DEEP SPOKEN KEYWORD SPOTTING

1. Introduction

Iván López-Espejo

Department of Electronic Systems, Aalborg University, Denmark

ivl@es.aau.dk

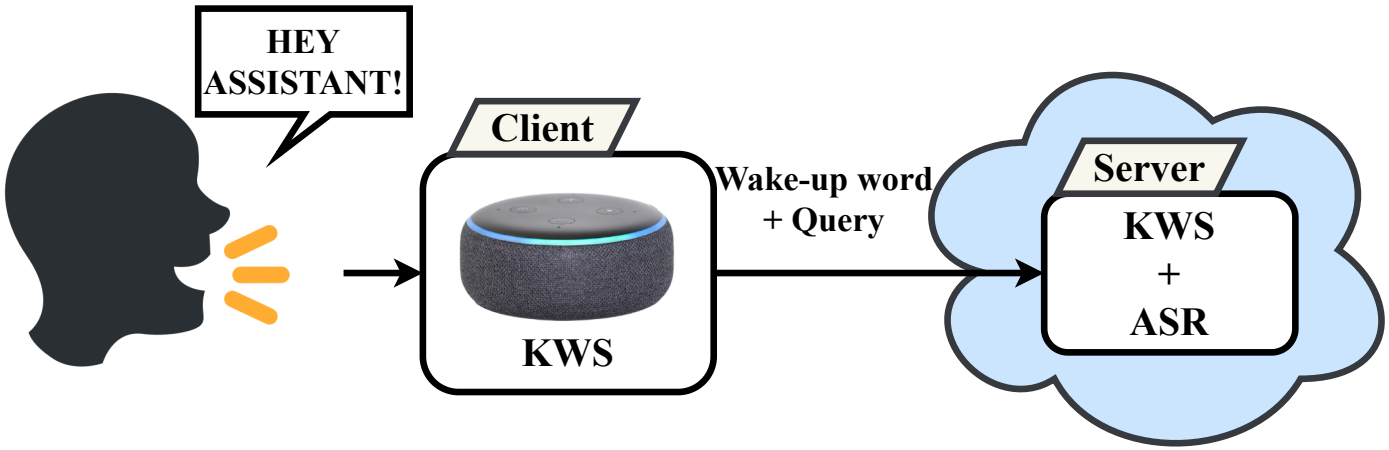
Sunday 18th September, 2022

INTERSPEECH 2022
September 18 - 22 • Incheon Korea

Overview

Introduction

- Speech technologies have become ubiquitous in nowadays society

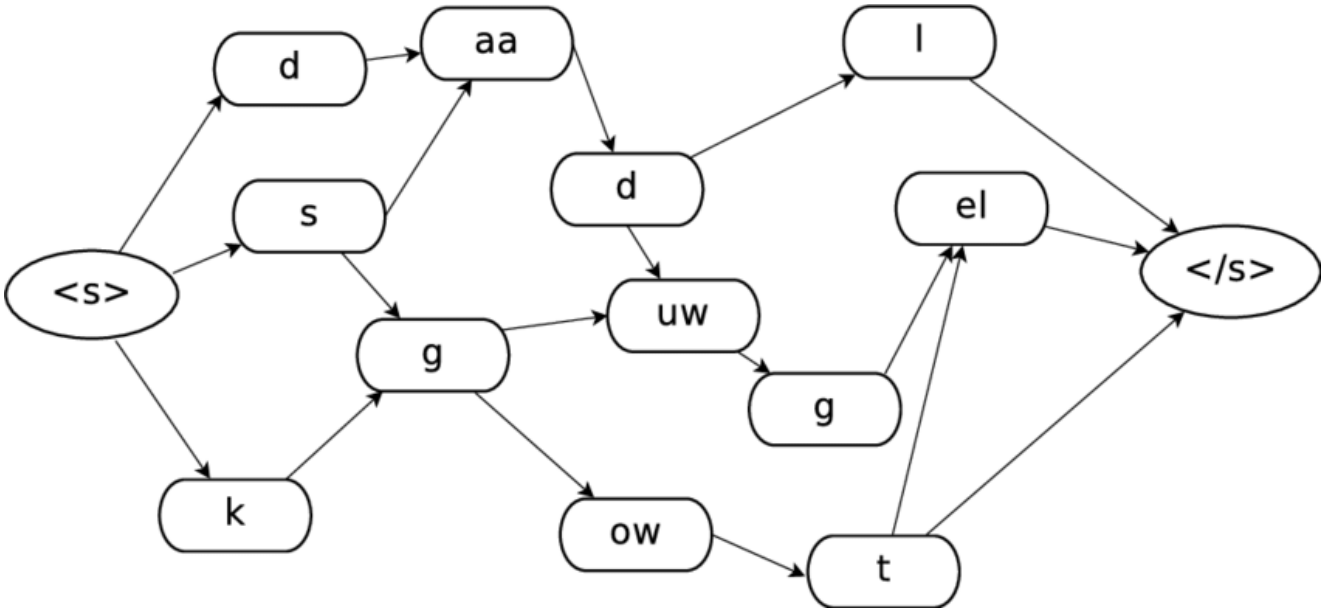


- Spoken **keyword spotting** (**KWS**) can be defined as the task of identifying keywords in audio streams comprising speech
- Applications of KWS: speech data mining, audio indexing, phone call routing, etc.

General Approaches

Over the years, different techniques have been explored for KWS:

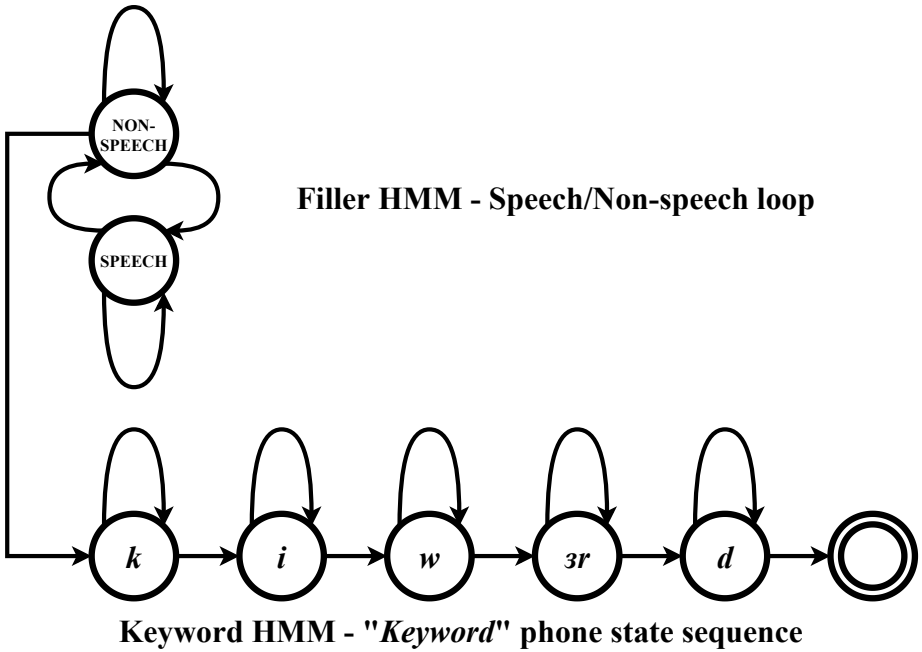
- **1) Large-vocabulary continuous speech recognition**
  - ☑ Flexibility to deal with non-predefined keywords
  - ☒ High computational complexity



Dong Wang, "Out-of-Vocabulary Spoken Term Detection". 2010

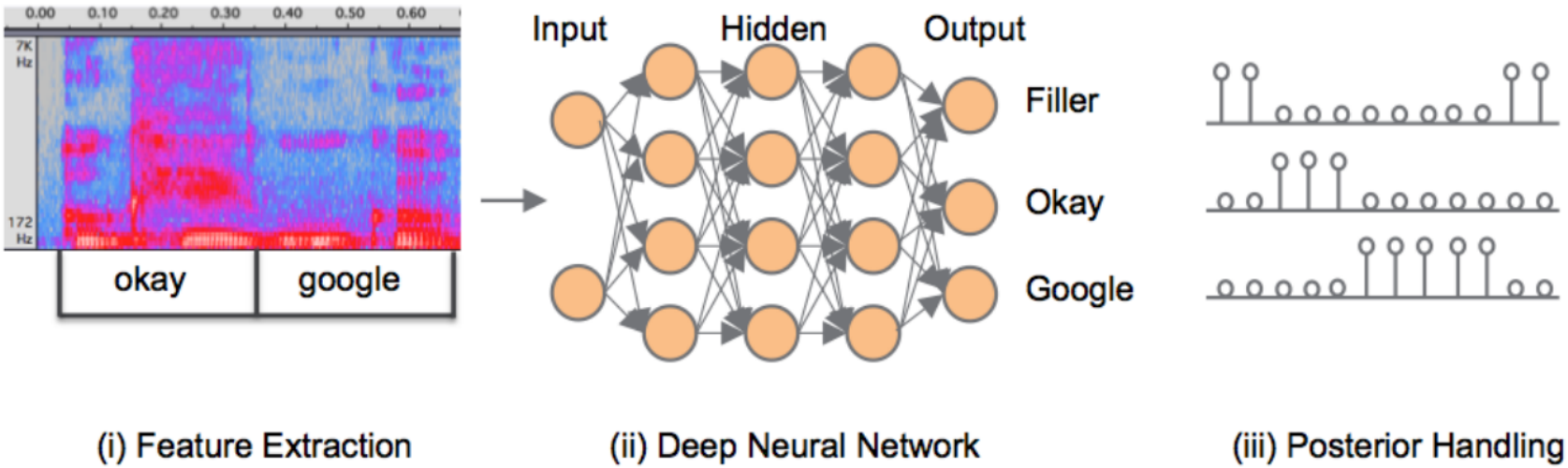
Over the years, different techniques have been explored for KWS:

- **2) Keyword/filler hidden Markov model (HMM)**
  - ☑ Good performance
  - ☒ Viterbi decoding is still needed



Over the years, different techniques have been explored for KWS:

- **3) Deep spoken keyword spotting**
  - ☑ No complicated sequence search algorithm
  - ☑ Adjustable complexity
  - ☑ Significant improvements over keyword/filler HMM in small-footprint scenarios



Guoguo Chen, Carolina Parada and Georg Heigold, "Small-footprint keyword spotting using deep neural networks". In Proc. of ICASSP 2014

Motivation and Goal of the Tutorial

- Deep KWS is very appealing to be deployed to a variety of consumer electronics with limited resources like earphones and headphones, smartphones, smart speakers and so on
- Much research on deep KWS has been conducted since 2014 until today
- We can expect that deep KWS will continue to be a hot topic in the future!
- Forecasts suggest that, by 2024, the number of voice assistant units will exceed that of world's population

For today

We will present a review into deep spoken KWS intended for practitioners and researchers who are interested in this technology

Organization of the Tutorial

1. **Introduction** (*Iván López-Espejo*)
2. **The Deep Spoken Keyword Spotting Approach** (*Iván López-Espejo*)
3. **Robustness in Keyword Spotting** (*John H. L. Hansen*)
4. **BREAK** (∼15 min)
5. **Audio-Visual Keyword Spotting** (*Zheng-Hua Tan*)
6. **Technology Applications** (*Zheng-Hua Tan*)
7. **Experimental Considerations** (*Iván López-Espejo*)
8. **Conclusions and Future Directions** (*Iván López-Espejo*)
9. **Q&A** (*Iván López-Espejo*, *Zheng-Hua Tan* and *John H. L. Hansen*)

Iván López-Espejo
Aalborg University (Denmark)

Zheng-Hua Tan
Aalborg University (Denmark)

John H. L. Hansen
The University of Texas at Dallas (USA)

DEEP SPOKEN KEYWORD SPOTTING

1. Introduction

Iván López-Espejo

Department of Electronic Systems, Aalborg University, Denmark

ivl@es.aau.dk

Sunday 18th September, 2022

INTERSPEECH 2022
September 18 - 22 • Incheon Korea

DEEP SPOKEN KEYWORD SPOTTING

2. The Deep Spoken Keyword Spotting Approach

Iván López-Espejo

Department of Electronic Systems, Aalborg University, Denmark

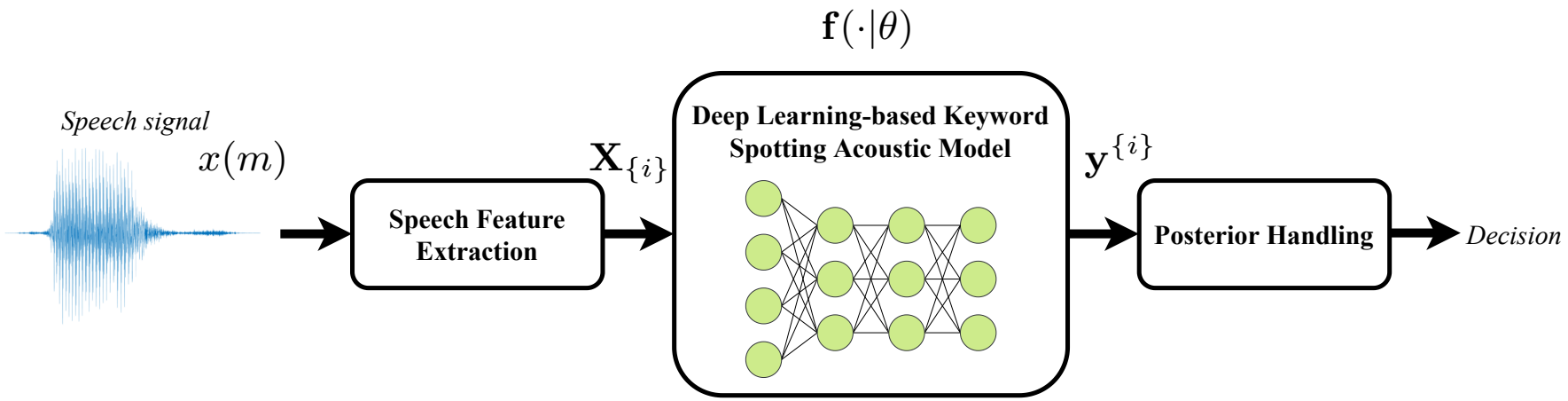
ivl@es.aau.dk

Sunday 18th September, 2022

INTERSPEECH 2022
September 18 - 22 • Incheon Korea

Overview

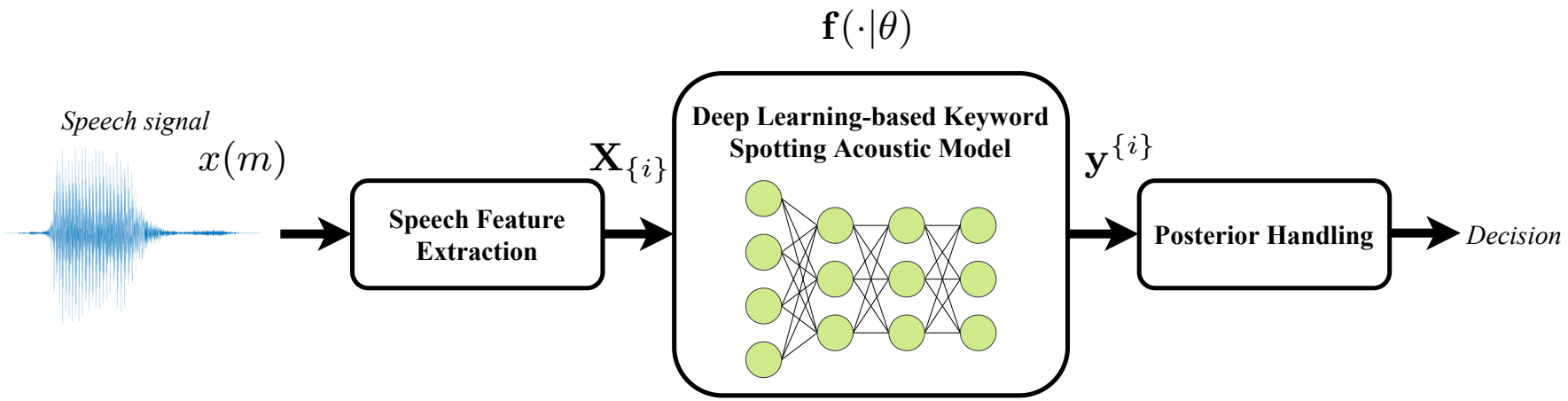
The General Pipeline



Three main blocks:

- Speech feature extractor
- Deep learning-based acoustic model
- Posterior handler

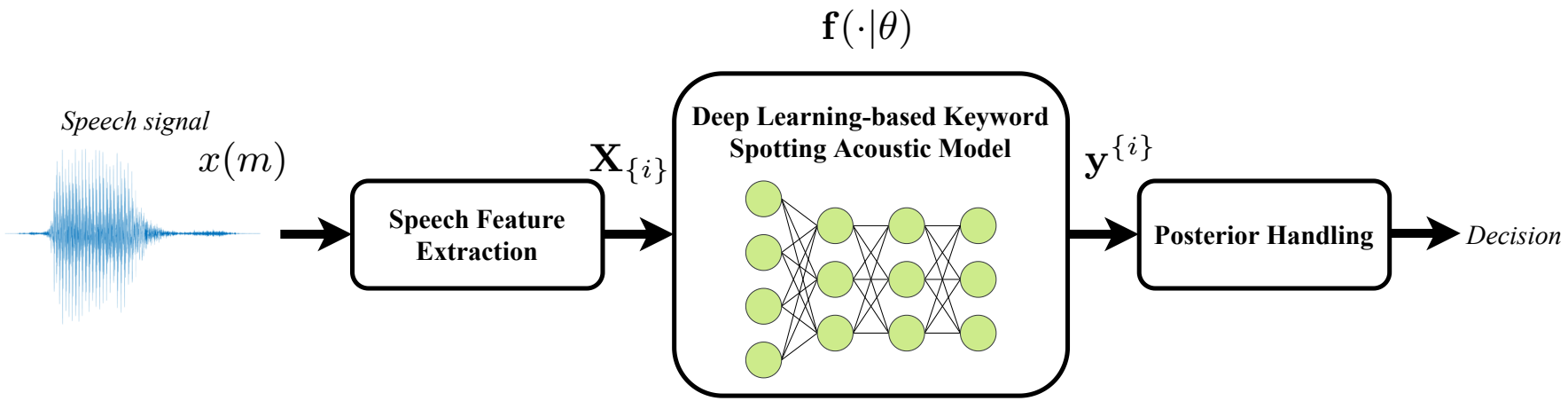
The General Pipeline



Speech feature extraction

- $x(m) \longrightarrow$ Speech feature extractor $\longrightarrow \mathbf{X} = (\mathbf{x}_0, ..., \mathbf{x}_t, ..., \mathbf{x}_{T-1}) \in \mathbb{R}^{K \times T}$
- $\mathbf{X}_{\{i\}} = (\mathbf{x}_{is-P}, ..., \mathbf{x}_{is}, ..., \mathbf{x}_{is+F}) \in \mathbb{R}^{K \times (P+F+1)}$, $i = \lceil \frac{P}{s} \rceil, ..., \lfloor \frac{T-1-F}{s} \rfloor$
- Typically, $F < P$

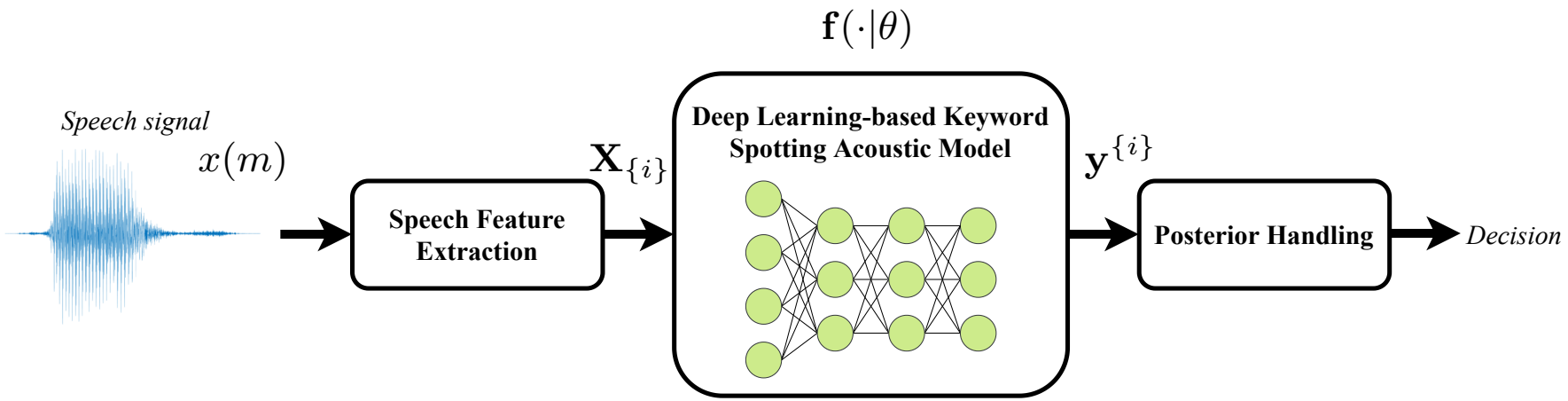
The General Pipeline



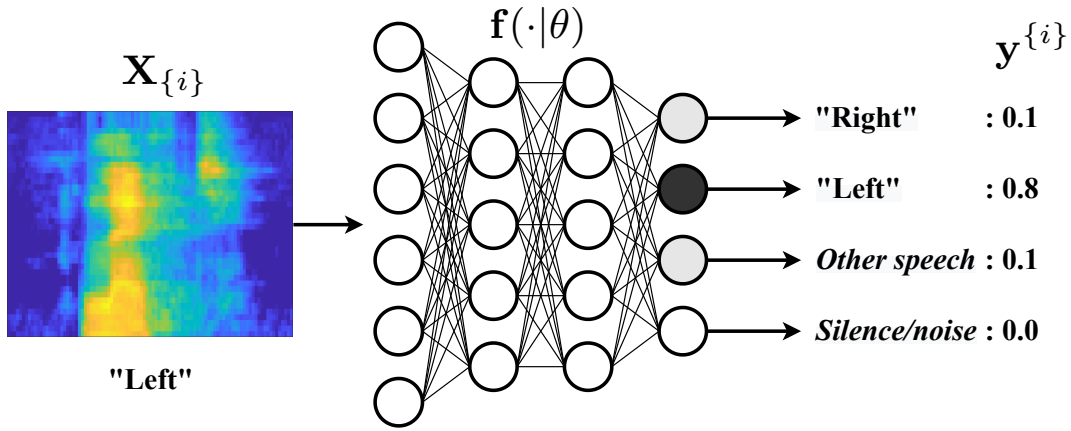
Deep learning-based acoustic modeling

- $\mathbf{f}\left(\cdot|\theta\right) : \mathbb{R}^{K\times(P+F+1)} \rightarrow [0,\ 1]^N$
- The N output nodes represent either words or subword units (e.g., context-independent phonemes)
- $\mathbf{y}_n^{\{i\}} = P\left(C_n|\mathbf{X}_{\{i\}}, \theta\right) = \mathbf{f}_n\left(\mathbf{X}_{\{i\}}|\theta\right),\ n = 1, ..., N$
- $\sum_{n=1}^{N} \mathbf{y}_n^{\{i\}} = 1,\ \forall i$ (fully-connected layer + softmax activation)

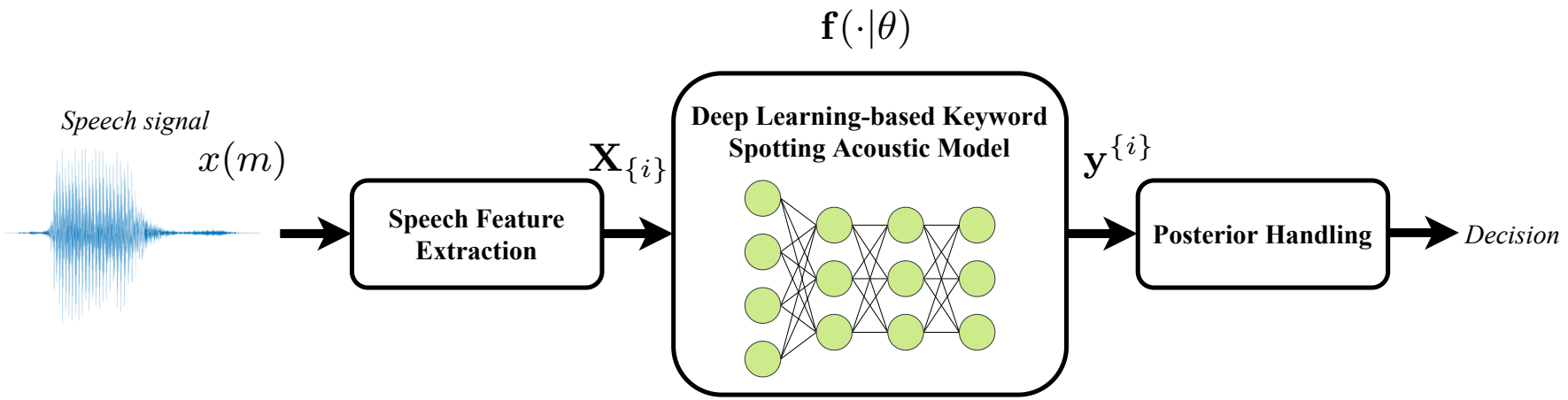
The General Pipeline



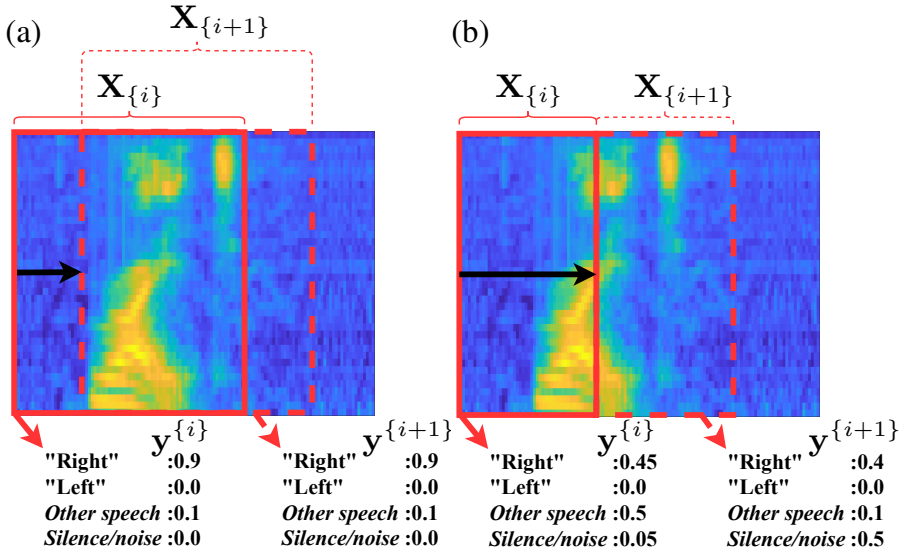
Deep learning-based acoustic modeling



The General Pipeline



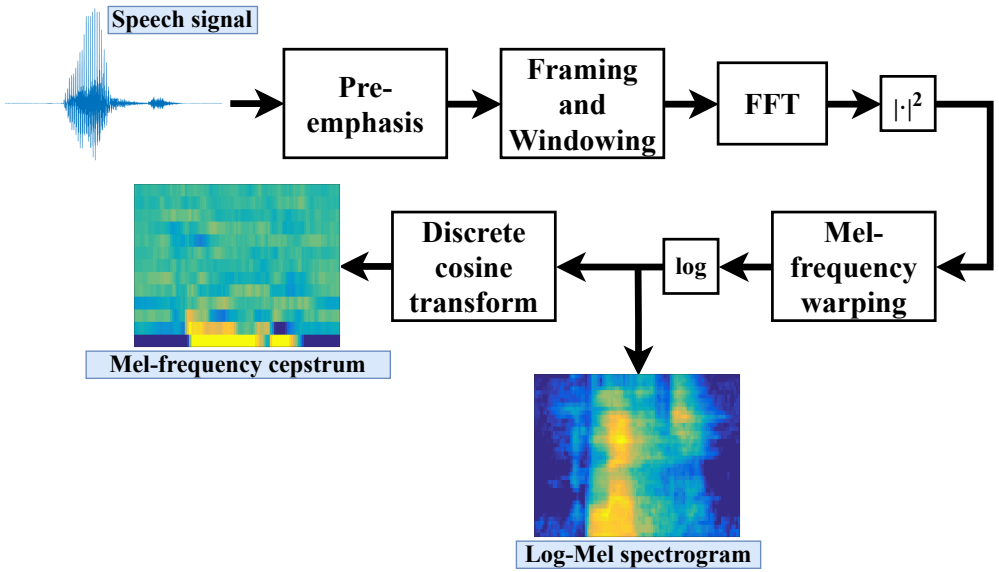
Posterior handling



- KWS is not a static task but a dynamic one
- $\hat{C}^{\{i\}} = \text{argmax}_{C_n} \mathbf{y}_n^{\{i\}} = \text{argmax}_{C_n} P\left(C_n|\mathbf{X}_{\{i\}}, \theta\right)$ (not the best way)

Speech Feature Extraction

Mel-Scale Related Features



- Log-Mel spectral coefficients and Mel-frequency cepstral coefficients (MFCCs) based on the perceptually-motivated Mel-scale filterbank
- A solid, competitive and safe choice $\longrightarrow$ Mel-scale-related features are, *by far*, the most widely used speech features in deep KWS
- Deep KWS performance is not significantly sensitive to the number of filterbank channels as long as the Mel-frequency resolution is not very poor (<10)

- A way to diminish the energy consumption and memory footprint of deep KWS systems consists of quantization of the acoustic model parameters
  - $\sim$ *performance of full-precision and 4-bit acoustic models*(1)
- The same philosophy can be applied to speech features
- Same performance by 8-bit (linearly-quantized) log-Mel spectra and full-precision MFCCs(2)
- Degradation is insignificant when exploiting 2-bit precision speech features(2)
- Much of the spectral information is superfluous when attempting to spot a set of keywords $\longrightarrow$ *Large room for future work on the design of new extremely-light and compact speech features for small-footprint KWS*

(1) Y. Mishchenko et al., "Low-bit quantization and quantization-aware training for small-footprint keyword spotting". In Proc. of ICMLA 2019
(2) A. Riviello and J. P. David, "Binary speech features for keyword spotting tasks". In Proc. of Interspeech 2019

- The development of E2E deep learning systems in which feature extraction is optimal in line with the task and training criterion is a recent trend
- **Optimal filterbank learning**
  - *SincConv: The acoustic model parameters are optimized jointly with the cut-off frequencies of a filterbank based on sinc-convolutions*(1)
  - *Filterbank matrix learning in the power spectral domain*(2)
  - *Parameter learning of a psychoacoustically-motivated gammachirp filterbank*(2)
- In (2), we found no statistically significant KWS accuracy differences between employing a learned filterbank and log-Mel features $\longrightarrow$ *Information redundancy?*

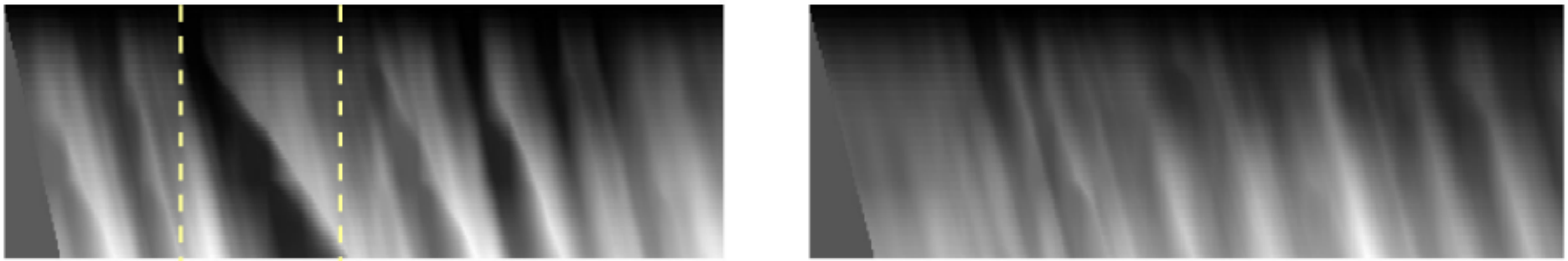
In conclusion

Handcrafted speech features currently provide state-of-the-art KWS performance at the same time that optimal feature learning requires further research to become the preferred alternative

(1) M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet". In Proc. of SLT 2018
(2) I. López-Espejo et al., "Exploring filterbank learning for keyword spotting". In Proc. of EUSIPCO 2021

