Speaker Verification System
Part A Final Presentation

Performed by: Barak Benita & Daniel Adler

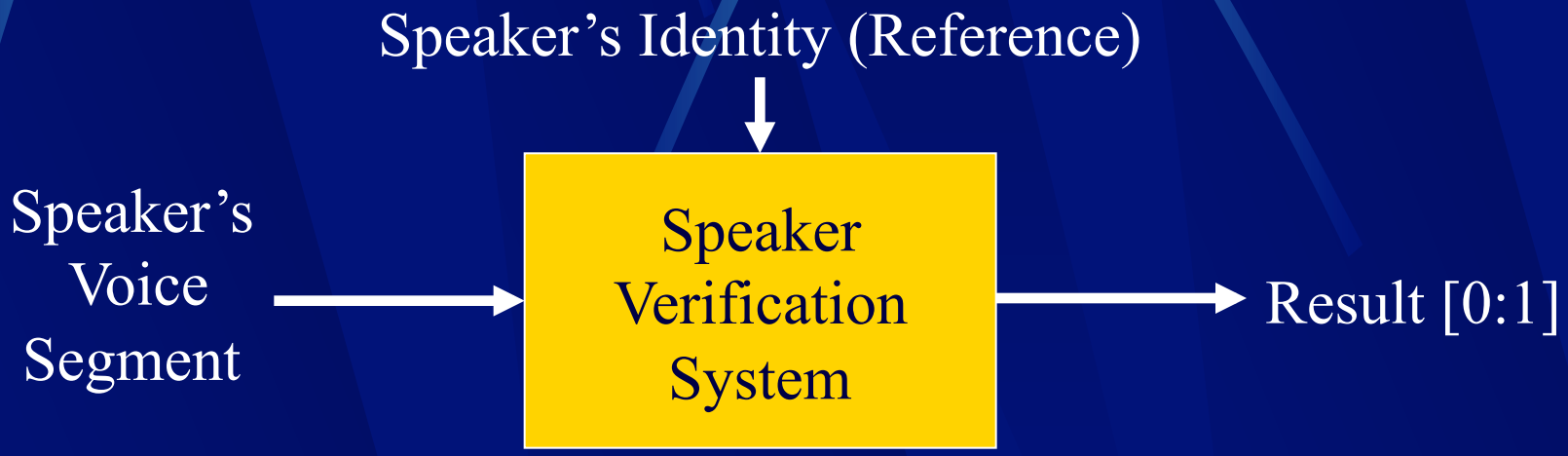
Instructor: Erez Sabbag

The Project Goal:

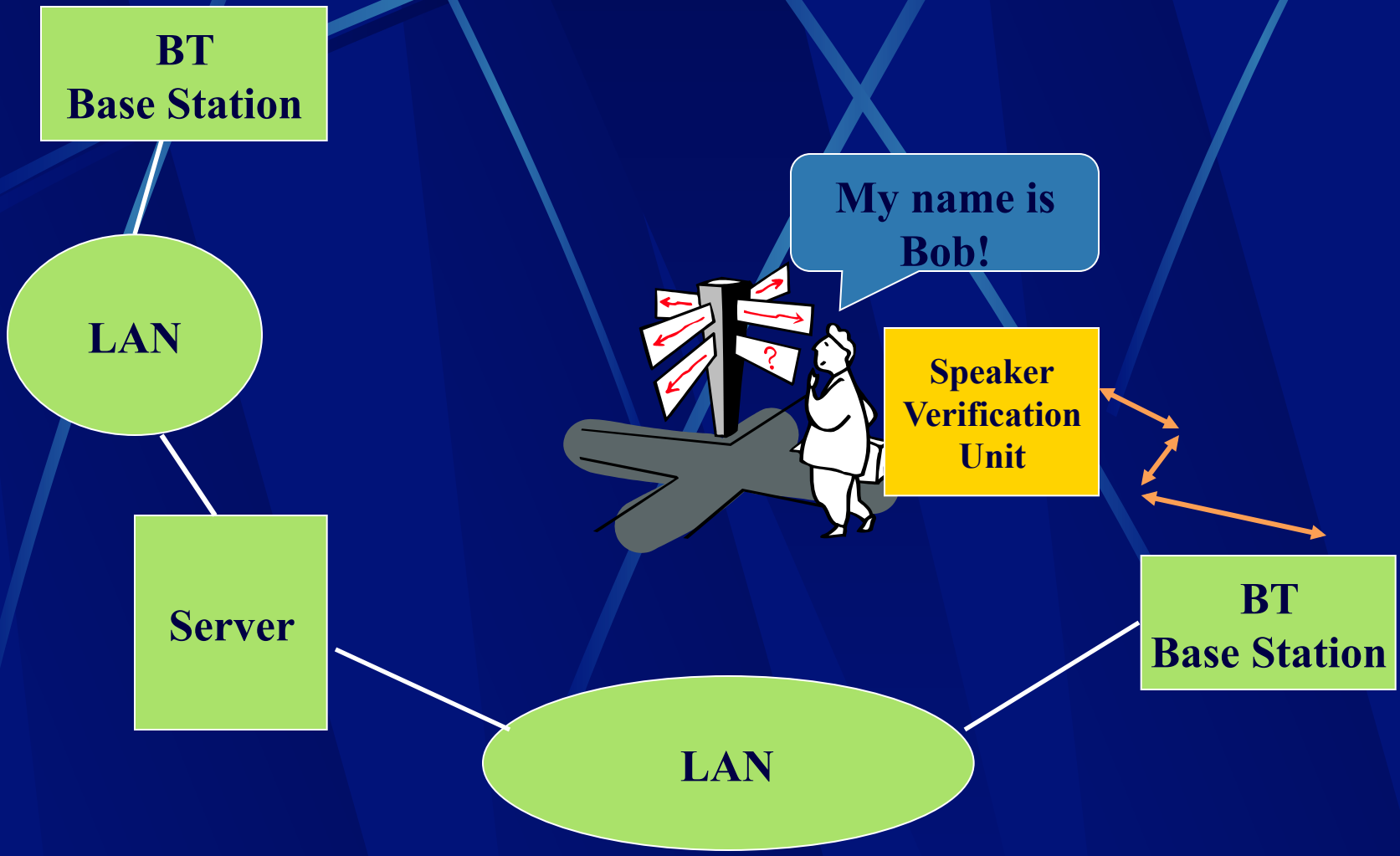
Implementation of a speaker verification algorithm on a TI 54X DSP

Introduction

Speaker verification is the process of automatically authenticating the speaker on the basis of individual information included in speech waves.



