UNIVERSITÀ
DEGLI STUDI
FIRENZE

Machine Learning

Automatic phoneme recognition on TIMIT Database

Giovanni Ortolani

Università degli studi di Firenze

Feature Extraction

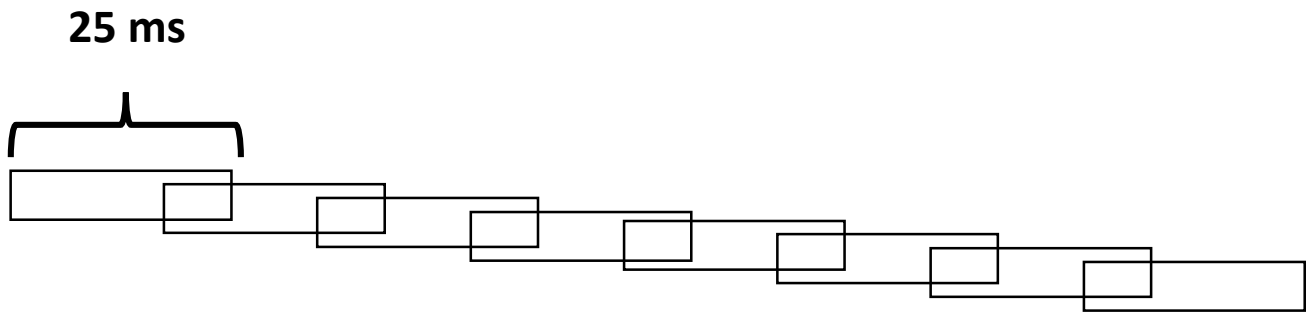
Features

- perceptual linear prediction (PLP) = 13 features
- First order derivative = 13 features
- Second order derivative = 13 features

