Deep Spoken Keyword Spotting

Iván López-Espejo[1], Zheng-Hua Tan[1] and Jesper Jensen[1,2]

[1]Department of Electronic Systems, Aalborg University, Denmark
[2]Oticon A/S, Denmark
{ivl,zt,jje}@es.aau.dk, jesj@oticon.com

AALBORG UNIVERSITY
DENMARK

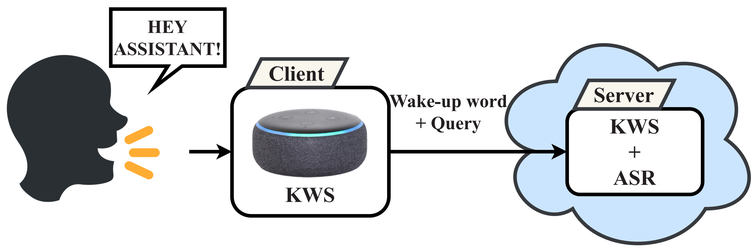
CASPR
Centre for Acoustic Signal Processing Research

Introduction

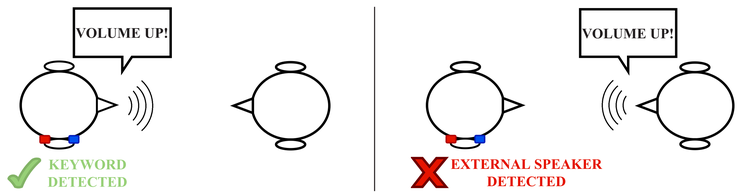
- **Keyword spotting (KWS):** Task of identifying keywords in audio streams comprising speech
- **Different KWS paradigms:** *1)* Large-vocabulary continuous speech recognition, *2)* keyword/filler hidden Markov model, and *3)* deep spoken KWS
- **Deep spoken KWS:**
  1. Simpler posterior handling instead of Viterbi decoding
  2. Easily adjustable DNN acoustic model complexity
  3. Superior performance in both clean and noisy conditions
- Deep spoken KWS is very appealing to be deployed on a variety of *consumer electronics with limited resources* like earphones, smartphones and smart speakers

Applications

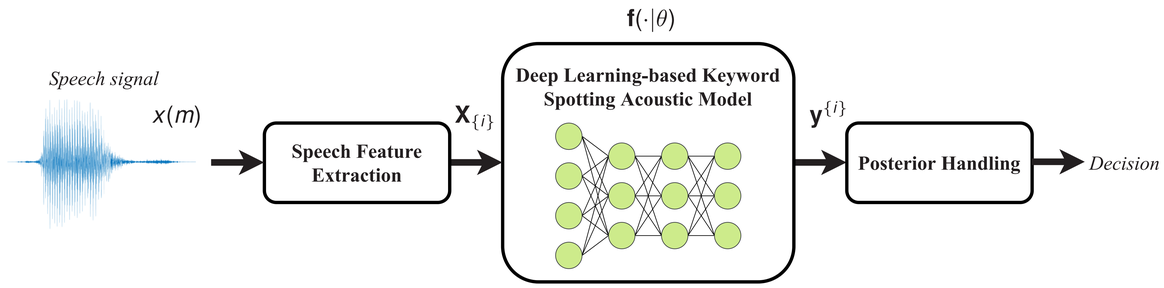
- Voice-dialing, interaction with a call center, speech retrieval, voice control of videogames and home automation, etc.
- **Personalized applications** by joint KWS and speaker verification
- **Activation of voice assistants** (*flagship application*):

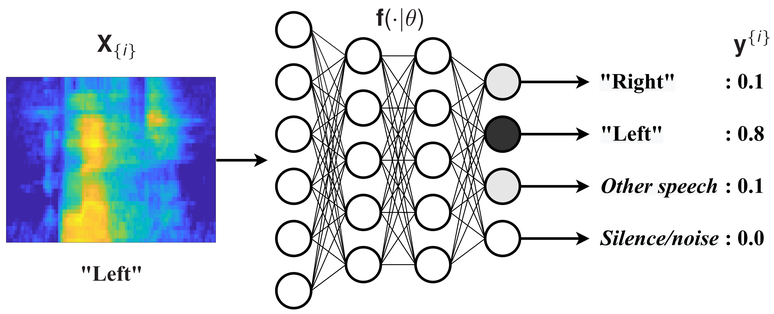


- Voice control of hearing assistive devices:



Deep Spoken Keyword Spotting Approach



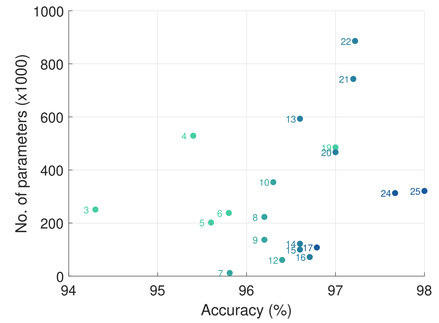


- **Mel-scale-related features** are, by far, the most widely used speech features in deep spoken KWS
- Many types of acoustic models have been studied: *Fully-connected feedforward neural networks (FFNNs)*, **convolutional neural networks (CNNs)**, *recurrent neural networks (RNNs)*...
- Non-streaming (static) and streaming (dynamic) posterior handling

Performance Comparison and Conclusions

Performance on the Google Speech Commands Dataset (**GSCD**) *v1*

Description	Year	Accuracy (%) GSCD *v1*	Complexity No. of params.
1 Standard FFNN with a pooling layer	2020	91.2	447k
4 CNN with striding	2018	95.4	529k
5 Bidirectional long short-term memory (BiLSTM) with attention	2018	95.6	202k
6 Residual CNN `res15`	2018	95.8 ± 0.484	238k
7 Time-delay neural network with shared weight self-attention	2019	95.81 ± 0.191	12k
8 DenseNet+BiLSTM with attention	2019	96.2	223k
9 Residual CNN with temporal convolutions `TC-ResNet14`	2019	96.2	137k
10 Single value decomposition filter	2019	96.3	354k
13 Gated recurrent unit (GRU) RNN	2020	96.6	593k
14 **SincConv+(DS-CNN)**	2020	**96.6**	**122k**
15 **Temporal CNN with depthwise convolutions** `TENet12`	2020	**96.6**	**100k**
16 **Residual DS-CNN with squeeze-and-excitation** `DS-ResNet18`	2020	**96.71 ± 0.195**	**72k**
17 `TC-ResNet14` **with neural architecture search** `NoisyDARTS-TC14`	2021	**96.79 ± 0.30**	**108k**
18 LSTM	2020	96.9	–
19 DS-CNN with striding	2018	97.0	485k
20 Convolutional recurrent neural network	2020	97.0	467k
21 BiGRU with multi-head attention	2020	97.2	743k
22 CNN with neural architecture search `NAS2_6_36`	2020	97.22	886k
23 Keyword Transformer `KWT-3`	2021	97.49 ± 0.15	5.3M
24 **Variant of TC-ResNet with self-attention** `LG-Net6`	2021	**97.67**	**313k**
25 **Broadcasted residual CNN** `BC-ResNet-8`	2021	**98.0**	**321k**



- State-of-the-art acoustic modeling is based on **CNNs**
- To reach a *high performance* with a *small computational footprint*, a CNN acoustic model should cover...
  1. A mechanism to exploit long time-frequency dependencies (e.g., dilated or temporal convolutions)
  2. Depthwise separable (**DS**) convolutions
  3. Residual connections