- Multi-frame shifted time similarity (MFSTS): Time-domain two-dimensional speech representation comprised of constrained-lag autocorrelation values
  - *Simple but low performing*

- Fusion of dynamic time warping and deep KWS:

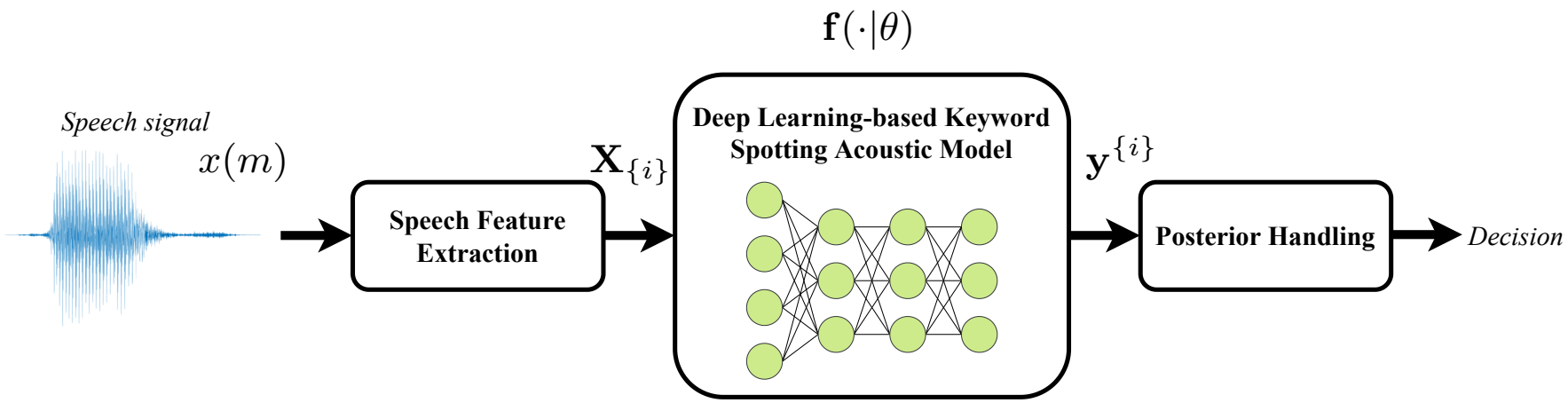


R. Shankar et al., "Spoken keyword detection using joint DTW-CNN". In Proc. of Interspeech 2018

☑ Open-vocabulary and language-independent scenarios

☒ It is prone to overfitting

Acoustic Modeling

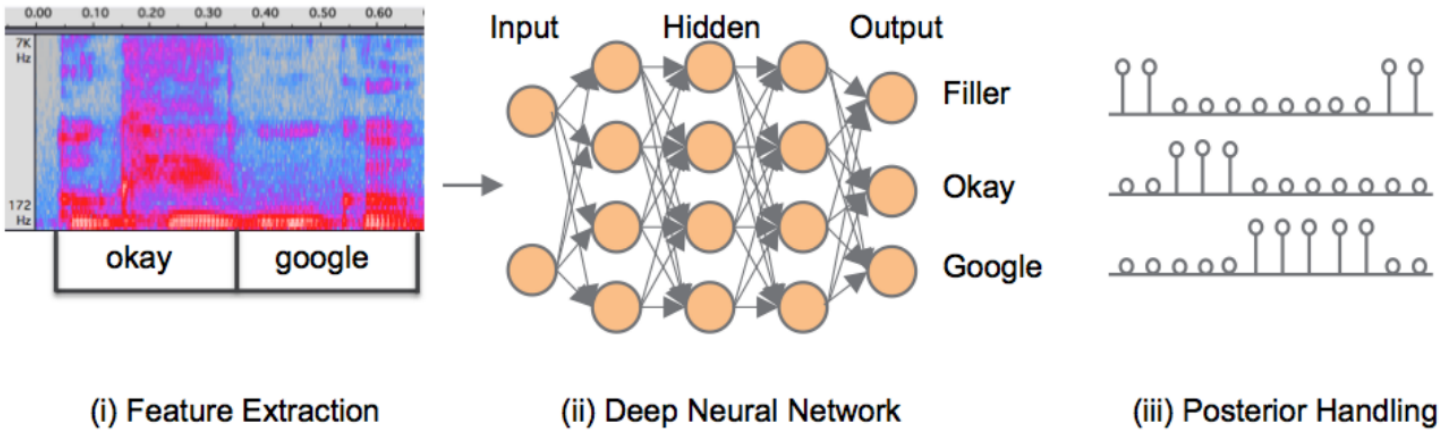


The natural trend is the design of increasingly accurate models while decreasing computational complexity

- Fully-connected feedforward neural networks
- **Convolutional neural networks**
- Recurrent and time-delay neural networks

Acoustic Modeling

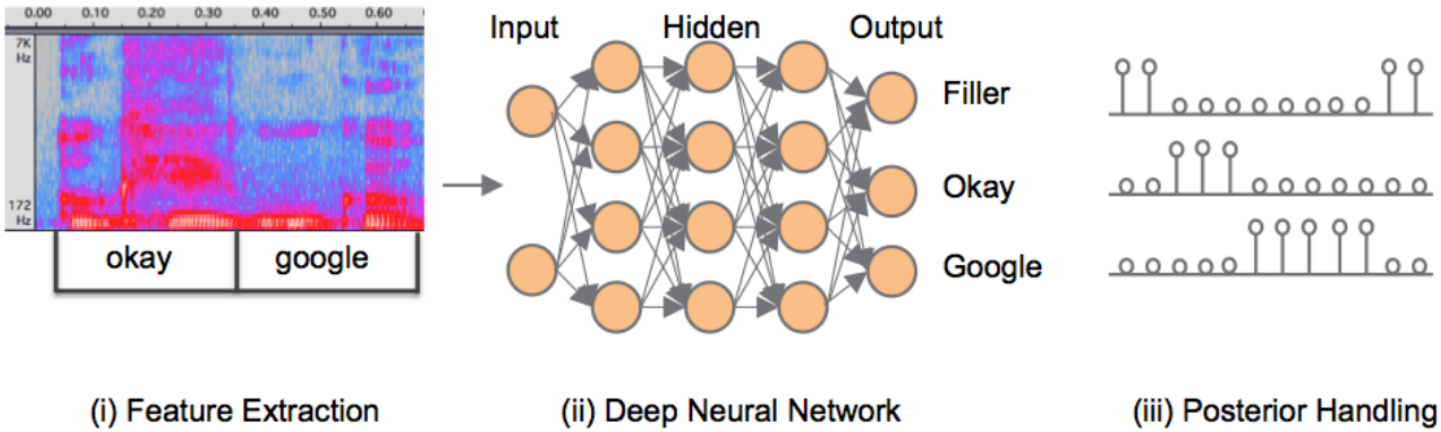
Fully-Connected Feedforward Neural Networks



G. Chen et al., "Small-footprint keyword spotting using deep neural networks". In Proc. of ICASSP 2014

- Three fully-connected layers with 128 neurons each
- Rectified linear unit (ReLU) activations
- Softmax output layer

☑ It outperforms, *with fewer parameters*, keyword/filler HMM in both clean and noisy conditions



G. Chen et al., "Small-footprint keyword spotting using deep neural networks". In Proc. of ICASSP 2014

- The use of fully-connected feedforward neural networks was quickly relegated to a secondary level

Nowadays

State-of-the-art acoustic models use convolutional and recurrent neural networks, since they can provide better performance with fewer parameters

There are closely related and computationally cheaper alternatives to fully-connected feedforward neural networks

1) Single value decomposition filter (SVDF) neural networks

- *They approximate fully-connected layers by low-rank approximations*
- *An SVDF neural network is a special case of a stacked one-dimensional CNN*

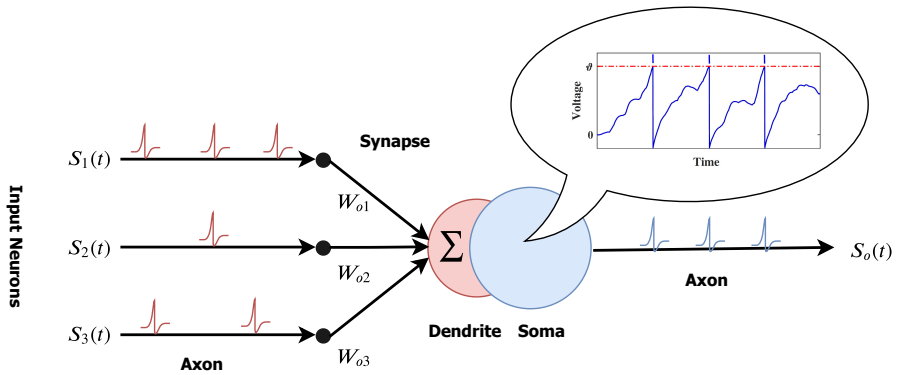
- SVDF achieved to reduce by 75% the acoustic model size of the first deep KWS system with no drop in performance(1)
- The performance of the first deep KWS system was improved by increasing the number of neurons while keeping the original number of multiplications(2)

(1) P. Nakkiran et al., "Compressing deep neural networks using a rank-constrained topology". In Proc. of Interspeech 2015
(2) G. Tucker et al., "Model compression applied to small-footprint keyword spotting". In Proc. of Interspeech 2016

There are closely related and computationally cheaper alternatives to fully-connected feedforward neural networks

2) Spiking neural networks (SNNs)

- *They are human brain-inspired and process the information in an event-driven manner*
- *The way information is processed alleviates the computational load when information is sparse*

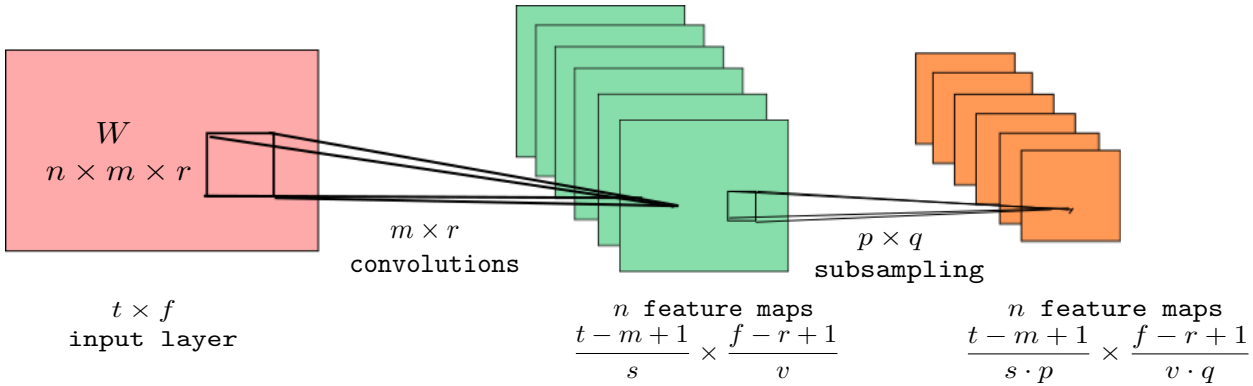


(1) E. Yılmaz et al., "Deep convolutional spiking neural networks for keyword spotting". In Proc. of Interspeech 2020

- Similar performance, >80% computational cost reduction(2), dozens of times energy saving(1)

(2) B. U. Pedroni et al., "Small-footprint spiking neural networks for power-efficient keyword spotting". In Proc. of BioCAS 2018

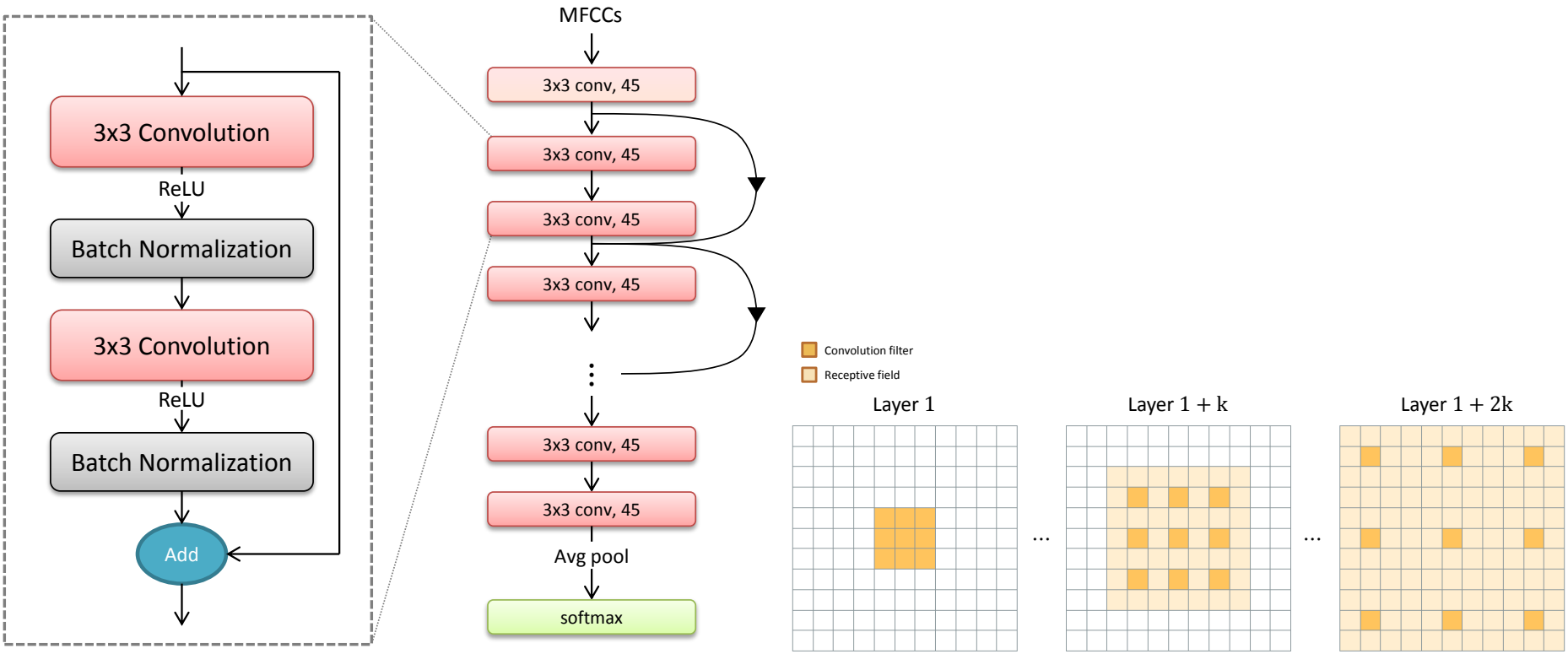
- From fully-connected feedforward to convolutional neural networks $\longrightarrow$ *A natural step taken back in 2015*(1)
  - *Exploitation of local speech time-frequency correlations, fewer parameters*



(1) T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting". In Proc. of Interspeech 2015

- The number of multiplications of the model can be easily limited to meet the computational constraints:
  - *Filter striding, kernel size, pooling size...*

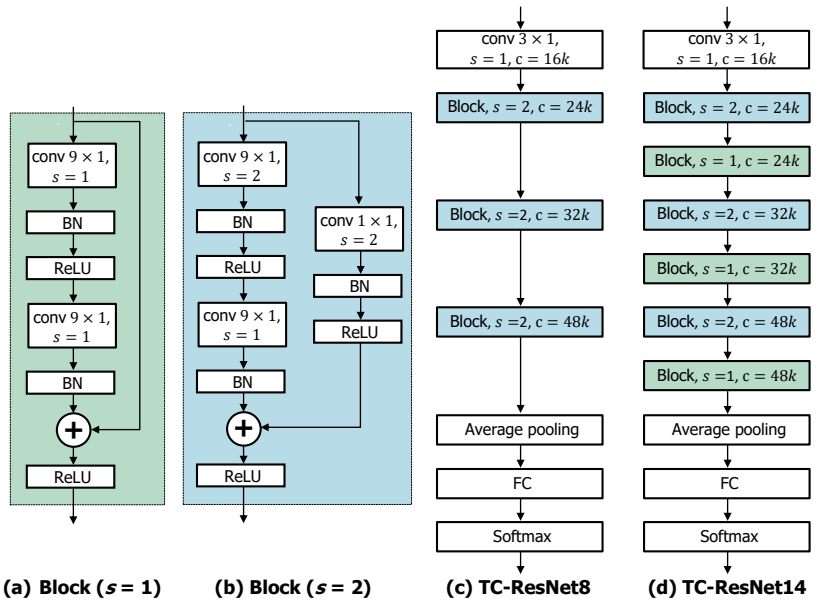
- **Residual learning** is widely considered to implement state-of-the-art acoustic models for deep KWS



(1) R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting". In Proc. of ICASSP 2018

- Tang and Lin established a new state-of-the-art back in 2018[(1)]

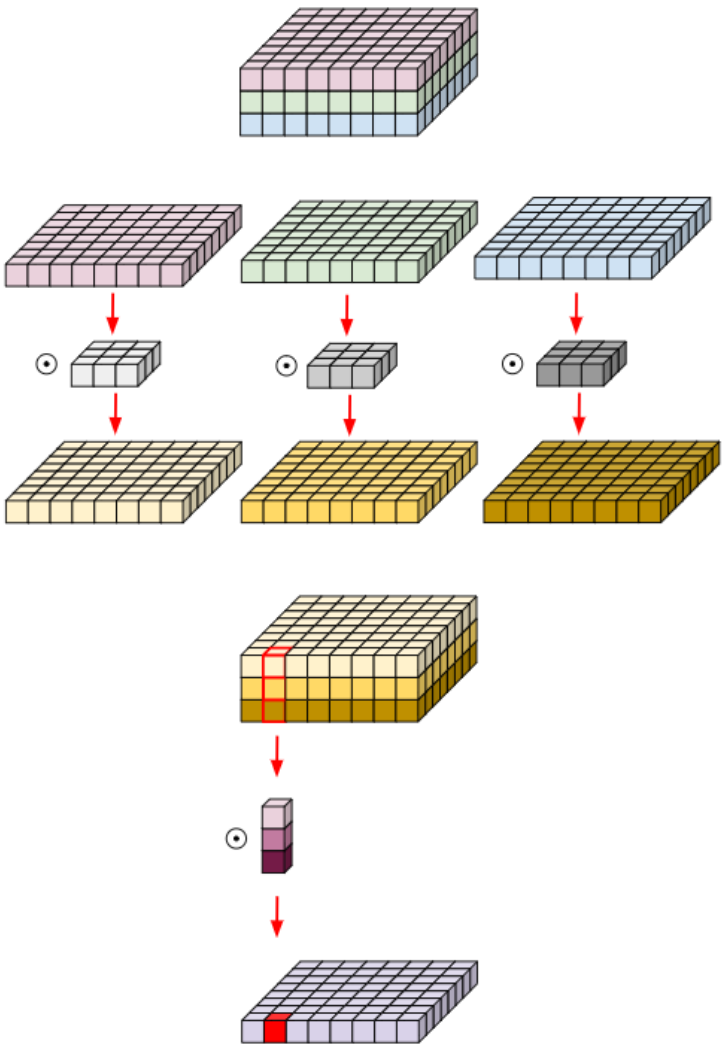
- **TC-ResNet**$^{(1)}$**:** One-dimensional convolutions along the time axis (temporal convolutions) while treating the (MFCC) features as input channels $\longrightarrow$ *Simultaneously capturing both high and low quefrency features*



(a) Block (*s* = 1) **(b) Block (*s* = 2)** **(c) TC-ResNet8** **(d) TC-ResNet14**

(1) S. Choi et al., "Temporal convolution for real-time keyword spotting on mobile devices". In Proc. of Interspeech 2019

- TC-ResNet matches Tang and Lin's KWS performance while dramatically decreasing both latency and the amount of FLOPs on a mobile device



E. Bendersky, "Depthwise separable convolutions for machine learning". 2018

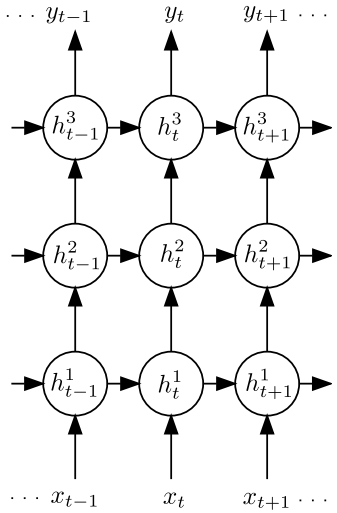
- **Depthwise separable (DS) convolutions** to reduce the computation and size of standard CNNs
- Reproducing the performance of TC-ResNet using less parameters(1)
- **Depthwise separable convolutions + residual learning:** It generally outperforms all standard residual networks, plain DS-CNNs and TC-ResNet with less computational complexity

(1) S. Mittermaier et al., "Small-footprint keyword spotting on raw audio data with sinc-convolutions". In Proc. of ICASSP 2020

- We believe that a modern CNN-based acoustic model should ideally encompass the following **three aspects**:

- ☑ A mechanism to exploit long time-frequency dependencies like, e.g., dilated convolutions
- ☑ Depthwise separable convolutions to substantially reduce both the memory footprint and computation of the model without sacrificing performance
- ☑ Residual connections to fast and effectively train deeper models providing enhanced KWS performance

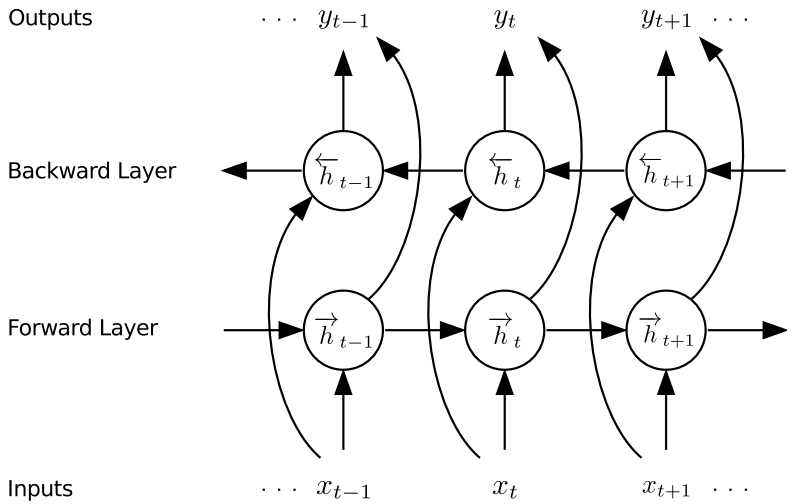
- Speech is a temporal sequence with strong time dependencies $\longrightarrow$ **Recurrent neural networks (RNNs)** and **time-delay neural networks (TDNNs)**
- Long short-term memory (LSTM) networks clearly outperform feedforward fully-connected neural networks for KWS acoustic modeling[(1)]



A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks". In Proc. of ICML 2014

(1) M. Sun et al., "Max-pooling loss training of long short-term memory networks for small-footprint keyword spotting". In Proc. of SLT 2016

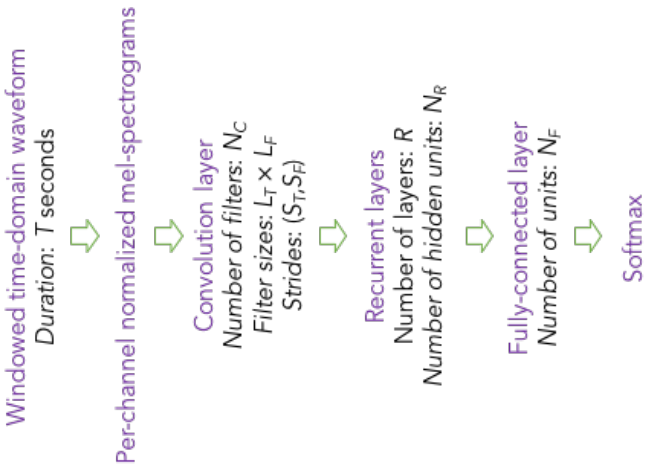
- When latency is not a strong constraint $\longrightarrow$ *Bidirectional RNNs* to capture causal and anticausal dependencies for improved KWS performance
- **Bidirectional LSTMs vs. bidirectional gated recurrent units (GRUs)**
  - *In KWS, there is no need to model very long time dependencies*
  - *GRUs demand less memory and are faster to train than LSTMs*
  - *GRUs perform similarly to or even better than LSTMs*[(1)]



A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks". In Proc. of ICML 2014

(1) S. O. Arik et al., "CRNNs for small-footprint keyword spotting". In Proc. of Interspeech 2017

- CNNs might have difficulties to model long time dependencies
- **Convolutional recurrent neural networks (CRNNs)** bring the best of two worlds:
  - **1)** *Convolutional layers model local spectro-temporal correlations of speech*
  - **2)** *Recurrent layers follow suit by modeling long-term time dependencies of speech*



☑ CRNNs generally outperform standalone CNNs and RNNs in KWS[(1)]

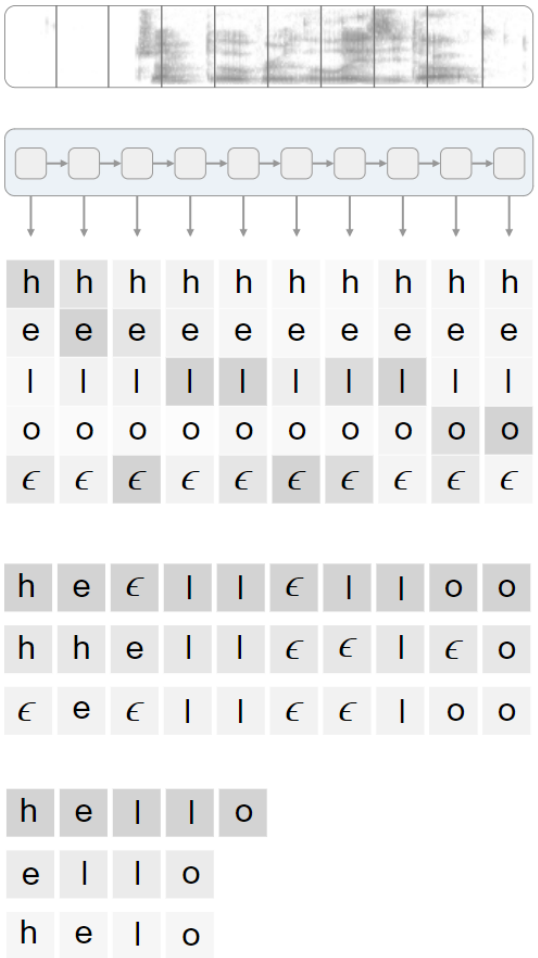
(1) M. Zeng and N. Xiao, "Effective combination of DenseNet and BiLSTM for keyword spotting". IEEE Access, 2019

- At training time, frame-level annotated data are typically required by, e.g., cross-entropy loss
- Frame-level annotated data may be cumbersome to get

For RNN acoustic modeling

Connectionist temporal classification (CTC) is an attractive alternative letting the model unsupervisedly locate and align the phonetic unit labels at training time![1]

(1) In other words, frame-level alignments of the target label sequences are not required for training



A. Hannun,
https://distill.pub/2017/ctc/

- Let $\mathbf{C} = (c_0, ..., c_{m-1})$ be the sequence of, e.g., characters corresponding to $\mathbf{X} = (\mathbf{x}_0, ..., \mathbf{x}_{T-1})$
- We ignore an accurate alignment between $\mathbf{C}$ and $\mathbf{X}$, and $m < T$
- CTC introduces the so-called blank token (ϵ)
- CTC is an *alignment-free* algorithm maximizing

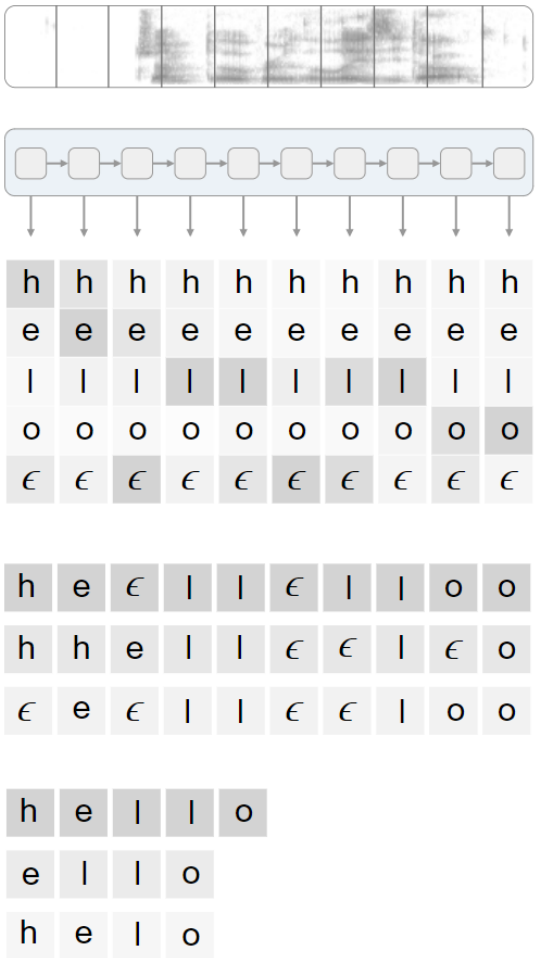
$$P(\mathbf{C}|\mathbf{X}) = \sum_{A \in \mathcal{A}_{X,C}} \prod_{t=0}^{T-1} P_t(\mathbf{c}|\mathbf{x}_0, ..., \mathbf{x}_t)$$

$$\text{e.g., } \mathbf{c} = \{h, e, l, o, \epsilon\}$$

- The acoustic model outputs can be understood as the probability distribution over all the possible label sequences given $\mathbf{X}$

Acoustic Modeling

Recurrent and Time-Delay Neural Networks



A. Hannun,
`https://distill.pub/2017/ctc/`

- **2007:** The very first attempt to apply CTC to KWS using a bidirectional LSTM(1)

  - *At training time, this system just needs the list of training words in order of occurrence in the speech signals*

- Different RNN **architectures** and **phonetic units** like phonemes and Mandarin syllables

- CTC KWS systems are superior to LVCSR and keyword/filler HMM with less or no additional computational cost

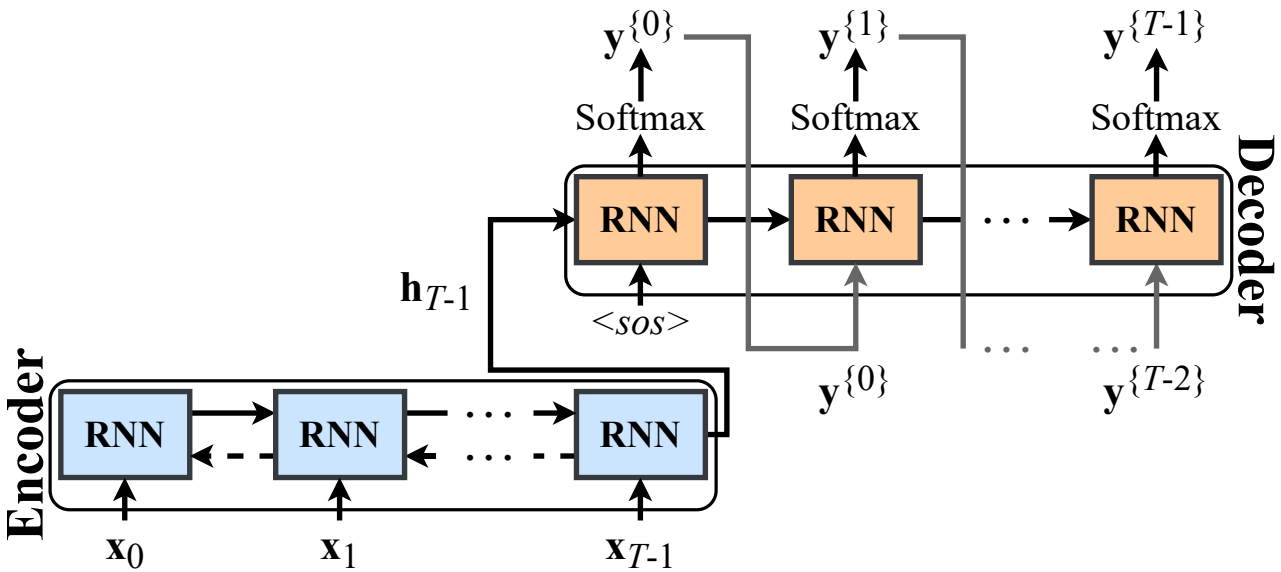
- ☑ CTC requires searching for the keyword phonetic unit sequence on a lattice $\longrightarrow$ Suitable for **open-vocabulary KWS**

