**INTRODUCTION**: This report covers the various tasks completed by me while working as a ‘Research Assistant’ for Dr. Catherine Watson on her one of the key projects named ‘Maori Pronunciation Aid (MPAid)’. As a startup, knowledge of Wavesurfer[1] software is necessary along with this, read chapter 3 (A tutorial example of using HTK) of HTK book[2] in order to understand the recognition process completely and also read the use/functions of various HTK tools like HVite, HErest, HCopy and impact of various parameters present in their commands on the output).

After this, start with the new version of the software, where different batch files (see folder ‘**Batches’** in ‘**HTK’** folder) were made in order to generate most of the files (.mlf’s, .mfc’s etc) automatically, whereas in Annie’s version we had to create most of the files manually. In every batch file, some comments are provided to increase the readability of the user.

[1] <http://sourceforge.net/projects/wavesurfer/>

[2] <http://htk.eng.cam.ac.uk/download.shtml>

Brief information about the **NEW VERSION of MPAid** is given as:

1. **‘DataPreparer’** is the batch file used to create the grammar files, dictionaries and mfc files, .mlf files which are necessary in order to generate HMM models. It needs the address of the audio recording of whose data has to be created (in this case it is ‘D:\UOA\MPAid\Audio’)
2. **‘HMMGenerator’** is the batch file used to generate all the HMM models starting from 0 to 15 by itself.
3. **‘ModelEvaluator’,** it needs address of the audio recording which is (‘D:\UOA\MPAid\Audio’) and after evaluation it stores results of static recognition (used by MPAid) in ‘Evaluations’ folder present in the ‘HTK’ folder.
4. **‘Livetest’** is the batch file used for live recognition, just double click it and start the live testing of the created HMM models.
5. ‘**RecordingRenamer**’ batch file renamed the old recording as follows:

**R001M 🡪 oldfemale-word-hau-R001M**

1. All parameter file needed for the HMM generation like configuration files(.conf), .led, .hed files are present in the **‘Params’** folder.
2. All ‘**.mfc’** files created by HCopy command present in the datapreparer batch file are present in the ‘**MFCs’** folder.
3. All **‘.mlf’** files created by various HTK tools are while HMM generation are present in the ‘**MLF’s’** folder.
4. All HTK tools are present in the ‘**TooLs’** folder.

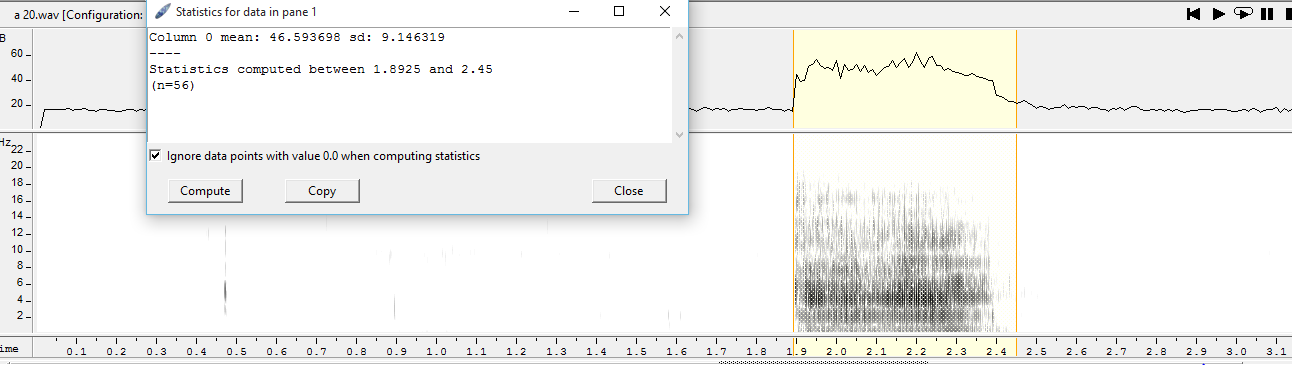
**TASK 1**

**CALCULATING THE OPTIMAL DYNAMIC RANGE OF THE MICROPHONE-COMPUTER PAIR FOR EFFICIENT RECOGNITION:**

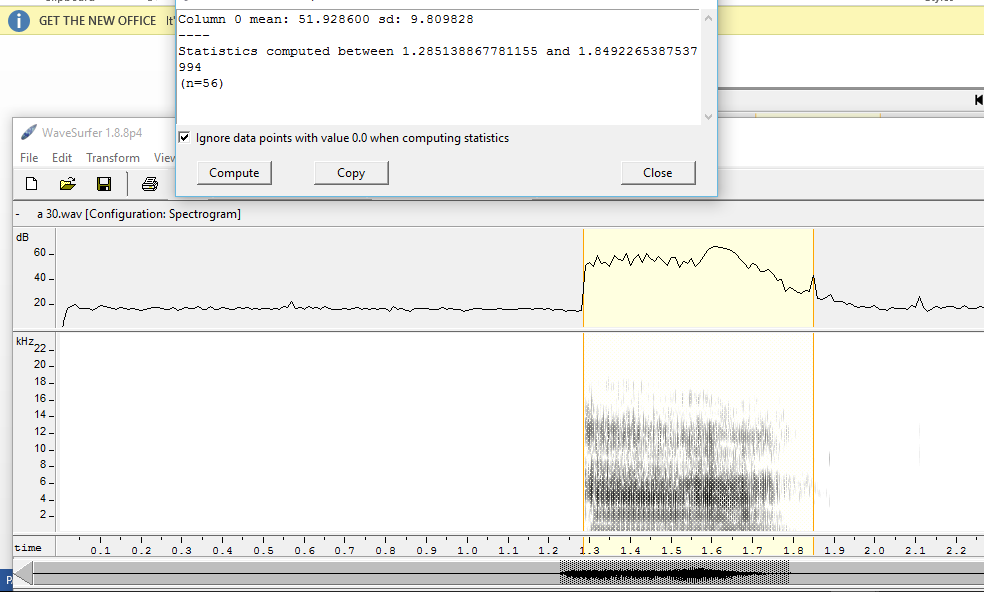
**INTRODUCTION**: In thisspeech analysis, I’ve considered three vowel sounds i.e. a, e and o and recorded them using Logitech headphones at various gain levels (100, 80, 60, 40, 30 and 20 according to the volume levels of computer screen). Then the analysis of these is done by comparing the values of mean power distribution and the standard deviation of the speech signal in the time window (n) (i.e. ‘n’ which is different for different vowel sounds) using wave-surfer software.

