Holly Quran-Based Arabic Text to Speech

By: Bana Akram Al-Sharif

Supervisor: Dr. Radwan Tahboub

Co-supervisor: Dr. Labib Arafeh

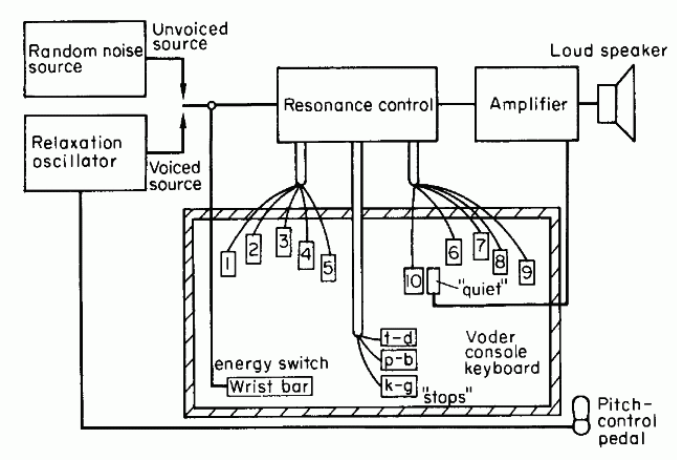
April,2014

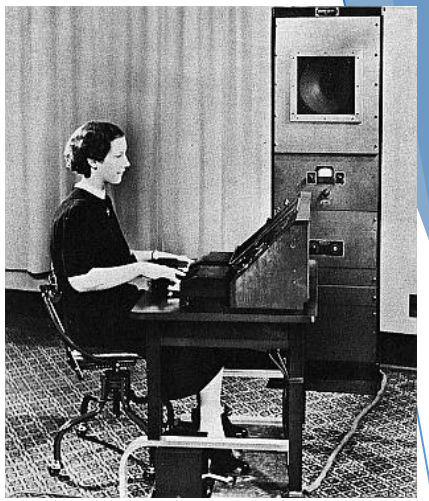
Outline

- Introduction
- Our System Description
- Evaluation
- To Do Procedure

Introduction

TTS History [18]





1939 Voder

Deutsches Museum (von Meisterwerken der Naturwissenschaft und Technik) in Munich, Germany
(Wolfgang von Kempelen)

- As a mechanic machine
  - (1003) Gerbert of Aurillac : "speaking heads“
  - (1198-1280) Albertus Magnus, and (1214–1294) Roger Bacon: Improvement in “speaking heads”
  - (1779) Christian Kratzenstein: Vocal tract that could produce the five long vowel sounds.
  - (1930) Bell Labs: Vocoder (voice coder)
- Electrical Speech Synthesizer
  - (1939) Homer Dudley: voder (Voice Demonstrator)
- Speech Synthesis by computers
  - (1970- up to now): concatenating (phones – diphones – etc.)
  - Since 10 years the English TTS development has greatest improvements, the challenge is to satisfying limited resource consumption (memory and CPU)[2]

Arabic TTS History

- The formal language in more than 24 countries[16,17]
- The fourth spoken language in world [17]
- Special and big religious value by being the language of the Holly Quran for more than 1.6 billion Muslims .[7]

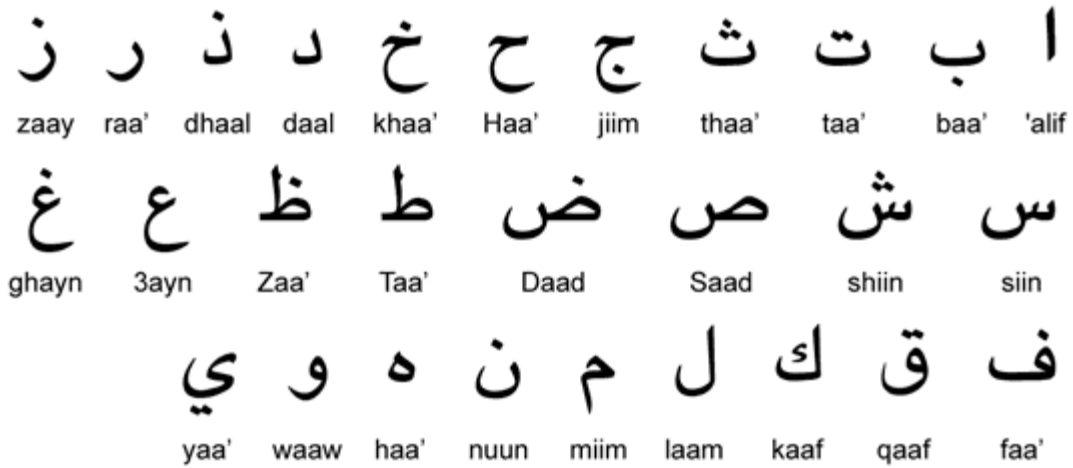
- Speech synthesis by concatenating sub-syllabic sound units

  El-Imam, Y.A. Publication Year: 1987 [7]

  ➔ 17 years after the birth of computers
- In 1997 publishing SAMPA (Speed Assessment Method Phonetic Alphabet) increase the interest in the ATTS[7]

Special properties in Arabic TTS

- Diacritic Arabic language considered as regular spelling language in comparison with French and English. [1][5][11]
- Milestones:
  - Corpus (data base)
  - Intonation and rhythm
  - Prosodic



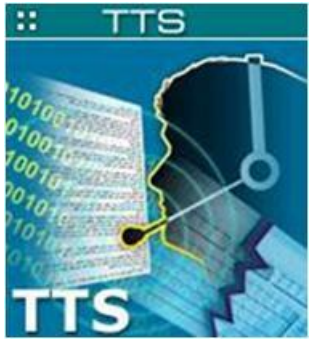
Comparison (Researches)

- From IEEE (recorded in April,10,2014)

	Arabic	English	*
Text to speech	108	359	4011
Tts	10	41	654
Text to voice	11	45	1012
Speech synthesis	53	285	8056
voice synthesis	9	45	1861
=	191	775	15594

