中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

A Survey on Keyword Spotting

Institute of Automation, Chinese Academy of Sciences, CASIA
National Laboratory of Pattern Recognition, NLPR
Intelligent Interaction Team
Ye Bai
baiye2016@ia.ac.cn

January 20, 2019

Outline

- Introduction
- Mainstream Approaches
  - HMM/Filler Models
  - Query-by-Example
  - LVCSR Based Methods
- Some Advances
- Take Home Messages

Background

- Keyword Spotting
- Keyword Search
- Spoken Term Detection

It focuses on detecting words which users choose in continues speech.

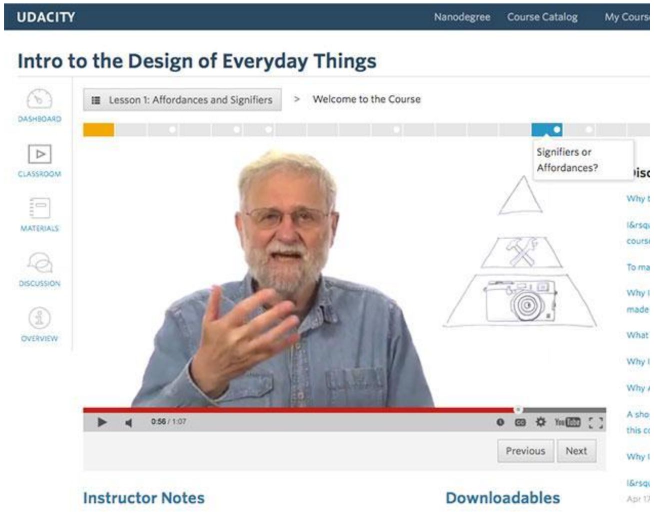
Typical application scenarios

Voice-controlled devices

- Voice-controlled devices active in terms of users' command words.
- A device activated when it receives some wake-up words.

Typical application scenarios

Searching keywords in audio



- For example, we have several hours of audio or video lectures.
- We are interested in some specific audio or video clips.
- We would like to retrieve the entire audio or video document in terms of some keywords.

Two different problems

- Keyword spotting
  - Keywords are usually fixed
  - Small-footprint
  - Efficient computation
  - Low-power consumption
- Spoken term detection (Keyword search)
  - Keywords are changeable
  - Need to locate the keywords in audio
  - Out-of-vocabulary

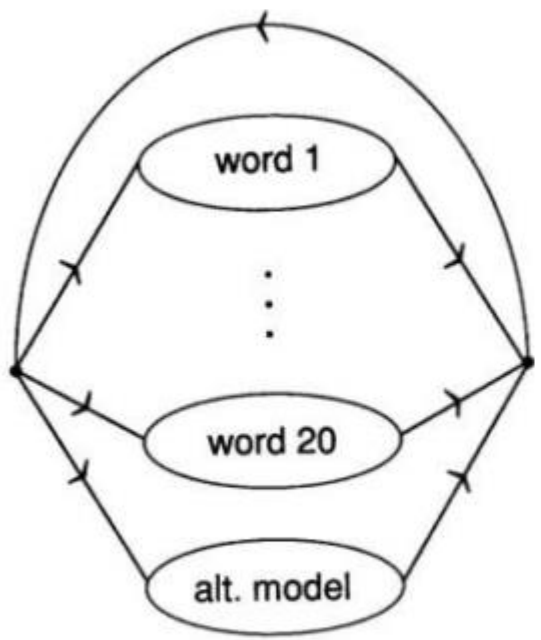
Outline

- Introduction
- Mainstream Approaches
  – HMM/Filler Models
  – Query-by-Example
  – LVCSR Based Methods
- Some Advances
- Take Home Messages

Filler Models

- The filler models are sometimes known as garbage models or acoustic keyword spotting.
- This model can be seen as a framewise sequence labelling problem.
- Keywords and non-keywords are modeled respectively in this approach.
- Filler models are a set of models which can match arbitrary non-keyword speech utterances.

HMM based filler models



- Each keyword and a filler are modeled using HMM respectively.
- Generative probability of a frame of speech parameters given a state of HMMs is estimated with GMMs or DNNs.

Wilpon J G, 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.

HMM based filler models

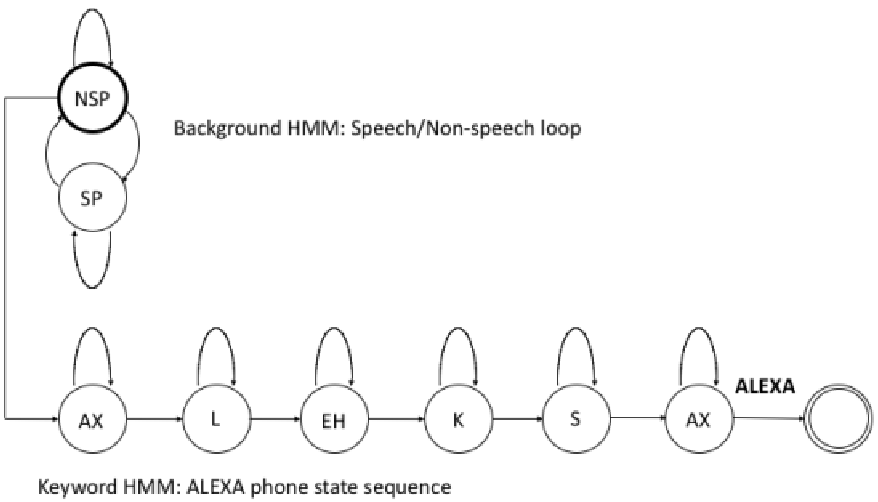
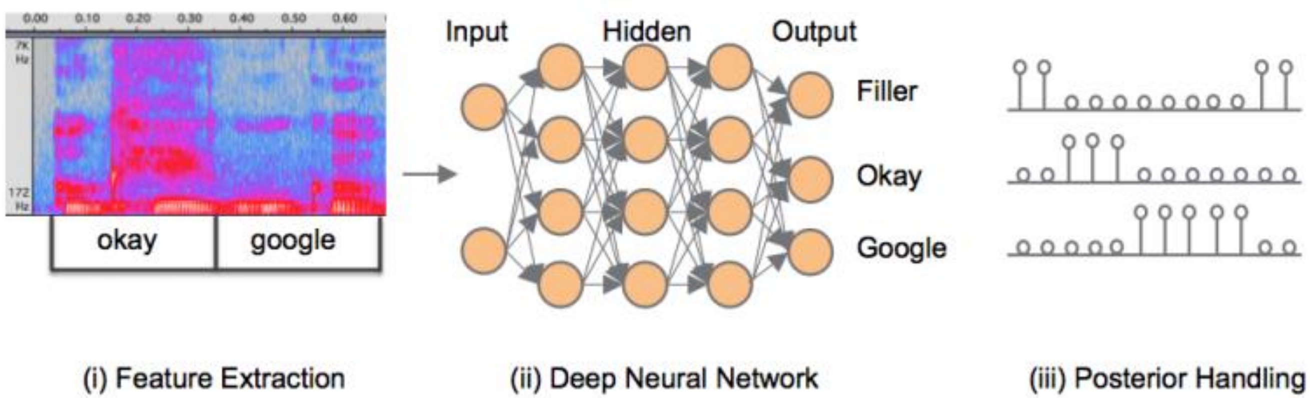


Figure 1: *HMM-based keyword spotting system*

- Each phone is modeled by an HMM model.
- Searching Graph is built with a handcraft phone-level grammar.

Sun M, Snyder D, Gao Y, et al. Compressed Time Delay Neural Network for Small-Footprint Keyword Spotting.[C]. conference of the international speech communication association, 2017: 3607-3611.

DNN based filler models



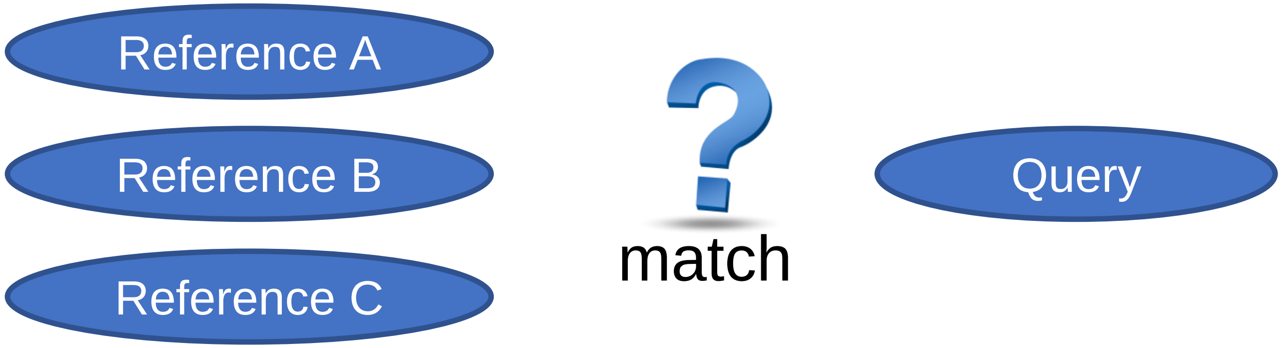
- DNN is used as a framewise classifier.
- Then the posteriors are smoothed with a window.
- The system is used in mobile devices.