(1) S. Fernández et al., "An application of recurrent neural networks to discriminative keyword spotting". In Proc of ICANN 2007

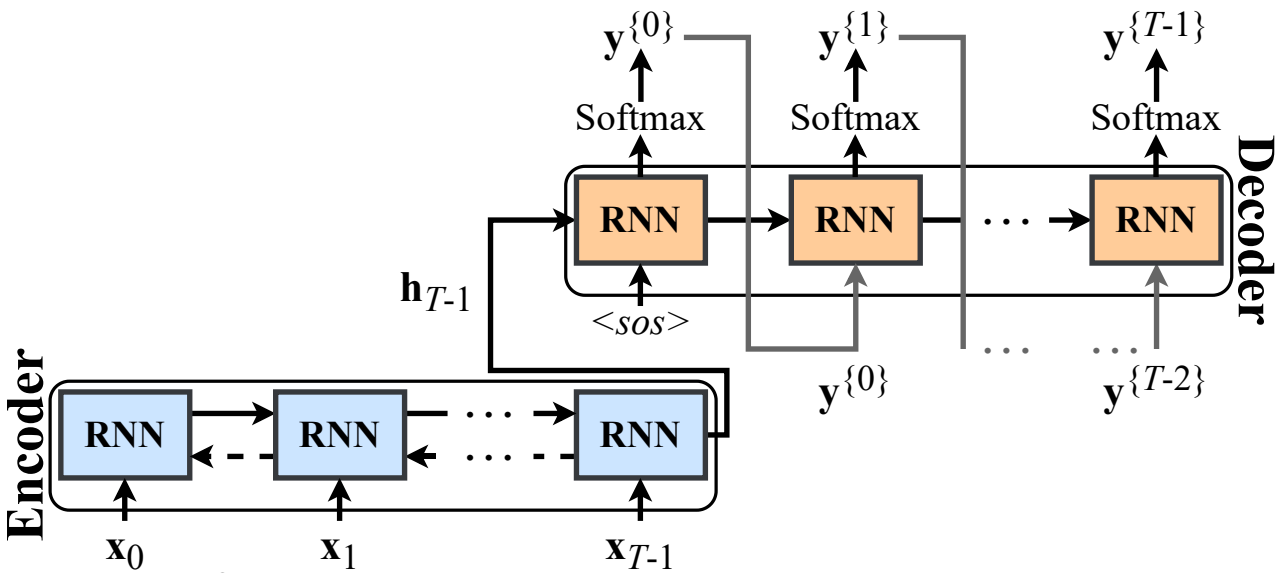
- CTC assumes conditional label independence, i.e., past model outputs do not influence current predictions:

$$P(\mathbf{C}|\mathbf{X}) = \sum_{A \in \mathcal{A}_{X,C}} \prod_{t=0}^{T-1} P_t(\mathbf{c}|\mathbf{x}_0, \ldots, \mathbf{x}_t)$$

- CTC may need an *external language model* to perform well
- **Seq2Seq:** A more convenient approach for KWS acoustic modeling

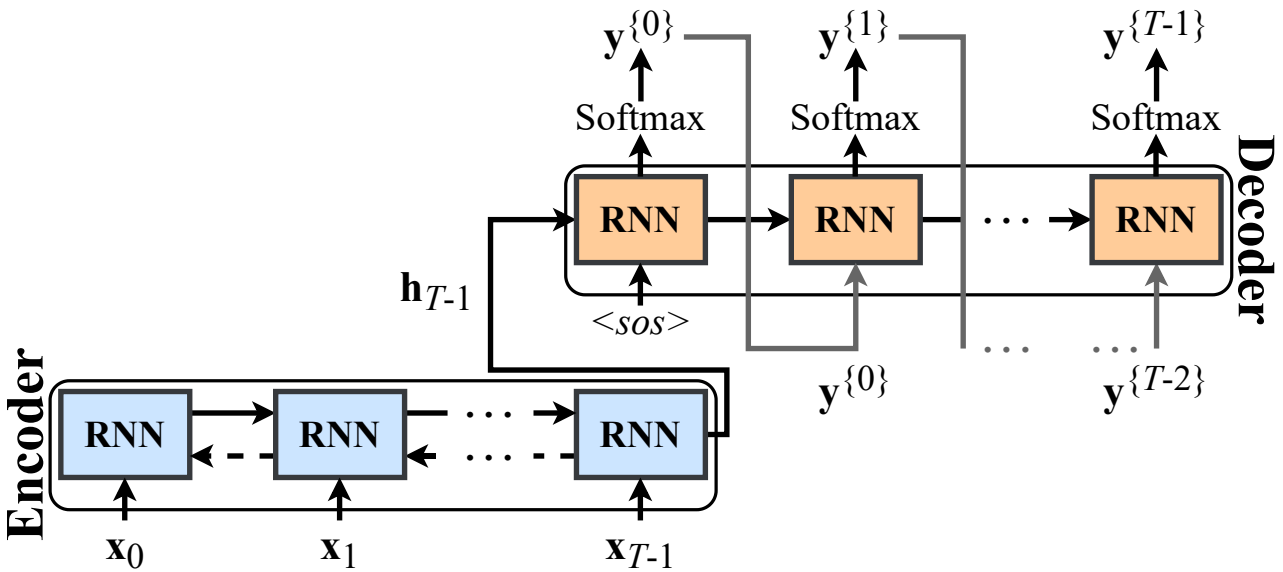


- **RNN-Transducer**, integrating both *acoustic and language models* (and predicting phonemes), is able to outperform a CTC KWS system even when the latter exploits an external phoneme N-gram language model(1)



(1) Y. He et al., "Streaming small-footprint keyword spotting using sequence-to-sequence models". In Proc. of ASRU 2017

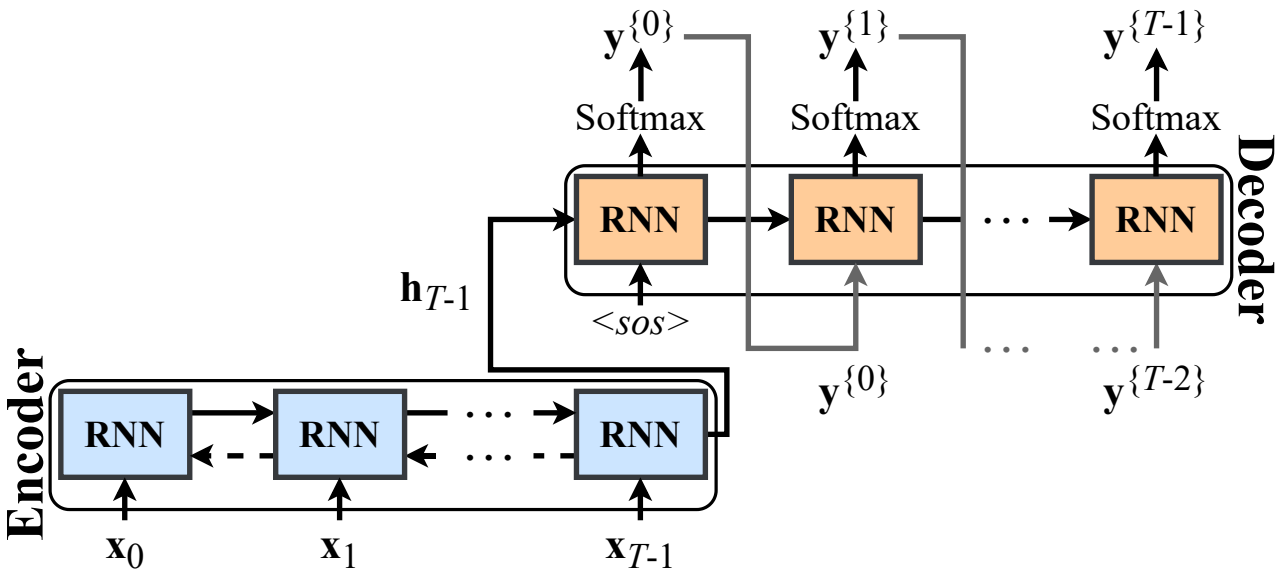
- The encoder has to condense all the needed information into a fixed-dimensional vector regardless the (*variable*) length of the input sequence
- The **attention mechanism** might assist by focusing on the speech sections that are more likely to comprise a keyword



- $\mathbf{h}_t = \text{Encoder}\left(\mathbf{x}_t, \mathbf{h}_{t-1}\right)$
- To assist the decoder, a context-relevant subset of $\{\mathbf{h}_0, ..., \mathbf{h}_{T-1}\}$ can be *attended* to yield $\mathbf{A}$ (*to be used instead of* $\mathbf{h}_{T-1}$):

$$\mathbf{A} = \sum_{t=0}^{T-1} \alpha_t \mathbf{h}_t$$

$$\alpha_t = \text{Attend}\left(\mathbf{h}_t\right) \qquad \sum_t \alpha_t = 1$$

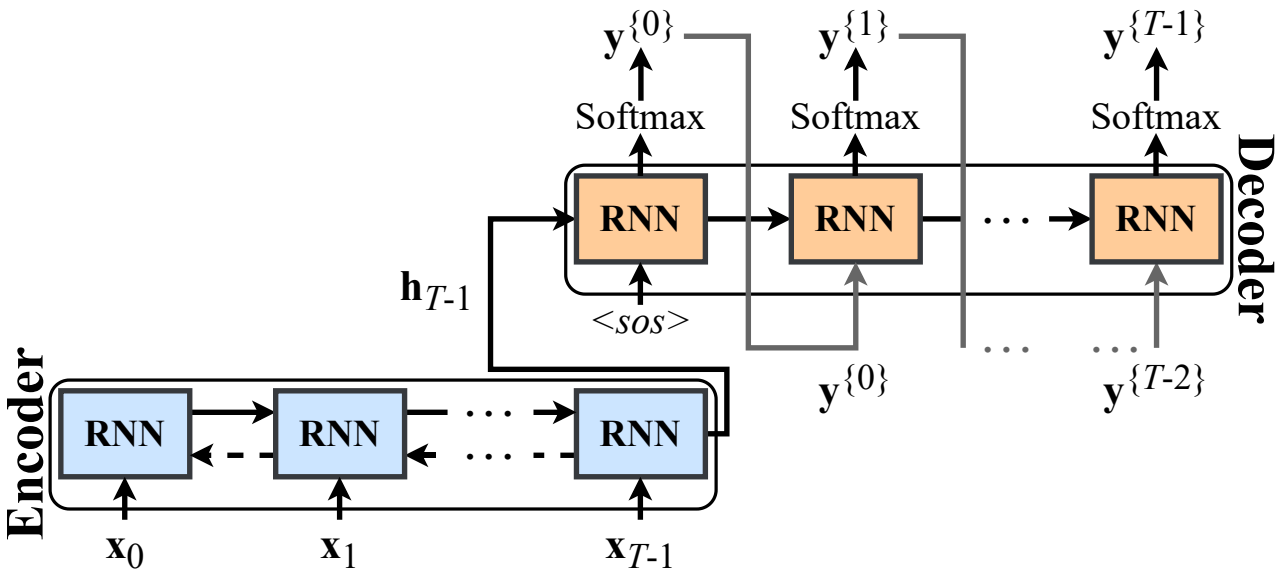


- Different works find that incorporating attention provides KWS performance gains with respect to counterpart Seq2Seq models without attention[(1−3)]

(1) D. C. de Andrade et al., "A neural attention model for speech command recognition". arXiv:1808.08929v1, 2018

(2) Z. Zhao and W.-Q. Zhang, "End-to-end keyword search based on attention and energy scorer for low resource languages". In Proc. of Interspeech 2020

(3) Z. Liu et al., "RNN-T based open-vocabulary keyword spotting in Mandarin with multi-level detection". In Proc. of ICASSP 2021



- Apart from CTC, **cross-entropy loss** is, *by far*, the most popular loss function for training deep spoken KWS acoustic models:

  - $l_n^{\{i\}}$: *Binary true (training) label corresponding to the input feature segment* $\mathbf{X}_{\{i\}}$

$$\mathcal{L}_{\text{CE}} = -\sum_{i}\sum_{n=1}^{N} l_n^{\{i\}} \log\left(\mathbf{y}_n^{\{i\}}\right)$$

- *Subword-level posteriors*: Training labels are generated by force alignment using an LVCSR system (*which will condition the subsequent KWS system performance*)

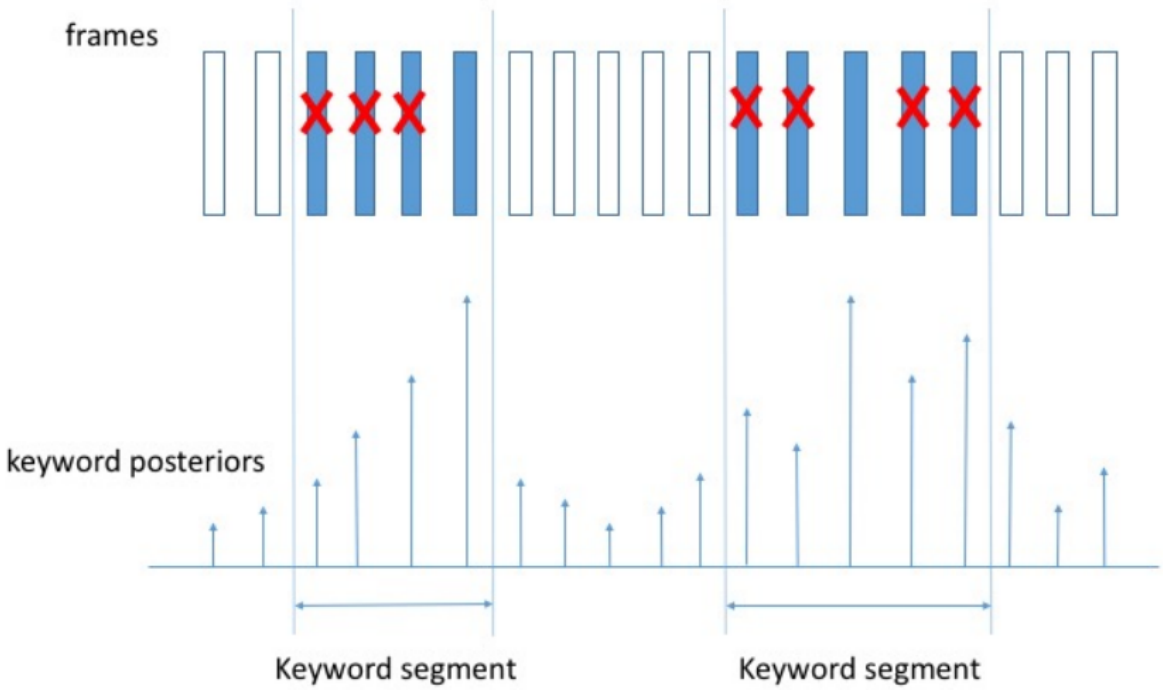
- **Max-pooling loss:** Teaching the acoustic model to only trigger at the highest confidence time near the end of the keyword:

  - $\hat{\mathbf{L}}$: *Set of all indices of the input feature segments in a minibatch belonging to any non-keyword class*
  - y_p^*: *Largest target posterior corresponding to the p-th keyword sample in the minibatch ($p = 1, ..., P$, and P is the total number of keyword samples in the minibatch)*

$$\mathcal{L}_{\text{MP}} = -\sum_{i \in \hat{\mathbf{L}}} \sum_{n=1}^{N} l_n^{\{i\}} \log\left(\mathbf{y}_n^{\{i\}}\right) - \sum_{p=1}^{P} \log\left(y_p^*\right)$$

- Max-pooling is superior to cross-entropy loss (especially when the acoustic model is initialized by cross-entropy loss training)[1]

(1) M. Sun et al., "Max-pooling loss training of long short-term memory networks for small-footprint keyword spotting". In Proc. of SLT 2016



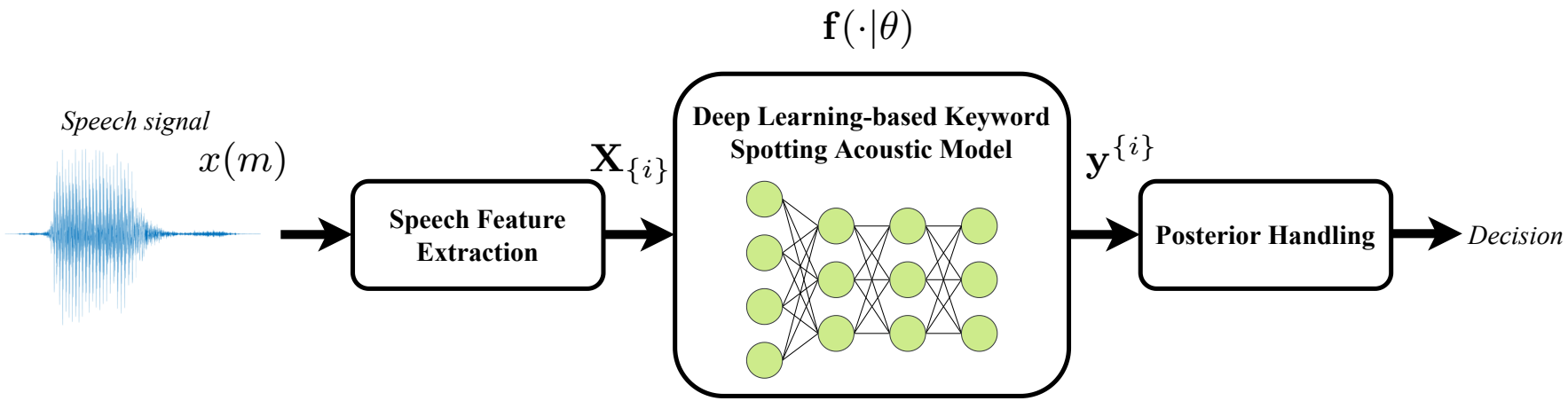
M. Sun et al., "Max-pooling loss training of long short-term memory networks for small-footprint keyword spotting". In Proc. of SLT 2016

- *Variants:* Weakly-constrained max-pooling, smoothed max-pooling...

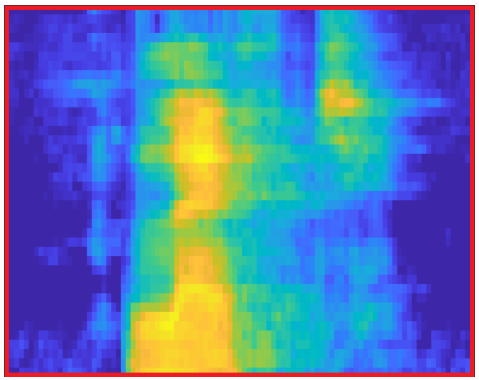
- Stochastic gradient descent (normally with momentum) and Adam
- Learning rate decay
- *Parameter regularization:* Weight decay, dropout...
- Initialization based on transfer learning from LVCSR acoustic models leads to better KWS models by, e.g., alleviating overfitting[(1)]

(1) Y. Tian et al., "Improving RNN transducer modeling for small-footprint keyword spotting". In Proc. of ICASSP 2021

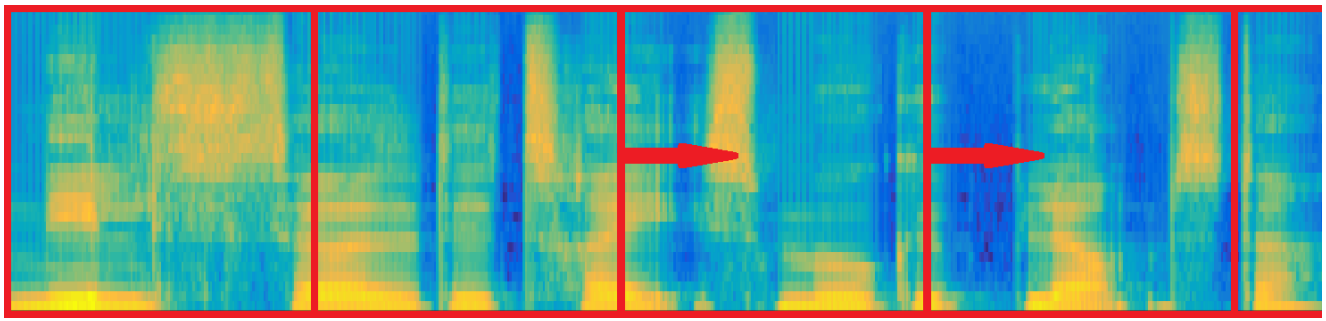
Posterior Handling



Non-streaming/static



Streaming/dynamic



- **Non-streaming mode:** Isolated word classification
- Since non-streaming deep KWS systems tend to produce very peaked posterior distributions (*no inter-class transition data*),

$$\hat{C}^{\{i\}} = \text{argmax}_{C_n}\, \mathbf{y}_n^{\{i\}} = \text{argmax}_{C_n}\, P\left(C_n|\mathbf{X}_{\{i\}}, \theta\right)$$

- Lack of realism from a practical point of view
- Non-streaming performance and streaming performance seem to be highly correlated, which makes non-streaming KWS research more relevant than it might look at first sight

- **Streaming mode:** Continuous processing (normally in real-time) of an input audio stream in which keywords are not isolated/segmented

$$\mathcal{Y} = \left\{..., \mathbf{y}^{\{i-1\}}, \mathbf{y}^{\{i\}}, \mathbf{y}^{\{i+1\}}, ...\right\}$$

- $\mathcal{Y}$ has strong local correlations
- $\mathcal{Y}$, which is inherently noisy, is typically smoothed over time on a class basis before further processing:

$$\mathcal{Y} \longrightarrow \text{Smoothing} \longrightarrow \bar{\mathcal{Y}} = \left\{..., \bar{\mathbf{y}}^{\{i-1\}}, \bar{\mathbf{y}}^{\{i\}}, \bar{\mathbf{y}}^{\{i+1\}}, ...\right\}$$

Case 1: Each of the N classes of a deep KWS system represents a whole word

- *1)* $\bar{\mathbf{y}}^{\{i\}}$ can be compared with a sensitivity threshold

OR

- *2)* The class with the highest posterior within a time sliding window can be picked from $\bar{\mathbf{y}}^{\{i\}}$

- Consecutive input segments $\left\{..., \mathbf{X}_{\{i-1\}}, \mathbf{X}_{\{i\}}, \mathbf{X}_{\{i+1\}}, ...\right\}$ may cover fragments of the same keyword realization $\longrightarrow$ *False alarms!*
- To prevent false alarms, a simple, yet effective mechanism consists of forcing the KWS system not to trigger for a short period of time right after a keyword has been spotted

Case 2: Each of the N classes still represents a whole word but keywords are composed of multiple words (e.g., "OK Google") **OR** each of the N classes represents a subword unit (e.g., a syllable) instead of a whole word

- Let us assume that the first class C_1 corresponds to the non-keyword class and that the remaining $N-1$ classes represent subunits of a **single** keyword(1):

$$S_c^{\{i\}} = \sqrt[N-1]{\prod_{n=2}^{N} \max_{h_{\max}(i) \leq k \leq i} \bar{\mathbf{y}}_n^{\{k\}}}$$

- A keyword is detected every time $S_c^{\{i\}}$ exceeds a sensitivity threshold to be tuned

(1) G. Chen et al., "Small-footprint keyword spotting using deep neural networks". In Proc. of ICASSP 2014

Case 2: Each of the N classes still represents a whole word but keywords are composed of multiple words (e.g., "OK Google") **OR** each of the N classes represents a subword unit (e.g., a syllable) instead of a whole word

- Let us assume that the first class C_1 corresponds to the non-keyword class and that the remaining $N-1$ classes represent subunits of a **single** keyword:

$$S_c^{\{i\}} = \sqrt[N-1]{\prod_{n=2}^{N} \max_{h_{\max}(i) \leq k \leq i} \bar{\mathbf{y}}_n^{\{k\}}}$$

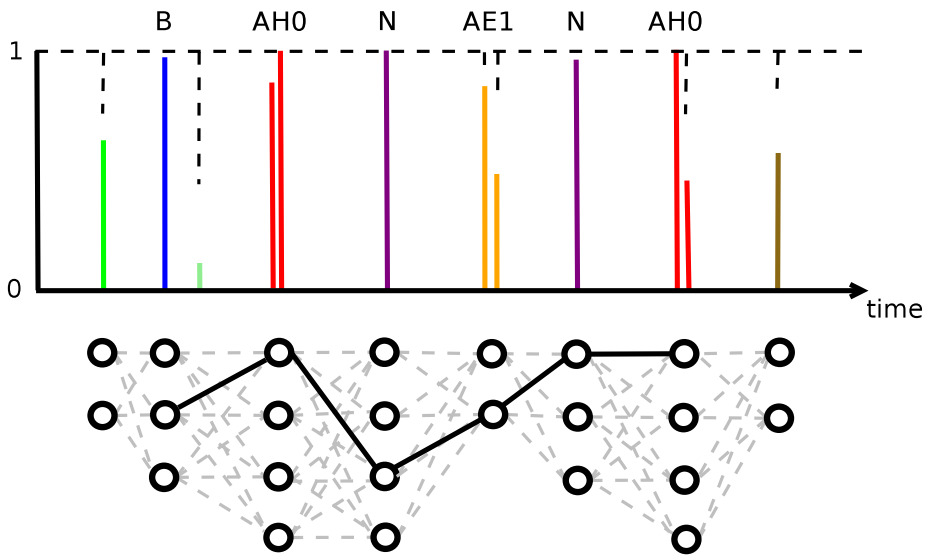
Decreasing false alarms

The above equation can be subject to the constraint that the keyword subunits trigger in the correct order of occurrence within the keyword(1)

(1) R. Prabhavalkar et al., "Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks". In Proc. of ICASSP 2015

Typically done in the context of CTC:

- From a posterior lattice, the goal is to find the most similar subword unit sequence to that of the target keyword
- A keyword is spotted if the score upon the search on the lattice is greater than a threshold

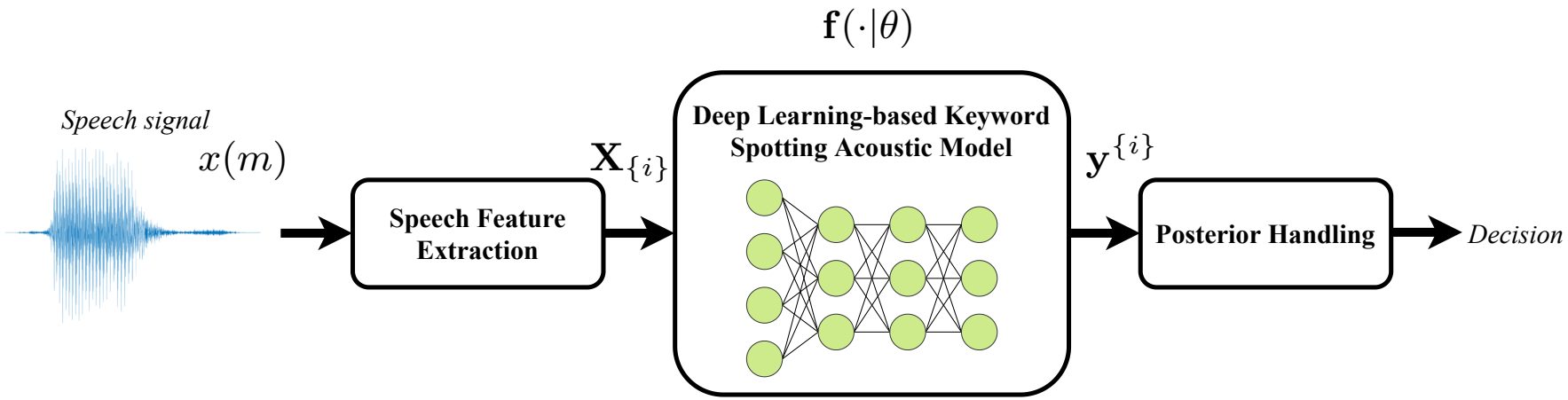


Y. Zhuang et al., "Unrestricted vocabulary keyword spotting using LSTM-CTC". In Proc. of Interspeech 2016

☑ Keyword personalization

☒ Higher computational complexity

Recap


$\mathbf{f}(\cdot|\theta)$
Speech signal
$x(m)$
Speech Feature Extraction
$\mathbf{X}_{\{i\}}$
Deep Learning-based Keyword Spotting Acoustic Model
$\mathbf{y}^{\{i\}}$
Posterior Handling
Decision

DEEP SPOKEN KEYWORD SPOTTING

2. The Deep Spoken Keyword Spotting Approach

Iván López-Espejo

Department of Electronic Systems, Aalborg University, Denmark

ivl@es.aau.dk

Sunday 18th September, 2022

INTERSPEECH 2022
September 18 - 22 • Incheon Korea

3. Robustness in Keyword Spotting: Deep Spoken KWS

John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
Department of Electrical Engineering
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.
(John.Hansen@utdallas.edu)

INTERSPEECH-2022 Tutorial

Sept. 18 – 22, 2022 Incheon, Korea

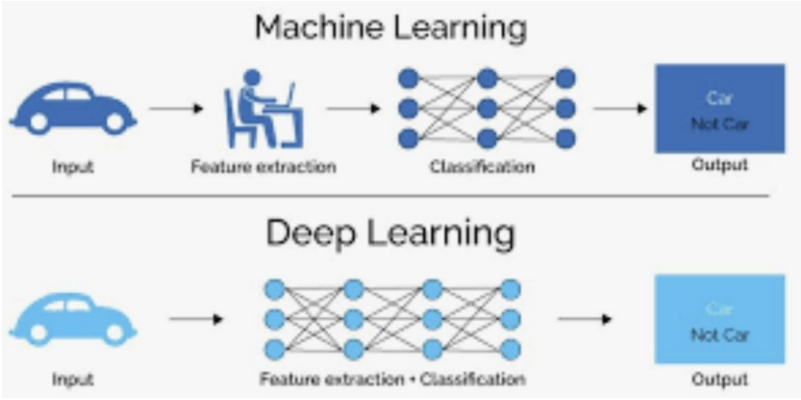
1. Outline

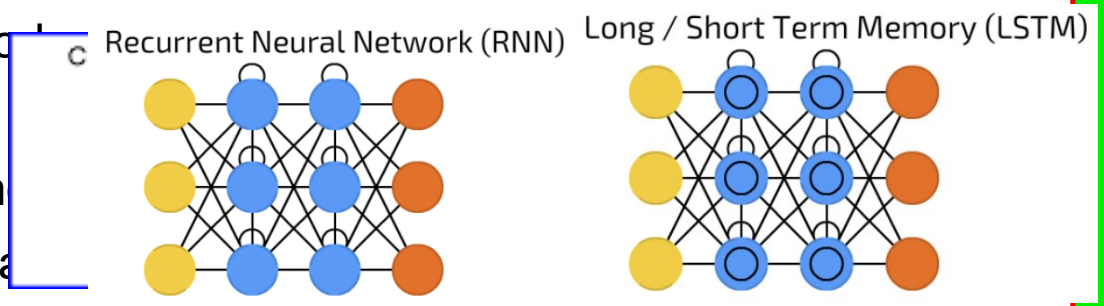
- **Introduction**: Robustness, Machine Learning & Data – Challenges
- **Speech Data**: Mismatch (Intrinsic, Extrinsic, and Context based issues)
- **Speech Task ML Challenges**: (*know your problem*, *know your data*)
- **KWS – Naturalistic Spaces; Spontaneous Speech, Distance Capture**
- **Example Studies**
  - Building ML Models: Dialect ID "Is the secret in the silence?"
  - Conversational Analysis: Prof-Life-Log: Word Count Estimation, KWS
  - KWS: Various Speech Corpora (Clean – Noisy – Naturalistic)
  - KWS: DARPA RATS example
  - KWS: Naturalistic Learning Spaces
- **Summary**

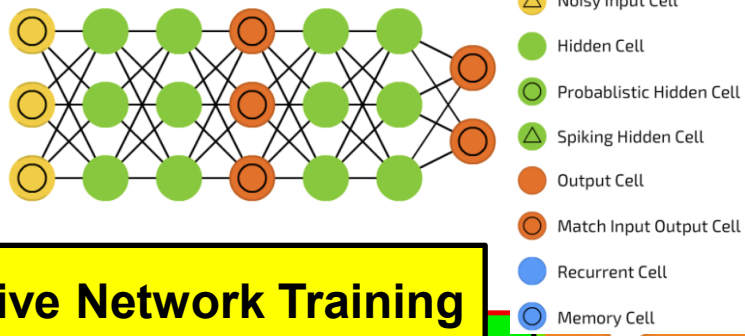
[1] J.H.L. Hansen, H. Boril, "On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks," Speech Communication, vol. 101, pp. 94-108, July 2018.
[2] M. Mirsamadi, J.H.L. Hansen, "Multi-domain adversarial training of neural network acoustic models for distant speech recognition," Speech Communication, vol. 106, pp. 21-30, Jan. 2020
[3] I. López-Espejo, Z.-H. Tan, J.H.L. Hansen, J. Jensen, "Deep Spoken Keyword Spotting: An Overview," IEEE Access, vol, 10, pp.4169 - 4199, 2022.
[4] J.H.L. Hansen, T. Hasan, "Speaker Recognition by Machines and Humans: A Tutorial Review," IEEE Signal Processing Magazine, pp. 74-99, Nov. 2015.

