Keyword Spotting

Zhiyong Zhang

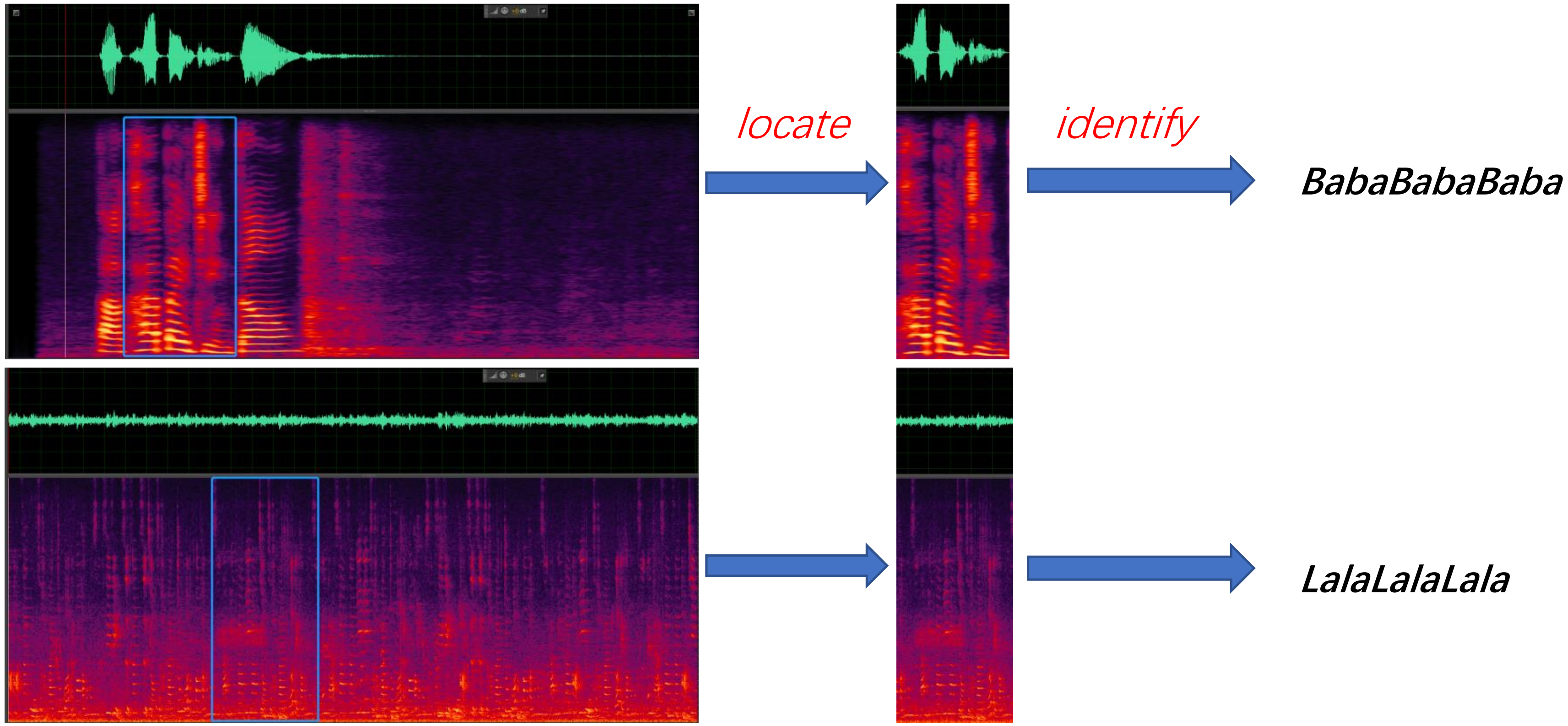
2022.01.10

Outline

- Introduction to online KWS
- Anchor-aware KWS

What’s KWS

- Problem statement
  - ✓Locate and identify interesting word in continuous speech signal