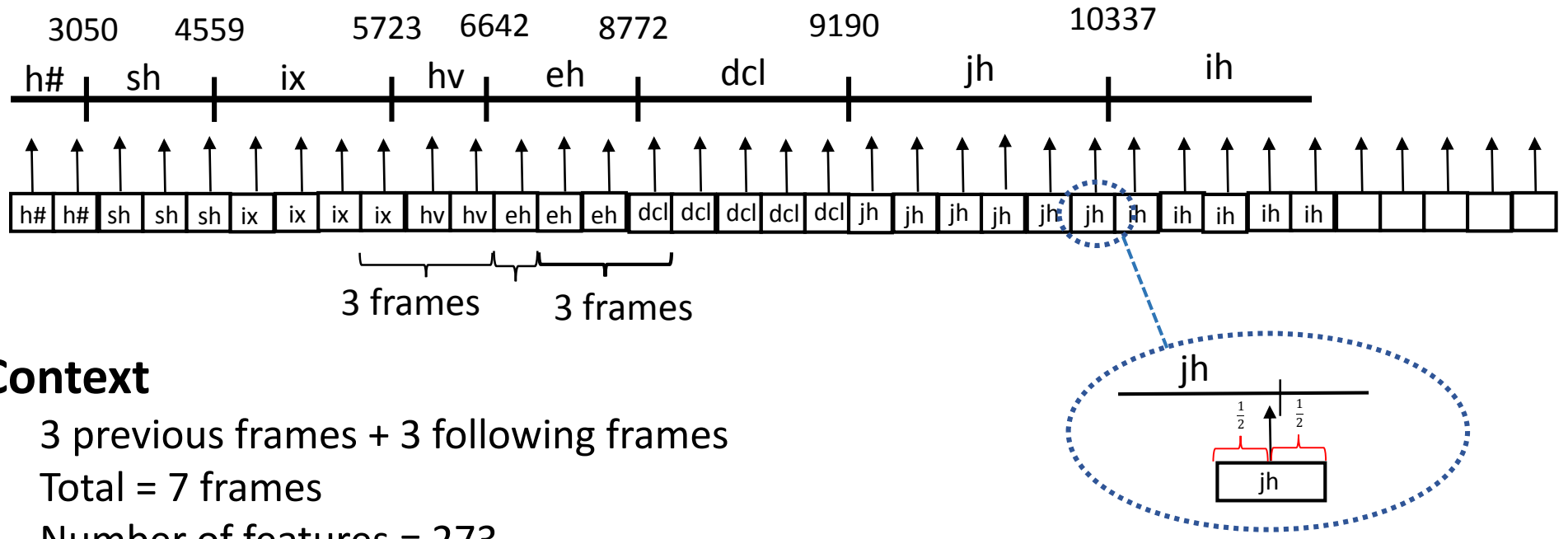
39 features each frame

Frames

- Frame size = 25 ms (400 samples with fs=16Khz)
- One frame each 10ms (160 samples with fs=16Khz)
- Overlapping = 15 ms (240 samples with fs=16Khz)



Feature Extraction



Context

- 3 previous frames + 3 following frames
- Total = 7 frames
- Number of features = 273

Improvements

- Added 13 MFCC features (without deltas)
- Tried frame with size of 20ms
- Tried to add energy of each frame

A frame is labeled with a certain phoneme, when at least half of it is contained in the phoneme window.

Datasets

Vowels

- Training set = 551932 samples
- Test set = 202952 samples
- Classes = 20

Initially

Train (half)
275966 samples
Valid (half)
275966 samples

Later

Train (3/4)
413949 samples
Valid (1/4)
137983 samples

Consontants

- Training set = 555172 samples
- Test set = 193047 samples
- Classes = 32

Initially

Train (half)
277586 samples
Valid (half)
277586 samples

Later

Train (3/4)
393879 samples
Valid (1/4)
131293 samples

PS: Vowels and consonants are 2 disjoint sets. Vowels set doesn’t contain consonants and vice versa.

