Lab 3A – Continuous Speech Recognition

Prepare the Database

Number of male and female in training and test set

$> cat train.lst| grep "/man"| cut -d "/" -f 6 | uniq -c| wc -l

	Male	Female
Training	55	57
Test	56	57

Number of utterances in training and test set

$> cat train_word.mlf |grep "\/" | wc -l

	Utterances
Training	8623
Test	8700

Number of phonemes

21 (phones0.lst)
22 (phones1.lst)

Number of nodes and arcs in the recognition network

16 nodes and 36 arcs

Train and test G-HMMs with different features

In the file .cfg we find :

- TARGETKIND

The type of feature we want to use. (MFCC or variants)

- TARGETRATE (HTK uses units of 100ns)

The frame period (100000 ~ 10ms)

- SAVECOMPRESSED

Compress the output features (No)

- SAVEWITHCRC

Add a CRC (resistance to data alteration) to file features (No)

- WINDOWSIZE

The size of the window where we will extract the MFCC features. (200000 ~ 20ms)

- USEHAMMING

Apply a hamming window on the window extracted for computing the FFT (Yes)

- USEPOWER

Take absolute value of FFT (Yes)

- PREEMCOEF

Value of the coefficent that compensates the db dropped due to radiation at lips. (0.97)

- NUMCHANS

Number of coefficients takes in the FFT (40)

- LOFREQ

Lowest frequency analyze in FFT (133.33 Hz)

- HIFREQ

Highest frequency analyze in FFT (6835 Hz)

- CEPLIFTER

Number of filter applied during the step of Mel Filterbank (21)

- NUMCEPS

Number of coefficients kept after the DCT (12)

- ENORMALISE

Energy normalisation on recorded audio files (No)

In the proto .mmf file :

- Size of the Vector is equal to the number of MFCC coefficients
- Number of States of the HMM model
- For each state specified :
  - Tell Means (same size as Vector)
  - Tell variances (same size as Vector)
- Show transition matrix

Differences between all the types of MFCC :

- MFCC : Keep 12 coefficients from the DCT (size 12)
- MFCC_0 : Keep the energy of the DCT (size 13)
- MFCC_0_D : Add delta vector (Delta is the derivative of the MFCC before and after) (size 26)
- MFCC_0_D_A : Add deltadelta vector (Delta^2 is the second derivative of the MFCC thus acceleration) (size 39)
- MFCC_0_D_A_Z : Remove the cepstral mean to the coefficients (size 39)

Globla Mean and Variance :
Compute it from the dataset (ex : training) and use it at inital value for all HMMs

Training

Difference between hmm 1-3 and 5-7

The iterations 1 to 3 contain 21 different HMM models since there are 21 phones.
The iterations 5 to 7 contain 22 HMM models. The 22th one is called 'sp' and shared a state with the phone 'sil'.

Difference between hmm 4 and 5

In HMM 4 the specific HMM for the phone 'sp' and 'sil' are not sharing a state. After using the command HHEd, in HMM 5 the phone 'sp' and 'sil' are sharing the state called 'silst'.

Testing

Simple gaussian

```
====================== HTK Results Analysis =====================
  Date: Mon May  9 17:33:31 2016
  Ref : workdir/test_word.mlf
  Rec : results_MFCC/recout_test.mlf
------------------------ Overall Results ------------------------
SENT: %Correct=78.30 [H=6812, S=1888, N=8700]
WORD: %Corr=92.33, Acc=91.20 [H=26392, D=868, S=1323, I=324, N=28
------------------------ Confusion Matrix -----------------------
       o   z   o   t   t   f   f   s   s   e   n
       h   e   n   w   h   o   i   i   e   i   i
           r   e   o   r   u   v   x   v   g   n
           o           e   r   e       e   h   e
                       e               n   t      Del [ %c / %e]
  oh 1965  12  18  11   1 157  33   1  29   7   3  320 [87.8/1.0]
zero    6 2512  1  24  14   1   0   1  34   4   1   11 [96.7/0.3]
 one   10   0 2419  8   2  12  10   0   1   3  64   72 [95.7/0.4]
 two    9   0   1 2489  7   0   0  22  16  33   0   48 [96.6/0.3]
thre    1   0   0  37 2538  0   0   1   1  18   0   11 [97.8/0.2]
four   27   0  19   0   0 2493 32   0   1   0   0   27 [96.9/0.3]
five   33   0   8   1  18  19 2467  0   0   2  34   17 [95.5/0.4]
 six    0   0   0  28   0   0   0 2557  3  16   0    6 [98.2/0.2]
seve   12   0   4  17   1   1   5  10 2528  8   5    9 [97.6/0.2]
eigh    3   0   6  88  93   5   1  12   3 2163  7  233 [90.8/0.8]
nine   22   1 126   3   7   2   7   1  18   0 2261 114 [92.4/0.7]
Ins    84   1   5  49   1   8  24  25   1 113  13
=================================================================
```

Illustration 1: Result on test set with single gaussian model

The average performance for finding a word is about 91.20 %. It's better to use this value since the Corr percentage are not removing the I (Insertion) value.