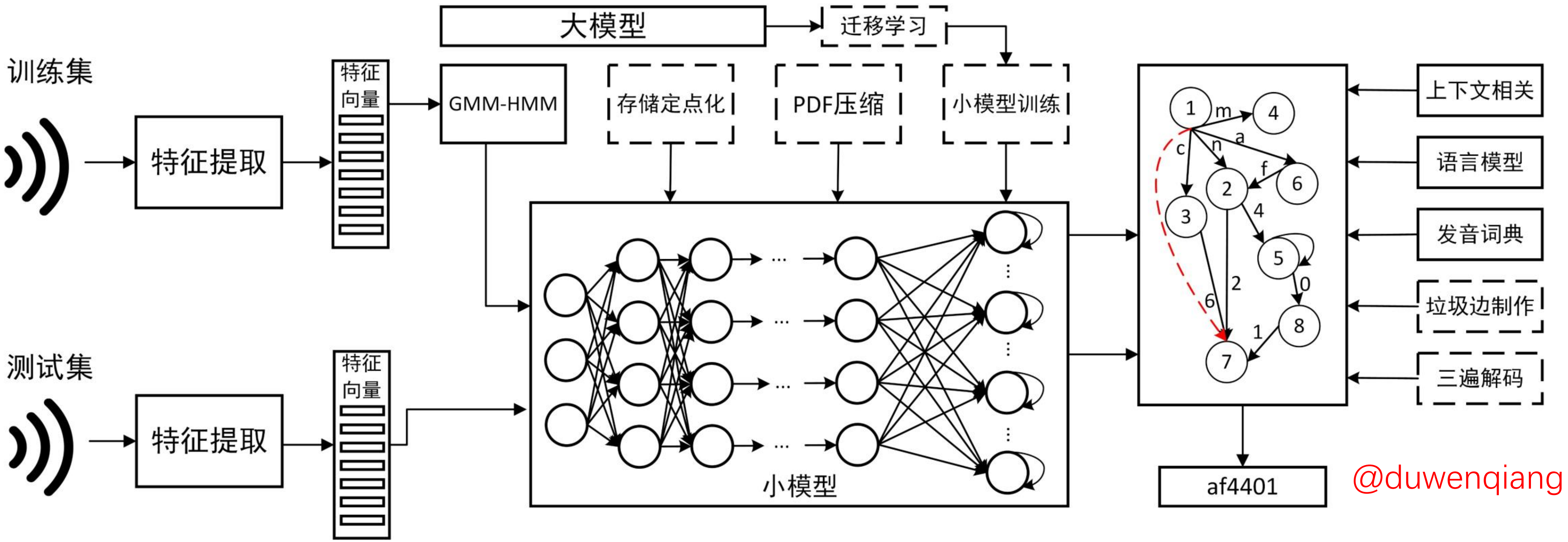


Good Properties of KWS

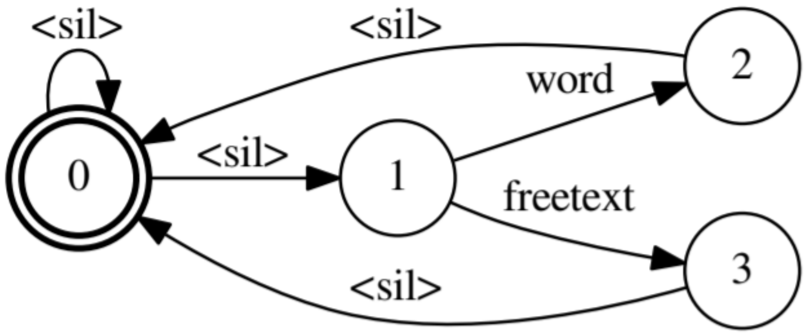
- Versatile
  - ✓Open vocabulary, keywords can be arbitrarily added or removed
- User-friendly
  - ✓Text or speech what you like
- Robust to OOVs
  - ✓Independent with training
- Computationally-efficient
  - ✓Low memory occupation, high computation efficient

Sacchi, N. , Nanchen, A. , Jaggi, M. , & Cernak, M. . (2019). Open-Vocabulary Keyword Spotting with Audio and Text Embeddings. Interspeech 2019.
https://publications.idiap.ch/attachments/papers/2019/Sacchi_INTERSPEECH_2019.pdf

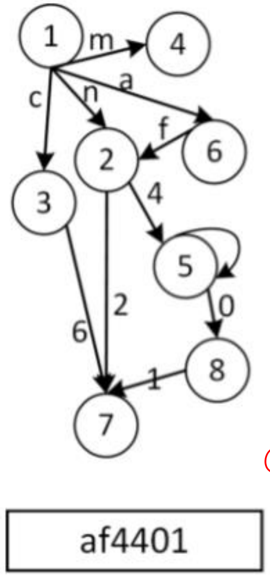
Framework of KWS

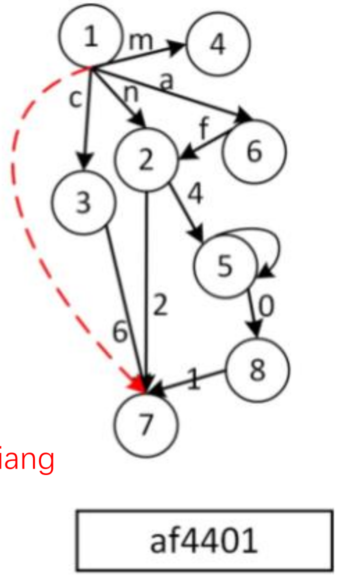


Keyword-filler KWS

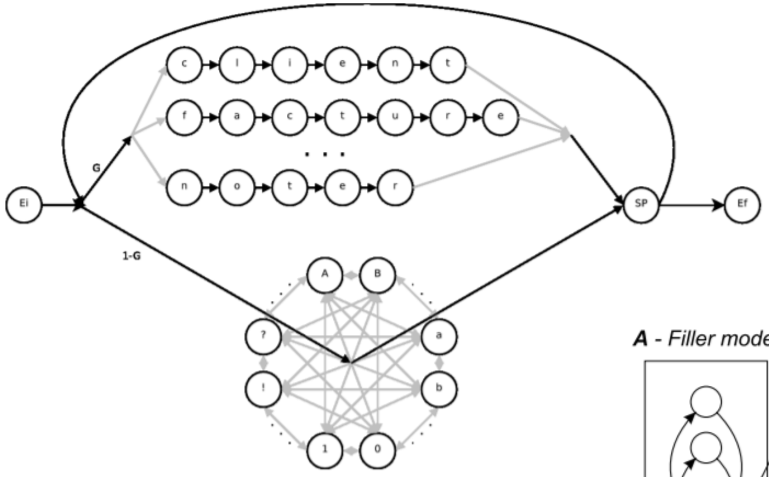


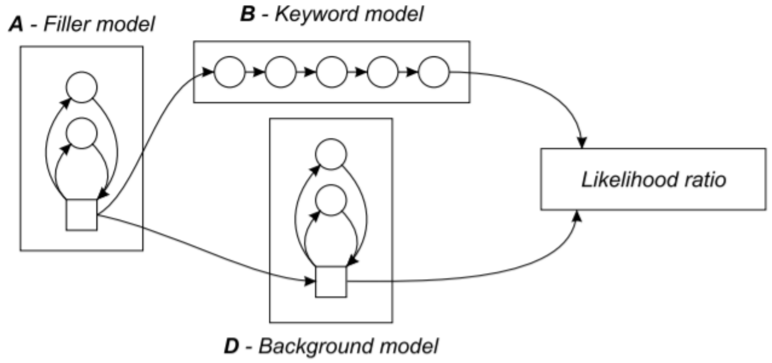
WAKE WORD DETECTION AND ITS APPLICATIONS
https://jscholarship.library.jhu.edu/bitstream/handle/1774.2/64380/WANG-DISSERTATION-2021.pdf?sequence=1&isAllowed=y