- English has about 4 times Arabic
- * is 82 times Arabic

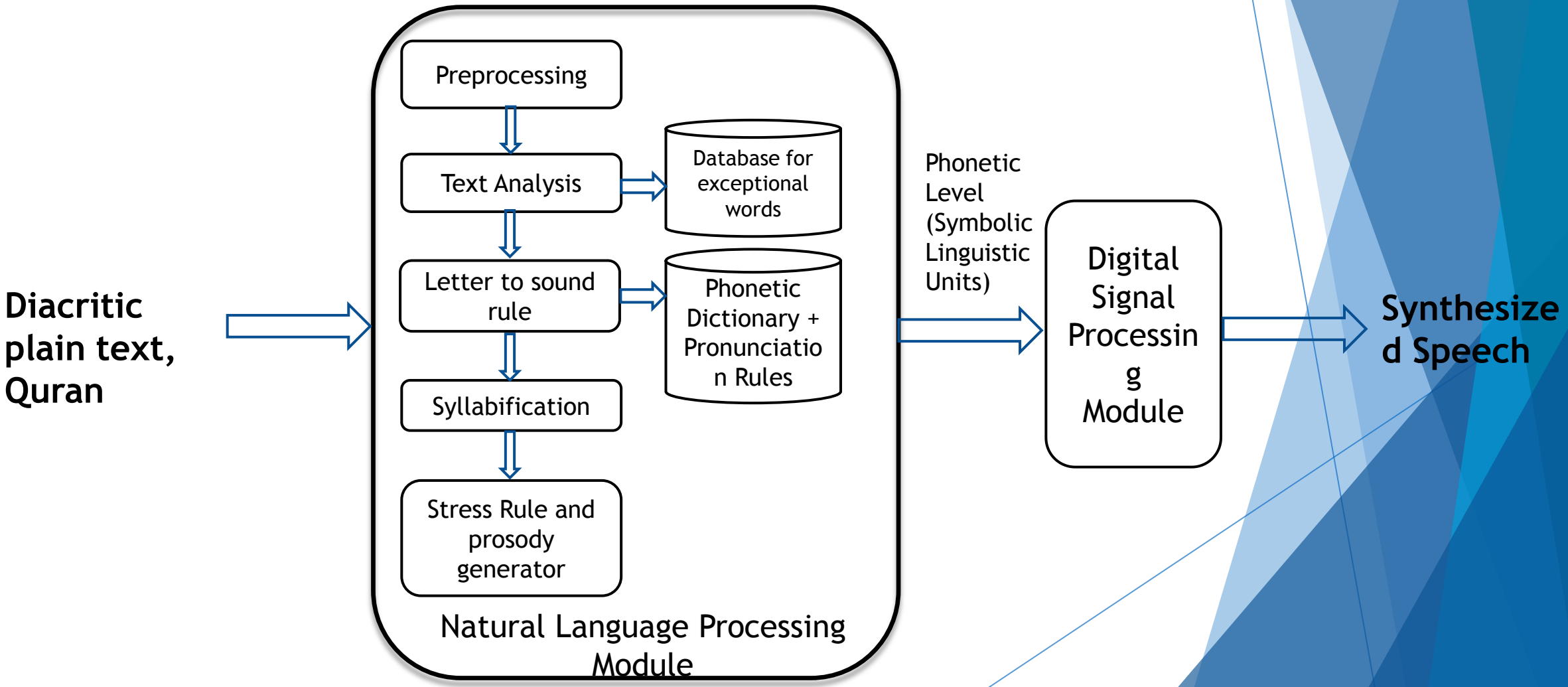
TTS



- “At the present state of the art, the limits of the achievable intelligibility and naturalness of synthetic speech are no longer set by technological factors, but rather by our limited knowledge about the acoustics and the perception of speech. In research, speech synthesis is used to test this knowledge.”[28]

- Science field: [2]
  - Artificial intelligence (AI) Computer science
  - Natural Language processing science
  - Algorithms (searching the corpus)

Typical TTS Components [2][5]



