End-to-End

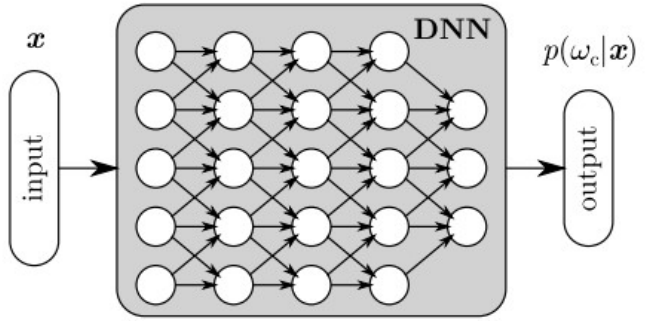
Keywords Spotting

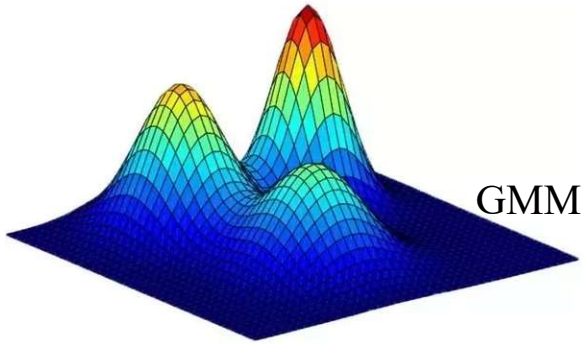
Ying Shi

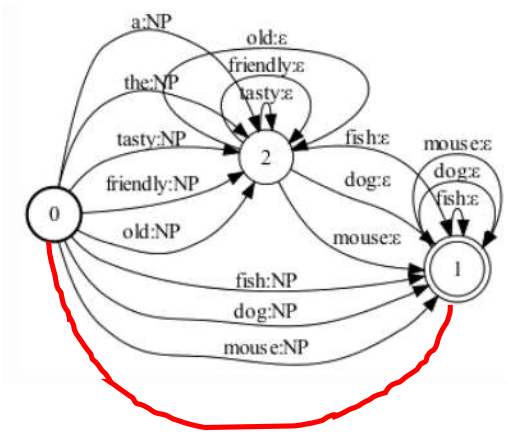
2021.11.29

Outline:

- Some Keywords Spotting methods
- Cross Modality view
- Cross Modality Attention E2E KWS







Garbage loop

Keywords

Hybrid System(AM+LM)

Overlap speech data?
Garbage loop weight?
Decoding window sensitive

Keywords Spotting

>>> e2e ASR System(CTC)

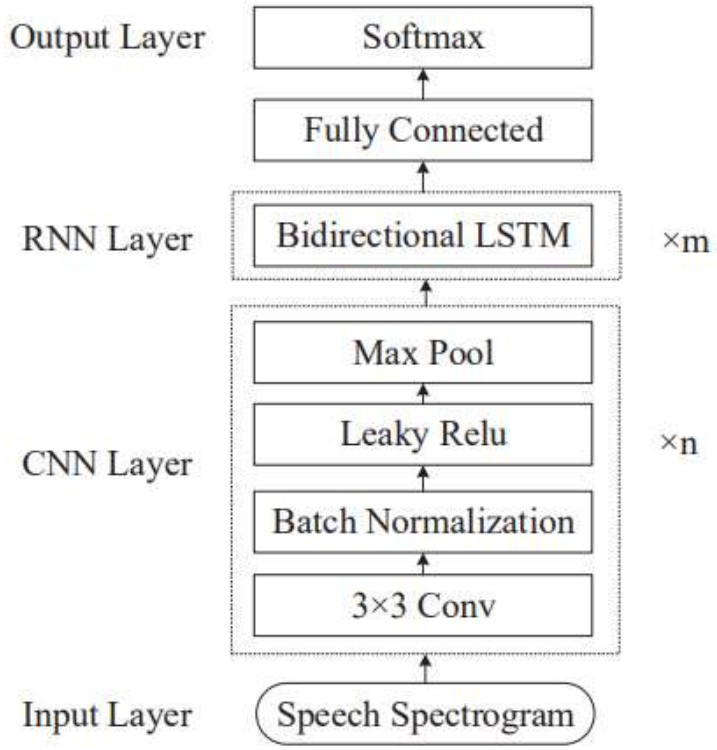


Fig. 1. The architecture of CRNN

In our work, all Chinese characters are first converted into tonal syllables. Then a mapping dictionary is created for all keywords character syllables. All syllables that are not in the dictionary are treated as the same label, which is defined as the filler symbol. Besides, to better represent pauses between words, a syllable-boundary is inserted between syllables. Finally the network output labels include all keyword character tonal syllables, a filler symbol, a syllable-boundary, and a CTC blank.

Table 2. Comparison of baseline and CRNN-CTC method

13 keywords	**FRR**	**FAR**
RNN-CTC	9.43%	0.47 times per hour
CRNN-CTC	5.35%	0.26 times per hour
20 keywords	**FRR**	**FAR**
RNN-CTC	9.99%	0.24 times per hour
CRNN-CTC	6.37%	0.17 times per hour

Table 3. Performances of CRNN-CTC based KWS using different modeling units

13 keywords	**FRR**	**FAR**
tonal syllables	5.35%	0.26 times per hour
characters	7.45%	0.27 times per hour
keywords	5.22%	0.84 times per hour
20 keywords	**FRR**	**FAR**
tonal syllables	6.37%	0.17 times per hour
characters	8.16%	0.20 times per hour
keywords	6.35%	0.55 times per hour

Overlap Speech data?
KWS Training Data?

CRNN-CTC Based Mandarin Keywords Spotting ICASSP 2020
Adaptation of RNN Transducer With Text-to-Speech Technology for Keywords Spotting

