Robustness Analysis of Arabic Dialect Speech Recognition Under Additive Noise Conditions

Erini Hosny Maher

Abstract— **Automatic speech recognition (ASR) systems for Arabic often exhibit limited robustness in the presence of dialectal variation and environmental noise. This work investigates the impact of different noise types and signal-to-noise ratio (SNR) levels on ASR performance across five Arabic dialects: Modern Standard Arabic, Egyptian, Saudi, Levantine, and Moroccan Arabic. Clean speech data are generated using a multilingual text-to-speech system with dialect-specific speaker reference recordings to preserve pronunciation and prosodic characteristics. Controlled noise, including white, pink, brown, street, and babble noise, is added to the synthesized speech at SNR levels ranging from 0 dB to 20 dB. Both clean and noisy speech samples are transcribed using an automatic speech recognition system, and performance is evaluated by comparing the ASR output with the original text. Word Error Rate (WER) is employed as the primary evaluation metric. The results demonstrate that ASR performance degrades significantly at low SNR levels, particularly under white noise, while pink and brown noise exhibit minimal impact relative to clean speech. Babble and street noise cause moderate, SNR-dependent degradation. In addition, notable performance differences are observed across dialects, with Modern Standard Arabic achieving the lowest WER and Moroccan Arabic exhibiting the highest error rates, highlighting the combined effects of acoustic noise and dialectal variability on Arabic ASR robustness.**

Index Terms—**Automatic Speech Recognition (ASR), speech in noise, text-to-speech, word error rate (WER).**

I. Introduction

Automatic speech recognition (ASR) has become a core technology in modern human–computer interaction, enabling applications such as voice assistants, transcription systems, and accessibility tools. While significant progress has been achieved for high-resource languages and controlled acoustic conditions, ASR performance remains highly sensitive to both **dialectal variation** and **environmental noise**, particularly for languages with rich linguistic diversity such as Arabic.

Arabic is characterized by a diglossic structure, where **Modern Standard Arabic (MSA)** is used in formal contexts, while everyday communication occurs through a wide range of **regional dialects**, including Egyptian, Saudi, Levantine, and Moroccan Arabic. These dialects differ substantially in phonology, vocabulary, and morphology, often to the extent that mutual intelligibility is reduced. Despite this diversity, many ASR systems are trained primarily on MSA or limited dialectal data, leading to degraded performance when exposed to non-standard varieties of Arabic.

In addition to dialectal variability, **real-world acoustic conditions** present a major challenge for ASR systems. Background noise such as street sounds, competing speakers, and environmental interference is unavoidable in practical deployments. The presence of noise can significantly distort speech signals and obscure phonetic cues, resulting in increased recognition errors. Although noise robustness has been widely studied, most evaluations focus on a limited set of noise conditions or treat Arabic as a single homogeneous language, without examining how different dialects respond to noise degradation.

This reveals an important gap in existing research: **the lack of a controlled, dialect-level analysis of ASR robustness under varying noise conditions**. Conducting such an analysis using real recorded speech is often costly and difficult due to data availability, speaker variability, and inconsistent recording conditions. Text-to-speech (TTS) systems offer a practical alternative, enabling the generation of clean, controlled speech data while preserving linguistic content and allowing systematic manipulation of acoustic conditions.

In this work, a multilingual zero-shot TTS model is employed to generate speech samples for multiple Arabic dialects using short reference recordings to capture speaker identity and prosodic characteristics. Controlled noise is then added at various signal-to-noise ratio (SNR) levels to simulate realistic acoustic environments. An ASR system is subsequently used to transcribe both clean and noisy speech, and performance is evaluated using Word Error Rate (WER). This framework allows a fair and systematic comparison of noise robustness across dialects while isolating the effects of noise type and SNR.

The main contributions of this work are summarized as follows:

Contributions:

- Construction of a controlled multi-dialect Arabic speech dataset using a text-to-speech system.
- Evaluation of ASR robustness across multiple noise

This work was conducted as part of an academic project in The DSP course at the Faculty of Engineering, Cairo University. The code and experimental scripts used in this study are publicly available at: **https://github.com/erinihosny/arabic-tts-asr-noise-robustness.git**

types and signal-to-noise ratio levels.

- Quantitative comparison of dialect-dependent performance degradation using Word Error Rate (WER).

II. Related work

A. Arabic Automatic Speech Recognition

Automatic speech recognition for Arabic presents unique challenges due to the language's diglossic nature and extensive dialectal diversity. While Modern Standard Arabic (MSA) is used in formal communication, spoken Arabic varies significantly across regions in terms of phonology, vocabulary, and morphology. These variations have been shown to cause substantial degradation in ASR performance when models trained on one variety are applied to another.

Several datasets have been developed to support Arabic ASR and dialect identification. Examples include the **King Saud University Arabic Speech Database [1]**, which focuses primarily on MSA, the **ADI-17 [3] and ADI-20 [4] datasets**, designed for Arabic dialect identification, and the **MGB-3 dataset [2]**, which contains broadcast speech covering multiple dialects. Additional resources such as **Common Voice (Mozilla) [6]** provide open-source Arabic speech data, while corpora like the **Arabic Speech Corpus** by Nawar Halabi [5] offer smaller, curated collections. Although these datasets have contributed significantly to Arabic ASR research, many are limited in terms of dialect coverage, speaker diversity, recording conditions, or public availability.

