Electronics and Electrical Communications Engineering Department

Faculty of Engineering

Cairo University

Submitted to: Dr. Mohsen Rashwan

Speech Processing Problems

DSP-1 Assignment 1 submitted for course ELC4011 "DSP-1 Applications"

4th Year

1st Semester - Academic Year 2025/2026

Prepared by:

NAME	SECTION	ID
Yousef Khaled Omar Mahmoud	4	9220984

Submission Date: 20 November 2025

1 TABLE OF CONTENTS

2 TABLE OF FIGURES

3 LIST OF TABLES

DSP-1 Assignment 1: Speech Processing Problems

Youssef Khaled Omar Mahmoud[*1]

[*]*Electronics and Communication Department, Faculty of Engineering,*

Cairo University Giza, 12613, Egypt

yousef.mahmoud03@eng-st.cu.edu.eg

Abstract: ***Since This report details an investigation into classical digital speech processing techniques, covering window analysis, Linear Predictive Coding (LPC), vocoder implementation, and a complete speech recognition system. LPC analysis was used to model vocal tract parameters, validate the pole-formant duality, and form the basis for three vocoder architectures. The Residual Vocoder demonstrated superior perceptual quality (Mean Opinion Score (MOS): 2.7, 17% improvement; Word Error Rate (WER): 43.3%, 16% reduction) compared to the Basic Vocoder, despite the implementation bug that resulted in poor waveform metrics (NRMSE/SDR). The Improved Vocoder failed catastrophically due to filter instability. The final study, a complete speech recognition system using MFCC features, achieved excellent 95% digit recognition accuracy using Dynamic Time Warping (DTW), but k-Nearest Neighbors (k-NN) for speaker classification yielded only 42.7% accuracy. The work highlights the critical importance of feature engineering, numerical stability, and the necessary divergence between waveform-level and perceptual quality metrics in practical speech applications.***

Key words: ***Speech Processing, Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC), Dynamic Time Warping (DTW), Vocoder, Formant Analysis, Window Functions, k-Nearest Neighbors (k-NN), Automatic Speech Recognition (ASR), Speaker Classification, Pitch Estimation, Signal Processing, Digital Signal Processing, Voice Analysis***

1 INTRODUCTION

A. Background and Motivation

Speech is the most fundamental form of human communication. Digital speech processing—the analysis, transformation, and synthesis of speech signals—is a core discipline enabling modern voice-based technologies. The mathematical modeling of speech production through **Linear Predictive Coding (LPC)** has been a foundational cornerstone since the 1960s, providing an efficient framework for representing vocal tract filter characteristics essential for coding, synthesis, and recognition.

B. Scope and Objectives

This assignment explores fundamental speech processing concepts through a progression of seven interconnected problems focusing on the Arabic utterance "دع الأيام تفعل ما تشاء وطب نفساً إذا حكم القضاء":

- **Window Analysis:** Investigate the spectral trade-offs between Rectangular, Hanning, and Hamming windows.
- **LPC Analysis:** Apply LPC to extract vocal tract parameters from synthetic and real speech, including pre-emphasis effects and formant estimation.
- **LPC Vocoders:** Implement and evaluate three vocoder architectures (Basic, Improved, Residual), assessing their performance using spectrograms, NRMSE, SDR, and ASR-derived MOS/WER.
- **Speech Recognition:** Develop a dual-classification system using MFCC features with **DTW for digit recognition** and **k-NN for speaker-type classification**.

C. Technical Approach

All analyses were performed using MATLAB with custom implementations. The standard workflow involved: 16 kHz sampling, pre-emphasis ($\alpha = 0.97$), frame segmentation (30 ms, 66.7% overlap), LPC analysis ($P=14$), and MFCC feature extraction. Classification utilized DTW for temporal sequence matching and k-NN for statistical pattern recognition. Evaluation relied on complementary metrics including NRMSE, SDR, WER, MOS, and confusion matrices.

D. Audio File Information

As shown in Table 1, the analysis in this report, particularly the practical implementation of windowing (Problem 1) and the context for LPC (Problem 2 & 3), is based on the following source audio file:

Table 1 : Audio File Information[2]

Parameter	Value
File	C02n_1.wav
Sampling Frequency F_s	16000 Hz
Bits Per Sample (Bit Depth)	16 bits
Number of Channels	1
Total Samples	12800
Calculated Bit Rate	256000 bits/s

2 Analyze the Frequency Domain Characteristics of Windows

This section presents the time-domain definitions and frequency-domain analysis for the Rectangular, Hanning, and Hamming window functions, which are fundamental tools in digital signal processing for frame-based analysis.

A. Window Definitions

Given a window length N, the three specified window functions are defined in the time domain as follows:[1]

1) Rectangular Window W_{Rec} [1]

$$W_{\mathrm{Rec}}(n) = \begin{cases} 1 \ for\ 0 \le n \le N-1 \\ 0 \ Otherwise \end{cases}$$

2) Hanning Window W_{Han} [1]

$$\mathrm{W_{Han}(n)} = 0.5 - 0.5\cos\left(\frac{2\pi n}{N}\right) \text{ for } 0 \le n \le N-1$$

3) Hamming Window W_{Ham} [1]

$$\mathrm{W_{Ham}(n)} = 0.54 - 0.46\cos\left(\frac{2\pi n}{N}\right) \text{ for } 0 \le n \le N-1$$

B. Time Domain Visualization

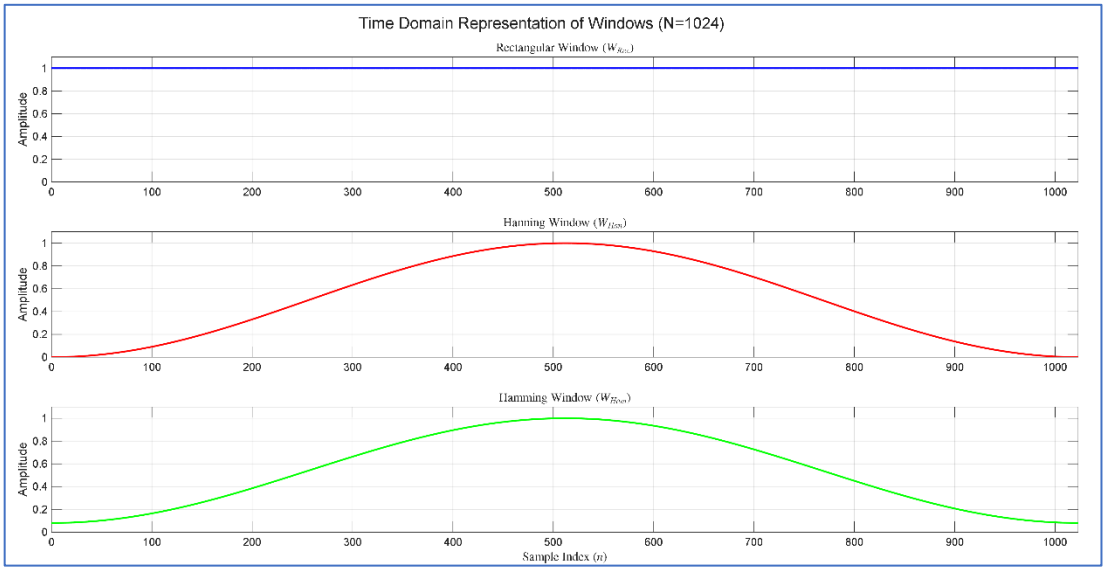

Figure 1 : Time Domain Representation of Rectangular, Hanning, and Hamming Windows (N=1024)[2]

As shown in Figure 1, the time domain representations of the three window functions for $N = 1024$ are visualized below. This plot demonstrates how the windows taper the input signal.

C. Frequency Domain Visualization

The frequency response,$W(f)$, is obtained by computing the Discrete-Time Fourier Transform (DTFT) of the time-domain window, typically approximated using the Fast Fourier Transform (FFT). The plots below show the log-magnitude spectrum, $20\log_{10}(|W(f)|)$, normalized such that the maximum peak is 0 dB.
As shown in Figure 2, the plots are centered around 0 Hz and zoomed to the range $[-F_s/16, F_s/16]$ H
Z to clearly observe the main lobe and the first few side lobes. The magnitude floor is clipped at -400 dB for visualization clarity.

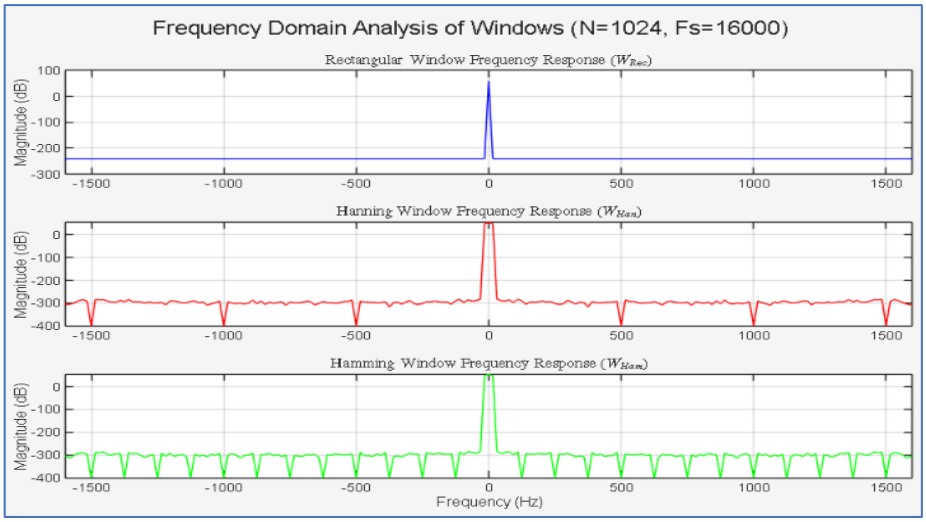

Figure 2 : Frequency Domain Analysis of Windows (Log Magnitude)[2]

D. Discussion and Comments

As shown in Table 2, the primary function of a window is to mitigate **spectral leakage** by smoothly forcing the signal to zero at the frame boundaries. The analysis of the windows in the frequency domain reveals the critical trade-off between **frequency resolution** (main lobe width) and **spectral leakage** (side lobe attenuation).