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.

Outline

- Introduction
- Mainstream Approaches
  - HMM/Filler Models
  - Query-by-Example
  - LVCSR Based Methods
- Some Advances
- Take Home Messages

Query-by-example methods



- Query-by-example is a task to detect some keywords in a speech signal, where the keywords are saved as patterns.
- Query-by-example methods allow users define their own keywords. It is more personalized for them to control their own devices.

Query-by-example methods

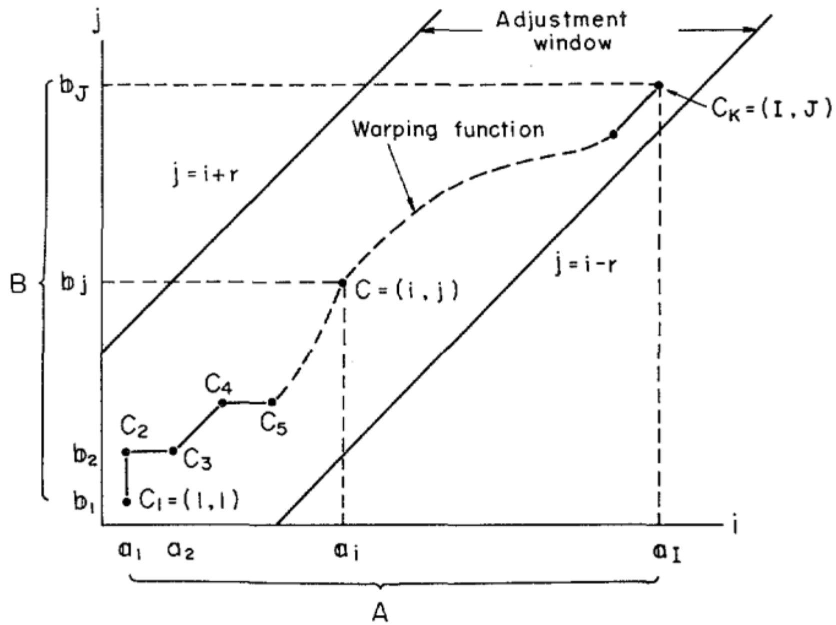
- DTW Based Methods
  - Extended from isolated word speech recognition.
  - The main difference is that the query is a word and the reference may be a longer sentence.
- Embedding Learning Based Method
  - Represent speech sequence of arbitrary length as a fixed-dimensional vector are used in KWS.

DTW Based Methods

- Compute similarity between two sequences of vectors.
- Two stages:
  - Convert the queries and target speech into same representations using acoustic models.
  - Compute confidence of appearance of the keywords to decide whether the keywords appear in speech stream.

Itakura F. Minimum prediction residual principle applied to speech recognition[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1975, 23(1): 154-158.
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.

DTW Based Methods



Formally

Given two sequences

$$X = x_1, \dots, x_N$$
$$Y = y_1, \dots, y_M.$$

Consider $c(k) = (i(k), j(k))$

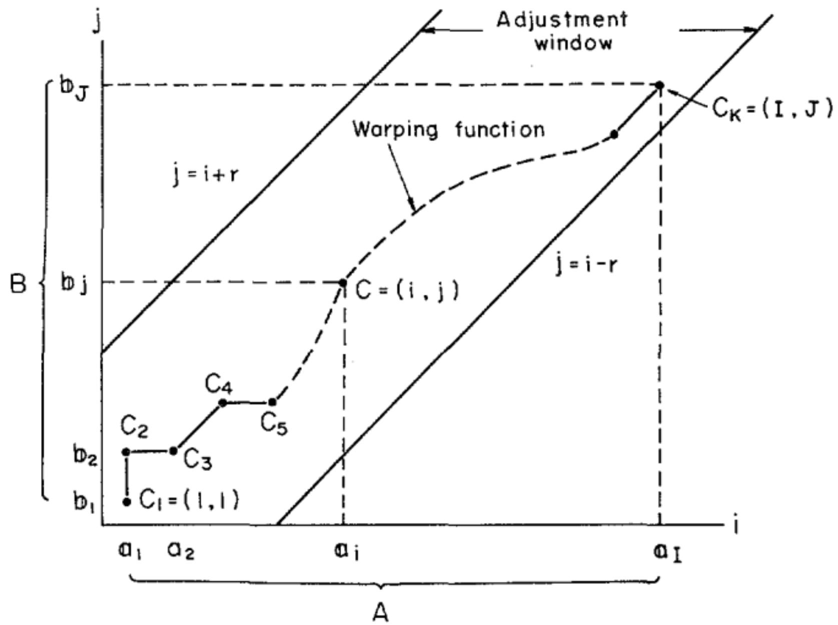
The matching pattern is a sequence of points

$$F = c(1), c(2), \dots, c(k), \dots, c(K),$$

The time-normalized distance is defined as

$$D(X,Y) = \underset{F}{Min}\left\{\frac{\sum_{k=1}^{K} d(c(k)) \cdot w(k)}{\sum_{k=1}^{K} w(k)}\right\}$$

DTW Based Methods



Five constraints:

a) Monotonicity

$$i(k-1) \le i(k) \text{ and } j(k-1) \le j(k)$$

b) Continuity

$$i(k)-i(k-1) \le 1 \text{ and } j(k)-j(k-1) \le 1$$

c) Boundary

$$i(1)=1, j(1)=1 \text{ and } i(K)=N, j(K)=M$$

d) Adjustment window

$$|i(k)-j(k)| \le R.$$

e) Slope constraint

DTW Based Methods

- Several variants of DTW for KWS
  - Segmental DTW
  - Segmented DTW
  - Non-segmental DTW
  - Subsequence DTW
  - Segmental local normalized DTW

Mantena G V, Achanta S, Prahallad 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.
Zhang Y, Glass J R. Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams[C]. ieee automatic speech recognition and understanding workshop, 2009: 398-403.