The performance for finding the good sentences of digits is about 78.3 %.

The worth case is the word « oh ». It generates 1 % of the global error and only 87.8 % of the retrieving « oh » are really a « oh ».
The word « six » is really a word « six » with the highest probability : 98.2 %.

GMM-2

```
===================== HTK Results Analysis =======================
  Date: Mon May  9 17:42:04 2016
  Ref : workdir/test_word.mlf
  Rec : results_MFCC_0_D_A/recout_test_nmix2.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=89.79 [H=7812, S=888, N=8700]
WORD: %Corr=98.46, Acc=96.37 [H=28142, D=117, S=324, I=596, N=28583]
------------------------ Confusion Matrix -------------------------
       o   z   o   t   t   f   f   s   s   e   n
       h   e   n   w   h   o   i   i   e   i   i
           r   e   o   r   u   v   x   v   g   n
           o           e   r   e       e   h   e
                       e               n   t      Del [ %c / %e]
  oh 2381  2   1   5   1  58  13   0   4   4  20   68 [95.7/0.4]
zero   3 2585  0  10   6   0   1   0   3   1   0    0 [99.1/0.1]
 one   1   0 2582  0   0   5   0   0   0   0   8    5 [99.5/0.0]
 two  15   1   0 2581  6   1   0   0   0   7   0   14 [98.9/0.1]
thre   0   0   0   6 2589  0   0   0   0   9   0    3 [99.4/0.1]
four   1   0   0   0   0 2586 11   0   0   0   0    1 [99.5/0.0]
five   2   0   1   2   0   1 2579  0   0   0  14    0 [99.2/0.1]
 six   0   0   0   1   0   0   0 2608  0   0   0    1 [100.0/0.0]
seve   2   0   1   0   1   1   0   0 2595  0   0    0 [99.8/0.0]
eigh   2   0   7  12  14   1   0   8   3 2535  8   24 [97.9/0.2]
nine   9   0  16   0   3   0  12   0   0   0 2521   1 [98.4/0.1]
Ins  244   0   6  96   3   8  10   4   2 204  19
===================================================================
```

Illustration 2: Result on test set with 2 gaussians model

We gain 5 % of accurancy for the words. Thus, the correct sentences increases of more than 10 %.

GMM-4

```
===================== HTK Results Analysis =======================
  Date: Mon May  9 17:45:21 2016
  Ref : workdir/test_word.mlf
  Rec : results_MFCC_0_D_A/recout_test_nmix4.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=93.20 [H=8108, S=592, N=8700]
WORD: %Corr=99.03, Acc=97.68 [H=28305, D=78, S=200, I=384, N=28583]
------------------------ Confusion Matrix -------------------------
       o   z   o   t   t   f   f   s   s   e   n
       h   e   n   w   h   o   i   i   e   i   i
           r   e   o   r   u   v   x   v   g   n
           o           e   r   e       e   h   e
                       e               n   t      Del [ %c / %e]
  oh 2467  4   0   2   0  18   5   0   5   3   8   45 [98.2/0.2]
zero   2 2598  0   3   3   0   0   0   3   0   0    0 [99.6/0.0]
 one   0   0 2596  0   0   2   0   0   0   0   3    0 [99.8/0.0]
 two   6   0   0 2605  4   0   0   0   0   4   0    6 [99.5/0.0]
thre   0   0   0   6 2595  0   0   0   0   4   0    2 [99.6/0.0]
four   5   0   0   0   0 2586  8   0   0   0   0    0 [99.5/0.0]
five   6   0   0   0   0   1 2576  0   0   0  16    0 [99.1/0.1]
 six   0   0   0   0   0   0   0 2605  4   0   0    1 [99.8/0.0]
seve   3   0   1   0   0   0   0   0 2596  0   0    0 [99.8/0.0]
eigh   2   0   9  10   7   0   0   7   3 2544  8   24 [98.2/0.2]
nine  11   0   7   0   0   0   7   0   0   0 2537   0 [99.0/0.1]
Ins  164   0   6  65   1   3   2   3   0 127  13
===================================================================
```

Illustration 3: Result on test set with 4 gaussians model

We gain 1.37 % of accurancy for the words. We still gain 3.5 % for the correct sentences.