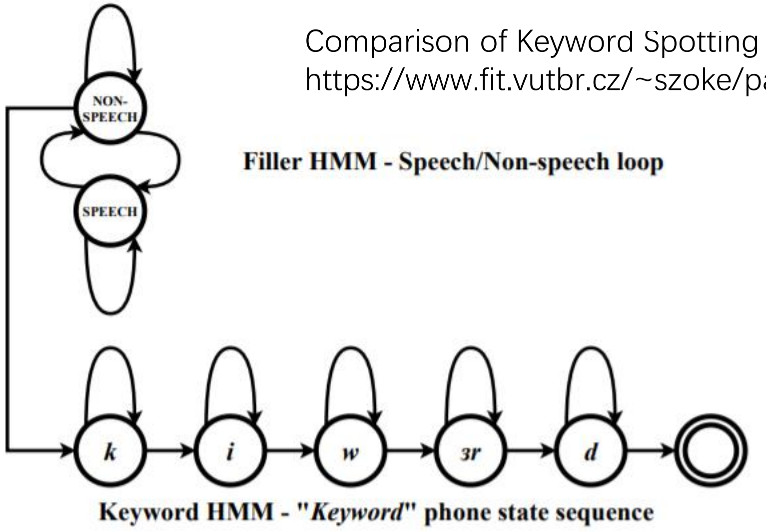


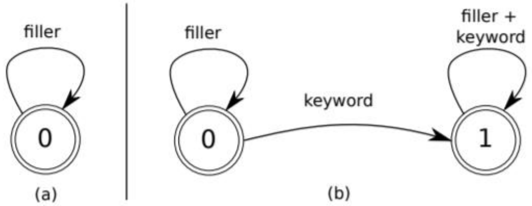
@duwenqiang





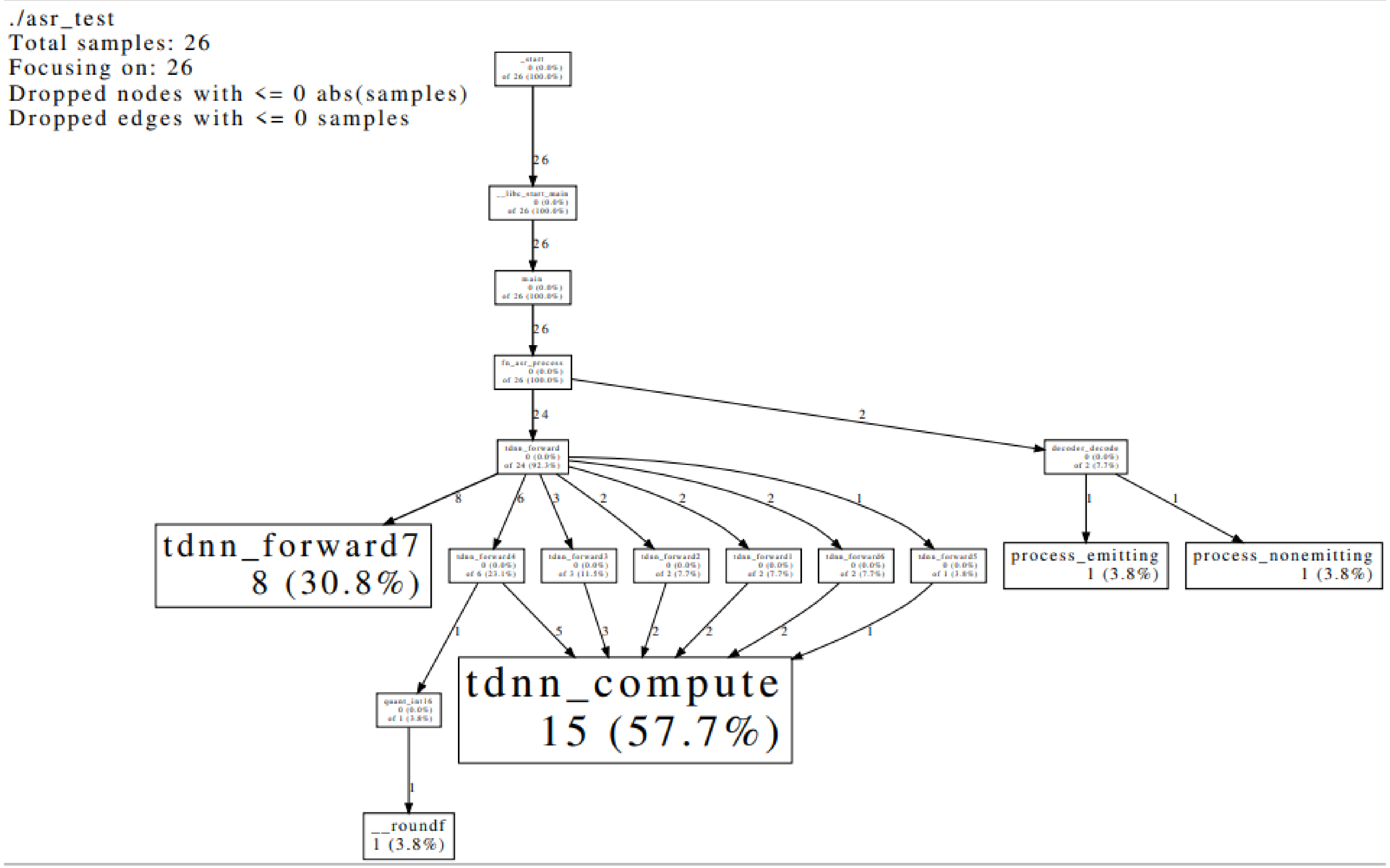
Comparison of Keyword Spotting Approaches for Informal Continuous Speech
https://www.fit.vutbr.cz/~szoke/papers/mlmi_2005.pdf