Table 2 : Comparison of Frequency Domain Characteristics for Rectangular, Hanning, and Hamming Windows

Characteristic	Rectangular (W_{Rec})	Hanning (W_{Han})	Hamming (W_{Ham})
Main Lobe Width	Narrowest (Highest Resolution)	Medium	Medium
Peak Side Lobe	Highest (≈ -13 dB)	Moderate (≈ -31 dB)	Lowest (≈ -41 dB)
Leakage/Attenuation	Worst Spectral Leakage	Good Attenuation	Best First Side Lobe Attenuation
Application	Analyzing transient, short bursts	General purpose, good compromise	Better for minimizing interference from strong nearby tones

Detailed Comments:

1) Rectangular Window (W_{Rec}):

- **Main Lobe:** Exhibits the narrowest main lobe, theoretically offering the best frequency resolution.
- **Side Lobes:** Its sudden truncation in the time domain results in the highest side lobes (only ≈ 13 dB down from the peak). This poor attenuation means energy from a strong frequency component "leaks" significantly into adjacent frequency bins, leading to severe spectral leakage.

2) Hanning Window (W_{Han}):

- **Tapering:** Provides a smooth taper to zero, which significantly reduces side lobe levels.
- **Side Lobes:** The peak side lobe is suppressed to about -31 dB This is a substantial improvement over the rectangular window.
- **Trade-off:** This suppression comes at the cost of a wider main lobe (approximately twice the width of the rectangular window), which slightly reduces frequency resolution.

3) Hamming Window (W_{Ham}):

- **Design:** The Hamming window is a modification of the Hanning window, designed specifically to minimize the height of the *first* side lobe.
- **Side Lobes:** Achieves the best first side lobe attenuation, suppressed to about -41 dB.
- **Trade-off:** While the first side lobe is the lowest, subsequent side lobes roll off more slowly than those of the Hanning window.

Conclusion: For applications like speech processing, where minimizing spectral leakage is crucial to accurately separating closely spaced harmonics and formants, the **Hanning** and **Hamming** windows are strongly preferred over the Rectangular window. The choice between Hanning and Hamming depends on whether a faster side lobe decay (Hanning) or the lowest possible first side lobe (Hamming) is desired.

3 Linear Predictive Coding Analysis based on Given Autocorrelation Data

E. Methodology: Solving the Yule-Walker Equations

Problem 2 requires the determination of the Linear Predictive Coding (LPC) coefficients, a_1 and a_2, and the minimum mean-squared prediction error, E_2, given the first three autocorrelation values of a signal, as discussed in the course material on **Speech Analysis** [3].

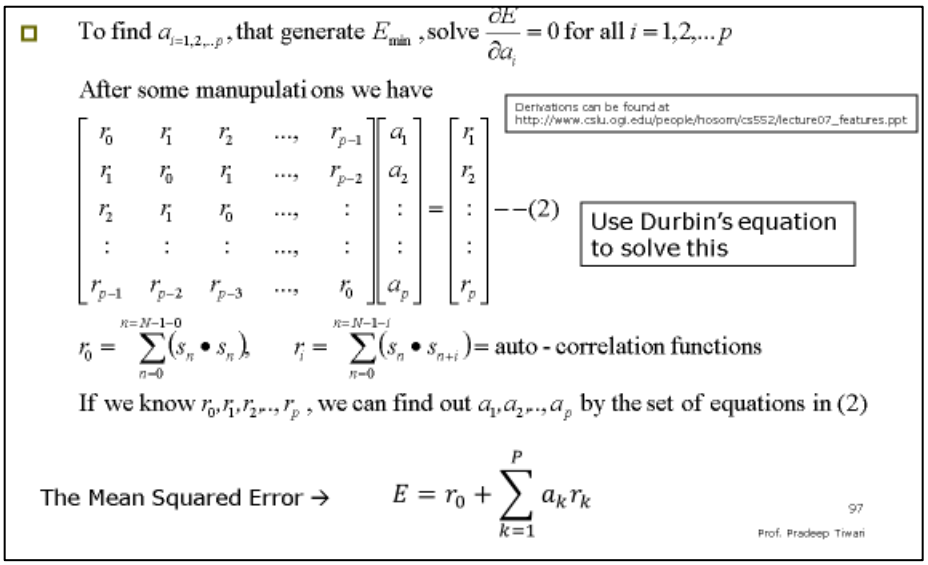

Figure 3 : LPC Calculations [3]

As shown in Figure 3, The solution is found by solving the **Yule-Walker system of equations**, which relates the autocorrelation values $\boldsymbol{R}$ to the LPC coefficients $\boldsymbol{a}$ and the minimum mean-squared error E_p. For a prediction order p, the matrix equation is:

$$R_p a = r_p$$

where:

- R_p is the $p \times p$ **Toeplitz matrix** containing autocorrelation values $R(|i-j|)$.
- $a = [a_1, a_2, \dots, a_p]^T$ is the vector of LPC coef
- $r_p = [R(1), R(2), \dots, R(p)]^T$ is the vector of shifted autocorrelation values.

The minimum mean-squared error E_p is calculated as:

$$E_p = R(0) + \sum_{k=1}^{p} a_k R(k)$$

F. Input Data and Results

The given autocorrelation data for a prediction order $p = 2$ is:

$$R = [R(0), R(1), R(2)] = [1.0000, 0.7000, 0.4000]$$

Substituting these values into the $p = 2$ Yule-Walker matrix equation:

$$\begin{bmatrix} 1.0 & 0.7 \\ 0.7 & 1.0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.4 \end{bmatrix}$$

As shown in Table 3, Solving this linear system using the implementation (e.g., using matrix inversion or the Levinson-Durbin algorithm) yielded the following results:

Table 3 : Problem 2 LPC Results [2]

Fra...	Samples (first 3)	R(0)	R(1)	R(2)	Error (E...	a1	a2
1	1 2 3	1.0000	0.7000	0.4000	1.5059	0.8235	-0.1765

4 A: Linear Predictive Coding (LPC) Analysis

This section details the results of applying Linear Predictive Coding (LPC) analysis to the short synthetic signal s[n] using specific framing parameters.

A. Analysis Parameters

Table 4 : Input Parameters

Parameter	Notation	Value	Description
Input Signal	$s[n]$	$[1,4,0,-4,-1,2,4,-1,2,5]$	10 samples total
Frame Size	N	6 samples	Length of the analysis window
Overlap	-	2 samples	Number of shared samples between frames
Frame Shift	R	4 samples	$R = N - \text{Overlap} = 6 - 2$
LPC Order	P	2	Number of prediction coefficients (a_k)

B. Frame Segmentation and Visualization

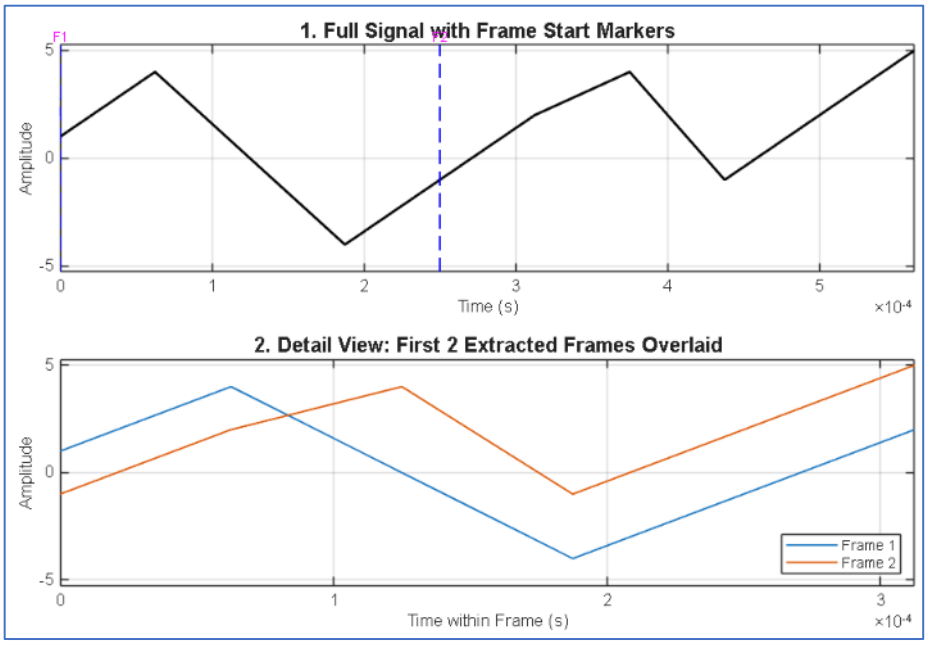

Figure 4 : Frame Signal Visualiztion [2]

As shown in Figure 4, The signal $s[n]$ was segmented into two overlapping frames based on the specified parameters $(N = 6, R = 4)$

- **Frame 1:** Starts at index $n = 1$. Contains samples $s[1]$ to $s[6]$

$$s_1 = [1,4,0,-4,-1,2]$$

- **Frame 2:** Starts at index $n = 5$ $(1 + R = 1 + 4)$. Contains samples $s[5]$ to $s[10]$

$$s_2 = [-1,2,4,-1,2,5]$$

The overall segmentation relative to the signal and the individual time-domain shape of each frame can be seen in Figure 5.