System Overview:



Project Description:

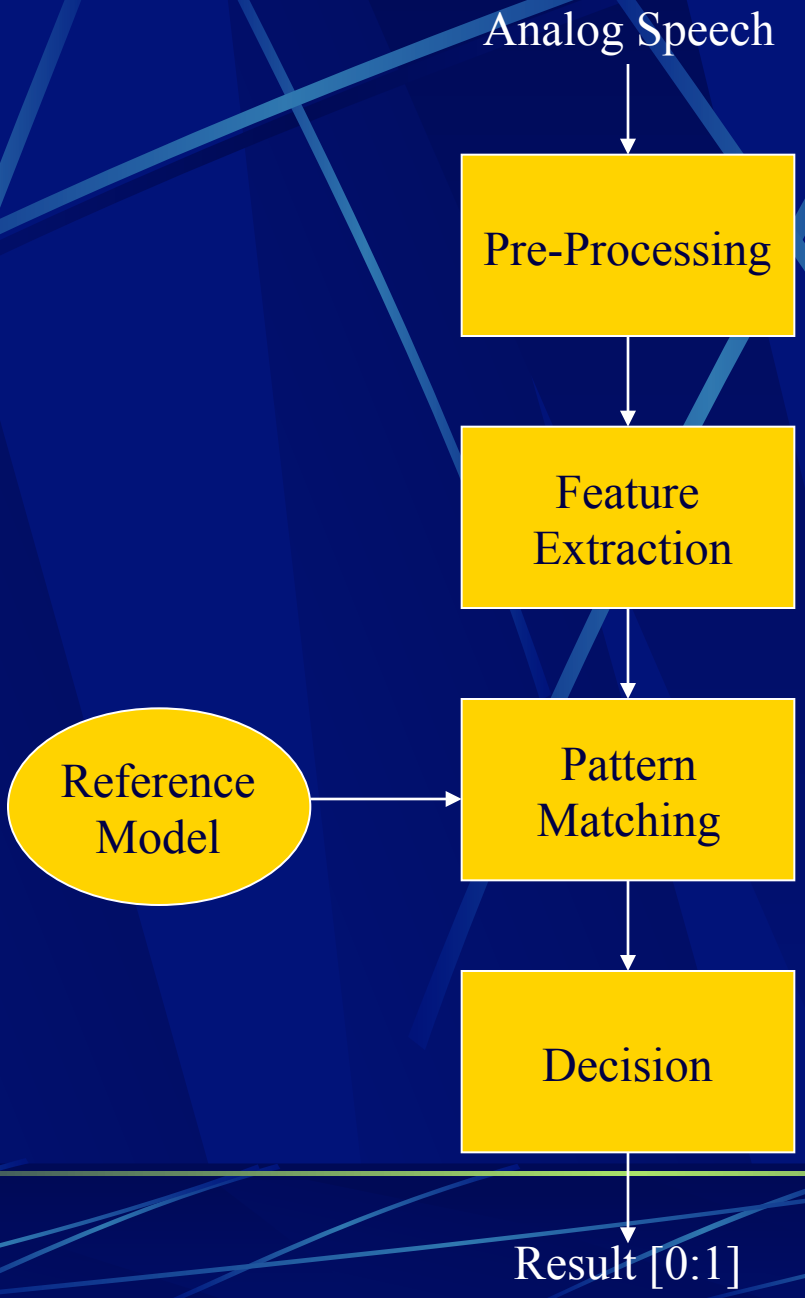
Part One:

- **Literature review**
- **Algorithms selection**
- **MATLAB implementation**
- **Result analysis**

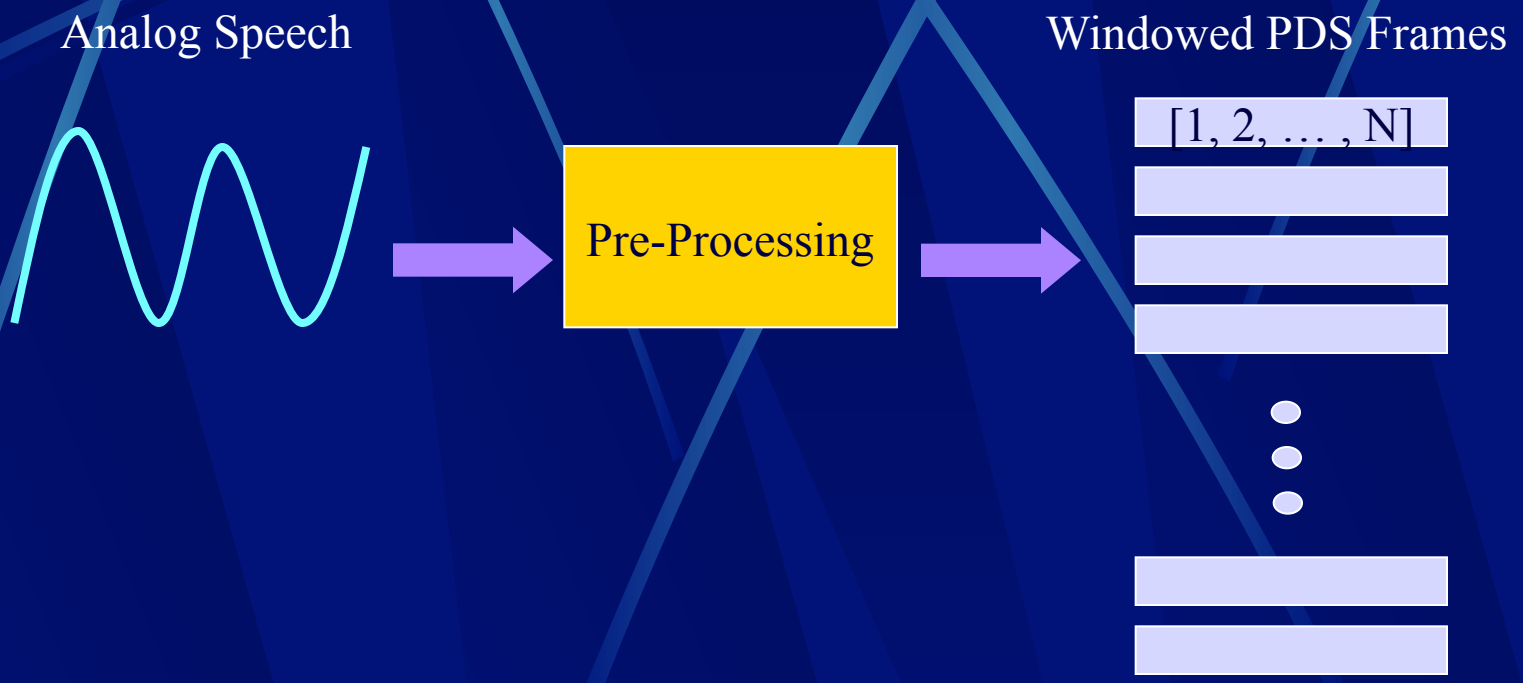
Part Two:

- **Implementation of the chosen algorithm on a DSP**

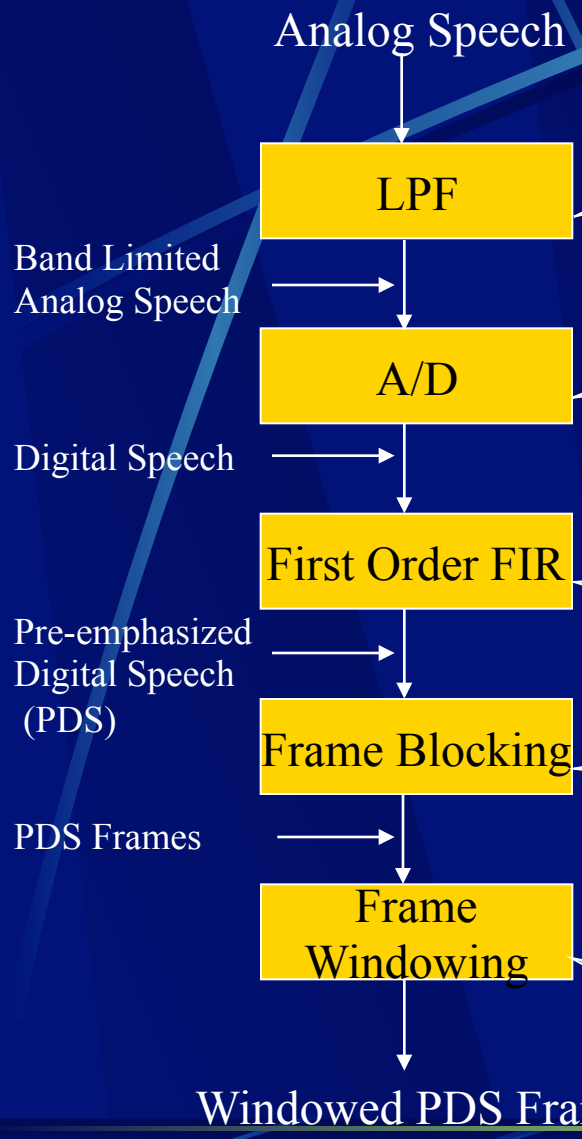
Speaker Verification System – Block Diagram



Pre-Processing (step 1)



Pre-Processing module



Anti aliasing filter to avoid aliasing during sampling. LPF [0, Fs/2]

Analog to digital converter with frequency sampling (Fs) of [10,16]KHz

Low order digital system to spectrally flatten the signal (in favor of vocal tract parameters), and make it less susceptible to later finite precision effects

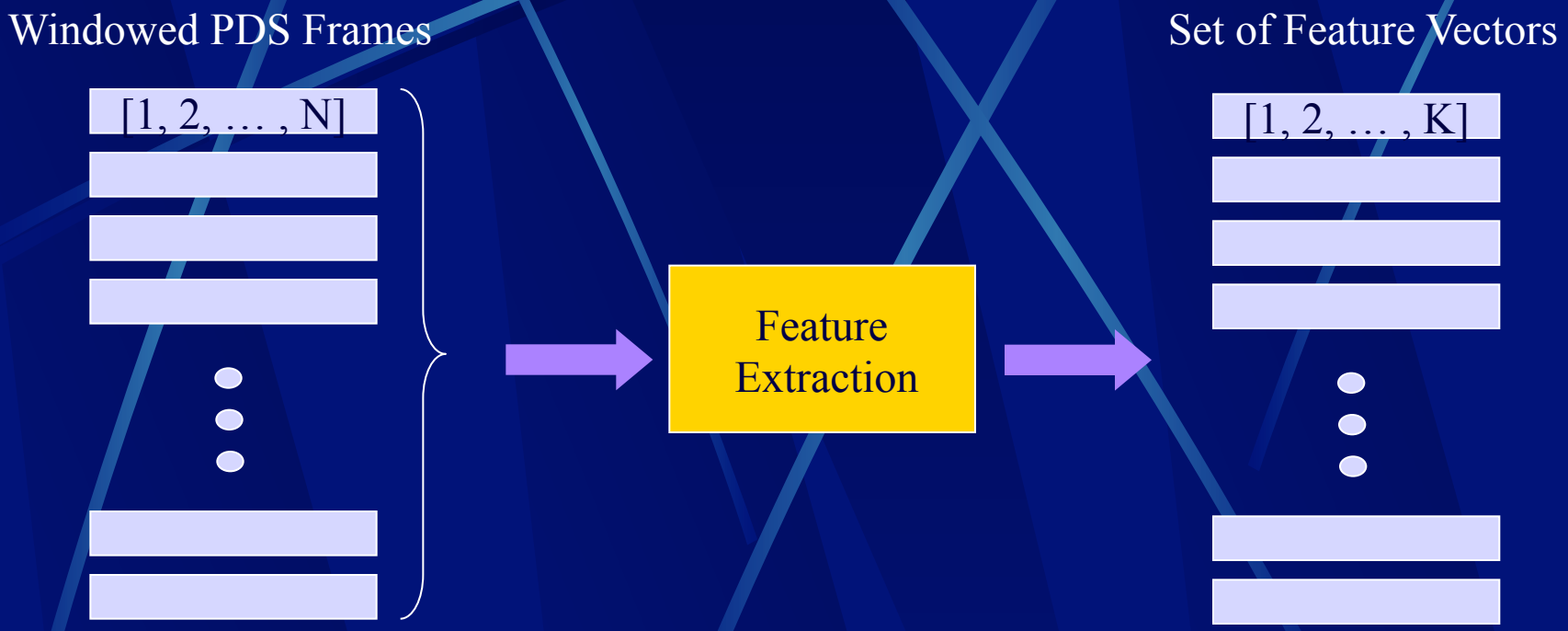
$$H(Z) = Z - \alpha Z^{-1}$$

Frame blocking of the sampled signal. Each frame is of N samples overlapped with N-M samples of the previous frame.
Frame rate ~ 100 Frames/Sec
N values: [200,300], M values: [100,200]

Using Hamming (or Hanning or Blackman) windowing in order to minimize the signal discontinuities at the beginning and end of each frame.