GMM-8

```
====================== HTK Results Analysis =======================
  Date: Mon May  9 17:50:33 2016
  Ref : workdir/test_word.mlf
  Rec : results_MFCC_0_D_A/recout_test_nmix8.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=94.95 [H=8261, S=439, N=8700]
WORD: %Corr=99.40, Acc=98.31 [H=28411, D=56, S=116, I=310, N=28583]
------------------------ Confusion Matrix -------------------------
       o   z   o   t   t   f   f   s   s   e   n
       h   e   n   w   h   o   i   i   e   i   i
           r   e   o   r   u   v   x   v   g   n
           o           e   r   e       e   h   e
                       e               n   t      Del [ %c / %e]
   oh 2490   3   0   2   0  16   1   0   4   4   3   34 [98.7/0.1]
 zero    0 2606   0   1   1   0   0   0   1   0   0    0 [99.9/0.0]
  one    0   0 2599   0   0   1   0   0   0   0   1    0 [99.9/0.0]
  two    1   0   0 2615   3   0   0   0   0   0   0    6 [99.8/0.0]
 thre    0   0   0   4 2602   0   0   0   0   1   0    0 [99.8/0.0]
 four    3   0   0   0   0 2594   2   0   0   0   0    0 [99.8/0.0]
 five    4   0   0   0   0   0 2588   0   0   0   7    0 [99.6/0.0]
  six    0   0   0   0   0   0   0 2607   2   0   0    1 [99.9/0.0]
 seve    2   0   2   0   0   0   0   0 2596   0   0    0 [99.8/0.0]
 eigh    1   0   7   3   6   0   0   4   2 2570   6   15 [98.9/0.1]
 nine   10   0   5   0   0   0   3   0   0   0 2544    0 [99.3/0.1]
 Ins   149   1   9  49   0   2   3   1   0  83  13
===================================================================
```

Illustration 4: Result on test set with 8 gaussians model

We gain 0.6 % of accurancy for the words. There is still a gain of almost 3 % for the sentences.

GMM-16

```
====================== HTK Results Analysis =======================
  Date: Mon May  9 17:59:20 2016
  Ref : workdir/test_word.mlf
  Rec : results_MFCC_0_D_A/recout_test_nmix16.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=96.01 [H=8353, S=347, N=8700]
WORD: %Corr=99.59, Acc=98.66 [H=28465, D=46, S=72, I=264, N=28583]
------------------------ Confusion Matrix -------------------------
       o   z   o   t   t   f   f   s   s   e   n
       h   e   n   w   h   o   i   i   e   i   i
           r   e   o   r   u   v   x   v   g   n
           o           e   r   e       e   h   e
                       e               n   t      Del [ %c / %e]
   oh 2501   2   0   3   0  10   0   0   5   3   2   31 [99.0/0.1]
 zero    0 2607   0   1   0   0   0   0   1   0   0    0 [99.9/0.0]
  one    0   0 2601   0   0   0   0   0   0   0   0    0
  two    0   0   0 2622   1   0   0   0   0   0   0    2 [100.0/0.0]
 thre    0   0   0   6 2601   0   0   0   0   0   0    0 [99.8/0.0]
 four    3   0   1   0   0 2593   2   0   0   0   0    0 [99.8/0.0]
 five    0   0   0   0   0   0 2594   0   0   0   5    0 [99.8/0.0]
  six    0   0   0   0   0   0   0 2610   0   0   0    0
 seve    2   0   1   1   0   0   0   0 2596   0   0    0 [99.8/0.0]
 eigh    0   0   3   3   0   0   0   1   2 2590   2   13 [99.6/0.0]
 nine    7   0   3   0   0   0   2   0   0   0 2550    0 [99.5/0.0]
 Ins   127   1   5  46   0   2   5   1   0  64  13
===================================================================
```

Illustration 5: Result on test set with 16 gaussians model

We gain 0.3 % of accurancy for the words. The percentage of good sentences increases of 1 %.