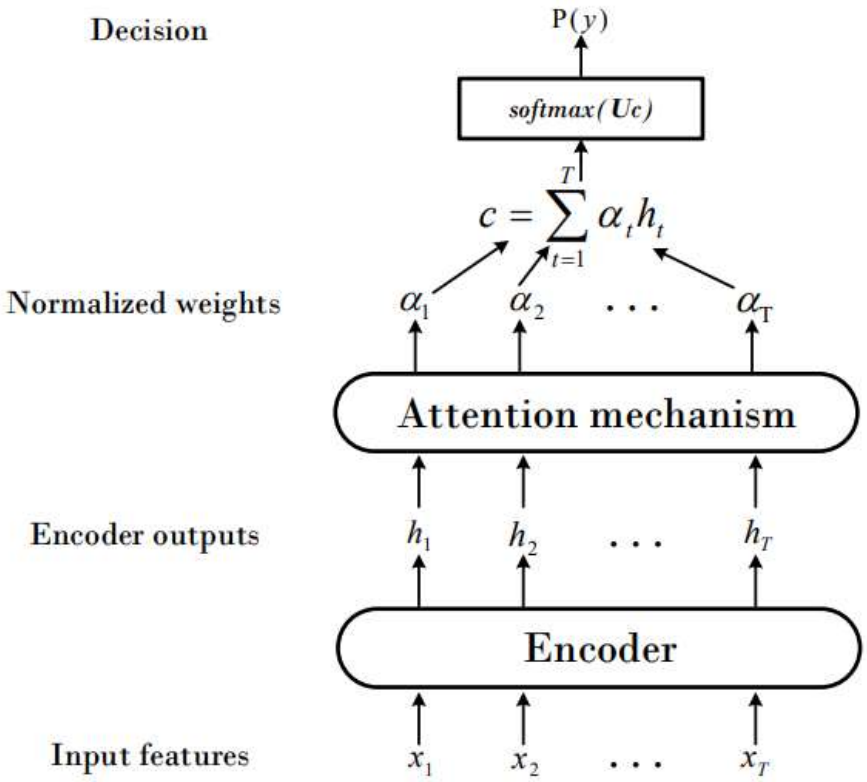


Figure 1: *Attention-based end-to-end model for KWS.*

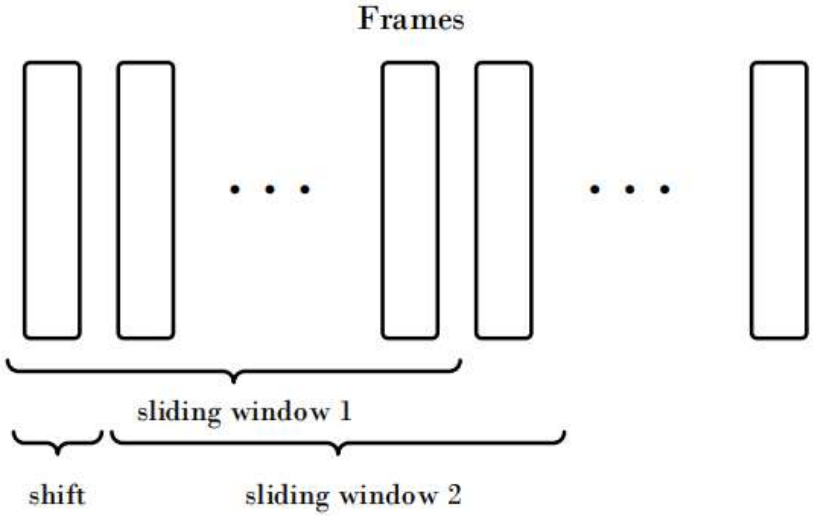


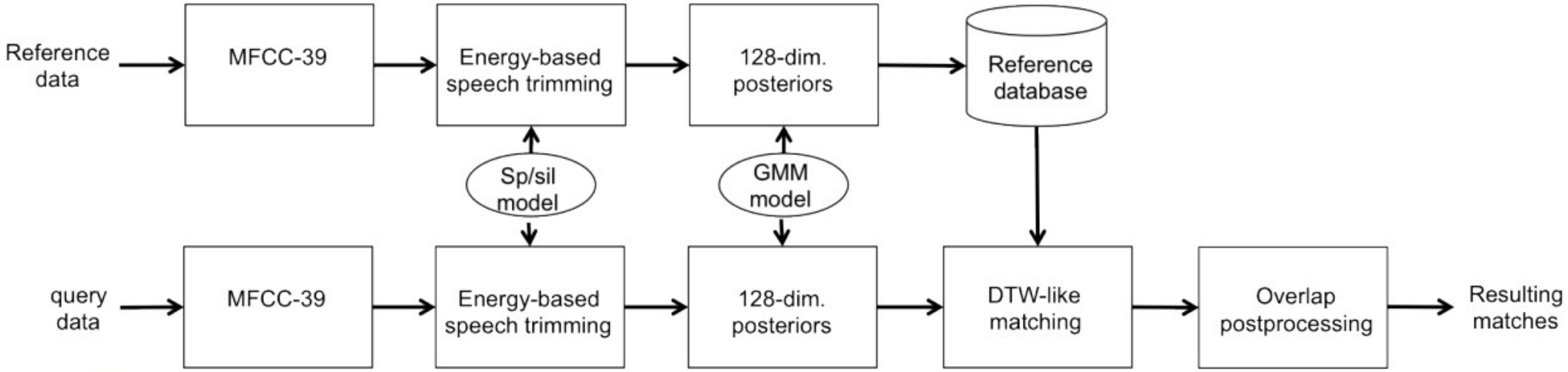
Figure 2: *Sliding windows used in decoding.*

Table 3: *Performance of adding convolutional layers in the GRU (CRNN) attention-based model with soft attention. FRR is at 1.0 false alarm (FA) per hour.*

Channel	Layer	Node	FRR (%)	Params (K)
8	1	64	2.48	52.5
8	2	64	1.34	77.3
16	1	64	**1.02**	84.1
16	2	64	1.29	109

KWS Training Data?