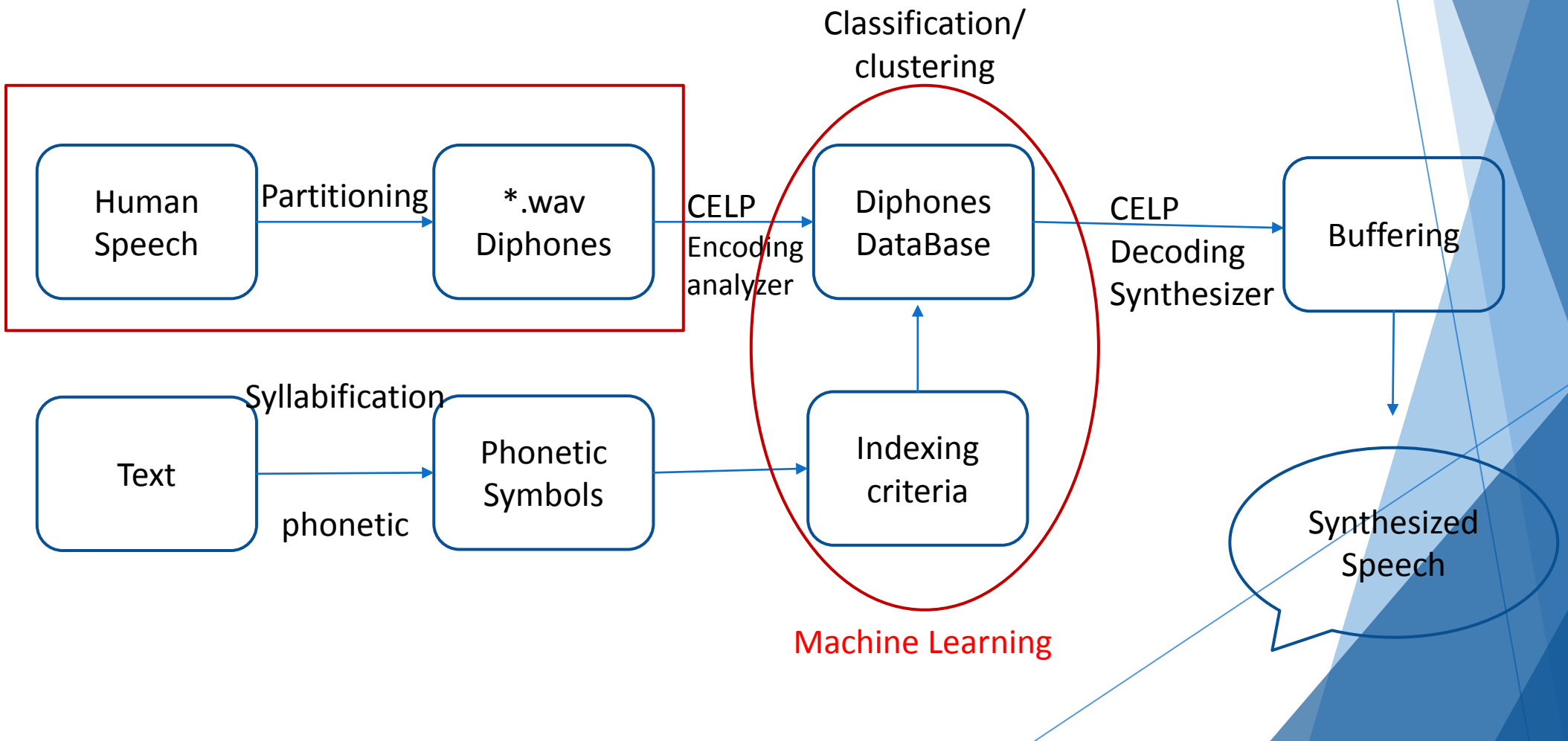
Our System Description

Characteristics in Holy Quran

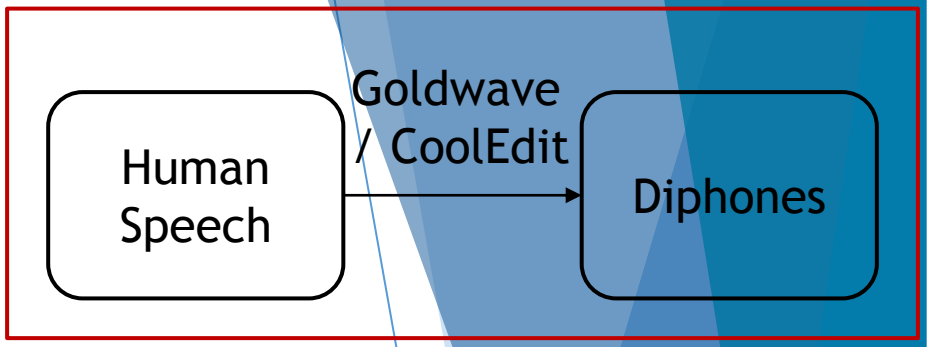
- The input text is from Holy Quran
- The output speech is a reciter's reading for verses

- Recitation full of features[13]
- Pure rule-based. [5,7]
- The Prosody are defines by Tajweed Rules.
- No breathing calculation
- No abbreviation, No annotation, No accents..[13]
- Well recorded speech for big data quantity, suitable for learning and testing

Our System Description

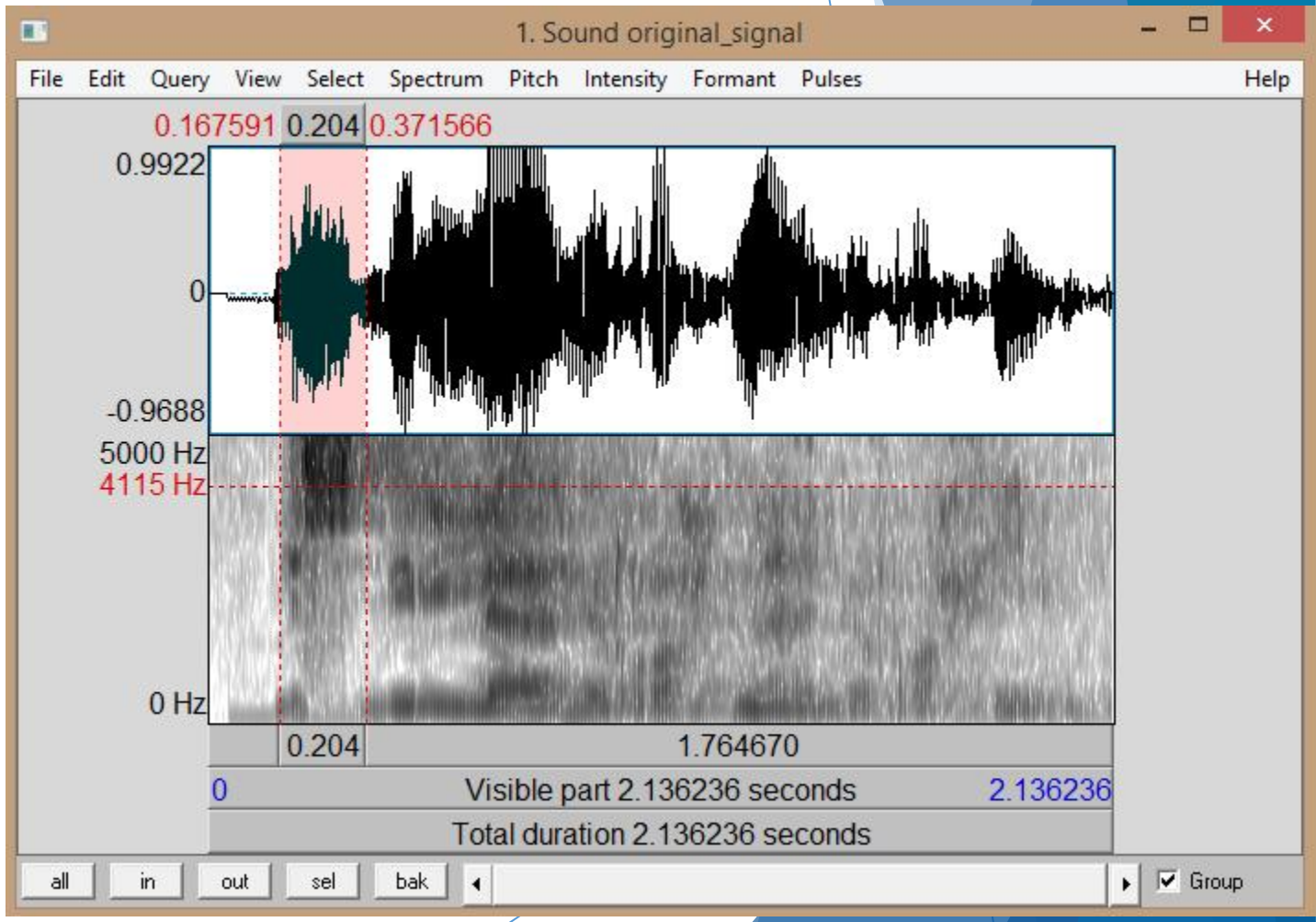


Build the diphones for the Data Base

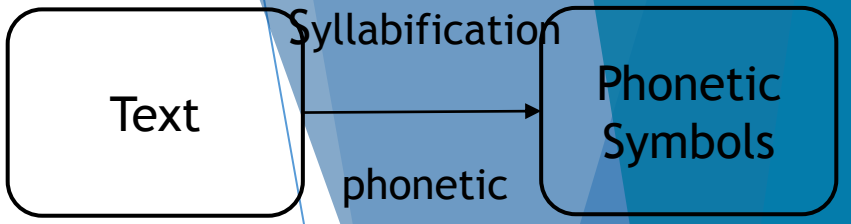


- We use the spectrogram of the signal to determine the diphones and cut it, then translating by CELP parameters, after that save it in a text file with its label.

