An Experimental Study on Light Speech Features for Small-Footprint 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@demant.com

AALBORG UNIVERSITY
DENMARK

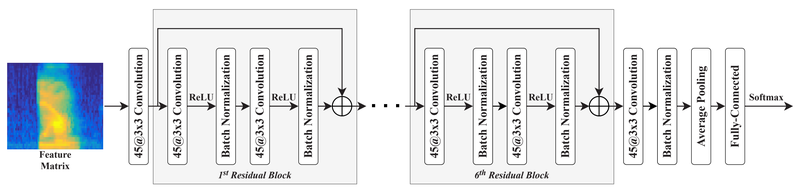
CASPR
Centre for Acoustic Signal Processing Research

Introduction

- **Keyword spotting (KWS)** enables voice interaction with small devices like smartphones, tablets and smartwatches
- *Due to computational and energy constraints*, embedding typical always-on KWS technology can pose a challenge
  1. Acoustic model parameter **quantization**
  2. **Reduction of the number of parameters and/or multiplications** of the acoustic model
- There is much **redundant information** in speech features fed into modern KWS acoustic models
- **Feature matrix size reduction** (2.) $\longrightarrow$ Remarkable **multiplication number reduction** in state-of-the-art acoustic models (*CNNs + residual connections*)

Experimental Framework

- We use the popular **Google Speech Commands Dataset** defining **10 keywords**: “yes”, “no”, “up”, “down”, “left”, “right”, “on”, “off”, “stop” and “go”
- Skip connections make the number of multiplications of the acoustic model heavily depend on the feature matrix size, since the successive feature maps have to preserve it for addition throughout the residual blocks

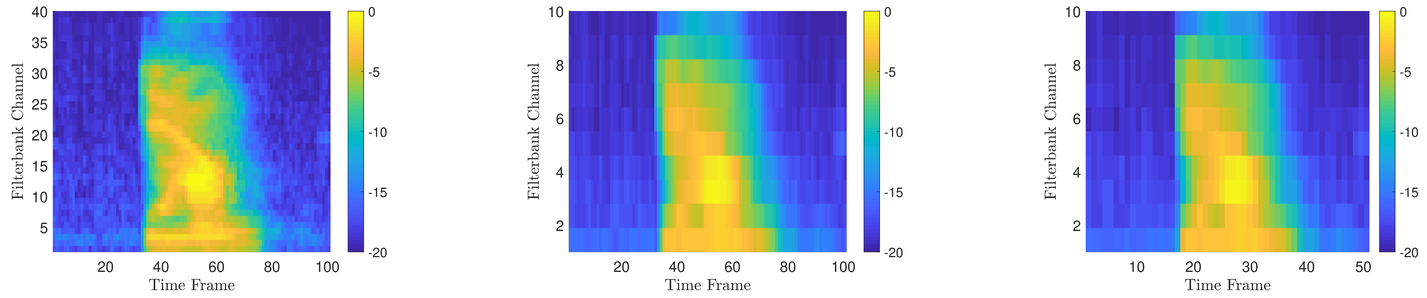


- **The final fully-connected layer has 11 nodes** corresponding to the 10 different keywords plus the non-keyword/filler class

Methodology

- **Log-Mel** and **Mel-frequency cepstral coefficients (MFCCs)**
- Smaller no. of filterbank channels/cepstral coefficients (**features**) and time resolution reduction by increasing the analysis window hop size

One-second long utterance comprising the word “*down*”



Results and Conclusions

The hop size/number of time frames is 10 ms/101:

No. of Features	No. of Mult.	Log-Mel			MFCC		
		Accuracy (%)	Training Time per Epoch (s)	Inference Time (μs)	Accuracy (%)	Training Time per Epoch (s)	Inference Time (μs)
40	895M	95.33 ± 0.28	98.7 ± 0.4	980 ± 7	95.24 ± 0.96	96.5 ± 0.4	980 ± 12
20	424M	95.70 ± 0.58	58.3 ± 0.7	590 ± 11	95.55 ± 0.65	55.2 ± 0.4	595 ± 7
10	188M	95.34 ± 0.76	36.2 ± 0.3	390 ± 21	95.24 ± 0.64	36.3 ± 0.3	385 ± 12
5	71M	93.00 ± 0.36	27.3 ± 0.5	292 ± 16	92.60 ± 0.87	26.3 ± 0.5	289 ± 10

The number of features is 10:

Hop Size (ms) / No. of Frames	No. of Mult.	Log-Mel			MFCC		
		Accuracy (%)	Training Time per Epoch (s)	Inference Time (μs)	Accuracy (%)	Training Time per Epoch (s)	Inference Time (μs)
10 / 101	188M	95.34 ± 0.76	36.2 ± 0.3	390 ± 21	95.24 ± 0.64	36.3 ± 0.3	385 ± 12
20 / 51	93M	94.63 ± 0.65	24.2 ± 0.5	265 ± 11	94.61 ± 0.89	24.1 ± 0.3	265 ± 9
30 / 34	61M	94.53 ± 0.47	19.2 ± 0.4	216 ± 10	93.50 ± 0.83	19.0 ± 0.4	216 ± 14
40 / 26	46M	93.24 ± 0.50	15.3 ± 0.4	197 ± 9	92.36 ± 0.55	15.2 ± 0.3	201 ± 8

- **Strong linear relationship** ($R^2 = 0.9641$, $p = 0.0001$) between the no. of multiplications of the acoustic model and its energy consumption
- The standard size of both **log-Mel and MFCC matrices can be reduced by a factor of 8** while essentially maintaining KWS performance
  1. $9.6\times$ number of multiplications/energy consumption reduction
  2. $4.0\times$ training time reduction
  3. $3.7\times$ inference time reduction
- This is a finding to bear in mind when designing *light and compact KWS systems* intended to be embedded on low-resource devices