Due to these constraints, generating controlled multi-dialect speech data remains challenging. Recent work has therefore explored **the use of text-to-speech systems to synthesize speech for ASR evaluation [10]**, enabling precise control over linguistic content and acoustic conditions while avoiding the variability inherent in real-world recordings.

B. Noise Robustness in ASR

Robustness to environmental noise has long been a central research topic in automatic speech recognition. Additive noise, such as white noise, background music, street sounds, and competing speakers, can significantly distort speech signals and reduce recognition accuracy. To study these effects, ASR systems are commonly evaluated under controlled noise conditions using predefined signal-to-noise ratio (SNR) levels [11].

Previous studies have shown that different noise types affect ASR performance to varying degrees, with stationary noises such as white noise often causing more severe degradation than structured or low-frequency-dominant noise. Speech-like noise, such as babble, has been found to be particularly challenging due to its similarity to the target speech signal. Despite extensive work in this area, relatively few studies explicitly examine how noise robustness varies across Arabic dialects [12], motivating further investigation.

C. Evaluation Metrics

Word Error Rate (WER) is the most widely used metric for evaluating ASR performance and is defined as the normalized sum of substitution, insertion, and deletion errors between a reference transcript and a hypothesis. Due to its simplicity and interpretability, WER is commonly adopted in both academic research and industrial benchmarks. [9]

However, WER has known limitations, particularly for morphologically rich languages such as Arabic. Minor phonetic or morphological variations can result in large word-level penalties [9], and pronunciation ambiguities are not always reflected in transcript comparisons. Despite these limitations, WER remains a standard baseline metric and provides a consistent basis for comparing ASR performance across different noise conditions and dialects, as adopted in this work.

III. Dataset and Experimental Setup

This section describes the dataset construction and experimental setup adopted in this work, including the selected Arabic dialects, text data preparation, speech generation process, noise modeling, and the automatic speech recognition system used for evaluation.

A. Dialects and Text Data

Five Arabic varieties were considered in this study: **Modern Standard Arabic (MSA), Egyptian Arabic, Saudi Arabic, Levantine Arabic,** and **Moroccan Arabic**. These dialects were selected to represent a diverse range of phonological, lexical, and morphological characteristics commonly encountered in spoken Arabic.

For each dialect, **20 sentences** were used, resulting in a total of **100 sentences**. The sentences were extracted from a single source document and manually grouped according to dialect. Sentence lengths varied moderately, containing a mix of short and medium-length utterances, and covered everyday vocabulary commonly used in spoken communication. Care was taken to preserve dialect-specific lexical choices and syntactic structures, particularly for non-MSA varieties, to ensure realistic representation of each dialect.

Using the same number of sentences per dialect ensures a balanced evaluation and allows fair comparison of ASR performance across dialects and experimental conditions.

B. Speech Generation Using Text-to-Speech

Due to the limited availability of publicly accessible, high-quality, multi-dialect Arabic speech datasets with consistent recording conditions, a **text-to-speech (TTS)** approach was adopted to generate clean speech samples. This strategy enables precise control over linguistic content while avoiding speaker and recording variability that commonly affects real-world speech data.

Speech synthesis was performed using XTTS v2, **a multilingual zero-shot TTS model [7] [8]**, which allows voice cloning from a short reference recording. For each dialect, a single speaker reference waveform was used to ensure speaker consistency across all generated utterances within that dialect. XTTS v2 conditions speech generation on a short reference

waveform, from which speaker identity, prosodic characteristics, and dialectal pronunciation patterns are implicitly captured.

The quality of voice cloning produced by XTTS-v2 improves as the duration of the reference audio increases, since longer samples provide more speaker-specific acoustic and prosodic information. While the model is capable of operating with very short reference clips, longer clean recordings allow better capture of **voice tone, accent, and expressive nuances**, resulting in higher speaker similarity and more consistent synthesized speech. However, the benefit of increasing reference duration saturates beyond a certain point, as the model processes the reference audio through a fixed-size speaker embedding mechanism. In practice, reference durations in the range of **15–30 seconds** are sufficient to achieve stable and natural-sounding synthesis without significant additional gains from longer recordings.

For each dialect, a single reference speaker was used to ensure speaker consistency across all synthesized utterances. Reference audio samples were selected to be **clean speech recordings**, free from background noise and music, and spoken entirely in the target dialect. Using a single speaker per dialect minimizes speaker-induced variability and allows observed differences in ASR performance to be attributed primarily to dialectal and acoustic factors rather than speaker identity.

The reference recordings were selected to reflect natural pronunciation and prosody characteristic of the target dialect. They were extracted from publicly available YouTube videos, with sources documented in the accompanying code repository.

All synthesized speech samples were generated at a **sampling rate of 16 kHz**, which is commonly used in speech processing and matches the requirements of the ASR system employed in this work. The generated speech files were stored in **WAV format**, using single-channel (mono) audio to maintain compatibility with subsequent processing stages.

No fine-tuning or speaker-specific training was performed in this work

C. Noise Types

To simulate realistic acoustic environments, five different noise types were considered:

- **White Noise:** A stationary noise with equal power across all frequencies, representing a highly disruptive and spectrally uniform interference.
- **Pink Noise:** A low-frequency-dominant noise with power inversely proportional to frequency, commonly used to model natural background noise.
- **Brown Noise:** An even stronger low-frequency noise with power inversely proportional to the square of frequency.
- **Street Noise:** Environmental noise recorded in outdoor settings, including traffic and ambient sounds.
- **Babble Noise:** Speech-like noise generated from multiple overlapping speakers, representing crowded conversational environments.