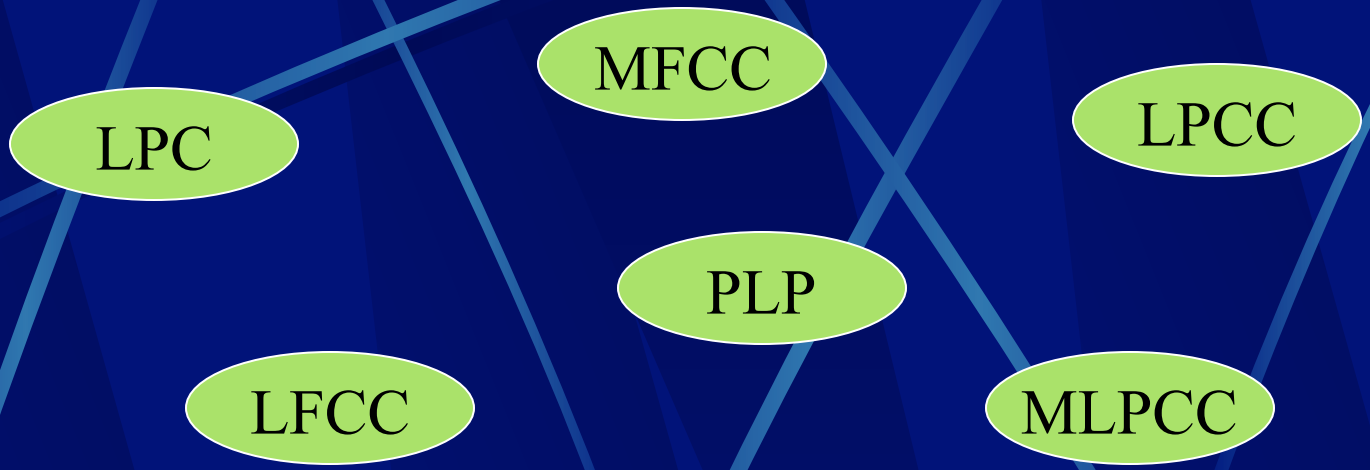
$$W(n) = 0.54 - 0.46\cos(\frac{2\Pi n}{N-1}),\ 0 \le n \le N-1$$

Feature Extraction (step 2)



Extracting the features of speech from each frame and representing it in a vector (feature vector).

Feature Extraction Methods



And the Winners are…

LPC

Linear Prediction Coeff

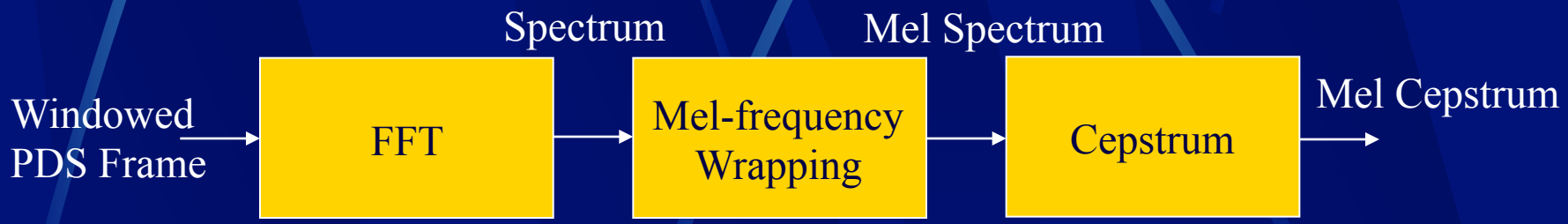
MFCC

Mel Freq Cepstral Coeff

For widely spread in many application and being prototypes for many other variant methods

Feature Extraction Module – MFCC

MFCC (Mel Frequency Cepstral Coefficients) is the most common technique for feature extraction. MFCC tries to mimic the way our ears work by analyzing the speech waves linearly at low frequencies and logarithmically at high frequencies.
The idea acts as follows:



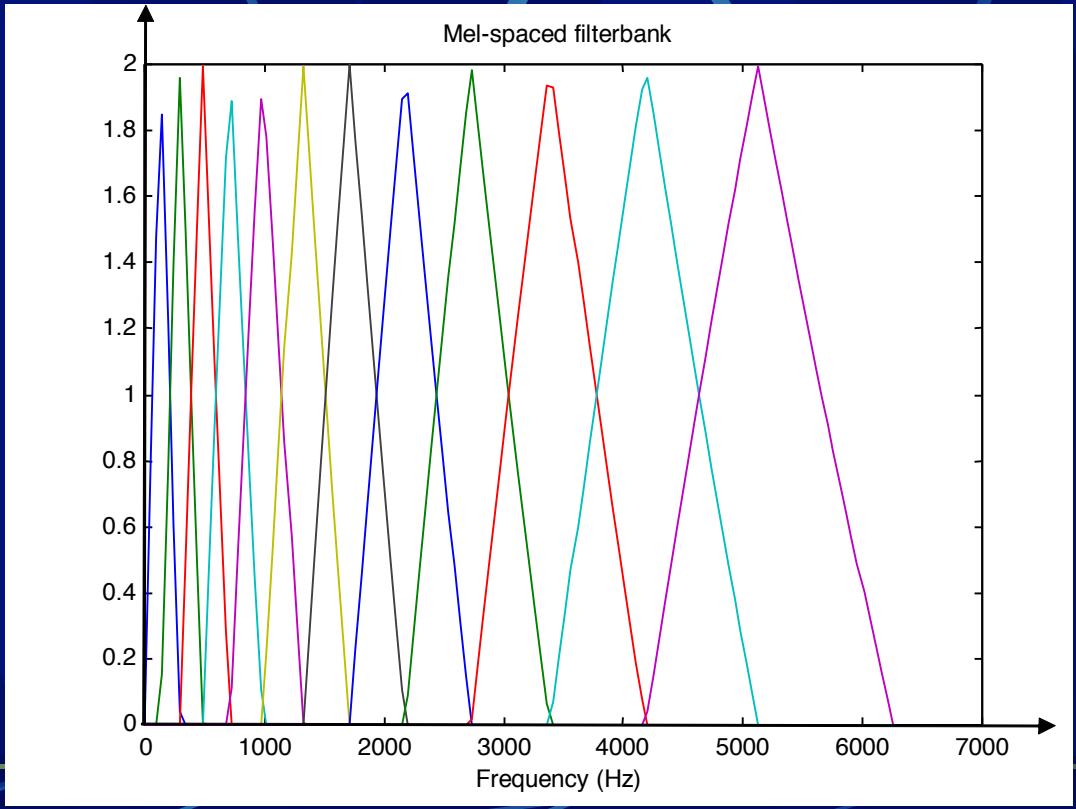
MFCC – Mel-frequency Wrapping

Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, *f*, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The *mel-frequency* scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Therefore we can use the following approximate formula to compute the mels for a given frequency *f* in Hz:

$$mel(f) = 2595 * \log_{10}(1 + f / 700)$$

MFCC – Filter Bank

One way to simulating the spectrum is by using a filter bank, spaced uniformly on the mel scale. That filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval.



MFCC – Cepstrum

Here, we convert the log mel spectrum back to time. The result is called the mel frequency cepstrum coefficients (MFCC). Because the mel spectrum coefficients are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT) and get a featured vector.

