University of Thi-Qar Journal for Engineering Sciences
Volume 16, Issue 1, June 2026
https://doi.org/10.32792/utq/utjes.16.1.742

SH_ArabicIraqiAccent: A Speech Corpus of Iraqi Arabic Speech in Baghdad's Northern Suburbs

Sarah I. Hasan[1*] Hasan M. Kadhim[2]

Received: 25th December, 2025/Accepted: 14th February 2026/ Published Online: 1st June 2026

Abstract

This paper presents the SH_ArabicIraqiAccent corpus, a newly developed speech corpus that captures spontaneous Iraqi Arabic from the northern suburbs of Baghdad. This corpus comprises spontaneous speech from 47 speakers (28 females aged 9–59 and 19 males aged 20–63) discussing both personal and general community topics relevant to the region. The recordings were made in a controlled environment at a sampling rate of 16 kHz, with 16-bit resolution and a mono channel, to ensure clarity and fidelity. The total duration is approximately 2.5 hours and is professionally recorded with minimal background noise. The corpus addresses the need for dialect-specific resources, as Modern Standard Arabic corpora do not adequately represent regional spoken varieties. The dataset was carefully designed to address key parameters, including speaker diversity (age and gender), phoneme coverage, and recording quality, and supports applications in natural language processing (NLP), automatic speech recognition (ASR), text-to-speech (TTS), and speaker identification. The corpus complements existing resources by focusing on underrepresented Iraqi Arabic dialects and offers detailed acoustic analyses, including fundamental frequency and intensity measurements. This work situates the SH_ArabicIraqiAccent alongside other primary Arabic linguistic resources and highlights its potential to enhance understanding of Arabic dialects and to support model development. The corpus and relevant statistical analyses are publicly available and contribute to Arabic speech and language research, particularly for dialectal NLP tasks.

Keywords— Speech Corpus, Iraqi Accents, Arabic Language, F0

1 Introduction

This article outlines the methodology used to conduct this research. It provides a detailed account of the data collection process, the technical specifications of the datasets, the rationale for selecting the tools and algorithms, the procedures for fundamental frequency tracking, and the statistical metrics used for comprehensive analysis and evaluation. The article presents an Iraqi-accented speech corpus. The researchers generated the corpus and uploaded it to GitHub. The paper describes the characteristics and specifications of that corpus. Technical parameters are also presented, along with the reasons for those choices. The corpus is similar to other Arabic-language corpora and identical to other Arabic-accent corpora for countries with Arabic culture. To that end, the researchers introduced corpora of Modern Standard Arabic, Arabic accents, and Iraqi Arabic.

Sarah I. Hasan
sara1999-10-22@uomustansiriyah.edu.iq
Hasan M. Kadhim
hasanalmgotir@uomustansiriyah.edu.iq
1 Electronic Technology, Technical Instructor Training Institute, Middle Technical University, Baghdad, Iraq.
2 Electrical Engineering Department, Mustansiriyah University, Baghdad, 10047, Iraq.

1.1 Official Modern Standard Arabic (MSA) Language Corpora

The MSA corpora are diverse and available in varying sizes, qualities, and purposes, whether for linguistic research or NLP applications. The main application is the extensive, freely available resources, which include a comprehensive study (48 freely available corpora) compiled from academic sources and research projects. Many of these resources/studies are dedicated to Modern Standard Arabic (or Classical Arabic), excluding various dialects (Abbas et al., 2023; Ibrahim et al., 2015).

Additionally, AraFast is a popular, modern corpus specifically designed for MSA and featuring high-quality content. It has been used for Transformer models, such as question-answer systems, and offers improved performance, especially in large, clean, and segmented versions (Alrayzah et al., 2024).

As a multipurpose corpus, the KACST Arabic Corpus (King Abdulaziz City for Science and Technology) comprises over 700 million words spanning several eras. It is a rich and free resource for exploration via chronological and domain-specific classification tools (Al-Thubaity, 2015). Another multi-purpose resource is the Kenten families (TenTen Corpus), which provides an Arabic version, arTenTen, an extensive web corpus accessible through the Sketch Engine platform (Bordonaba-Plou & Jreis-Navarro, 2023).

