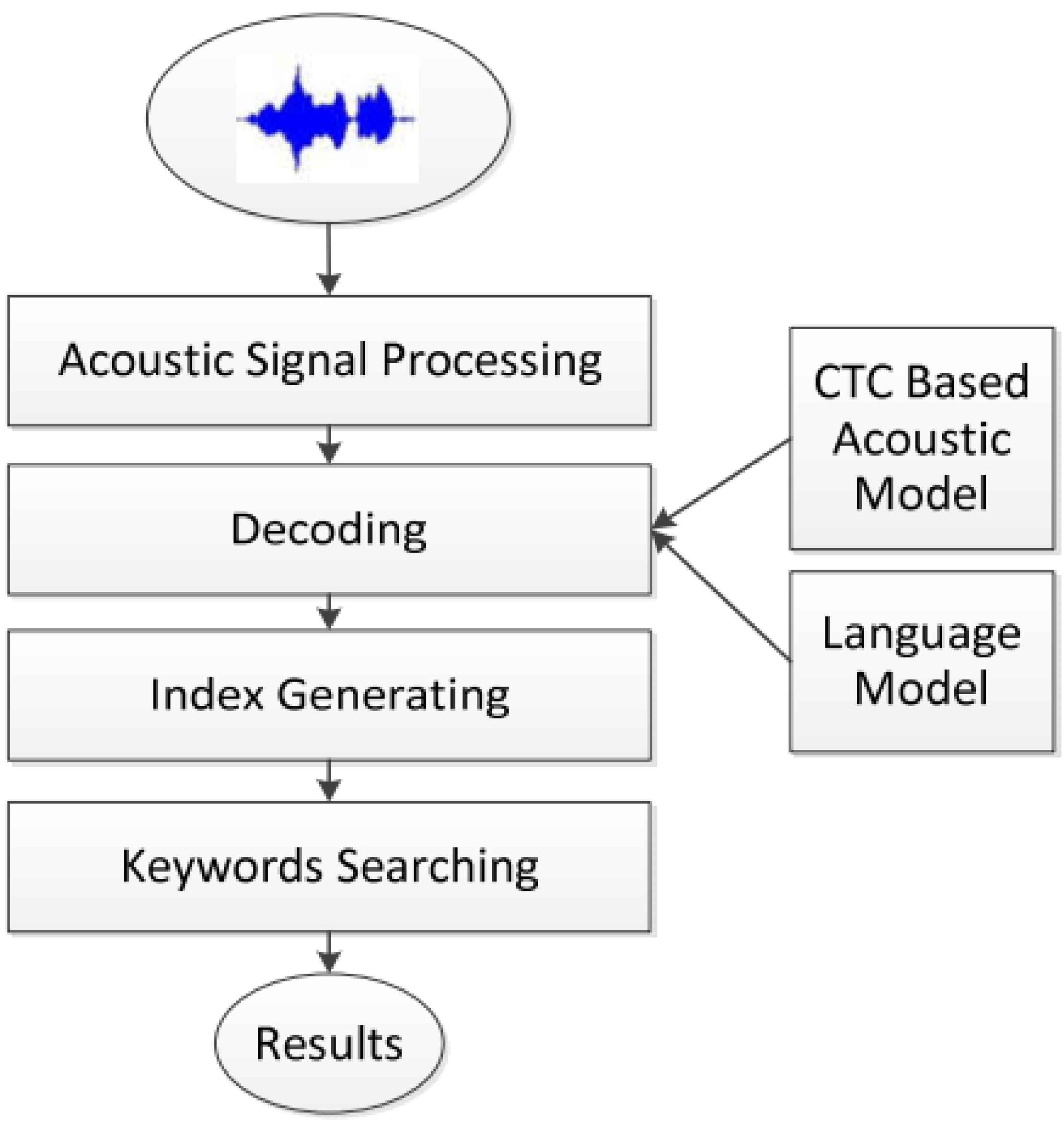

Figure 2: *Illustration of proposed KWS system*

The front-end of keywords spotting system is an ASR system. Then the candidate results of ASR will be converted to an index for searching keywords.

The search index is constructed with timed factor transducer algorithm. The weight of an arc of a timed factor transducer is a three tuple which saves score, start time, and end time.

Searching is divided into two steps. First, compiling the query string to an linear finite state acceptor. And then compose the acceptor with the index. The time information of where the keywords occur can be obtained by projecting the WFST.

Backgrounds

- In non-specific tasks for the KWS, the LVCSR based approach is widely used since that it does not require any prior knowledge about speech for searching the keywords.
- Traditional hybrid model LVCSR system is complicated. The construction of acoustic model is divided into several stages. State level model is constructed without actual meaning in phonetics. It is difficult to bring in knowledge of phonetics for specific language to acoustic model.
- CTC is a direct method for sequence labelling tasks with recurrent neural network model. It can simplify the architecture of LVCSR with a single recurrent neural network (RNN).
- We construct our keywords spotting system based on CTC acoustic model for Mandarin.
- The model is constructed for the initials and finals of Mandarin syllables.