Feature Extraction Module – LPC

LPC (Linear Prediction Coefficients) is a method of extracting the features of speech from a speech signal .LPC encodes a signal by finding a set of weights on earlier signal values that can predict the next signal value:

$$S(k) = \sum_{m=1}^{K_{LPC}} a_{LPC(m)} \cdot S(k-m) + Error(k)$$

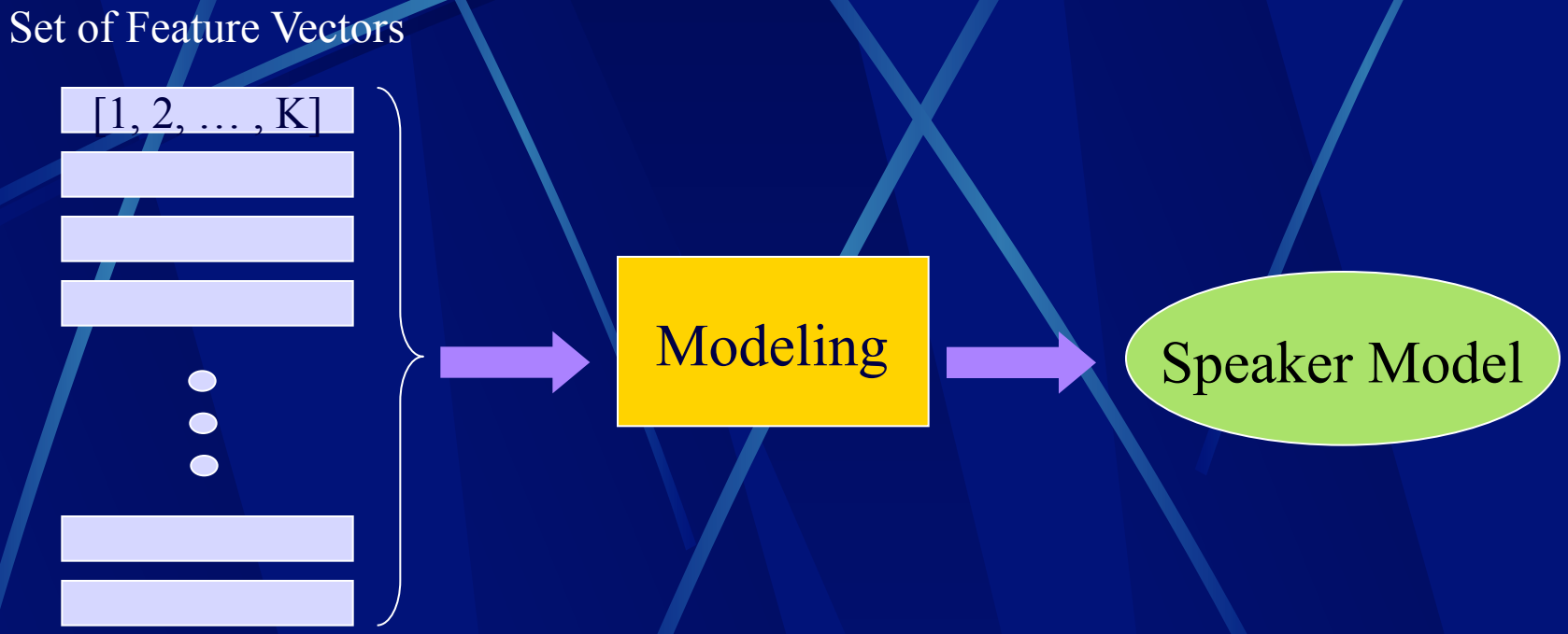
If values for a[1..k] can be found such that Error[k] is very small for a stretch of speech (say one analysis window), then we can represent the speech features with a[1..k] instead of the signal values in the window. The result of LPC analysis then is a set of coefficients a[1..k] and an error signal Error[k].

Pattern Matching Modeling (step 3)

The pattern matching modeling techniques is divided into two sections:

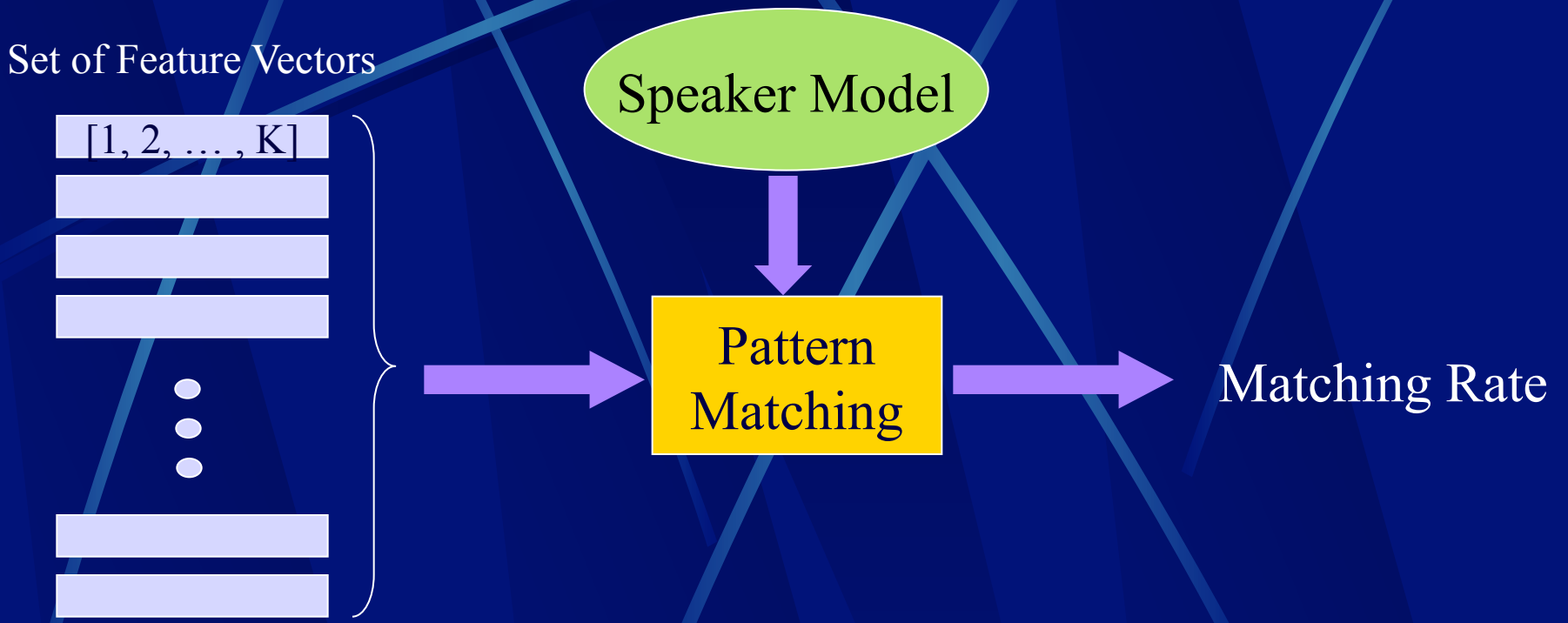
- The enrolment part, in which we build the reference model of the speaker.
- The verifications (matching) part, where the users will be compared to this model.