For specialized corpora and auxiliary systems, the Quranic Arabic corpus, titled Linguistic Resources for the Quranic Text (77,430 words) and featuring morphological and syntactic analysis, falls within the scope of classical Arabic (Sherif & Ngonga Ngomo, 2015). For special use, there is the Arabic speech corpus for MSA, which includes audio recordings and their detailed transcriptions, for non-commercial use in speech generation and audio processing (El-Khair, 2016). A parallel corpus (paragraphs from UN texts) known as EAPCOUNT, containing millions of words distributed between English and Arabic (Alotaibi, 2016). Corpus catalogs, such as the Masader, a research interface that aggregates over 500 Arabic datasets for language processing and audio text, are handy for researchers and developers (Alyafeai et al., 2022). Tools for Arabic Language Processing, such as libraries and software frameworks like MADAMIRA, Farasa, SAFAR, and CAMeL Tools, support tasks like text segmentation, morphological analysis, syntactic tagging, and others, specifically for Modern Standard Arabic (Darwish et al., 2021).

1.2 Corpus for Arabic accents

The most prominent Arabic dialect corpora widely used in language research and natural language processing are the MADAR Arabic Dialect Corpus and Lexicon, which contain translated parallel sentences in 25 dialects from different Arab cities, as well as in English, French, and MSA (Bouamor et al., 2018). Multi-Dialectal Parallel Corpus, which includes 2,000 linked sentences between Classical Arabic and the following dialects: Egyptian, Tunisian, Jordanian, Palestinian, and Syrian, in addition to English. PADIC is the Parallel Arabic Dialect Corpus, which contains approximately 6,400 sentences from the dialects of the Arabic Maghreb (Algeria, Tunisia, Morocco) and the Levant (Syria, Palestine), with alignment to classical Arabic (Meftouh et al., 2015). Lisan Corpora Multiple Arabic Dialects, which includes the dialects of Yemeni (about 1.05 million Yemeni words), Iraqi (approximately 45,000 Iraqi words), Sudanese (approximately 52,000 Sudanese words), and Libyan (approximately 51,000 Libyan words), with detailed morphological and syntactic instructions (Jarrar et al., 2023). Gumar Corpus Arabian Gulf Dialect, which is the most extensive corpus for the Arabic Gulf dialect, containing approximately 110 million words from forum novels by writers from the Arabic Gulf, with subdialects identified and morphological instructions provided (Khalifa et al., 2016). The Multi-Genre Dialect Corpus (Zaidan & Callison-Burch) covers Egyptian, Gulf, Levantine, Maghrebi, and Iraqi dialects. Includes online newspaper comments and tweets, with human input via Mechanical Turk (Zaidan & Callison-Burch, 2014). The Currasat Corpus at Birzeit University comprises Arabic dialect corpora developed by the Sina Institute for NLP in Palestine. It is a crucial resource for NLP research in the Arabic-speaking world. The Curras Baladi (Lebanese dialect) contains approximately 9,600 words, morphologically and syntactically labeled to represent Levantine dialects (Al-Haff et al., 2022). The above corpora have been summarized in Table 1.

Table 1*:* Summarizes several corpora of Arabic-accented languages. The references for the table details are provided in the subsections 1.1, 1.2, and 1.3

Corpus	Dialects Covered	Type / Notes
MADAR	25 dialects from Arab cities	Parallel sentences with MSA, English, and French
Multi-Dialectal Parallel Corpus (LREC'14)	Egyptian, Tunisian, Jordanian, Palestinian, Syrian	Sentences aligned with MSA, dialects, and English
PADIC	Algeria, Tunisia, Morocco, the Levant	Aligned with MSA
Lisan Corpora	Yemeni, Iraqi, Sudanese, Libyan	Content from social media, morphologically annotated
Gumar Corpus	Gulf dialect	Texts from forum novels, large-scale data
Zaidan & Callison-Burch Corpus	Egyptian, Levantine, Maghrebi, Iraqi Gulf,	Newspaper comments and Twitter, human annotations
Curras + Baladi	Lebanese dialect (general Levantine)	Parallel sentences with MSA, English, and French