Segmental DTW

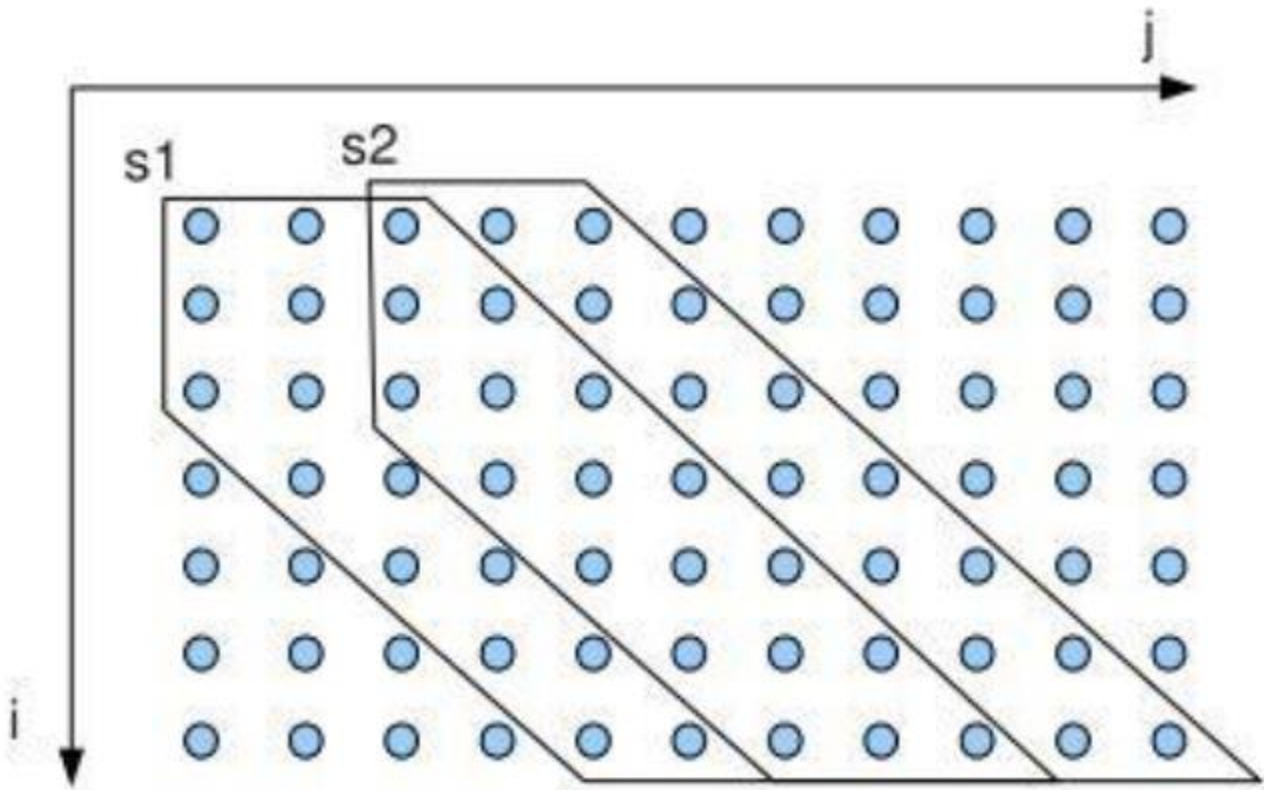
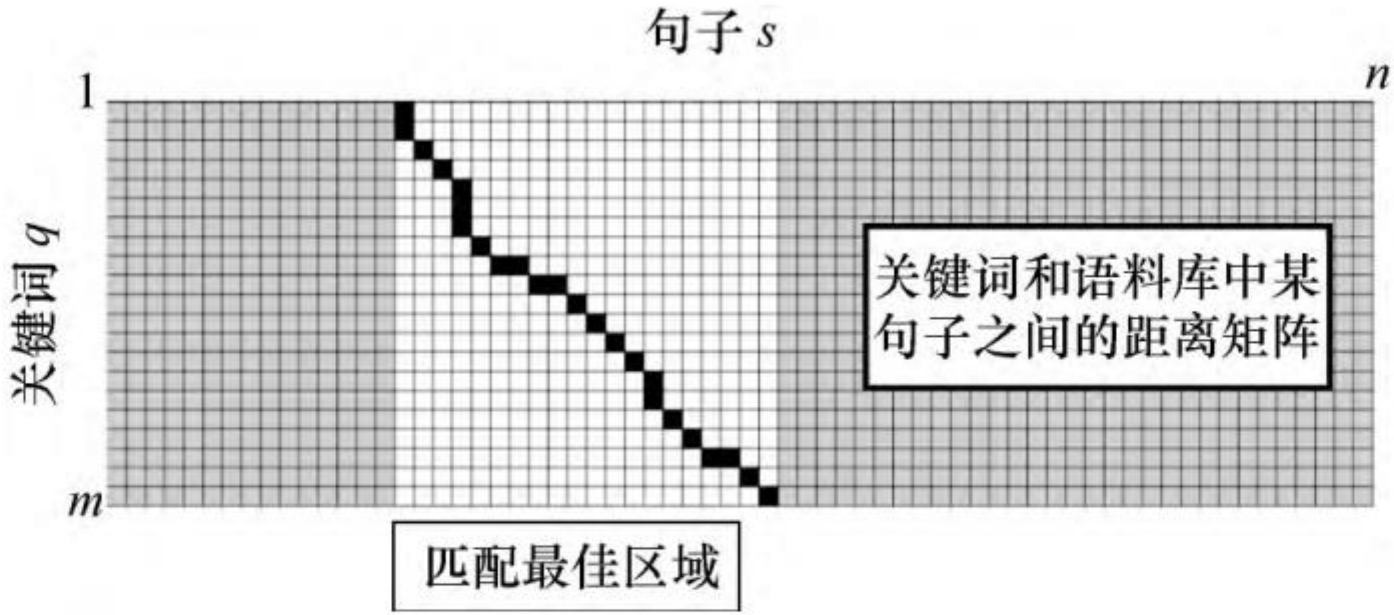


Fig. 1. S-DTW path

Segmental local normalized DTW

- Time complexity of SLN-DTW is O(mnd)



Feature representations and distance computation

- Main feature representations
  - Acoustic parameters (MFCC, FBANK)
  - Posteriorgram (GMM, DNN)
  - Bottleneck feature (DNN, Autoencoder)
- Distance computation
  - Compute similarity at each DTW step
  - Euclid distance
  - $-\log(\boldsymbol{x} \cdot \boldsymbol{y})$
  - $1 - \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}$

Some drawbacks of DTW

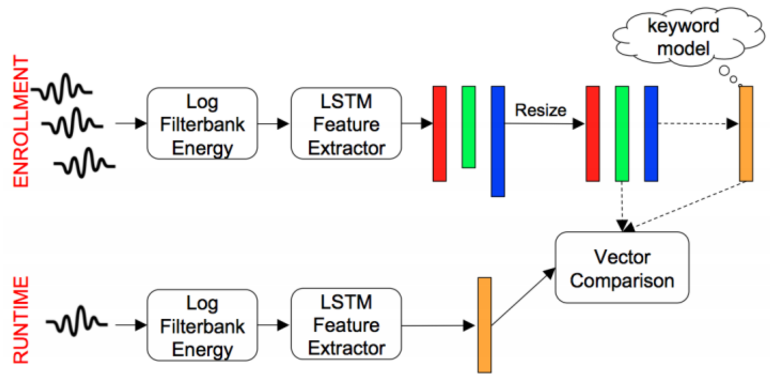
- Comparing two sequences using DTW based methods costs polynomial time.
- DTW is often oversensitive to longer phonetic segments.

Embedding Learning Based Method

- General ideas of non-DTW methods are based on to construct a fixed-dimensional vector to represent a speech segment of arbitrary length.

- In this case, common distances such as Euclid or cosine can be used to measure similarity between two sequences.

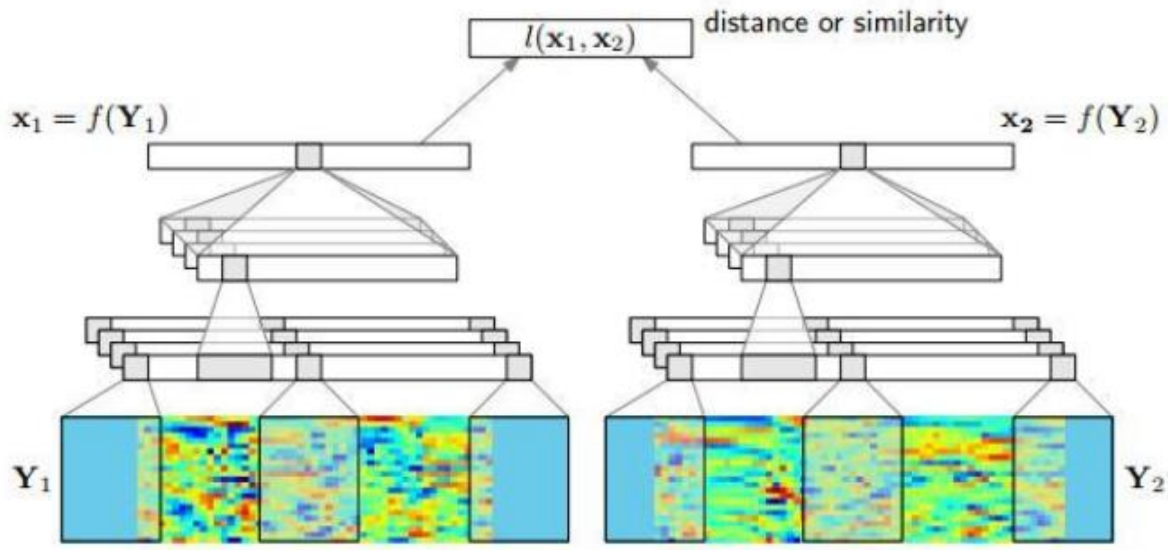
Embedding learning using LSTM



- Audio is preprocessed by a voice activity detection system.
- For speech regions, 40-dimensional mel-filterbank features are generated.
- 15k output targets represent whole word units.
- A fixed-length representation f is created by choosing the last k state vectors.

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.

Siamese networks based on CNN



- Weakly supervised: the transcripts of training data and testing data are unknown.

Kamper H, 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.

Outline

- Introduction
- Mainstream Approaches
  - HMM/Filler Models
  - Query-by-Example
  - LVCSR Based Methods
- Some Advances
- Take Home Messages

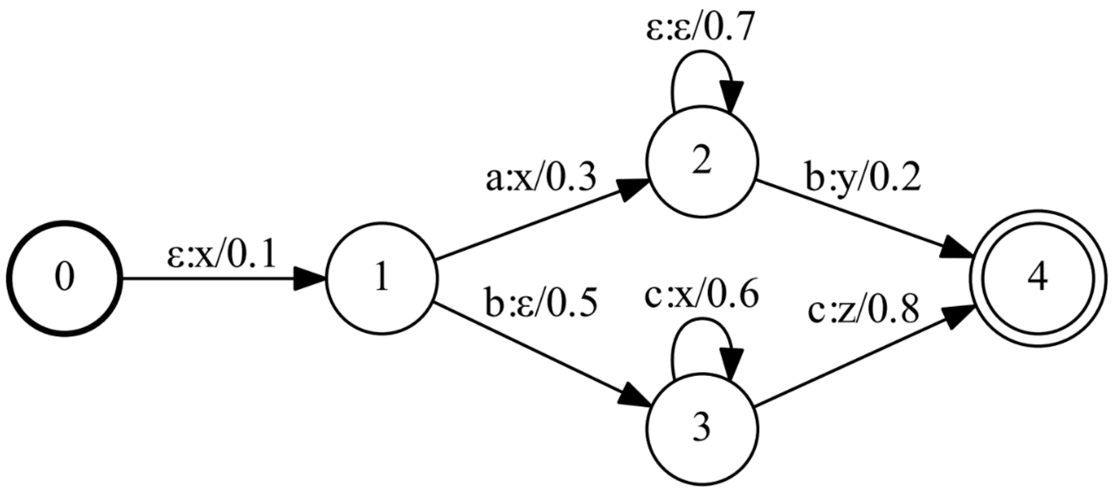
LVCSR based methods

LVCSR based methods

- The recognition results of LVCSR may contain errors, which will hurt the keyword spotting effect.