1. Machine Learning & Data

- **Machine Learning vs. Deep Learning**
- **Data** drives models and solutions for ML/Deep Learning
- **Architectures include:**
  - CNN: convolutional neural network
  - DNN: deep neural network
  - RNN: recurrent neural netwo
  - AE : auto-encoder
  - LSTM: long-short term mem
  - GANs: generative adversaria
- **Speech Technology & ML/Deep Learnin**
  - ASR/KWS, SID, Diarization, LID/DID/A (Language, dialect, accent), Emotion / Stress Recognition, Conversational Analysis, etc.







Speech Data NEEDED for effective Network Training

1. Speech Communications

- Historically, speech communication and engineering design for electronic communications has focused on 1-on-1 scenario

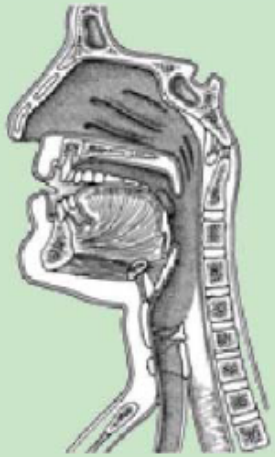
- Most human-to-human speech studies, have also focused on 1-on-1 context.

vs.

Future Speech Communication Research Directions:
(1) Naturalistic Data; (2) Multi-Speaker Context; (3) Massive "Big" Data;
(4) STEM / Team-Based communications; (5) Voice Enabled & Distant

1. Speech Data:
Mismatch is Everywhere, and Growing

INTRINSIC
Speaker Based



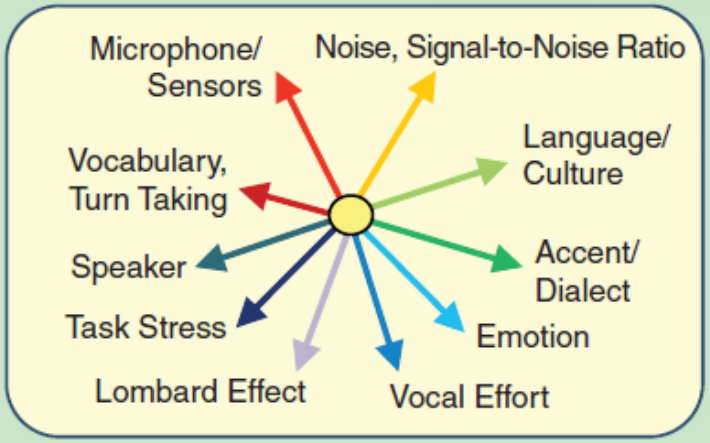
EXTRINSIC
Technology Based

CONTEXT
Conversation Based

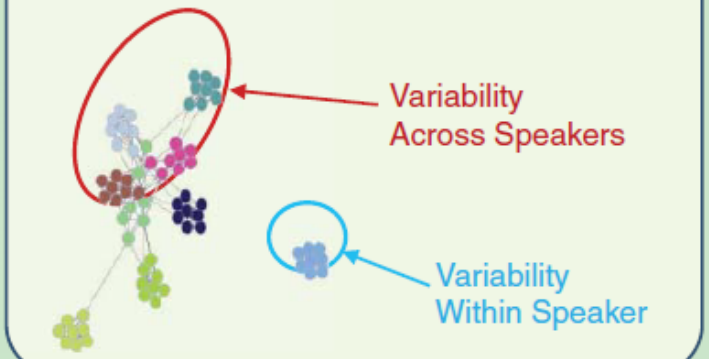
Human-to-Human
Human-to-Machine

Prompted/Read Speech
Spontaneous Speech

Monologue
Two-Way Conversation
Group Discussion

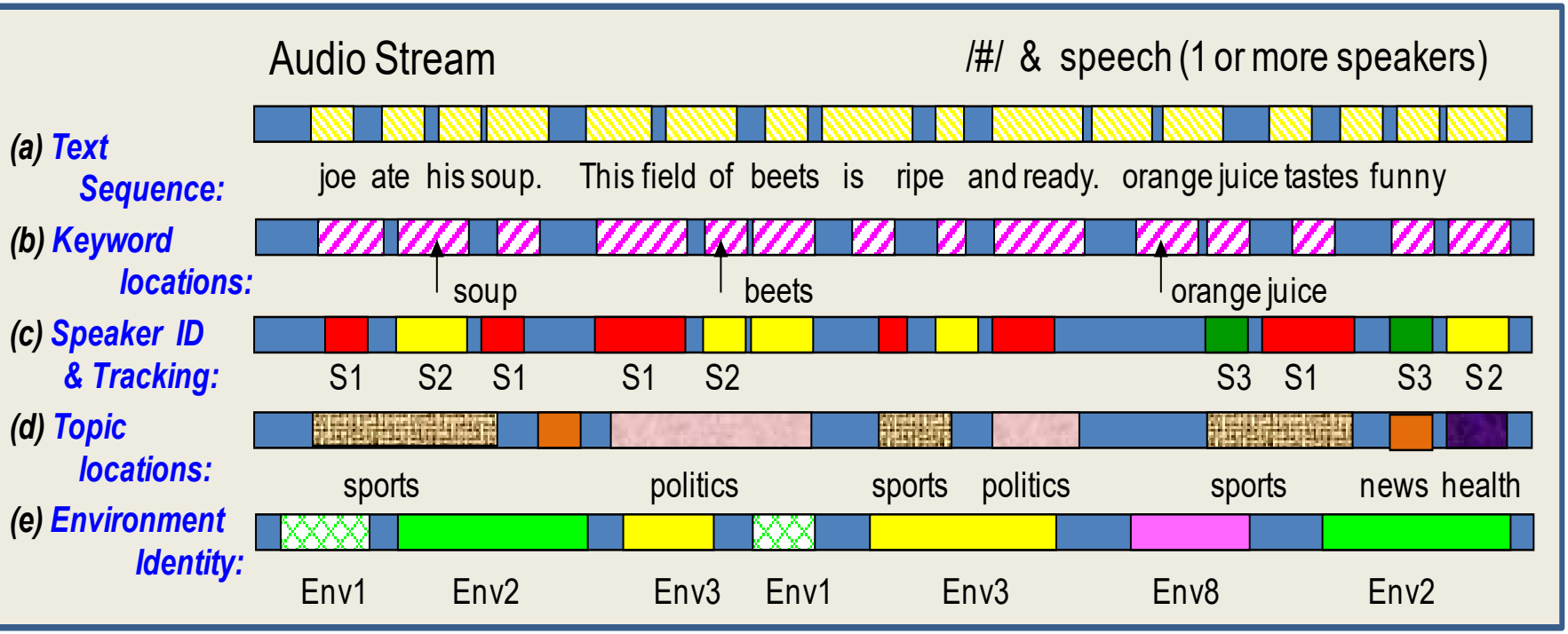


Speech Utterance Space (Two Dimensional)



1. Audio Diarization: Naturalistic Data

Diarization can be MUCH Richer for Knowledge Extraction



- Speech Recognition – challenges in spontaneous conversational speech (not "prompted" like Apple SIRI,etc), coarticulation issues, group/overlap, etc.
- ML/Deep Learning Models: SAD, ASR, KWS, SID, LID/DID/AID, Emotion/Stress/Sentiment recognition, Conversational Analysis, etc.

2. Naturalistic Data Streams

- **Challenges in ML Speech Tasks** (know your problem, know your data)
- Speech Data – know context, speakers, scenarios for data collection
- Obtain as much "Meta-Data" as possible for corresponding audio
- ML training methods explore: (a) Data Augmentation (e.g., expanding data by adding noise/reverb/distortion to "clean" data); (b) use Meta-Data for tiered training (e.g., curriculum training, "Student-Teacher" modeling, etc.)
- RISKY to use "found data" in a blind manner!!!
  - Building ML Models on Found Data: Dialect ID "Is the secret in the silence?"


Speech Communication 101 (2018) 94–108

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks☆

John H.L. Hansen[*,a], Hynek Bořil[b,b]

[a] Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering, University of Texas at Dallas, Richardson, TX, USA
[b] Pioneer Speech Signal Processing Laboratory (PSSPL), Electrical and Computer Engineering Department, University of Wisconsin-Platteville, USA

ARTICLE INFO

Keywords:
Realism
Speech recognition
Human-computer interaction
Computational paralinguistics

ABSTRACT

Recent years have witnessed notable advancements in the areas of speech, speaker and language/dialect recognition. However, many of the emerging scientific principles appear to be drifting to the sidelines with the assumption that access to larger amounts of data is all that is required to address a growing range of issues relating to new scenarios. This study surveys several challenging domains in formulating effective solutions in realistic speech data, and in particular the notion of using naturalistic data to better reflect the potential effectiveness of new algorithms. Our main focus is on intra-speaker mismatch and speech variability issues due to (i) differences in noisy speech with and without Lombard effect and a communication factor, (ii) realistic field data in noisy and increased cognitive load conditions, (iii) speech variability introduced by whispered speech, and (iv) dialect identification using found data. Finally, we study speaker–environment and speaker–speaker interactions in a newly established, fully naturalistic Prof-Life-Log corpus. The specific outcomes from this study include an analysis of the strengths and weaknesses of simulated vs. actual speech data collection for research.

- J.H.L. Hansen, H. Boril, "On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks," Speech Communication, vol. 101, pp. 94-108, July 2018.
- H. Boril, A. Sangwan, J.H.L. Hansen, "Arabic Dialect Identification - 'Is the Secret in the Silence?' and Other Observations," ISCA Interspeech-2012, Portland, OR, Sept. 9-13, 2012

2.1 Data Example: Dialect ID

Building ML Models: Dialect ID "Is the secret in the silence?"

Linguistic Data Consortium (LDC) Corpora

- Conversational telephone speech (CTS)
- 4 dialects – Gulf, Iraqi, Egyptian, Levantine
- LDC sets: (1) Gulf Arabic CTS; (2) Iraqi Arabic CTS;
  (3) CALLHOME & CALLFRIEND Egyptian Arabic Speech
  (4) Arabic CTS Levantine Fisher Training Data Set 3
- Past studies [1–3] used Levantine Arabic CTS instead of Fisher corpus

[1] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling," in *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Athens, Greece, 2009.

[2] F. Biadsy, J. Hirschberg, and D. P. W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," in *INTERSPEECH'11*, Florence, Italy, 2011.

[3] M. Akbacak, D. Vergyri, A. Stolcke, N. Scheffer, and A. Mandal, "Effective Arabic dialect classification using diverse phonotactic models," *INTERSPEECH'11*, Florence, Italy, 2011.

2.2 Arabic Corpora for DID

In-House Pan-Arabic Corpus [4] (UTD-CRSS)

- Conversational speech; lapel microphone (hands-free)
- 5 dialects – United Arab Emirates, Egyptian, Iraqi, Palestinian, Syrian
- 100 speakers per dialect (gender balanced)
- Each session – 2 speakers, 4 combined conversational recordings
- 4 dialects (blue) used in current experiments

	LDC Corpora				In-House Pan-Arabic			
	GLF	IRQ	LEV	EGY	PS	IRQ	SY	EGY
Train Set (Hrs)	32.7	16.1	11.9	33.9	10.6	9.3	10.8	9.9
Test Set (Hrs)	2.0	2.3	1.6	10.1	2.8	2.7	2.5	2.6
Avg. Chunk Length	11.3 sec				11.9 sec			

[4] Y. Lei and J. H. L. Hansen, "Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese," *IEEE Trans. on Audio, Speech, Lang. Proc.* 19:1, pp. 85 –96, Jan. 2011.

2.2 Speech – Based DID

DID on LDC Corpora – Speech Chunks

- Initial 'Toy' Experiment
- Naïve GMM ML classifier; 32 mixtures, modified MFCC front-end (20 rectangular non-overlapping filters), 25/10 ms windowing, static+Δ+ΔΔ
- Closed in-set DID task (pick 1-of-4 dialects)

Ground Truth	Assigned Dialect (Speech Chunks)				Acc (%) [Avg 82.0]
	Gulf	Iraqi	Levantine	Egyptian	
Gulf	**510**	120	4	1	80.3
Iraqi	184	**527**	1	2	73.8
Levantine	120	10	**370**	0	74.0
Egyptian	8	0	0	**3174**	99.7

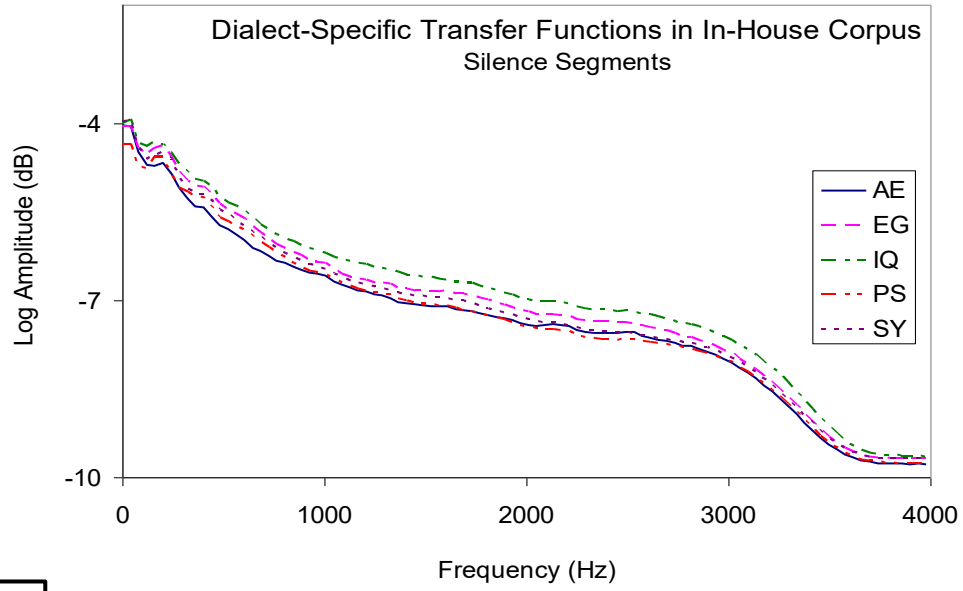
- Suspiciously high accuracy
  - do we detect dialects or something else…  …?

H. Boril, A. Sangwan, J.H.L. Hansen, "Arabic Dialect Identification - 'Is the Secret in the Silence?' and Other Observations," ISCA Interspeech-2012, Portland, OR, Sept. 9-13, 2012

2.2 Silence – Based DID

DID on Pan-Arabic Corpus – Silence Chunks

- Repeated GMM-based DID experiment on silence chunks as in LDC case → Acc = 24.7% (chance) → **speech needed for DID here!**
- Long-term channel characteristics – similar trend across dialects → **meaningful corpus for Arabic DID**



2.2 Silence – Based DID

DID on LDC Corpora – Silence Chunks

- Suspiciously high accuracy in previous case – do we detect dialects or something else…  …?
- Task 2 – similar to previous task, but now on **silence chunks**
- The overall DID accuracy is higher on silence segments (82.0→83.3%)
- For our naïve classifier, **the presence of speech actually hurts DID**

DID on "Silence" segments

Ground Truth	Assigned Dialect (Silence Chunks)				Acc (%) [Avg 83.3]
	Gulf	Iraqi	Levantine	Egyptian	
Gulf	**260**	78	0	0	76.9
Iraqi	96	**228**	0	0	70.4
Levantine	24	1	**158**	1	85.9
Egyptian	0	0	0	**1973**	100

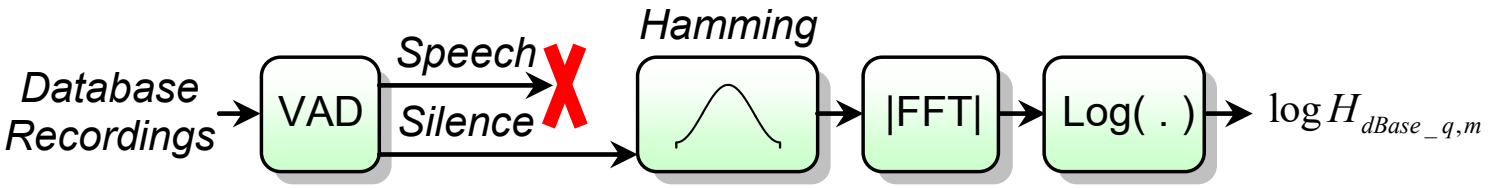
DID on "Speech"

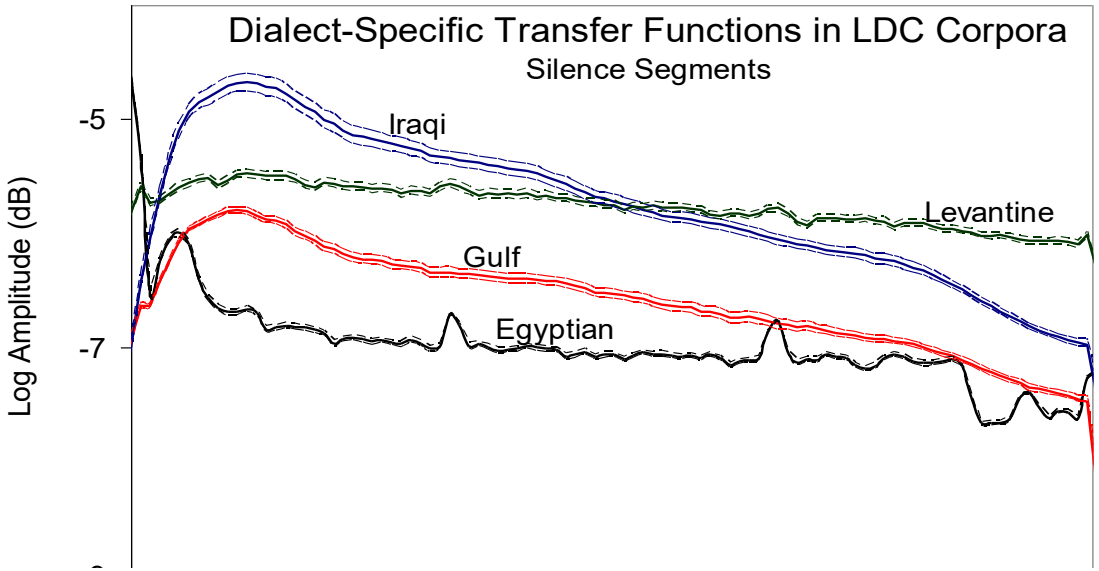
Acc (%) [Avg 82.0]
80.3
73.8
74.0
99.7

2.2 Silence – Based DID

Search for Non-Linguistic Cues

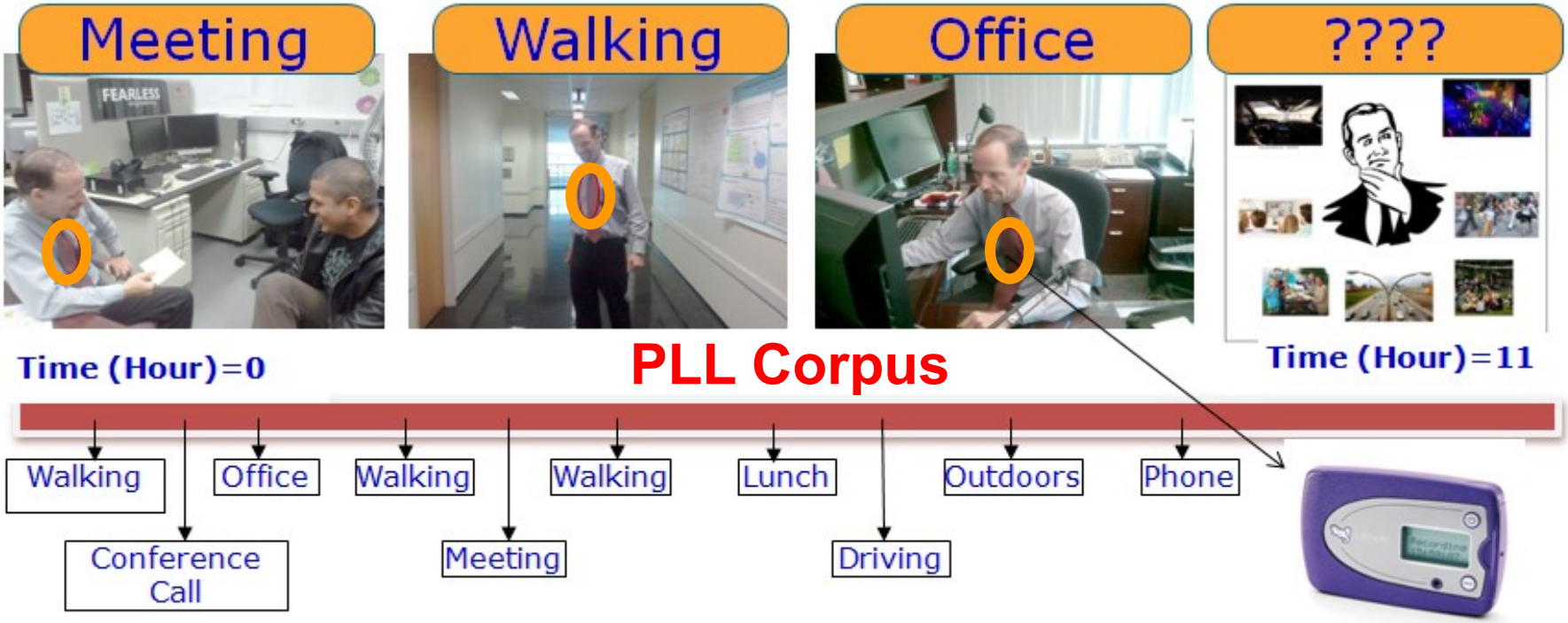
- Analysis of long-term channel characteristics in silence segments





For KWS network model training,
Know your problem, ***Know your data***

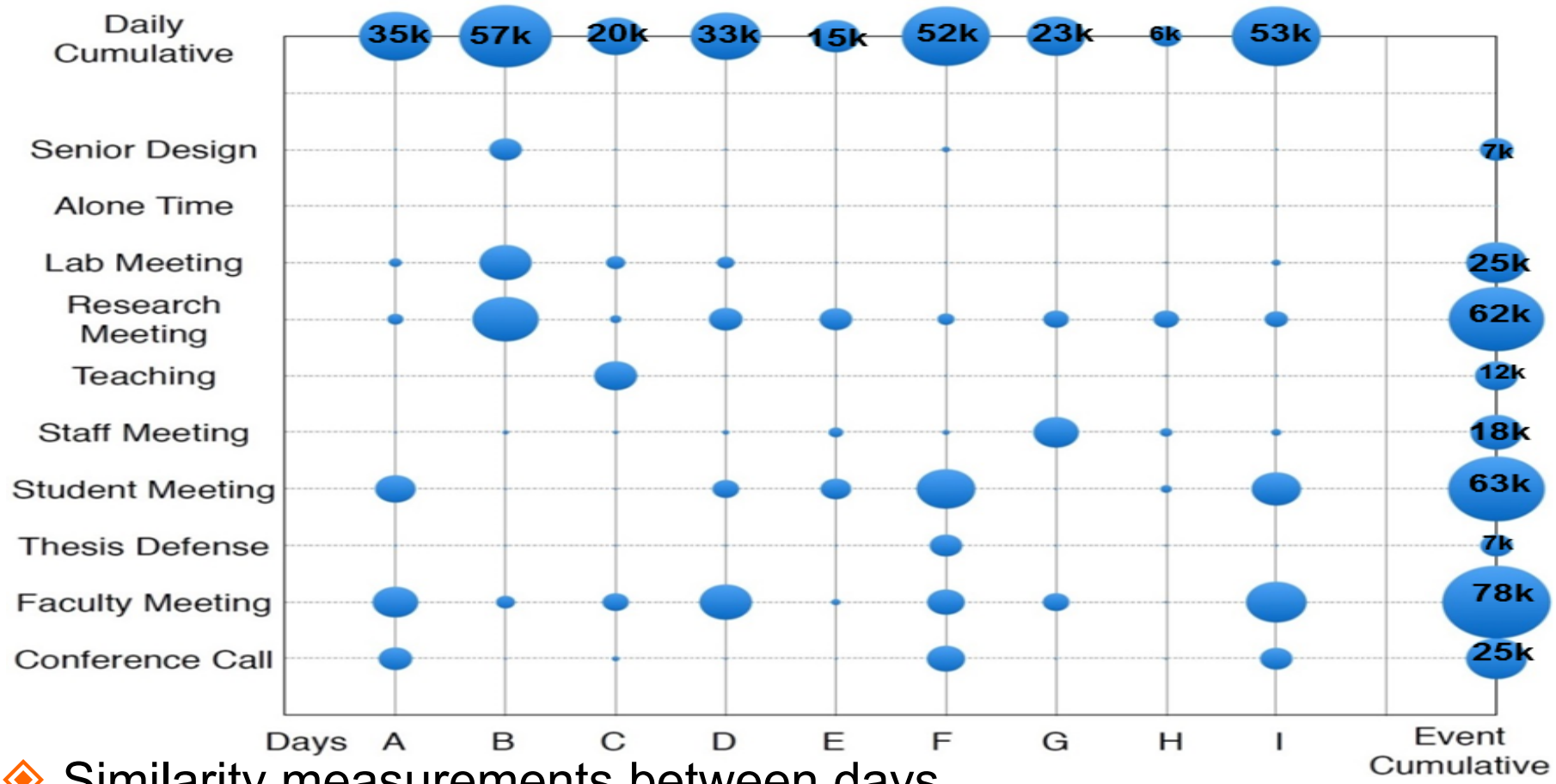
2.3 Prof-Life-Log: Massive Data



- Unscripted speech collection in natural environments
- Excellent Naturalistic Audio for DIARIZATION advancements
- Unrestricted topics, vocabulary, language; 8-16hrs/day; +100days
- Good for: Diarization; SID; KWS; Co-Speaker research

2010-2019

2.3 Prof-Life-Log: Daily Word Count



- Similarity measurements between days….
- Most similar: A & I (ρ=0.87); Most diverse : B & F (ρ=0.27)

2.3 Prof-Life-Log: Keyword Recognition

Phone Confusion Network (PCN) based Search Strategy

Abhijeet Sangwan and John H.L. Hansen

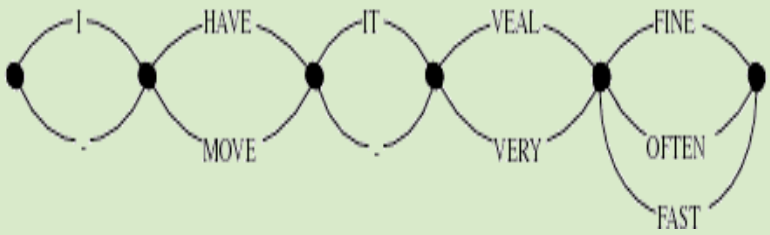
- **Phone Confusion Network (PCN) based Keyword Recognition**
  - **Advantages of phone based approach**
    - Faster than LVCSR (large vocab. continuous speech recognition)
    - No issues with OOV (out of vocabulary) queries (unlike LVCSR)
    - Flexibility in dealing with pronunciation variations (unlike LVCSR)
    - Useful where LMs (language models) are hard to build (LM required in LVCSR)
  - **Disadvantages**
    - High false-alarm rates
    - Cannot take advantage of higher level lexical knowledge (due to lack of both pronunciation & language models)
- **Phone Confusion Network based Keyword Recognition**
  - Search for keywords inside PCNs using maximum likelihood criterion
  - Keyword represented as a phone graph; Algorithm framework capable of incorporating pronunciation variations

2.3 Keyword Recognition (KWS):
KWS System Description: PCN-KWS vs. DM-PLS

Abhijeet Sangwan and John H.L. Hansen

Speech Signal

Confusion Network



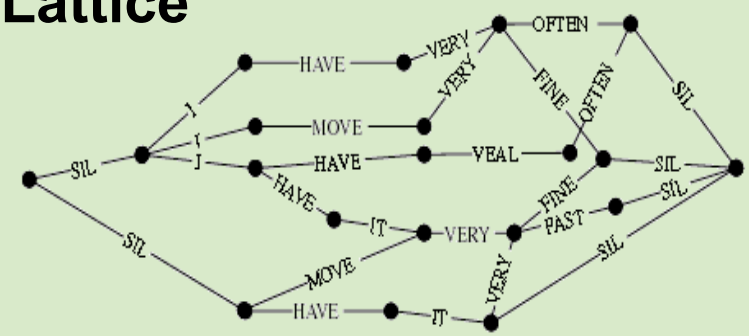
PCN-KWS

PCN = Phone Confusion Network

DM-PLS = Dynamic Match-Phone Lattice Search

ASR

Lattice

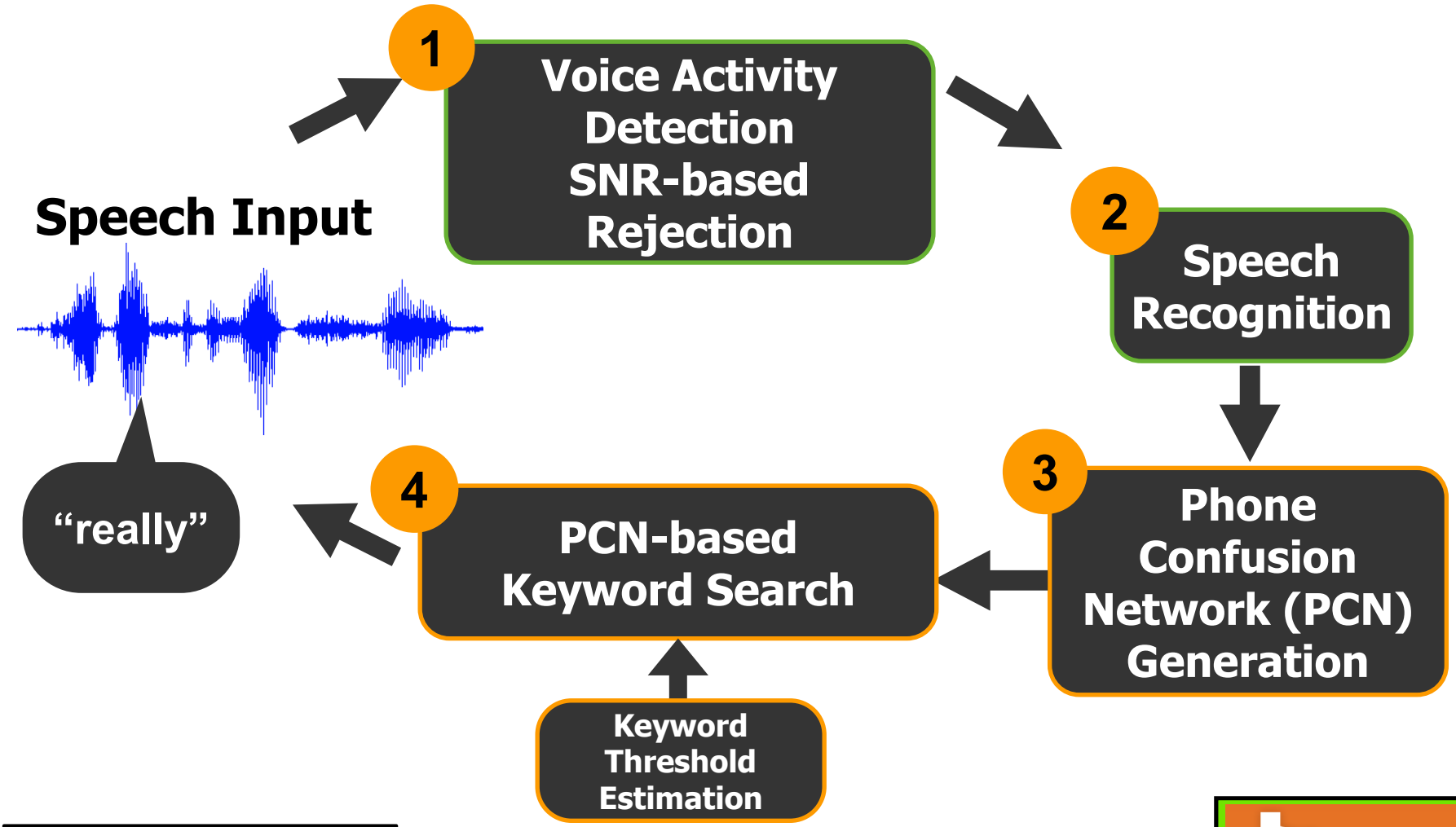


DM-PLS

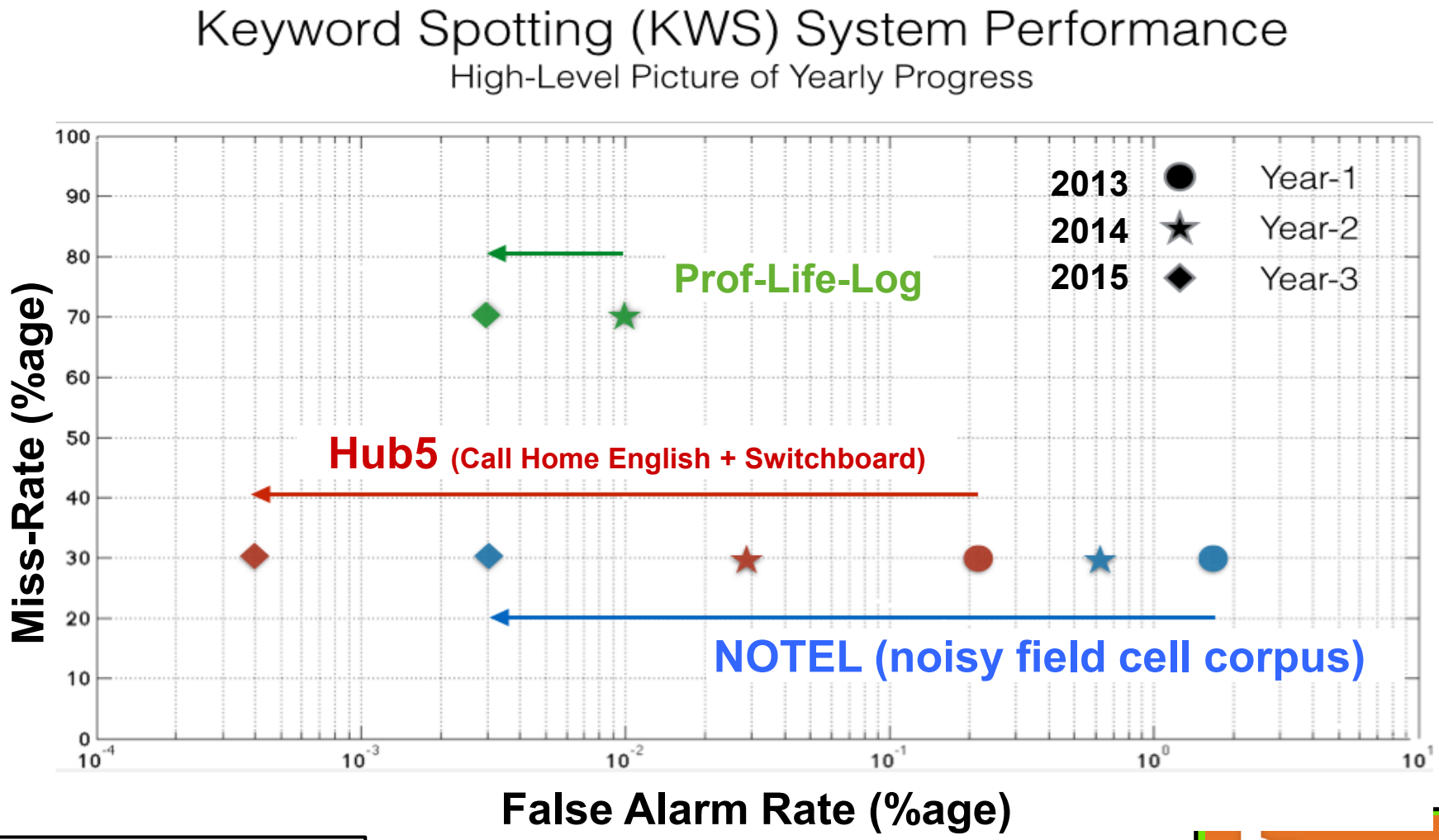
INTERSPEECH 2022
September 18 - 22 • Incheon Korea