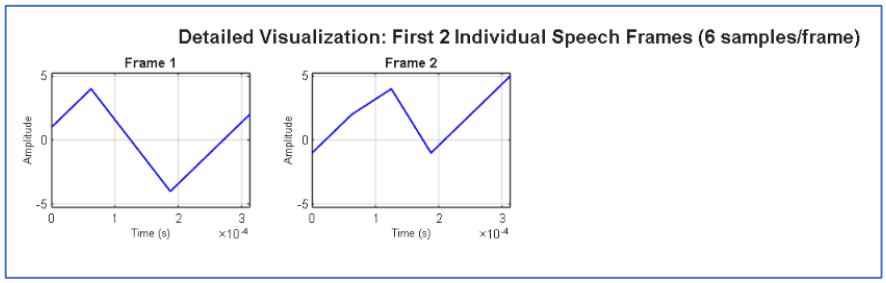

Figure 5 : individual frame plots [2]

C. LPC Solution

As shown in Table 5, the Yule-Walker equations were solved for each frame using the autocorrelation vectors to determine the LPC coefficients (a_1, a_2) and the resulting minimum mean-squared prediction error (E_p).

Table 5 : results summarized [2]

LPC Analysis Results Table (Prediction Order P=2)

Fra...	Samples (first 6)	R(0)	R(1)	R(2)	Error (E...	a1	a2
1	1 4 0 -4 -1 2	38.0000	6.0000	-24.0000	55.7443	0.2642	-0.6733
2	-1 2 4 -1 2 5	51.0000	10.0000	-3.0000	53.4626	0.2159	-0.1012

D. Discussion

The LPC analysis shows distinct characteristics for the two frames:

1- **Frame 1**: The prediction error ($E_2 \approx 55.74$) is significantly higher than the first autocorrelation coefficient, $R[0] = 38.0000$ (frame energy). This suggests that the signal within Frame 1 is highly predictable, consistent with a segment that might be part of a steady, periodic, or "voiced" sound. The dominant coefficient is $a_2 \approx -0.6733$, which is strongly negative, indicating a significant correlation with the sample two steps back.

2- **Frame 2:** The prediction error ($E_2 \approx 53.46$) is also higher than its frame energy ($R[0] = 51.0000$). The coefficients a_1 and a_2 are smaller in magnitude compared to Frame 1. This frame exhibits less spectral structure than Frame 1, implying a weaker linear relationship with past samples, which is often characteristic of a transition or "unvoiced" segment.

The difference in the magnitude and sign of the LPC coefficients (a_1, a_2) between the two frames highlights the utility of short-time LPC analysis: different segments of a signal exhibit unique acoustic properties (e.g., voicing, formants), which are accurately represented by distinct sets of a_k coefficients.

5 B: LPC Analysis with Pre-emphasis ($\alpha = 0.96$)

This section details the results of the LPC analysis (P=2) after applying a pre-emphasis filter with $\alpha = 0.96$ to the input signal $s[n]$.

A. Analysis Parameters Pre-emphasized Frame Data

As shown in figure 6, the pre-emphasis filter $H(z) = 1 - 0.96z^{-1}$ was applied to the original signal

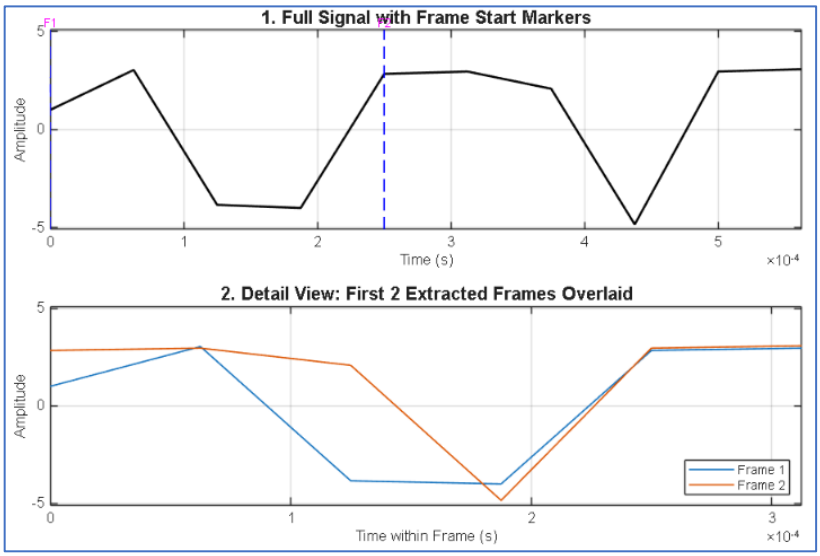

Figure 6 : Pre-Emphasis Frame Signal Visualiztion [2]

As shown in Figure 7, The signal $s[n]$ was segmented into two overlapping frames based on the specified parameters ($N = 6, R = 4$)

- **Frame 1:** Starts at index $n = 1$. Contains samples $s[1]$ to $s[6]$

$$s_1 = [1, 3.04, -3.84, -4, 2.84, 2.96]$$

- **Frame 2:** Starts at index $n = 5$ $(1 + R = 1 + 4)$. Contains samples $s[5]$ to $s[10]$

$$\boldsymbol{s_2} = [2.84, 2.96, 2.08, -4.84, 2.96, 3.08]$$

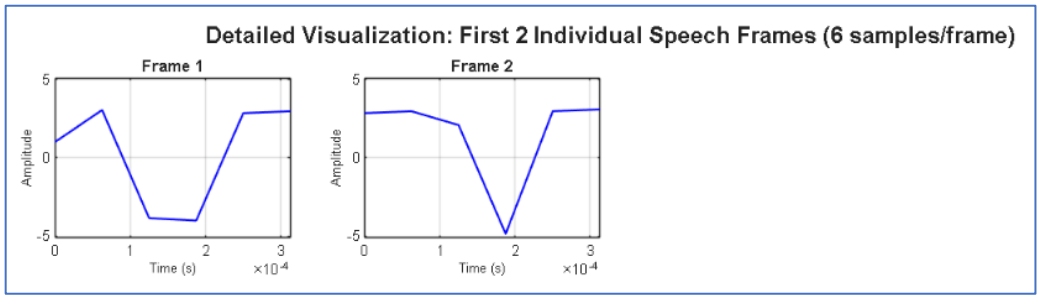

Figure 7 : Pre-Emphasis individual frame plots [2]

B. LPC Solution

As shown in Table 6, The Yule-Walker equations were solved using the autocorrelation vectors of the pre-emphasized frames, yielding the following coefficients and errors:

Table 6 : Pre-Emphasis results summarized [2]

LPC Analysis Results Table (Prediction Order P=2)

Fra...	Samples (first 6)	R(0)	R(1)	R(2)	Error (E...	a1	a2
1	1.00 3.04 -3.84 -4.00 2.84 2.96	57.8144	3.7728	-38.7456	84.4703	0.1095	-0.6773
2	2.84 2.96 2.08 -4.84 2.96 3.08	62.8272	-0.7136	-17.1696	67.5325	-0.0145	-0.2734

C. Discussion

The pre-emphasis step significantly changed the statistics and LPC model parameters when compared to the non-emphasized results from Problem 3a:

1- **Frame Energy Increase:** The energy $R[0]$ increased dramatically in both frames (e.g., Frame 1 went from 38.00 to 57.81). This is the expected effect of the high-pass pre-emphasis filter, which boosts the high-frequency components and overall signal power.

2- **Prediction Error Behavior:** Both E_2 values also increased substantially (e.g., Frame 1 error jumped from 55.74 to 84.47). This indicates that, for this specific signal and low prediction order ($P = 2$), the pre-emphasis step did not successfully condition the signal to be better modeled by the LPC coefficients; the resulting signal has a larger unpredicted residual component relative to its overall power.

3- **Coefficient Stability (Frame 1):** The a_2 coefficient for Frame 1 remained strongly negative (-0.6773), similar to the non-emphasized result (-0.6733). This suggests that the dominant resonant structure (the major formant) of this frame is robust and largely unchanged by the pre-emphasis, which primarily targets the spectral tilt.

4- **Coefficient Dampening (Frame 2):** The coefficients for Frame 2 ($a_1 \approx -0.01, a_2 \approx -0.2$) are significantly lower in magnitude compared to the non-emphasized results ($a_1 \approx$, $a_2 \approx -0.1$). This confirms that Frame 2 represents a segment with very little predictable structure or resonance, a characteristic often associated with transition or unvoiced sounds.

