End-to-End Framework for Keyword Spotting

Students:

Ivancich Stefano

Masiero Luca

Supervisors:

Prof. Rossi Michele

Meneghello Francesca

DEPARTMENT OF INFORMATION ENGINEERING

UNIVERSITY OF PADOVA

22 September 2020

Outline

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

- The Problem
- The Solution
- What we tried
- Architecture 1: 1DCNN on raw data
- Architecture 2: DSConv
- Architecture 3: Ensemble
- Performance comparisons vs. other papers
- Conclusions

The Problem

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

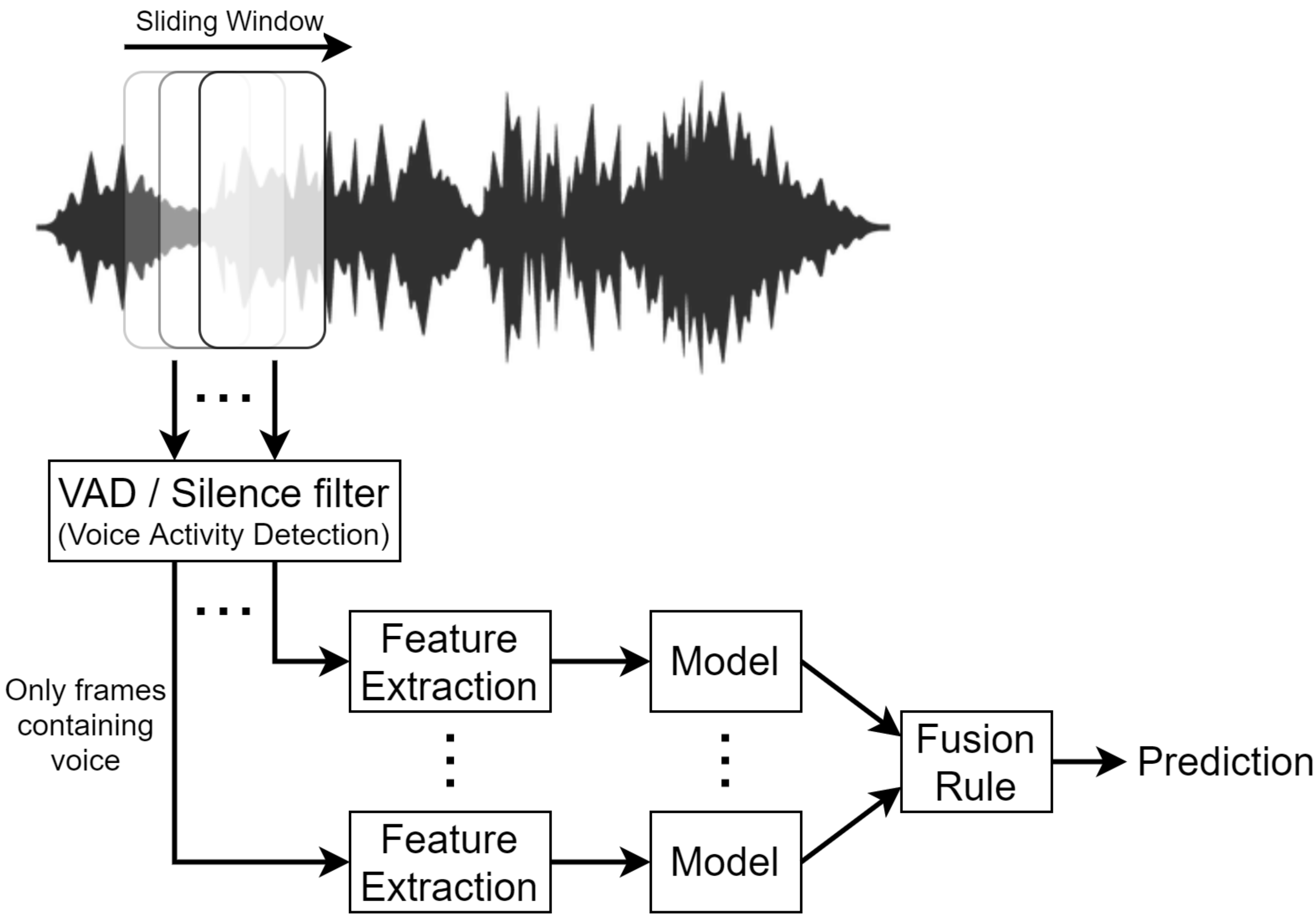
Keyword Spotting

- Detect a relatively small set of predefined keywords (10 or 21) in a stream of user utterances.
- **Application**: Mobile phone, smart home device, consumer and robotics.
- **Constraints**: small footprint and fast (Real Time).