UT DALLAS

2.3 Keyword Recognition (KWS / ASR):
Phone Confusion Network (PCN): Prof-Life-Log:



2.3 Keyword Spotting/Recognition: High Level View (PCN-KWS)



2.4 KWS/ASR: DARPA RATS Task
Communications Noise Field (Farsi, Arabic, Dari, Urdu, Pashtu)

(2012)

DARPA RATS SCENIC TEAM

John H.L. Hansen, Abhijeet Sangwan, Wooil Kim, Omid Sadjadi, Keith Godin

Center for Robust Speech Systems
University of Texas at Dallas
http://crss.utdallas.edu/

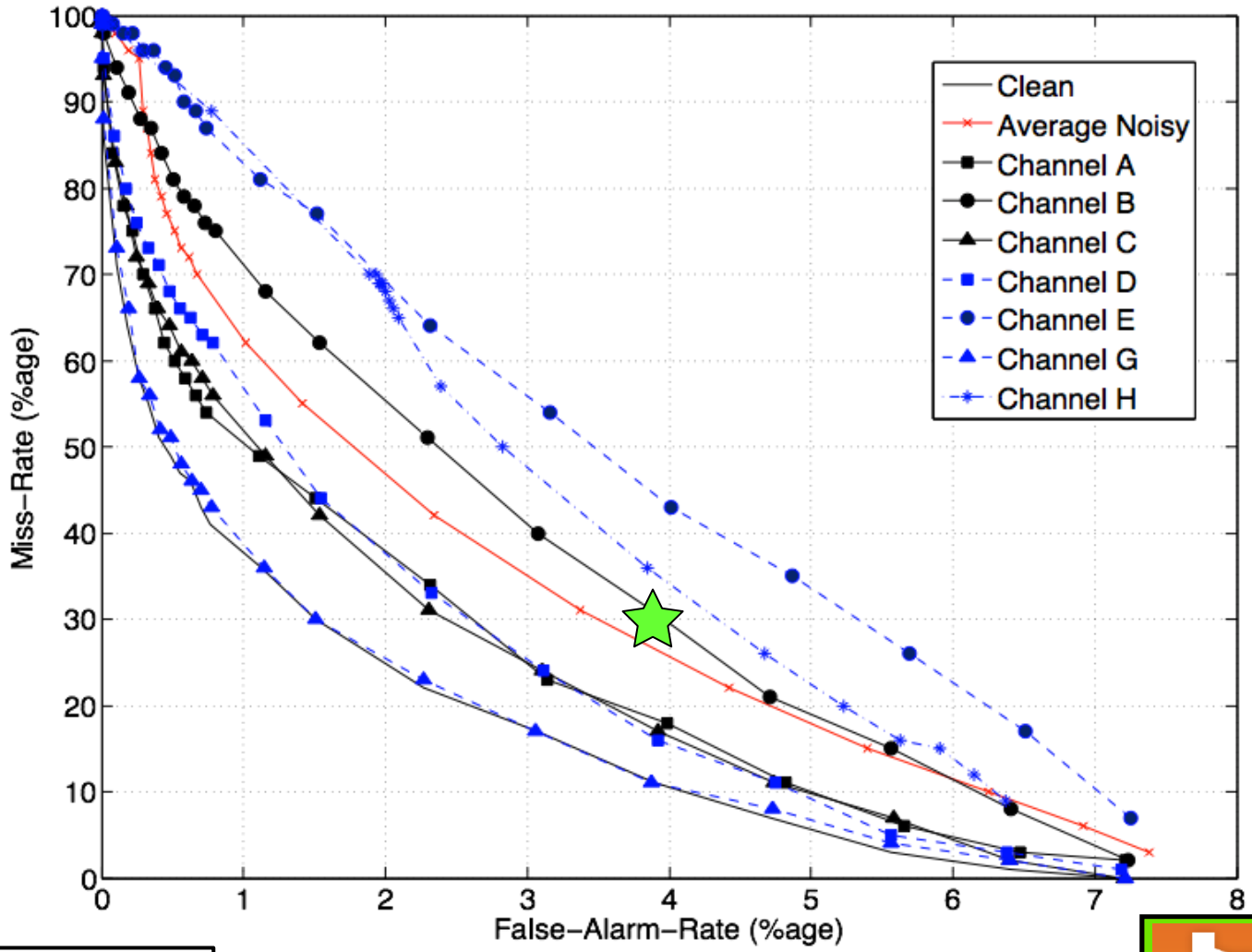
KWS

- 1. PCN: Phone Confusion Networks – explore Query Expansion
- 2. Things that work: Cepstral Based Normalization methods, Query Length, SNR based KW rejection
- 3. Ideas: Hybrid Sub-Word solution: between "word" and "phone" level sub-word system; On Demand Discriminative Keyword Modeling; Front-End Processing

2.4 Keyword Spotting: DARPA RATS

Baseline Performance (PCN-KWS) on RATS-Farsi

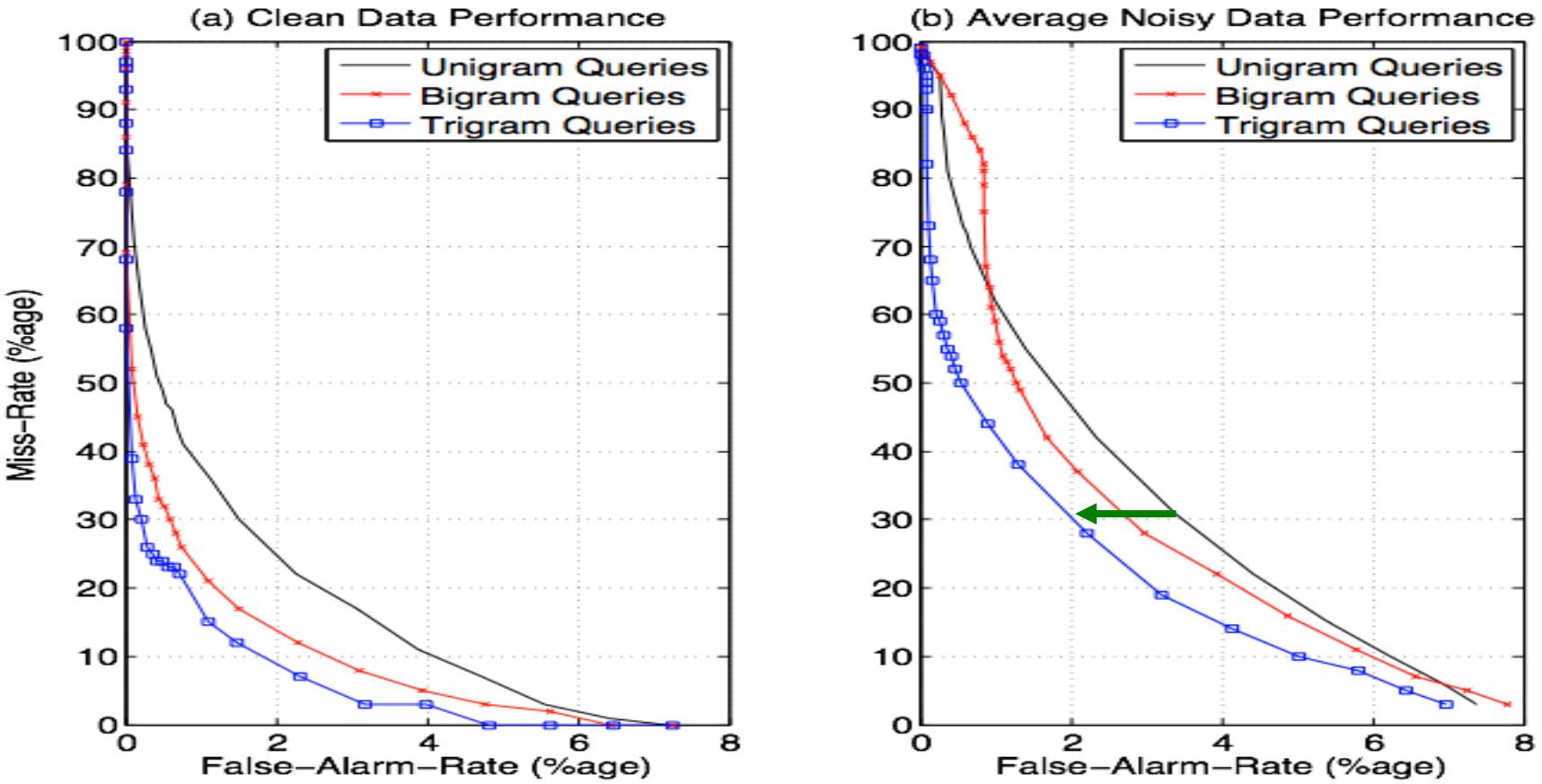
(2012)



2.4 Keyword Spotting: DARPA RATS

Impact of Query Length: (PCN-KWS) on RATS-Farsi

(2012)



- As expected, better KWS results with longer queries
- Automatic Query Expansion also considered to reduce False-Alarms

2.5 Child Speech Diarization Technology: Motivation, System & Results: WH-words

Child – Adult Vocal Engagement: "Hot Spot" Detection

- Engagement: Quality interactions of children during formative years crucial for development
- Corpus: day-long speech recordings of 33 children (2.5-5 yrs) using LENA devices
- Tagging: communication metric calculation (word count, talk time for child w/ adult, etc.)
- RESULTS: Diarization w/ CRSS-diar v1 toolkit provides diarization error rate of 40.44% in tagging adult vs. children's speech

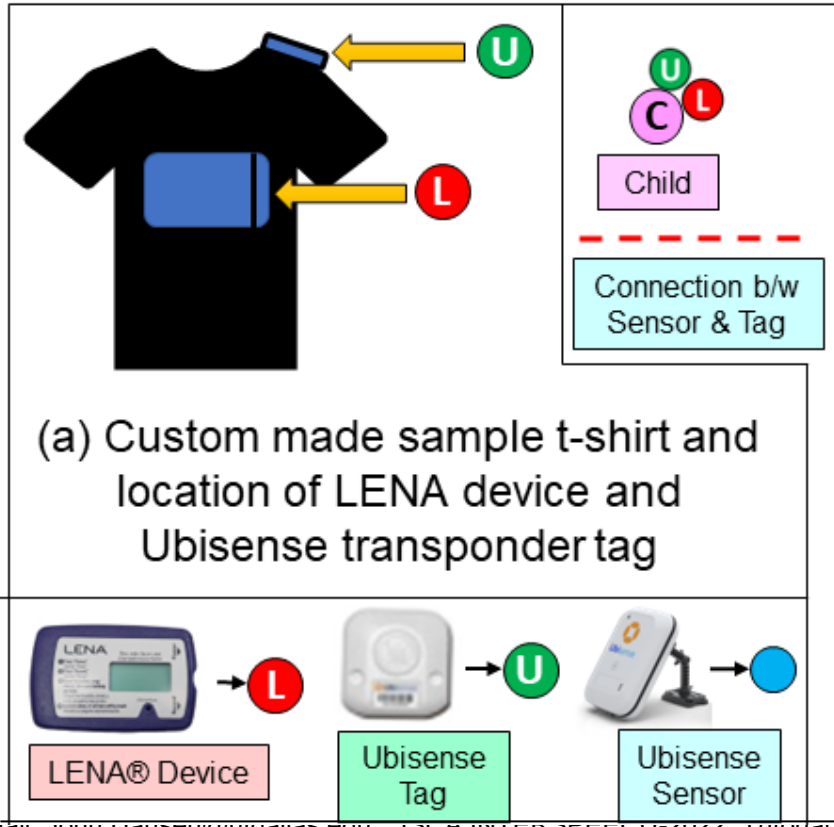
Ubisense device

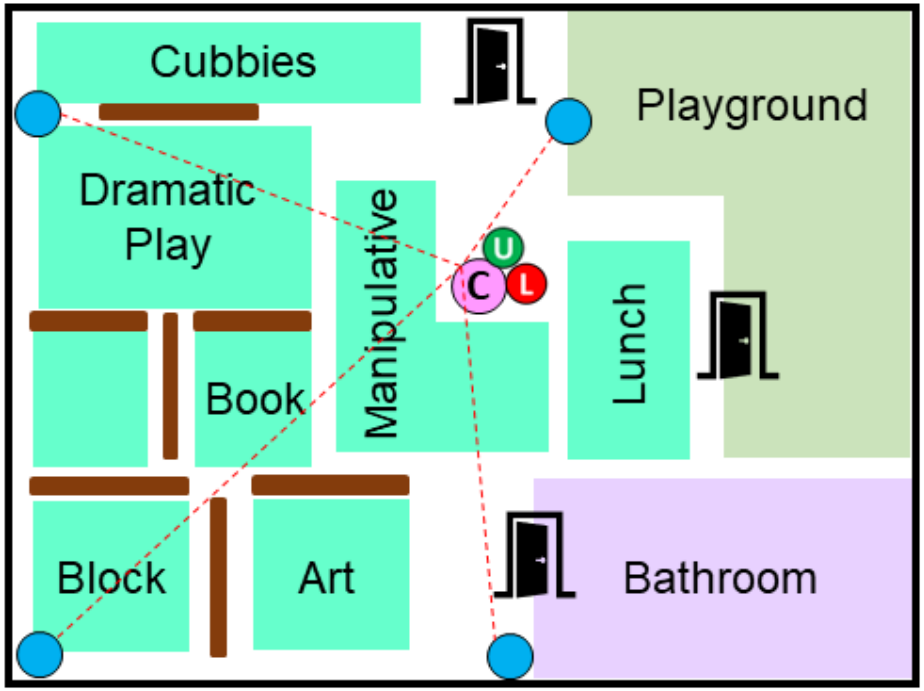
LENA device

2.5 Child Learning Spaces: WH-words & Verbs

WH-words (who, what, when, where, why, how)

- **WH-WORDS & VERBS:** language learning milestones established by ASLHA; adopted by CDC's; (**WH = child curiosity; Verbs = Grammar knowledge**)
- GOAL: capture **young children's interactions** with teachers in naturalistic **preschool classroom** environments
- ASR/KWS in preschool child classrooms -> **Challenging** task





2.5 KWS Child Learning Spaces: WH-words, Verbs
Impact of Child ASR on Language Learning Milestones

- **ASR Acoustic Model**: TDNN-F + CNN + Attention
- **ASR Language Model**: RNN

Child ID and type	% correctly identified by ASR		
	Verbs	WH-words	Sentences*
#1 + P	75	95	52
#2 + P	77	81	35
#3 + P	69	82	45
#4 + P	66	41	33
#5 + P	58	54	27
#6 + P (delayed)	51	37	28
#7 + S	64	77	38
#8 + S	69	80	49
#9 + S	69	66	40
#10 + S	69	82	43
#11 + S	61	86	36
#12 + S	61	50	39
#13 + S	66	100	41
#14 + S (delayed)	50	50	26

P = Primary, S = Secondary
*Recognized without any insertion/deletion/substition

- Test Split: 14 preschool children (3-5 yrs) with & w/o speech/language delays
- Primary child wears LENA, Secondary children are background speakers

[1] S. Dutta, S.A. Tao, J.C. Reyna, R.E. Hacker, D.W. Irvin, J.F. Buzhardt, J.H.L. Hansen. "Challenges remain in Building ASR for Spontaneous Preschool Children Speech in Naturalistic Educational Environments." ISCA Interspeech-2022, Sept. 2022.

2.6 Child Speech Recognition: WH-words

KWS Recognition Results (Child vs Adult)

Adult WH-word metrics

WH-Word /Location	WHAT	WHEN	WHERE	HOW	WHY	WHO
Science	65.2%	**100.0%**	50.0%	83.3%	**100.0%**	66.7%
Reading	**85.3%**	92.3%	**81.8%**	**94.1%**	92.3%	**75.0%**
All	72.2%	69.0%	71.0%	72.8%	52.4%	73.2%

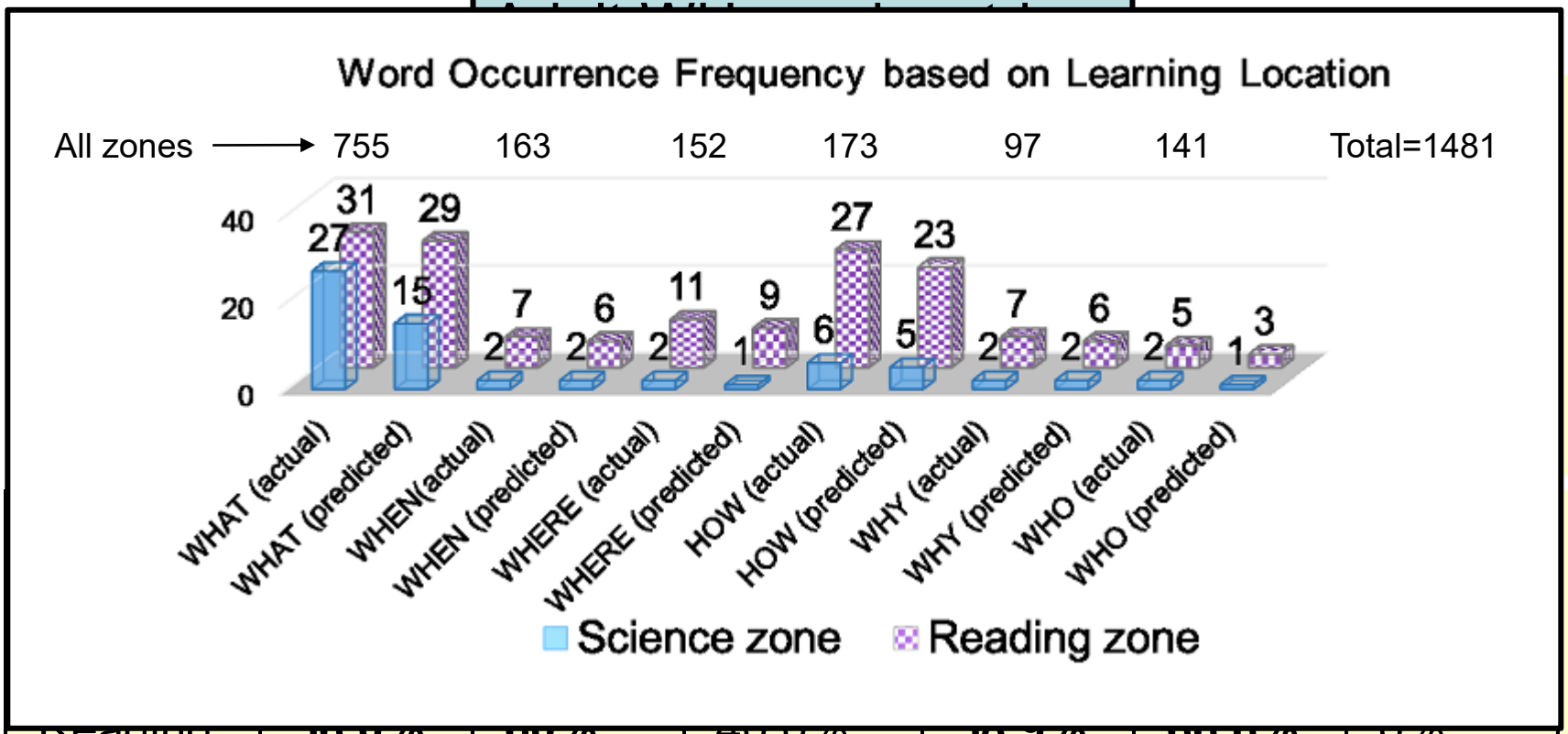
Child WH-word metrics

WH-Word /Location	WHAT	WHEN	WHERE	HOW	WHY	WHO
Science	41.5%	0%	**53.3%**	28.6%	50.0%	0%
Reading	**56.0%**	**60%**	40.0%	**58.9%**	**66.6%**	0%
All	47.3%	39.1%	30.2%	44.5%	56.0%	**39.8%**

Most F1-scores better in reading vs. science; suggested due to better audio environmental conditions

2.6 Child Speech Recognition: WH-words

KWS Recognition Results (Child vs Adult)



Reading	56.0%	60%	40.0%	58.9%	66.6%	0%
All	47.3%	39.1%	30.2%	44.5%	56.0%	**39.8%**

Most F1-scores better in reading vs. science; suggested due to better audio environmental conditions

2.7 Fearless Steps Challenges (2019/22)

Annotated (speaker label) 100-hours of audio from 5 critical loops

- **3 Apollo-11 Mission Events:**
  - Lift-off: 25 hours
  - Lunar Landing: 50 hours
  - Lunar Walking: 25 hours
- Dataset used for various tasks including speech activity detection, speaker recognition, and diarization; ***"human team assessment"***

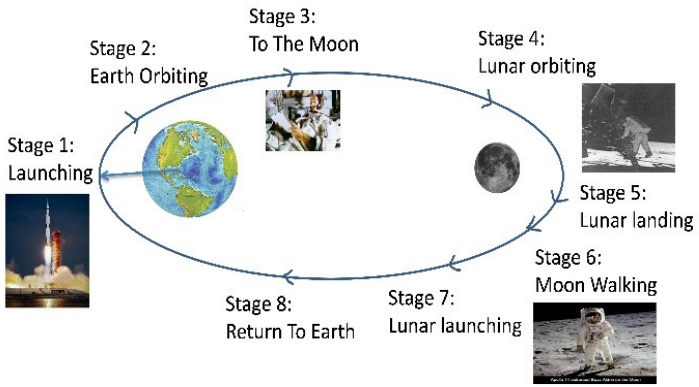
Challenge

- **+160** Academic & Industry Organizations from **35** **Countries** have participated!!
- 110 System Submissions from 15 teams received and evaluated (FS-3)!

Results** announced at the exact time of the **First Step taken on the Moon!

2.7 Fearless Steps Challenges (2019/22)

- **Five Challenges:**
  - **SAD**: Speech Activity Detection
  - **SD**: Speaker Diarization
  - **SID**: Speaker Identification
  - **ASR**: Automatic Speech Recognition
  - **SENT**: Sentiment Detection
- **Three Apollo-11 Mission Stages:**
  - *Lift-Off*
  - *Lunar Landing*
  - *Lunar Walking*

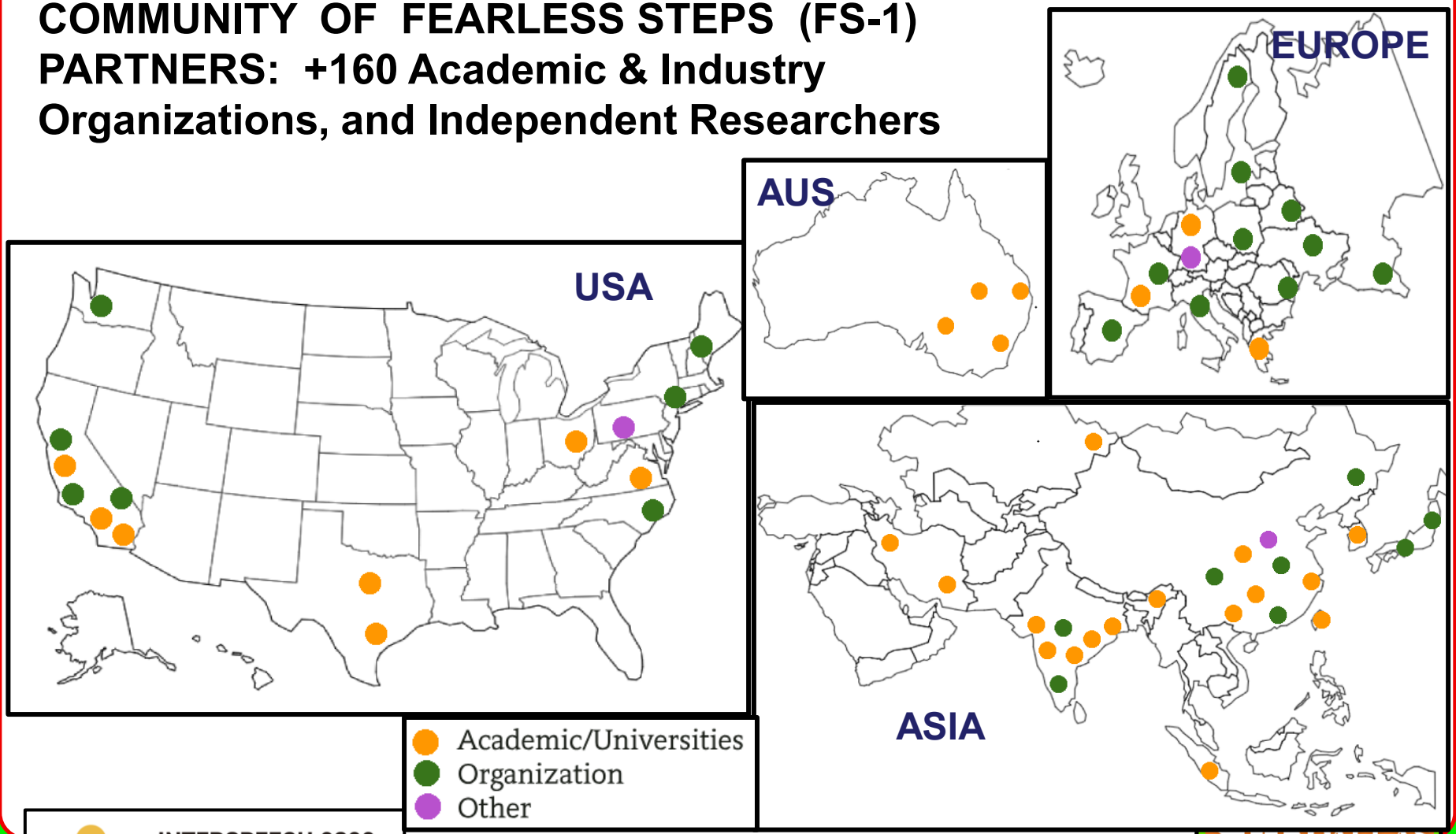


- **Data from Five NASA Audio Channels:**
  - **MOCR**: Mission Operations Control Room
  - **FD**: Flight Director
  - **EECOM**: Electrical, Environmental, and Consumables Manager
  - **GNC**: Guidance, Navigation, and Control
  - **NTWK**: Network Controller

FS Phase-1 Challenge Dates: February 11 to June 28 2019

2.7 Fearless Steps Challenges (2019/22)

COMMUNITY OF FEARLESS STEPS (FS-1)
PARTNERS: +160 Academic & Industry Organizations, and Independent Researchers



3. Robustness in Keyword Spotting

Summary

- **Introduction**: Robustness, Machine Learning & Data – Challenges
- **Speech Data**: Mismatch (Intrinsic, Extrinsic, and Context based issues)
- **Speech Task ML Challenges**: (know your problem, know your data)
- **KWS – Robustness (Distance based Speech Capture, Data, Systems)**
- **Case Studies**
  - Conversational Analysis: Prof-Life-Log: Word Count Estimation, KWS
  - Building ML Models: Dialect ID "Is the secret in the silence?"
  - KWS: Various Speech Corpora (Clean – Noisy – Naturalistic)
  - KWS: DARPA RATS; Naturalistic Learning Spaces(Child-Teacher; Apollo, etc.)

[1] J.H.L. Hansen, H. Boril, "On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks," Speech Communication, vol. 101, pp. 94-108, July 2018.
[2] M. Mirsamadi, J.H.L. Hansen, "Multi-domain adversarial training of neural network acoustic models for distant speech recognition," Speech Communication, vol. 106, pp. 21-30, Jan. 2020
[3] I. López-Espejo, Z.-H. Tan, J.H.L. Hansen, J. Jensen, "Deep Spoken Keyword Spotting: An Overview," IEEE Access, vol, 10, pp.4169 - 4199, 2022.
[4] J.H.L. Hansen, T. Hasan, "Speaker Recognition by Machines and Humans: A Tutorial Review," IEEE Signal Processing Magazine, pp. 74-99, Nov. 2015.

References (JH)

[1] J.H.L. Hansen, A. Stauffer, W. Xia, "Nonlinear Waveform Distortion: Assessment and Detection of Clipping on Speech Data and Systems," Speech Communication. vol. 134, pp. 20-31, Sept. 2021. (https://doi.org/10.1016/j.specom.2021.07.007)

[2] J.H.L. Hansen, H. Boril, "On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks," Speech Communication, vol. 101, pp. 94-108, July 2018.

[3] J.H.L. Hansen, H. Boril, “Robustness in Speech, Speaker, and Language Recognition: “You’ve Got to Know Your Limitations,” ISCA INTERSPEECH-2016, Paper ID: 1395, pp. 2766-2770, San Francisco, CA, Sept. 8-12, 2016.