1.3 Iraqi-accents corpus

There are several corpora dedicated to the Iraqi dialect (Arabic Iraqi Dialect Corpora), some of which are independent, and others are part of projects that cover multiple Arabic dialects. The most important of these is the Lisan Iraqi Corpus, part of the Lisan Corpora project (Birzeit University), which contains approximately 45,000 words in the Iraqi dialect, drawn from social media and conversations, and is accompanied by precise morphological and syntactic annotations (Jarrar et al., 2023). The MADAR Corpus is a large-scale project that covers 25 Arabic dialects from various cities (including Baghdad, Iraq's capital), as well as Modern Standard Arabic, English, and French. Includes parallel sentences, facilitating comparison between dialects (Bouamor et al., 2018). The PADIC Covers dialects from the Maghreb and Levant, as well as some other dialects, including Iraqi, to a limited extent (Meftouh et al., 2015). Contains texts aligned to MSA. The Zaidan & Callison-Burch Corpus includes dialects of Egyptian, Gulf, Levantine, Maghrebi, and Iraqi Arabic (Zaidan & Callison-Burch, 2014). The text is drawn from comments in online newspapers and on Twitter, dialect-labeled data, other Speech & NLP Resources (including some audio resources for Iraqi speech datasets used in ASR projects), and from university research or funded projects in NLP for Arabic dialects.

2 Main Parameters of the Speech Corpus

To create and record a high-quality speech corpus usable for NLP, ASR, or Text-To-Speech (TTS) research, there is a set of key parameters that must be carefully defined and adjusted (Sachdeva et al., 2022):

2.1 Recording setup of the Corpus

For the recording quality, the sampling rate is typically 16 kHz or 22.05 kHz for speech recognition, and 44.1/48 kHz for high-quality TTS applications. For bit depth: at least 16-bit PCM (24-bit is preferred). The audio channel is typically mono, unless the target is a specialized application. For the equipment, a professional microphone (condenser or dynamic) of suitable quality, a sound card/audio interface to reduce distortion, and an acoustically treated room (to reduce echo and noise). For the recording conditions, constant loudness, reduced background noise, and a constant distance between the speaker and the microphone are required.

2.2 Corpus design

According to the stated goal, these include speech recognition, TTS, speaker recognition, and accent recognition, among others. For types of transcripts, sentences, isolated words, natural dialogue, and free syllables. For

Linguistic distribution, Phoneme coverage, syllables, syntactic, and semantic diversity. For the number of speakers: (single/multiple), with gender (male/female), age, and accents specified. And for data size, how many hours of recording are required?

2.3 Metadata of the corpus

Include the speaker identification, the age, gender, accent, and education level, the recording date, transcript type, and the sentence descriptions (language category, emotion, etc., if necessary). In addition to that: Number of clips, hours recorded, and release version. To protect the rights, the license type can be added.

2.4 Post-processing of the corpus

Include segmentation (trimming sentences/ words), annotation (such as text transcripts and phonetic transcription, if necessary), and forced alignment. Data cleaning by removing errors and poor-quality recordings. File formats are WAV (preferably uncompressed) for audio, and texts and tags are in CSV, JSON/ XML formats. The storage and the pickups with the organization are considered.

2.5 Corpus management of the corpus

Includes the standardized file naming system, the comprehensive documentation of the collection and recording process, and the usage rights and licenses. For annotation, transcribe, align timestamps, and add prosodic/emotional markers. The distribution is considered.

3 Data Preparation and Collection

This study relies on the analysis of vocal corpora from four distinct groups: a custom-collected dataset and three pre-existing corpora (Meyer & Nelson, 2020) (Table 2).

- Iraqi Corpus: This dataset was collected locally using Audacity and a high-quality mono microphone. It includes 19 voice recordings from males aged 20 to 63 and 28 recordings from females aged 9 to 59. The recorded tapes were spontaneous and natural, free of background noise and minimal acoustic artifacts, such as echo or interference.
- Saudi Corpus: A pre-existing dataset of voice recordings in the Saudi dialect was utilized.
- Modern Standard Arabic Corpus: A pre-existing dataset of speech in Modern Standard Arabic was used.
- TIMIT Corpus: The standard TIMIT corpus, a globally recognized dataset for speech processing research, was adopted.

Table 2*:* Comparisons between the number of files and the length of the total files

Female				
	KSA	**TIMIT**	**Arabic**	**Iraqi**
Number of files	10	170	148	28
Length of total files (minutes)	11:18	9:38	32:48	90:00
Size (MB)	20.7	17.6	60.0	176
Male				
	KSA	**TIMIT**	**Arabic**	**Iraqi**
Number of files	153	225	140	19
Length of total files (minutes)	44:34	14:59	34:15	60:00
Size (MB)	81.6	27.4	62.7	110

4 SH_ArabicIraqiAccent Corpus

The researcher generated a speech corpus to assist users and researchers focused on Arabic with an Iraqi accent. The suburbs around Baghdad exhibit the accent of Iraq's capital. The speakers of the corpus are 28 females and 19 males. The ages of the females range from 9 to 59 years, and the ages of the males range from 20 to 63

years. In their speeches, the speakers discussed various general topics related to the community in the region surrounding Baghdad, the capital. Some of these subjects are personal, while others concern general community issues.