Attention-based End-to-End Models for Small-Footprint Keyword Spotting Xiaomi



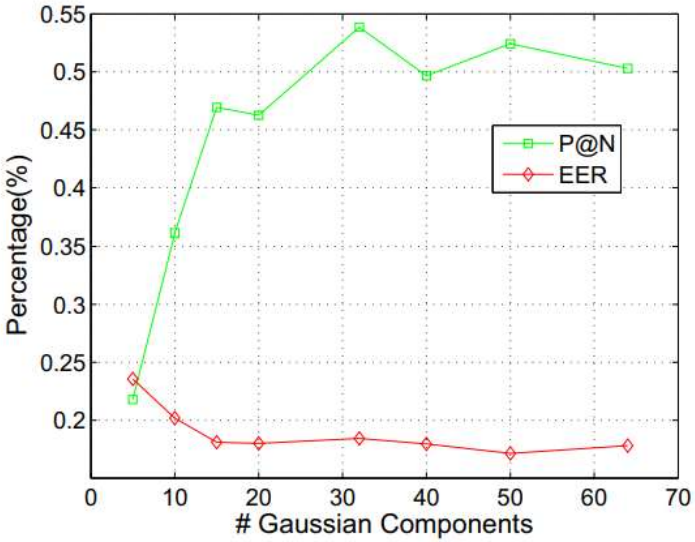


Fig. 5. Effect of different numbers of Gaussian components

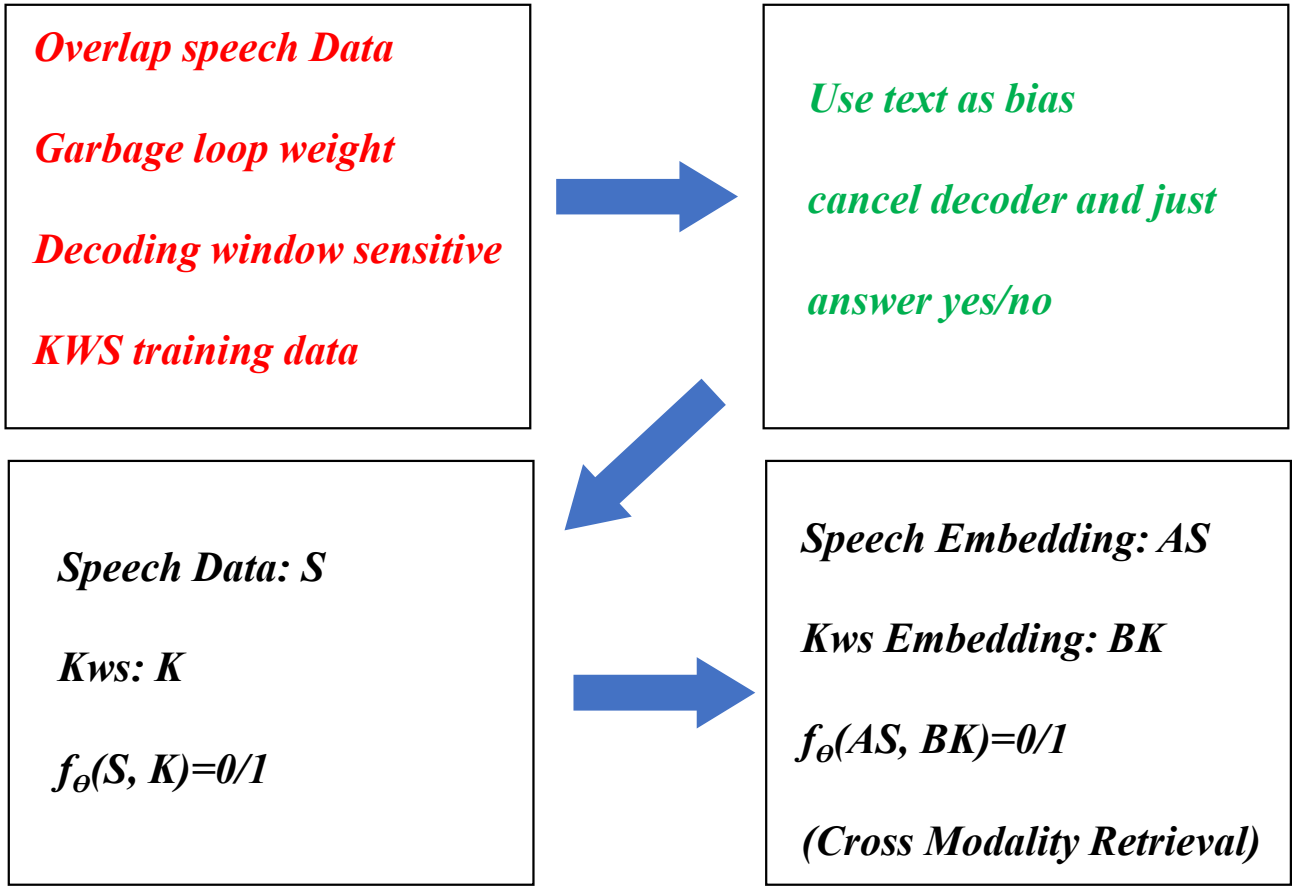
TABLE III
EFFECT OF DIFFERENT NUMBERS OF KEYWORD SAMPLES

# Examples	P@10	P@N	EER
1	27.0%	17.3%	27.0%
5	61.3%	33.0%	16.8%
10	68.3%	39.3%	15.8%

Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posteriorgrams MIT
Memory Efficient Sub-Sequence DTW for Query-By-Example Spoken Term Detection

Cross Modality View

>>>



CCA & KCCA: Canonical correlation analysis; An overview with application to learning methods

CFA & LSI: Multimedia Content Processing through Cross-Modal Association

JLR: Learning Cross-Media Joint Representation With Sparse and Semi-supervised Regularization

LGCFL: Learning Consistent Feature Representation for Cross-Modal Multimedia Retrieval

DCCA: Deep Correlation for Matching Images and Text

Corr-AE: Cross-modal Retrieval with Correspondence Autoencoder

CM-GANs: Cross-modal Generative Adversarial Networks for Common Representation Learning

MMCA: Multi-Modality Cross Attention Network for Image and Sentence Matching

…

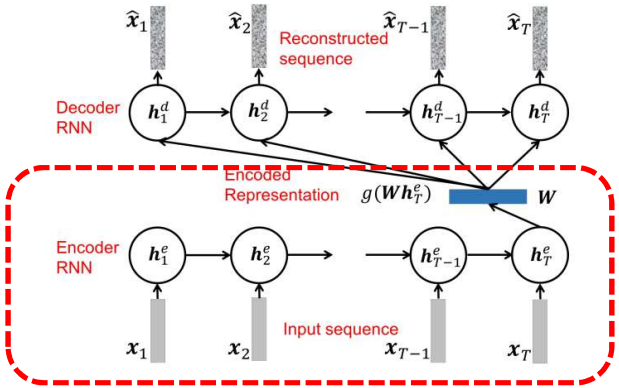
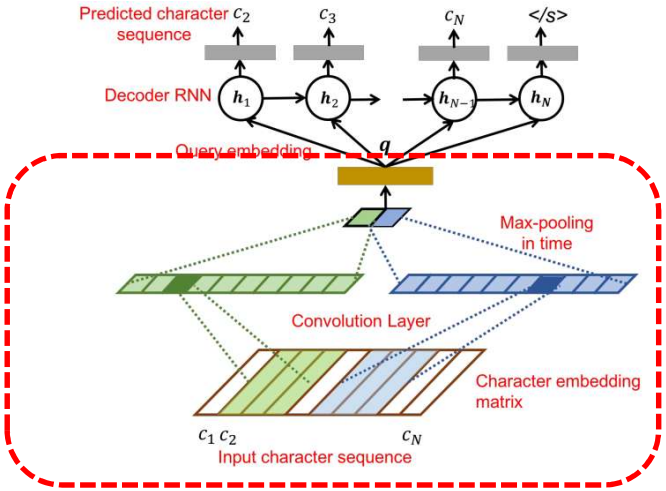


Fig. 1. This figure shows an RNN acoustic auto-encoder for a T-length input sequence of acoustic feature vectors through a D-dimensional encoded representation $g(\mathbf{W}\mathbf{h}_T^e)$, where g denotes a ReLU activation function.