The selected noise types cover a range of stationary and non-stationary conditions and allow evaluation of ASR robustness under diverse acoustic scenarios.

D. Noise Injection Procedure

Noise was added to the clean synthesized speech at controlled **signal-to-noise ratio (SNR)** levels of **0, 5, 10, 15, and 20 dB**. These levels span from extremely noisy to near-clean conditions and are commonly used in ASR robustness studies.

For a clean speech signal $s[n]$and noise signal $n[n]$, the noise scaling factor was computed to satisfy the desired SNR according to:

$$SNR_{dB} = 10\log_{10}(\frac{P_s}{P_n})$$

where P_s and P_n denote the average power of the speech and noise signals, respectively. The noise signal was normalized and scaled before being added to the speech waveform.

To ensure reproducibility and systematic evaluation, the noisy audio files were organized using a hierarchical directory structure based on **dialect**, **noise type**, and **SNR level**. Each clean utterance therefore had multiple noisy counterparts corresponding to different noise conditions.

E. Automatic Speech Recognition System

Automatic transcription was performed using a **pre-trained multilingual ASR model [9]** capable of recognizing Arabic speech without dialect-specific fine-tuning. The model operates on 16 kHz mono audio and performs end-to-end speech-to-text decoding.

In this work, speech transcription was performed using the **Whisper** "base" model, a multilingual automatic speech recognition system trained on large-scale, weakly supervised speech data covering multiple languages and acoustic conditions. The base model was selected as a trade-off between computational efficiency and recognition accuracy, making it suitable for large-scale evaluation while remaining representative of commonly deployed ASR systems.

The model performs end-to-end speech-to-text decoding without requiring language-specific fine-tuning. The ASR system was configured to explicitly transcribe Arabic speech, and default decoding parameters were used to ensure consistency across all experiments. This choice reflects a realistic evaluation scenario in which off-the-shelf ASR systems are applied to speech data without domain adaptation.

The selected ASR model is well-suited for this study due to its strong multilingual capabilities, robustness to moderate noise levels, and widespread use in recent speech recognition research. Its performance under controlled noise conditions provides meaningful insight into the combined effects of dialectal variation and acoustic degradation.

It should be noted that the Whisper base model does not explicitly model Arabic morphology or dialectal variation. Instead, it relies on learned acoustic–textual correlations, which may contribute to higher error rates for dialects that are underrepresented in the training data.

IV. EVALUATION METHODOLOGY

This section describes the evaluation methodology adopted in this work, including text normalization procedures applied to both reference and hypothesis transcripts and the use of Word Error Rate (WER) as the primary evaluation metric.

A. Text Normalization

Text normalization is a crucial preprocessing step when evaluating automatic speech recognition systems for Arabic. Due to the language's rich morphology and orthographic variation, minor spelling differences can lead to disproportionately large penalties during transcript comparison if left unaddressed. To ensure a fair and consistent evaluation, both the original reference texts and the ASR-generated hypotheses were normalized prior to WER computation.

The normalization process applied in this work consists of the following steps:

1. **Whitespace Normalization:**
   Leading and trailing whitespace is removed, and multiple consecutive spaces are collapsed into a single space. This prevents artificial word boundary mismatches caused by inconsistent spacing.
2. **Removal of Non-Arabic Characters:**
   All non-Arabic characters, including punctuation marks, numerals, and characters outside the Arabic Unicode range, are removed. This step ensures that evaluation focuses solely on Arabic lexical content and avoids penalties caused by transcription artifacts or model hallucinations in other scripts.
3. **Orthographic Normalization of Alef Variants:**
   Different forms of the Alef character (أ، إ، آ، ٱ) are normalized to a single canonical form (ا). This is particularly important in Arabic ASR evaluation, as these variants are frequently interchangeable in writing and are not always consistently predicted by ASR systems.
4. **Removal of Diacritics:**
   All Arabic diacritics (tashkīl), including short vowels and other pronunciation markers, are removed from both reference and hypothesis transcripts. Since the ASR system produces unvowelized text and most Arabic datasets are represented without diacritics, this step prevents unfair penalization due to diacritic mismatches that do not affect lexical identity.

After normalization, all text is represented using a consistent Arabic-only format, reducing sensitivity to orthographic variation while preserving the lexical and semantic content of the utterances. This preprocessing step significantly improves the reliability of WER as a comparative metric across dialects and noise conditions.

B. Word Error Rate

Word Error Rate (WER) is employed as the primary metric for evaluating ASR performance. WER is defined as the normalized sum of substitution (S), deletion (D), and insertion (I) errors between a reference transcript and a hypothesis transcript, and is computed as:

$$WER = \frac{S + D + I}{N}$$

where N is the total number of words in the reference transcript.

In this work, WER is computed using the 'jiwer' library, which provides a standardized and widely adopted implementation for ASR evaluation. WER values are calculated at the sentence level and then averaged across all utterances within each dialect, noise type, and signal-to-noise ratio (SNR) condition.

Despite its widespread use, WER has known limitations, particularly for morphologically rich languages such as Arabic. The metric relies on **exact word matching**, meaning that minor spelling differences or morphological variations are penalized in the same manner as complete word mismatches. For example, variations in verb inflection, clitic attachment, or orthographic conventions may result in high WER despite the underlying semantic correctness of the transcription.