Conclusion :

The more you add gaussians, the better are the results. However, the derivative of the percentage by the number of gaussians is decreasing.
Since the sentences are a mix between different words, a small increase in the words accuracy gives a bigger increase for the sentences accuracy. Thus, even the smaller increase in the words accuracy can be important.

Tune insertion penalty parameter

For the one gaussian model, we have a balance favorable for the deletions.
For the GMM model, we have always a balance favorable for the insertions.

The default value for the gaussion model is -20.

Test with value 0 :

```
====================== HTK Results Analysis =======================
  Date: Mon May  9 19:15:32 2016
  Ref : workdir/test_word.mlf
  Rec : results_MFCC/recout_test.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=73.70 [H=6412, S=2288, N=8700]
WORD: %Corr=93.55, Acc=88.82 [H=26739, D=440, S=1404, I=1351, N=28583]
------------------------ Confusion Matrix -------------------------
       o   z   o   t   t   f   f   s   s   e   n
       h   e   n   w   h   o   i   i   e   i   i
           r   e   o   r   u   v   x   v   g   n
           o           e   r   e       e   h   e
                       e               n   t      Del [ %c / %e]
  oh 2130   3  19   6   1 145  29   1  23  12   5  183 [89.7/0.9]
zero   42 2489  2  19  12   0   0   1  31   4   4    5 [95.6/0.4]
 one   17   0 2448  7   2  12   7   0   0   5  77   26 [95.1/0.4]
 two   10   0   0 2527  5   0   0  20  15  30   0   18 [96.9/0.3]
thre    1   0   0  38 2536  0   0   2   1  23   0    6 [97.5/0.2]
four   28   0  17   2   0 2507 28   0   0   0   0   17 [97.1/0.3]
five   38   0   8   4  12  16 2476  1   0   9  30    5 [95.5/0.4]
 six    1   0   0  32   0   0   0 2545  4  27   0    1 [97.5/0.2]
seve   17   0   7  17   3   1   4   7 2519 10  11    4 [97.0/0.3]
eigh    5   0   4 110  89   2   1  13   5 2249 11  125 [90.4/0.8]
nine   35   0 121  14   7   1   6   1  11   3 2313  50 [92.1/0.7]
Ins   440   3  26 232  12  41  68  73   5 411  40
===================================================================
```

Illustration 6: Insertion penalty set to 0 for MFCC model

We have more insertion and a lower accuracy.

Test with value -60 :

```
===================== HTK Results Analysis =======================
  Date: Mon May  9 19:19:29 2016
  Ref : workdir/test_word.mlf
  Rec : results_MFCC/recout_test.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=70.97 [H=6174, S=2526, N=8700]
WORD: %Corr=86.96, Acc=86.81 [H=24856, D=2318, S=1409, I=43, N=28583]
------------------------ Confusion Matrix -------------------------
       o   z   o   t   t   f   f   s   s   e   n
       h   e   n   w   h   o   i   i   e   i   i
           r   e   o   r   u   v   x   v   g   n
           o           e   r   e       e   h   e
                       e               n   t      Del [ %c / %e]
   oh 1654 74  23  12   0  150 34   1  28   4   2  575 [83.5/1.1]
 zero   2 2468  1  20  16   1   0   2  31   6   2   60 [96.8/0.3]
  one  11   4 2217  5   2  20  14   0   3   1  54  270 [95.1/0.4]
  two   9   3   1 2277  8   1   0  19  19  50   0  238 [95.4/0.4]
 thre   0   0   1  30 2474  0   1   0   1  19   1   80 [97.9/0.2]
 four  26   0  18   2   0 2445 29   1   1   0   0   77 [96.9/0.3]
 five  28   0   8   1  18  17 2411  2   2   0  34   78 [95.6/0.4]
  six   0   1   0  25   0   0   0 2518  3  15   0   48 [98.3/0.2]
 seve   5   1   3  20   1   0   5  14 2493  6   4   48 [97.7/0.2]
 eigh   1   1  12  61  98  13   5  16  11 1906  4  486 [89.6/0.8]
 nine   8   8 122   6   7   4  16   2  38   0 1993 358 [90.4/0.7]
 Ins    6   1   1   3   0   2   3   8   3  12   4
===================================================================
```

Illustration 7: Insertion penalty set to -60 for MFCC model

Even worth, the deletions are bigger and the accuracy is smaller than with p = 0 that was already smaller.