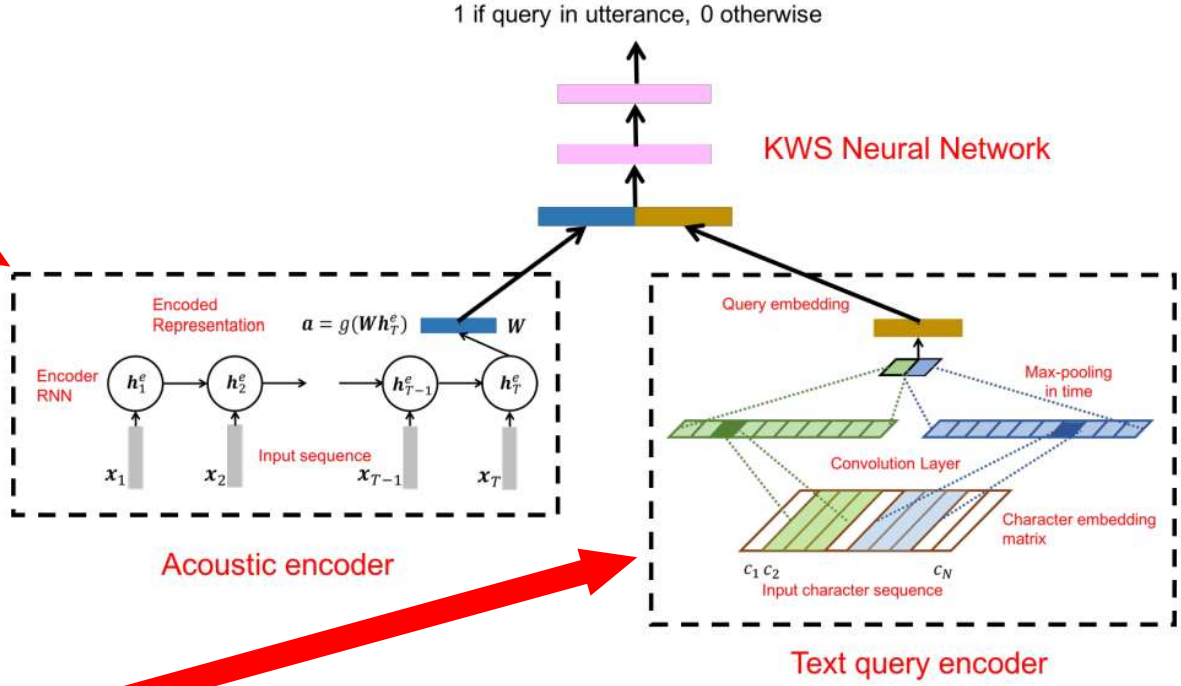


Fig. 2. This figure shows an character CNN-RNN LM for encoding text queries. We show two convolutional masks for simplicity.

Table 2. This table compares the KWS accuracy of the E2W KWS and DNN-HMM hybrid ASR systems for IV and OOV queries.

Query Type →	**IV**	**OOV**
DNN-HMM (2gm word LM)	76.7	50.0 (chance)
DNN-HMM (4gm grapheme LM)	70.7	55.5
E2E ASR-free	55.6	57.7

Table 2 shows the classification accuracies of the DNN-HMM ASR system and the proposed E2E ASR-free KWS system. We obtain a classification accuracy of 55.6% on IV and 57.7% on OOV queries, which is significantly above chance. As expected, the IV performance is lower than that of the hybrid ASR system using 2-gm word LM. But it is interesting to note that the E2E ASR-free and hybrid system using 4-gm grapheme LM have closer accuracies, especially for OOV queries, where the E2E KWS system performs better by 2.2% absolute. This result is encouraging, since the hybrid system uses word-level transcriptions for training the acoustic model and 36 times more training time than the E2E ASR-free KWS system. We performed further analysis of the dependence of KWS performance on query length. Table 1 shows the classification accuracy as a function of number of graphemes in the query. We observe that both the ASR-based and E2E KWS systems have difficulty detecting short queries. In case of the E2E system, this is because it is difficult to derive a reliable representation for short queries due to the lack of context. A key advantage of the E2E KWS system is that it takes 36 times less time to train than the DNN-HMM system.

Table 1. This table compares the KWS accuracy of the E2W KWS and DNN-HMM hybrid ASR systems for different IV query lengths.

Query Length →	$\leq$3	4	5	6	7	8	9	10	11	12	13	14	$\geq$15
DNN-HMM (2gm word LM)	69.8	72.5	74.6	77.9	77.3	78.8	76.7	80.0	78.7	78.9	74.5	77.1	78.6
DNN-HMM (4gm grapheme LM)	70.6	74.7	71.1	72.8	71.9	70.1	66.4	68.4	67.3	65.2	65.6	65.4	65.3
E2E ASR-free	51.8	56.4	56.5	55.6	55.3	55.1	55.7	52.1	53.5	58.4	55.8	56.7	60.0

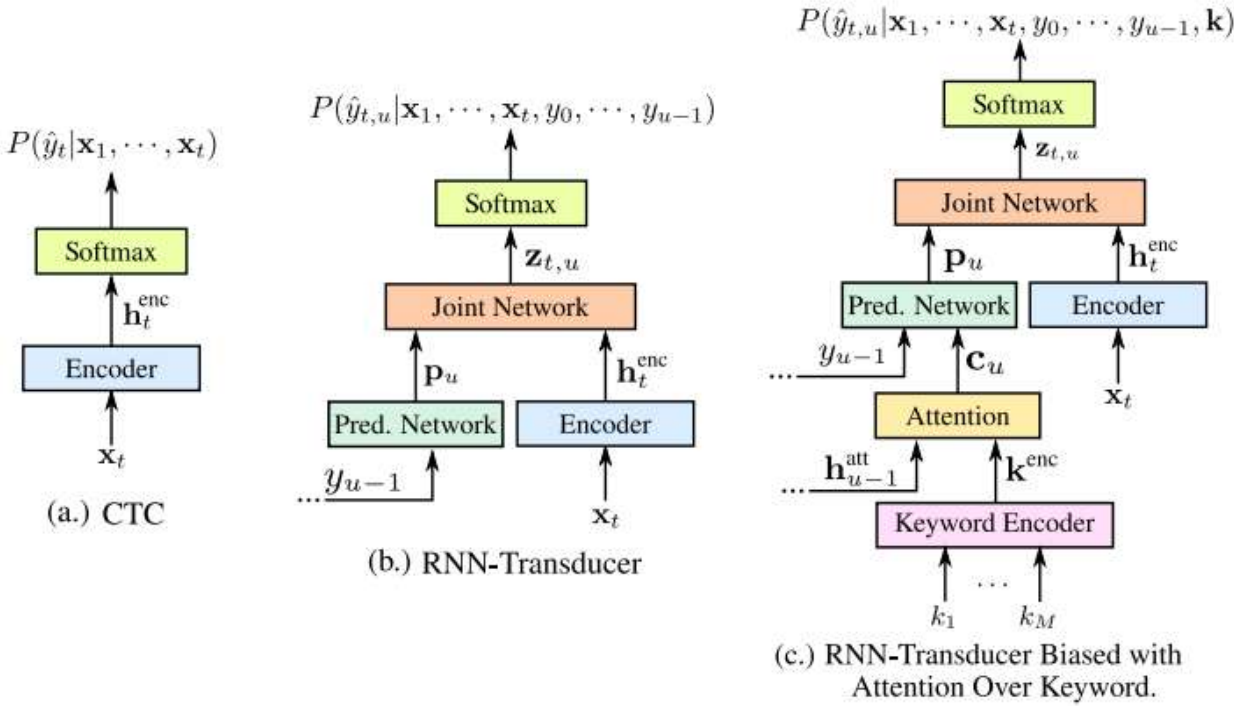
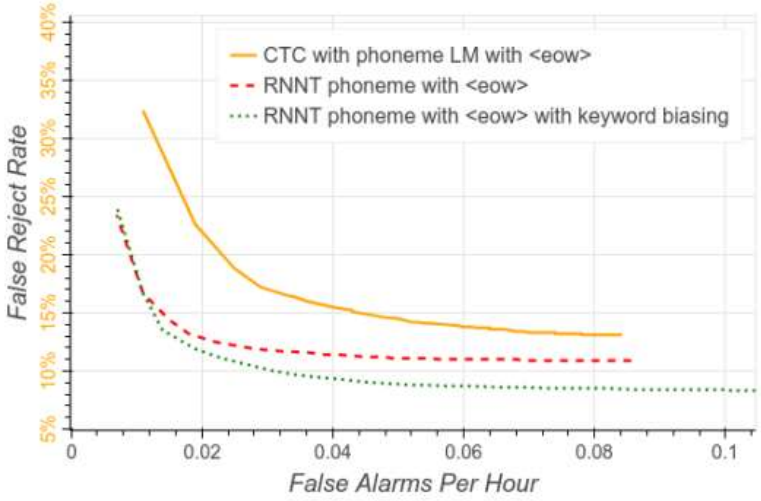