Nevertheless, WER remains a standard and interpretable metric for ASR evaluation and provides a consistent basis for comparing system performance across different experimental conditions. Its use in this study enables direct comparison of noise robustness and dialectal effects under controlled settings, while the aforementioned limitations are acknowledged and discussed in the analysis of results.

While transcript-level metrics such as WER cannot capture all aspects of pronunciation or prosodic correctness, they provide a practical and widely accepted measure for evaluating ASR performance under varying acoustic conditions.

V. Results

A. Qualitative Observations

It was observed that the TTS model does not consistently disambiguate **morphologically ambiguous Arabic words when diacritics are absent**. For example, the word تعلم may be pronounced as تَعْلَم "you know" or "she knows") or as تَعَلُّم ("learning"), while قرأت may be realized as قَرَأْتُ ("I read") or قَرَأَتْ ("she read"). In such cases, the model selects a pronunciation based on learned statistical patterns from the training data rather than performing explicit semantic or grammatical inference from context. As a result, different valid pronunciations may be produced for identical written forms, even when the surrounding sentence does not strongly constrain the intended meaning.

This behavior cannot be detected through transcript comparison alone, as the textual outputs are identical across interpretations; instead, it was identified through manual inspection of the generated audio samples.

The generated speech exhibited consistent dialectal phonetic patterns and prosodic features matching the reference speaker. Beyond segmental pronunciation, the model also reproduced suprasegmental characteristics, including pause placement, intonation contours, speech rhythm, and overall speaking attitude. These prosodic cues influenced how morphologically ambiguous words were realized in speech. In the absence of diacritics, the selected pronunciation was often accompanied by corresponding prosodic patterns, such as subtle pauses, stress placement, and intonational emphasis, which conveyed different grammatical roles or speaker attitudes. These variations suggest that the model relies on learned correlations between acoustic prosody and linguistic context, inferred from the reference speaker and training data, rather than explicit morphological disambiguation. Consequently, multiple valid interpretations of the same written form may be rendered differently in speech, even when transcript-level evaluation remains unchanged.

B. Quantitative Approach

1. Per-dialect WER-vs-SNR Tables

Modern Standard Arabic

Noise Type	0 dB	5 dB	10 dB	15 dB	20 dB	Clean
White	93.1%	72.5%	56.8%	44.0%	38.2%	32.9%
Pink	32.6%	34.2%	34.7%	32.1%	33.8%	32.9%
Brown	35.1%	35.1%	33.6%	32.4%	32.4%	32.9%
Street	56.9%	44.9%	36.4%	33.5%	32.1%	32.9%
Babble	55.5%	42.4%	37.1%	33.2%	34.7%	32.9%

Table 1: WER vs. different noise levels for modern standard Arabic

Egyptian dialect

Noise Type	0 dB	5 dB	10 dB	15 dB	20 dB	Clean
White	94.1%	81.6%	66.0%	56.7%	52.7%	52.8%
Pink	53.9%	53.0%	53.7%	54.5%	53.3%	52.8%
Brown	52.3%	55.6%	56.2%	55.1%	55.1%	52.8%
Street	66.1%	59.6%	54.8%	50.0%	54.4%	52.8%
Babble	69.1%	55.7%	55.9%	49.9%	54.5%	52.8%

Table 2: WER vs. different noise levels for Egyptian Arabic dialect

Saudi dialect

Noise Type	0 dB	5 dB	10 dB	15 dB	20 dB	Clean
White	96.8%	86.3%	74.1%	60.0%	51.9%	48.6%
Pink	48.7%	52.1%	50.3%	49.2%	50.3%	48.6%
Brown	50.5%	49.9%	49.2%	49.2%	49.2%	48.6%
Street	73.0%	56.1%	54.5%	50.5%	49.5%	48.6%
Babble	68.5%	57.3%	52.5%	53.0%	52.4%	48.6%

Table 3: WER vs. different noise levels for Saudi Arabic dialect

Levantine dialect

Noise Type	0 dB	5 dB	10 dB	15 dB	20 dB	Clean
White	96.2%	77.8%	67.9%	54.9%	51.5%	49.6%
Pink	54.9%	52.4%	48.5%	48.2%	49.4%	49.6%
Brown	52.1%	51.5%	48.8%	48.3%	488%	49.6%
Street	66.6%	56.6%	55.4%	55.4%	54.6%	49.6%
Babble	69.5%	61.8%	59.3%	54.7%	54.8%	49.6%

Table 4: WER vs. different noise levels for Levantine Arabic dialect

Moroccan dialect

Noise Type	0 dB	5 dB	10 dB	15 dB	20 dB	Clean
White	101.3%	97.6%	91.0%	78.7%	78.1%	70.5%
Pink	76.1%	72.9%	69.9%	69.0%	69.6%	70.5%
Brown	68.4%	68.4%	69.0%	69.6%	69.6%	70.5%
Street	96.1%	77.8%	71.7%	72.9%	72.0%	70.5%
Babble	91.2%	75.8%	72.6%	70.9%	67.5%	70.5%

Table 5: WER vs. different noise levels for Moroccan Arabic dialect

The clean condition represents a noise-free baseline and is therefore identical across all noise types.

Across all dialects, white noise consistently results in the highest WER, especially at low SNR levels:

- At 0 dB, WER exceeds 0.93 for all dialects and reaches 1.01 for Moroccan Arabic.
- Even at 20 dB, white-noise WER remains significantly worse than clean performance.