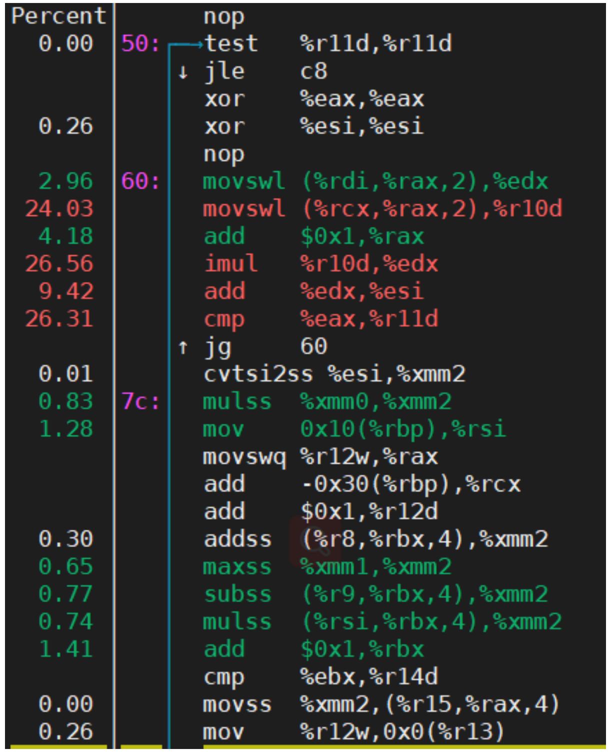


STREAMING SMALL-FOOTPRINT KEYWORD SPOTTING USING SEQUENCE-TO-SEQUENCE MODELS
https://arxiv.org/pdf/1710.09617.pdf

Engine Performance Analysis



- 88.5% computation ratio in NN
- Data move and mul. opt.
- 7.6% in decoding
- 2% in residual operation



Results

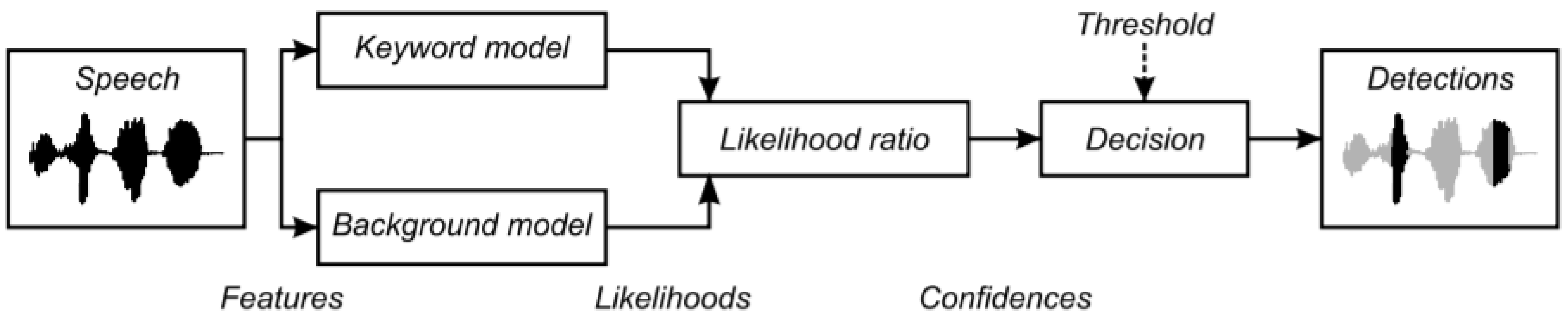
模型	字错误率	大小（M）	pdf
1000-tdnn-f-chain-6layer_dim512_pdf4000	16.55%	20	3360
1000h-cmd-dim128-5layer_pdf500_outdim500	29.68%	2.9	440
1000h-cmd-dim128-5layer_pdf500_outdim300	29.65%	2.0	440
1000h-cmd-dim128-5layer_pdf2232	26.48%	3.1	2232
1800h-cmd-dim128-5layer_pdf500_outdim500	28.51%	2.9	448
1800h-cmd-dim128-5layer_pdf500_outdim300	28.20%	2.1	448

模型		AirportDaxingTest1	AirportDaxingTest2	size(M)
M1	1800h_tdnn_dim256-5layer_pdf856_outdim500	27.07%	28.69%	8.3
M2	1800h_tdnn_dim256-5layer_pdf448_outdim500	24.09%	26.64%	6.1
M3	1800h_tdnn_dim368-5layer_pdf448_outdim500	24.44%	26.99%	8.9
M4	1800h_tdnn_dim512-5layer_pdf448_outdim500	23.20%	26.87%	14
M5	1800h_tdnn-f_dim256_dim512_7layer_pdf3256_outdim800	18.97%	24.87%	22
M6	1800h_tdnn_dim128-5layer_pdf448_outdim300	27.65%	26.73%	2.1
M7	1800h_tdnn_dim256-5layer_pdf448_outdim300	28.73%	30.18%	3.9
M8	1800h_tdnn_dim368-5layer_pdf448_outdim300	24.01%	25.37%	8.9
M9	1800h_tdnn_dim512-5layer_pdf448_outdim300	22.23%	25.48%	14
M10	1800h_tdnn_dim1024-5layer_pdf448_outdim300	23.05%	26.03%	41