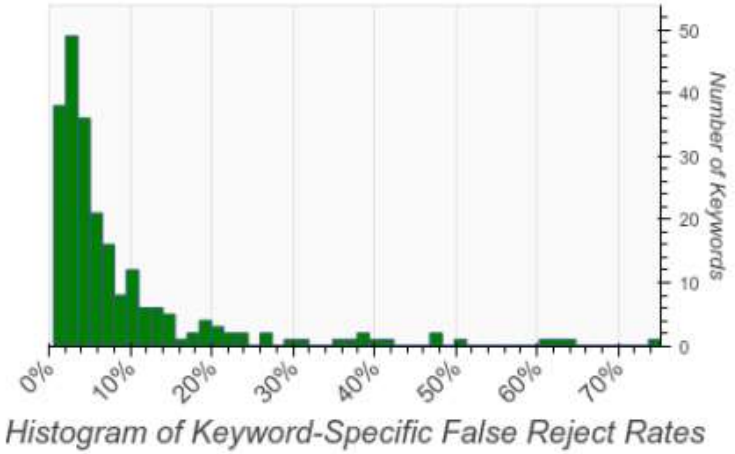
Fig. 1: A schematic representation of the models used in this work.

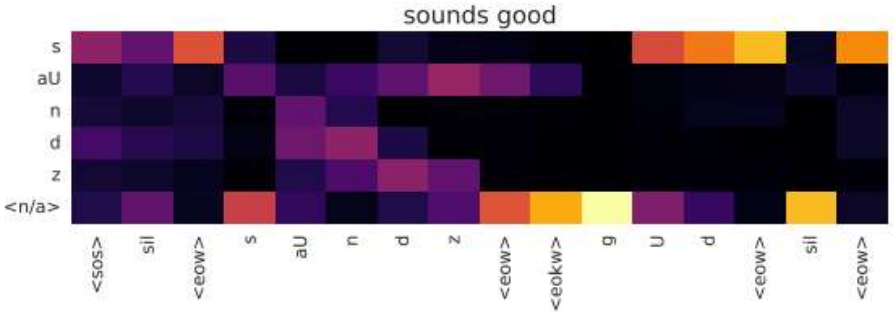
Unlike the RNN-T model, which can be trained given pairs of input and output sequences $(\mathbf{x}, \mathbf{y})$, in order to train the RNN-T model with keyword biasing, we need to also associate a keyword phrase, $\mathbf{k}$, with the training instance. We create examples where the keyword, $\mathbf{k}$, is present in $\mathbf{x}$, as well as examples where the keyword is absent in $\mathbf{x}$ as follows: with probability p^{kw} we uniformly sample one of the words in $\mathbf{x}$ as the keyword, $\mathbf{k}$, and with probability $1 - p^{\text{kw}}$ we uniformly sample a word which is not in $\mathbf{x}$ as the keyword, $\mathbf{k}$. If we select one of the words in $\mathbf{x}$ as the target, we modify the target labels $\mathbf{y}$ by inserting a special symbol ⟨eokw⟩ after the occurence of the keyword. For example, when training with phoneme targets, for the utterance `the cat sat`, (which corresponds to the phoneme sequence[3] [D V ⟨eow⟩ k { t ⟨eow⟩ s { t ⟨eow⟩]), if we sampled $\mathbf{k}$ =`cat` as the keyword, then we would modify the target labels as, $\mathbf{y}$ = [D V ⟨eow⟩ k { t ⟨eow⟩ ⟨eokw⟩ s { t ⟨eow⟩]. Note that the ⟨eow⟩ token marks the end of each word token (see Section 3.2). The intuition behind adding the ⟨eokw⟩ at the end of the keyword phrase in the transcript, is that it might serve as a marker that the model should attend to the targets in the keyword phrase. As a final note, the training and inference algorithms for this model are similar to the standard RNN-T model.

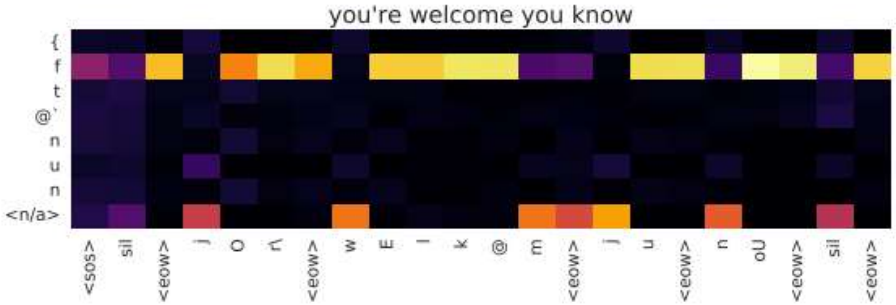
[3]We use X-SAMPA to denote phonemes throughout the paper.

Keywords Spotting
>>> *Cross Modality*









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

(b) Attention matrix of a negative utterance for the keyword "afternoon", with the transcript "you're welcome you know".