[4] H. Boril, A. Sangwan, J.H.L. Hansen, "Arabic Dialect Identification - 'Is the Secret in the Silence?' and Other Observations," ISCA Interspeech-2012, Mon-O1b-01, pg. 1-4,, Portland, OR, Sept. 9-13, 2012

[5] V. Kothapally, J.H.L. Hansen, "SkipConvGAN: Monaural Speech Dereverberation using Generative Adversarial Networks via Complex Time-Frequency Masking," IEEE Trans. Audio, Speech, Lang. Proc., . vol, 30, pp. 1600-1613, Mar.. 2022

[6] R. Lileikyte, D. Irvin, J.H.L. Hansen, "Assessing Child Communication Engagement and Statistical Speech Patterns for American English via Speech Recognition in Naturalistic Active Learning Spaces," Speech Communication, pp. 98-108, 2022

[7] I. López-Espejo, Z.-H. Tan, J.H.L. Hansen, J. Jensen, "Deep Spoken Keyword Spotting: An Overview," IEEE Access, vol, 10, pp.4169 - 4199, 2022.

[8] S. Ranjan, C. Zhang, J.H.L. Hansen, "Curriculum Learning Based Approaches for Robust End-to-End Far-Field Speech Recognition," Speech Communication, vol. 132, pp. 123-131, Sept. 2021

[9] D.W. Irvin, Y. Luo, J.M. Huffman, N. Grasley-Boy, B. Rous, J.H.L. Hansen, “Capturing talk and proximity in the classroom: Advances in measuring features of young children's friendships", Early Childhood Research Quarterly, pp. 102-109, 2021.

[10] M. Yousefi, J.H.L. Hansen, “Block-based high performance CNN architectures for frame-level Overlapping Speech Detection", IEEE Trans. Audio, Speech and Language Processing, vol, 29, pp. 28-40, Nov. 2020.

[11] M. Mirsamadi, J.H.L. Hansen, "Multi-domain adversarial training of neural network acoustic models for distant speech recognition," Speech Communication, vol. 106, pp. 21-30, Jan. 2020

[12] J.H.L. Hansen, M. Najafian, R. Lileikyte, D. Irvin, B. Rous, "Speech and Language Processing for Assessing Child-Adult Interaction Based on Diarization and Location," Inter. Journal of Speech Tech., Vol. 22, pp. 697-709, June 2019.

[13] L.N. Kaushik, A.Sangwan, J.H.L. Hansen, "Speech Activity Detection In Naturalistic Audio Environments: Fearless Steps Apollo Corpus," IEEE Signal Processing Letters, vol. 25, no. 9, pp. 1290-1294, Sept. 2018.

[14] Q. Zhang, J.H.L. Hansen,"Language/Dialect Recognition based on Unsupervised Deep Learning," IEEE Trans. on Audio, Speech and Lang. Proc., vol. 26, no. 5, pp. 873-882, May 2018.

References (JH)

[15] S. Ghaffarzadegan, H. Boril, J.H.L. Hansen, "Deep Neural Network Training for Whisper Speech Recognition using Small Databases and Generative Model Sampling," Inter. Journal of Speech Technology, Vol. 20, Issue 4, pp. 1063-1075, Dec. 2017

[16] C. Yu, J.H.L. Hansen, "A study of voice production characteristics of astronaut speech during Apollo-11 for long-term speaker modeling in space," Journal of the Acoustical Society of America, Vol. 141, No. 3, pp.1605-1614, March 2017

[17] F. Weng, P. Angkititrakul, E. Shriberg, L. Heck, S. Peters, J.H.L. Hansen, "Conversational In-Vehicle Dialog Systems: The past, present, and future," IEEE Signal Processing Magazine: Special Issue - Signal Processing for Smart Vehicle Technologies, vol. 33, no. 6, , pp. 49-60, Nov. 2016.

[18] A. Ziaei, A. Sangwan, J.H.L. Hansen, "Effective word count estimation for long duration daily naturalistic audio recordings," Speech Communication, vol. 84, pp. 15-23, Nov. 2016.

[19] S.M. Mirsamadi, J.H.L. Hansen, "A Generalized Nonnegative Tensor Factorization Approach for Distant Speech Recognition with Distributed Microphones" IEEE Trans. Audio, Speech, Lang. Proc., vol. 24, No.10, pp. 1721-1731, Oct. 2016.

[20] J.H.L. Hansen, T. Hasan, "Speaker Recognition by Machines and Humans: A Tutorial Review," IEEE Signal Processing Magazine, pp. 74-99, Nov. 2015.

[21] S.O. Sadjadi, J.H.L. Hansen, "Blind Spectral Weighting for Robust Speaker Identification under Reverberation Mismatch Conditions," IEEE Trans. Audio, Speech Lang. Proc., vol. 22, no. 5, pp. 935-943, May. 2014

[22] J.H.L. Hansen, J.-W. Suh, P. Angkititrakul, Y. Lei, "Effective Background Data Selection for SVM-Based Speaker Recognition for Unseen Test Environments: More is Not Always Better," Inter. Journal Speech Technology, vol. 17, issue 3, pp. 211-221, Sept. 2014.

[23] S.O. Sadjadi, J.H.L. Hansen,"Unsupervised Speech Activity Detection using Voicing Measures and Perceptual Spectral Flux," IEEE Signal Processing Letters, vol. 20, no. 3, pp. 197-200, March 2013

[24] M. Akbacak, J.H.L. Hansen, "Spoken Proper Name Retrieval for Limited Resource Languages Using Multilingual Hybrid Representations," IEEE Trans. Audio, Speech and Language Processing, vol. 18, no. 6, pp. 1486-1495, Aug. 2010

[25] J.H.L. Hansen, V.S.Varadarajan, "Analysis and Compensation of Lombard Speech Across Noise Types and Levels with Application to In-Set/Out-of-Set Speaker Recognition, IEEE Trans. Audio, Speech & Lang. Proc., pp. 366-378, Feb. 2009

[26] S. Ranjan, G. Liu, J.H.L. Hansen, "An i-Vector PLDA based Gender Identification Approach for Severely Distorted and Multilingual DARPA RATS Data," IEEE ASRU-2015: Workshop, Paper#1243, Scottsdale, Arizona (USA), Dec.13-17, 2015

[27] H. Dubey, A. Sangwan, J.H.L. Hansen,"Using Speech Technology for Quantifying Behavioral Characteristics in Peer-Led Team Learning Sessions," Computer Speech & Language, vol. 46, pp. 343-366, 2017.

References (JH)

[28] J.H.L. Hansen, A. Joglekar, M. Chandra Shekhar, V. Kothapally, C. Yu, L. Kaushik, A. Sangwan "The 2019 Inaugural Fearless Steps Challenge: A Giant Leap for Naturalistic Audio," ISCA INTERSPEECH-2019, pp. 1851-1855, Sept.15-19, 2019.
[29] J.H.L. Hansen, A. Sangwan, A. Joglekar, A.E. Bulut, C. Yu and L. Kaushik, "Fearless Steps: Apollo-11 Corpus Advancements for speech technologies from Earth to the Moon," ISCA INTERSPEECH-2018, pp. 2758-2762, Sept. 2-6, 2018.
[30] V. Mitra, et al J.H.L. Hansen, R. Stern, A. Sangwan, N. Morgan, "Fusion Strategies for Robust Speech Recognition and Keyword Spotting for Channel- and Noise-Degraded Speech," ISCA INTERSPEECH-2016, pp. 3683-3687,Sept. 8-12, 2016.

[31] C. Hart, B., & Risley, T. R. (1995). Meaningful differences in the everyday experience of young American children. Paul H Brookes Publishing.
[32] C. Eberbach, K. Crowley."From seeing to observing: How parents and children learn to see science in a botanical garden," Journal of the Learning Sciences, vol. 26, no. 4, pp. 608–642, 2017.
[33] P. J Yoder, B. Davies, K. Bishop, L. Munson. 1994. Effect of adult continuing wh-questions on conversational participation in children with developmental disabilities. Journal of Speech, Language, and Hearing Research 37,1 (1994), 193–204.
[34] M. L Rowe, K.A Leech, N. Cabrera, "Going beyond input quantity: Wh-questions matter for toddlers' language and cognitive development," Cognitive Science, vol. 41, pp. 162–179, 2017.
[35] URL: https://www.lenafoundation.org
[36] M. Woźniak, W. Odziemczyk, K.l Nagórski, "Investigation of practical and theoretical accuracy of wireless indoor positioning system Ubisense," Reports on Geodesy and Geoinformatics, vol. 95, no. 1, 36–48., 2013.
[37] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," ISCA Interspeech-2018, pp. 3743–3747, 2018.
[38] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, Nelson Enrique Yalta Soplin, et al., "A comparative study on transformer vs RNN in speech applications. IEEE ASRU-2019: Automatic Speech Recog. and Understanding Workshop. pp. 449–456, 2019.

DEEP SPOKEN KEYWORD SPOTTING

5. Audio-Visual Keyword Spotting

Zheng-Hua Tan

Department of Electronic Systems, Aalborg University, Denmark

zt@es.aau.dk

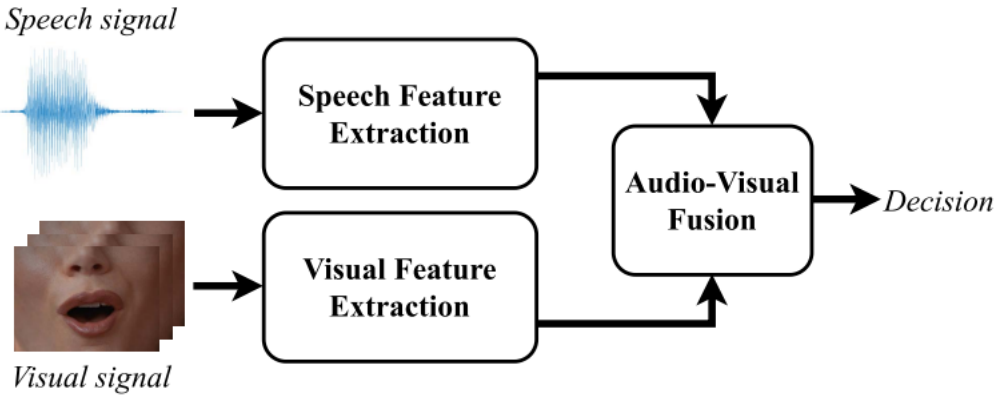
Sunday 18th September, 2022

Outline

- AV KWS Framework
- Visual feature extraction
- Audio-visual fusion
- Noise-robustness
- Benchmark datasets

Introduction and framework

- Human speech perception uses both auditory and visual information (e.g., lips)
  - KWS can benefit from visual info., esp. in noisy conditions
- Audio visual KWS framework [1]
  - Speech & visual feature extraction, and audio-visual fusion



on whether and where a user specified keyword occurs

[1] López-Espejo, I., Tan, Z. H., Hansen, J., & Jensen, J. (2021). Deep spoken keyword spotting: An overview. *IEEE Access*.

Deep learning setup

Visual feature extraction

- A two-step approach:
  - Face detection and lip localization (via landmark estimation)
  - (Classical) visual feature extraction itself from the lips crop, e.g., in [1]
- Alternatively, a deep learning model can take as input raw images containing the uncropped speaker's face (as a preferred approach)
  - E.g., in [2], a clip of talking face is fed into 18-layer spatio-temporal ResNet for visual feature extraction
  - 3D CNN is used in [3] as well

[1] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, ''A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion,'' IEEE Trans. Multimedia, vol. 18, no. 3, pp. 326–338, Mar. 2016
[2] L. Momeni, T. Afouras, T. Stafylakis, S. Albanie, and A. Zisserman, ''Seeing wake words: Audio-visual keyword spotting,'' in Proc. Brit. Mach. Vis. Virtual Conf., Sep. 2020.
[3] R. Ding, C. Pang, and H. Liu, ''Audio-visual keyword spotting based on multidimensional convolutional neural network,'' in Proc. IEEE Int. Conf. Image Process., Athens, Greece, Aug. 2018

Fusion strategies

- Feature-level fusion: Speech and visual features are concatenated before their joint classification
- Decision-level fusion: The final decision is made by combining decisions from separate speech and visual classifiers (preferred)
  - In [1], the softmax outputs of the audio and visual networks are combined through a summation, with fixed weights, to estimate the posterior probability of each keyword

$$P(x_l|A, \boldsymbol{V}, \boldsymbol{W}) = \alpha P(x_l|A, \boldsymbol{W}_a) + (1-\alpha) P(x_l|\boldsymbol{V}, \boldsymbol{W}_v)$$

  - Similarly in [2]. Adaptive weights based on the reliabilities of two modalities are used in [3]

[1] R. Ding, C. Pang, and H. Liu, ‘‘Audio-visual keyword spotting based on multidimensional convolutional neural network,’’ in Proc. IEEE Int. Conf. Image Process., Athens, Greece, Aug. 2018
[2] L. Momeni, T. Afouras, T. Stafylakis, S. Albanie, and A. Zisserman, ‘‘Seeing wake words: Audio-visual keyword spotting,’’ in Proc. Brit. Mach. Vis. Virtual Conf., Sep. 2020.
[3] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, ‘‘A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion,’’ IEEE Trans. Multimedia, vol. 18, no. 3, pp. 326–338, Mar. 2016

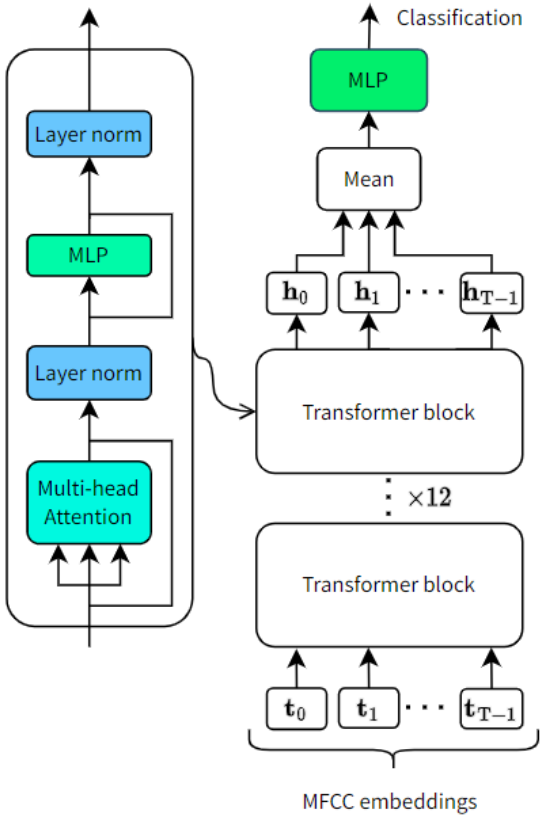
Visual keyword spotting with attention

- Video frames -> a visual front-end (e.g., CNN) (to extract low-level visual features) -> N_v Transformer layers (to encode temporal information)
- The phoneme sequence of the keyword -> N_t Transformer layers
- Text embedding + visual embedding -> a joint multi-modal Transformer to predict:
  - the probability the keyword occurs in the video
  - frame-level probabilities indicating the location of the word
- It outperforms the prior state-of-the-art methods on the challenging LRW, LRS2, LRS3 datasets by a large margin

[1] Prajwal, K. R., Momeni, L., Afouras, T., & Zisserman, A. (2021). Visual keyword spotting with attention. BMVC 2021.

Self-supervised learning for KWS

- Keyword transformer (KWT) model



Results for the three KWT models on Google Speech Commands V2 data set. Baseline (Full) indicates models trained on the full training set.

Model	Test accuracy		
	Baseline	Baseline (Full)	Data2Vec
KWT-1	0.7428	**0.9638**	0.9294
KWT-2	0.8584	0.9498	**0.9507**
KWT-3	0.8411	0.9079	**0.9529**
	20% trn	100% trn	80% + 20%

[1] H. S. Bovbjerg, Z.-H. Tan (2022). Improving Label-Deficient Keyword Spotting Using Self-Supervised Pretraining.

Noise robustness

- Audio-visual KWS achieves the greatest relative improvements with respect to audio-only KWS at lower SNRs, while it improves the performance at high SNRs as well, as consistently found in the literature, e.g., those in the previous slide

- AV KWS surpasses the performance of both video-only and audio-only KWS within a wide range of SNRs

- Robustness against lighting conditions and head pose variation has been less studied systematically

Benchmark data sets

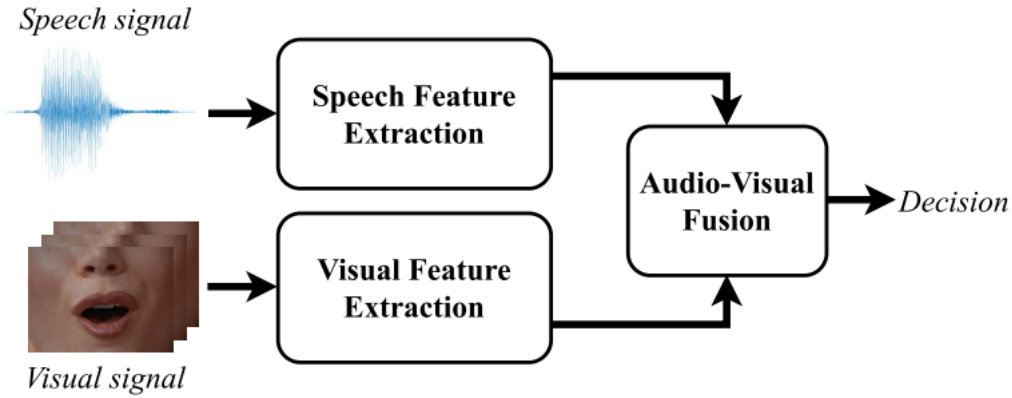
Realistic and challenging audio-visual benchmarks

- Lip Reading in the Wild (LRW) [1]
  - One of the first visual speech databases in-the-wild, ca 170h
  - Single-word utterances from BBC TV broadcasts
  - Over a million word instances, spoken by 1000+ people
- Lip Reading Sentences 2 (LRS2) [2]
  - 100,000+ natural spoken sentences from BBC TV
- Lip Reading Sentences 3 (LRS3) [3]
  - 400+ hours from TED(x) talks

[1] J. S. Chung and A. Zisserman, ‘‘Lip reading in the wild,’’ in Proc. Asian Conf. Comput. Vis., Taipei, Taiwan, 2016.
[2] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, ‘‘Lip reading sentences in the wild,’’ in Proc. Conf. Comput. Vis. Pattern Recognit., Honolulu, HI, USA, 2017.
[3] T. Afouras, J. Son Chung, and A. Zisserman, ‘‘LRS3-TED: A large-scale dataset for visual speech recognition,’’ 2018, arXiv:1809.00496.

Recap

acoustic noise



on whether and where a user specified keyword occurs

lighting condition and pose

DEEP SPOKEN KEYWORD SPOTTING

6. Technology Applications

Zheng-Hua Tan

Department of Electronic Systems, Aalborg University, Denmark

zt@es.aau.dk

Sunday 18th September, 2022

INTERSPEECH 2022
September 18 - 22 • Incheon Korea

Outline

- Applications in general
- Voice activation of voice assistants
- Personalized keyword spotting systems
- Voice control of hearing assistive devices

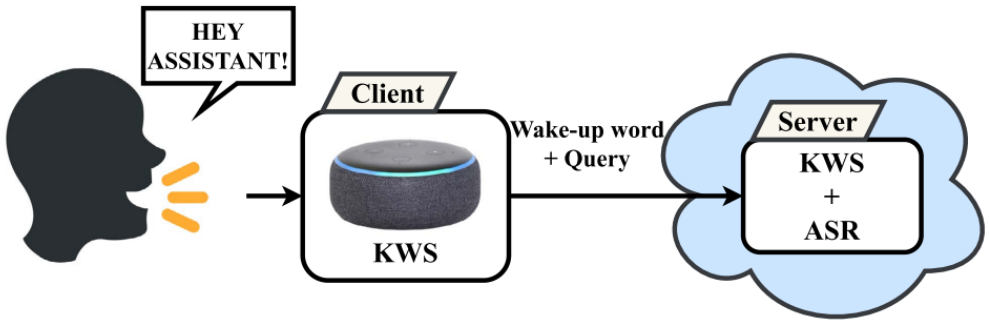
Applications

- the activation of voice assistants
- speech retrieval
- voice-dialing, and interaction with a call center
- assistive technology for vision-impaired people in special scenarios, e.g., the activation of pedestrian call buttons in crosswalks
- hands-free voice control in in-vehicle systems, videogames, home automation
- human robot interaction
- etc.

López-Espejo, I., Tan, Z. H., Hansen, J., & Jensen, J. (2021). Deep spoken keyword spotting: An overview. *IEEE Access*.

Voice activation of voice assistants

- The flagship application of (deep) KWS
  - By 2024, the number of voice assistant units is expected to reach 8.4b, exceeding the world's population [1].
- Typical voice assistant client-server framework [2]
  - The client device has an always-on KWS system to detect whether a user utters a wakes-up keyword/phrase
  - When the keyword is spotted, the supposed wake-up word audio and subsequent query audio are sent to a server to be processed by LVCSR



[1] https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/
[2] López-Espejo, I., Tan, Z. H., Hansen, J., & Jensen, J. (2021). Deep spoken keyword spotting: An overview. *IEEE Access*.

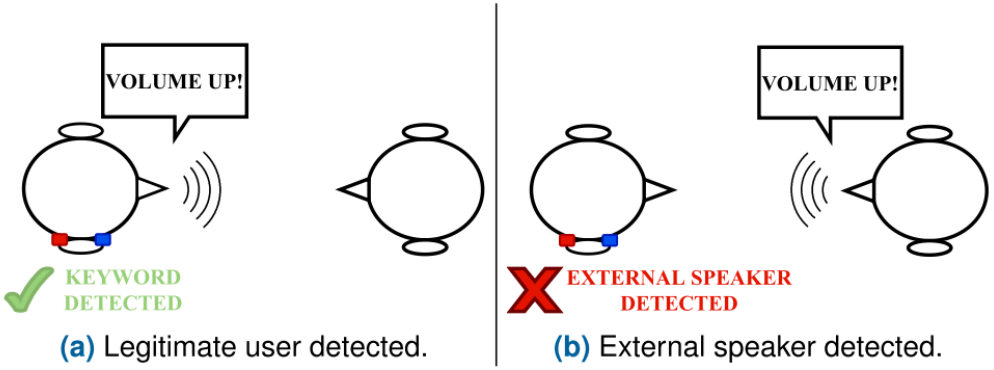
Personalized keyword spotting systems

- Personalization of the KWS system can be desirable
  - Personalized queries
  - Personalized control of devices like hearing aids
- Combining KWS and speaker verification
  - Independently trained deep learning models to perform both tasks.
    - In [1], d-vector based TI-SV (separately trained) is applied to largely reduce the false acceptence rate
  - A multi-task learning scheme, joint KWS and speaker verification
    - In [2], d-vector based TD-SV is jointly trained with KWS by sharing one convolutional layer operating on log filter bank energy

[1] Rikhye, R., Wang, Q., Liang, Q., He, Y., Zhao, D., Narayanan, A., & McGraw, I. Personalized keyphrase detection using speaker and environment information. *Interspeech 2021*.
[2] Kumar, R., Yeruva, V., & Ganapathy, S. On Convolutional LSTM Modeling for Joint Wake-Word Detection and Text Dependent Speaker Verification. In *Interspeech* 2018.

Voice control of hearing assistive devices

- Manually operating small, body-worn devices like hearing aids is not always feasible or can be cumbersome.
- An alternative way to speaker verification to provide personalization in KWS for hearing aids [1]:
  - Multitask learning
  - Exploiting GCC-PHAT coefficients from dual-microphone hearing aids, achieves almost flawless users' own voice/external speaker detection (reducing FAR)



[1] López-Espejo, I., Tan, Z. H., & Jensen, J. (2020). Improved external speaker-robust keyword spotting for hearing assistive devices. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Summary

- Applications in general
- Voice activation of voice assistants
- Personalized keyword spotting systems
- Voice control of hearing assistive devices

Thank you for your attention!

DEEP SPOKEN KEYWORD SPOTTING

7. Experimental Considerations

Iván López-Espejo

Department of Electronic Systems, Aalborg University, Denmark

ivl@es.aau.dk

Sunday 18th September, 2022

INTERSPEECH 2022
September 18 - 22 • Incheon Korea

Overview

Datasets

- **Data** are an essential ingredient of any machine learning system for both training the parameters of the algorithm and validating it

- Corpora used over the years in ASR are now also being employed for deep KWS:
  - *LibriSpeech*
  - *TIDIGITS*
  - *TIMIT*
  - *Wall Street Journal (WSJ) corpus*
  - ...

- ☒ These corpora do not standardize a way of utilization facilitating KWS technology reproducibility and comparison (*e.g., the set of considered keywords*)

Datasets

Name	Developer	P.A.?	Language	Noisy?	No. of KW	Training set			Test set		
						Size	*+ sampl.*	*- sampl.*	*Size*	*+ sampl.*	*- sampl.*
-	Alibaba	N	Mandarin	Y	1	24k h	-	-	-	12k	600 h
-	Baidu	N	English	Y	1	12k	-	-	2k	-	-
-	Chinese Academy of Sciences	N	Mandarin	Y	2	47.8k	8.8k	39k	-	1.7k	-
-	Fluent.ai	N	English	Y	1	50 h	5.9k	-	22 h	1.6k	-
-	Google	N	English	Y	10	>3k h	60.7k	133k	81.2k	11.2k	70k
-	Google	N	English	Y	14	326.8k	10k	316.8k	61.3k	1.9k	59.4k
-	Harbin Institute of Technology	N	Mandarin	-	1	115.2k	19.2k	96k	28.8k	4.8k	24k
-	Logitech	N	English	-	14	-	-	-	-	-	-
-	Mobvoi	N	Mandarin	Y	1	67 h	20k	54k	7 h	2k	5.9k
-	Sonos	Y	English	Y	16	0	0	0	1.1k	1.1k	0
-	Tencent	N	Mandarin	Y	1	339 h	224k	100k	-	-	-
-	Tencent	N	Mandarin	Y	1	65.9 h	6.9 h	59 h	8.7 h	0.9 h	7.8 h
-	Tencent	N	Mandarin	Y	42	22.2k	15.4k	6.8k	10.8k	7.4k	3.4k
-	Xiaomi	N	Mandarin	-	1	1.7k h	188.9k	1M	52.2 h	28.8k	32.8k
AISHELL-2 (13)	AISHELL	Y	Mandarin	N	13	24.8 h	>24k	-	16.7 h	>8.4k	-
AISHELL-2 (20)	AISHELL	Y	Mandarin	N	20	35 h	>34k	-	23.9 h	>12k	-
"Alexa"	Amazon	N	English	Y	1	495 h	-	-	100 h	-	-
Google Speech Commands Dataset *v1*	Google	Y	English	Y	10	51.7k	18.9k	32.8k	6.5k	2.4k	4.1k
Google Speech Commands Dataset *v2*	Google	Y	English	Y	10	84.6k	30.8k	53.8k	10.6k	3.9k	6.7k
"Hey Siri"	Apple	N	English	Y	1	500k	250k	250k	-	6.5k	2.7k h
Hey Snapdragon Keyword Dataset	Qualcomm	Y	English	N	4	-	-	-	4.3k	4.3k	-
Hey Snips	Snips	Y	English	Y	1	50.5 h	5.9k	45.3k	23.1 h	2.6k	20.8k
"Narc Ya"	Netmarble	N	Korean	Y	1	130k	50k	80k	800	400	400
"Ok/Hey Google"	Google	N	English	Y	2	-	1M	-	>3k h	434k	213k
"Ok/Hey Google"	Google	N	English	Y	2	-	-	-	247 h	4.8k	7.5k
Ticmini2	Mobvoi	N	Mandarin	Y	2	157.5k	43.6k	113.9k	72.9k	21.3k	51.6k

- Datasets are normally comprised of hundreds or thousands of speakers who do not overlap across sets

Datasets

Name	Developer	P.A.?	Language	Noisy?	No. of KW	Training set			Test set		
						Size	*+ sampl.*	*- sampl.*	*Size*	*+ sampl.*	*- sampl.*
-	Alibaba	N	Mandarin	Y	1	24k h	-	-	-	12k	600 h
-	Baidu	N	English	Y	1	12k	-	-	2k	-	-
-	Chinese Academy of Sciences	N	Mandarin	Y	2	47.8k	8.8k	39k	-	1.7k	-
-	Fluent.ai	N	English	Y	1	50 h	5.9k	-	22 h	1.6k	-
-	Google	N	English	Y	10	>3k h	60.7k	133k	81.2k	11.2k	70k
-	Google	N	English	Y	14	326.8k	10k	316.8k	61.3k	1.9k	59.4k
-	Harbin Institute of Technology	N	Mandarin	-	1	115.2k	19.2k	96k	28.8k	4.8k	24k
-	Logitech	N	English	-	14	-	-	-	-	-	-
-	Mobvoi	N	Mandarin	Y	1	67 h	20k	54k	7 h	2k	5.9k
-	Sonos	Y	English	Y	16	0	0	0	1.1k	1.1k	0
-	Tencent	N	Mandarin	Y	1	339 h	224k	100k	-	-	-
-	Tencent	N	Mandarin	Y	1	65.9 h	6.9 h	59 h	8.7 h	0.9 h	7.8 h
-	Tencent	N	Mandarin	Y	42	22.2k	15.4k	6.8k	10.8k	7.4k	3.4k
-	Xiaomi	N	Mandarin	-	1	1.7k h	188.9k	1M	52.2 h	28.8k	32.8k
AISHELL-2 (13)	AISHELL	Y	Mandarin	N	13	24.8 h	>24k	-	16.7 h	>8.4k	-
AISHELL-2 (20)	AISHELL	Y	Mandarin	N	20	35 h	>34k	-	23.9 h	>12k	-
"Alexa"	Amazon	N	English	Y	1	495 h	-	-	100 h	-	-
Google Speech Commands Dataset *v1*	Google	Y	English	Y	10	51.7k	18.9k	32.8k	6.5k	2.4k	4.1k
Google Speech Commands Dataset *v2*	Google	Y	English	Y	10	84.6k	30.8k	53.8k	10.6k	3.9k	6.7k
"Hey Siri"	Apple	N	English	Y	1	500k	250k	250k	-	6.5k	2.7k h
Hey Snapdragon Keyword Dataset	Qualcomm	Y	English	N	4	-	-	-	4.3k	4.3k	-
Hey Snips	Snips	Y	English	Y	1	50.5 h	5.9k	45.3k	23.1 h	2.6k	20.8k
"Narc Ya"	Netmarble	N	Korean	Y	1	130k	50k	80k	800	400	400
"Ok/Hey Google"	Google	N	English	Y	2	-	1M	-	>3k h	434k	213k
"Ok/Hey Google"	Google	N	English	Y	2	-	-	-	247 h	4.8k	7.5k
Ticmini2	Mobvoi	N	Mandarin	Y	2	157.5k	43.6k	113.9k	72.9k	21.3k	51.6k

- The advancement of the KWS technology is led by the private sector of the United States of America and China

Datasets

Name	Developer	P.A.?	Language	Noisy?	No. of KW	Training set			Test set		
						Size	*+ sampl.*	*- sampl.*	*Size*	*+ sampl.*	*- sampl.*
-	Alibaba	N	Mandarin	Y	1	24k h	-	-	-	12k	600 h
-	Baidu	N	English	Y	1	12k	-	-	2k	-	-
-	Chinese Academy of Sciences	N	Mandarin	Y	2	47.8k	8.8k	39k	-	1.7k	-
-	Fluent.ai	N	English	Y	1	50 h	5.9k	-	22 h	1.6k	-
-	Google	N	English	Y	10	>3k h	60.7k	133k	81.2k	11.2k	70k
-	Google	N	English	Y	14	326.8k	10k	316.8k	61.3k	1.9k	59.4k
-	Harbin Institute of Technology	N	Mandarin	-	1	115.2k	19.2k	96k	28.8k	4.8k	24k
-	Logitech	N	English	-	14	-	-	-	-	-	-
-	Mobvoi	N	Mandarin	Y	1	67 h	20k	54k	7 h	2k	5.9k
-	Sonos	Y	English	Y	16	0	0	0	1.1k	1.1k	0
-	Tencent	N	Mandarin	Y	1	339 h	224k	100k	-	-	-
-	Tencent	N	Mandarin	Y	1	65.9 h	6.9 h	59 h	8.7 h	0.9 h	7.8 h
-	Tencent	N	Mandarin	Y	42	22.2k	15.4k	6.8k	10.8k	7.4k	3.4k
-	Xiaomi	N	Mandarin	-	1	1.7k h	188.9k	1M	52.2 h	28.8k	32.8k
AISHELL-2 (13)	AISHELL	Y	Mandarin	N	13	24.8 h	>24k	-	16.7 h	>8.4k	-
AISHELL-2 (20)	AISHELL	Y	Mandarin	N	20	35 h	>34k	-	23.9 h	>12k	-
"Alexa"	Amazon	N	English	Y	1	495 h	-	-	100 h	-	-
Google Speech Commands Dataset *v1*	Google	Y	English	Y	10	51.7k	18.9k	32.8k	6.5k	2.4k	4.1k
Google Speech Commands Dataset *v2*	Google	Y	English	Y	10	84.6k	30.8k	53.8k	10.6k	3.9k	6.7k
"Hey Siri"	Apple	N	English	Y	1	500k	250k	250k	-	6.5k	2.7k h
Hey Snapdragon Keyword Dataset	Qualcomm	Y	English	N	4	-	-	-	4.3k	4.3k	-
Hey Snips	Snips	Y	English	Y	1	50.5 h	5.9k	45.3k	23.1 h	2.6k	20.8k
"Narc Ya"	Netmarble	N	Korean	Y	1	130k	50k	80k	800	400	400
"Ok/Hey Google"	Google	N	English	Y	2	-	1M	-	>3k h	434k	213k
"Ok/Hey Google"	Google	N	English	Y	2	-	-	-	247 h	4.8k	7.5k
Ticmini2	Mobvoi	N	Mandarin	Y	2	157.5k	43.6k	113.9k	72.9k	21.3k	51.6k