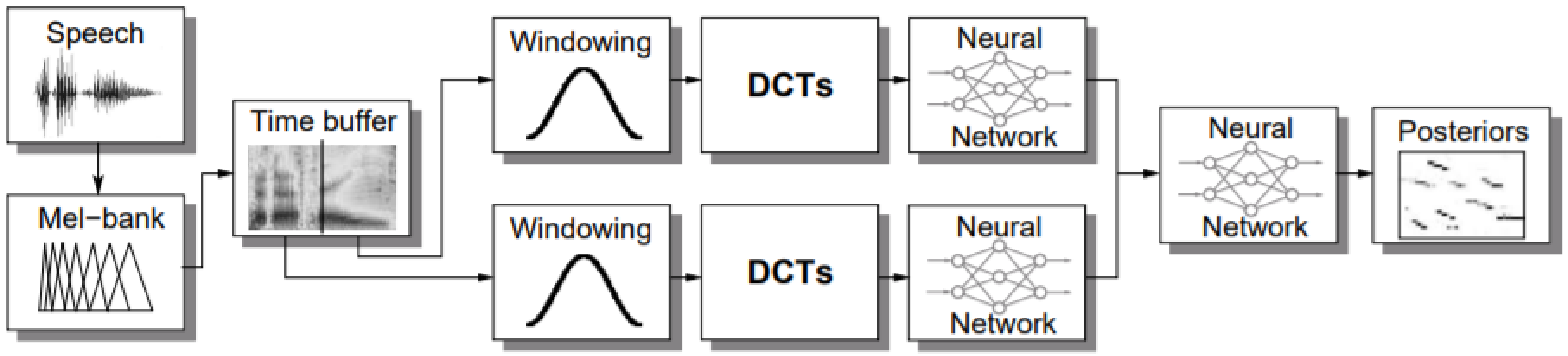
- Some conclusion
  - Recognition accuracy is highly related with parameter size.
  - Big-parameter does not mean best result.
  - Performance is data-driven

- How to select appropriate model structure and leverage its capacity
  - Data distribution
  - Local or global property of task
  - Chain/CNN/CRNN/Transformer/Conformer/Audiomer

Different Keyword Spotting Systems



Keyword-filler KWS

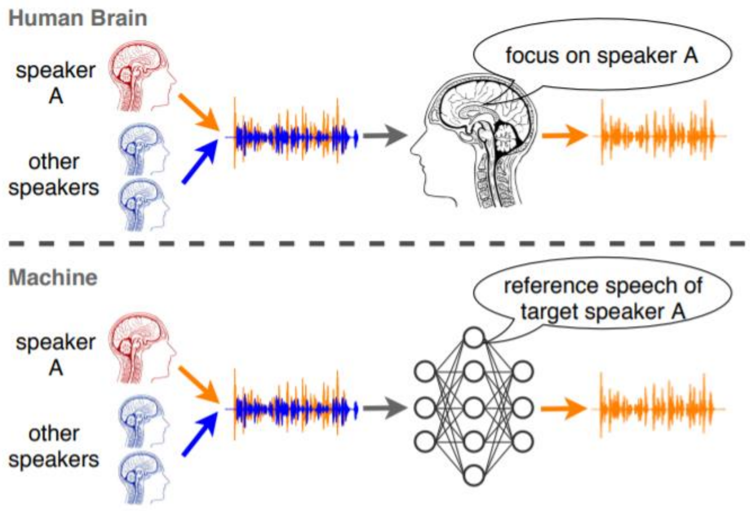


Neural Network KWS

Comparison of Keyword Spotting Approaches for Informal Continuous Speech
https://www.fit.vutbr.cz/~szoke/papers/mlmi_2005.pdf

Anchor-Aware Keyword Spotting

- Selective Auditory Attention
  - These remarkable abilities are implemented with accurate processing of low-level stimulus attributes, segregation of auditory information into coherent voices, and selectively attending to a voice at the exclusion of others to facilitate higher level processing
  - Attention is not a static, one way information distillation process. It is believed to be a modulation of focus between the bottom-up sensory-driven factors…



What's Anchor

- Domain
- Noise
- Speaker
- Text
- Sound event

Domain aware KWS

*Domain Aware Training for Far-field Small-footprint Keyword Spotting

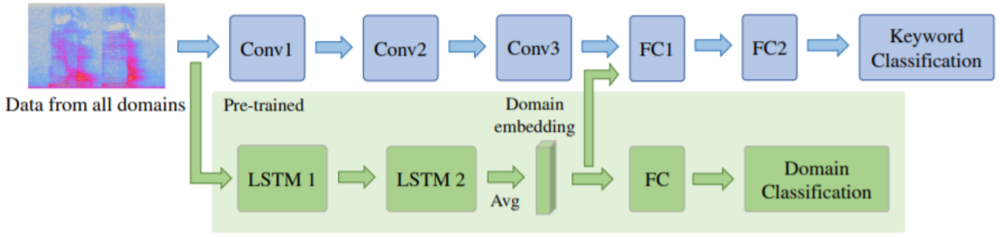


Figure 1: *Framework of the domain embedding system.*

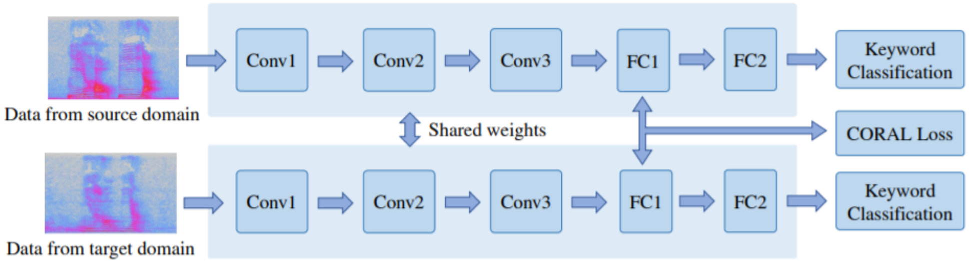


Figure 2: *Framework of the CORAL system.*

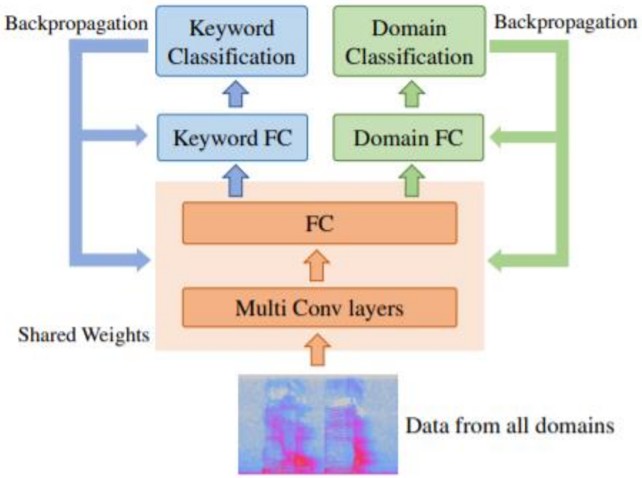


Figure 3: *Framework of the MTL system.*

Table 3: *Performance of the baseline system (the false reject (FR) rate (%) under one false alarm (FA) per hour)*