Streaming Small-Footprint Keyword Spotting Using Sequence-to-Sequence Models Google

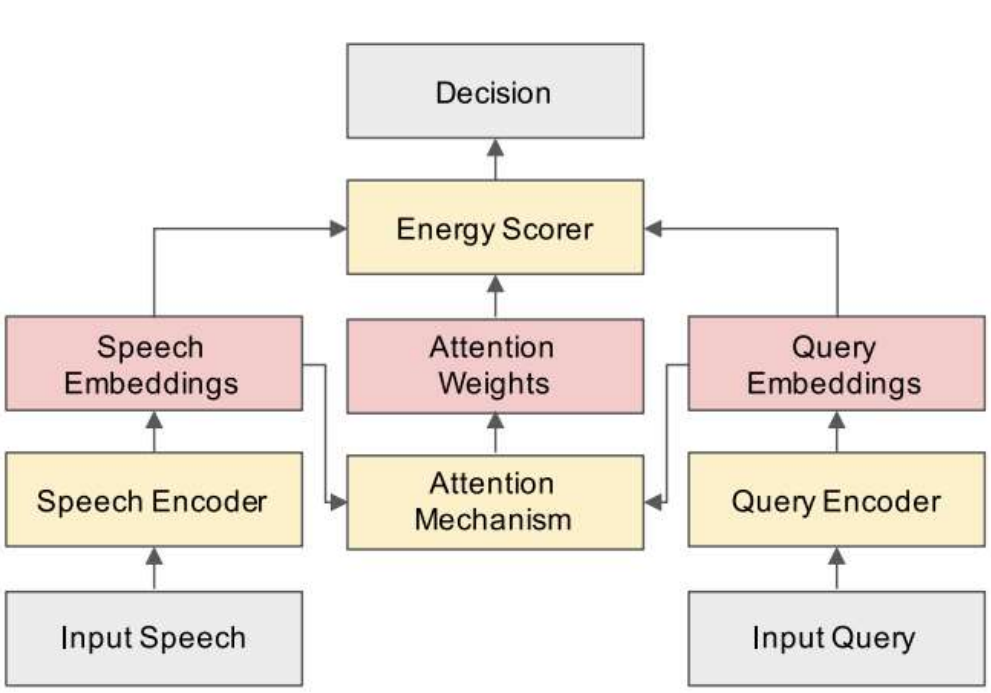
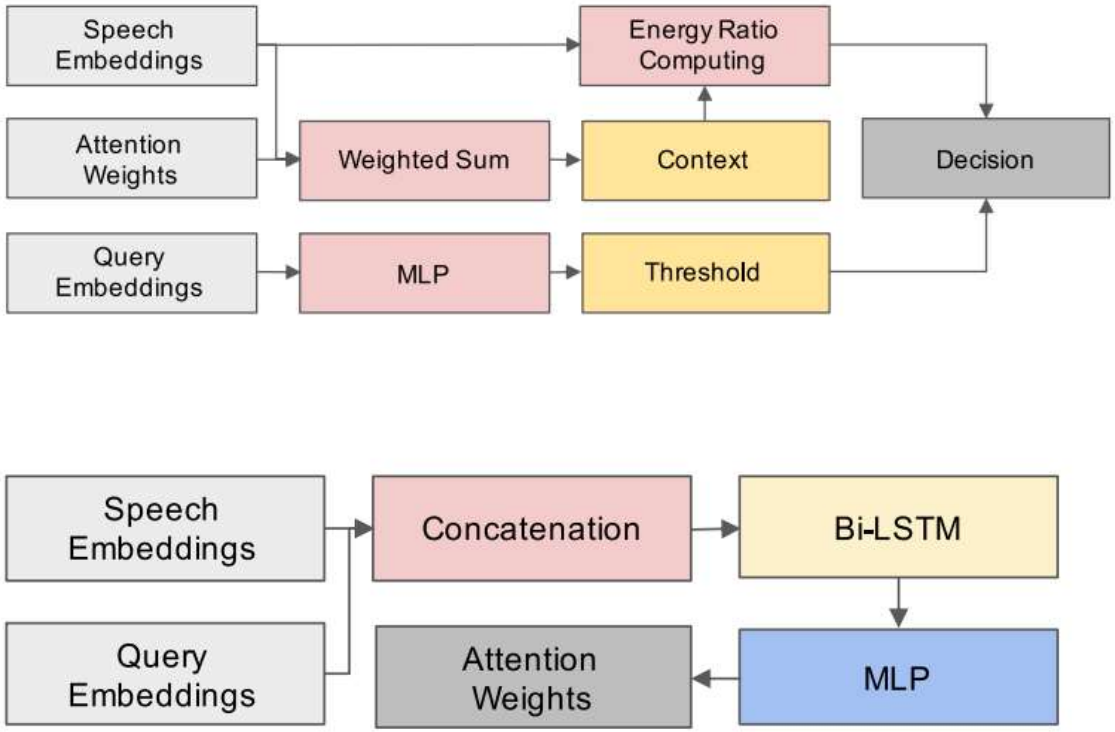


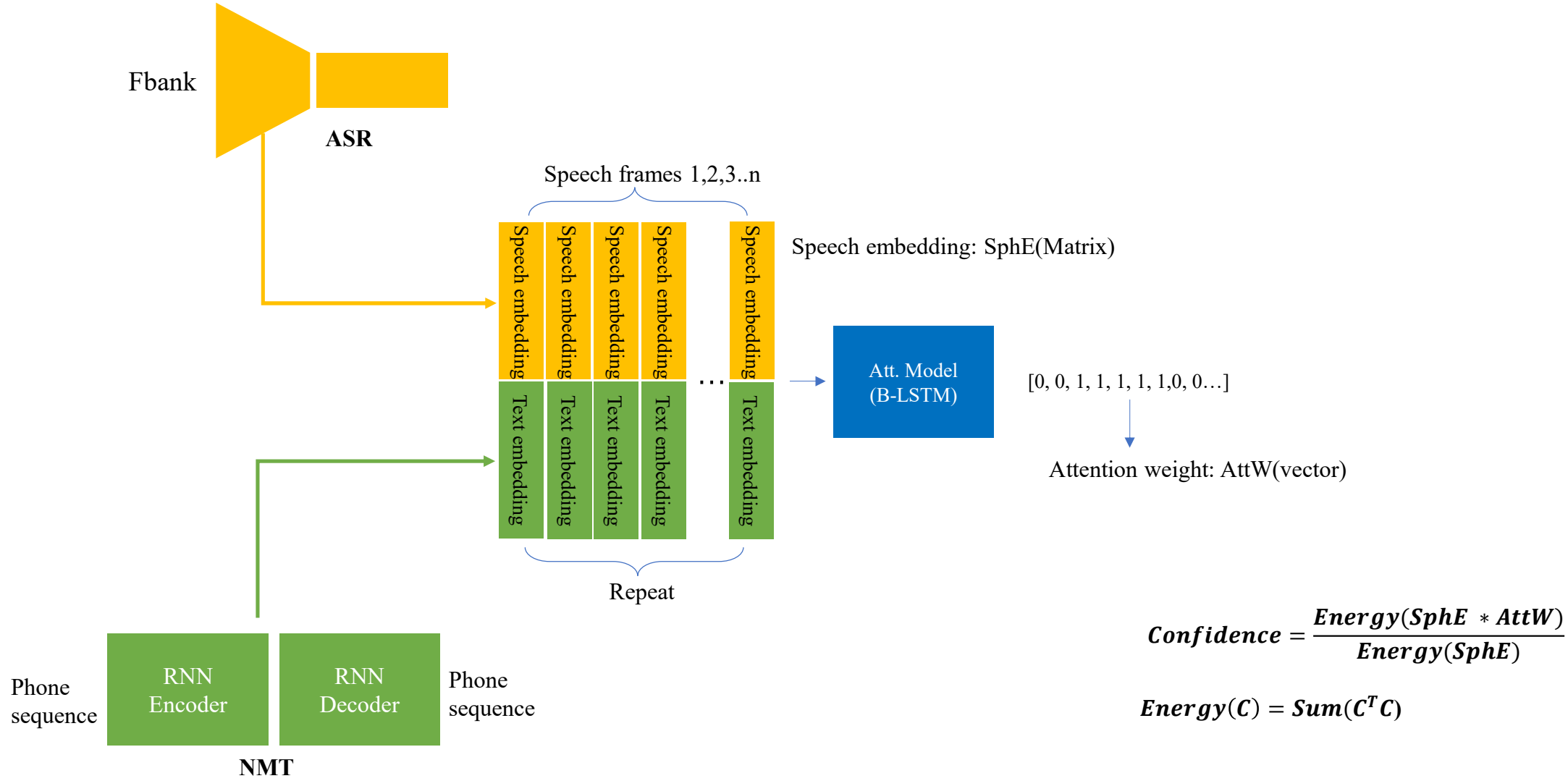
Fig. 1. The overall structure of our E2E KWS system.