Technically, the sampling rate of the recorded speech is 16 kHz (i.e., the bandwidth of the speech is limited to 8 kHz according to the Nyquist-Shannon rate). Each sample of that speech has 16-bit resolution (i.e., approximately 64 k quantization levels). The speech is recorded in mono (i.e., one channel) (Rabiner & Schafer, 2007). The duration of the female speech was 1:30 (i.e., one and a half hours), while the duration of the male speech was about 1:00 (one hour). The recorded speech was professional, with no background noise and minimal other acoustic effects, such as echo and interference from nearby machines and devices. The speakers were talking spontaneously about their speech.

To implement this task, the Audacity and Speech Analyzer players were used. The software was beneficial in removing many silent periods from the speakers' speeches in the corpus. The speech is stored as *.wav* files on the PC (Madisetti, 2018). The files have been uploaded to the corpus repository on GitHub for the researchers. The files have meaningful, sequential names: F001.wav to F028.wav and *M001.wav* to *M019.wav*. Other essential files have also been uploaded to the repository. Monthly, the researchers intend to update and improve the repository's content. Further details about the corpus website are available on GitHub (Hasan & Kadhim, 2024,).

The researchers intentionally uploaded these enriched files to support research hypotheses for both Iraqi and non-Iraqi researchers focusing on the nature of Iraqi accents across different languages in Iraq, such as Arabic, Kurdish, and others. The researchers in this paper are prepared to support any of them. The Corpus maintains datasets for different Arabic language accents, such as Egyptian and Saudi, as well as for other countries where Arabic is the official language in governmental institutions. The researchers have chosen three females and three males to describe the following parameters and characteristics of speech:

4.1 Statistical Analysis of F0

The F0 of the corpus's overall female and overall male speech has been tracked and calculated using the following well-known, reliable Pitch Detection Algorithms (PDA): RAPT, PRAAT, SHR, and YIN. These results have been detailed in our article (Hasan & Kadhim, 2025). Briefly, the average values of these analyses are tabulated in Table 3.

According to the analysis, the Mean, Standard Deviation, Variance, and Mode of F0 are higher for females than for males. Because the YIN algorithm is designed for noisy background speech, its analysis results differ from those of the other three algorithms. In the article, that analysis is compared with analyses of the Arabic Saudi accent, the official Arabic language, and the TIMIT English language corpus.

4.2 Different Ages

4.2.1 Time, Spectrum, and Spectrogram

For F, we selected three ages: 9, 25, and 53 years. For M, we selected the following age groups: 9, 26, and 52 years. Individually (each age alone), illustrated in the following Figures 1-3 and Table 4. All the speeches of: Our corpus, the Saudi corpus, the MSA speech corpus, and the TIMIT corpus are in Table 5. The spectrums of our corpus "SH_ArabicIraqiAccent" are described in the third row of Figures 1-3.

Table 3: Statistical analysis of F0 for our SH_ArabicIraqiAccent corpus: Females (F) and Males (M) speakers. The PART, PRAAPT, SHR, and YIN algorithms tracked and then estimated the F0 frequency. More details are available in our article (Hasan & Kadhim, 2025).

	RAPT		PRAAPT		SHR		YIN	
	F	**M**	**F**	**M**	**F**	**M**	**F**	**M**
Mean (Hz)	198	131	211	133	221	138	207	145
Standard Deviation (Hz)	63	52	56	41	62	42	93	97
Variance (Hz^2)	3935	2674	3087	1654	3843	1728	8598	9439
Mode (Hz)	186	105	185	103	186	102	50	101
Skew	0.65	2.7	1.3	3.0	0.86	0.82	1.6	3.0
Kurtosis	5.6	16	7.3	23	3.2	3.1	8.1	13

4.2.2 Raw Pitch and Auto Pitch

Raw Pitch denotes the direct, unprocessed measurement of the fundamental frequency (F0) of the human voice from the acoustic signal. It is a physical measurement, typically expressed in Hertz (Hz), of the rate at which the vocal folds vibrate. Auto Pitch is a processed or corrected version of the pitch, adjusted using software to align notes with a specific musical scale or key. The fourth row of Figures 1-3, and Table 4 contain the F0 illustrations.