- Except for the "Narc Ya" corpus (in Korean), all the datasets shown are in either English or Mandarin Chinese

Datasets

Name	Developer	P.A.?	Language	Noisy?	No. of KW	Training set			Test set		
						Size	*+ sampl.*	*- sampl.*	*Size*	*+ sampl.*	*- sampl.*
-	Alibaba	N	Mandarin	Y	1	24k h	-	-	-	12k	600 h
-	Baidu	N	English	Y	1	12k	-	-	2k	-	-
-	Chinese Academy of Sciences	N	Mandarin	Y	2	47.8k	8.8k	39k	-	1.7k	-
-	Fluent.ai	N	English	Y	1	50 h	5.9k	-	22 h	1.6k	-
-	Google	N	English	Y	10	>3k h	60.7k	133k	81.2k	11.2k	70k
-	Google	N	English	Y	14	326.8k	10k	316.8k	61.3k	1.9k	59.4k
-	Harbin Institute of Technology	N	Mandarin	-	1	115.2k	19.2k	96k	28.8k	4.8k	24k
-	Logitech	N	English	-	14	-	-	-	-	-	-
-	Mobvoi	N	Mandarin	Y	1	67 h	20k	54k	7 h	2k	5.9k
-	Sonos	Y	English	Y	16	0	0	0	1.1k	1.1k	0
-	Tencent	N	Mandarin	Y	1	339 h	224k	100k	-	-	-
-	Tencent	N	Mandarin	Y	1	65.9 h	6.9 h	59 h	8.7 h	0.9 h	7.8 h
-	Tencent	N	Mandarin	Y	42	22.2k	15.4k	6.8k	10.8k	7.4k	3.4k
-	Xiaomi	N	Mandarin	-	1	1.7k h	188.9k	1M	52.2 h	28.8k	32.8k
AISHELL-2 (13)	AISHELL	Y	Mandarin	N	13	24.8 h	>24k	-	16.7 h	>8.4k	-
AISHELL-2 (20)	AISHELL	Y	Mandarin	N	20	35 h	>34k	-	23.9 h	>12k	-
"Alexa"	Amazon	N	English	Y	1	495 h	-	-	100 h	-	-
Google Speech Commands Dataset *v1*	Google	Y	English	Y	10	51.7k	18.9k	32.8k	6.5k	2.4k	4.1k
Google Speech Commands Dataset *v2*	Google	Y	English	Y	10	84.6k	30.8k	53.8k	10.6k	3.9k	6.7k
"Hey Siri"	Apple	N	English	Y	1	500k	250k	250k	-	6.5k	2.7k h
Hey Snapdragon Keyword Dataset	Qualcomm	Y	English	N	4	-	-	-	4.3k	4.3k	-
Hey Snips	Snips	Y	English	Y	1	50.5 h	5.9k	45.3k	23.1 h	2.6k	20.8k
"Narc Ya"	Netmarble	N	Korean	Y	1	130k	50k	80k	800	400	400
"Ok/Hey Google"	Google	N	English	Y	2	-	1M	-	>3k h	434k	213k
"Ok/Hey Google"	Google	N	English	Y	2	-	-	-	247 h	4.8k	7.5k
Ticmini2	Mobvoi	N	Mandarin	Y	2	157.5k	43.6k	113.9k	72.9k	21.3k	51.6k

- The majority of the speech corpora of interest are for **company internal use only** (*Mobvoi's TicKasa Fox, Google's Google Home and Xiaomi's AI Speaker smart speakers*)

Datasets

Name	Developer	P.A.?	Language	Noisy?	No. of KW	Training set			Test set		
						Size	*+ sampl.*	*- sampl.*	*Size*	*+ sampl.*	*- sampl.*
-	Alibaba	N	Mandarin	Y	1	24k h	-	-	-	12k	600 h
-	Baidu	N	English	Y	1	12k	-	-	2k	-	-
-	Chinese Academy of Sciences	N	Mandarin	Y	2	47.8k	8.8k	39k	-	1.7k	-
-	Fluent.ai	N	English	Y	1	50 h	5.9k	-	22 h	1.6k	-
-	Google	N	English	Y	10	>3k h	60.7k	133k	81.2k	11.2k	70k
-	Google	N	English	Y	14	326.8k	10k	316.8k	61.3k	1.9k	59.4k
-	Harbin Institute of Technology	N	Mandarin	-	1	115.2k	19.2k	96k	28.8k	4.8k	24k
-	Logitech	N	English	-	14	-	-	-	-	-	-
-	Mobvoi	N	Mandarin	Y	1	67 h	20k	54k	7 h	2k	5.9k
-	Sonos	Y	English	Y	16	0	0	0	1.1k	1.1k	0
-	Tencent	N	Mandarin	Y	1	339 h	224k	100k	-	-	-
-	Tencent	N	Mandarin	Y	1	65.9 h	6.9 h	59 h	8.7 h	0.9 h	7.8 h
-	Tencent	N	Mandarin	Y	42	22.2k	15.4k	6.8k	10.8k	7.4k	3.4k
-	Xiaomi	N	Mandarin	-	1	1.7k h	188.9k	1M	52.2 h	28.8k	32.8k
AISHELL-2 (13)	AISHELL	Y	Mandarin	N	13	24.8 h	>24k	-	16.7 h	>8.4k	-
AISHELL-2 (20)	AISHELL	Y	Mandarin	N	20	35 h	>34k	-	23.9 h	>12k	-
"Alexa"	Amazon	N	English	Y	1	495 h	-	-	100 h	-	-
Google Speech Commands Dataset *v1*	Google	Y	English	Y	10	51.7k	18.9k	32.8k	6.5k	2.4k	4.1k
Google Speech Commands Dataset *v2*	Google	Y	English	Y	10	84.6k	30.8k	53.8k	10.6k	3.9k	6.7k
"Hey Siri"	Apple	N	English	Y	1	500k	250k	250k	-	6.5k	2.7k h
Hey Snapdragon Keyword Dataset	Qualcomm	Y	English	N	4	-	-	-	4.3k	4.3k	-
Hey Snips	Snips	Y	English	Y	1	50.5 h	5.9k	45.3k	23.1 h	2.6k	20.8k
"Narc Ya"	Netmarble	N	Korean	Y	1	130k	50k	80k	800	400	400
"Ok/Hey Google"	Google	N	English	Y	2	-	1M	-	>3k h	434k	213k
"Ok/Hey Google"	Google	N	English	Y	2	-	-	-	247 h	4.8k	7.5k
Ticmini2	Mobvoi	N	Mandarin	Y	2	157.5k	43.6k	113.9k	72.9k	21.3k	51.6k

- The great majority of datasets are **noisy** (*signals are distorted by, e.g., natural and realistic background acoustic noise or room acoustics*)

Datasets

We will want to minimize the mismatch between the KWS performance at the lab phase and that one observable in the inherently-noisy real-life conditions

- **1) Natural noisy speech:** Some datasets were created from natural noisy speech recorded, many times in far-field conditions, by smart speakers, smartphones and tablets
  - *Home environments with background music or TV sound*

- **2) Simulated noisy speech:** Some datasets were generated by artificially distorting clean speech signals through *data augmentation*
  - *Noise types:* Babble, café, car, music, street...
  - *SNR levels commonly within the range [-5, 20] dB (Filtering and Noise-adding Tool*[1]*)*
  - *Noise datasets: TUT, DEMAND, MUSAN, NOISEX-92, CHiME...*
  - *Alteration of room acoustics, e.g., to simulate far-field conditions from close-talk speech*

(1) H. G. Hirsch, "FaNT - Filtering and noise adding tool". `https://github.com/i3thuan5/FaNT`

Datasets

- Collecting a good amount of natural noisy speech data in the desired acoustic conditions is not always feasible!

Alternative

Simulation of noisy speech is a smart and cheaper alternative allowing us for obtaining similar technology performance(1)

(1) T. Ko et al., "A study on data augmentation of reverberant speech for robust speech recognition". In Proc. of ICASSP 2017

Datasets

Name	Developer	P.A.?	Language	Noisy?	No. of KW	Training set			Test set		
						Size	*+ sampl.*	*- sampl.*	*Size*	*+ sampl.*	*- sampl.*
-	Alibaba	N	Mandarin	Y	1	24k h	-	-	-	12k	600 h
-	Baidu	N	English	Y	1	12k	-	-	2k	-	-
-	Chinese Academy of Sciences	N	Mandarin	Y	2	47.8k	8.8k	39k	-	1.7k	-
-	Fluent.ai	N	English	Y	1	50 h	5.9k	-	22 h	1.6k	-
-	Google	N	English	Y	10	>3k h	60.7k	133k	81.2k	11.2k	70k
-	Google	N	English	Y	14	326.8k	10k	316.8k	61.3k	1.9k	59.4k
-	Harbin Institute of Technology	N	Mandarin	-	1	115.2k	19.2k	96k	28.8k	4.8k	24k
-	Logitech	N	English	-	14	-	-	-	-	-	-
-	Mobvoi	N	Mandarin	Y	1	67 h	20k	54k	7 h	2k	5.9k
-	Sonos	Y	English	Y	16	0	0	0	1.1k	1.1k	0
-	Tencent	N	Mandarin	Y	1	339 h	224k	100k	-	-	-
-	Tencent	N	Mandarin	Y	1	65.9 h	6.9 h	59 h	8.7 h	0.9 h	7.8 h
-	Tencent	N	Mandarin	Y	42	22.2k	15.4k	6.8k	10.8k	7.4k	3.4k
-	Xiaomi	N	Mandarin	-	1	1.7k h	188.9k	1M	52.2 h	28.8k	32.8k
AISHELL-2 (13)	AISHELL	Y	Mandarin	N	13	24.8 h	>24k	-	16.7 h	>8.4k	-
AISHELL-2 (20)	AISHELL	Y	Mandarin	N	20	35 h	>34k	-	23.9 h	>12k	-
"Alexa"	Amazon	N	English	Y	1	495 h	-	-	100 h	-	-
Google Speech Commands Dataset *v1*	Google	Y	English	Y	10	51.7k	18.9k	32.8k	6.5k	2.4k	4.1k
Google Speech Commands Dataset *v2*	Google	Y	English	Y	10	84.6k	30.8k	53.8k	10.6k	3.9k	6.7k
"Hey Siri"	Apple	N	English	Y	1	500k	250k	250k	-	6.5k	2.7k h
Hey Snapdragon Keyword Dataset	Qualcomm	Y	English	N	4	-	-	-	4.3k	4.3k	-
Hey Snips	Snips	Y	English	Y	1	50.5 h	5.9k	45.3k	23.1 h	2.6k	20.8k
"Narc Ya"	Netmarble	N	Korean	Y	1	130k	50k	80k	800	400	400
"Ok/Hey Google"	Google	N	English	Y	2	-	1M	-	>3k h	434k	213k
"Ok/Hey Google"	Google	N	English	Y	2	-	-	-	247 h	4.8k	7.5k
Ticmini2	Mobvoi	N	Mandarin	Y	2	157.5k	43.6k	113.9k	72.9k	21.3k	51.6k

- Datasets mainly fit the application of KWS that, lately, is boosting research on this technology: *wake-up word detection for voice assistants*

Datasets

Name	Developer	P.A.?	Language	Noisy?	No. of KW	Training set			Test set		
						Size	*+ sampl.*	*- sampl.*	*Size*	*+ sampl.*	*- sampl.*
-	Alibaba	N	Mandarin	Y	1	24k h	-	-	-	12k	600 h
-	Baidu	N	English	Y	1	12k	-	-	2k	-	-
-	Chinese Academy of Sciences	N	Mandarin	Y	2	47.8k	8.8k	39k	-	1.7k	-
-	Fluent.ai	N	English	Y	1	50 h	5.9k	-	22 h	1.6k	-
-	Google	N	English	Y	10	>3k h	60.7k	133k	81.2k	11.2k	70k
-	Google	N	English	Y	14	326.8k	10k	316.8k	61.3k	1.9k	59.4k
-	Harbin Institute of Technology	N	Mandarin	-	1	115.2k	19.2k	96k	28.8k	4.8k	24k
-	Logitech	N	English	-	14	-	-	-	-	-	-
-	Mobvoi	N	Mandarin	Y	1	67 h	20k	54k	7 h	2k	5.9k
-	Sonos	Y	English	Y	16	0	0	0	1.1k	1.1k	0
-	Tencent	N	Mandarin	Y	1	339 h	224k	100k	-	-	-
-	Tencent	N	Mandarin	Y	1	65.9 h	6.9 h	59 h	8.7 h	0.9 h	7.8 h
-	Tencent	N	Mandarin	Y	42	22.2k	15.4k	6.8k	10.8k	7.4k	3.4k
-	Xiaomi	N	Mandarin	-	1	1.7k h	188.9k	1M	52.2 h	28.8k	32.8k
AISHELL-2 (13)	AISHELL	Y	Mandarin	N	13	24.8 h	>24k	-	16.7 h	>8.4k	-
AISHELL-2 (20)	AISHELL	Y	Mandarin	N	20	35 h	>34k	-	23.9 h	>12k	-
"Alexa"	Amazon	N	English	Y	1	495 h	-	-	100 h	-	-
Google Speech Commands Dataset *v1*	Google	Y	English	Y	10	51.7k	18.9k	32.8k	6.5k	2.4k	4.1k
Google Speech Commands Dataset *v2*	Google	Y	English	Y	10	84.6k	30.8k	53.8k	10.6k	3.9k	6.7k
"Hey Siri"	Apple	N	English	Y	1	500k	250k	250k	-	6.5k	2.7k h
Hey Snapdragon Keyword Dataset	Qualcomm	Y	English	N	4	-	-	-	4.3k	4.3k	-
Hey Snips	Snips	Y	English	Y	1	50.5 h	5.9k	45.3k	23.1 h	2.6k	20.8k
"Narc Ya"	Netmarble	N	Korean	Y	1	130k	50k	80k	800	400	400
"Ok/Hey Google"	Google	N	English	Y	2	-	1M	-	>3k h	434k	213k
"Ok/Hey Google"	Google	N	English	Y	2	-	-	-	247 h	4.8k	7.5k
Ticmini2	Mobvoi	N	Mandarin	Y	2	157.5k	43.6k	113.9k	72.9k	21.3k	51.6k

- As a trend, publicly available datasets tend to be smaller than in-house ones

Datasets

Name	Developer	P.A.?	Language	Noisy?	No. of KW	Training set			Test set		
						Size	*+ sampl.*	*- sampl.*	*Size*	*+ sampl.*	*- sampl.*
-	Alibaba	N	Mandarin	Y	1	24k h	-	-	-	12k	600 h
-	Baidu	N	English	Y	1	12k	-	-	2k	-	-
-	Chinese Academy of Sciences	N	Mandarin	Y	2	47.8k	8.8k	39k	-	1.7k	-
-	Fluent.ai	N	English	Y	1	50 h	5.9k	-	22 h	1.6k	-
-	Google	N	English	Y	10	>3k h	60.7k	133k	81.2k	11.2k	70k
-	Google	N	English	Y	14	326.8k	10k	316.8k	61.3k	1.9k	59.4k
-	Harbin Institute of Technology	N	Mandarin	-	1	115.2k	19.2k	96k	28.8k	4.8k	24k
-	Logitech	N	English	-	14	-	-	-	-	-	-
-	Mobvoi	N	Mandarin	Y	1	67 h	20k	54k	7 h	2k	5.9k
-	Sonos	Y	English	Y	16	0	0	0	1.1k	1.1k	0
-	Tencent	N	Mandarin	Y	1	339 h	224k	100k	-	-	-
-	Tencent	N	Mandarin	Y	1	65.9 h	6.9 h	59 h	8.7 h	0.9 h	7.8 h
-	Tencent	N	Mandarin	Y	42	22.2k	15.4k	6.8k	10.8k	7.4k	3.4k
-	Xiaomi	N	Mandarin	-	1	1.7k h	188.9k	1M	52.2 h	28.8k	32.8k
AISHELL-2 (13)	AISHELL	Y	Mandarin	N	13	24.8 h	>24k	-	16.7 h	>8.4k	-
AISHELL-2 (20)	AISHELL	Y	Mandarin	N	20	35 h	>34k	-	23.9 h	>12k	-
"Alexa"	Amazon	N	English	Y	1	495 h	-	-	100 h	-	-
Google Speech Commands Dataset *v1*	Google	Y	English	Y	10	51.7k	18.9k	32.8k	6.5k	2.4k	4.1k
Google Speech Commands Dataset *v2*	Google	Y	English	Y	10	84.6k	30.8k	53.8k	10.6k	3.9k	6.7k
"Hey Siri"	Apple	N	English	Y	1	500k	250k	250k	-	6.5k	2.7k h
Hey Snapdragon Keyword Dataset	Qualcomm	Y	English	N	4	-	-	-	4.3k	4.3k	-
Hey Snips	Snips	Y	English	Y	1	50.5 h	5.9k	45.3k	23.1 h	2.6k	20.8k
"Narc Ya"	Netmarble	N	Korean	Y	1	130k	50k	80k	800	400	400
"Ok/Hey Google"	Google	N	English	Y	2	-	1M	-	>3k h	434k	213k
"Ok/Hey Google"	Google	N	English	Y	2	-	-	-	247 h	4.8k	7.5k
Ticmini2	Mobvoi	N	Mandarin	Y	2	157.5k	43.6k	113.9k	72.9k	21.3k	51.6k

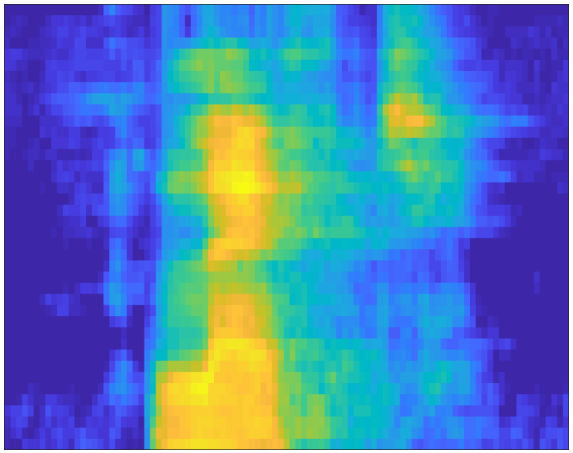
- $\frac{\textit{- sampl.}}{\textit{+ sampl.}} > 1$ is to accurately reflect potential scenarios of use consisting of always-on KWS applications like wake-up word detection

Datasets

Google Speech Commands Dataset

- The publicly available Google Speech Commands Dataset[(1)] has become the *de facto* open benchmark for (deep) KWS development and evaluation
  - *Sampling rate of 16 kHz*
  - *Recorded by phone and laptop microphones*
  - *Noisy to some extent*

One-second long speech segments covering one word each



Version	# of Speakers	# of Words	# of Utt.
v1	1,881	30	64,727
v2	2,618	35	105,829

(1) P. Warden, "Speech Commands: A dataset for limited-vocabulary speech recognition". *arXiv:1804.03209v1*, 2018

Version 1 ($v1$)	Version 2 ($v2$)	yes	no	up	down	left	KW
		right	on	off	stop	go	
		zero	one	two	three	four	Non-KW
		five	six	seven	eight	nine	
		bed	bird	cat	dog	happy	
		house	Marvin	Sheila	tree	wow	
		backward	forward	follow	learn	visual	

This benchmark also standardizes...

- ...the training, development and test sets
- ...a training data augmentation procedure involving background noises
- ...

Datasets

Google Speech Commands Dataset

We can raise two relevant points of criticism:

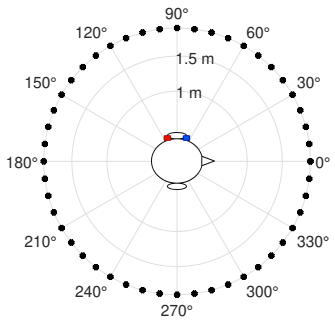
- **1) Class balancing:** The different keyword and non-keyword classes are rather balanced, which is generally not realistic
- **2) Non-streaming mode:** In multi-class classification of independent short input segments, a full keyword or non-keyword is surely present within every segment. However, real-life KWS involves the continuous processing of an input audio stream!

A few works generate streaming versions of this database by concatenation of one-second long utterances[(1)] in such a manner that the resulting word class distribution is unbalanced ⟵ *This point should be standardized for the sake of reproducibility and comparison!*

(1) I. López-Espejo et al., "A novel loss function and training strategy for noise-robust keyword spotting". IEEE/ACM TASLP, 2021

We produced three outcomes revolving around the Google Speech Commands Dataset *v2*:

- **1)** A variant of it emulating hearing aids as a capturing device(1)



- **2)** Another noisier variant with a diversity of noisy conditions(2) (i.e., types of noise and SNR levels)
- **3)** Manually-annotated speaker gender labels(3)

(1) I. López-Espejo et al., "Keyword spotting for hearing assistive devices robust to external speakers". In Proc. of Interspeech 2019
(2) `http://ilopez.es.mialias.net/misc/NoisyGSCD.zip`
(3) `https://ilopezes.files.wordpress.com/2019/10/gscd_spk_gender.zip`

Evaluation Metrics

- The gold plate test of any speech communication system is a test with relevant end-users $\longleftarrow$ Costly and time-consuming!
- Objective evaluation metrics must allow us to determine the goodness of a system and be highly correlated to the subjective user experience
- We review and provide some criticism of the most common **binary classification** metrics for KWS
- In the event of having multiple keywords, a common approach consists of applying the metric computation for every keyword and, then, the result is averaged

- **Accuracy:** The ratio between the number of correct predictions/classifications and the total number of them

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \in [0, 1]$$

- Accuracy tends to be an *unsuitable* evaluation metric yielding potentially misleading conclusions!

Ground truth	**NK**	**NK**	**KW**	**NK**	**NK**	**KW**	**NK**	**NK**	**NK**	**NK**
SYS1	**NK**	**NK**	**KW**	**NK**	**NK**	**NK**	**KW**	**NK**	**NK**	**NK**
SYS2	**NK**	**NK**	**NK**	**NK**	**NK**	**NK**	**NK**	**NK**	**NK**	**NK**

Evaluation Metrics

Accuracy

- Accuracy is a **widely** used evaluation metric for deep KWS $\longleftarrow$ *Google Speech Commands Dataset in non-streaming mode*
- Word classes are rather balanced in the Google Speech Commands Dataset $\longrightarrow$ Accuracy can still be considered a meaningful metric
- We have experimentally observed[(1)] for KWS a strong correlation between accuracy on a quite balanced scenario and more suitable metrics like F-score on a more realistic, unbalanced scenario
- Although not ideal, the employment of *accuracy can still be useful under certain experimental conditions*

(1) I. López-Espejo et al., "Improved external speaker-robust keyword spotting for hearing assistive devices". IEEE/ACM TASLP, 2020

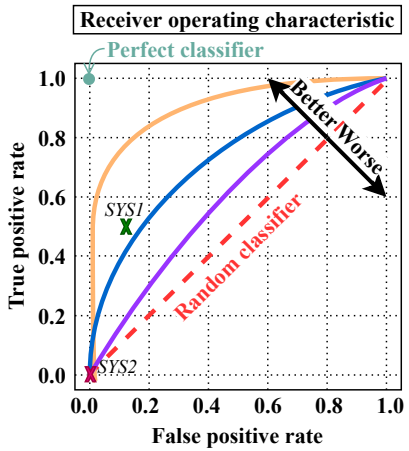
- The probability that a positive sample is correctly detected as such:

$$\text{True Positive Rate (TPR)} = \text{Recall} = \frac{TP}{TP + FN}$$

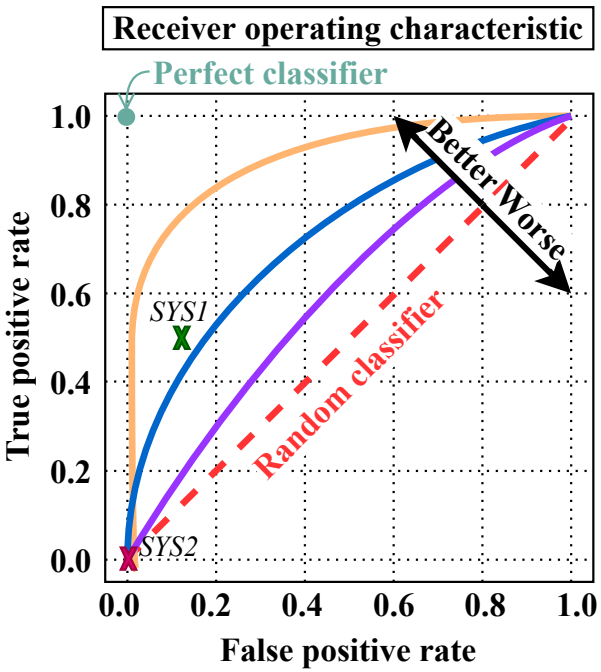
- The probability that a negative sample is wrongly classified as a positive one:

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN}$$

- The **receiver operating characteristic (ROC)** curve is obtained by sweeping the sensitivity (decision) threshold:



- **Area under the ROC curve ($\mathbf{AUC_{ROC}} \in [0, 1]$):** The probability that a classifier ranks a randomly-chosen positive sample higher than a randomly-chosen negative one

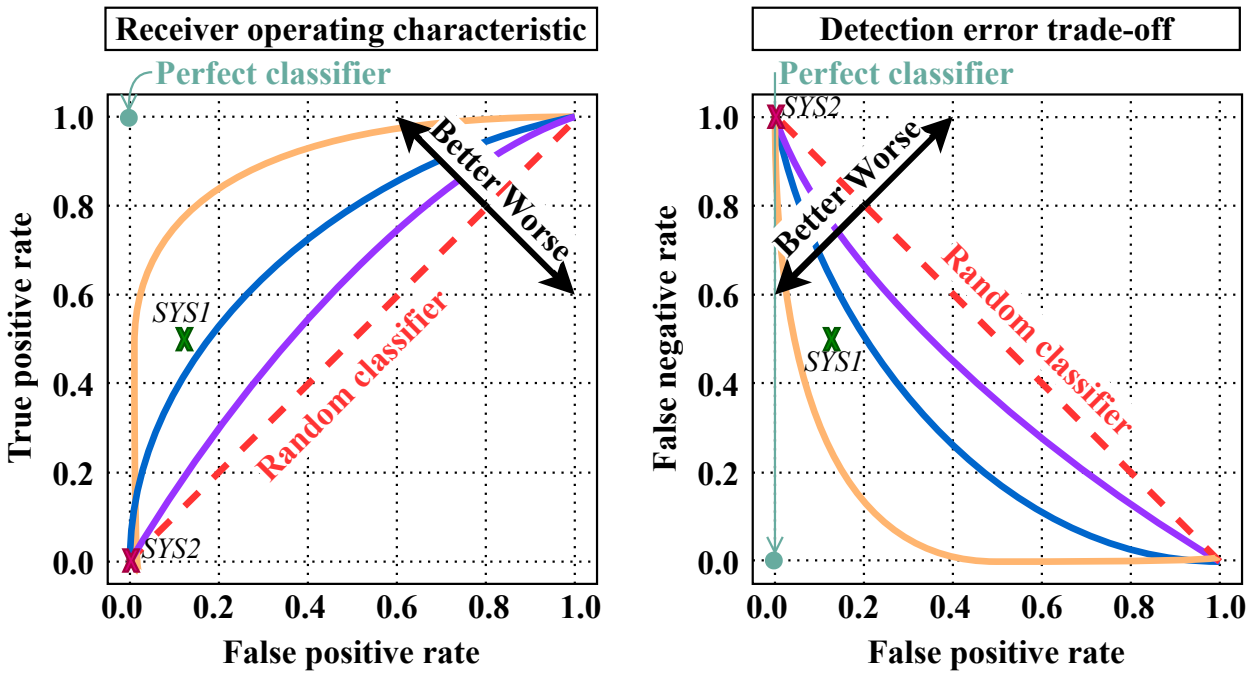


Ground truth	NK	NK	KW	NK	NK	KW	NK	NK	NK	NK
SYS1	NK	NK	KW	NK	NK	NK	KW	NK	NK	NK
SYS2	NK	NK	NK	NK	NK	NK	NK	NK	NK	NK

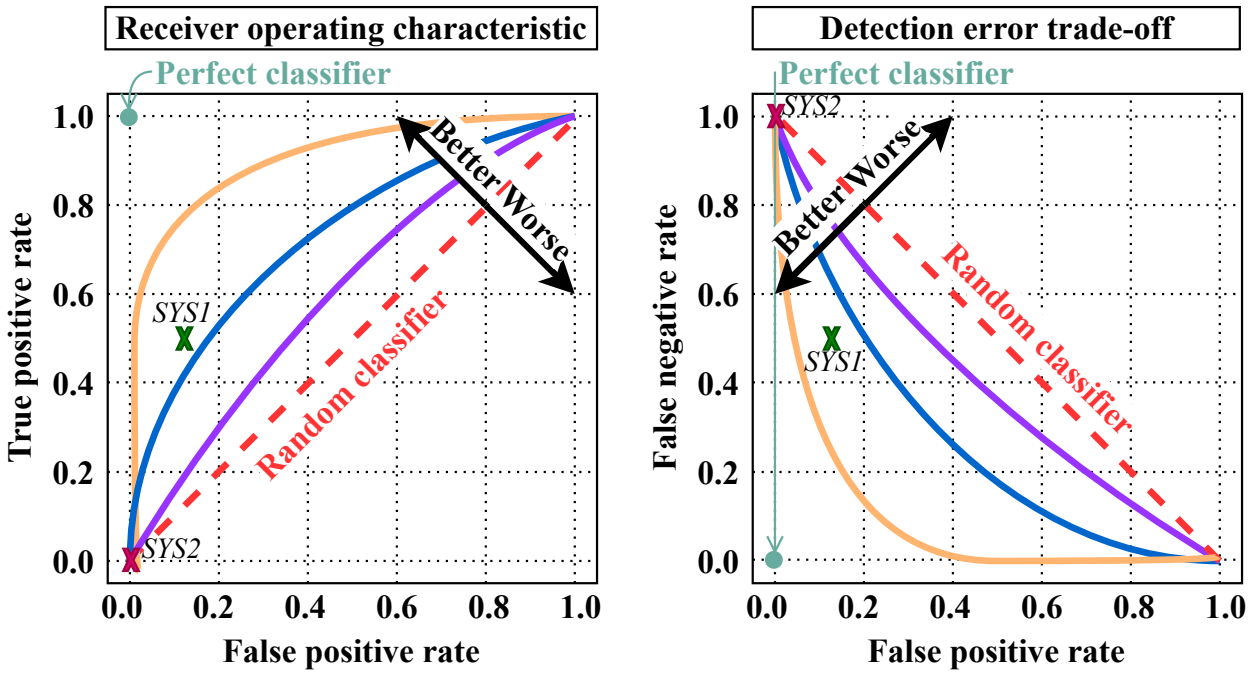
- Because

$$\text{False Negative Rate (FNR)} = \frac{FN}{FN + TP} = 1 - \text{TPR},$$

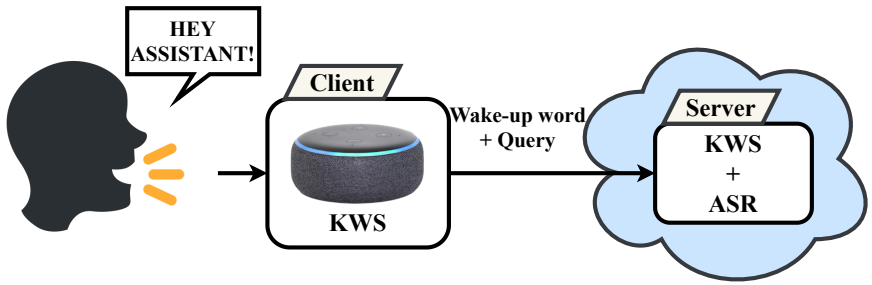
the **detection error trade-off (DET)** curve is nothing else but a vertically-flipped version of the ROC curve:



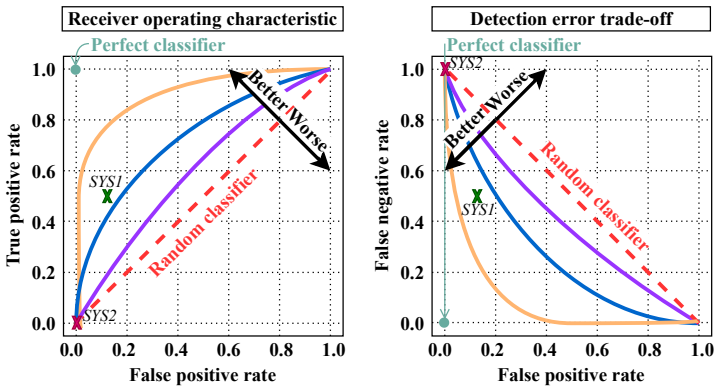
- **Area under the DET curve ($\mathbf{AUC_{DET}} \in [0, 1]$):** The smaller, the better
- **Equal error rate (EER):** The intersection point between the identity line and the DET curve (i.e., the point at which FNR = FPR)



- In voice activation of voice assistants, **privacy** is a major concern $\longrightarrow$ EER is not a good metric, since the cost of a false alarm is significantly greater than that of a miss detection!