- Programs Goldwave, Praat



Phonetization



- Example Phonetic symbols with prosody and stress level information[3]

(بِسْمِ اللَّهِ الرَّحْمنِ الرَّحِيم)

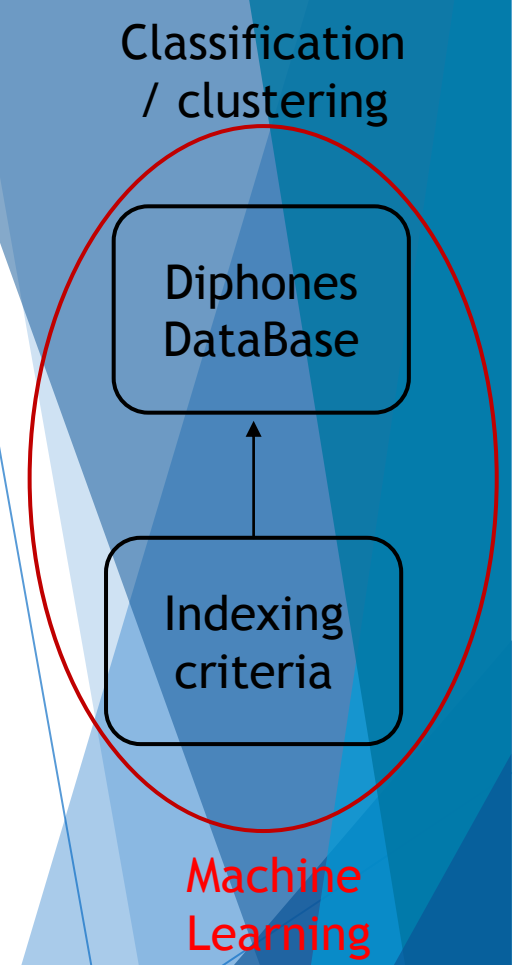
- Bis mil la:hir raX ma:nir raXi:m
- (CVV|CVV:CVC:CVC|CVV|CV|CVV|CV :CVV|CV|CVC:CV|CVV|CVC)
- CVV(WS)|CVV(PS):CVC(PS):CVC(WS)| CVV(SS)|CV(WS)||CVV(PS)|CV(WS)|: CVV(WS)|CV(WS)||CVC(PS):CV(WS)| |CVV(WS)|CVC(PS).

The Speech Assessment Methods Phonetic Alphabet (SAMPA) for Arabic

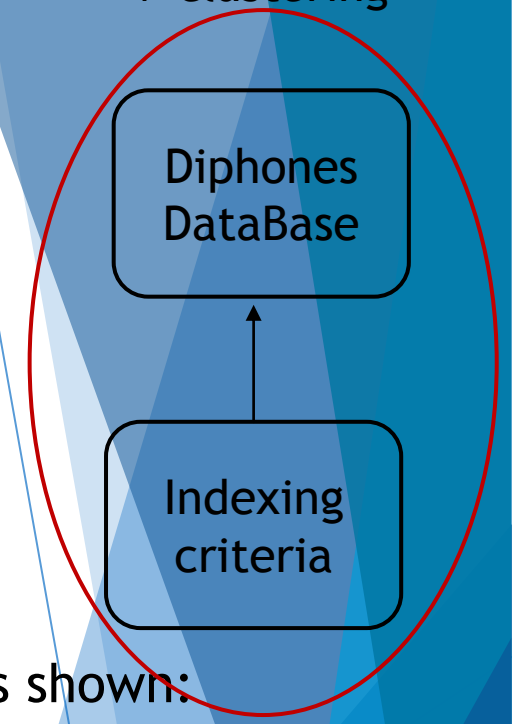
Arabic grapheme	Phonemic symbol	Arabic grapheme	Phonemic symbol
Consonants			
ء	/ʔ/	ض	/dˤ/
ب	/b/	ط	/tˤ/
ت	/t/	ظ	/Dˤ/
ث	/T/	ع	/ʔˤ/, /ʔ\/
ج	/g/	غ	/G/
خ	/x/	ف	/f/
ح	/X	ق	/q/
د	/d/	ك	/k/
ذ	/D/	ل	/l/
ر	/r/	م	/m/
ز	/z/	ن	/n/
س	/s/	ه	/h/
ش	/S/	و	/w/
ص	/sˤ/	ي	/j/
Vowels		Diphthongs	
...ً	/a/		/aj/
	/a:/		/aw/
...ٍ	/i/		
	/i:/		
...ٌ	/u/		
	/u:/		

Labeling

- Concatenating the units from the database of labeled speech needs effective way, some used
  - Comprehensive searching
  - Classification and clustering dependent on syllable cc,vv,..etc.



Labeling



0000	000	001	010	011	100	101	110
00010	بَ	بُ	بِ	بْ	با	بو	بي
00100							

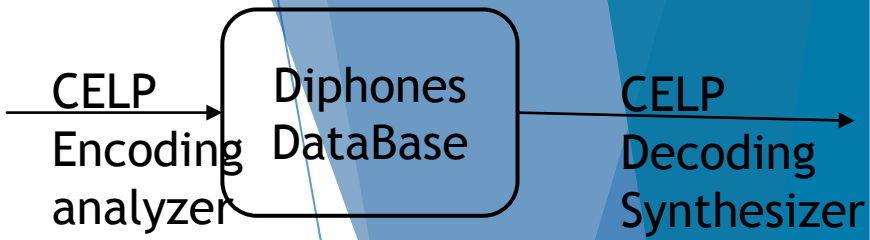
We suggest a data structure(like table) contains all Diphones and arranged as shown:
The first 5 bits (from left) point to the phonemic symbol (as in SAMBA table)
The rest 3 bits point to the syllable, prosody..