Training CTC Based Acoustic Model

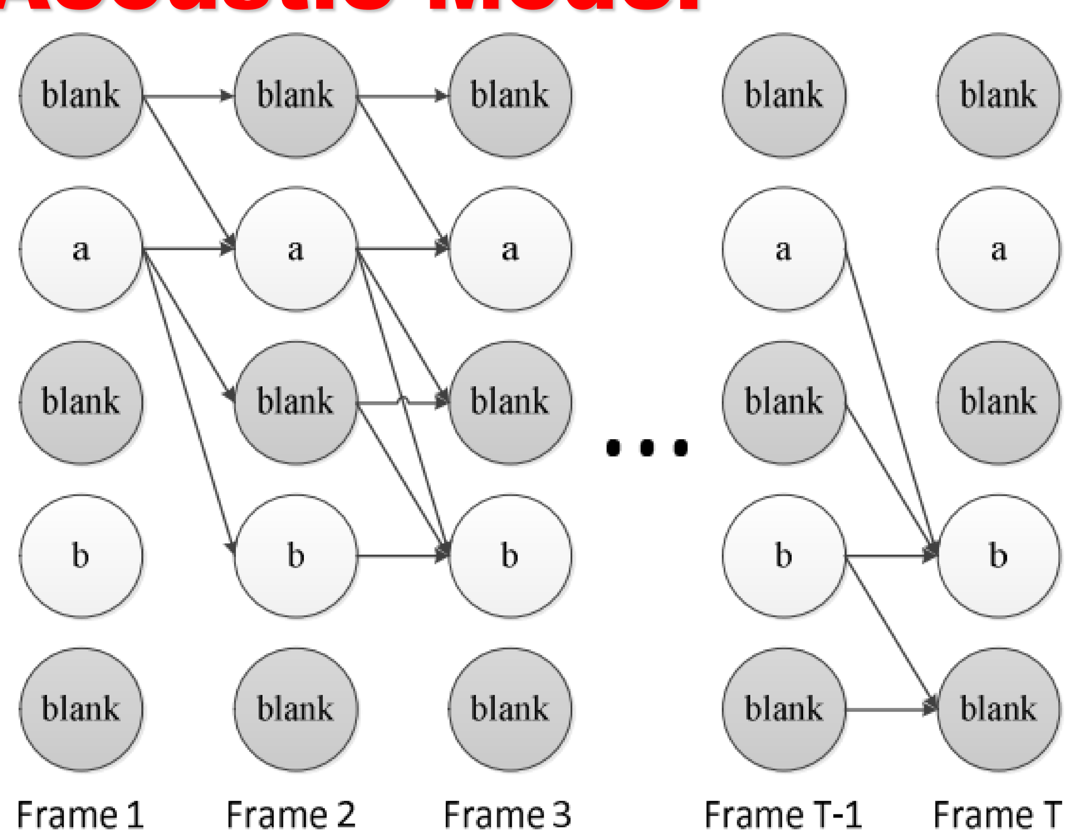

Figure 1: *Trellis of the labelling "ab"*

The objective function is the probability of symbol sequence respect to feature sequence.

$$P(\mathbf{l}|\mathbf{X}) = \sum_{\pi \in B^{-1}(\mathbf{l})} P(\pi|\mathbf{X})$$

Using Forward-Backward algorithm, the probability can be calculated effectively.

$$P(\mathbf{l}|\mathbf{X}) = \sum_{u=1}^{2U+1} \alpha_t(\pi_u)\,\beta_t(\pi_u)$$

The partial derivative of the objective is

$$\frac{\partial \ln P(\mathbf{l}|\mathbf{X})}{\partial y_k^t} = \frac{1}{P(\mathbf{l}|\mathbf{X})}\frac{1}{y_k^t}\sum_{\pi_u \in \{u|\mathbf{l}_u = k\}} \alpha_t(\pi_u)\,\beta_t(\pi_u)$$

Evaluations

Table 1. *Comparison of WER between baseline and CTC approach.*

Model	WER
DNN-HMM	7.12%
CTC(FBANK)	2.60%
CTC(MFCC)	2.06%

Table 2. Comparison of ATWV and MTWV between baseline and CTC approach.

Model	ATWV	MTWV
DNN-HMM	0.7816	0.7853
CTC(FBANK)	0.8225	0.8268
CTC(MFCC)	0.8310	0.8328

- The network consists of four unidirectional LSTM layers, each layer has 320 cells.
- The models are trained on RASC863: 863 annotated 4 regional accent speech corpora [19]. The corpora contains 250 hours of speech in Mandarin. The speech is sampled at 16kHZ.

Acknowledgements

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No.2015AA016305).