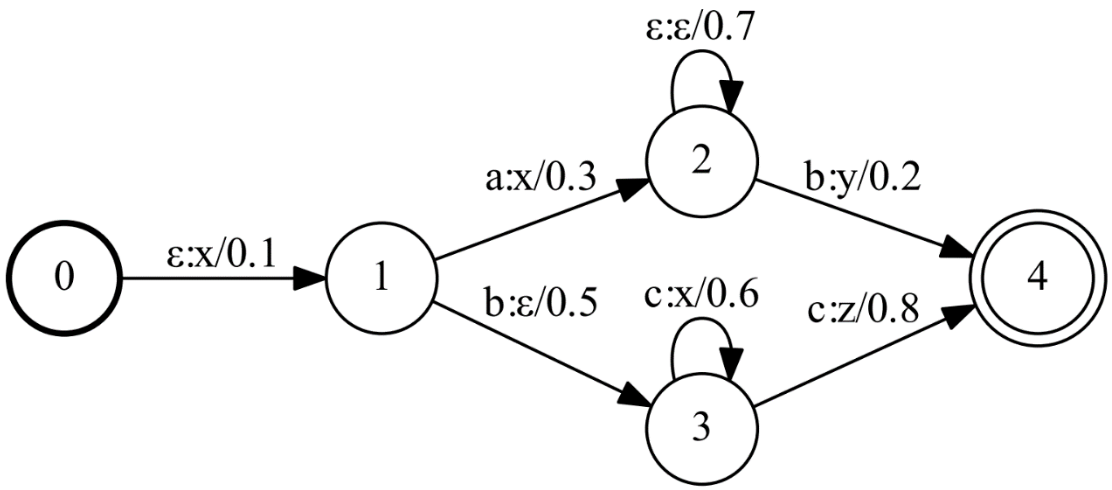
- How to index raw result of ASR?
  - Location of each word
  - Lattice

Prerequisite: WFST



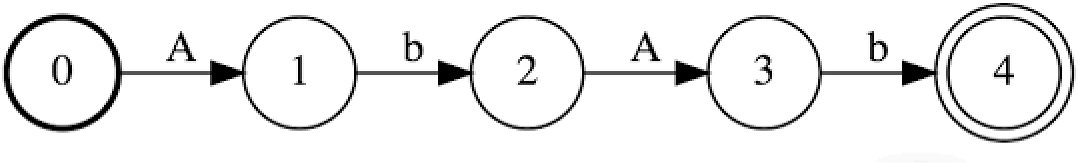
- Weighted Finite State Transducer (WFST) is a graph.

Prerequisite: WFST



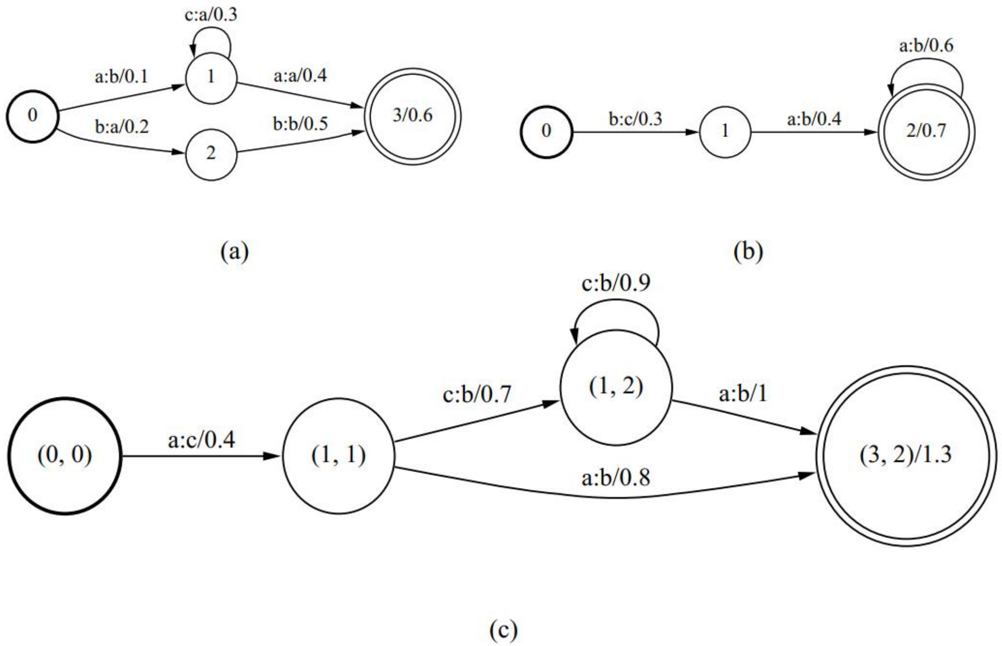
- It can be used to map a sequence to another, e.g., bcc to xxz.

Prerequisite: WFST



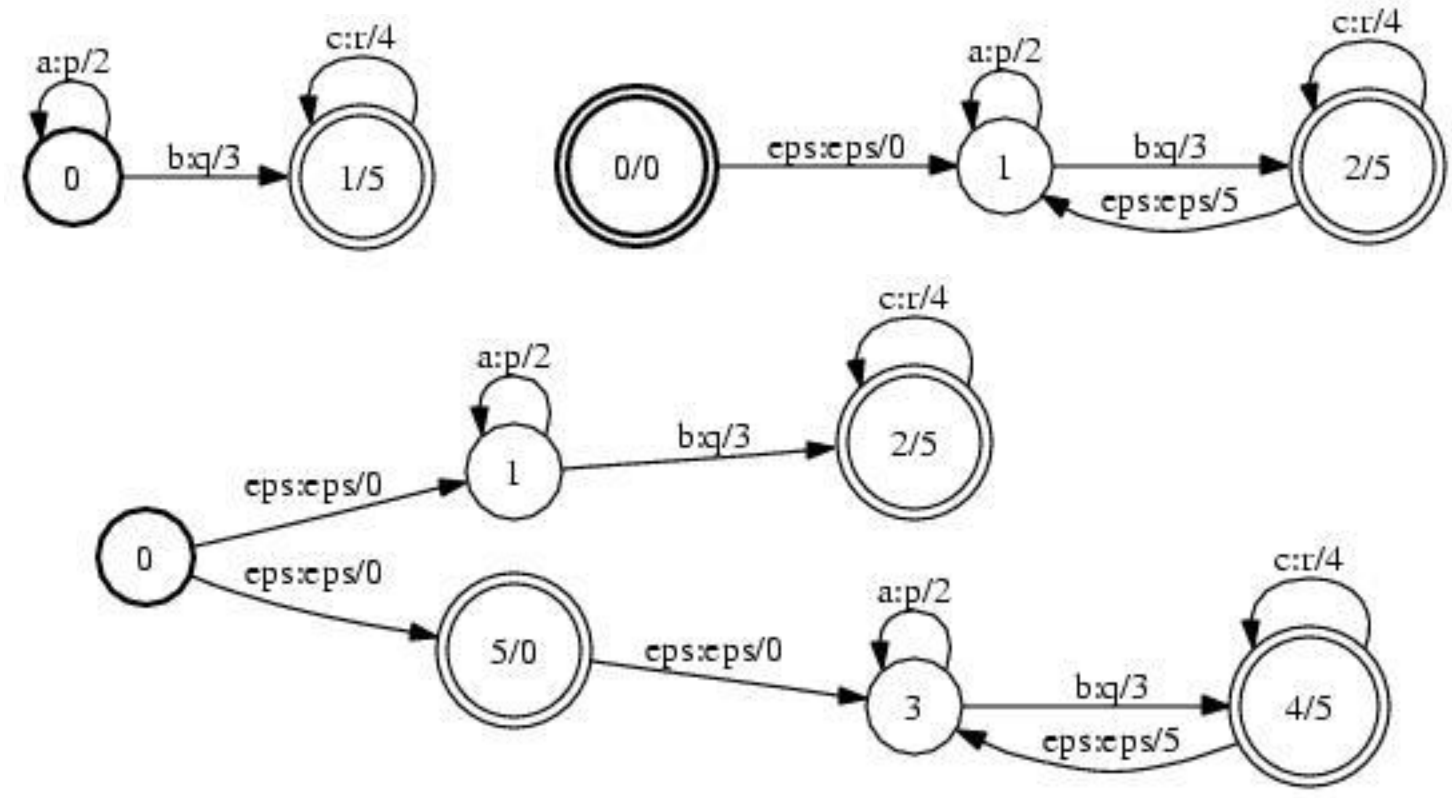
- A WFST can also be used to represent a string.