Using 8 bits we have indexed 256 entry.

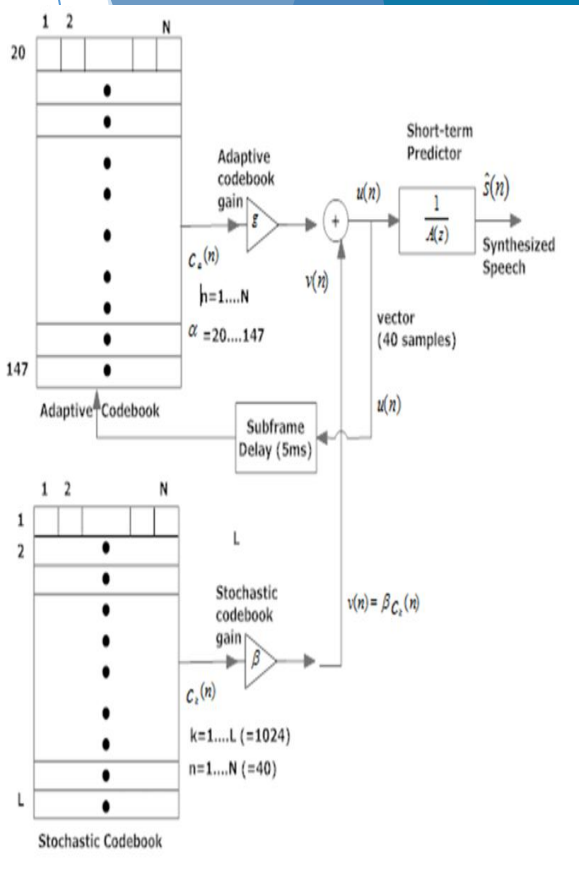
So as بِin بِسْمwe already know that its in 00011010

If this tested and proved, it will be the first, otherwise statistical model(machine learning) is an alternative

CELP



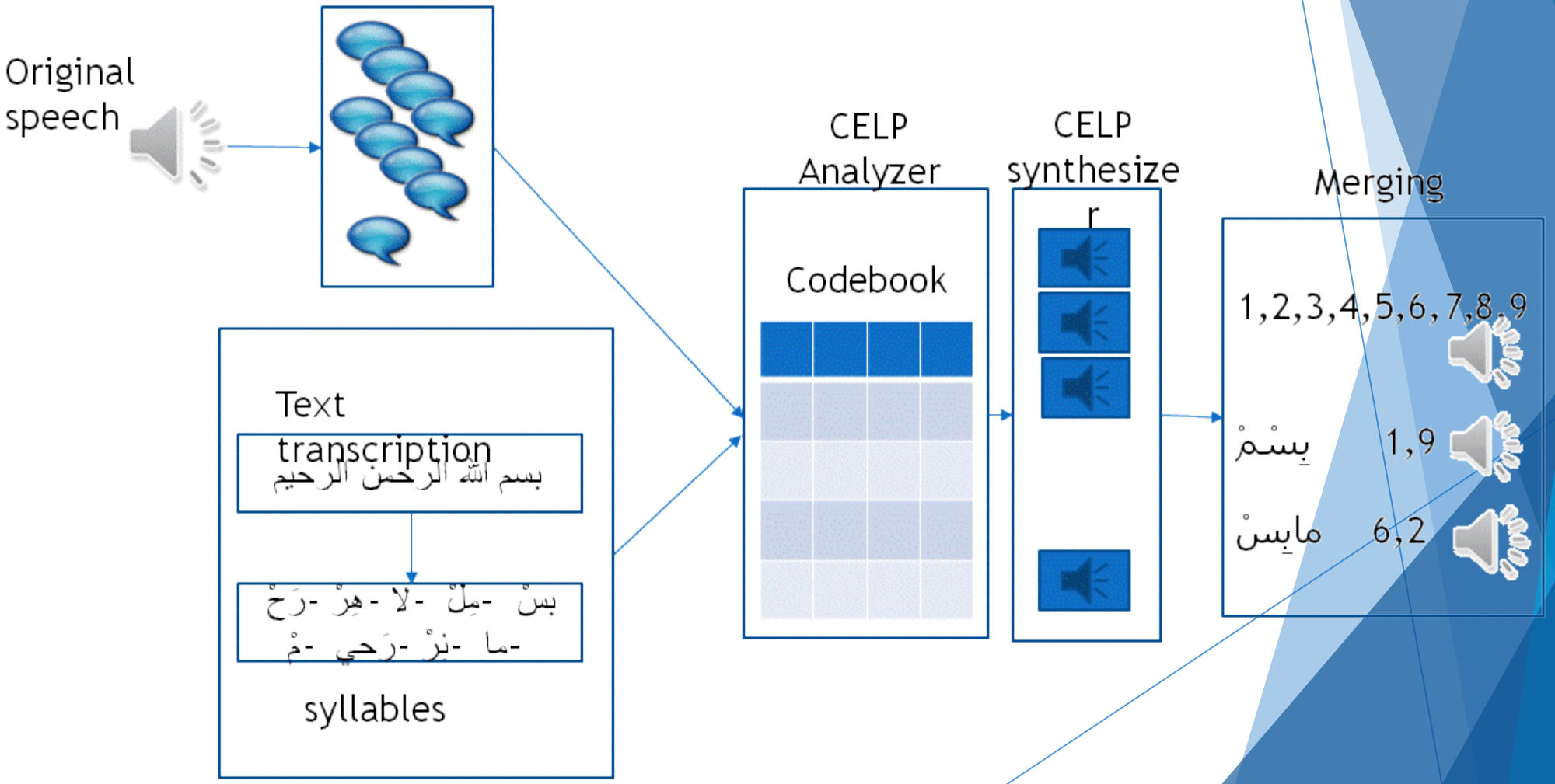
- Code Excited Linear Prediction (CELP) CELP Coding is an analysis by synthesis technique. The idea behind analysis-by-synthesis at the encoder is to analyze a short-time frame (or more) of speech, and extract parameters from this. These parameters are then used to create a frame of reconstructed speech.