Metrics

- Accuracy
- Number of parameters
- Prediction speed (milliseconds)

The Solution



What we tried

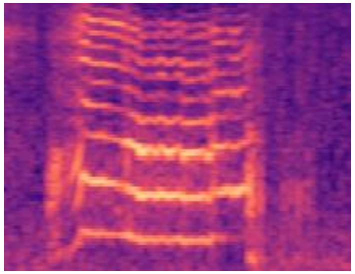
Datasets (Google Speech Dataset V2)

- **10-commands** ("yes", "no", "up", "down", "left", "right", "on", "off", "stop", "go");
- **21-commands** ([…], "zero", "one", "two", "three", "four", "five", "six", "seven", "eight", "nine", **unknown**).

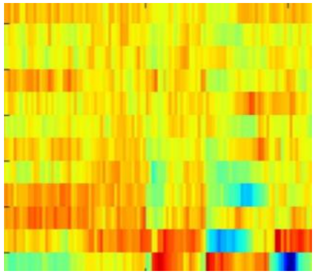
Data Preprocessing techniques

- No preprocessing (Raw Waveform)
- 80 Mel spectrogram
- 40MFCC
- 40MFCC + 40Δ + 40Δs (=120)

80 Mels



40 MFCC

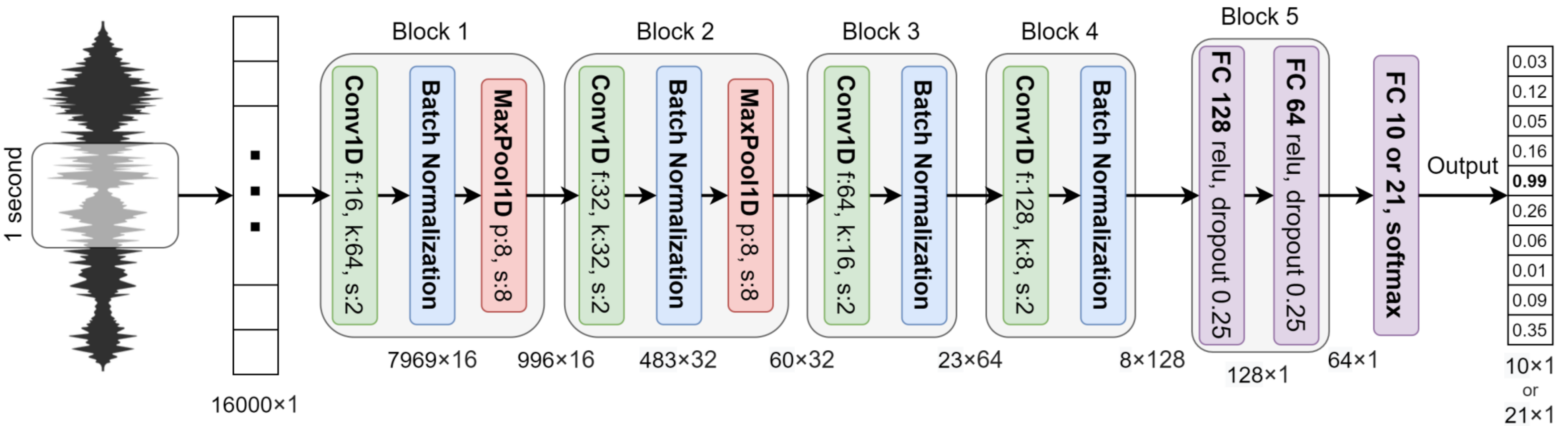


Learning Architectures

- 1DCNN on raw data
- DSConv (Small - Medium - Large)
- Ensemble (Small - Medium - Large)

1D CNN on raw data

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE



	10-commands (30K – 3K – 3K)	21-commands (84K – 9K – 11K)
Accuracy %	93.0	89.1
# parameters	257,018	257,733
Speed (ms)	28.71	28.25

Separable CONV

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Separable convolution performs better than regular convolutional layers.