End-to-end keyword search system based on attention mechanism and energy scorer for low resource languages THU

Keywords Spotting

>>> *Cross Modality*



$$Confidence = \frac{Energy(SphE * AttW)}{Energy(SphE)}$$

$$Energy(C) = Sum(C^T C)$$

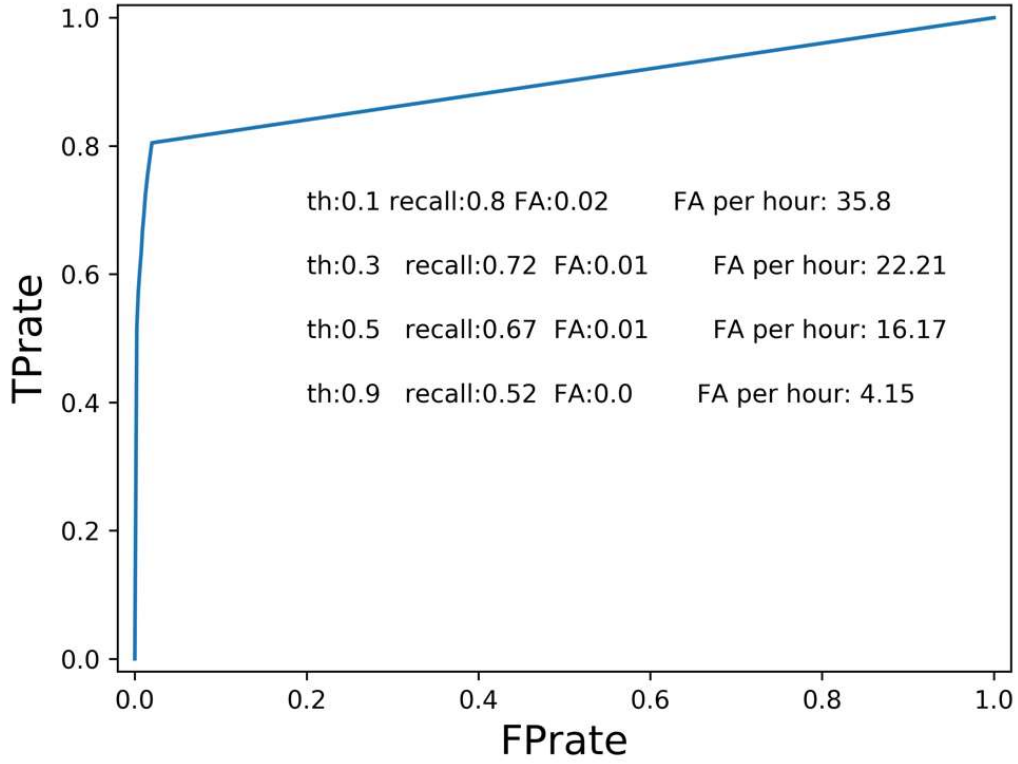
End-to-end keyword search system based on attention mechanism and energy scorer for low resource languages THU

Table 1
The KWS performance of ACC and AUC with different speech decoders for Assamese IV and OOV.

		CTC	Attention Seq2Seq	Baseline
ACC	IV	0.7061	0.7343	0.6135
	OOV	0.7049	0.7072	0.6042
AUC	IV	0.7737	0.7787	0.6384
	OOV	0.7715	0.7577	0.6320