This behavior is expected, as white noise uniformly masks the speech spectrum, severely degrading phonetic cues critical for ASR.

Pink and brown noise show WER values very close to the clean condition, across all SNR levels and dialects:

- For most dialects, WER under pink/brown noise fluctuates within ±0.02 of clean WER.
- In some cases, noisy WER is even marginally lower than clean WER, likely due to statistical variation or regularization effects.

This indicates that low-frequency-dominant noise has limited impact on ASR performance, particularly when speech energy dominates the spectrum.

Babble and street noise demonstrate clear but gradual degradation trends:

- At 0–5 dB, WER increases substantially relative to clean speech.
- Performance improves steadily as SNR increases, approaching clean WER at 15–20 dB.

Babble noise consistently results in slightly higher WER than street noise, likely due to speech-like interference, which is more confusable for ASR systems than stationary environmental noise.

For all dialects and noise types:

- WER under noisy conditions **never outperforms clean WER in a sustained manner**.
- As SNR increases, noisy WER consistently converges toward the clean baseline.

This confirms that the clean condition serves as a **performance lower bound**, validating its use as a shared reference across noise types.

These results highlight several important implications:

- **WER alone may not capture pronunciation or morphological ambiguities**, especially in Arabic, as discussed in the qualitative analysis.
- Dialectal diversity plays a major role in ASR robustness and must be explicitly evaluated.
- Noise-aware training or dialect-specific adaptation would likely yield substantial improvements, particularly for under-resourced dialects such as Moroccan Arabic.

Although higher SNR generally leads to improved ASR performance, the observed WER does not strictly decrease monotonically with SNR. This behavior is attributed to multiple sources of variability:

- Finite sample size, which causes statistical variance.
- Stochastic noise generation, which breaks monotonicity
- Probabilistic ASR decoding effects: Modern ASR systems (including Whisper) [9]: use beam search or sampling, rely on probabilistic token selection and are sensitive to small acoustic differences. A slightly different waveform can cause different token boundaries, repeated words, and insertions or deletions even if audio quality improves.
- Particularly in morphologically rich languages such as Arabic. Arabic is especially sensitive to

prefixes (و، ف، ب، ل), clitics and verb inflections. A single phonetic change can split one word into two, merge two words or alter word boundaries, which WER penalizes this heavily. So a "better" acoustic signal can still produce a worse WER.

2. Clean WER per Dialect

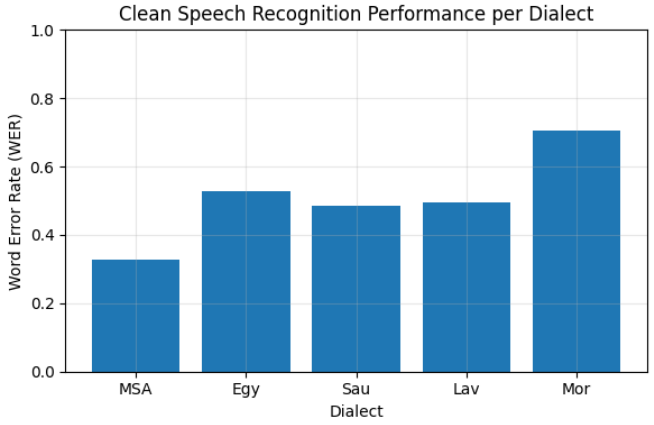

Figure 1:Average word error rate (WER) for clean speech across the evaluated Arabic dialects. This figure highlights intrinsic recognition difficulty differences between dialects in the absence of additive noise.

As shown in Fig. 1, clean-speech ASR performance varies significantly across dialects. The clean-condition results reveal substantial variation in baseline ASR performance across Arabic dialects:

- MSA achieves the lowest clean WER (0.329), indicating the highest transcription accuracy.
- Moroccan Arabic exhibits the highest clean WER (0.705), followed by Egyptian, Saudi, and Levantine dialects.

This trend aligns with expectations, as modern ASR systems are typically trained on more standardized and resource-rich varieties, such as MSA, while underperforming on dialects with higher phonological and lexical divergence, particularly Moroccan Arabic.

3. WER vs SNR per Dialect

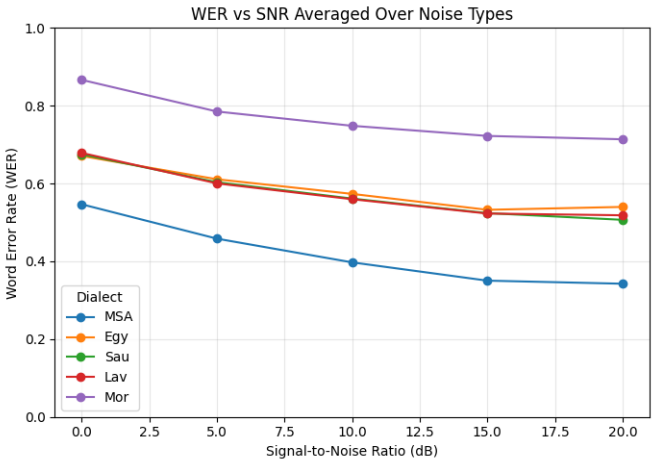

Figure 2: Average word error rate (WER) as a function of signal-to-noise ratio (SNR), averaged over all noise types, for each Arabic dialect. Higher WER at lower SNRs indicates reduced robustness to noise.