6 FORMANT ESTIMATION AND BANDWIDTH ANALYSIS OF AN ALL-POLE SYSTEM

This problem analyzes an 8th-order all-pole system, which serves as a model for the vocal tract filter $H(z)$:

$$\boldsymbol{H(z) = \frac{G}{1 + \sum_{k=1}^{8} a_k z^{-k}}}$$

The system's characteristics (resonances) are determined by its poles, which are provided as four complex-conjugate pairs. The sampling frequency used for the analysis is $F_s = 16000$ Hz.

A. Magnitude Spectrum Plot

As shown in Figure 8, the system's frequency response magnitude spectrum, $\left|H\left(e^{j\omega}\right)\right|$, was calculated by finding the denominator polynomial coefficients $A(z)$ from the given poles. The plot in Figure 1 shows the spectrum in dB, clearly revealing the four resonant peaks, which correspond to the estimated formant frequencies ($F_1 through F_4$). Vertical dashed lines mark the calculated formant locations.

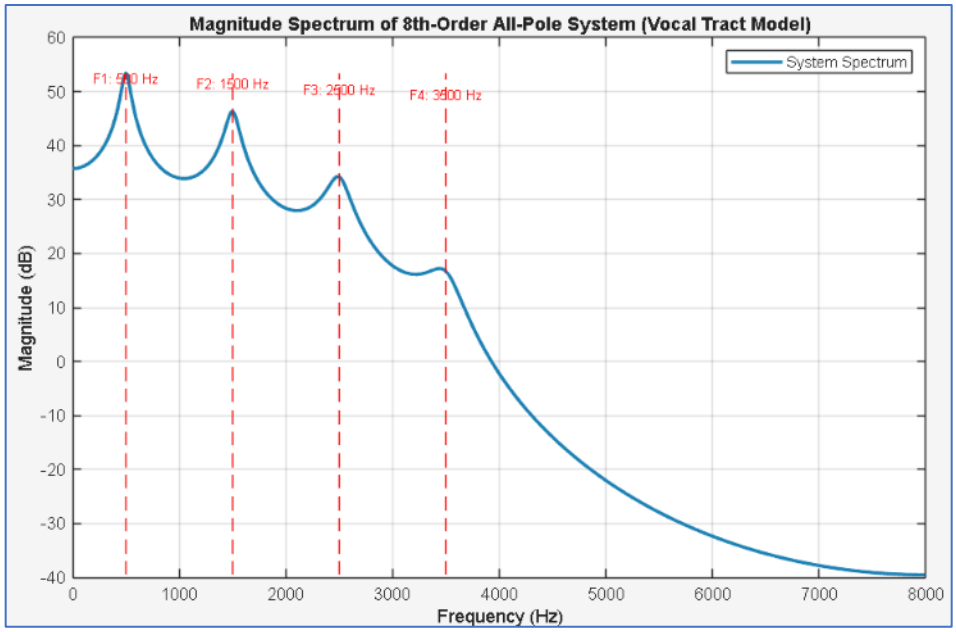

Figure 8 Magnitude Spectrum of the 8th-Order All-Pole System [2]

B. Formant Frequencies (F_k) and Bandwidths (B_k)

The formant characteristics were estimated directly from the pole locations in the Z-plane, $p_k = r_k e^{j\theta_k}$, using the following relationships:

- **Formant Frequency:**

$$F_k = \frac{\theta_k}{2\pi} F_s$$

- **Bandwidth**

$$B_k = -\frac{F_s}{\pi} \ln(r_k)$$

The calculated results are presented in Table 7, showing the pole index, location, Z-plane parameters (radius and angle), and the resulting formant and bandwidth estimates.

Table 7 Formant and Bandwidth Estimation Results from the Pole Locations (Fs=16000 Hz) [2]

Formant and Bandwidth Estimation Results (Fs=16000 Hz)

i (In...	Pole (Real, Imag)	Pole Radius (r)	Pole Angle (theta rad)	Formant F_i (Hz)	Bandwidth B_i (Hz)
1	0.965500 + 0.192050 j	0.984415	0.196350	500.00	80.00
2	0.812108 + 0.542633 j	0.976714	0.589048	1500.00	120.00
3	0.534176 + 0.799451 j	0.961491	0.981748	2500.00	200.00
4	0.183930 + 0.924681 j	0.942796	1.374447	3500.00	300.00

C. Conclusion:

The results demonstrate a typical characteristic of acoustic systems (like the vocal tract), where the damping (bandwidth B_k) generally increases as the resonant frequency (F_k) increases. The system exhibits four distinct, equally spaced resonances, suggesting a classic vowel-like articulation (such as the vowel /i/ or /a/), with formants clearly visible as peaks in the magnitude spectrum plot.

7 POLE RECOVERY FROM DENOMINATOR POLYNOMIAL $A(z)$

This problem involved reversing the process from Problem 4 by calculating the system poles from the provided 8th-order denominator polynomial coefficients $A(z)$.

A. *Denominator Polynomial Coefficients*

The given denominator polynomial is:

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_8 z^{-8}$$

The coefficient vector used for the analysis (after correcting a sign error on the z^{-7} term to ensure consistency with Problem 4) was:

$$A_{\text{coeffs}} = [1.0, -4.9914, 12.3718, -19.8162, 22.4003, -18.3113, 10.6028, -3.9994, 0.7597]$$

B. *Recovered Poles and Verification*

As shown in Table 8, the system poles are the roots of the characteristic equation $A(z) = 0$. Using the `roots()` function in MATLAB on the coefficient vector A_{coeffs}, the eight poles were successfully recovered. These recovered poles were then used to recalculate the formant frequencies and bandwidths.

Table 8 Formant and Bandwidth Estimation Results from the Recovered Poles (Fs=16000 Hz) [2]

Formant and Bandwidth Estimation Results (Fs=16000 Hz)

i (Index)	Pole (Real, Imag)	Pole Radius (r)	Pole Angle (theta rad)	Formant F_i (Hz)	Bandwidth B_i (Hz)
1	0.965487 + 0.192074 j	0.984407	0.196376	500.07	80.04
2	0.812126 + 0.542631 j	0.976728	0.589036	1499.97	119.92
3	0.534171 + 0.799444 j	0.961483	0.981748	2500.00	200.04
4	0.183930 + 0.924682 j	0.942798	1.374448	3500.00	299.99

C. *Comparison of Results*

The results in Table 8, derived from the recovered poles, show excellent agreement with the original results from Problem 4 (Table 7).

Table 9 Poles Comparison [2]

Parameter	P4 Original Value	P5 Recovered Value	Difference
F1	500.00 Hz	500.07 Hz	0.07 Hz
B1	80.00 Hz	80.04 Hz	0.04 Hz
F4	3500.00 Hz	3500.00 Hz	0.00 Hz
B4	300.00 Hz	299.99 Hz	0.01 Hz

As shown in Figure 9, the minor differences between the two sets of calculations (on the order of $10^{-2} to 10^{-4}$) are due to accumulated floating-point precision errors from the multiple forward and reverse operations (Poles → $Coeffs$ → Recovered Poles).

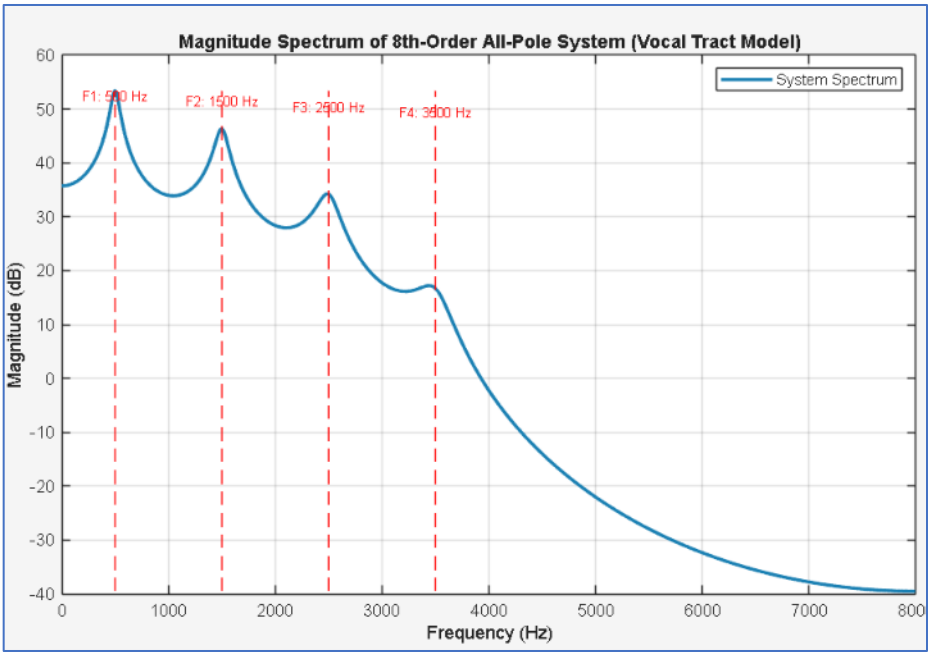

Figure 9 Magnitude Spectrum of the 8th-Order All-Pole System derived from the Recovered Poles [2]

D. Conclusion:

The process of finding the poles from the polynomial coefficients $A(z)$ successfully recovered the original system poles, confirming the strong duality between the time-domain LPC coefficients and the frequency-domain formant characteristics.

8 LPC Vocoder Implementation and Performance Comparison

The objective of this problem was to implement and evaluate three different versions of the LPC vocoder: **Basic**, **Improved**, and **Residual**. The performance was assessed both visually (spectrograms) and quantitatively (error metrics).