- Sample rate at 8 kHz, the frame size is 20 ms = 160 samples), and
- the block duration for the excitation sequence selection is 5 ms (40 samples).
- 40* 1024 matrix: creating the Gaussian codebook
- 10 bits index, 8 bits pith filter, 12 bits LPC parameter (inverse sine), 3 bits the gain, 7 bits for pitch filter coeff

- MATLAB

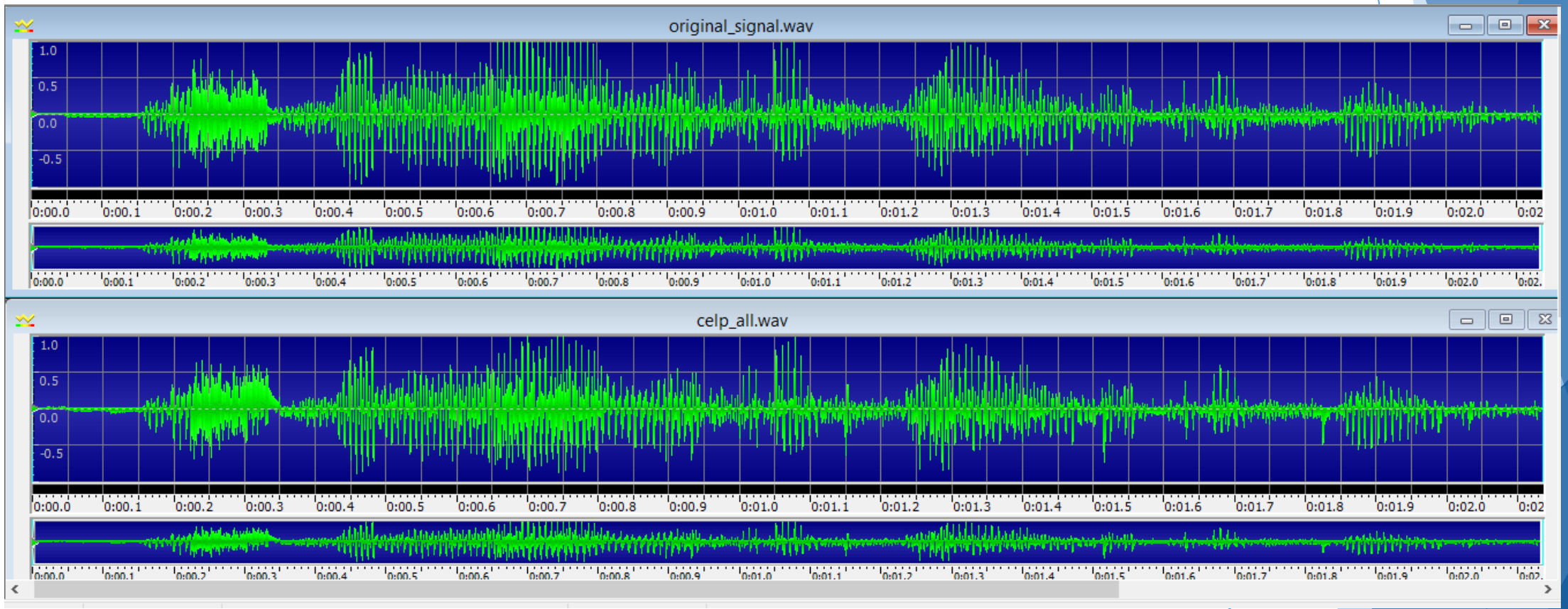
CELP Experiment Procedure



CELP Experiment Results

Speech sample	Signal length	Frame #	Max original	Max Synthesized CELP	t-shift (original-synthesized)
بسم الله الرحمن الرحيم	23552	147	0.99	1.07	0
بس	2159	13	0.73	0.73	0
مل	2424	15	0.82	0.87	0
لا	2951	18	0.99	1.05	0
هر	2141	13	0.69	0.66	0
رح	1979	12	0.99	1..02	0
ما	2645	16	0.99	0.91	0
نر	1799	11	0.38	0.36	0
رحي	3959	24	0.57	0.68	0
م	1655	10	0.15	0.14	0

CELP, original signal and synthesized signal

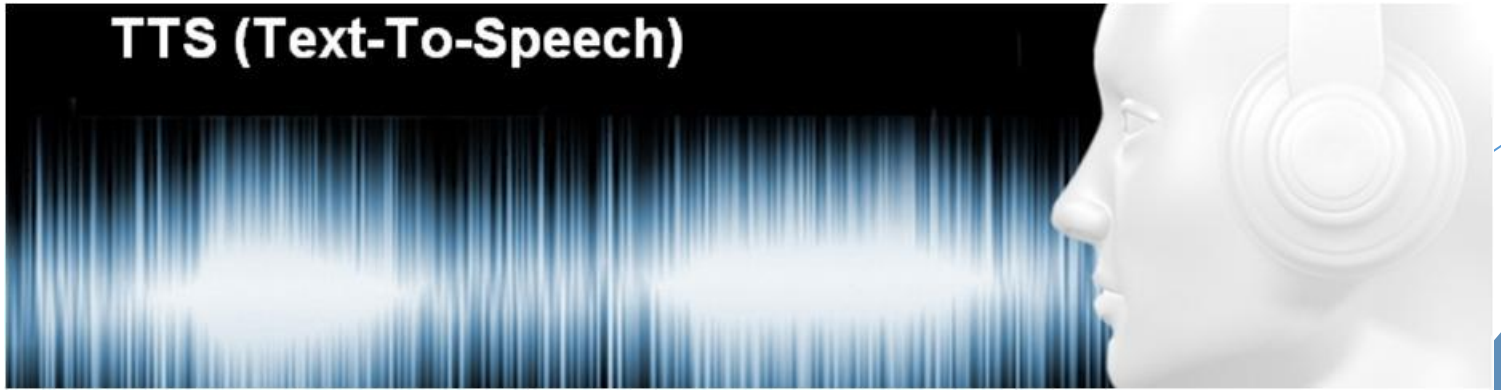


Evaluation

Evaluation

- Is a novel contribution in our thesis at the level of ATTS in concept and in the TTS synthesis in the criteria

- A high quality text to speech system should produce synthesized speech whose spectrograms should nearly match with the natural speech.[22]



Evaluation (signal processing parameters)

- ***cross-correlate*** signals to determine if there is a match, deal with data difference not time difference

$$(f \star g)(t) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} f^*(\tau)\ g(t+\tau)\,d\tau,$$