1. **Vowel “a” analysis:**

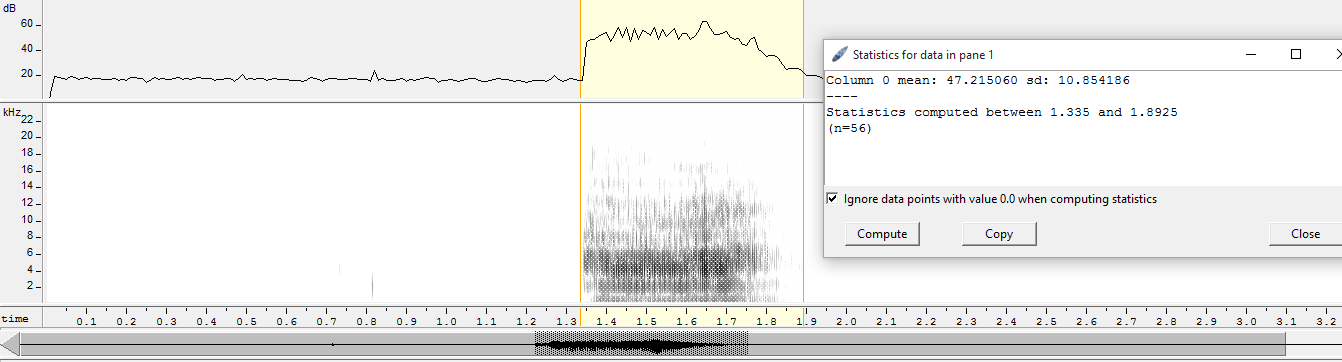
**1.a** Vowel ‘a’ at value 20



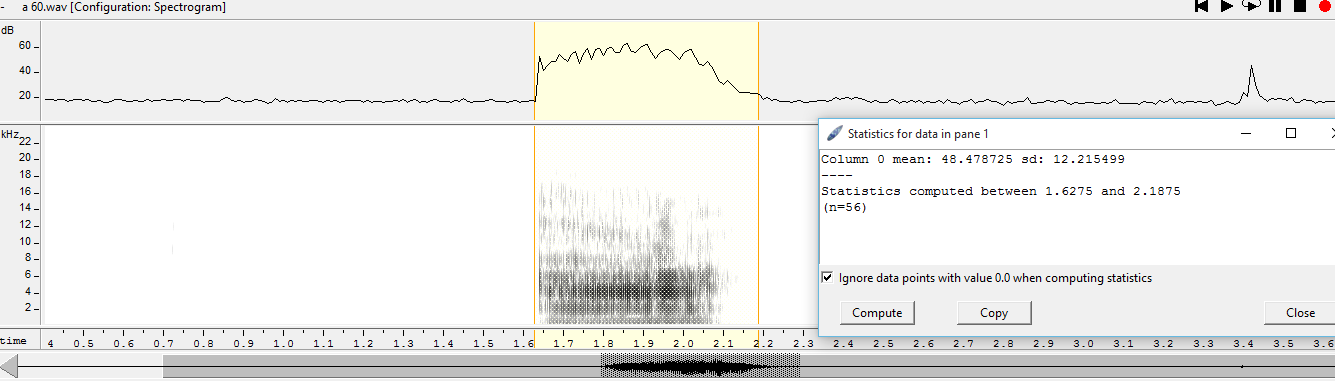
**1.b vowel ‘a’ at gain =30**



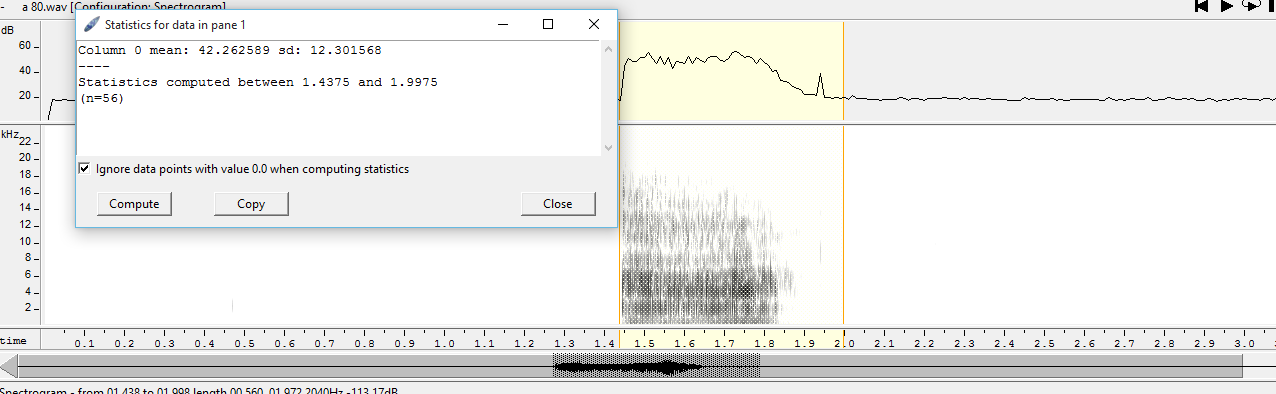
**1.c vowel ‘a’ at gain =40**



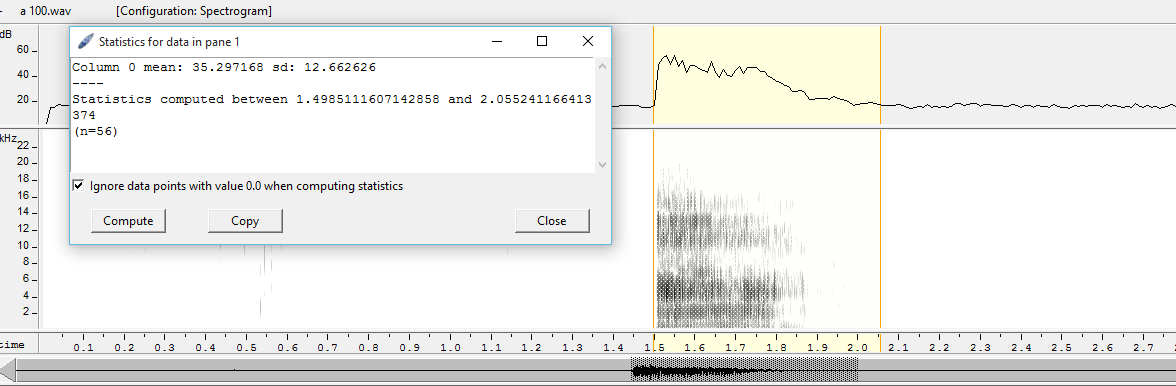
**1.d vowel ‘a’ at gain =60**



**1.e vowel ‘a’ at gain =80**



**1.f vowel ‘a’ at gain =100**



1. **Analysis of vowel ‘a’: n =56**

**‘n’** is the window of time (\*100) within which we are doing this statistical analysis.

|  |  |  |
| --- | --- | --- |
| **Gain of microphone increased in steps** | **Mean value of signal power (in db)** | **Standard deviation** |
| 20 | 46.59 | 9.15 |
| 30 | 51.92 | 9.80 |
| 40 | 47.21 | 10.85 |
| 60 | 48.48 | 12.21 |
| 80 | 42.26 | 12.30 |
| 100 | 35.29 | 12.66 |

1. **Analysis of vowel ‘e’:** Using the time window (n=52), previous exercise is performed that we did for vowel ‘a’. So the tabular representation of mean signal power, standard deviation vs gain of the microphone is given as follows

|  |  |  |
| --- | --- | --- |
| Gain of microphone increased in steps | Mean value of signal power (in db) | Standard deviation |
| 20 | 45.13 | 14.44 |
| 30 | 52.89 | 9.69 |
| 40 | 52.88 | **12.64** |
| 60 | 53.57 | 9.22 |
| 80 | 46.95 | 12.92 |
| 100 | 50.68 | 12.74 |

1. **Analysis of vowel ‘o’:** Using the time window (n=54), previous exercise is performed that we did for vowel ‘a’. So the tabular representation of mean signal power, standard deviation vs gain of the microphone is given as follows

|  |  |  |
| --- | --- | --- |
| Gain of microphone increased in steps | Mean value of signal power (in db) | Standard deviation |
| 20 | 36.58 | 5.91 |
| 30 | 40.38 | 7.47 |
| 40 | 40.068 | 6.014 |
| 60 | 39.27 | 4.85 |
| 80 | 36.82 | 9.83 |
| 100 | 34.26 | 6.82 |

**Combined Analysis:**

Plot1: Comparing mean power vs microphone gain

Plot2: Comparing standard deviation vs microphone gain