4.2.3 Intensity and Linear magnitude

The intensity of speech is the physical measurement of sound energy, typically ranging from 30 to 60 dB for an everyday conversation and varying significantly for other vocal efforts. While physical measurements are often expressed in decibels (dB), perceived intensity (loudness) is a subjective experience. A typical conversation is considered most intelligible when it is about 10-15 dB above the ambient noise level. The "linear magnitude" of speech refers to the direct, instantaneous physical measurement of the sound wave's displacement or pressure, typically represented in units like pascals (Pa) or as a unitless value in a digital system. This contrasts with the more common logarithmic representation (decibels, dB) used in acoustics and auditory perception. The intensity is shown in the fifth row of all figures (Figures 1-3) and Table 4, individually and for all speeches.

4.2.4 Melogram and Semitone

A melogram (or melograph) is a visual representation of musical sound, typically used in ethnomusicology and music analysis to study the nuances of performance, especially in vocal music. Semitones measure Melogram. The semitone, also called a half step or half tone, is the smallest interval between two adjacent notes in a standard 12-tone scale. The Melogram is shown in the sixth row of all figures (Figures 1-3), and Table 4 individually and for all speeches.

4.3 Different Corpora

Our Corpus has been compared with the Arabic MSA and TIMIT standard corpora, as well as the Saudi Arabian accent corpus. Table 5 illustrates the following parameters for that comparison: the time-domain waveform; raw F0 (40 to 500 Hz); F0 (50 to 500 Hz); percentage intensity; and melogram from minimum to maximum semitones. The table for female and male speakers across all corpora.