Forced Alignment

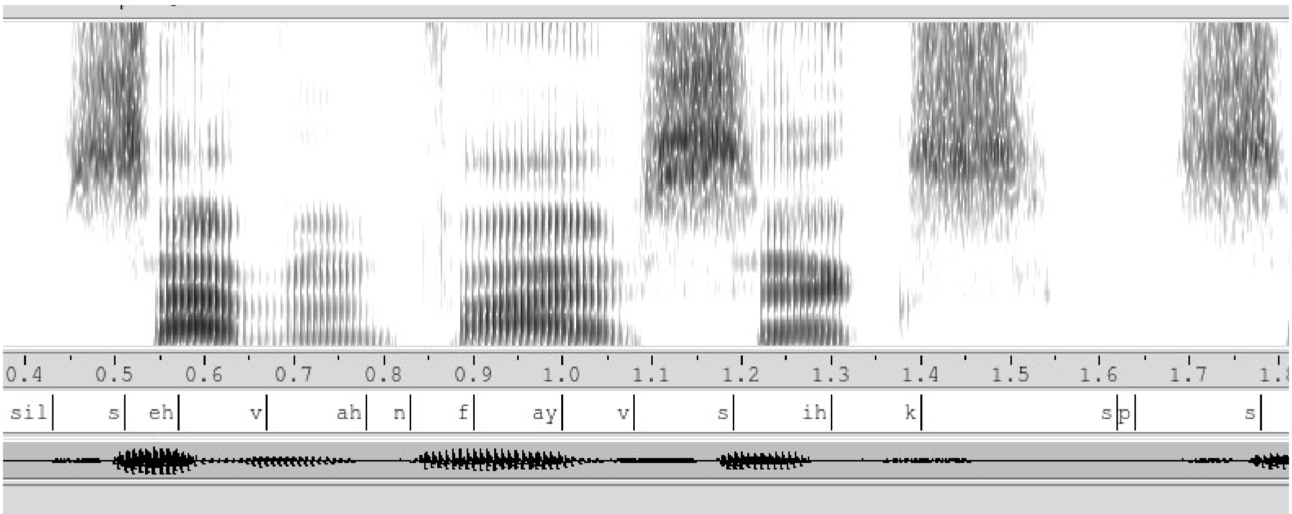

Illustration 8: An utterance where we see the frequencies and the phones associated

The 's' and the 'f' are well aligned. A 'f' is a gray block in the bottom and a 's' is a gray block in the top. The big white gap is a 'sp' and I have heard a break.

Analysis

Parameter evolution at different iterations

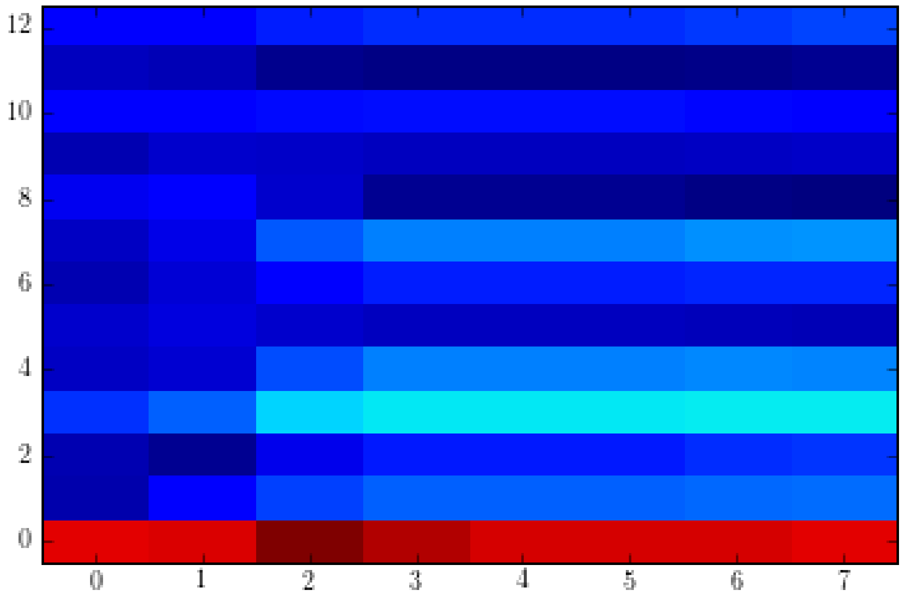

Illustration 9: Mean parameters evolution of state 2 in 'n' (MFFC0)

The mean parameters almost converge in the forth iteration and stop changing after the seventh. The convergence is reached. The coeffcient 0 that correspond to the DCT energy is really bigger than the other ones.