**CONCLUSION**: As it can be seen that the maximum signal (i.e. maximum mean powers) levels are in experienced in the range 30 to 60 in terms of microphone gain, which is also confirmed by minimum levels of the standard deviation (fairly low than the values of std deviation outside the range 30 to 60 in terms of microphone gain) of the signal levels in that range. So the best dynamic range for the efficient speech recognition is in between 30 and 60. And this dynamic range wouldn’t change even if the word/vowel selection is from different language for example vowels of Maori language.

**Common Flaw’s/trends founded on older version of MPAid:**

**Microphone Settings:** The microphone gain is maintained at 45 (best level for recognition) and to have the same orientation as I had, speaker should be able to see the microphone at same level as that to nose while looking at both with just the left eye. (Receiver part of microphone should be on the left side)

1. Just ‘**inhalation’** of air by the speaker at the very beginning is recognized as **‘tae/tai’.**
2. Just the ‘**exhalation’** of the air also results in **‘Kї’.**
3. Even the actof ‘**clearing throat’** before speaking resulted in **‘tai’** recognition.
4. There is ‘**no null**’ option in the recognition, even for any rubbish input recognizer still pops up with **some uncertain/unknown output**.
5. For one word, the recognizer most of the times comes up with multiple words or set of tri-phones and bi-phones, in short it ‘**lacks one to one mapping**’ which is a real need.
6. The usual recording (with windows recorder) has **alot less/marginal** **noise** in it as compared to the MPAi software in which recording are enriched with noise.

**TASK 2**

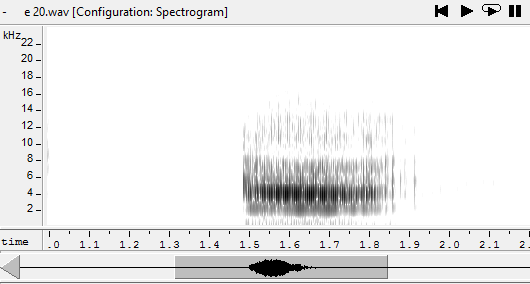
**FORMANT ANALYSIS:**

**FORMANT**: Formant is the concentration of Acoustic energy over a frequency in a speech wave. Each Formants corresponds to a resonance frequency of vocal tract (cavity). Darker the Formant in the Spectrogram more energy is present at that frequency means more audible/stronger it is.[a]

[a] Retrieved from <http://person2.sol.lu.se/SidneyWood/praate/whatform.html>

**SPECTROGRAM**: ‘Kay Sona’ was earlier used to get the spectrograms for the sound signals. Spectrogram represents the spectral data over time with the **amplitude shown in different shades of grey colour**.

The degree of darkness represents the various resonance frequencies.