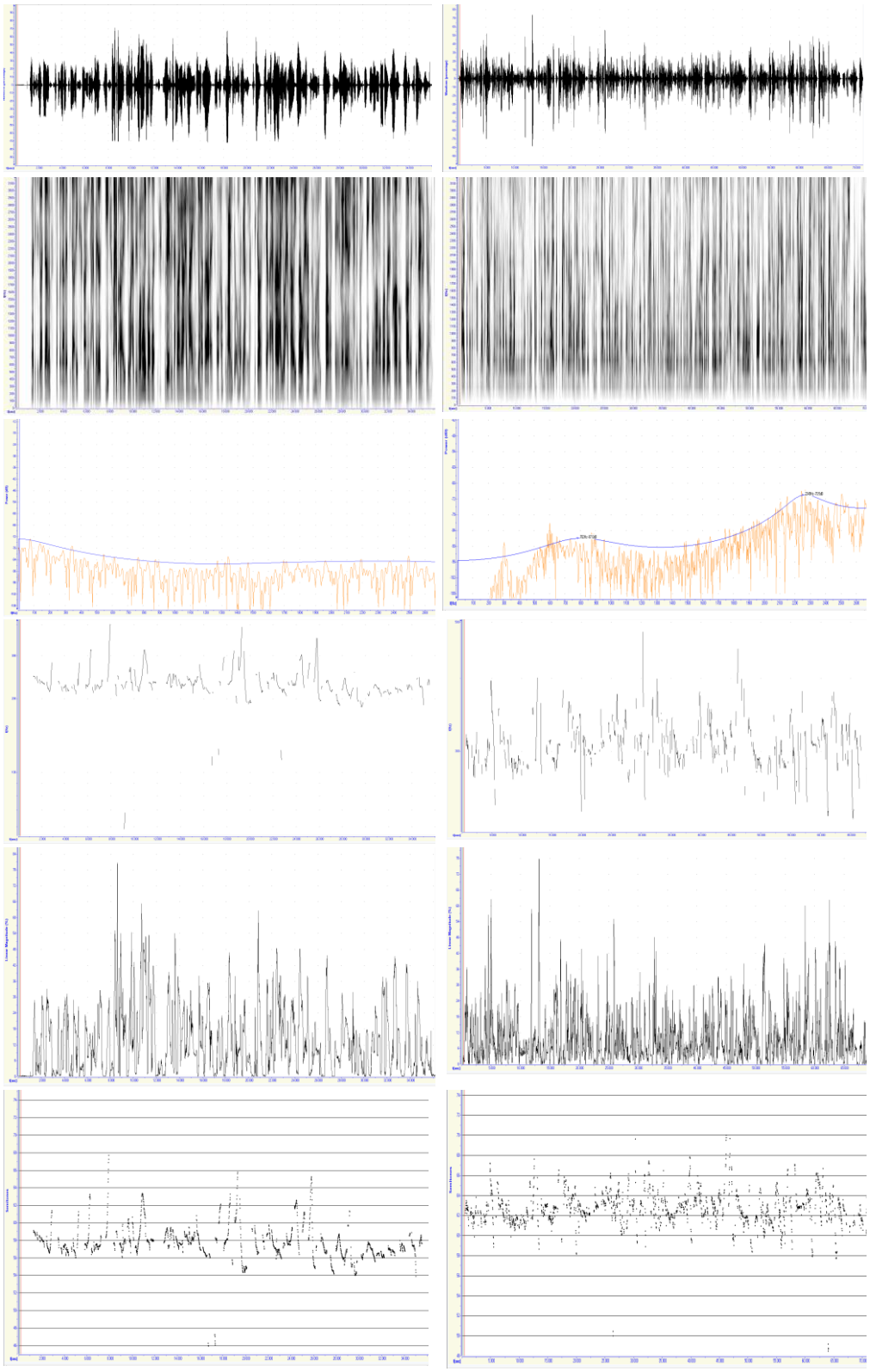

Figure 1: Our corpus. LHS is a 9-year-old female: (1st row) Time domain 36 s; (2nd row) Linear-scale spectrogram; (3rd row) Spectrum -108 to -37 dB; (4th row) F0 200 to 340 Hz; (5th row) Intensity per 80%; and (6th row) Melogram 54 to 64 semitones. RHS is a 9-year-old male: (1st) Time domain 71 s; (2nd) Linear-scale spectrogram; (3rd) Spectrum -108 to -70.5 dB; (4th) F0 around 300 Hz; (5th) Intensity per 74%; and (6th) Melogram 58 to 70 semitones.