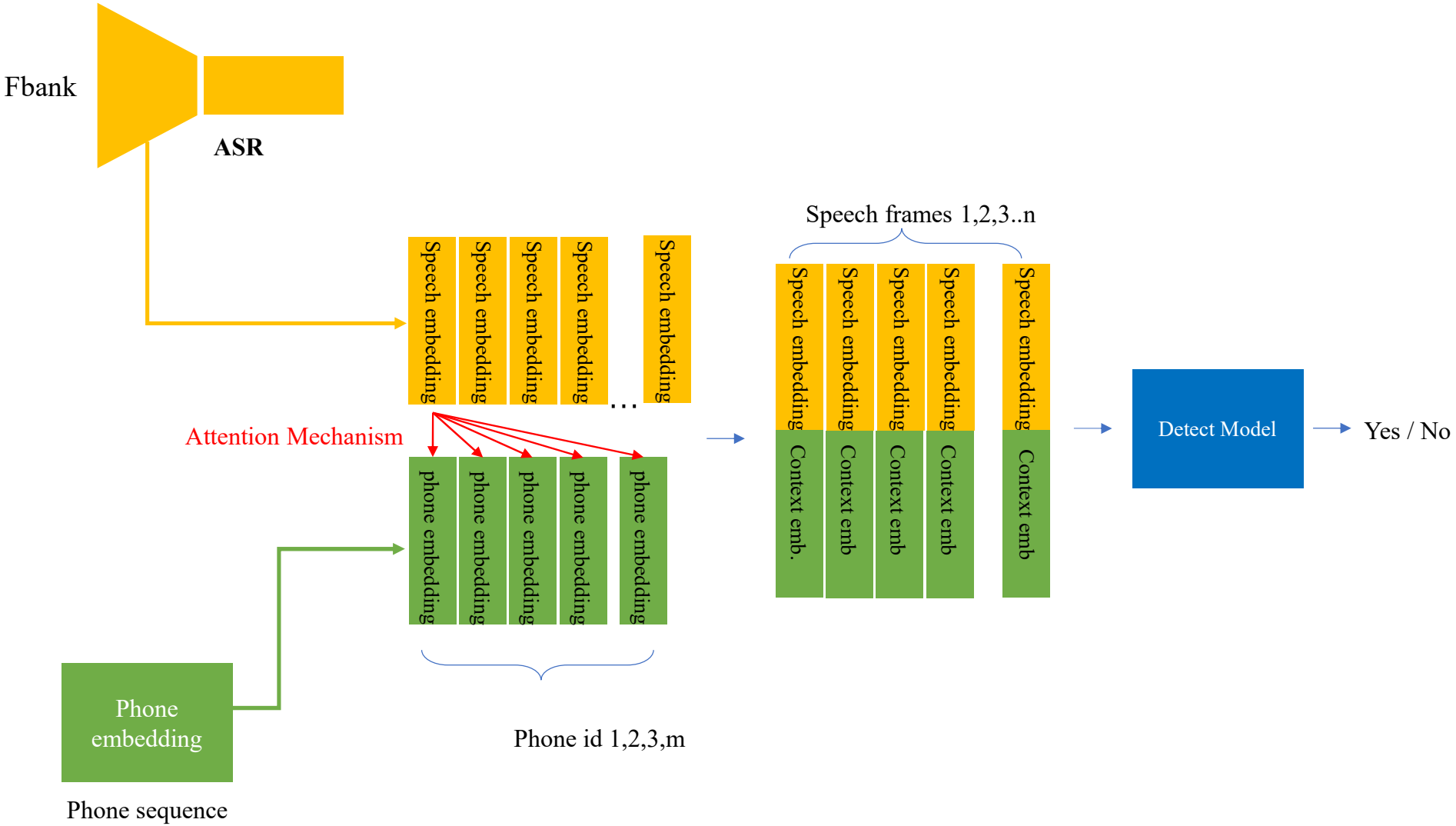
Table 2
The performance of ACC and AUC with pre-trained and un-pre-trained speech encoder–decoders for Assamese IV and OOV.

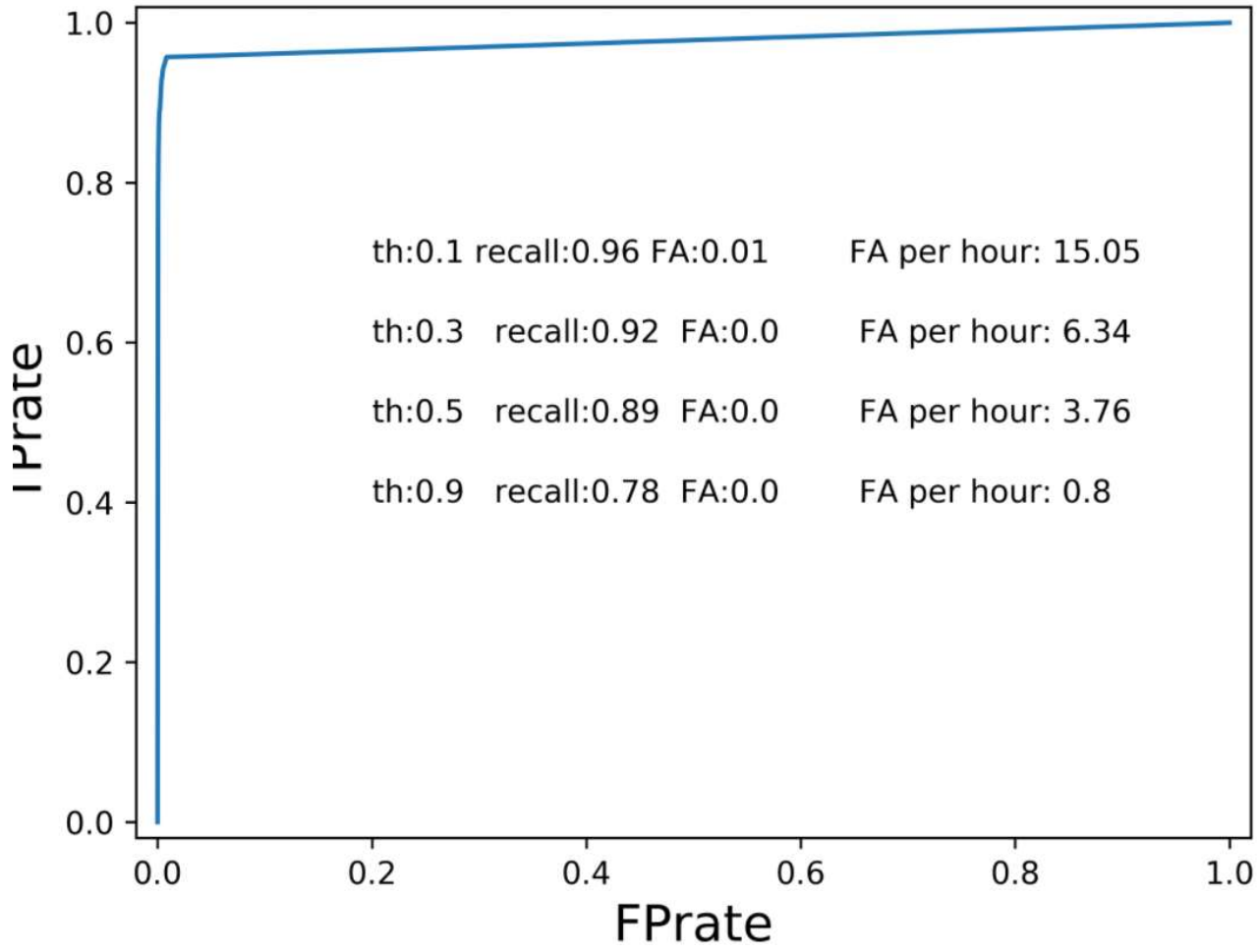
		Pre-trained	Un-pre-trained	Baseline
ACC	IV	0.7343	0.6380	0.6135
	OOV	0.7072	0.6318	0.6042
AUC	IV	0.7787	0.6945	0.6384
	OOV	0.7577	0.6930	0.6320



Keywords Spotting

>>> *Cross Modality Attention E2E KWS*




1.0
0.8
0.6
0.4
0.2
0.0
TPrate
th:0.1 recall:0.96 FA:0.01 FA per hour: 15.05
th:0.3 recall:0.92 FA:0.0 FA per hour: 6.34
th:0.5 recall:0.89 FA:0.0 FA per hour: 3.76
th:0.9 recall:0.78 FA:0.0 FA per hour: 0.8
0.0
0.2
0.4
0.6
0.8
1.0
FPrate

- Control FA
- Test on Overlap Speech

Thanks