Training set	0.25M	1M	3M
Only 0.25M	1.29	2.91	11.6
Only 1M	2.03	1.58	7.77
Only 3M	10.9	8.00	10.6
Mix of 0.25M and 1M	**0.91**	**1.38**	6.06
Mix of 0.25M and 3M	1.54	1.97	**5.60**
Mix of all distances	1.41	1.64	6.33

Table 4: *Performances of models trained with different methods on the test sets*

Model name	0.25M	1M	3M
EMB1	1.11	1.59	4.99
EMB2	1.21	**1.02**	**4.11**
CORAL1	1.37	1.05	4.69
CORAL2	1.19	1.41	5.02
CORAL3	**1.09**	1.52	5.97
CORAL4	1.27	1.47	5.21
CORAL5	1.21	1.41	4.78
MTL	1.70	1.44	5.15

Acoustic-feature: x = {x1, x2, · · · xTs}
WUW: w = {w1, w2 · · · wM}

$$s_{w_i}(\boldsymbol{x}_t) = \frac{1}{L} \sum_{j=t-L-1}^{t} p_{w_i}(\boldsymbol{x}_j),$$

$$h(\boldsymbol{x}) = \left[\max_{1 \le t_1 < \cdots < t_M \le T_s} \prod_{i=1}^{M} s_{w_i}(\boldsymbol{x}_{ti}) \right]^{\frac{1}{M}}$$

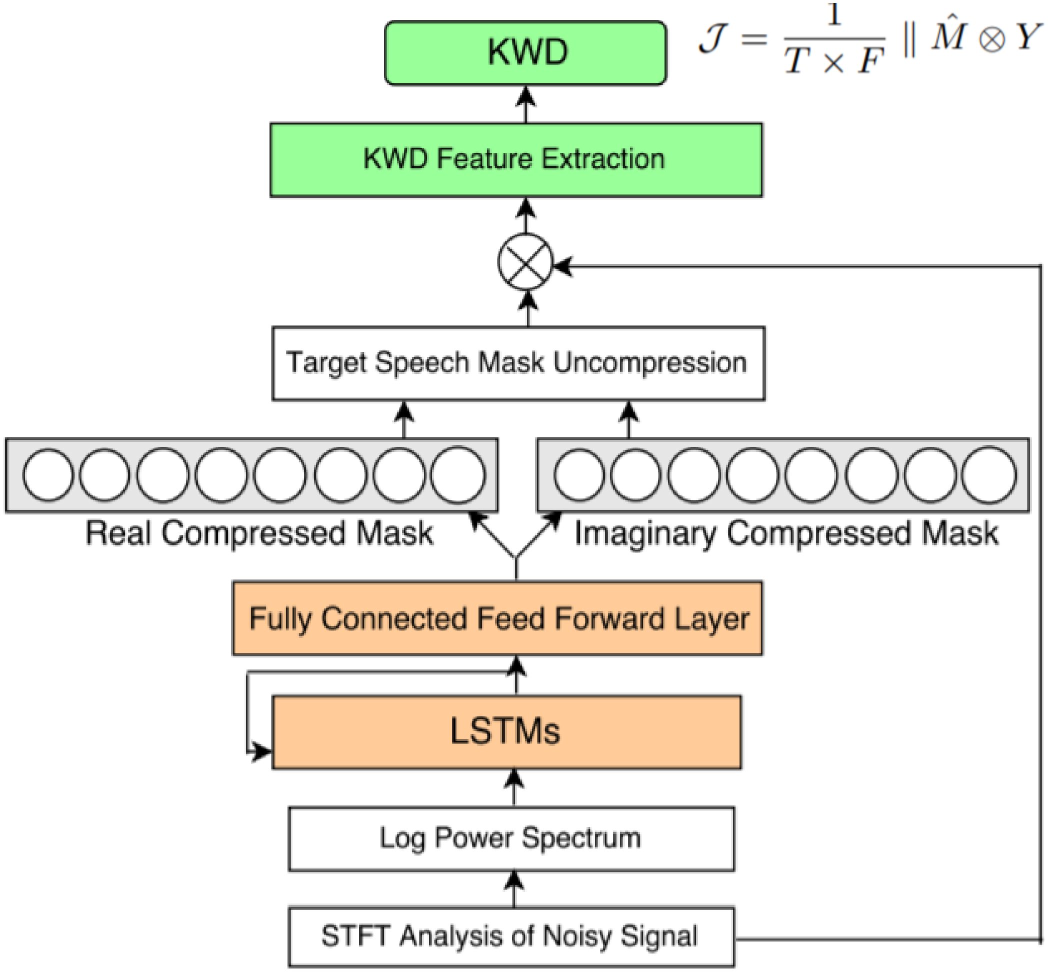
Simple is best?