As shown in Fig. 2, Across all noise types and SNR levels, dialect robustness closely follows the clean-condition ranking: **MSA > Saudi ≈ Levantine > Egyptian > Moroccan**

This suggests that noise sensitivity amplifies existing weaknesses rather than introducing new ones. Dialects with higher baseline WER degrade more severely under adverse acoustic conditions.

A clear overall trend is observed: WER decreases as SNR increases for all noise conditions, confirming that improved acoustic quality generally leads to better recognition accuracy.

4. WER vs SNR per Noise Level

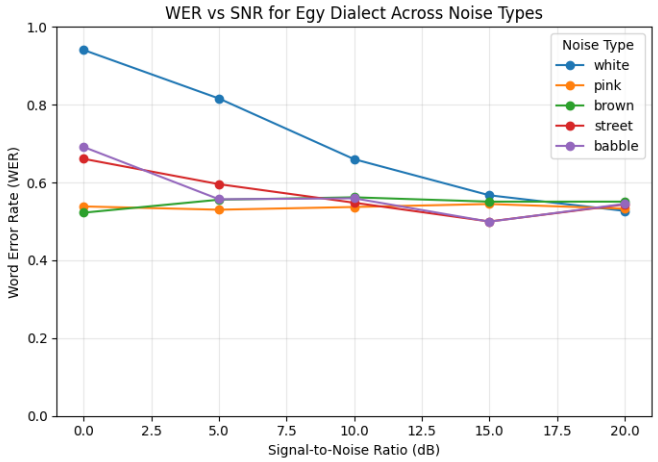

Figure 3: Word error rate (WER) versus signal-to-noise ratio (SNR) for the Egyptian Arabic dialect under different noise types. Babble and street noise exhibit stronger degradation compared to stationary noise sources.

Figure 3 illustrates the effect of different noise types and signal-to-noise ratio (SNR) levels on the ASR performance for the Egyptian dialect, measured in terms of Word Error Rate.

White noise results in the most severe degradation, particularly at low SNRs, with WER close to 0.95 at 0 dB. This behavior is expected, as white noise uniformly corrupts the entire frequency spectrum of speech, making it especially harmful for ASR systems. As SNR increases, the WER under white noise decreases steadily, approaching the performance of other noise types at higher SNRs.

Pink and brown noise exhibit much smaller sensitivity to SNR changes, with relatively flat WER curves across all tested levels. Their impact remains close to the clean-speech baseline, suggesting that low-frequency-dominant or spectrally shaped noises are less disruptive to the Whisper model for this dialect.

Street and babble noise show intermediate behavior. At low SNRs, both significantly degrade performance, though less severely than white noise. Babble noise, in particular, reflects the challenge posed by speech-like interference, which competes directly with the target speech signal. As SNR increases beyond 10–15 dB, WER under both noise types converges toward similar values, indicating partial robustness in moderate noise conditions.

5. Heatmap of WER Across Arabic Dialects as a single SNR level

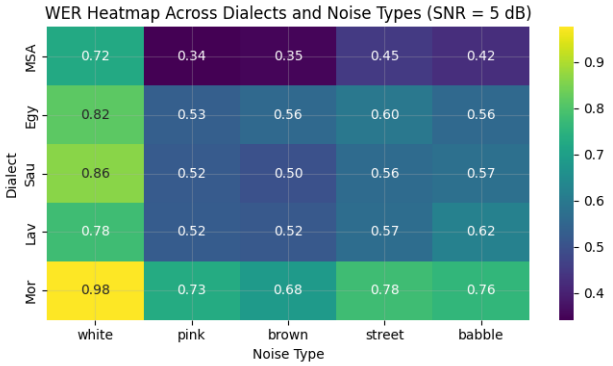

Figure 4:Heatmap of word error rate (WER) across Arabic dialects and noise types at 5 dB SNR. Darker colors indicate higher error rates, revealing strong interactions between dialectal variability and noise characteristics.

Fig. 4 demonstrates that ASR performance is highly dependent on both dialect and noise type, with white noise causing the most severe degradation and pink/brown noise having minimal impact. While MSA remains the most robust dialect, Moroccan Arabic consistently exhibits the highest error rates, highlighting the need for dialect-aware and noise-robust ASR models.

6. Grouped bar chart for ONE dialect

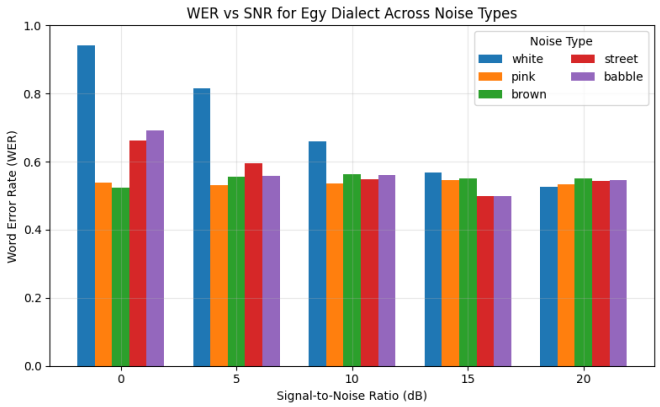

Figure 5: Word error rate (WER) for the Egyptian Arabic dialect under different noise types at varying signal-to-noise ratio (SNR) levels. Each SNR group shows the relative impact of stationary and non-stationary noise sources on recognition performance.