**2.a FORMANT ANALYSIS OF ‘e’ vowel:**

Analysis based on the recordings provided by Dr. Peter J Keegan (See the folder Dr. Peter J Keegan Recordings)

LTAS: Long term average spectrum

Order = 20

n=24 (time window)

FFT points =512

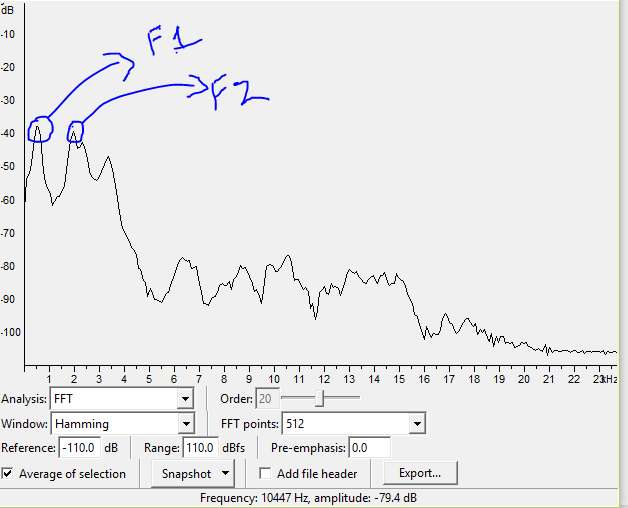


Fig: Formants (formant1 F1, formant2 F2) magnitude plot vs frequency (KHz).

|  |  |  |
| --- | --- | --- |
| Gain | **First Formant (F1/red) power magnitude in dB** | **Second Formant (F2/green) Power magnitude in dB** |
| 10 | -37.6 | -39.8 |
| 20 | -43.2 | -46.6 |
| 30 | -37.1 | -38.6 |
| 40 | -33.2 | -35.5 |
| 50 | -31.4 | -32.7 |
| 60 | -29.9 | -31.2 |
| 70 | -27 | -30 |
| 80 | -26.9 | -28.0 |
| 90 | -26.7 | -28.2 |
| 100 | -26.4 | -28 |

Table: Formant magnitudes vs gain increase.

Plot 1: Formant power vs increase in gain plot.

**CONCLUSION:** Value of the magnitude of the first and the second formants increases with the increase in the reference gain of microphone and remains consistent after value of 70. (It’s obvious)

**Mean and standard deviation analysis for ‘e’: n=52**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gain | Mean first formant  (frequencies) | Mean Second formant (frequencies) | Std deviation first formant | Std deviation second formant |
| 10 | 576.905720 | 1964.493215 | 79.569739 | 37.573991 |
| 20 | 527.216958 | 1965.588811 | 116.625216 | 48.425084 |
| 30 | 536.641136 | 1911.931875 | 80.598981 | 62.850665 |
| 40 | 521.851669 | 1981.082713 | 69.287782 | 55.046553 |
| 50 | 537.896117 | 1973.477239 | 46.299399 | 57.639913 |
| 60 | 545.098813 | 1998.203181 | 59.090249 | 49.024963 |
| 70 | 541.02 | 1970.32 | 64.56 | 53.67 |
| 80 | 535.782448 | 1979.813466 | 63.787189 | 58.947309 |
| 90 | 585.852610 | 2003.668635 | 163.789247 | 49.707947 |
| 100 | 563.671250 | 2003.809077 | 62.886793 | 58.076119 |

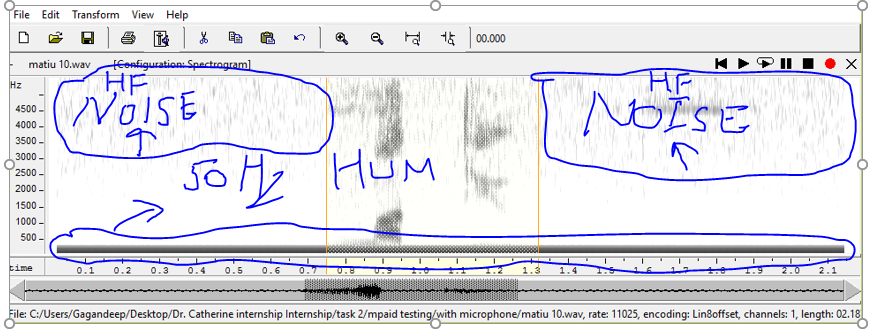
**CONCLUSIONS:** Even if we keep on increasing the gain levels which in turns increase the magnitude (dB) of the formants continuously, the mean formant frequencies for both the formants remains the same.

**2.b. POTENTIAL FINDINGS:**

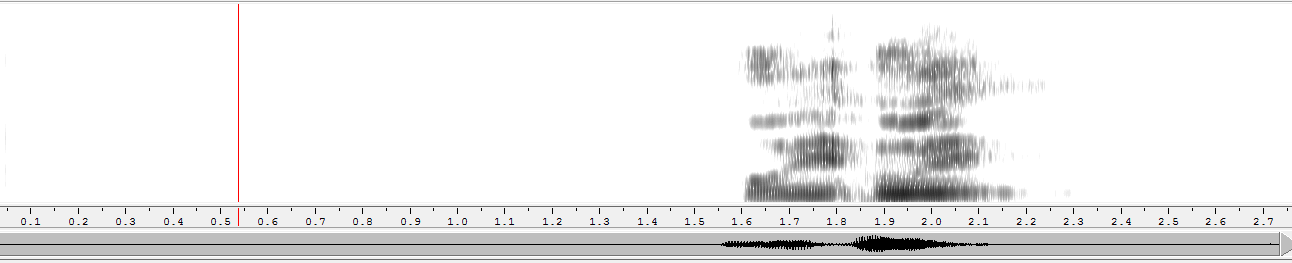
1. The live recognition ability of the HTK is **far better than the** MPAid tool.
2. The recordings of HTK are **noise free**. (See cmd live recognition part)
3. (Ans to Dr. Catherine’s ques of last meeting) Even the recordings done with the microphone array of this pc are pretty much noiseless recordings, which are far better than the MPAid.

(check recordings ‘a\_window\_recorder’ and ‘a\_MPAid\_recorder’)

1. Recording of the MPAid carries **high frequency noise but low energy** in it (see diagram below).
2. The MPAid recordings have a 20-30Hz **hum** (of amplitude around -40 dB) **present, a high energy peak** in it. (**OUR RECORDER PERFORMANCE IS POOR - WE NEED TO CHANGE IT (means it was not effectively used/linked in the last version, we replaced it with VLC recorder but the response became slower because of the bigger size of the recordings (they were of very high quality) done by VLC, and also it is meant for videos**), **SO WE AGAIN REPLACED VLC WITH N-AUDIO RECODER BUT USED/LINKED IT PROPERLY TO THE MPAID).**



1. Hum (mammoth peak) and High frequency noise can be seen in the recordings of the older version of the N-Audio Recorder.