Two types of Separable convolution: Spatial and Depthwise.

Spatial separable convolution divides a kernel into two smaller kernels.

E.g. division of a 3x3 kernel into a 3x1 and 1x3 kernel.

$$\begin{bmatrix} 3 & 6 & 9 \\ 4 & 8 & 12 \\ 5 & 10 & 15 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

Problem: not all kernels can be "separated" (mathematically) into two.

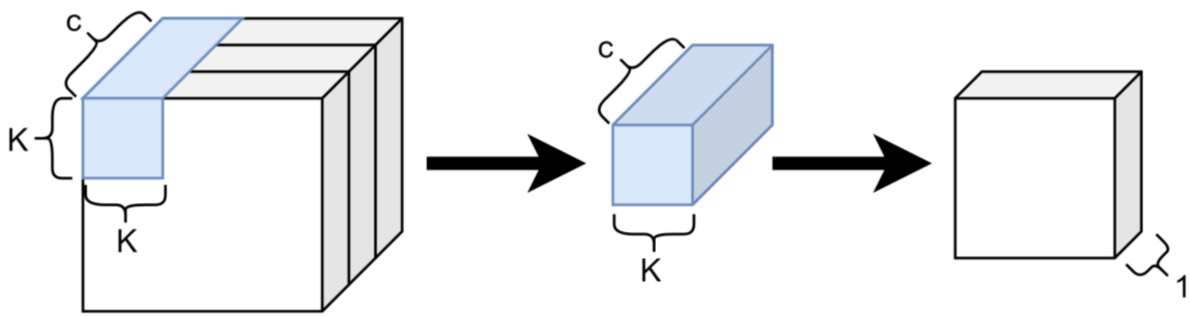
Depthwise Separable convolution uses kernels that cannot be "factored" into two smaller kernels.

It splits a kernel into two separate kernels that do two convolutions:

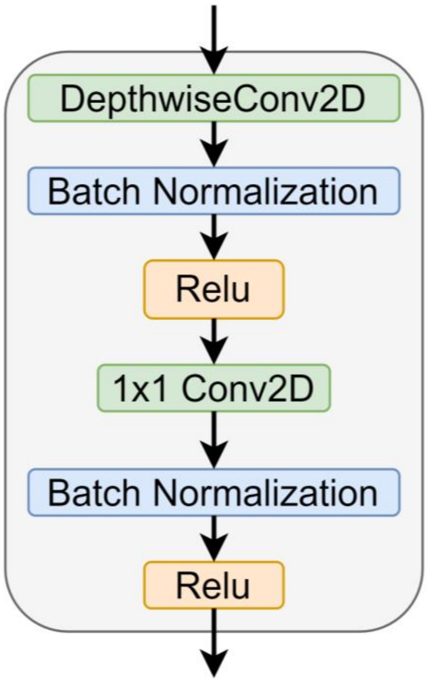
- the depthwise convolution;
- the pointwise (1x1) convolution.