As shown in Fig. 5, white noise causes the most severe degradation in ASR performance across all dialects, particularly at low SNR levels. In contrast, babble and street noise introduce moderate, SNR-dependent degradation, with WER values gradually converging toward clean-speech performance as SNR increases.

This behavior can be attributed to the uniform spectral masking introduced by white noise, which disrupts phonetic cues across the entire frequency range, whereas structured environmental noise preserves portions of the speech spectrum.

VI. Discussion

The experimental results highlight the combined impact of **dialectal variation** and **acoustic noise** on Arabic automatic speech recognition performance. The observed trends provide insight into both the linguistic characteristics of Arabic dialects and the limitations of contemporary ASR models when operating under adverse conditions.

A. Dialect-Dependent ASR Performance

A clear performance gap is observed across dialects, with **Modern Standard Arabic (MSA)** consistently achieving the lowest Word Error Rate (WER) under clean and noisy conditions, while **Moroccan Arabic** exhibits the highest error rates. Egyptian, Saudi, and Levantine Arabic occupy intermediate positions. This ordering reflects differences in data availability and linguistic standardization rather than purely acoustic factors.

MSA benefits from extensive representation in training data for modern ASR systems, as it is the dominant form used in formal speech, news broadcasts, and written content. In contrast, dialects such as Moroccan Arabic diverge significantly from MSA in phonology, lexicon, and syntax, and are underrepresented in most large-scale training corpora. As a result, ASR models trained predominantly on MSA or limited dialectal data struggle to generalize effectively to these varieties.

The fact that dialectal ranking remains consistent across noise conditions suggests that **noise amplifies existing recognition weaknesses** rather than introducing new error patterns. Dialects with higher baseline WER degrade more severely as acoustic conditions worsen, indicating that robustness to noise is closely tied to the quality of linguistic modeling.

B. Impact of Phonetic and Morphological Variability

Arabic dialects exhibit substantial **phonetic variability**, including vowel reduction, consonant shifts, and dialect-specific realizations of common phonemes. These variations pose challenges for ASR systems that rely on learned acoustic–phonetic mappings, particularly when such mappings are biased toward MSA pronunciations.

In addition, Arabic's **rich morphology** contributes to higher error rates under both clean and noisy conditions. Small acoustic distortions can lead to incorrect recognition of prefixes, suffixes, or clitics, resulting in word-level substitutions or deletions that significantly affect WER. This effect is especially pronounced in dialectal speech, where morphological constructions may differ from standardized forms and are less consistently represented in training data.

These linguistic properties help explain why improvements in signal quality, such as increasing the signal-to-noise ratio (SNR), do not always produce strictly monotonic reductions in WER. Even when acoustic clarity improves, morphological ambiguity and phonetic variation can still lead to recognition errors.

C. Effect of Noise Type and SNR

The experimental results demonstrate that **noise type plays a critical role** in ASR degradation. White noise consistently causes the most severe performance degradation, particularly at low SNR levels. This behavior is expected, as white noise

uniformly masks the speech spectrum and disrupts phonetic cues across all frequencies.

In contrast, **pink and brown noise** have relatively minor impact on ASR performance, with WER values often remaining close to those observed under clean conditions. These noise types are dominated by low-frequency components, which appear to interfere less with the spectral regions most critical for speech recognition.

Babble and street noise produce moderate but noticeable degradation, particularly at low SNR levels. Babble noise is especially challenging due to its speech-like characteristics, which increase the likelihood of confusion between target speech and background interference. As SNR increases, performance under these noise conditions gradually converges toward the clean baseline, confirming that the ASR system retains some robustness to structured environmental noise.

D. Clean Speech as a Performance Baseline

Across all dialects and noise conditions, clean-speech performance serves as a lower bound toward which noisy-condition WER converges as SNR increases. This behavior validates the use of the clean condition as a shared reference across noise types and supports the interpretation of noise-induced degradation as a relative loss from baseline performance rather than an independent phenomenon.

However, the relatively high clean WER observed for some dialects, particularly Moroccan Arabic, indicates that improving noise robustness alone is insufficient to achieve reliable ASR performance. Instead, improved dialectal coverage and linguistic modeling are necessary to address the root causes of recognition errors.

E. Implications for Arabic ASR Systems

The findings of this study suggest that current ASR systems remain **highly sensitive to both dialectal variation and acoustic noise**, with dialectal mismatch often dominating performance degradation. While noise-robust modeling can mitigate some errors, it cannot compensate for limited dialect representation in training data.

These results emphasize the need for:

- Larger and more diverse dialectal Arabic speech corpora
- Dialect-aware or dialect-adaptive ASR models
- Evaluation frameworks that consider both linguistic and acoustic variability.

Furthermore, they demonstrate the value of controlled experimental setups using synthesized speech, which allow systematic analysis of individual factors affecting ASR performance while maintaining consistent linguistic content.

VII. Limitations and Future Work

Although the proposed experimental framework enables controlled and systematic evaluation of Arabic ASR performance under varying noise conditions and dialects, several limitations should be acknowledged, which also motivate directions for future work.

A. Limitations

First, the speech data used in this study are generated using a text-to-speech (TTS) system rather than recorded from real speakers. While synthesized speech allows precise control over linguistic content, speaker consistency, and acoustic conditions, it may not fully capture the variability present in real-world speech, such as spontaneous speaking styles, disfluencies, microphone characteristics, and recording artifacts. As a result, the reported performance may differ from that observed with naturally recorded speech.