This is the spectrogram of a recording done with the new version of the MPAid. You can see that there is no high frequency noise and the absence of the HUM (mammoth peak) in the spectrogram.

**2.c CHECKING DYNAMIC RANGE DIIFERENCE OF ‘WITH’ AND ‘WITHOUT MICROPHONE (HeadPhones)’ RECORDINGS USING MPAid:**

1. **With Micro-phones: n=55**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Mean signal power** | **Standard deviation** | **First formant**  **mean** | **First formant sd** | **Second formant mean** | **Second formant sd** |
| Gain 10 | 8.484 | 1.155 | 606.97 | 418.072 | 1738.365 | 424.045 |
| 20 | 8.789 | 1.609 | 668.838 | 438.055 | 1730.487 | 489.504 |
| 30 | 9.893 | 3.354 | 613.651 | 376.956 | 1767.579 | 459.068 |
| 40 | 9.609 | 3.005 | 520.702 | 324.356 | 1723.842 | 405.006 |
| 50 | 9.251 | 2.586 | 607.317 | 370.607 | 1627.992 | 445.922 |
| 60 | 8.506 | 1.235 | 546.504 | 335.367 | 1599.476 | 351.905 |
| 70 | 9.357 | 2.517 | 549.056 | 328.733 | 1569.752 | 360.876 |
| 80 | 9.427 | 2.497 | 554.674 | 333.411 | 1641.001 | 472.631 |
| 90 | 9.234 | 2.427 | 611.633 | 369.234 | 1846.266 | 546.554 |
| 100 | 9.353 | 2.872 | 585.826 | 355.842 | 1699.785 | 405.827 |

1. **Without Micro-phones.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gain | **Mean\_without** | **Std\_without** | **Mean\_with** | **Std\_with** |
| 10 | 10.105704 | 3.033926 | 8.484 | 1.155 |
| 20 | 10.117326 | 3.266212 | 8.789 | 1.609 |
| 30 | 10.028006 | 3.397025 | 9.893 | 3.354 |
| 40 | 10.141270 | 3.003101 | 9.609 | 3.005 |
| 50 | 10.439425 | 3.354275 | 9.251 | 2.586 |
| 60 | 10.242525 | 3.546539 | 8.506 | 1.235 |
| 70 | 9.907931 | 2.639088 | 9.357 | 2.517 |
| 80 | 10.727633 | 3.878564 | 9.427 | 2.497 |
| 90 | 9.978643 | 2.610157 | 9.234 | 2.427 |
| 100 | 9.851663 | 2.695750 | 9.353 | 2.872 |

**CONCLUSION:**

1. Standard deviation in the signal is proportional to the mean signal power in as it can be seen in both the plots, that std deviation and mean curves are identical.
2. Best optimal values of gain would be **between 30 to 50** somewhere near value of 30, as the value of signal strength is maximum. (same as the results of task 1)
3. There is **difference** in the mean signal power and std deviation of the recordings recorded **with or without headphones** using MPAid.

**TASK 3**

**PHONEME’S STRING ANALYSIS:**

**COMMANDS USED:**

**1.a** HVite -C user/config1 -H hmm15/macros -H hmm15/hmmdefs -S user/testL1YoungFemale\_gagan.scp -l \* -T 4 -i analysisOutputOriginal/recout0aYoungfemale1.mlf -w user/wordNetwork -p 0.0 -s 5.0 user/dictionary user/tiedList>analysisOutputOriginal/HViteOutYoungfemale

**1.b** HVite -C user/config1 -H hmm15/macros -H hmm15/hmmdefs -S user/testL1YoungFemale\_gagan.scp -l \* -a -f -i analysisOutputOriginal/recout0bYoungfemale1.mlf -w user/wordNetwork -p 0.0 -s 5.0 user/dictionary user/tiedList

**1.c** HResults -t -I user/wordTranscript\_gagann.mlf user/tiedList analysisOutputOriginal/recout0aYoungfemale1.mlf>analysisOutputOriginal/HResultsOutYoungfemale\_gagan

{See Annie’s Report Page 16, for the commands in order to analyse ‘Young Male, Old Male and Old Female’ recordings. You can also do it by yourself by editing the ‘testL1YoungFemale\_gagan.scp’ according to Youngmale, old male and old female test data.}

**WAY 2:** There’s a batch file called ‘**ModelEvaluator’** in ‘**Batches**’ folder in **‘HTK’** folder, after running that go to the ‘**MLFs**’ folder in the **‘HTK’** folder and open file named ‘**RecMLF**’ to get the results of the static recognition by the Htk for all the recordings. (Results are better than later way because I’ve prepared the dictionary and all other files using double characters for the long vowels like ä🡪aa, ö🡪oo etc. Reason being the HTK works on ASCII and these are special characters not present in ASCII so are supported by HTK, so HTK performs internal character set conversions and while representing the results back the words having special characters lose their uniqueness when HTK performs reconversion. For example: During live recognition (in older version) kë and kї were recognized as k’).

**L group (Young Female (H) group) (for results, see file HResultsOutYoungFemale\_gagan)**

**‘GREEN**’ colour shows one vowel of the diphthong is correctly recognised,

‘**BLUE**’ colour shows particular vowel is correctly recognised. where ‘**LIGHT BLUE’** shading of the box represents exact word matching.