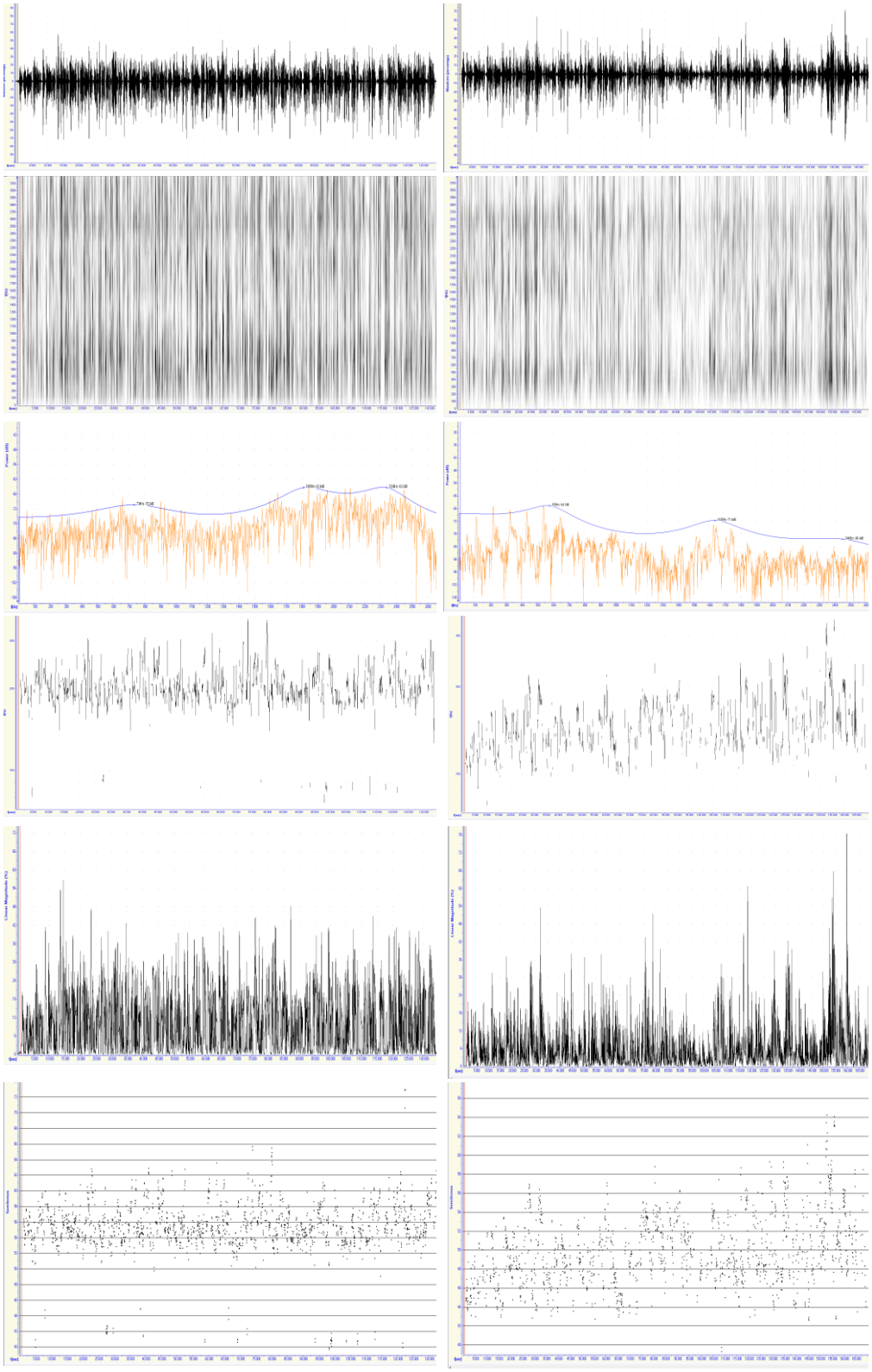

Figure 2: Our corpus. LHS is a 25-year-old female: (1st row) Time domain 135 s; (2nd row) Spectrogram; (3rd row) Spectrum -108 to -63.3 dB; (4th row) F0 150 to 350 Hz; (5th row) Intensity per 57%; and (6th row) Melogram 50 to 66 semitones. RHS is a 26-year-old male: (1st) Time domain 171 s; (2nd) Spectrogram; (3rd) Spectrum -108 to -64.7 dB; (4th) F0 100 to 300 Hz; (5th) Intensity per 66%; and (6th) Melogram from 42 to 64 semitones.