Enrollment part – Modeling



This part is done outside the DSP and the DSP receives only the speaker model (calculated offline in a host).

Pattern Matching



Pattern Matching - Modeling Methods

The most common methods for pattern matching - modeling in the speaker recognition world

	VQ	DTW	CHMM variants	VQ + DHMM	Neural Network
Implementation	Simple	Simple	Complex	Complex	Medium
Text Dependent / Independent	TI / TD	TD	TI / TD	TI / TD	TI / TD
Popularity	High	Medium	High	Medium	Low (growing)
*** Performance** (according to research reports)	Medium / High	Low / Medium	Medium / High	Medium / High	Medium
Challenge	Simple	Simple	High	High	High

Pattern Matching - Modeling Methods Cont.

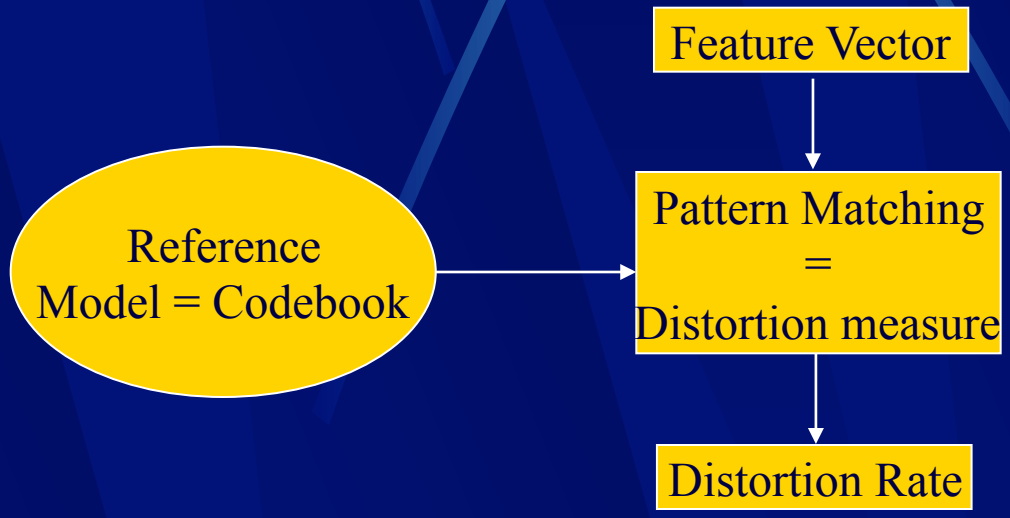
And the Oscar goes to …

- VQ (Vector Quantization)
  Model = Codebook
  Matching Criteria = Distance between vector and the nearest codebook centroid.

- CHMM (Continuous Hidden Markov Model)
  Model = HMM
  Matching Criteria = Probability score

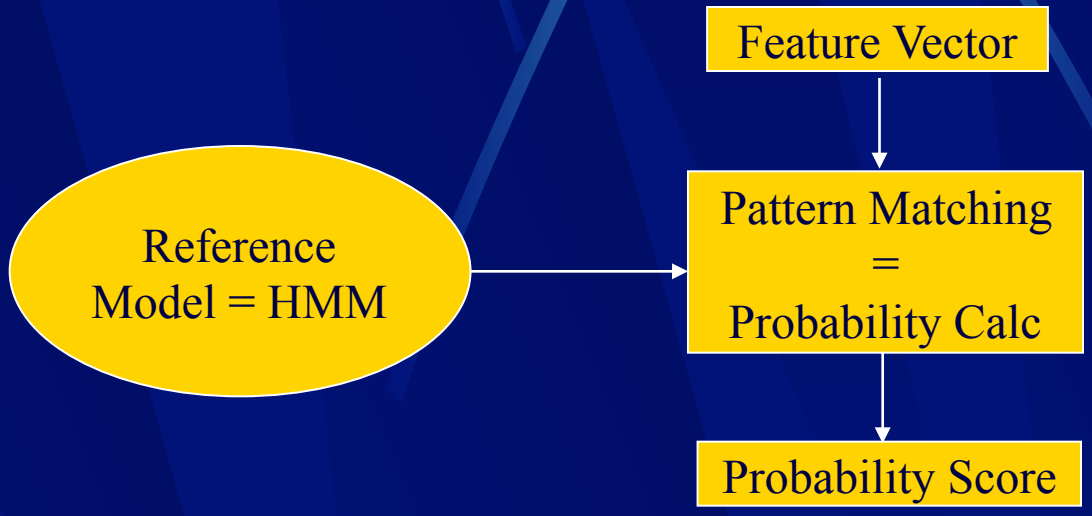
Pattern Matching Modeling Module – Vector Quantization (VQ)

In the enrolment part we build a codebook of the speaker according to the LBG (Linde, Buzo, Gray) algorithm, which creates an N size codebook from set of L feature vectors.
In the verification stage, we are measuring the distortion of the given sequence of the feature vectors to the reference codebook.



Pattern Matching Modeling Module – Hidden Markov Model (HMM)

In the enrolment stage we build an HMM for the specific speaker (this procedure creates the following outputs: A and B matrix, vector). The building of the model is done by using the Baum-Welch algorithm. In the matching procedure, we compute the matching probability of the current speaker with the model. This is done by the Viterbi algorithm.



Decision Module (Optional)

In VQ the decision is based on checking if the distortion rate is higher than a preset threshold:
if distortion rate > t, Output = Yes,
else Output = No.

In HMM the decision is based on checking if the probability score is higher than a preset threshold:
if probability scores > t, Output = Yes,
else Output = No.

Experiment Description

The Voice Database:

- Two reference models were generated (one male and one female), each model was trained in 3 different ways:
  - repeating the same sentence for 15 seconds
  - repeating the same sentence for 40 seconds
  - reading random text for one minute
- The voice database is compound from 10 different speakers (5 males and 5 females), each speaker was recorded in 3 ways:
  - repeating the reference sentence once (5 seconds)
  - repeating the reference sentence 3 times (15 seconds)
  - speaking a random sentence for 5 seconds

Experiment Description Cont.

The Tested Verification Systems:

- System one: LPC + VQ
- System two: MFCC + VQ
- System three: LPC + CHMM
- System four: MFCC + CHMM

Experiment Description Cont.

Model Parameters:

- Number of coefficients of the feature vector (12 or 18)
- Frame size (256 or 330)
- Offset size (128 or 110)
- Sampling rate is 11025Hz
- Codebook size (64 or 128)
- Number of iterations for codebook creation (15 or 25)

Experiment Description Cont.