- ***Time delay*** and ***maximum/minimum*** **amplitude**

- ***Covariance*** of signal : A measure of how much the deviations of two or more variables or processes match. (cov α similarity)

$$\sigma(x, y) = \mathrm{E}\left[(x - \mathrm{E}[x])(y - \mathrm{E}[y])\right],$$

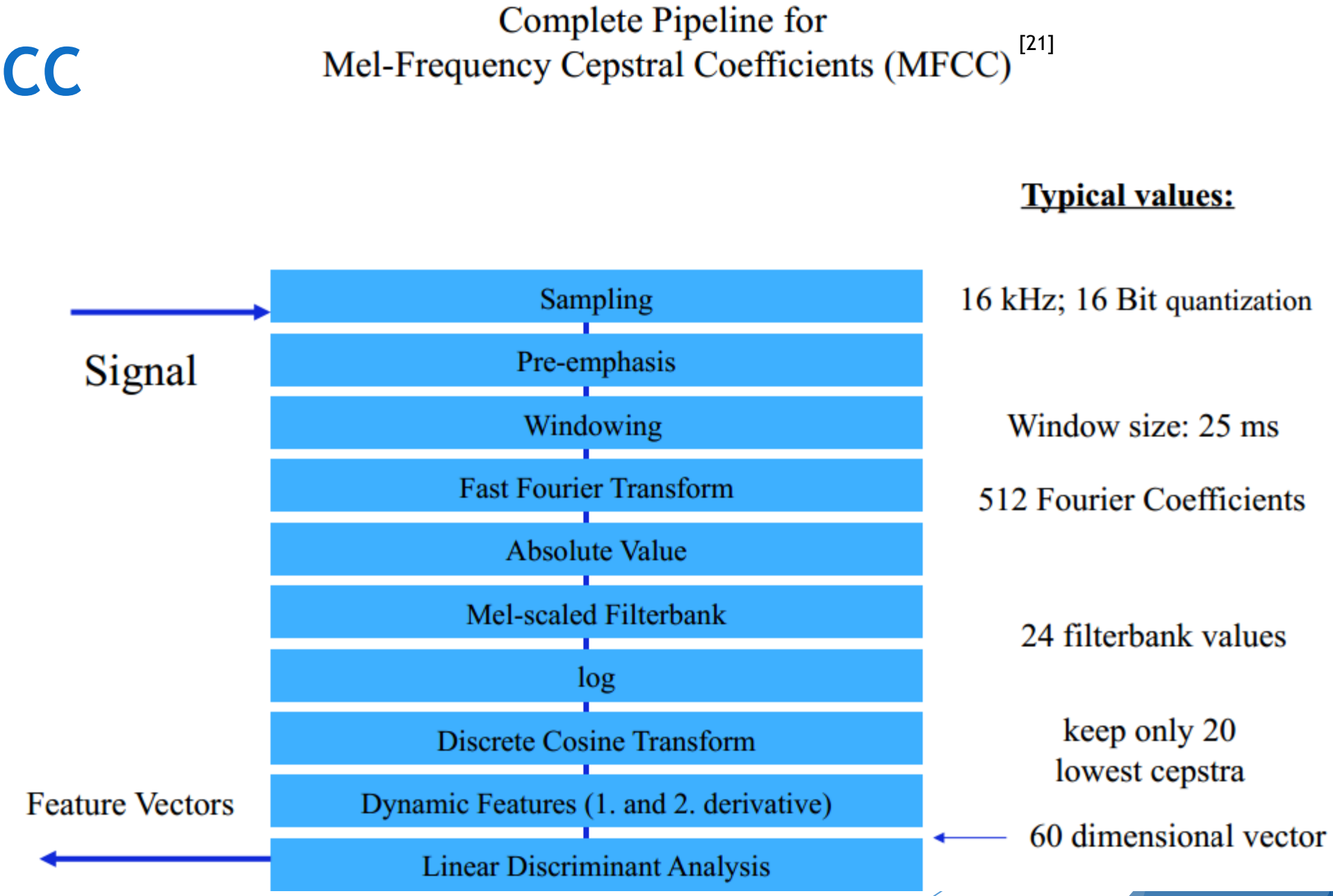
- ***Spectral coherence*** identifies frequency-domain correlation between signals. $C_{xy} = \frac{|G_{xy}|^2}{G_{xx} G_{yy}}$
- ***mean square error*** and ***spectral distances***. The spectral distances are defined by

$$d_p = \left[\sum_{k=1}^{K} |x_k - y_k|^p\right]^{\frac{1}{p}} = \|x - y\|_p$$

$$d = \frac{1}{16}\left(\sum_{k=1}^{16} |x_k - y_k|^2\right)$$

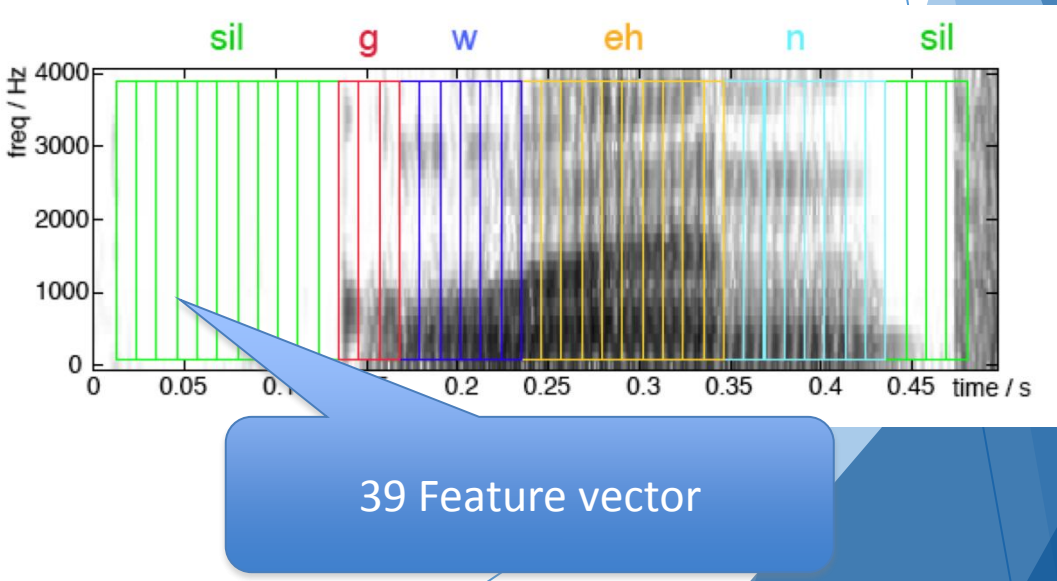
- **diff** = Simulink.sdi.compareSignals(signalID1,signalID2) to find the data match and the tolerance