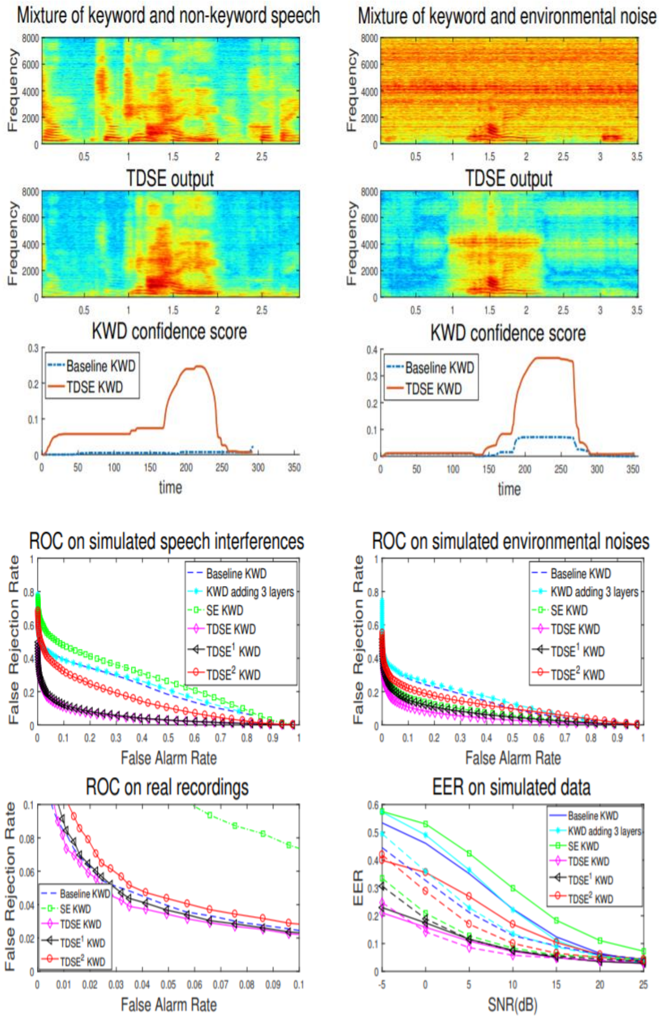
Wu, H. , Jia, Y. , Nie, Y. , & Li, M. . (2020). Domain Aware Training for Far-field Small-footprint Keyword Spotting. Interspeech 2020.
https://indico2.conference4me.psnc.pl/event/35/contributions/3468/attachments/1026/1067/Wed-2-2-4.pdf

Text-dependent KWD

*Text-Dependent Speech Enhancement for Small-Footprint Robust Keyword Detection

$$\mathcal{J} = \frac{1}{T \times F} \parallel \hat{M} \otimes Y - S \parallel_F^2,$$