- A popular variant of the ROC and DET curves is that one replacing FPR along the x-axis by the *number of false alarms per hour*



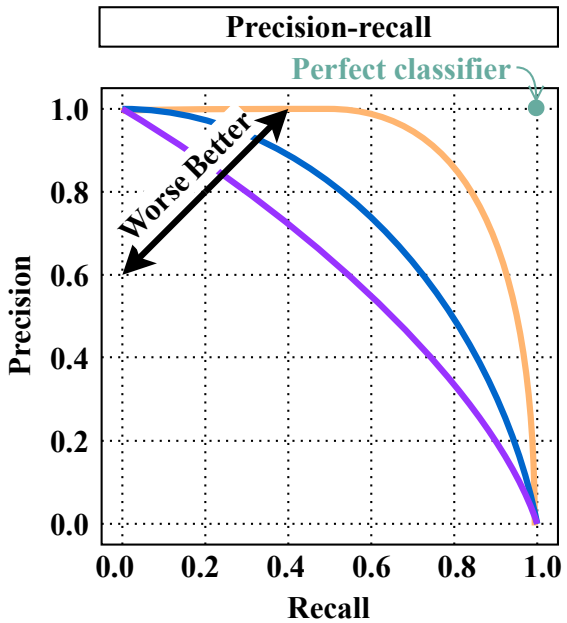
Term-weighted value (TWV)

$$\text{TWV} = 1 - (\text{FNR} + \beta\text{FPR}), \quad \beta \gg 1 \text{ (e.g., } \beta = 999.9)$$

- The probability that a sample that is classified as positive is actually a positive sample:

$$\text{Precision} = \frac{TP}{TP + FP}$$

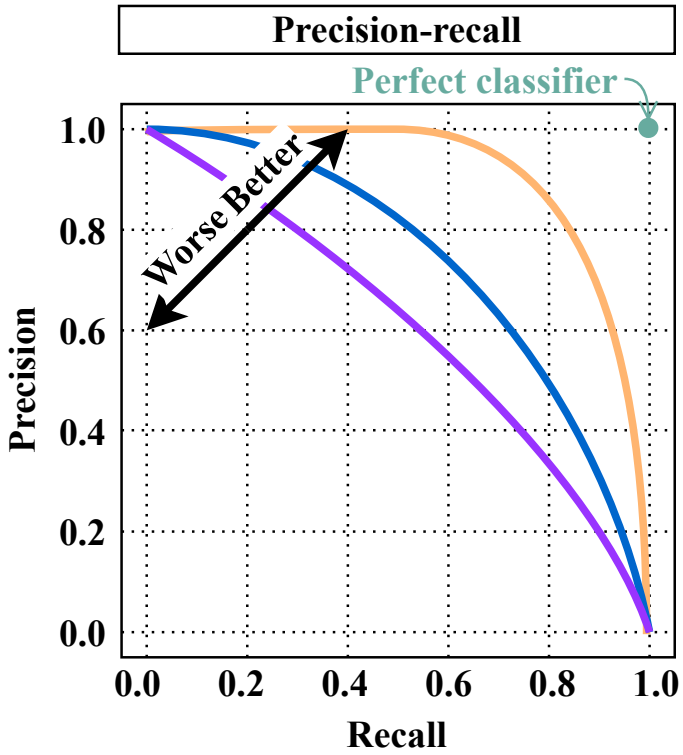
- The **precision-recall (PR)** curve is again obtained by sweeping the sensitivity threshold:



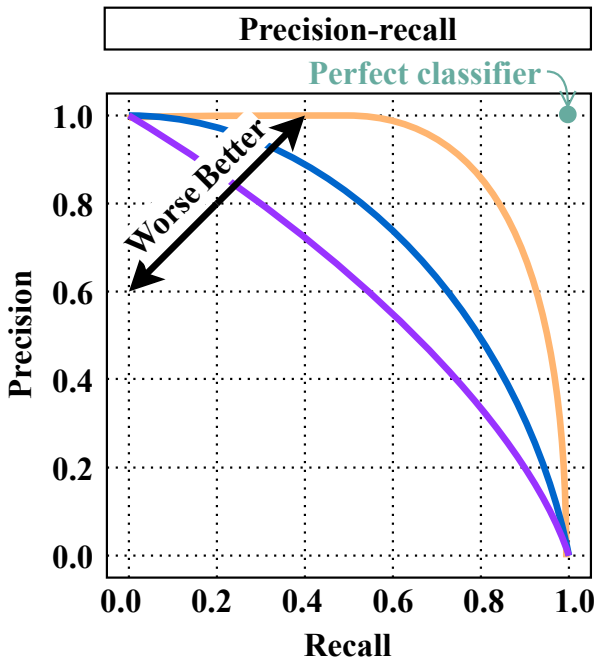
Reminder:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **Area under the PR curve ($\mathbf{AUC_{PR}} \in [0, 1]$):** The larger, the better
- The random guessing line depends on the proportion of the positive class within both classes
  - *Balanced scenario: Horizontal line at a precision of 0.5*
  - *In KWS: Horizontal line closer to 0*



- One-to-one correspondence between the PR and ROC/DET curves[(1)]
- However, the PR curve is considered to be a more informative visual analysis tool ⟵ We can better focus on the *minority* positive (i.e., keyword) class of interest (Precision $= TP/(TP + FP)$)



Ground truth	NK	NK	KW	NK	NK	KW	NK	NK	NK	NK
SYS1	NK	NK	KW	NK	NK	NK	KW	NK	NK	NK
SYS2	NK	NK	NK	NK	NK	NK	NK	NK	NK	NK

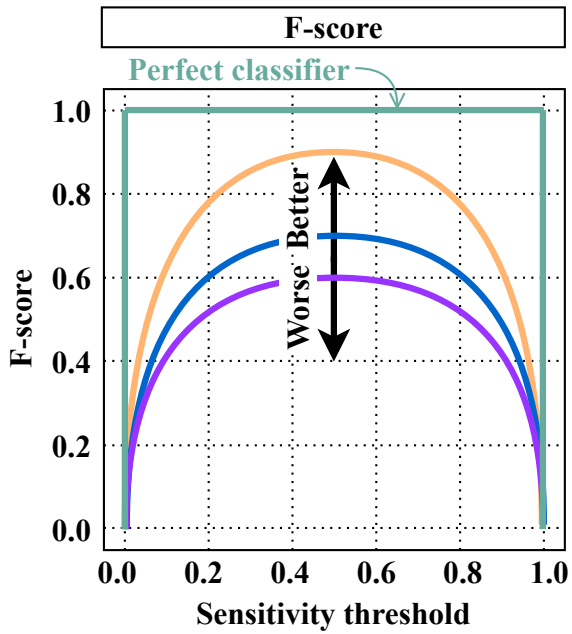
SYS1 ⟶ (Recall $= 0.5$, Precision $= 0.5$)
SYS2 ⟶ Precision $= 0/0$ is undefined

(1) J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves". In Proc. of ICML 2006

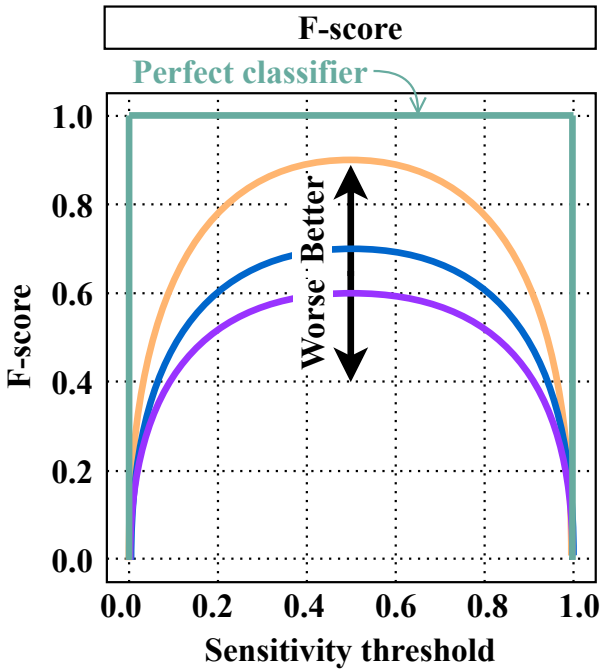
- F-score, F_1, is the harmonic mean of precision and recall:

$$F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}} = \frac{2TP}{2TP + FP + FN}$$

- The larger $F_1 \in [0, 1]$, the better
- F-score can be calculated as a function of the sensitivity threshold:



- **Area under the F_1 curve ($\text{AUC}_{F_1} \in [0, 1]$):** The larger, the better
- As for the PR curve, the random guessing line depends on the proportion of the positive class within both classes



Ground truth	NK	NK	KW	NK	NK	KW	NK	NK	NK	NK
SYS1	NK	NK	KW	NK	NK	NK	KW	NK	NK	NK
SYS2	NK	NK	NK	NK	NK	NK	NK	NK	NK	NK

$SYS1 \longrightarrow F_1 = 0.5$
$SYS2 \longrightarrow F_1 = 0$

Performance Comparison

ID	Description	Year	Accuracy (%)		Computational complexity	
			GSCD *v1*	GSCD *v2*	No. of params.	No. of mults.
1	Standard FFNN with a pooling layer [32]	2020	91.2	90.6	447k	–
2	DenseNet with trainable window function and mixup data augmentation [67]	2018	92.8	–	–	–
3	Two-stage TDNN [58]	2018	94.3	–	251k	25.1M
4	CNN with striding [32]	2018	95.4	95.6	529k	–
5	BiLSTM with attention [133]	2018	95.6	96.9	202k	–
6	Residual CNN `res15` [30]	2018	95.8 ± 0.484	–	238k	894M
7	TDNN with shared weight self-attention [16]	2019	95.81 ± 0.191	–	12k	403k
8	DenseNet+BiLSTM with attention [48]	2019	96.2	97.3	223k	–
9	Residual CNN with temporal convolutions `TC-ResNet14` [50]	2019	96.2	–	137k	–
10	SVDF [32]	2019	96.3	96.9	354k	–
11	SincConv+(Grouped DS-CNN) [70]	2020	96.4	97.3	62k	–
12	Graph convolutional network `CENet-40` [49]	2019	96.4	–	61k	16.18M
13	GRU [32]	2020	96.6	97.2	593k	–
14	SincConv+(DS-CNN) [70]	2020	96.6	97.4	122k	–
15	Temporal CNN with depthwise convolutions `TENet12` [52]	2020	96.6	–	100k	2.90M
16	Residual DS-CNN with squeeze-and-excitation `DS-ResNet18` [51]	2020	96.71 ± 0.195	–	72k	285M
17	`TC-ResNet14` with neural architecture search `NoisyDARTS-TC14` [146]	2021	96.79 ± 0.30	97.18 ± 0.26	108k	6.3M
18	LSTM [32]	2020	96.9	97.5	–	–
19	DS-CNN with striding [32]	2018	97.0	97.1	485k	–
20	CRNN [32]	2020	97.0	97.5	467k	–
21	BiGRU with multi-head attention [32]	2020	97.2	98.0	743k	–
22	CNN with neural architecture search `NAS2_6_36` [125]	2020	97.22	–	886k	–
23	Keyword Transformer `KWT-3` [90]	2021	97.49 ± 0.15	98.56 ± 0.07	5.3M	–
24	Variant of TC-ResNet with self-attention `LG-Net6` [91]	2021	97.67	96.79	313k	–
25	Broadcasted residual CNN `BC-ResNet-8` [100]	2021	98.0	98.7	321k	89.1M

- Google Speech Commands Dataset *v1* and *v2* in non-streaming mode

Performance Comparison

ID	Description	Year	Accuracy (%)		Computational complexity	
			GSCD *v1*	GSCD *v2*	No. of params.	No. of mults.
1	Standard FFNN with a pooling layer [32]	2020	91.2	90.6	447k	–
2	DenseNet with trainable window function and mixup data augmentation [67]	2018	92.8	–	–	–
3	Two-stage TDNN [58]	2018	94.3	–	251k	25.1M
4	CNN with striding [32]	2018	95.4	95.6	529k	–
5	BiLSTM with attention [133]	2018	95.6	96.9	202k	–
6	Residual CNN `res15` [30]	2018	95.8 ± 0.484	–	238k	894M
7	TDNN with shared weight self-attention [16]	2019	95.81 ± 0.191	–	12k	403k
8	DenseNet+BiLSTM with attention [48]	2019	96.2	97.3	223k	–
9	Residual CNN with temporal convolutions `TC-ResNet14` [50]	2019	96.2	–	137k	–
10	SVDF [32]	2019	96.3	96.9	354k	–
11	SincConv+(Grouped DS-CNN) [70]	2020	96.4	97.3	62k	–
12	Graph convolutional network `CENet-40` [49]	2019	96.4	–	61k	16.18M
13	GRU [32]	2020	96.6	97.2	593k	–
14	SincConv+(DS-CNN) [70]	2020	96.6	97.4	122k	–
15	Temporal CNN with depthwise convolutions `TENet12` [52]	2020	96.6	–	100k	2.90M
16	Residual DS-CNN with squeeze-and-excitation `DS-ResNet18` [51]	2020	96.71 ± 0.195	–	72k	285M
17	`TC-ResNet14` with neural architecture search `NoisyDARTS-TC14` [146]	2021	96.79 ± 0.30	97.18 ± 0.26	108k	6.3M
18	LSTM [32]	2020	96.9	97.5	–	–
19	DS-CNN with striding [32]	2018	97.0	97.1	485k	–
20	CRNN [32]	2020	97.0	97.5	467k	–
21	BiGRU with multi-head attention [32]	2020	97.2	98.0	743k	–
22	CNN with neural architecture search `NAS2_6_36` [125]	2020	97.22	–	886k	–
23	Keyword Transformer `KWT-3` [90]	2021	97.49 ± 0.15	98.56 ± 0.07	5.3M	–
24	Variant of TC-ResNet with self-attention `LG-Net6` [91]	2021	97.67	96.79	313k	–
25	Broadcasted residual CNN `BC-ResNet-8` [100]	2021	98.0	98.7	321k	89.1M

- The number of parameters and multiplications of the acoustic model are solid proxies predicting the power consumption of these systems(1)

(1) R. Tang et al., "An experimental analysis of the power consumption of convolutional neural networks for keyword spotting". In Proc. of ICASSP 2018

Performance Comparison

ID	Description	Year	Accuracy (%)		Computational complexity	
			GSCD *v1*	GSCD *v2*	No. of params.	No. of mults.
1	Standard FFNN with a pooling layer [32]	2020	91.2	90.6	447k	–
2	DenseNet with trainable window function and mixup data augmentation [67]	2018	92.8	–	–	–
3	Two-stage TDNN [58]	2018	94.3	–	251k	25.1M
4	CNN with striding [32]	2018	95.4	95.6	529k	–
5	BiLSTM with attention [133]	2018	95.6	96.9	202k	–
6	Residual CNN `res15` [30]	2018	95.8 ± 0.484	–	238k	894M
7	TDNN with shared weight self-attention [16]	2019	95.81 ± 0.191	–	12k	403k
8	DenseNet+BiLSTM with attention [48]	2019	96.2	97.3	223k	–
9	Residual CNN with temporal convolutions `TC-ResNet14` [50]	2019	96.2	–	137k	–
10	SVDF [32]	2019	96.3	96.9	354k	–
11	SincConv+(Grouped DS-CNN) [70]	2020	96.4	97.3	62k	–
12	Graph convolutional network `CENet-40` [49]	2019	96.4	–	61k	16.18M
13	GRU [32]	2020	96.6	97.2	593k	–
14	SincConv+(DS-CNN) [70]	2020	96.6	97.4	122k	–
15	Temporal CNN with depthwise convolutions `TENet12` [52]	2020	96.6	–	100k	2.90M
16	Residual DS-CNN with squeeze-and-excitation `DS-ResNet18` [51]	2020	96.71 ± 0.195	–	72k	285M
17	`TC-ResNet14` with neural architecture search `NoisyDARTS-TC14` [146]	2021	96.79 ± 0.30	97.18 ± 0.26	108k	6.3M
18	LSTM [32]	2020	96.9	97.5	–	–
19	DS-CNN with striding [32]	2018	97.0	97.1	485k	–
20	CRNN [32]	2020	97.0	97.5	467k	–
21	BiGRU with multi-head attention [32]	2020	97.2	98.0	743k	–
22	CNN with neural architecture search `NAS2_6_36` [125]	2020	97.22	–	886k	–
23	Keyword Transformer `KWT-3` [90]	2021	97.49 ± 0.15	98.56 ± 0.07	5.3M	–
24	Variant of TC-ResNet with self-attention `LG-Net6` [91]	2021	97.67	96.79	313k	–
25	Broadcasted residual CNN `BC-ResNet-8` [100]	2021	98.0	98.7	321k	89.1M

- The second version of this dataset has more word samples $\longrightarrow$ Better trained acoustic models

Performance Comparison

ID	Description	Year	Accuracy (%)		Computational complexity	
			GSCD *v1*	GSCD *v2*	No. of params.	No. of mults.
1	Standard FFNN with a pooling layer [32]	2020	91.2	90.6	447k	–
2	DenseNet with trainable window function and mixup data augmentation [67]	2018	92.8	–	–	–
3	Two-stage TDNN [58]	2018	94.3	–	251k	25.1M
4	CNN with striding [32]	2018	95.4	95.6	529k	–
5	BiLSTM with attention [133]	2018	95.6	96.9	202k	–
6	Residual CNN `res15` [30]	2018	95.8 ± 0.484	–	238k	894M
7	TDNN with shared weight self-attention [16]	2019	95.81 ± 0.191	–	12k	403k
8	DenseNet+BiLSTM with attention [48]	2019	96.2	97.3	223k	–
9	Residual CNN with temporal convolutions `TC-ResNet14` [50]	2019	96.2	–	137k	–
10	SVDF [32]	2019	96.3	96.9	354k	–
11	SincConv+(Grouped DS-CNN) [70]	2020	96.4	97.3	62k	–
12	Graph convolutional network `CENet-40` [49]	2019	96.4	–	61k	16.18M
13	GRU [32]	2020	96.6	97.2	593k	–
14	SincConv+(DS-CNN) [70]	2020	96.6	97.4	122k	–
15	Temporal CNN with depthwise convolutions `TENet12` [52]	2020	96.6	–	100k	2.90M
16	Residual DS-CNN with squeeze-and-excitation `DS-ResNet18` [51]	2020	96.71 ± 0.195	–	72k	285M
17	`TC-ResNet14` with neural architecture search `NoisyDARTS-TC14` [146]	2021	96.79 ± 0.30	97.18 ± 0.26	108k	6.3M
18	LSTM [32]	2020	96.9	97.5	–	–
19	DS-CNN with striding [32]	2018	97.0	97.1	485k	–
20	CRNN [32]	2020	97.0	97.5	467k	–
21	BiGRU with multi-head attention [32]	2020	97.2	98.0	743k	–
22	CNN with neural architecture search `NAS2_6_36` [125]	2020	97.22	–	886k	–
23	Keyword Transformer `KWT-3` [90]	2021	97.49 ± 0.15	98.56 ± 0.07	5.3M	–
24	Variant of TC-ResNet with self-attention `LG-Net6` [91]	2021	97.67	96.79	313k	–
25	Broadcasted residual CNN `BC-ResNet-8` [100]	2021	98.0	98.7	321k	89.1M

- The most frequently used acoustic model type is based on **CNN** ⟵ Highly competitive performance and lesser computational complexity!

Performance Comparison

ID	Description	Year	Accuracy (%)		Computational complexity	
			GSCD *v1*	GSCD *v2*	No. of params.	No. of mults.
1	Standard FFNN with a pooling layer [32]	2020	91.2	90.6	447k	–
2	DenseNet with trainable window function and mixup data augmentation [67]	2018	92.8	–	–	–
3	Two-stage TDNN [58]	2018	94.3	–	251k	25.1M
4	CNN with striding [32]	2018	95.4	95.6	529k	–
5	BiLSTM with attention [133]	2018	95.6	96.9	202k	–
6	Residual CNN `res15` [30]	2018	95.8 ± 0.484	–	238k	894M
7	TDNN with shared weight self-attention [16]	2019	95.81 ± 0.191	–	12k	403k
8	DenseNet+BiLSTM with attention [48]	2019	96.2	97.3	223k	–
9	Residual CNN with temporal convolutions `TC-ResNet14` [50]	2019	96.2	–	137k	–
10	SVDF [32]	2019	96.3	96.9	354k	–
11	SincConv+(Grouped DS-CNN) [70]	2020	96.4	97.3	62k	–
12	Graph convolutional network `CENet-40` [49]	2019	96.4	–	61k	16.18M
13	GRU [32]	2020	96.6	97.2	593k	–
14	SincConv+(DS-CNN) [70]	2020	96.6	97.4	122k	–
15	Temporal CNN with depthwise convolutions `TENet12` [52]	2020	96.6	–	100k	2.90M
16	Residual DS-CNN with squeeze-and-excitation `DS-ResNet18` [51]	2020	96.71 ± 0.195	–	72k	285M
17	`TC-ResNet14` with neural architecture search `NoisyDARTS-TC14` [146]	2021	96.79 ± 0.30	97.18 ± 0.26	108k	6.3M
18	LSTM [32]	2020	96.9	97.5	–	–
19	DS-CNN with striding [32]	2018	97.0	97.1	485k	–
20	CRNN [32]	2020	97.0	97.5	467k	–
21	BiGRU with multi-head attention [32]	2020	97.2	98.0	743k	–
22	CNN with neural architecture search `NAS2_6_36` [125]	2020	97.22	–	886k	–
23	Keyword Transformer `KWT-3` [90]	2021	97.49 ± 0.15	98.56 ± 0.07	5.3M	–
24	Variant of TC-ResNet with self-attention `LG-Net6` [91]	2021	97.67	96.79	313k	–
25	Broadcasted residual CNN `BC-ResNet-8` [100]	2021	98.0	98.7	321k	89.1M

- Neural architecture search (17-`NoisyDARTS-TC14` vs. 9-`TC-ResNet14`; 22-`NAS2_6_36`)

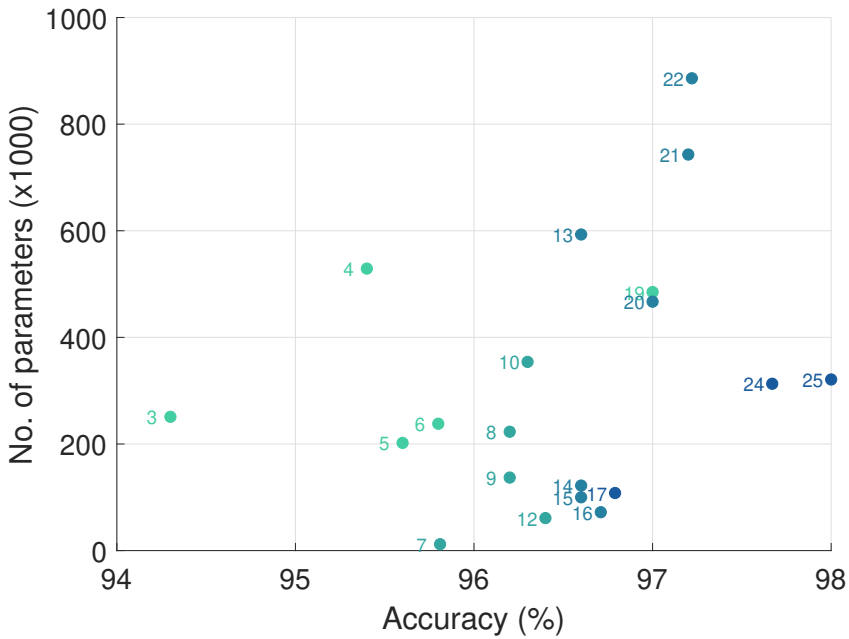
Performance Comparison

ID	Description	Year	Accuracy (%)		Computational complexity	
			GSCD *v1*	GSCD *v2*	No. of params.	No. of mults.
1	Standard FFNN with a pooling layer [32]	2020	91.2	90.6	447k	–
2	DenseNet with trainable window function and mixup data augmentation [67]	2018	92.8	–	–	–
3	Two-stage TDNN [58]	2018	94.3	–	251k	25.1M
4	CNN with striding [32]	2018	95.4	95.6	529k	–
5	BiLSTM with attention [133]	2018	95.6	96.9	202k	–
6	Residual CNN `res15` [30]	2018	95.8 ± 0.484	–	238k	894M
7	TDNN with shared weight self-attention [16]	2019	95.81 ± 0.191	–	12k	403k
8	DenseNet+BiLSTM with attention [48]	2019	96.2	97.3	223k	–
9	Residual CNN with temporal convolutions `TC-ResNet14` [50]	2019	96.2	–	137k	–
10	SVDF [32]	2019	96.3	96.9	354k	–
11	SincConv+(Grouped DS-CNN) [70]	2020	96.4	97.3	62k	–
12	Graph convolutional network `CENet-40` [49]	2019	96.4	–	61k	16.18M
13	GRU [32]	2020	96.6	97.2	593k	–
14	SincConv+(DS-CNN) [70]	2020	96.6	97.4	122k	–
15	Temporal CNN with depthwise convolutions `TENet12` [52]	2020	96.6	–	100k	2.90M
16	Residual DS-CNN with squeeze-and-excitation `DS-ResNet18` [51]	2020	96.71 ± 0.195	–	72k	285M
17	`TC-ResNet14` with neural architecture search `NoisyDARTS-TC14` [146]	2021	96.79 ± 0.30	97.18 ± 0.26	108k	6.3M
18	LSTM [32]	2020	96.9	97.5	–	–
19	DS-CNN with striding [32]	2018	97.0	97.1	485k	–
20	CRNN [32]	2020	97.0	97.5	467k	–
21	BiGRU with multi-head attention [32]	2020	97.2	98.0	743k	–
22	CNN with neural architecture search `NAS2_6_36` [125]	2020	97.22	–	886k	–
23	Keyword Transformer `KWT-3` [90]	2021	97.49 ± 0.15	98.56 ± 0.07	5.3M	–
24	Variant of TC-ResNet with self-attention `LG-Net6` [91]	2021	97.67	96.79	313k	–
25	Broadcasted residual CNN `BC-ResNet-8` [100]	2021	98.0	98.7	321k	89.1M

- Effectiveness of **CRNNs** (8-DenseNet+BiLSTM with attention vs. 2-DenseNet and 5-BiLSTM with attention; 20-CRNN)

Performance Comparison

On the Google Speech Commands Dataset *v1*

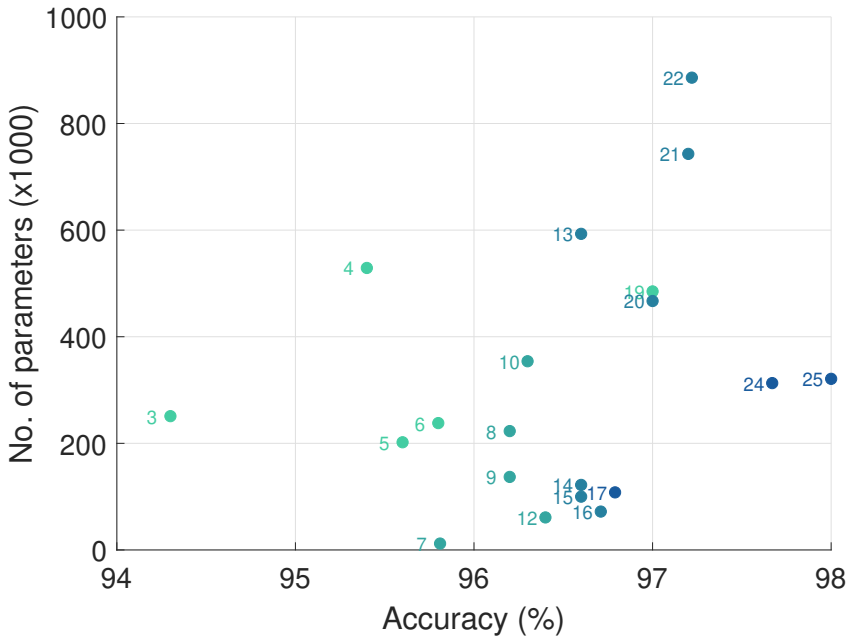


- **Systems with IDs 14, 15, 16 and 17:** Based on CNNs
  - *Most of them integrate residual connections and/or depthwise separable convolutions*
  - *Systems 15, 16 and 17 integrate either dilated or temporal convolutions to exploit long time-frequency dependencies*

- **Systems with IDs 24 and 25:** Based on CNNs with residual connections
  - *System 24 has temporal convolutions and self-attention layers, and System 25 has dilated convolutions*
  - *System 25 incorporates depthwise separable convolutions*

Performance Comparison

On the Google Speech Commands Dataset *v1*



A **state-of-the-art** KWS system comprising a **CNN acoustic model** should cover...

- ☑ A mechanism to exploit long time-frequency dependencies
- ☑ Depthwise separable convolutions
- ☑ Residual connections

DEEP SPOKEN KEYWORD SPOTTING

7. Experimental Considerations

Iván López-Espejo

Department of Electronic Systems, Aalborg University, Denmark

ivl@es.aau.dk

Sunday 18th September, 2022

INTERSPEECH 2022
September 18 - 22 • Incheon Korea

DEEP SPOKEN KEYWORD SPOTTING

8. Conclusions and Future Directions

Iván López-Espejo

Department of Electronic Systems, Aalborg University, Denmark

ivl@es.aau.dk

Sunday 18th September, 2022

INTERSPEECH 2022
September 18 - 22 • Incheon Korea

Conclusions and Future Directions

Conclusions

- Speech features $\longrightarrow$ DNN-based acoustic model (**core**) $\longrightarrow$ Posterior probability processing
- Deep spoken KWS has revitalized KWS research by enabling a massive deployment of this technology for real-world applications (e.g., **voice assistant activation**)
- Advances in ASR research will continue impacting the field of KWS (e.g., optimal feature learning for end-to-end ASR)

- Advancing **acoustic modeling** towards two goals simultaneously:
  1) Improving KWS performance in real-life acoustic conditions
  2) Computational complexity reduction
- Development of novel and efficient **convolutional blocks**
- Neural architecture search
- **Acoustic model compression** (parameter quantization, network pruning, knowledge distillation...):
  1) Reduced memory footprint
  2) Decreased inference latency
  3) Less energy consumption

- Semi-supervised learning for KWS:
  - *Industrial environment*
  - *Hybrid learning based on both small (labeled) and big (unlabeled) volumes of data*

- Personalization:
  1) Efficient open-vocabulary (personalized) KWS
  2) Joint KWS and speaker verification

- Multi-channel KWS for robustness purposes

DEEP SPOKEN KEYWORD SPOTTING

8. Conclusions and Future Directions

Iván López-Espejo

Department of Electronic Systems, Aalborg University, Denmark

ivl@es.aau.dk

Sunday 18th September, 2022

INTERSPEECH 2022
September 18 - 22 • Incheon Korea