Depth-wise Sep. CONV

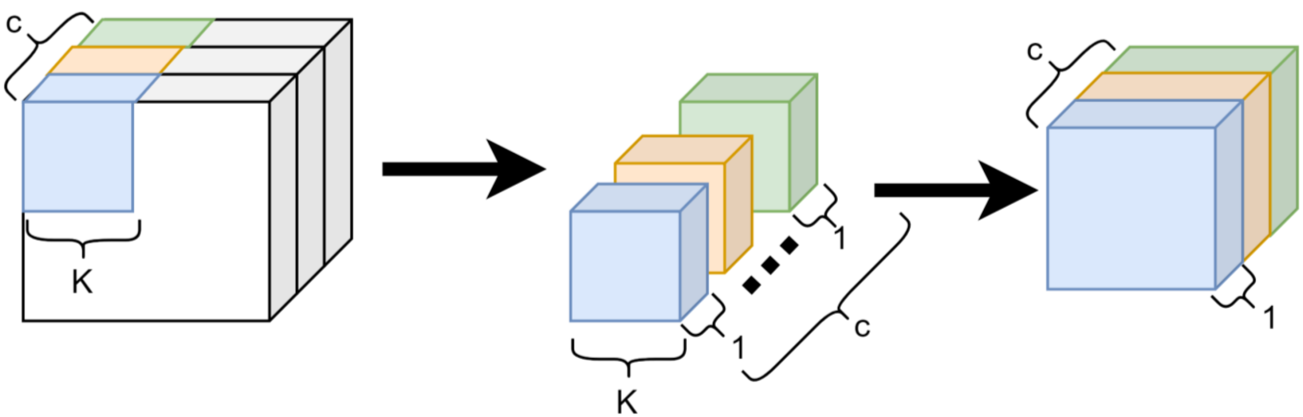
1 Filter Normal Convolution



Separaple Depthwise Convolution

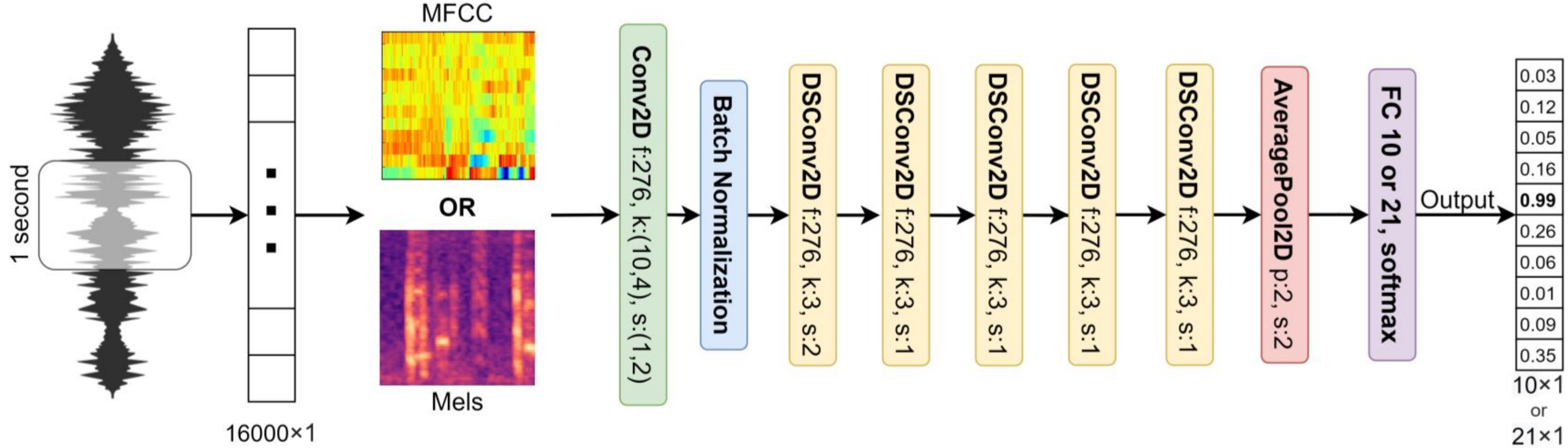


DepthWise Convolution = 1 filter KxKx1 for each channel



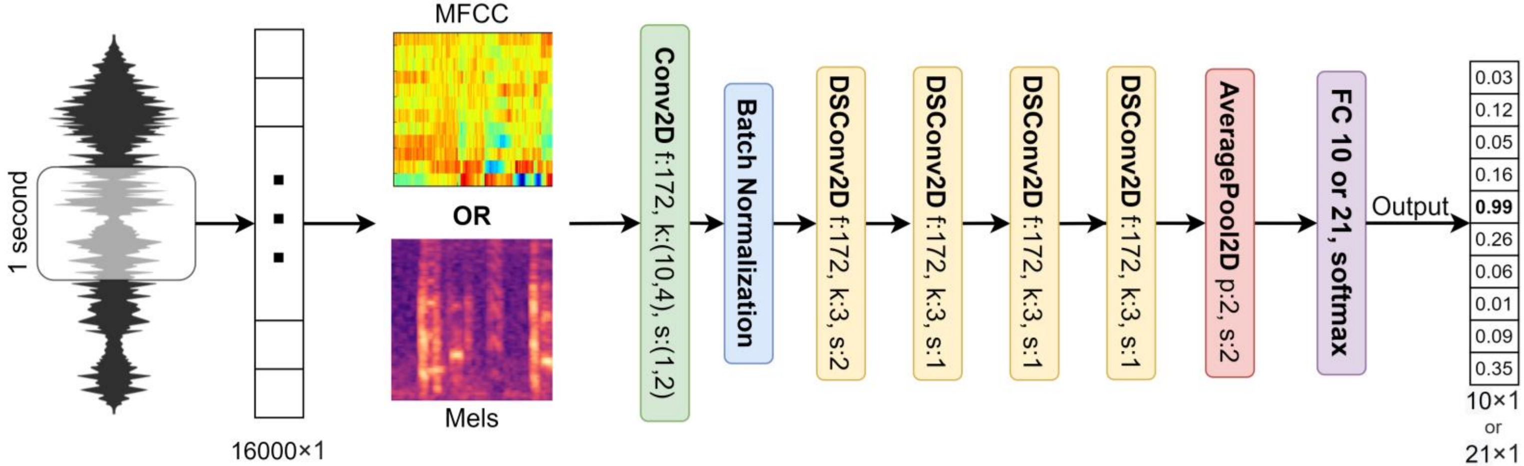
DSConv Model - *Large*

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE



	10-commands (30K-3K-3K)			21-commands (84K-9K-11K)		
	80 Mels	*40MFCC*	*40 + Δs MFCC*	*80 Mels*	*40MFCC*	*40 + Δs MFCC*
Accuracy	**96.0**	95.3	95.5	93.4	**93.7**	92.7