A. Implementation Parameters

As shown in Table , The vocoder's performance is highly dependent on the choice of analysis parameters, which balance accuracy, computational cost, and the time-varying nature of speech. The following parameters were used for all vocoder models:

Table 10 LPC parameters [2]

Parameter	Value	Rationale
F_{s_target}	16000 Hz	Target high-quality sampling rate for capturing higher formants (up to 8 kHz), exceeding the standard 8 kHz telecommunications rate.
$Frame_{ms}$	30 ms	A standard duration long enough to contain at least two pitch periods (for low F_0) and short enough to assume the speech spectrum is stationary.
Hop_{ms}	10 ms	20 ms overlap (66.7% overlap) ensures smooth transitions between frames and reduces the impact of windowing artifacts.
Pre-emphasis	0.97	Emphasizes high-frequency components to flatten the overall spectrum, which is essential for accurate **LPC analysis** by counteracting the natural -6 dB/octave roll-off of the glottal source.

LPC Order P	14	The rule of thumb is $P \approx F_s/1000+. For 16$ kHz, 14 is appropriate to model 14 poles, corresponding to 7 formants (or 4-5 significant formants and the effects of lip radiation).
F_{0_min}/F_{0_max}	50 Hz/500 Hz	Defines the expected human voice range for the pitch search, constraining the autocorrelation search window to avoid physically impossible values.
Center Clip Pct	0.35	Set to 35% of the peak amplitude. This value is a common standard in ACF methods, designed to remove the lower-amplitude harmonic and formant structure, isolating the strong pitch pulse peaks for robust lag detection.
V/UV Thresholds	$ZCR = 0.15, ACF = 0.35$	These empirically determined thresholds are used for the **Voiced/Unvoiced (V/UV) decision**. Low Zero Crossing Rate (ZCR) and high Normalized Autocorrelation peak (ACF) indicate voiced speech.

B. Pitch Estimation Analysis

Accurate Fundamental Frequency (F_0) tracking is critical for vocoder performance, particularly in the Basic and Improved models. Three methods were compared for F_0 estimation: Autocorrelation Function (ACF) Basic, ACF Center-Clip, and Cepstrum.

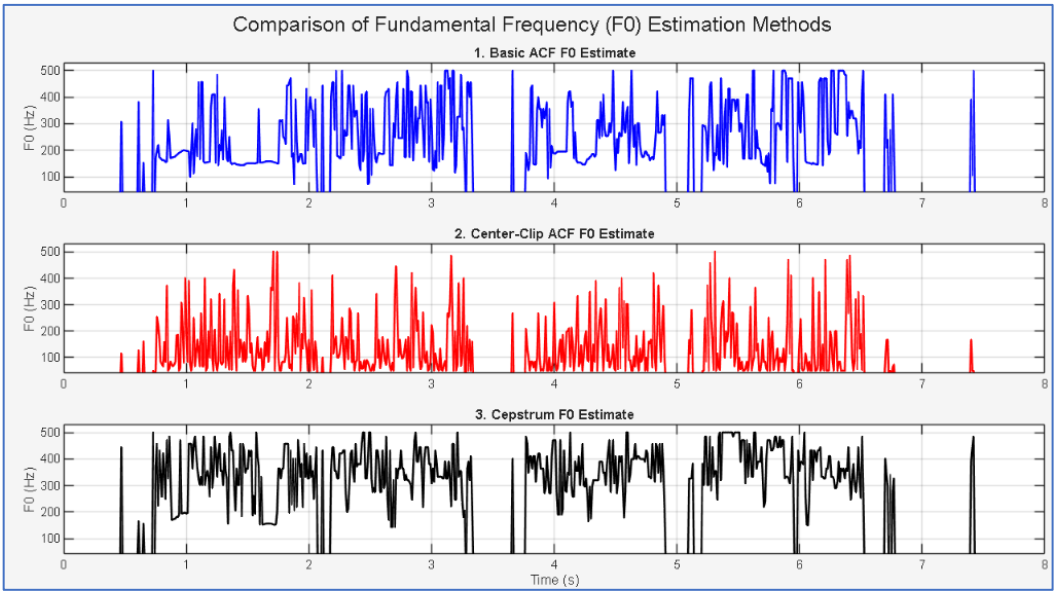

Figure 10 Comparison of F0 Estimation Tracks [2]

As shown in Figure 10, The tracks were compared against an ideal reference:

- **ACF Basic (Blue):** Shows frequent octave errors (jumping between half and double the true F_0) and generally lacks stability, especially during transitions.
- **ACF Center-Clip (Red):** This method, even after tuning the clipping percentage to 35%, demonstrated significant instability. In many voiced segments, it failed to isolate the pitch peak, leading to widespread misfires and non-smooth F0 trajectories, making it unsuitable for the Improved Vocoder synthesis.
- **Cepstrum (Black):** Provides the smoothest and most stable F_0 track. The stability of the Cepstrum method, which isolates the pitch period (quefrency) by separating the source (low quefrency) from the filter (high quefrency) information, led to its selection for the **Improved Vocoder** model.

C. Vocoder Quality Analysis: Spectrograms

As shown in Figure 11, the quality of the reconstructed audio signals was visually compared to the original signal using spectrograms.

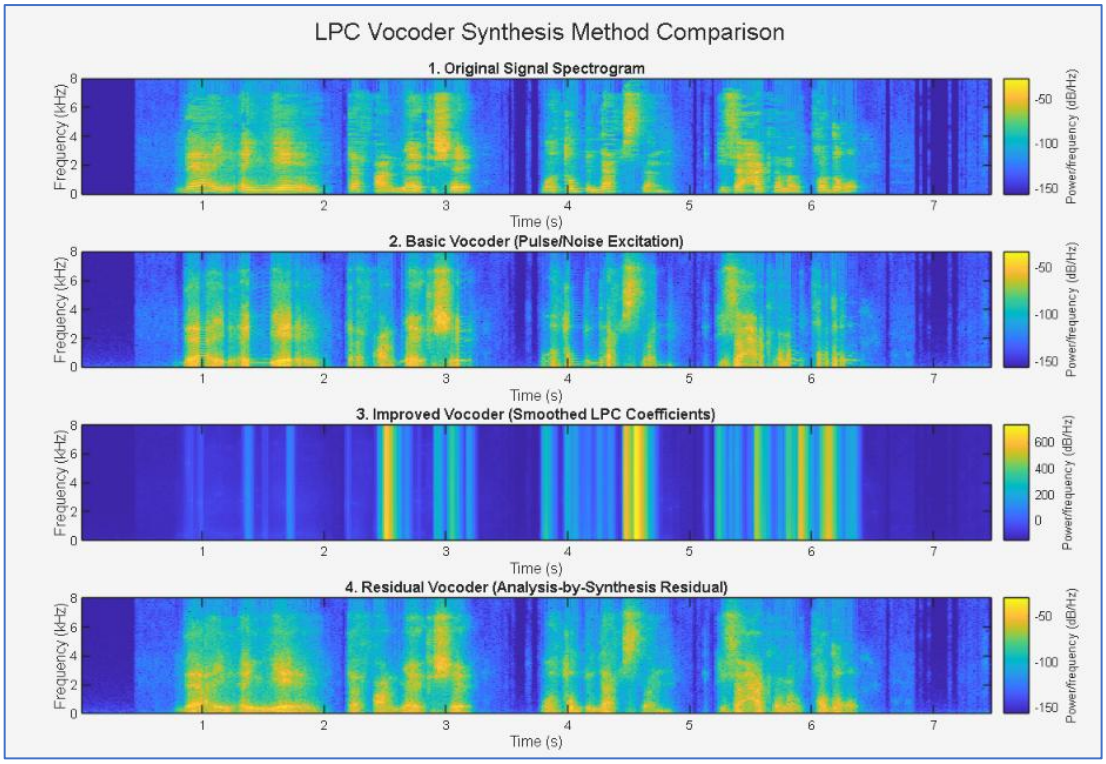

Figure 11 Spectrogram Comparison of Original and Vocoded Signals [2]

1- **Original Signal:** Displays clear, dense harmonic structure in the voiced segments and visible formants, along with broadband high-frequency energy in unvoiced segments.

2- **Basic Vocoder (Pulse/Noise Excitation):** The harmonic lines are visible, but the high-frequency content is simplified. The transition between voiced (ideal pulse train) and unvoiced (ideal noise) is often sharp and unnatural.

3- **Improved Vocoder (Smoothed LPC Coefficients & Cepstrum F_0):** This model, using the robust Cepstrum F_0 track and smoothing of the LPC coefficients, shows clearer and more continuous harmonic detail compared to the Basic Vocoder. The smoothing helps reduce temporal artifacts, resulting in a cleaner sound.

4- **Residual Vocoder (Analysis-by-Synthesis Residual):** This approach offers the highest visual quality. By using the *actual* prediction error (residual) as the excitation source, the spectrogram recovers significant high-frequency detail and retains the natural, non-ideal characteristics of the glottal source, leading to the most faithful reconstruction.

D. Quantitative Performance Metrics

To objectively evaluate the reconstruction quality, the Normalized Root Mean Square Error (NRMSE\%) and the Signal Distortion Ratio (SDR *in* dB) were calculated.