MFCC


Complete Pipeline for
Mel-Frequency Cepstral Coefficients (MFCC) [21]
Typical values:
Signal
Sampling
16 kHz; 16 Bit quantization
Pre-emphasis
Windowing
Window size: 25 ms
Fast Fourier Transform
512 Fourier Coefficients
Absolute Value
Mel-scaled Filterbank
24 filterbank values
log
Discrete Cosine Transform
keep only 20
lowest cepstra
Feature Vectors
Dynamic Features (1. and 2. derivative)
60 dimensional vector
Linear Discriminant Analysis

MFCC : Feature Extraction

- Extract a feature vector from each frame
  - 12 MFCCs (Mel frequency cepstral coefficient) + 1 normalized energy = 13 features
  - Delta MFCC = 13
  - Delta-Delta MCC = 13
- Total: 39 features



Comparison

For Evaluation we have

- MFCC for the original signal & the MFCC for the synthesized one
- With mathematical differences between MFCC's will reflect the naturalness taking into account the signal processing parameters: cross-correlate, Covariance of signal, Max Amplitude, mean square error and spectral distances, and maximum amplitude for each.

The percentage for each parameter is not determined yet.

To Do procedure

- Implement a mini-database for diphones, selecting the verses is important.
- Fix the index procedure and test it, by decode some from the Gaussian codebook in some order.
- Determine the percentage for evaluation parameters.
- Evaluate our system and others.

Bibliography

1. Al-Ghamdi M et all, "**Phonetic Rules in Arabic Approach**", accepted in 2002, مجلة جامعة الملك سعود، م 16 ص ص 1-25، الرياض(1424ه\2004م)
2. Al-Saud N. and Al-Khalifa H., "**An Initial Comparative Study of Arabic Speech Synthesis Engines in iOS and Android**", iiWAS2012, 3-5 December, 2012, Bali, ACM 411
3. Assaf M., **"A prototype of an Arabic Diphone Speech Synthesizer in Festival"**, Master Thesis, 2005
4. Black A. and Taylor P., "Automatically Clustering Similar Units For Unit Selection In Speech Synthesis", 1997, The Festival Speech Synthesis System: system documentation. Technical Report HCRC/TR-83.
5. Chabchoub A. et all, "Di-Diphone Arabic Speech Synthesis Concatenation", International Journal of Computers & Technology. Council for Innovative Research: www.ijctonline.com ISSN: 2277-3061International Journal of Computers & Technology: Volume 3. No. 2, OCT, 2012
6. Dey S. et all, "Architectural Optimizations for Text to Speech Synthesis in Embedded Systems", 1-4244-0630-7/07/$20.00 ©2007 IEEE.
7. El-Imam Y., "Phonetization of Arabic:rules and algorithms", Science Direct, accepted in 2003, Computer Speech and Language 18 (2004) 339–373
8. Elothmany A., "Arabic Text-To-Speech Including Prosody (ATTSIP) for Mobile Devices", AlQuds University , 2013
9. Elshafie A, "Automaticall Y Clustering Similar Units For Unit Selection In Speech Synthesis toward an Arabic Text to Speech", 1991, The Arabian Journal for Science and Engineering. Volume 16, Number 4B.
10. Fulcher j. et all, "A Neural Network, Speech-based Approach to Literacy", 2002
11. http://en.wikipedia.org/wiki/Text-To-Speech accessed on 1/3/2014
12. http://www.azlifa.com/pp-lecture-8/ accessed on 20/4/2013
13. Ibrahim N., "Automated Tajweed Checking Rules Engine For Quranic Verse Recitation", Phd thesis, 2010
14. Odeh N., "Diphone-Based Arabic Speech Synthesizer for Limited Resources Systems", AlQuds University, Fall 2012,2013

Bibliography

15. Sarma Ch. et all, "A Rule Based Algorithm for Automatic Syllabification of a Word of Bodo Language", 2012, IJCCN, ISSN 2319-2720
16. Sassi S. et all, "A Text to Speech System for Arabic Using Neural Networks", ISSN :1098-7576, ISBN:0-7803-5529-6, IEEE pages 3030 - 3033 vol.5, 1999
17. Sassi S. et all, "Neural Speech Synthesis System for Arabic Language using CELP Algorithm", ISBN:0-7695-1165-1 , IEEE pages (119 – 121) 2001
18. Springer D.,"An Introduction to Text to Speech Synthesis", Book, ISBN 1-4020-0369-2 , 2001
19. Tabbal, H., et all, '**Analysis and Implementation of a "Quranic" verses delimitation system in audio files using speech recognition techniques**', Proceeding of the IEEE Conference of 2nd Information and Communication Technologies, 2006. ICTTA '06.Volume 2, pp. 2979 – 2984
20. Zhang M et all, "Phoneme Cluster Based State Mapping Fortext-Independent Voice Conversion", 978-1-4244-2354-5/09/$25.00 ©2009 IEEE, ICASSP 2009
21. http://www.lsv.uni-saarland.de/Vorlesung/Digital_Signal_Processing/Summer13/DSP_13_Chap6.pdf
22. Speech Technology - Kishore Prahallad (skishore@cs.cmu.edu)
23. Anees I. et all, "Al-Aswat Al-Arabia", 1990
24. Elshafei M. etall, "Tichniques for high quality Arabic Speech Synthesis", Elsevier Science 2002
25. Elshafie A. et all," Toward an Arabic Text-To-Speech system." The Arabic Journal Science and Engine, 1991.
26. Abu Alyazeed M. et all, "Comparison of Syllables and Sub-Syllable Methods for Speech Synthesis", 1989
27. Abu Ghattas N. and Abdel Nour H., "Text-to-Speech Synthesis by Diphones for Modern Standard Arabic", An-Najah Univ. J. Res. (N. Sc.), Vol. 19, 2005
28. Elshafei M. etall, "Tichniques for high quality Arabic Speech Synthesis", Elsevier Science 2002

Thanks for listening