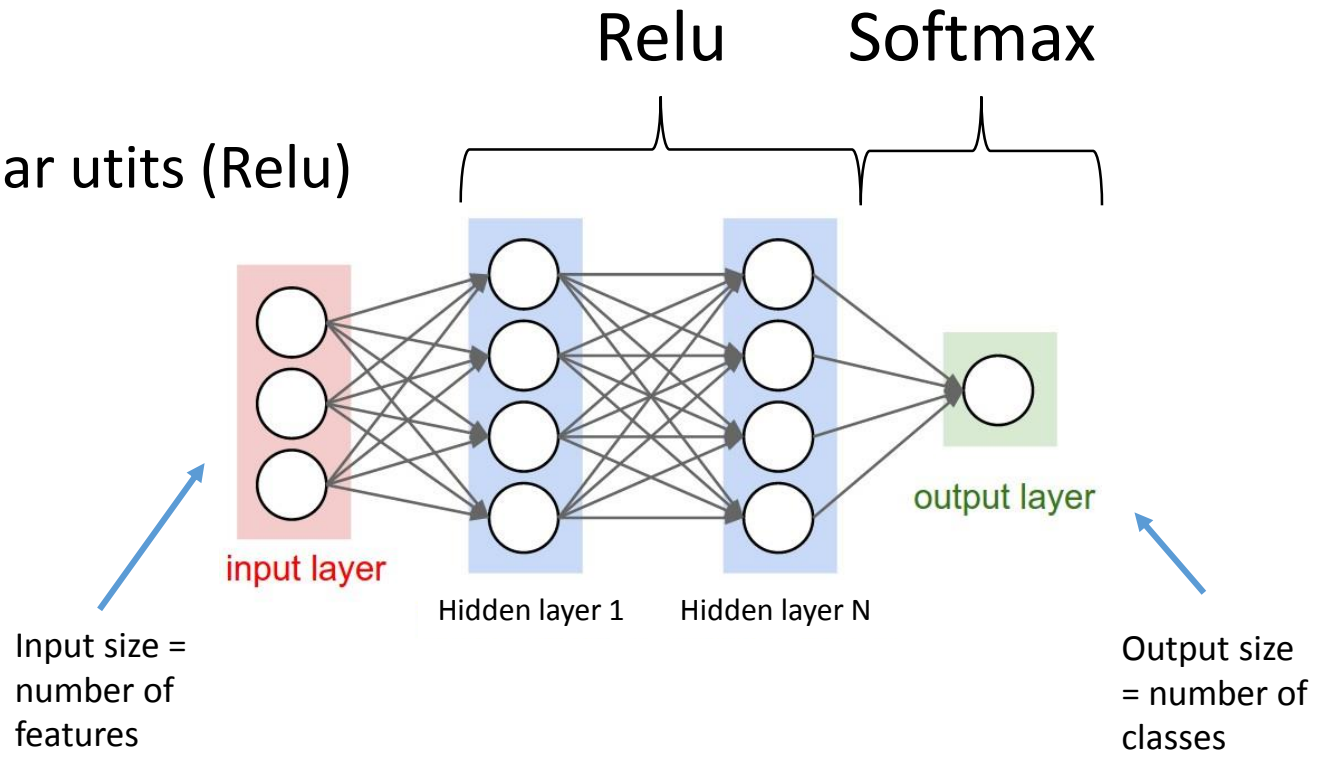
Neural Network

Language and library

- Python
- Pylearn2

Configuration

- Hidden units = Rectified linear utits (Relu)
- Output layef = SoftMax
- Sparse_init = 7
- Irange (output) = 0.01
- Momentum = 0.5



Neural Network (vowels)

Vowels

Layers X Units	Batch size	Learning rate	Epochs	Accuracy %
3X400	100	.001	500	54.9864007253
3X400	100	.001	1000	54.8883479838
3X500	100	.001	500	55.2125625764
3X1000	100	.001	500	56.2679845481
3X500	100	.0001	500	59.2859395325
3X500	100	.00001	500	57.9284756987
3X500	2000	.0001	500	53.5752296109
3x500 (sparse init 14)	100	.0001	500	57.8240175017
3X500	100	.00001	1000	57.9284756987
3X500 (standard scaler)	100	.0001	500	59.7525523276
3X500 (train 3/4 valid 1/4)	100	.0001	500	60.7276597422

Consonants

Consonants:

It has been chosen the best configuration (for vowels)

- Hidden layers = 3X500 Rectified linear units (Relu)
- Output layer = Softmax
- Batch size = 100
- Learning rate = .0001
- Feature scaling = Standard scaler (mean = 0, std deviation = 1)

Accuracy = 74.2449986014

Improvements

All the following tests have been done scaling the features with mean 0 and standard deviation 1

Features	Vowels – Accuracy	Consonants - Accuracy
9 Frames	61.7200126138	
Frame size 20 ms	61.0779888841	
RASTA filtering	58.9701013047	
Energy	61.9230162797	
9 Frames + MFCC	63.3277819386	
9 Frames + MFCC + 5 layers X 500		75.5327745719
9 Frames + MFCC + 5 layers X 1000		72.3844452387

MaxOut + DropOut

Hidden Layers (MaxOut) + Output Layer (SoftMax)

- Num_pieces = 10
- Irange = .005
- max_col_norm = 1.9365
- Batch size = 100
- Learning rate = .1
- Termination criterion = Monitor based (it stops if the error decrease less than 0.001 for 10 epochs)
- Momentum = starting from .5 to .7 (after 250 epochs)

Units are dropped out with probability 0.8. Initially this technique has been applied only to the first layer.

MaxOut + DropOut

Features	Model	Vowels – Accuracy	Consonants - Accuracy
9 Frames MFCC	2X500	50.9578619575	
9 Frames MFCC	3X500	53.5308841499	68.0003729681
9 Frames MFCC Frame size 20 ms	4X1000 – DropOut on 3 layers	59.6623832236	
9 Frames MFCC	4X500		68.1759787823
9 Frames MFCC	4X500 – DropOut on 3 layers		77.634346218
9 Frames MFCC Frame size 20 ms	4X1000 – DropOut on 3 layers		60.2576587963