‘**BROWN**’ colour shows same vowel recognised.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Word No.** | **Word** | **Speaker No** 1  (L1H01M\_) | **Speaker No 2**  (L1H02M\_) | **Speaker No 3**  (L1H03M\_) | **Speaker No 4**  (L1H04M\_) | **Speaker No 5**  (L1H05M\_) | **Speaker No 6**  (L1H06M\_) |
| 01 | tënei | kë tae |  |  | kë mau | kë toi |  |
| 02 | täne | pau hë | täne kë | täne kë | hau hë | tae tai | tae tënei |
| 03 | hau | pae |  | hau tü | pau | pae | tae |
| 04 | hou |  | pou | hau | pou | hoe | hei |
| 05 | pao |  |  | mätao pou | pao tae |  | pao pou |
| 06 | pau | hou |  | hou | hau | hau | tae |
| 07 | pou | hou |  | tü | hou | hou | tae |
| 08 | pö | pö pou |  | hou | hou | hau pao | pao |
| 09 | pai | hoe | pae pou | pai kei |  | pou | pae tae |
| 10 | pae |  |  |  |  | pae pao |  |
| 11 | kë |  |  |  | hë pae | hë |  |
| 12 | kei | hei | kë | hei | hei | hei |  |
| 13 | kї | kë | kї | kë | kë | kë | kei |
| 14 | hë | kë | kë | täne kë | hë mau |  | kë |
| 15 | hei |  | hau |  | hau | hau |  |
| 16 | hї | kї | kї | kë | kї | kї | kei |
| 17 | tae | pae | pae | pae kë | pae | pae | pae |
| 18 | tai | tai pou | pai | pae kei | pai | pai pau | pae |
| 19 | mätao | mätao pae | mätou pao | mätao pou | mätao pou |  | mätao pou |
| 20 | mätau | mätou | mätau pae | mätou | hou mätau | mätau pae | mätau pae |
| 21 | mätou | mätao hoe |  | mätau | mätau | mätao hoe | mätau tae |
| 22 | toetoe | hoe hoe | toi hoe pou | toi hoe kë | toi toi | hoe hoe | hoi tënei |
| 23 | toi | hoihoi |  | hoi kei | hoe kei | toi hau | hoe kei |
| 24 | hoihoi | hoe hoe | toi toi |  | hoihoi kë | hoihoi pae | hoe hoe kei |
| 25 | hoe |  |  | hoe kei | hoihoi | hoe pao | hoe toi |
| 26 | mao |  |  | hou mao | hou  pao |  | mao pou |
| 27 | mau | hou pae tü |  | tü mau tü |  | mau tü | kë pae |
| 28 | moutere | mau | mao hou kë kë | mau toi kei | hou mau tënei | hoihoi tai kë | tai tënei |
| 29 | tü | mau kë kë | kë | hë | tae hou | kei | kei |
| 30 | matiu | tü matiu tae pae | mau kë |  | mätau | mätau | mätao kë |

**CONCLUSIONS:**

1. Main patterns of misrecognition in terms of ‘**CONSONANTS**’ are:

t🡪h/p, p🡪h/t, h🡪k/p, m🡪h

**NOTE:** The misrecognised words depends on the factor that before which vowel the consonant is present. For ex: kei🡪hei, kë🡪kї

1. Main patterns of misrecognition in terms of ‘**VOWELS--DIPHTHONGS**’ are:

oe🡪oi, oi🡪oe/ei, iu🡪au/ao/ë, ei🡪ae/oi/au, ї🡪ë, ou🡪ao, a🡪ä

1. ‘ä’ is perfectly recognised.
2. pae, pou are getting added (may be because of the inhalation or exhalation). (need to check)
3. For the shaded regions, after checking corresponding recording there were some sounds like watery mouth, knocking the bench at the end of the recording, infant crying (mätao L1H01M\_19) which led to the

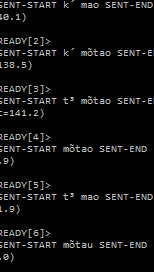
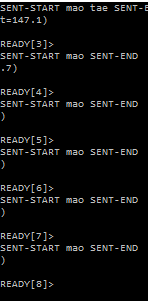
**TASK 4**

**PREEMPHASIS COEFFCIENT AND SAMPLE RATE CHANGE**

In this task, I changed the PREEMCOEF and SOURCERATE (standard values 0.90 and 625 respectively) in the configuration files (HMMs.conf, MFCs.conf, test.conf according to new version and these files are present in the ‘**Params**’ folder in the ‘**HTK**’ folder or config0, config1, config2, config3, config4, config5, 6 configuration files present in the **‘user’** folder in the **‘Annie’** folder in the previous version)in order to check if there is any improvement in the recognition.

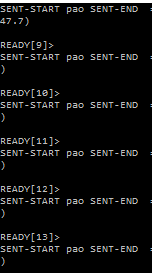
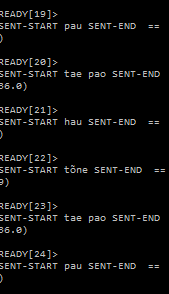
**mao (Microphone volume level-- 44)**

**625/0.97 (old parameters)** **525/0.90(new parameters)**

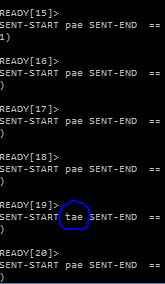
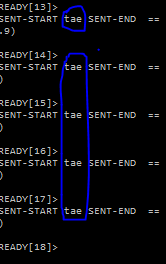
**Pao**

**625/0.97 (old parameters)** **525/0.90(new parameters)**



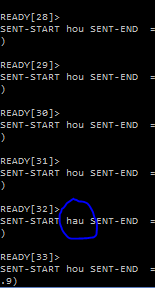
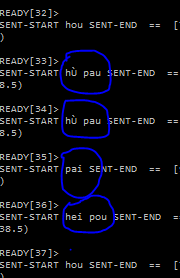
**pae**

**625/0.97 (old parameters)** **525/0.90(new parameters)**



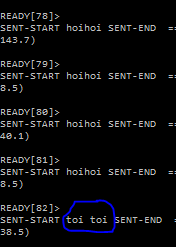
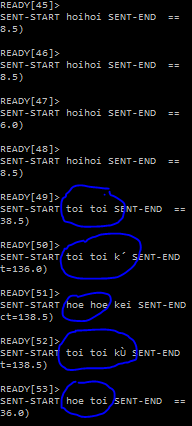
**hou**

**625/0.97 (old parameters)** **525/0.90(new parameters)**



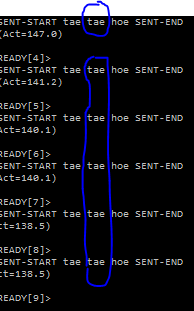
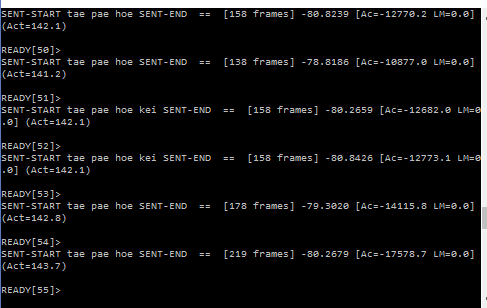
**Hoihoi**