# params	874,930	**571,330**	1,178,530	1,375,881	**738,321**	2,013,441
Speed (ms) **With FE**	**33.39** **41.44**	30.24 45.32	33.42 44.25	33.79 41.62	**30.87** **39.66**	33.13 44.62

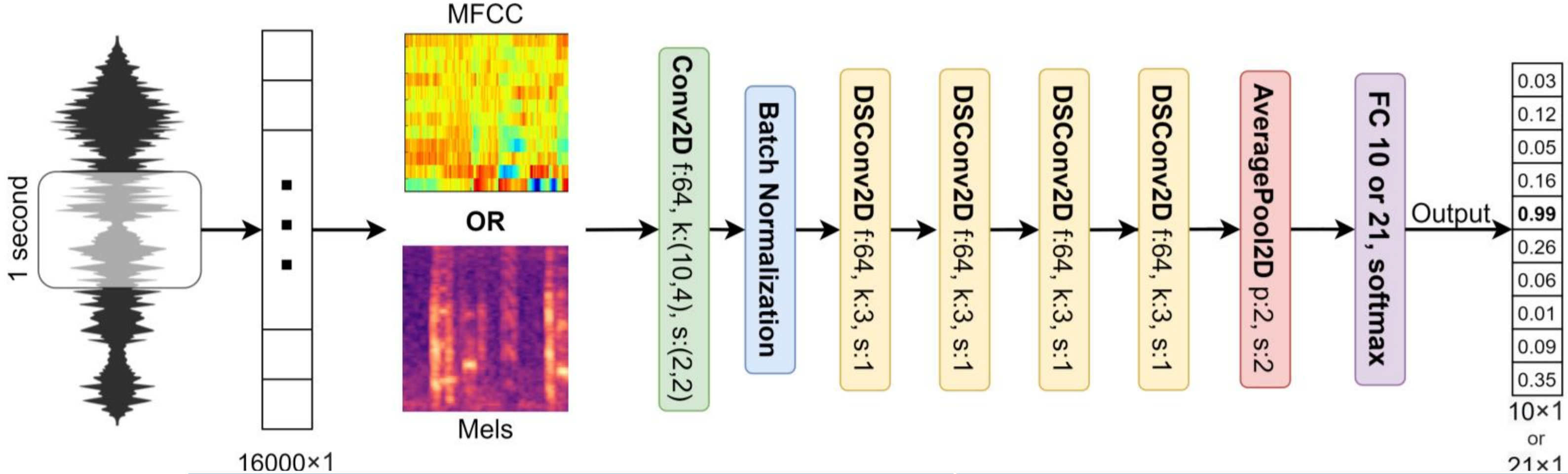
DSConv Model - *Medium*



	10-commands (30K-3K-3K)			21-commands (84K-9K-11K)		
	80 Mels	*40MFCC*	*40 + Δs MFCC*	*80 Mels*	*40MFCC*	*40 + Δs MFCC*
Accuracy	94.3	**95.0**	94.8	**92.7**	92.2	91.7
# params	469,398	**262,998**	675,798	832,673	**399,233**	1,266,113
Speed (ms) With FE	30.75 38.49	**29.55** **38.23**	30.76 41.82	**32.25** **39.72**	30.01 46.76	33.03 47.08