Prerequisite: WFST



- Composition is an operation of WFST.
  - T1: A to B
  - T2: B to C
  - T1 * T2: A to C

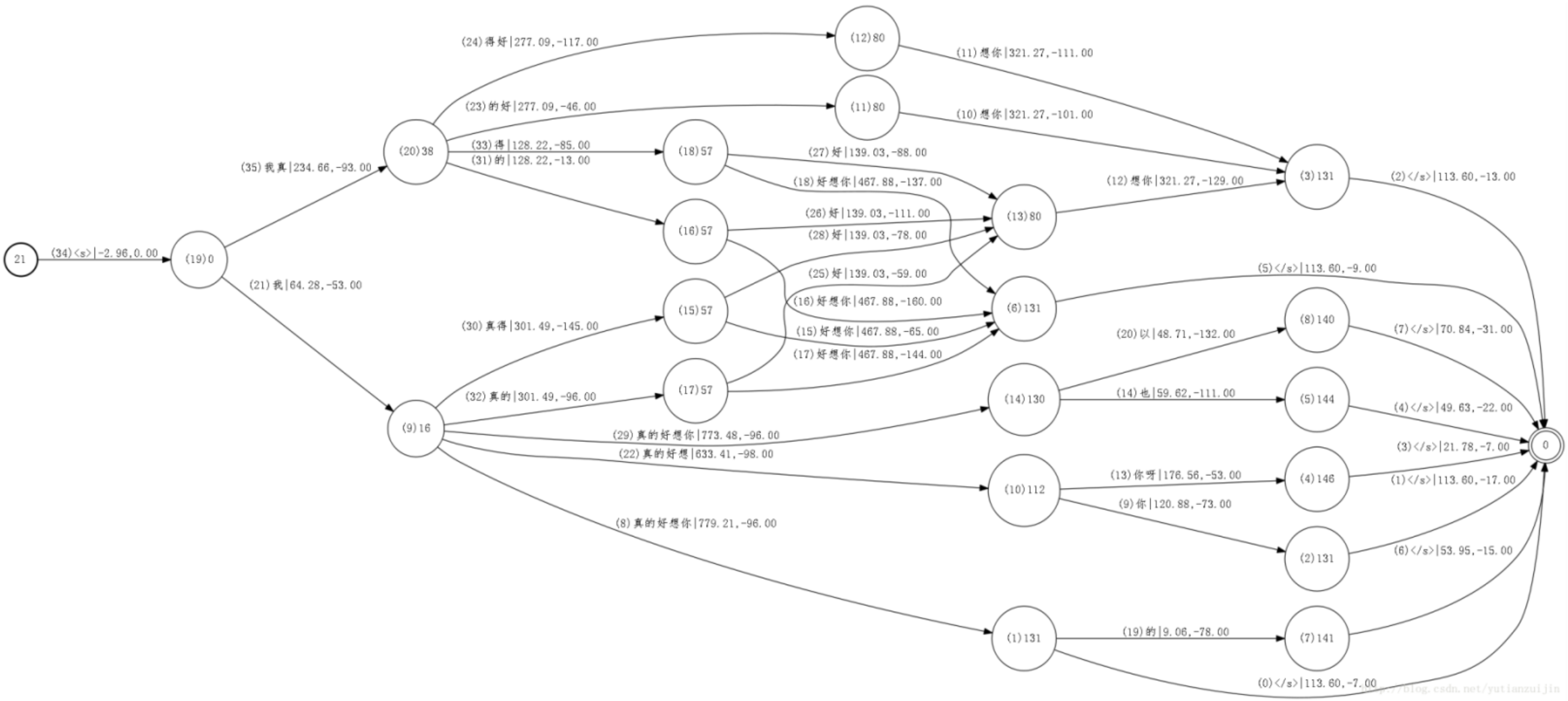
Prerequisite: WFST



- Union is also an operation of WFST.

Prerequisite: Lattice

- A lattice is a compact representation of ASR results.



Prerequisite: Factor Automata

- v is a factor of u if $u=xvy$, where u,x,v,y is strings.

- A Factor Automaton $F(u)$ of a string u is an automaton which can recognize factors of u.

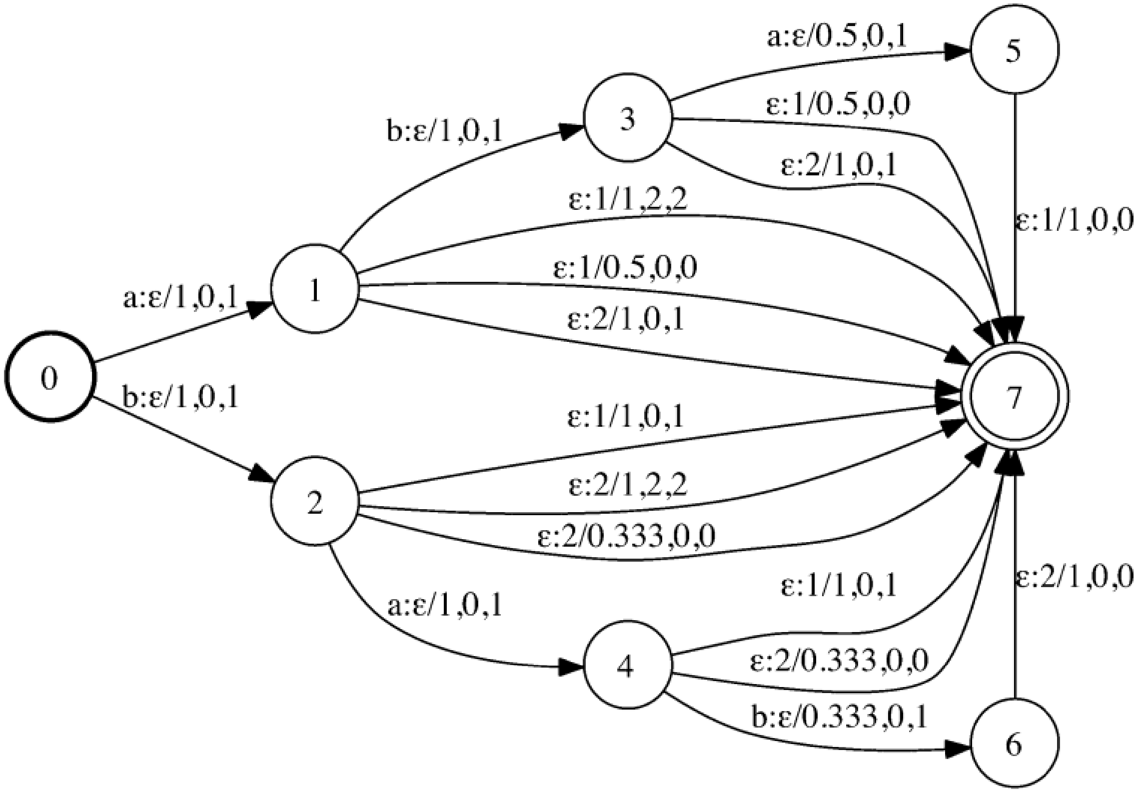
Timed Factor Transducer

- A TFT is a WFST mapping each factor x:
  - the set of automata in which x appears;
  - start-end times of the intervals where appears in each automaton;
  - the posterior probabilities of actually occurring in each automaton.

Can D, Saraclar M. Lattice Indexing for Spoken Term Detection[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(8): 2338-2347.

TFT for Lattice Indexing

- Indexing
  - Convert lattices to TFTs
  - Union
  - Optimize



TFT for Lattice Indexing

- Searching
  - Convert query to a linear acceptor X
  - Compose X and T: R
  - Each successful path in R is a single arc, the label is the automaton id, and a (LogP, start-time, end-time) triplet.

OOV problem

- The out-of-vocabulary problem is more important in KWS than in ASR.
- Users often would like to search names or new words which are out-of-vocabulary.
- A basic approach to tackle OOV problem is using sub-word units such as phones or syllables as results of the LVCSR system.

Proxy word:
a unified process method

- Proxy words are IV keywords which are acoustically similar as OOV keywords.

- In spotting stage, proxy words are searched in the index instead of original out-of-vocabulary query.

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.