**625/0.97 (old parameters)** **525/0.90(new parameters)**



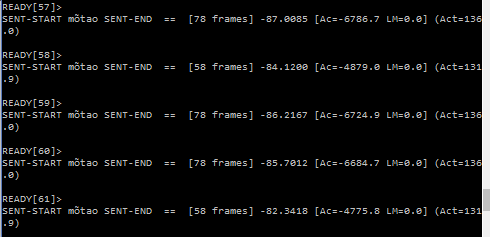
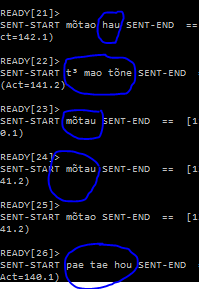
**Tae pae hoe** (just to check recognition with combination of words of wordlist)

**625/0.97 (old parameters)** **525/0.90(new parameters)**

**Mätao**

**625/0.97 (old parameters)** **525/0.90(new parameters)**



**CONCLUSIONS**: By changing the sampling rate (SOURCRATE) from 625 to 525 and pre-emphasis coefficient (PREEMCOEF) from 0.97 to 0.90 following conclusions has been generated:

**1.** **For 625/0.97** the recognition **results** are good in the beginning but they **degrade** as we proceed further however **for 525/0.90** they are **consistent and accurate.**

**2. Recognition of consonants** like p/h/t with parameters 525/0.90 is far better than 625/0.97.

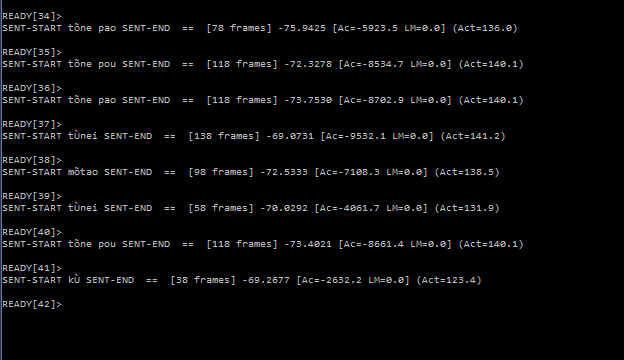
**3**. **No word breakage at the diphthongs** with these new settings.

**4**. **If spoken sensibly with the new parameters the air is recognized less** as compared to the old parameters. Basically the **recognizer became less sensitive** to minor air spikes.

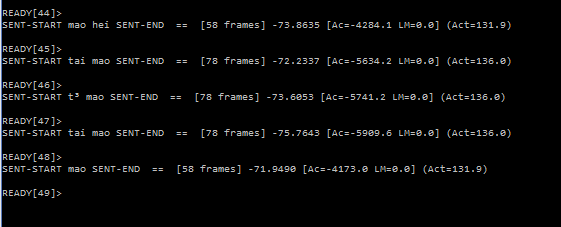
**5**. The **recognizer became less sensitive to disturbance**, so external noise (not inhalation and exhalation yet) very marginally effects the recognition, but it should not override the required sound.

1. **SOURCERATE: 525, PREEMCOEF: 0.99**

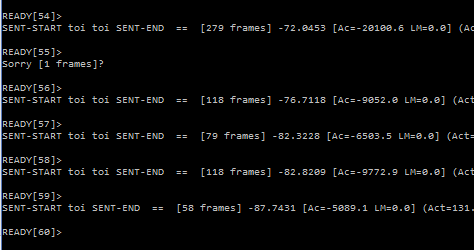
This is done just to check if there is any improvement in the recognition by increasing the pre-emphasis coefficient from 0.97 to 0.99.



Recognized words for ‘**pao’**



Recognized words for **‘mao’**



Recognized words for **‘hoihoi’**

**CONCLUSION**: 1. For simple words like ‘pao’/pou/mao **some additional words has been recognized** and **recognition for consonants is off.** So that’s why combination 525/0.99 was not selected because it’s not that efficient as compared to the 525/0.90 combination.

**2. Recognizer inefficient** to recognize **consonants like ‘p’ (pae🡪tae) and ‘h’ (hoihoi🡪toitoi).**

**WHAT REALLY HAPPENED:**

1. **SOURCE RATE DECREASE:** The previous sampling rate (SOURCERATE) was 625 (scale is 100 nanoseconds) which is equal to 62.5 micro seconds.

So sampling frequency(Fold)= 1/(62.5\*10^-6) =16,000Hz

Now with decrease (Fnew) =1/(52.5\*10^-6) =19,048 Hz

SOURCRATE (time) decreases means the sampling frequency increases.

**NOTE:** Human voice lies in range (300Hz to 8Khz), so earlier we did sampling twice the signal frequency i.e 16KHz, now we raised it to 19KHz or simply we increased the extent of the OVERSAMPLING.

Now the merits of the OVERSAMPLING are:

1. Improves resolution.
2. Avoid aliasing and noise at high frequencies.
3. Avoid Phase distortion.
4. Relax Anti-Aliasing filter requirements.
5. **PRE-EMPHASIS COEFFICIENT DECREASE**: The two main phenomenon’s linked with the pre-emphasis coefficient (PREEMCOEFF (k) in range 0≤k<1)are:
6. Boosting high frequencies, more the value more it boosts up the higher frequencies in order to increase the SNR.

**NOTE:** The higher value of this coefficient (i.e 0.97) was causing the recognition of the rubbish components like air spikes & nearby distortion, because they are short duration (instant phenomena’s (on time scale)) which means of higher the frequency.