Conclusions number 1:

MFCC performs better than LPC

SYSTEM 1: MFCC + VQ (frame size=330, offset=110, vector length=12,booksize=64,it=15)

Test format	Ref format	W1	W2	W3	W4	W5	Ref model	M2	M3	M4	M5
5 fix	15 fix	0.3257	0.3253	0.2953	0.2199	0.5733	0.2694	0.5022	0.3478	0.485	0.5636
15 fix	15 fix	0.4222	0.3287	0.2394	0.1942	0.3954	0.0451	0.4326	0.3968	0.6347	0.3767
5 dif	15 fix	0.6627	0.231	0.4556	0.1385	0.4211	0.3326	0.6125	0.281	0.5472	0.6305
5 fix	40 fix	0.2669	0.2347	0.2337	0.1833	0.3575	0.2069	0.3207	0.2278	0.3587	0.3137
15 fix	40 fix	0.3247	0.1191	0.1993	0.1294	0.2941	0.078	0.2907	0.2469	0.3995	0.2566
5 dif	40 fix	0.4498	0.1785	0.2647	0.1232	0.2542	0.2178	0.403	0.2338	0.3616	0.3439
5 fix	60 dif	0.2427	0.2357	0.2766	0.2056	0.3593	0.2232	0.2596	0.1849	0.3338	0.2583
15 fix	60 dif	0.3138	0.1865	0.202	0.1316	0.2943	0.0933	0.2403	0.212	0.3397	0.2226
5 dif	60 dif	0.4636	0.1744	0.2777	0.1202	0.2651	0.1688	0.3145	0.1776	0.2629	0.3052

SYSTEM 1: LPC + VQ (frame size=330, offset=110, vector length=12,booksize=64,it=15)

Test format	Ref format	W1	W2	W3	W4	W5	Ref model	M2	M3	M4	M5
5 fix	15 fix	0.0057	0.0327	0.0194	0.0023	0.2407	0.005	0.0147	0.0258	0.0207	0.0376
15 fix	15 fix	0.0238	0.1967	0.0122	0.0538	0.0586	0.000269	0.0155	0.0299	0.0235	0.0248
5 dif	15 fix	0.0184	0.0025	0.0238	0.0013	0.1276	0.1941	0.0093	0.0013	0.0177	0.0277
5 fix	40 fix	0.005	0.016	0.0136	0.0024	0.0814	0.0036	0.0058	0.0146	0.0076	0.0082
15 fix	40 fix	0.0142	0.123	0.007	0.0203	0.0175	0.000882	0.0114	0.0143	0.0137	0.0056
5 dif	40 fix	0.0122	0.0025	0.0102	0.0014	0.0204	0.0714	0.0057	0.000937	0.0138	0.0082
5 fix	60 dif	0.003	0.0107	0.01027	0.0024	0.0243	0.0035	0.0056	0.0113	0.0074	0.006
15 fix	60 dif	0.0115	0.0178	0.0059	0.0137	0.0188	0.0014	0.0069	0.0109	0.0105	0.0052
5 dif	60 dif	0.0109	0.0037	0.0066	0.0015	0.0327	0.0098	0.0046	0.000875	0.007	0.0078

Experiment Description Cont.

Conclusions number 2:

Window size of 330 and offset of 110 samples performs better than window size of 256 and offset of 128 samples

SYSTEM 1: MFCC + VQ (frame size=256, offset=128, vector length=12,booksize=64,it=15)

Test format	Ref format	W1	W2	W3	W4	W5	Ref model	M2	M3	M4	M5
5 fix	15 fix	0.3972	0.5082	0.3322	0.3381	0.3375	0.5723	0.3627	0.3023	0.461	0.4081
15 fix	15 fix	0.414	0.4068	0.2701	0.1831	0.4785	0.0855	0.3758	0.3757	0.5243	0.4477
5 dif	15 fix	0.5166	0.2434	0.4849	0.3541	0.2993	0.3071	0.6007	0.4247	0.5278	0.5504
5 fix	40 fix	0.3925	0.4425	0.3125	0.3254	0.3363	0.4058	0.3455	0.2278	0.3986	0.3978
15 fix	40 fix	0.4023	0.335	0.2459	0.1775	0.4116	0.1566	0.3401	0.3027	0.4528	0.392
5 dif	40 fix	0.5176	0.2325	0.4807	0.3438	0.3004	0.2659	0.593	0.423	0.4778	0.4996
5 fix	60 dif	0.3794	0.3988	0.29	0.2981	0.33	0.4658	0.2691	0.2488	0.3809	0.3496
15 fix	60 dif	0.3673	0.3355	0.2636	0.1623	0.3596	0.1632	0.2769	0.2962	0.4433	0.3515
5 dif	60 dif	0.4826	0.1973	0.4427	0.2659	0.2679	0.2132	0.4195	0.3158	0.35	0.4068

SYSTEM 1: MFCC + VQ (frame size=330, offset=110, vector length=12,booksize=64,it=15)

Test format	Ref format	W1	W2	W3	W4	W5	Ref model	M2	M3	M4	M5
5 fix	15 fix	0.3257	0.3253	0.2953	0.2199	0.5733	0.2694	0.5022	0.3478	0.485	0.5636
15 fix	15 fix	0.4222	0.3287	0.2394	0.1942	0.3954	0.0451	0.4326	0.3968	0.6347	0.3767
5 dif	15 fix	0.6627	0.231	0.4556	0.1385	0.4211	0.3326	0.6125	0.281	0.5472	0.6305
5 fix	40 fix	0.2669	0.2347	0.2337	0.1833	0.3575	0.2069	0.3207	0.2278	0.3587	0.3137
15 fix	40 fix	0.3247	0.1191	0.1993	0.1294	0.2941	0.078	0.2907	0.2469	0.3995	0.2566
5 dif	40 fix	0.4498	0.1785	0.2647	0.1232	0.2542	0.2178	0.403	0.2338	0.3616	0.3439
5 fix	60 dif	0.2427	0.2357	0.2766	0.2056	0.3593	0.2232	0.2596	0.1849	0.3338	0.2583
15 fix	60 dif	0.3138	0.1865	0.202	0.1316	0.2943	0.0933	0.2403	0.212	0.3397	0.2226
5 dif	60 dif	0.4636	0.1744	0.2777	0.1202	0.2651	0.1688	0.3145	0.1776	0.2629	0.3052

Experiment Description Cont.