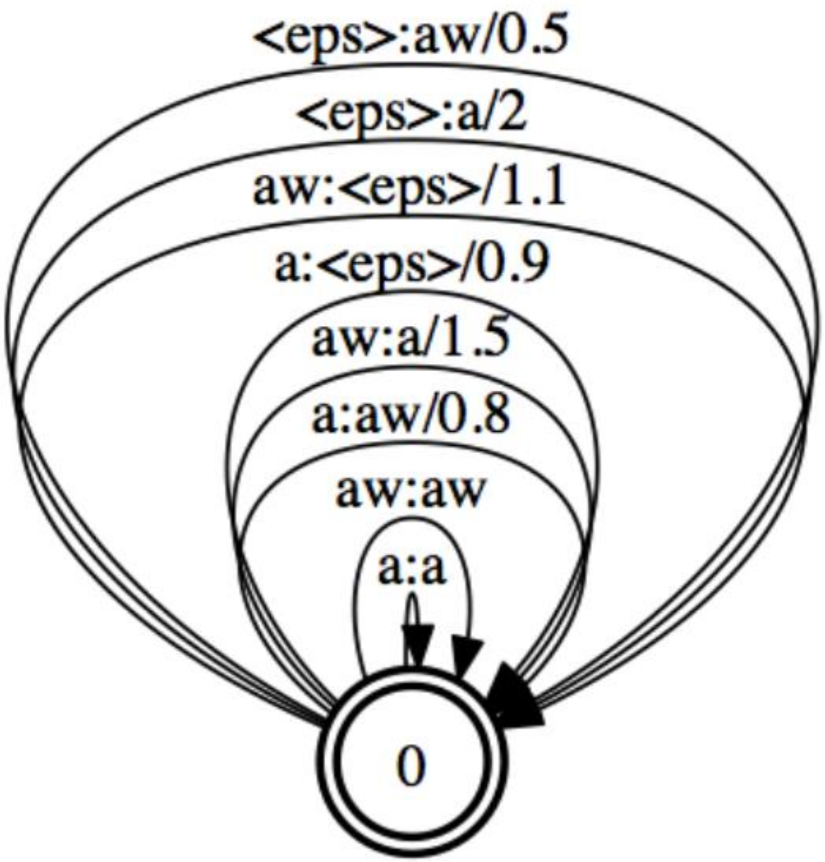
Proxy words generation

- Proxy words are generated based on WFST.

$$K' = \mathrm{Project}\left(\mathrm{ShortestPath}\left(K \circ L_2 \circ E \circ (L_1^*)^{-1}\right)\right).$$

- where K is a FSA for an OOV word;
- L_2 is a FST for pronunciation of the OOV word;
- E is an edit-distance transducer;
- L_1 denote the pronunciation lexicon of LVCSR.
- K' is a FSA corresponding to proxy words.

Phone confusion matrix estimation



- The phone confusion matrix is generated using maximum likelihood estimation.
- The pronunciations of the words are obtained using G2P software.

Outline

- Introduction
- Mainstream Approaches
  – HMM/Filler Models
  – Query-by-Example
  – LVCSR Based Methods
- Some Advances
- Take Home Messages

Model Compression

- Alexa

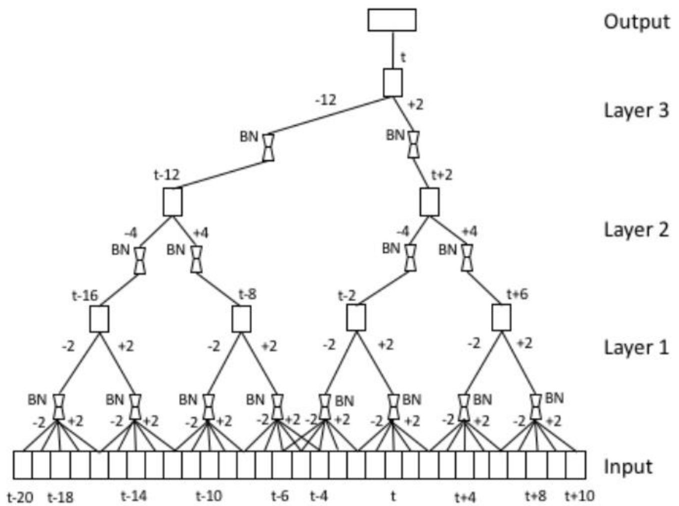
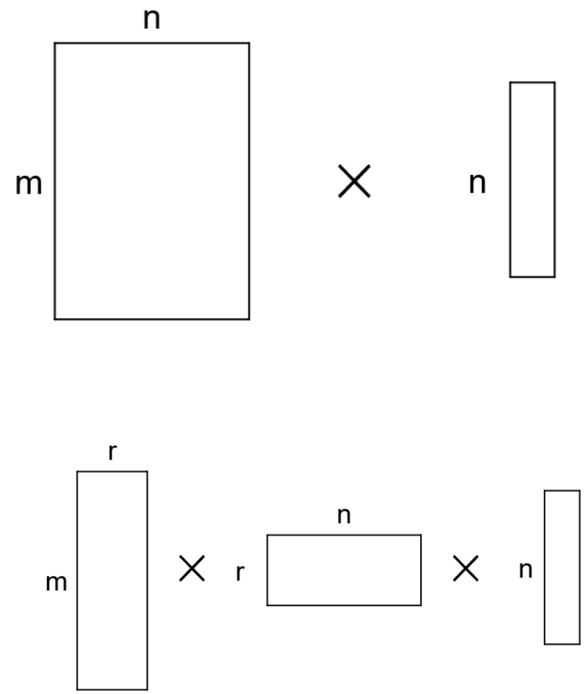


Figure 2: *TDNN architecture with SVD compression. ‘BN’ labels linear bottleneck layers.*



$m \times r + r \times n = (m+n)r$ parameters

$m \times r + r \times n = (m+n)r$ multiplications

If $r \ll \frac{mn}{m+n}$, it makes sense.

Sun M, Snyder D, Gao Y, et al. Compressed Time Delay Neural Network for Small-Footprint Keyword Spotting.[C]. conference of the international speech communication association, 2017: 3607-3611.