**Which got sorted out with the lesser value of this coefficient i.e. 0.90.**

1. More Pre-emphasis also flattens the PHASE-RESPONSE.

**TASK 5**

**NEW MODEL GENERATION TO SOLVE THE RECOGNITION PROBLEM FOR EXTENDED VOWELS**

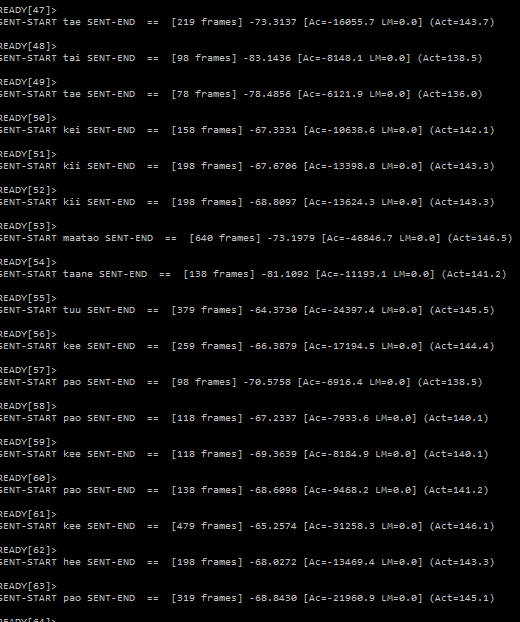
**NEED:** Please note that, the HTK works on basic ASCII (not extended) and characters like ë, ü, and other extended vowel representations are not supported by that which are either UTF-8 or ANSI, so HTK do character set conversions and in order to process these characters and generate them during live recognition so that’s why for kë and kϊ these were shown as k’ or ku’ so words lose uniqueness during recognition as we couldn’t tell if k’ corresponds to kë or kϊ.

So in order to solve that I changed all the extended vowels as follows

ë 🡪 ee, ü 🡪 uu , ϊ 🡪 ii and similarly for the other extended vowels.

**Modifications**: Changed all files that needed to be in order to train the new model set using the parameters of 0.90/525 (PREEMCOEF and SOURCERATE) for configuration file, all master label files, dictionaries, word list, grammar, script files, monophone’s list, triphone’s list and every other file having dependency on extended vowels were changed with this new format.

**Results**



**Merits of this new set:**

1. Increased readability as there are no random/rubbish characters like t3, u’, ku’, k’ etc recognized during the live recognition which HTK used to have.
2. Better recognition of all the extended vowels.
3. The PREEMCOEF coefficient taken for live recognition is 0.90 because the recognition was very sensitive to the high frequencies so in order to tackle that, I reduced the high frequency boost.
4. Same idea was then used for the newer version of the MPAid and resulted in improved accuracy (76%) for both genders. And solved the character set problem as well.

**TASK 6**

**Impact of NUMCHANS (FILTER BANKS) in order to improve the recognition rate for Females**

As we know that the women voice lies in high frequency range as compared to the men, so the filter banks (see page 65 of HTK book) are spread over the frequency range in the such a way that they are much more close in the lower frequency range and the separation as we move towards higher frequency range. So, we thought that by increasing the filter banks we could decrease the separation between filter banks at higher frequency region resulting in improved recognition but it did not work that well. (Blue boxes in the following table represents the accuracy values for standard values of NUMCHANS).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SENTENTENCE ACCURACY** | **SENTENCE ACCURACY** | **WORD**  **ACCURACY** | **WORD**  **ACCURACY** |
|  | **PREEMPHASIS=**0.90 | **PREEMPHASIS=**0.80 | **PREEMPHASIS=**0.90 | **PREEMPHASIS=**0.80 |
| **NUMCHANS=30** |  |  |  |  |
| Female | 59.61 | 60.34 | 66.26 | 66.75 |
| Male | 66.14 | 67.41 | 68.67 | 70.25 |
| Both | 62.47 | 63.43 | 67.31 | 68.28 |
| **NUMCHANS=28** |  |  |  |  |
| Female | 61.82 | 60.59 | 67.49 | 67.49 |
| Male | 68.35 | 69.30 | 70.25 | 72.15 |
| Both | 64.68 | 64.40 | 68.70 | 69.53 |
| **NUMCHANS=26** |  |  |  |  |
| Female | 63.05 | 61.82 | 67.98 | 67.49 |
| Male | 69.30 | 67.41 | 71.20 | 69.49 |
| Both | 65.79 | 64.27 | 69.39 | 68.56 |
| **NUMCHANS=24** |  |  |  |  |
| Female | 61.82 | 61.82 | 67.98 | 67.73 |
| Male | 69.30 | 68.67 | 71.20 | 70.89 |
| Both | 65.10 | 64.82 | 69.39 | 69.11 |
| **NUMCHANS=22** |  |  |  |  |
| Female | 63.30 | 63.08 | 68.23 | 67.98 |
| Male | 68.67 | 68.67 | 70.57 | 70.57 |
| Both | 65.50 | 65.51 | 69.25 | 69.11 |

**CONCLUSIONS:**

1. Word and Sentence accuracies of the models trained on female data decreases with the increase and decrease in the value of the NUMCHANS (i.e. number of filter banks).
2. The accuracies are more as we increase the PREEMPHASIS coefficient.
3. Even on the female-data trained model set the accuracies are more for the male test recordings.

**WORD N-BEST RECOGNITION:**

HVite -H hmm15/macros -H hmm15/hmmdefs -C user/config6 -w user/wordNetwork -t 250 **-n 3** **5** -e -i output.mlf -g -p 0.0 -s 5.0 user/dictionary user/tiedList

This is the command to get best ‘n’ words (n=5 here, for 3 tokens) for one utterance during live recognition, along with their corresponding average log likelihood values. Fro this, run the above command which will start the live recognizer with which you can perform few live recordings and check the corresponding results (The results shown during live recognition would of the higher average log likelihood values (means lesser negative value) from the n-best list generated for each recording that you can check in ‘output.mlf’ file).

**Way 2:** Run the batch file **‘livetest’** present in the **‘Batches’** folder of **‘HTK’** folder and see the results of n-best recognition in the **‘output.mlf’** file. (if ‘-n 3 5’ is not present in the batch file, just add it by yourself).

**HOW TO TRAIN HMM MODELS:**

See Annie’s report (Page No. 14, APPENDIX A- HTK Commands used) in order to start with the HMM model generation. The only hard part is the generation of first model ie. Hmm0. In order to create this a very good tutorial is provided by voxforge.[a] After running the ‘HCompv’ command the next steps can be followed as given in the link.

[a]<http://www.voxforge.org/home/dev/acousticmodels/linux/create/htkjulius/tutorial/monophones/step-6>