DSConv Model - *Small*

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE



	10-commands (30K-3K-3K)			21-commands (84K-9K-11K)		
	80 Mels	*40MFCC*	*40 + Δs MFCC*	*80 Mels*	*40MFCC*	*40 + Δs MFCC*
Accuracy	92.5	**92.9**	92.5	**90.0**	89.2	86.5
# params	300,618	**127,818**	473,418	604,757	**241,877**	967,637
Speed (ms) With FE	31.00 38.31	**29.27** **38.23**	30.54 40.23	**32.86** **37.97**	29.48 48.15	30.31 40.99

Ensemble: 10-commands

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Ensemble between the best models.

	10-commands (30K-3K-3K)		
	Large	**Medium**	**Small**
	DSConv L 80 Mels + *DSConv L 40 MFCC* + *DSConv L 40 MFCC Δ*	*DSConv L 40 MFCC* + *DSConv M 80 Mels* + *DSConv M 40 MFCC Δ*	*DSConv M 40 MFCC* + *DSConv S 80 Mels* + *DSConv S 40 MFCC*
Accuracy	96.8	96.4	95.6
# params	2,624,790	1,303,726	691,463
Speed (ms)	131.01	122.04	106.09

Ensemble: 21-commands

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Ensemble between the best models.

	21-commands (84K-9K-11K)		
	Large	**Medium**	**Small**
	DSConv L 80 Mels + *DSConv L 40 MFCC* + *DSConv L 40 MFCC Δ*	*DSConv L 40 MFCC* + *DSConv M 80 Mels* + *DSConv M 40 MFCC*	*DSConv L 40 MFCC* + *DSConv M 40 MFCC* + *DSConv S 40 MFCC*
Accuracy	95.2	95.0	94.2
# params	2,498,019	1,970,227	1,379,431
Speed (ms)	125.9	114.21	109.03

Performances: 10-commands

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

	Accuracy %	# Parameters	Speed (ms)
SincConv [Mittermaier et al.]	97.4	162K	40.35
Our Ensemble Large	**96.8**	2,624,790	131.01
Our Ensemble Medium	96.4	1,303,726	122.04
Our DSConvLarge + 80Mels	96.0	874,930	41,44
Our Ensemble Small	95.6	691,463	106.09
Our DSConvMedium + 40MFCC	95.0	262,998	38,23
Our 1DCNN on raw data	93,0	257,018	**28,71**
Our DSConvSmall + 40MFCC	92,9	**127,818**	38,23

- **Best**: Ensemble Large
- **Smaller**: DSConvSmall + 40MFCC
- **Fastest**: 1DCNN on raw data

Performances: 21-commands

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

	Accuracy %	# Parameters	Speed (ms)
SincConv [Mittermaier et al.]	97.4	162K	40.35
Our Ensemble Large	**95.2**	2,498,019	125.9
Our Ensemble Medium	95.0	1,970,227	114.21
Our Ensemble Small	94.2	1,379,431	109.03
Our DSConvLarge + 40MFCC	93.7	738,321	39.66
Our DSConvMedium + 80Mels	92,7	832,673	39,72
Our DSConvSmall + 80Mels	90,0	604,757	37,97
Our 1DCNN on raw data	89,1	**257,733**	**28,25**

- **Best**: Ensemble Large
- **Smaller**: 1DCNN on raw data
- **Fastest**: 1DCNN on raw data

Conclusions

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Conclusions

- Tests: our models are **very good at classifying keywords**;
- We did not beat the state-of-the-art models;
- We found that the number of convolutional layers played a key role in detecting high-level concepts;
- No difference between 80 Mels or 40 MFCCs;
- Different model sizes in order to fit different devices.

Future Work

- Try different hyperparameters during training;
- Change the **structure** of the **network** using:
    - SincConv;
    - GDSConv.
    - Attention model