Model Compression

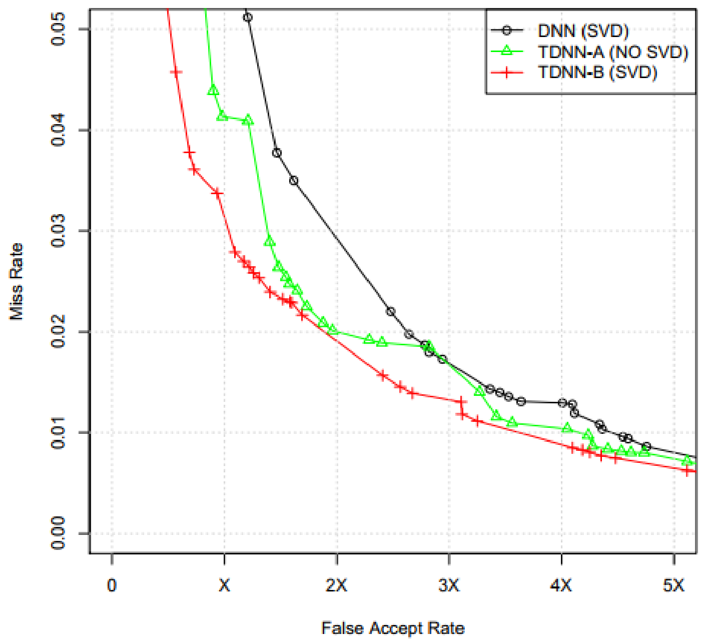


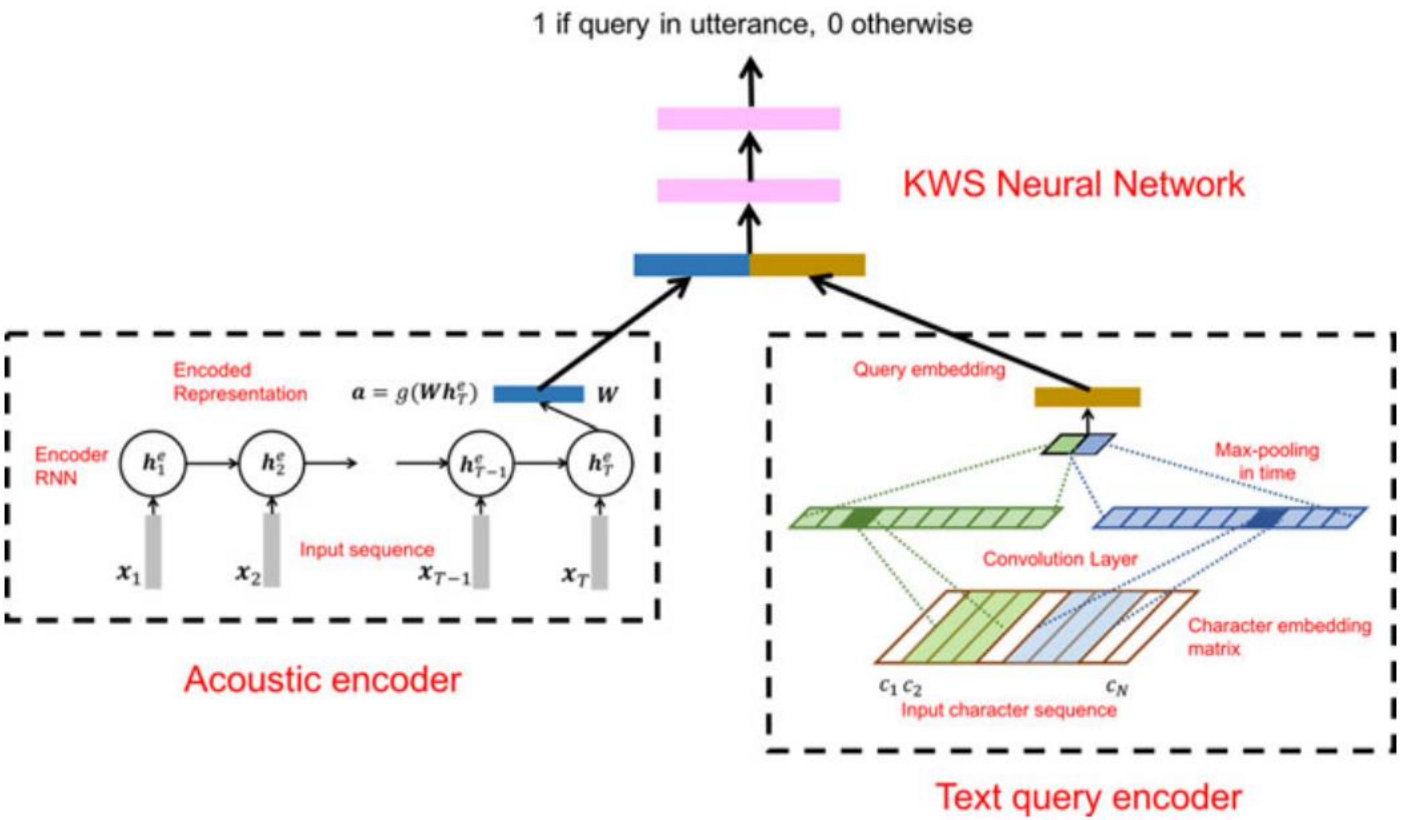
Figure 4: *DNN/TDNN-HMM DET curves for 'Alexa' detection.*

Table 2: *Relative change of HMM DET AUC for TDNN models without SVD compression (TDNN-A) and with SVD compression (TDNN-B), compared to the baseline SVD compressed DNN. All three models have comparable number of parameters ($\leq 100k$). Lower AUC indicates better performance*

Model	DNN	TDNN-A	TDNN-B
AUC Relative Change	0%	−19.7%	−37.6%

Sun M, Snyder D, Gao Y, et al. Compressed Time Delay Neural Network for Small-Footprint Keyword Spotting.[C]. conference of the international speech communication association, 2017: 3607-3611.

Compute similarities between heterogeneous patterns



Audhkhasi K, Rosenberg A, Sethy A, et al. End-to-end ASR-free keyword search from speech[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1351-1359.

Similarity Image Classification For Query-by-Example KWS

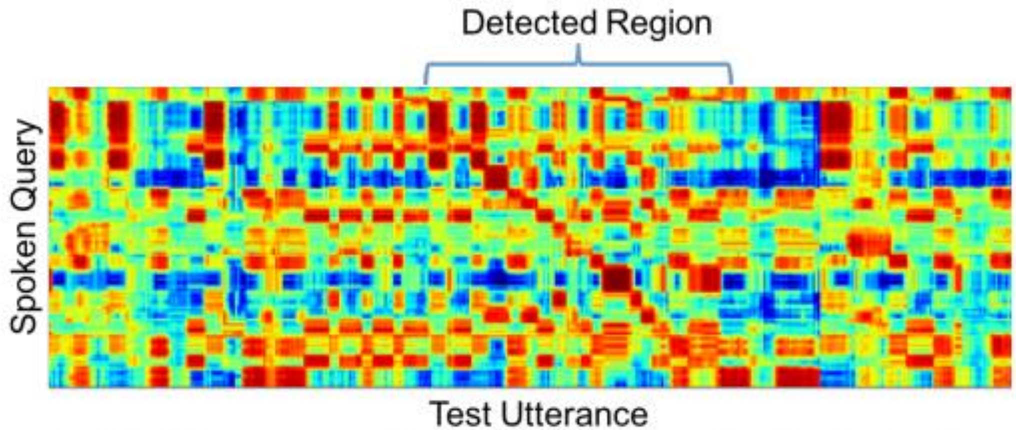


Figure 1: *Positive case: the query occurs in the test utterance*

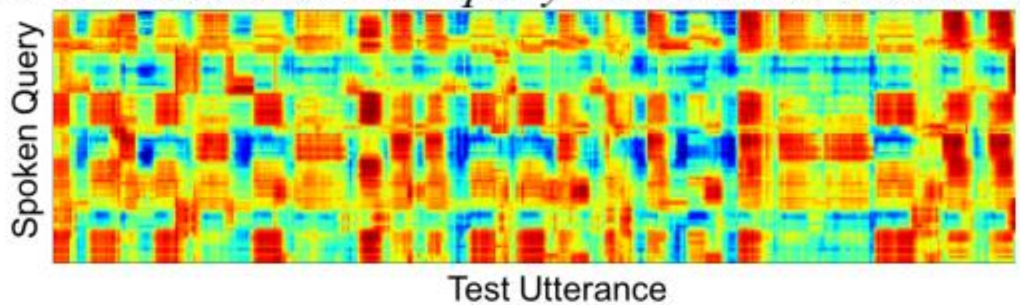


Figure 2: *Negative case: the query does not occur in the test utterance*

Ram D, Miculicich L, Bourlard H. CNN based query by example spoken term detection[C]//Proceedings of the Nineteenth Annual Conference of the International Speech Communication Association (INTERSPEECH). 2018.

Streaming Seq2Seq Models for KWS

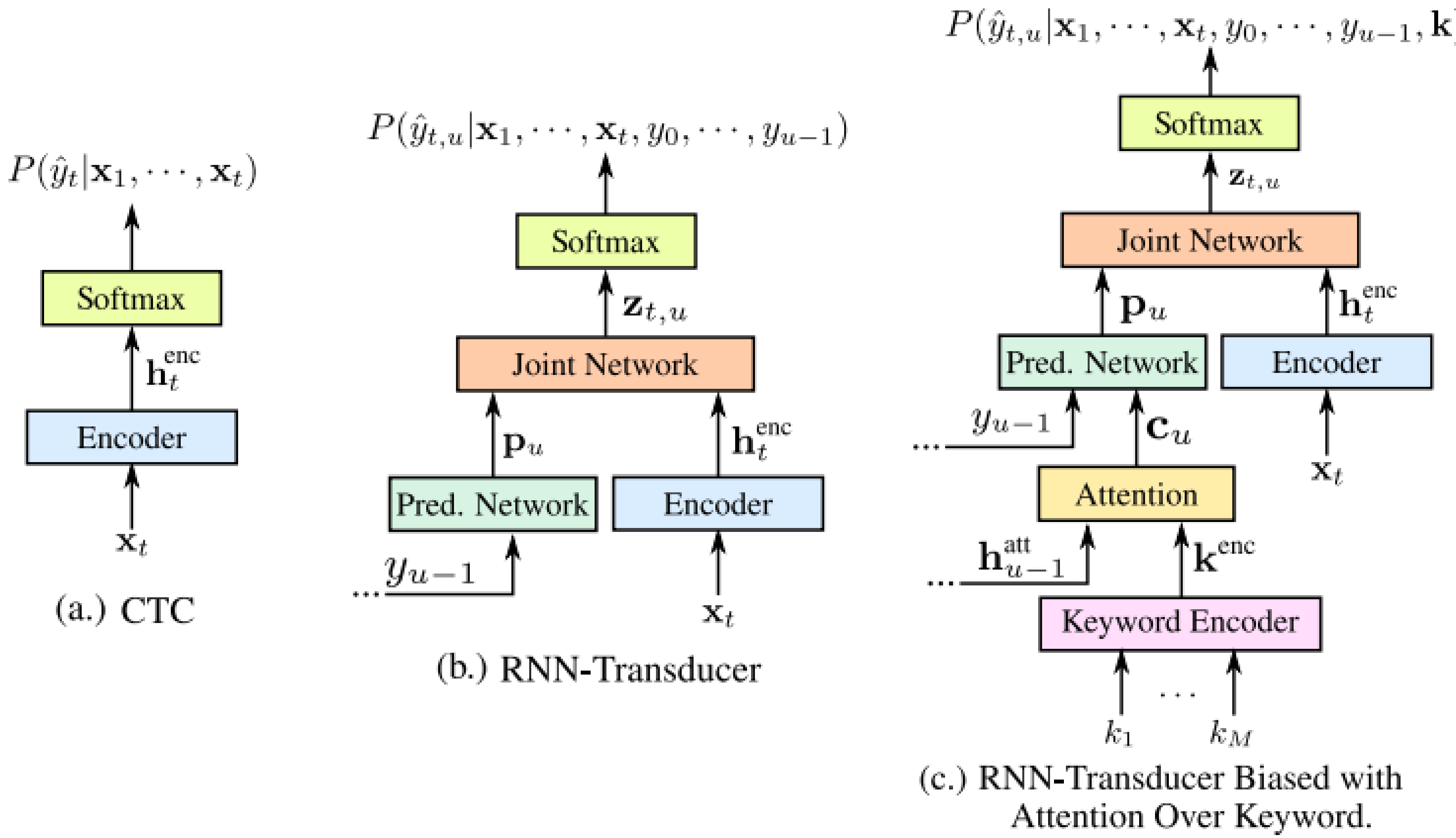
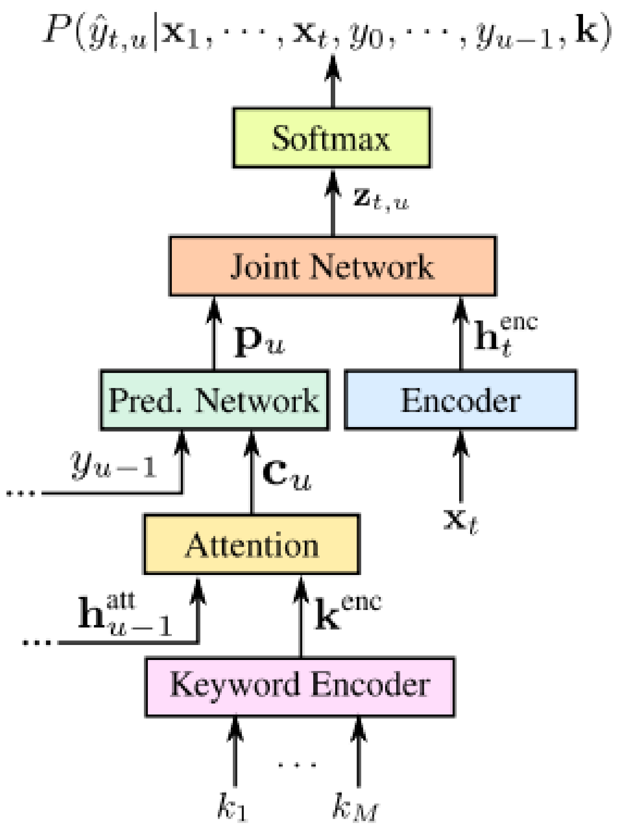