Conclusions number 3:

Feature vector of 18 coeffs is better than feature vector of 12 coeffs

SYSTEM 1: MFCC + VQ (frame size=256, offset=128, vector length=12,booksize=64,it=15)

Test format	Ref format	W1	W2	W3	W4	W5	Ref model	M2	M3	M4	M5
5 fix	15 fix	0.3972	0.5082	0.3322	0.3381	0.3375	0.5723	0.3627	0.3023	0.461	0.4081
15 fix	15 fix	0.414	0.4068	0.2701	0.1831	0.4785	0.0855	0.3758	0.3757	0.5243	0.4477
5 dif	15 fix	0.5166	0.2434	0.4849	0.3541	0.2993	0.3071	0.6007	0.4247	0.5278	0.5504
5 fix	40 fix	0.3925	0.4425	0.3125	0.3254	0.3363	0.4058	0.3455	0.2278	0.3986	0.3978
15 fix	40 fix	0.4023	0.335	0.2459	0.1775	0.4116	0.1566	0.3401	0.3027	0.4528	0.392
5 dif	40 fix	0.5176	0.2325	0.4807	0.3438	0.3004	0.2659	0.593	0.423	0.4778	0.4996
5 fix	60 dif	0.3794	0.3988	0.29	0.2981	0.33	0.4658	0.2691	0.2488	0.3809	0.3496
15 fix	60 dif	0.3673	0.3355	0.2636	0.1623	0.3596	0.1632	0.2769	0.2962	0.4433	0.3515
5 dif	60 dif	0.4826	0.1973	0.4427	0.2659	0.2679	0.2132	0.4195	0.3158	0.35	0.4068

SYSTEM 1: MFCC + VQ (frame size=256, offset=128, vector length=18,booksize=64,it=25)

Test format	Ref format	W1	W2	W3	W4	W5	Ref model	M2	M3	M4	M5
5 fix	15 fix	0.8154	0.9922	0.7438	0.8032	0.9057	1.1391	0.7442	0.6448	0.9821	0.8873
15 fix	15 fix	0.8557	0.7991	0.6167	0.4397	0.9995	0.1782	0.7672	0.7925	1.11	0.8639
5 dif	15 fix	1.0977	0.5512	1.0694	0.7901	0.681	0.6533	1.1692	0.8597	0.9929	1.0311
5 fix	40 fix	0.8647	0.8619	0.7368	0.7597	0.7895	0.789	0.7112	0.5384	0.8551	0.8543
15 fix	40 fix	0.818	0.6769	0.5871	0.4312	0.8438	0.3345	0.7109	0.6691	0.9276	0.8141
5 dif	40 fix	1.047	0.4799	0.9706	0.7869	0.6297	0.5145	.1.2128	0.8389	0.8875	0.9378
5 fix	60 dif	0.781	0.7955	0.7016	0.7342	0.7508	0.8907	0.6077	0.5457	0.7809	0.7303
15 fix	60 dif	0.7371	0.6291	0.5792	0.4213	0.7758	0.3423	0.5887	0.6486	0.9136	0.7131
5 dif	60 dif	0.9588	0.4184	0.9513	0.6973	0.5561	0.4494	0.9532	0.7278	0.7144	0.7833

Experiment Description Cont.

Conclusions number 4:

Worst combinations:

- 5 seconds of fixed sentence for testing with an enrolment of 15 seconds of the same sentence.
- 5 seconds of fixed sentence for testing with an enrolment of 40 seconds of the same sentence.

Best combinations:

- 15 seconds of fixed sentence for testing with an enrolment of 40 seconds of the same sentence.
- 15 seconds of fixed sentence for testing with an enrolment of 60 seconds of random sentences.
- 5 seconds of a random sentence with an enrolment of 60 seconds of random sentences.

Experiment Description Cont.

The Best Results:

SYSTEM 1: MFCC + VQ (frame size=330, offset=110, vector length=18,booksize=128,it=25

Test format	Ref format	W1	W2	W3	W4	W5	Ref model	M2	M3	M4	M5
5 fix	15 fix	0.7202	0.6562	0.5986	0.4898	1.0638	0.5237	0.9571	0.6818	0.8953	1.1104
15 fix	15 fix	0.853	0.6342	0.4921	0.402	0.7028	0.0769	0.8601	0.7728	1.1329	0.7399
5 dif	15 fix	1.249	0.414	0.8638	0.3263	0.7745	0.6421	1.17	0.5797	0.9824	1.2382
5 fix	40 fix	0.5855	0.4604	0.5109	0.4563	0.6556	0.3783	0.5971	0.4195	0.6024	0.6319
15 fix	40 fix	0.6436	0.4041	0.4023	0.2978	0.5307	0.1541	0.5615	0.4826	0.665	0.4974
5 dif	40 fix	0.9107	0.3176	0.533	0.3009	0.5291	0.3862	0.7507	0.459	0.5913	0.6839
5 fix	60 dif	0.5269	0.4063	0.5026	0.4619	0.6194	0.3991	0.4757	0.3567	0.5282	0.4885
15 fix	60 dif	0.5792	0.3451	0.3821	0.2992	0.4961	0.1702	0.4617	0.405	0.5601	0.4162
5 dif	60 dif	0.7762	0.2962	0.5423	0.3012	0.4555	0.3324	0.5866	0.3973	0.4531	0.5716

Test format	Ref format	W1	Ref model	W3	W4	W5	M1	M2	M3	M4	M5
5 fix	15 fix	0.4671	0.4142	0.4239	0.4225	0.6485	0.4675	0.5539	0.4258	0.6562	0.6273
15 fix	15 fix	0.4924	0.116	0.358	0.3013	0.5259	0.1985	0.5313	0.4394	0.8425	0.4712
5 dif	15 fix	0.6923	0.3158	0.5259	0.3002	0.4525	0.4424	0.7459	0.5024	0.7325	0.6348
5 fix	40 fix	0.4709	0.3605	0.4633	0.4873	0.6435	0.4718	0.5098	0.3917	0.643	0.5728
15 fix	40 fix	0.4708	0.2923	0.3733	0.3038	0.4717	0.1919	0.4713	0.4161	0.7254	0.4085
5 dif	40 fix	0.6697	0.2852	0.6129	0.3169	0.4306	0.4295	0.5837	0.4088	0.6289	0.6273
5 fix	60 dif	0.4581	0.3597	0.4116	0.3662	0.6174	0.5032	0.4536	0.3593	0.6336	0.5008
15 fix	60 dif	0.4583	0.3358	0.374	0.2248	0.5238	0.1965	0.4333	0.3886	0.7011	0.3922
5 dif	60 dif	0.6575	0.2675	0.4952	0.2306	0.479	0.3746	0.5879	0.3704	0.625	0.5559

Time Table – First Semester

14.11.01 – Project description presentation

15.12.01 – completion of phase A: literature review and algorithm selection

25.12.01 – Handing out the mid-term report

25.12.01 – Beginning of phase B: algorithm implementation in MATLAB

10.04.02 – Publishing the MATLAB results and selecting the algorithm that will be implemented on the DSP

Time Table – Second Semester

10.04.02 – Presenting the progress and planning of the project to the supervisor

17.04.02 – Finishing MATLAB Testing

17.04.02 – The beginning of the implementation on the DSP

Summer – Project presentation and handing the project final report