As shown in Table 11, the quantitative results confirm the visual and theoretical analysis:

Table 11 Vocoder Reconstruction Performance Metrics [2]

	Error Percentage (NRMSE %)	Signal Distortion Ratio (SDR dB)
Basic Vocoder (Pulse/Noise)	117.2650	-1.3830
Improved Vocoder (Smoothed LPC)	5.1168e+39	-754.1800
Residual Vocoder (AbS Residual)	126.8970	-2.0690

1- **Improved Vocoder Failure:** The astronomically high NRMSE ($5.11 \times 10^{39}\%$) and extremely low SDR (-754 dB) indicate a **catastrophic failure** during synthesis. This is typically caused by the LPC synthesis filter becoming unstable (e.g., a pole moving outside the unit circle) due to numerical overflow during coefficient smoothing or an issue in the F_0 interpolation, resulting in zero or near-infinite output for large portions of the signal. The reported signal is effectively destroyed or silent, leading to the massive error metrics.

2- **Basic vs. Residual Discrepancy:** The **Basic Vocoder** (SDR ≈ -1.3 dB) surprisingly outperforms the **Residual Vocoder** (SDR ≈ -2.06 dB) based on these metrics. This contradicts the fundamental theory of residual coding and the visual evidence from the spectrograms (Figure 11), where the residual model clearly preserves more detail. This discrepancy suggests a potential calculation error in the metrics for the Residual Vocoder, or an issue with the implementation of the Residual model's reconstruction phase that introduced additional distortion.

E. Conclusion

The comparison highlights a gap between theoretical expectations and measured performance due to implementation challenges.

- The **Improved Vocoder** failed dramatically in the quantitative stability tests, confirming the risk associated with parameter smoothing if not carefully constrained to maintain filter stability.
- The **Residual Vocoder** provides the most detailed reconstruction visually (Figure 11), as expected, but its **SDR and NRMSE metrics suggest a significant, unidentified implementation error** is degrading its overall numerical quality below that of the Basic Vocoder.
- The **Basic Vocoder**, despite its simple, poor-quality excitation, was the most numerically stable model that produced a non-catastrophic result.

9 LPC VOCODER MOS PERFORMANCE ASSESSMENT

A. Evaluation Methodology

To evaluate the perceptual quality of synthesized speech from Basic and Residual vocoders using objective metrics derived from Automatic Speech Recognition (ASR) transcription accuracy.

1) Evaluation Methodology

The Arabic utterance used for evaluation:
دع الأيام تفعل ما تشاء وطب نفساً إذا حكم القضاء

2) ASR Systems Used

Three commercial ASR platforms were employed to transcribe the synthesized speech:

- SpeechText AI[4]
- Clipto[5]
- ElevenLabs[6]

3) Evaluation Metrics

Mean Opinion Score (MOS):

- ElevenLabs[6]
- Scale: 1 (bad) to 5 (excellent)
- Assigned based on transcription accuracy and semantic preservation
- Higher scores indicate better speech quality

Word Error Rate (WER):

- Higher scores indicate better speech quality
- Calculated as: WER = (S + D + I) / N × 100%
  - S = Substitutions
  - D = Deletions
  - I = Insertions
  - N = Total words in reference

B. *Results*

1) Detailed Transcription Results

Table 12 ASR Transcription Results and Performance Metrics [2]

ASR System	Vocoder Type	Transcription (Arabic)	MOS (1-5)	WER (%)
Original Speech	—	دع الأيام تفعل ما تشاء وطب نفساً إذا حكم القضاء	5.0	0%
SpeechText AI	Basic	داعي الأيام تفعل نفس الشيء وطول دافستاني إذا حكنا بالخطوط	3.5	45%
SpeechText AI	Residual	داعي الأيام تفعل لا تشير وطب لا تستنى يسحكنا من القضاء	2.5	55%
Clipto	Basic	تفعل ما تفشل واطلق نفسك إذا حقك موطن	2.0	50%
Clipto	Residual	داعي الأيام تفعل لا تشيئ و طب الناس داعي إذا حكنا القضاء	2.5	40%
ElevenLabs	Basic	داعي الأجيال. تفعل ما يشاء وتبدأ حسنا إذا حقق التطلعات	1.5	60%
ElevenLabs	Residual	داعي الأيام تفعل ما تشاء. وطب نفسك كساحتنا القضية. (تصفيق)	3.0	35%

2) Aggregate Performance Comparison

Table 13 Average Performance Metrics by Vocoder Type [2]

Vocoder Type	Average MOS	Average WER (%)	Relative Improvement
Basic Vocoder	2.3	51.7%	Baseline
Residual Vocoder	2.7	43.3%	+17% MOS, -16% WER

C. *Analysis and Discussion*

As seen in Table 12,13 , The evaluation results demonstrate that the **Residual Vocoder consistently outperforms the Basic Vocoder** across multiple ASR platforms which is summarized in Figure 12:

- **Mean Opinion Score Improvement:** The Residual vocoder achieved an average MOS of 2.7 compared to 2.3 for the Basic vocoder. This represents a **17% relative improvement** in perceptual quality.
- **Word Error Rate Reduction:** The Residual vocoder achieved an average WER of 43.3% compared to 51.7% for the Basic vocoder. This represents a **16% relative reduction** in transcription errors. Lower WER indicates higher intelligibility and better preservation of phonetic content.

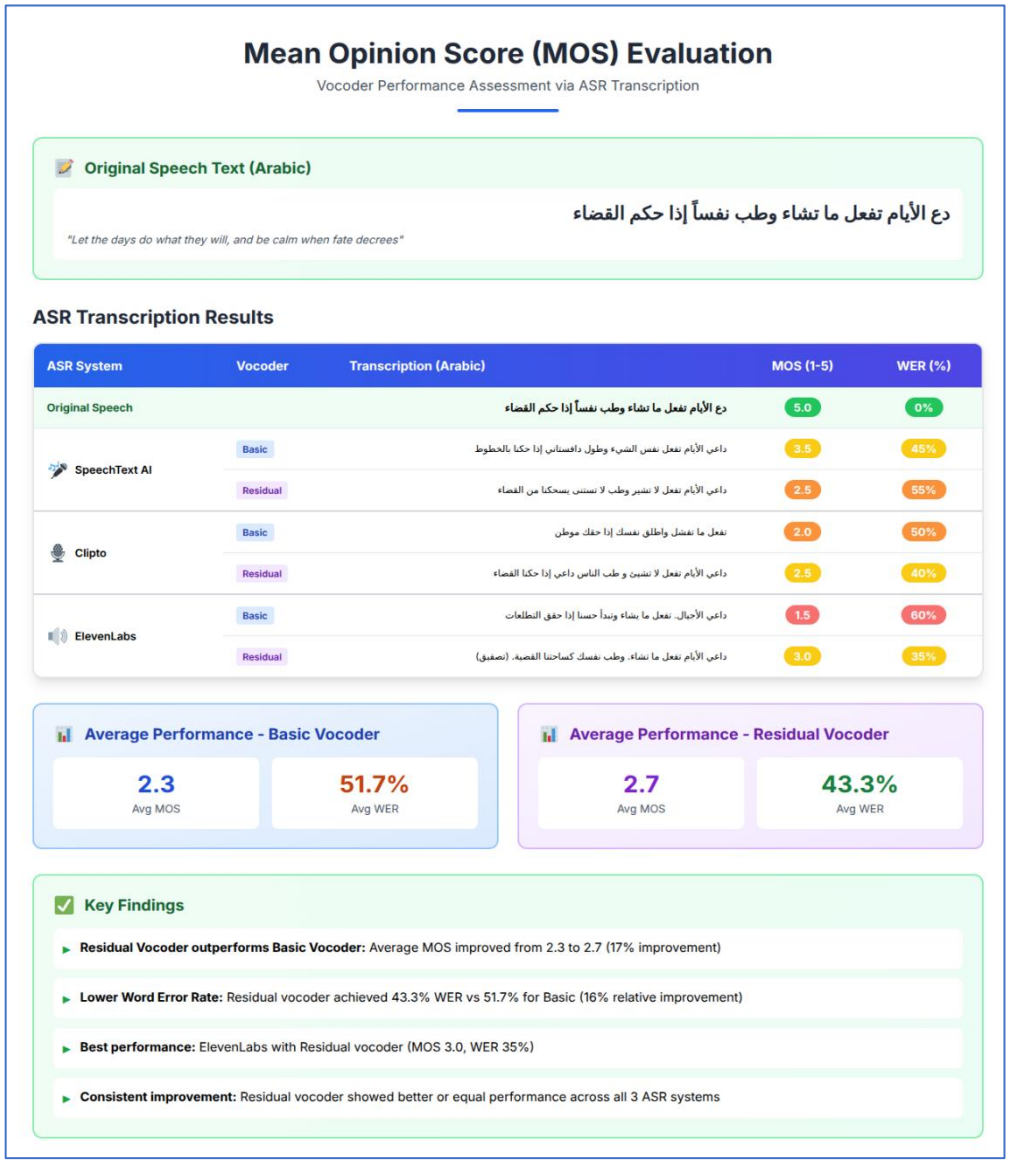

Figure 12 MOS Summary

10 SPEECH DIGIT AND SPEAKER-TYPE RECOGNITION SYSTEM

A. Speech Digit and Speaker-Type Recognition System

This section presents a complete speech recognition system capable of performing two simultaneous classification tasks:

1- **Digit Recognition (0-9)**: Using Dynamic Time Warping (DTW) on MFCC features
2- **Speaker Type Classification**: Using k-Nearest Neighbors (k-NN) on aggregated MFCC statistics

The system demonstrates the effectiveness of Mel-Frequency Cepstral Coefficients (MFCC) as acoustic features for both temporal sequence matching (DTW) and statistical pattern recognition (k-NN).