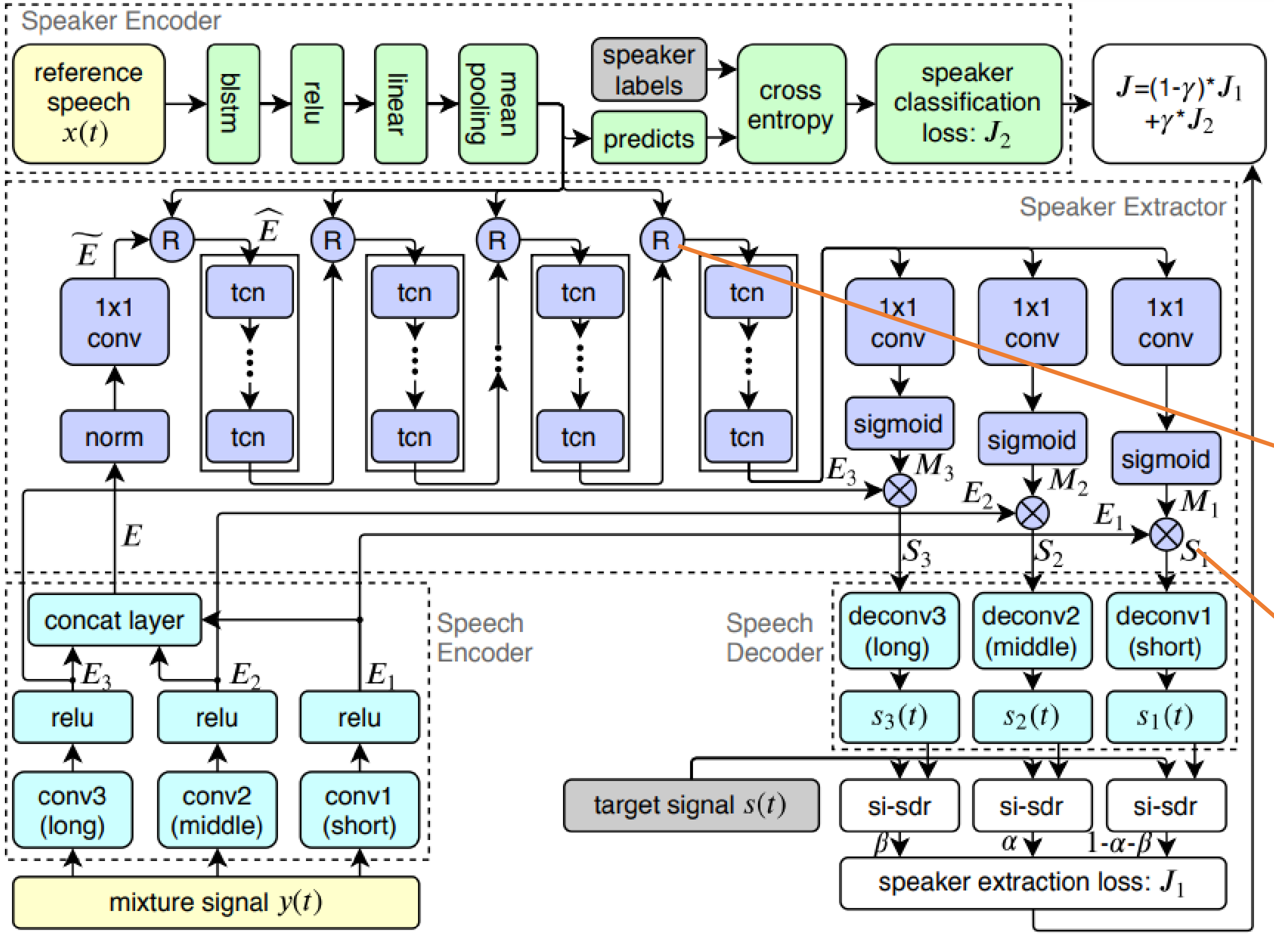


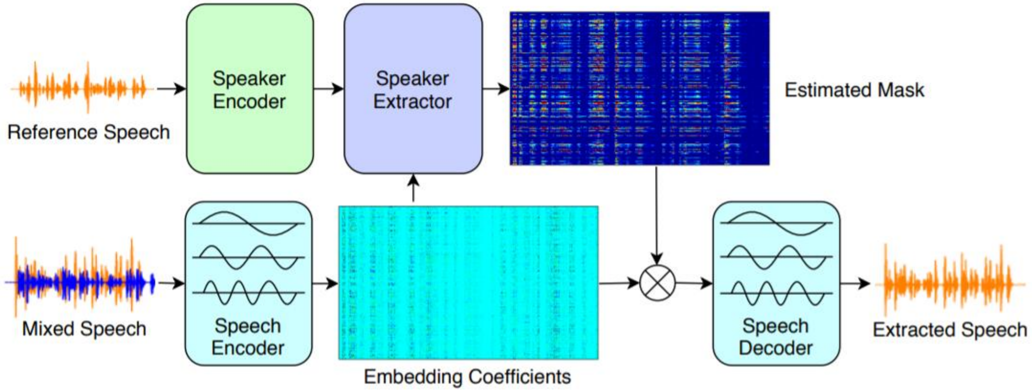


https://ai.tencent.com/ailab/media/publications/Text-DependentSpeechEnhancementforSmall-FootprintRobustKeyword_Detection.pdf

Speaker extraction

*SpEx: Multi-Scale Time Domain Speaker Extraction Network





Concatenate Speaker vector repeatedly to the intermediate representations along channel dimension.

$$
\begin{aligned}
S_i &= M_i \otimes E_i \\
&= f(E, g(x)) \otimes E_i
\end{aligned}
$$

Speaker encoder serves as the TOP-DOWN voluntary focus in selective auditory attention.

https://arxiv.org/pdf/2004.08326.pdf

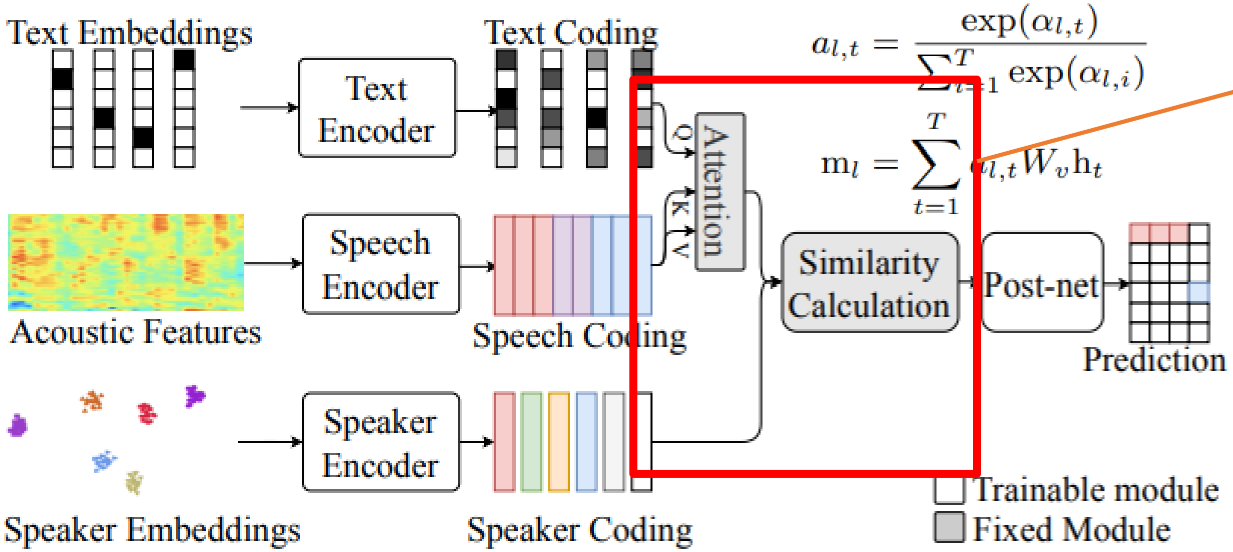
Speaker & Text-Aware diarization

*Speaker embedding-aware neural diarization for flexible number of speakers with textual information

$$\alpha_{l,t} = \textbf{dot}\left(W_q \mathrm{u}_l, W_k \mathrm{h}_t\right)$$

$$a_{l,t} = \frac{\exp(\alpha_{l,t})}{\sum_{i=1}^{T} \exp(\alpha_{l,i})}$$

$$\mathrm{m}_l = \sum_{t=1}^{T} a_{l,t} W_v \mathrm{h}_t$$



How to modulate the relation-ship of selective information

$$A_{t,n} = \textbf{dot}(\mathrm{h}_t, \mathrm{e}_n), A_{t,n} \in [-\infty, \infty]$$

$$A_{t,n} = \textbf{dot}(\sigma(\mathrm{h}_t), \sigma(\mathrm{e}_n)), A_{t,n} \in [-D, D]$$

$$A_{t,n} = \textbf{dot}(\textbf{Norm}(\mathrm{h}_t), \textbf{Norm}(\mathrm{e}_n)), A_{t,n} \in [-1, 1]$$

Table 3: The world-level DER (%) of different models on the simulation set.

Model	SC	Training Text	Grand	Recognition
Exp 1	×	Recognition	3.12	3.28
Exp 2	×	Grand	2.97	3.19
Exp 3	√	Recognition	1.82	2.08
Exp 4	√	Grand	**1.66**	**1.93**

Table 1: The DERs (%) of different similarity metrics on the simulation set.

Metrics	DER(Con.)	DER(Olp.)
cosine	6.23	12.62
dot	**3.63**	8.42
σ-dot	4.23	**7.87**

https://arxiv.org/pdf/2111.13694.pdf

Thanks