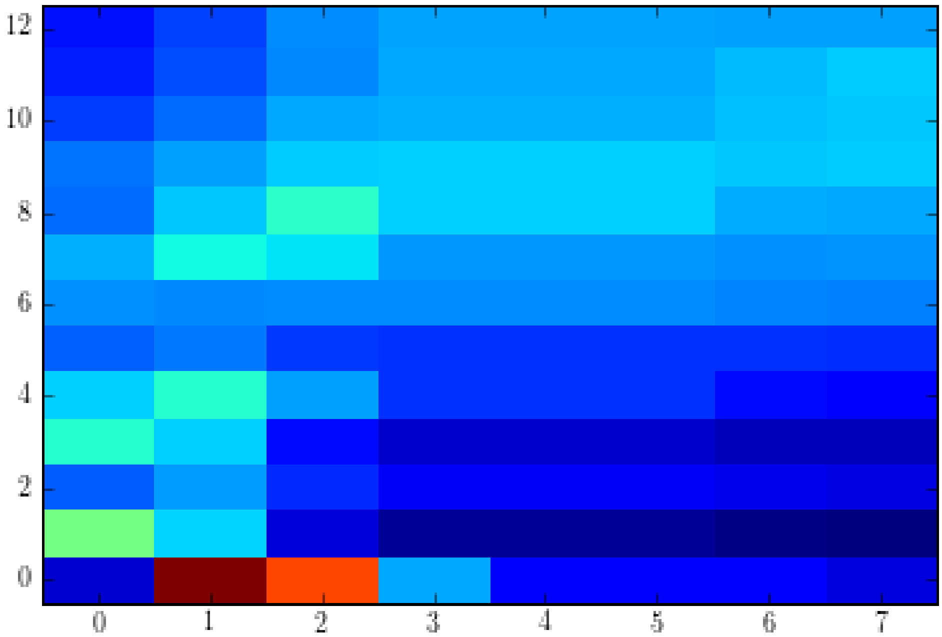

Illustration 10: Variance parameters evolution of state 2 in 'n' (MFCC0)

As for the mean, the variance parameters change a lot before the forth iteration and some minor changes are still visible in the last iteration.

Parameters for different phonemes

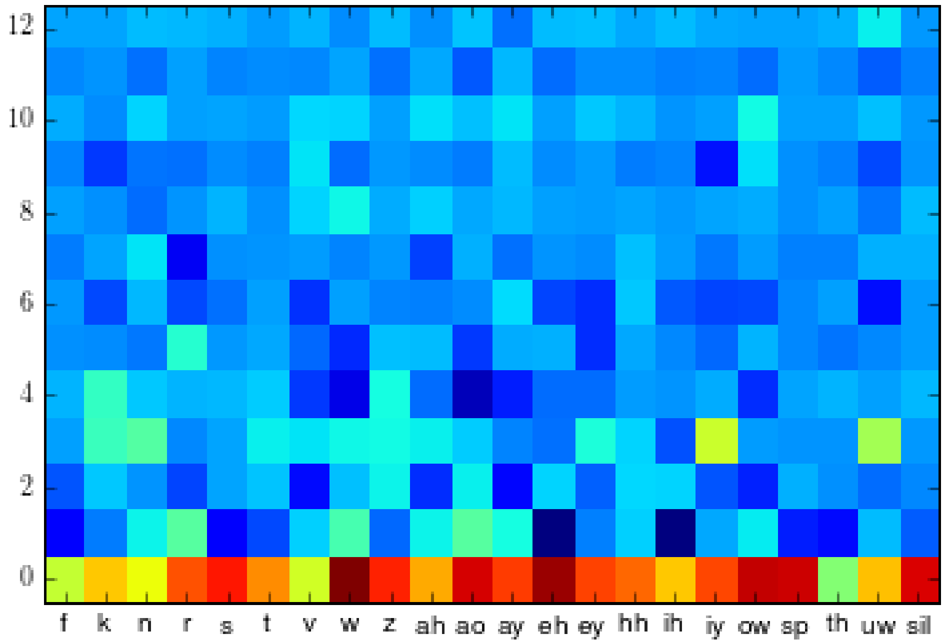

Illustration 11: Mean of state 0 for each phonemes in hmm7 in MFCC0

A 's' and a 'sil' are very similar. The energy is really big and all the other coefficients are almost constant.

Usually the fricatives have a lower energy ('f', 'k', 'n'...).

The vowels have a biger energy and the lower coefficients (2-7) are more variable and sometimes very low.