Fig. 1: A schematic representation of the models used in this work.

He Y, Prabhavalkar R, Rao K, et al. Streaming small-footprint keyword spotting using sequence-to-sequence models[C]//Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE

Streaming Seq2Seq Models for KWS



$$\mathbf{k}^{\text{enc}} = [k_1^{\text{enc}}, \cdots, k_M^{\text{enc}}, k_{M+1}^{\text{enc}}]$$

is one-hot encodings of M+1 phonemes of a keyword.

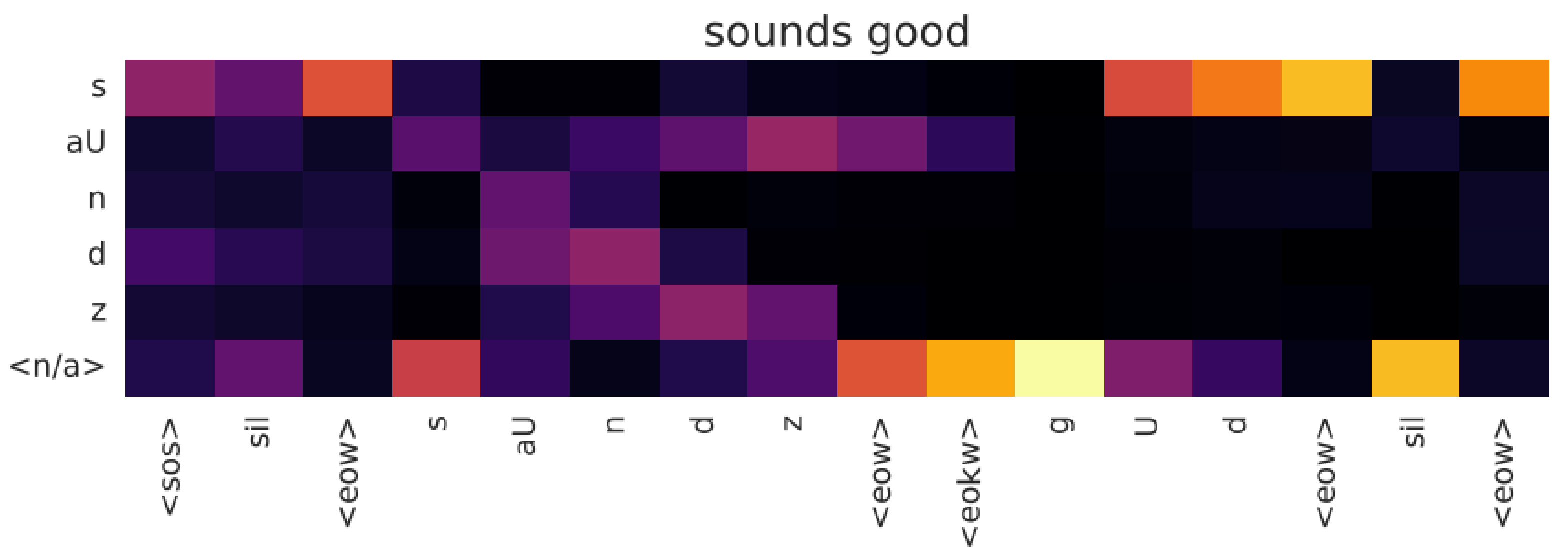
$$\beta_{j,u} = \langle \phi(k_j^{\text{enc}}), \psi(\mathbf{h}_{u-1}^{\text{att}}) \rangle$$

$$\alpha_{j,u} = \frac{e^{\beta_{j,u}}}{\sum_{j'=1}^{M+1} e^{\beta_{j',u}}}$$

$$\mathbf{c}_u = \sum_{j=1}^{M+1} \alpha_{j,u} k_j^{\text{enc}}$$

He Y, Prabhavalkar R, Rao K, et al. Streaming small-footprint keyword spotting using sequence-to-sequence models[C]//Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE

Streaming Seq2Seq Models for KWS



(a) Attention matrix of a positive utterance for the keyword "sounds", with the transcript "sounds good".

He Y, Prabhavalkar R, Rao K, et al. Streaming small-footprint keyword spotting using sequence-to-sequence models[C]//Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE

Streaming Seq2Seq Models for KWS

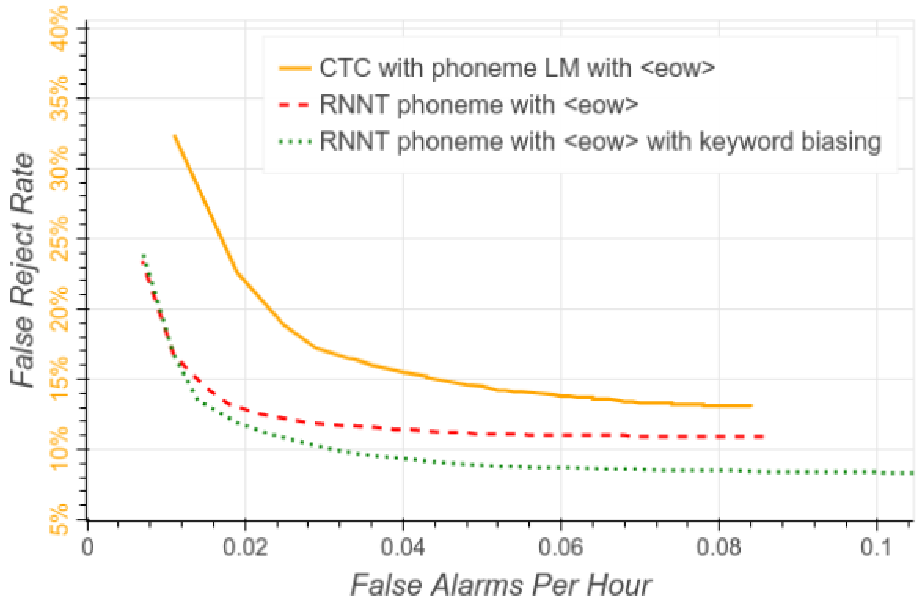


Fig. 6: A comparison of the RNN-T phoneme model with keyword biasing against the best CTC baseline and the RNN-T phoneme system without biasing on the test set. All systems use the $\langle \texttt{eow} \rangle$ token.

He Y, Prabhavalkar R, Rao K, et al. Streaming small-footprint keyword spotting using sequence-to-sequence models[C]//Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE

Outline

- Introduction
- Mainstream Approaches
  - HMM/Filler Models
  - Query-by-Example
  - LVCSR Based Methods
- Some Advances
- Take Home Messages

Take Home Messages

- Keyword spotting focuses on detecting keywords in computation constrained conditions.

- The out-of-vocabulary keywords are problems of spoken term detection.

Thank you!