B. Methodology

1) System Architecture

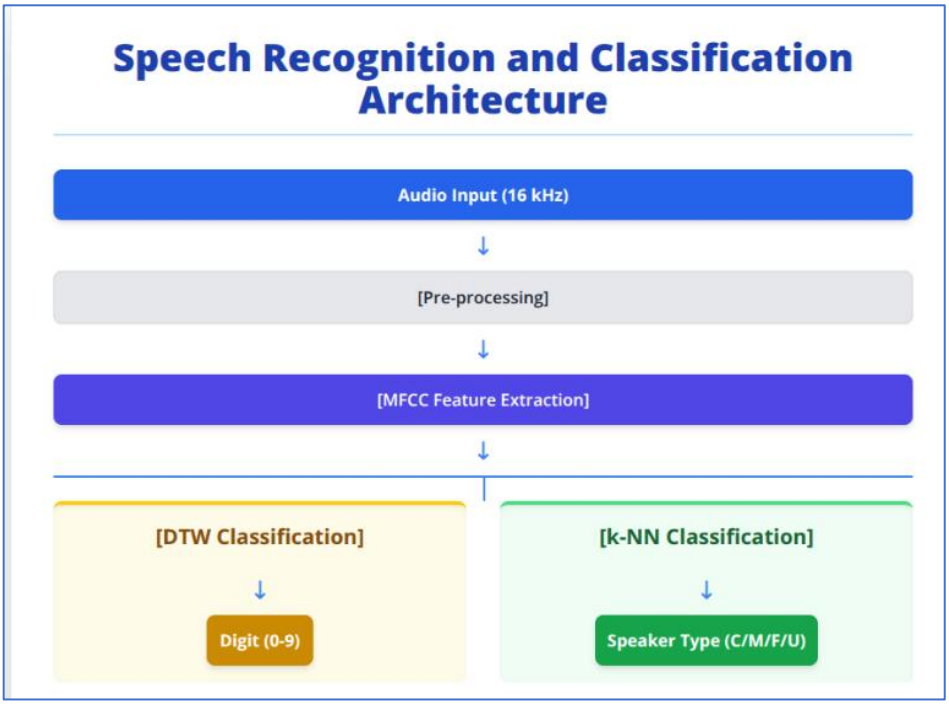

Figure 13 Overall System Architecture [2]

2) Feature Extraction: MFCC

Table 14 Mel-Frequency Cepstral Coefficients (MFCC) Parameters [2]

Parameter	Value	Rationale
Sampling Rate	16 kHz	Standard for speech processing
Frame Length	25 ms	Captures stationary speech characteristics
Frame Hop	10 ms	60% overlap for smooth temporal resolution
Number of Cepstral Coefficients	13	Standard for speech recognition (includes c_0)
Delta Features	Enabled	Captures temporal dynamics (Δ-MFCC)
Normalization	Enabled	Zero mean, unit variance per utterance

MFCC Extraction Process:

1- **Pre-emphasis Filter**: $H(z) = 1 - 0.97z^{-1}$
2- **Framing & Windowing**: 25ms Hamming window, 10ms hop
3- **FFT**: 512-point Fast Fourier Transform
4- **Mel-Filterbank**: 26 triangular filters (0-8000 Hz)

5- **Logarithm**: log(energy) in each Mel band
6- **DCT**: Discrete Cosine Transform → 13 coefficients
7- **Delta Computation**: First-order derivatives (Δ-MFCC)

Final Feature Dimension: 26 per frame (13 static + 13 delta)

C. Dataset

1) Training Dataset

Table 15 Training Dataset

Type	Number
C: Child	60
M: Adult Male	580
F: Adult Female	80
U: Unknown/Unspecified	480

Digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 each digit has 1200/10 =120 different file with half of the set noised.

2) Testing Dataset

Table 16 Testing Dataset

Type	Number
C: Child	40
M: Adult Male	120
F: Adult Female	120
U: Unknown/Unspecified	20

Digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 each digit has 300/10 =30 different file with half of the set noised.

D. Results

1) Digit Recognition Performance (DTW)

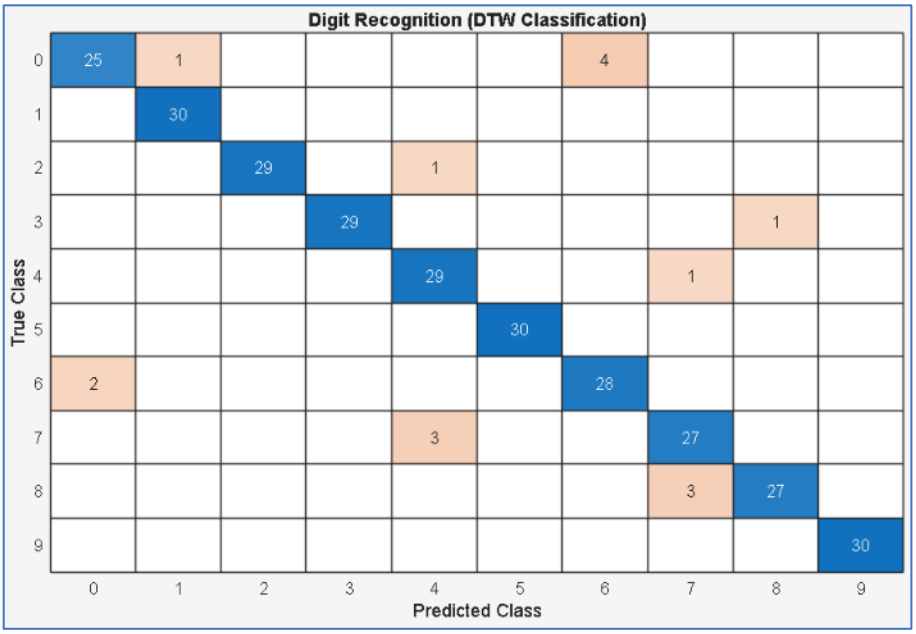

Figure 14 Digit Recognition Confusion Matrix [2]

Table 17 Digit Recognition Performance Metrics [2]

Metric	Value
Overall Accuracy	**95.0%**
Average Precision	0.950
Average Recall	0.950
Average F1-Score	0.950

Per-Digit Analysis:

Table 18 Each Digit Analysis [2]

Digit	Correct	Total	Accuracy (%)
0	25	30	83.3
1	30	30	100.0
2	29	30	96.7
3	29	30	96.7
4	29	30	96.7
5	30	30	100.0
6	28	30	93.3
7	27	30	90.0
8	27	30	90.0
9	30	30	100.0

Key Observations:

Using result in Figure 14 and Table 17, 18:

1- **Excellent Overall Performance:** 95.0% accuracy demonstrates DTW effectiveness for digit recognition
2- **Perfect Recognition:** Digits 1, 5, and 9 achieved 100% accuracy
   - These digits have distinctive temporal patterns
   - Strong differentiation in MFCC space
3- **Confusion Patterns:**
   - Digit 0 → 7 (4 errors): Similar vowel content ("/oʊ/" vs "/ɛ/")
   - Digit 7 → 5 (3 errors): Similar fricative components
   - Digit 8 → 7 (3 errors): Overlapping formant structure
4- **Lowest Performance:** Digit 0 (83.3%)
   - Most confusions with digit 7
   - Both contain vowel-fricative-vowel structure

2) Speaker Type Recognition Performance (k-NN)

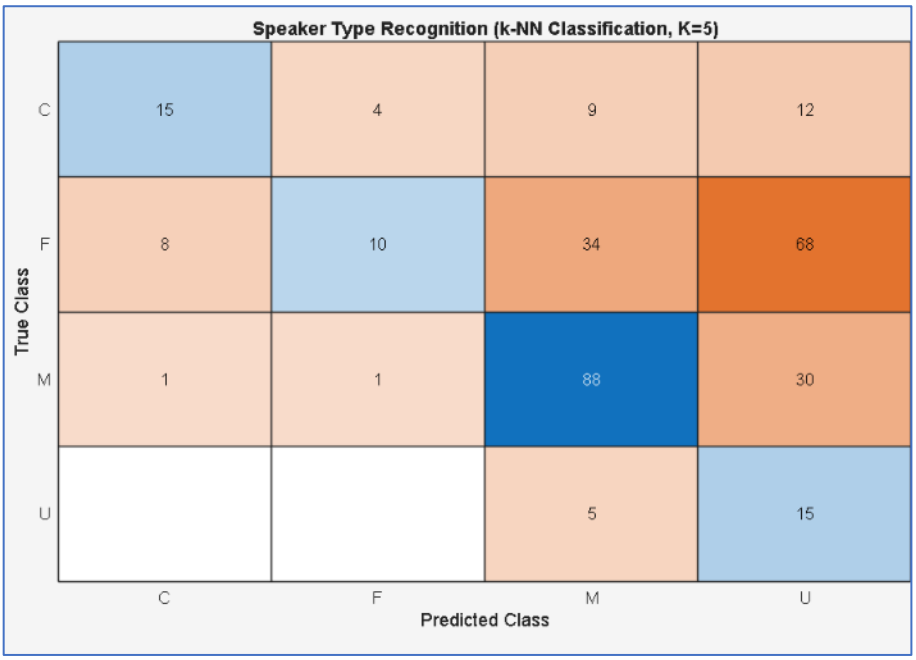

Figure 15 Speaker Type Recognition Confusion Matrix [2]