Second, the evaluation is conducted using a single ASR model without dialect-specific fine-tuning or noise-aware adaptation. Although this choice reflects a realistic use case involving off-the-shelf ASR systems, it limits the generalizability of the findings to other architectures or training paradigms. Different ASR models may exhibit varying degrees of robustness to dialectal variation and noise, and the observed trends may not fully transfer across systems.

Finally, performance evaluation relies primarily on Word Error Rate (WER), which has known limitations, particularly for morphologically rich languages such as Arabic. WER penalizes all word-level mismatches equally and does not account for partial correctness, phonetic similarity, or semantic equivalence. Consequently, some recognition outputs that are perceptually close to the reference may still result in high WER values.

B. Future Work

Several extensions can be explored to address the aforementioned limitations. Future work may incorporate real recorded speech data collected from native speakers across different dialects to better reflect real-world acoustic and linguistic variability. Combining synthesized and real speech could further improve evaluation realism while retaining experimental control.

In addition to WER, alternative evaluation metrics such as Character Error Rate (CER) may provide more fine-grained insight into ASR performance, particularly in cases where word segmentation errors dominate. Furthermore, the use of fuzzy or phonetic-aware evaluation metrics, which account for phonetic similarity rather than exact word matching, could offer a more perceptually meaningful assessment of recognition quality for Arabic.

Finally, future studies may investigate noise-robust and dialect-adaptive ASR models, including systems trained with explicit noise augmentation or dialectal supervision. Such models have the potential to significantly improve recognition accuracy under adverse conditions and reduce the performance gap observed across dialects.

VIII. Coonclusion

This work presented a systematic evaluation of Arabic automatic speech recognition performance under varying noise conditions and across multiple dialects. Using a controlled experimental framework, speech samples were generated for five Arabic dialects using a text-to-speech system, corrupted with different noise types at multiple signal-to-noise ratio (SNR) levels, and subsequently transcribed using an automatic speech recognition model. Performance was assessed by

comparing ASR outputs with the original text using Word Error Rate (WER).

The results demonstrate that ASR performance is strongly influenced by both **dialectal variation** and **acoustic noise**. Modern Standard Arabic consistently achieved the lowest error rates, while Moroccan Arabic exhibited the highest WER under both clean and noisy conditions. Noise type was shown to play a critical role, with white noise causing severe degradation at low SNR levels, whereas pink and brown noise had relatively minor impact. Babble and street noise produced moderate, SNR-dependent degradation, with performance gradually approaching clean-speech levels as SNR increased. Across all conditions, noise was found to amplify existing dialect-dependent weaknesses rather than altering the relative performance ranking among dialects.

These findings underscore the importance of evaluating ASR systems beyond clean and standardized speech conditions. They highlight the need for improved dialectal coverage, noise-aware training strategies, and more robust evaluation frameworks for Arabic speech technologies. Overall, the results emphasize that achieving reliable Arabic ASR in real-world environments requires systems that are both **dialect-aware and noise-robust**, particularly for under-resourced spoken varieties.

IX. References

[1] A. Ali, Y. Zhang, P. Cardinal, N. Dahak, S. Vogel, and J. Glass, "A complete KSU Arabic speech database for automatic speech recognition," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Miami, FL, USA, 2012, pp. 1–6.

[2] A. Ali, S. Vogel, and S. Renals, "Speech recognition challenge in the wild: Arabic MGB-3," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, 2017, pp. 316–322.

[3] M. Najafian, S. Marcel, and S. Bengio, "ADI-17: A speech corpus for Arabic dialect identification," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 2804–2808.

[4] M. Najafian, S. Marcel, and S. Bengio, "ADI-20: Arabic dialect identification in the wild," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 1–5.

[5] N. Halabi, "Arabic speech corpus," 2016. [Online]. Available: https://sourceforge.net/projects/arabic-speech-corpus/

[6] Mozilla Foundation, "Common Voice: A multilingual speech corpus," 2023. [Online]. Available: https://commonvoice.mozilla.org

[7] E. Casanova, J. Weber, C. Shulby, A. Gölge, and M. Müller, "Coqui TTS: A deep learning toolkit for text-to-speech," 2022. [Online]. Available: https://github.com/coqui-ai/TTS

[8] E. Casanova, J. Weber, L. Fernandez, and M. Müller, "XTTS: A multilingual zero-shot text-to-speech model," 2023. [Online]. Available: https://github.com/coqui-ai/TTS

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," arXiv preprint arXiv:2212.04356, 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[10] Z. Huang, G. Keren, Z. Jiang, S. Jain, D. Goss-Grubbs, N. Cheng, F. Abtahi, D. Le, D. Zhang, A. D'Avirro et al., "Text generation with speech synthesis for ASR data augmentation," arXiv preprint arXiv:2305.16333, 2023.

[11] A. Varga et al., "The Aurora 2 database and an evaluation methodology for noise robust speech recognition," Speech Commun., vol. 40, no. 1–2, pp. 83–97, Apr. 2003.

[12] A. Waheed, B. Talafha, P. Sullivan, A. Elmadany, and M. Abdul-Mageed, "A Robust Dialect-Aware Arabic Speech Recognition System," 2023. Accessed: Dec. 15, 2025. [Online]. Available: https://aclanthology.org/2023.arabicnlp-1.38.pdf