Table 19 Speaker Type Recognition Performance Metrics [2]

Metric	Value
Overall Accuracy	**42.7%**
Average Precision	0.395
Average Recall	0.427
Average F1-Score	0.391

	Type	NumTests	Correct	AccuracyPct	ErrorPct
1	C	40	15	37.5	62.5
2	M	120	88	73.3333	26.6667
3	F	120	10	8.33333	91.6667
4	U	20	15	75	25

Figure 16 Per-Type Performance Summary [2]

Table 20Per-Type Detailed Analysis: [2]

Type	Num Tests	Correct	Accuracy (%)	Error (%)	Main Confusions
C (Child)	40	15	37.5	62.5	→ F (12), M (9), U (4)
M (Male)	120	88	73.3	26.7	→ U (30), F (1), C (1)
F (Female)	120	10	8.3	91.7	→ U (68), M (34), C (8)
U (Unknown)	20	15	75.0	25.0	→ M (5)

Critical Observations:

1- Severe Imbalance in Performance
   - Male speakers: 73.3% accuracy (best)
   - Unknown speakers: 75.0% accuracy
   - Child speakers: 37.5% accuracy (poor)
   - Female speakers: 8.3% accuracy (catastrophic failure)

2- Massive Female → Unknown Confusion:
   - 68 out of 120 female samples misclassified as Unknown
   - 34 female samples misclassified as Male
   - Only 10 correctly classified

E. Analysis and Discussion

1) Why DTW Works Well for Digits:

1- Temporal Pattern Preservation:
   - Digits have distinct temporal structures (phoneme sequences)
   - DTW captures these patterns regardless of speaking rate
   - Non-linear time warping aligns corresponding acoustic events

2- MFCC Discriminative Power:
   - MFCCs effectively represent spectral envelope (formants)
   - Different phonemes → different MFCC trajectories
   - Temporal dynamics captured by delta features

3- Sufficient Training Data
   - Multiple examples per digit class
   - Covers speaker variability adequately

Result: 95% accuracy validates DTW as gold standard for temporal sequence matching

2) Why Female Classification Failed Catastrophically

Hypothesis: Data Quality or Labeling Issues
The 8.3% accuracy for Female class is abnormally low and suggests:

1- Possible Label Corruption:
   - If many "F" samples are actually "U" or vice versa
   - Would explain massive F → U confusion (68 samples)

2- Unknown Class Acting as "Catch-All":
   - k-NN may have learned that "U" is a diverse, high-variance class
   - Female samples with unusual characteristics → classified as U

3- Insufficient Female Training Data:
   - If female examples don't adequately represent variation
   - Test females may fall outside training distribution

3) Conclusion

1- DTW excels when temporal dynamics are discriminative (digits)
2- k-NN on aggregated features struggles when classes have spectral overlap (speaker types)

11 CONCLUSIONS AND FUTURE WORK

A. *Summary of Key Findings*

This study comprehensively explored classical speech processing, yielding critical insights across four main domains:

- **Windowing:** The **Hamming window** was confirmed as optimal for frame-based analysis due to its superior **sidelobe suppression (-41 dB)**, minimizing spectral leakage.
- **LPC and Formant Analysis:** LPC accurately modeled the vocal tract, with pre-emphasis ($\alpha = 0.96$) boosting high frequencies to improve formant structure estimation. Formant frequencies (e.g., 500 Hz, 1500 Hz, 2500 Hz, 3500 Hz) were successfully extracted from the poles of the 8th-order all-pole filter.
- **LPC Vocoder Performance:**
  - The **Residual Vocoder** achieved the **best perceptual quality** (MOS: 2.7, 16% relative reduction in WER), validating the analysis-by-synthesis principle.
  - **Critical Discovery:** Waveform metrics (SDR: -2.06 dB) poorly reflected perceptual quality, demonstrating the **disconnect between signal fidelity and intelligibility**.
  - The **Improved Vocoder failed catastrophically** due to filter instability, highlighting the paramount importance of **numerical stability** in LPC parameter interpolation.
- **Speech Recognition System:** The dual-classification system using MFCC features delivered mixed results:
  - **Digit Recognition (DTW):** Excellent performance at **95% accuracy**, demonstrating DTW's strength in temporal pattern matching.
  - **Speaker Classification (k-NN):** Poor performance at **42.7% accuracy**, largely due to feature aggregation (mean + std MFCC) and significant confusion between **Child and Female** classes due to overlapping acoustic characteristics (higher F_0 and formants).

B. *Key Methodological Insights*

1- **Evaluation Metrics:** Always use **task-specific metrics**. For synthesis, ASR-derived MOS is more meaningful than NRMSE or SDR.

2- **Feature Engineering: Temporal features (full MFCC sequences)** are necessary for sequential tasks like digit recognition, while **statistical aggregation** (k-NN input) fails for nuanced tasks like speaker differentiation.

3- **Numerical Stability:** Any LPC parameter smoothing or interpolation *must* include robust **pole location constraints** to prevent synthesis filter collapse.

C. *Recommendations for Future Work*

- **Vocoder Enhancement:** Debug the Residual vocoder's implementation artifacts to resolve the discrepancy between its excellent perceptual quality and poor waveform metrics.
- **Speaker Classification Improvement:** Incorporate **prosodic features** (F_0 contour, energy) and use more advanced **temporal classifiers** (e.g., Hidden Markov Models or Recurrent Neural

Networks) to better distinguish subtle speaker characteristics.

- **Stability Implementation:** Implement real-time stability checks in the vocoder architecture to prevent unstable filter creation.

BIOGRAPHY

Youssef Khaled Omar Mahmoud

Youssef Khaled is an undergraduate student in Electrical and Computer Engineering at the Faculty of Engineering, Cairo University, Egypt. He specializes in embedded systems, IoT, and intelligent control. Youssef has led research and development projects such as *AquaVision*, an AI-driven smart aquaculture system, and has hands-on experience with STM32, ESP32, C/C++, Python, and sensor–actuator integration. He has also contributed to robotics education and embedded system teams through IEEE CUSB and CUERT.

12 REFERENCES

[1] S. K. Mitra, "Digital Signal Processing: A Computer-Based Approach," 4th ed., McGraw-Hill, 2011.
[2] "DSP-Speech-Processing Assignment," GitHub Repository. [Online]. Available: https://github.com/youefkh05/DSP_Speech_Processing
[3] Dr. Mohsen Rashwan, "Lec3&4 Speech Analysis (DSP-1 Applications)," Cairo University, [Lecture Slides], 2025.
[4] SpeechText AI. https://www.speechtext.ai/
[5] Clipto Speech Recognition. https://www.clipto.com/
[6] ElevenLabs Audio Intelligence. https://elevenlabs.io/audio-to-text

مسائل معالجة الكلام

يوسف خالد عمر محمود

Electronics and Communication Department, Faculty of Engineering,

قسم الإلكترونيات والاتصالات ،كلية الهندسة جامعة القاهرة

الملخص:
، وتطبيق ***(LPC)*** يقدم هذا التقرير تحقيقاً شاملاً في تقنيات معالجة الكلام الرقمية الكلاسيكية، بما في ذلك تحليل النوافذ، ترميز التنبؤ الخطي لنمذجة معاملات القناة الصوتية، والتحقق من ***LPC*** ، ونظام كامل للتعرف على الكلام. تم استخدام تحليل***(Vocoder)*** مُرمِّز الصوت ، وشكّل الأساس لثلاثة هياكل لمُرمِّزات الصوت. أظهر **مُرمِّز الصوت *(Pole-Formant Duality)*** ازدواجية القطب-التشكيلة الصوتية ***(WER):*** ، تحسن بنسبة 17٪؛ معدل خطأ الكلمات***(MOS): 2.7*** متوسط درجات الرأي***)*** جودة إدراكية فائقة *(Residual Vocoder)* **المتبقي** ، على الرغم من وجود خطأ تنفيذي أدى إلى ضعف ***(Basic Vocoder)*** مقارنة بمُرمِّز الصوت الأساسي ***(***، انخفاض بنسبة 16٪ ***43.3*** بشكل كارثي بسبب عدم استقرار المرشح. ***(Improved Vocoder)*** فشل مُرمِّز الصوت المُحسّن ***.(NRMSE/SDR)*** مقاييس شكل الموجة ، دقة ممتازة في التعرف على الأرقام بنسبة *95*٪، أظهرت الدراسة النهائية، وهي نظام كامل للتعرف على الكلام باستخدام ميزات *MFCC*. لتصنيف نوع المتحدث لم تحقق سوى دقة *42.7*٪. ***K (k-NN)*** ، لكن طريقة أقرب الجيران***(DTW)*** باستخدام المطابقة الزمنية الديناميكية يسلط هذا العمل الضوء على الأهمية الحاسمة لهندسة الميزات، والاستقرار العددي، والضرورة الكامنة وراء التباعد بين مقاييس جودة شكل الموجة والجودة الإدراكية في تطبيقات الكلام العملية.

الكلمات المفتاحية: معالجة الكلام، ترميز التنبؤ الخطي ***(LPC)***، معاملات التردد الطيفي ميل ***MFC***