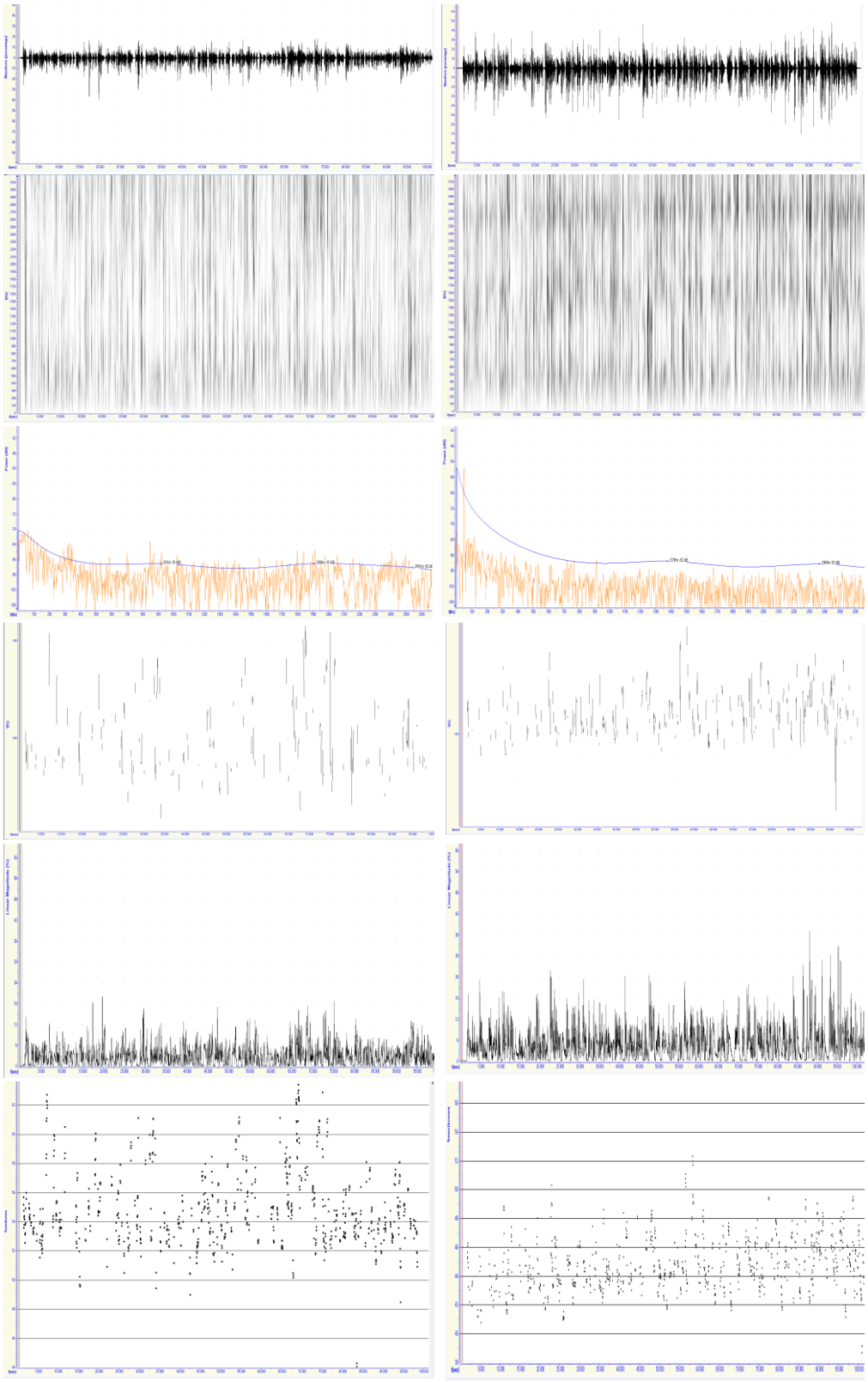

Figure 3. Our corpus. LHS is a 53-year-old female: (1st row) Time domain 101 s; (2nd row) Spectrogram; (3rd row) Spectrum -108 to -79 dB; (4th row) F0 150 to 300 Hz; (5th row) Intensity per 18%; and (6th row) Melogram 49 to 63 semitones. RHS is a 52-year-old male: (1st) Time domain 103 s; (2nd) Spectrogram; (3rd) Spectrum -108 to -56 dB; (4th) F0 90 to 150 Hz; (5th) Intensity per 32%; and (6th) Melogram from 40 to 52 semitones.

Table 4: Summarizes the acoustical parameters of the researcher's corpus speech.

	9-year-old		around 25-year-old		around 50-year-old	
	F	**M**	**F**	**M**	**F**	**M**
Time domain (sec)	36	71	135	171	101	103
Spectrum(dB)	-108	-108	-108	-108	-108	-108
	to -37	to -70.5	to -63.3	to -64.7	to -79	to -56
F0 (Hz)	200	around 300	150	100	150	9
	to 340		to 350	to 300	to 300	to 125
Intensity (%)	80	74	57	66	18	32
Melogram (semitones)	64	58	50	42	49	40
	to 54	to 70	to 66	to 64	to 63	to 52

Table 5: Acoustical parameters comparisons between the researcher's corpus, the Saudi accents corpus, the MAS corpus, and the TIMIT corpus speech.

	Our corpus		Saudi corpus		MSA corpus		TIMIT corpus	
gender	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**
Time (minutes)	90:00	60:00	11:18	44:35	32:48	34:15	9:37	14:59
Raw F0(Hz)	40	40	40	40	40	40	170	80
	to 500	to 500	to 500	to 500	to 500	to 500	to 300	to 225
F0 (Hz)	70	50	100	50	50	50	130	100
	to 500	to 300	to 500	to 500	to 500	to 400	to 325	to 200
Intensity (%)	98	84	72	75	78	100	72	36
Melogram	40	30	40	28	30	30	46	40
(semitones)	to 84	to 70	to 72	to 71	to 72	to 71	to 66	to 58

5 Conclusions

This paper introduces the SH_ArabicIraqiAccent corpus, a novel speech dataset designed to capture the characteristics of the Iraqi Arabic accent as spoken in the northern suburbs of Baghdad. The corpus aims to support research and applications in NLP, ASR, and text-to-speech systems that focus on Arabic dialects, particularly Iraqi Arabic.

The corpus was locally collected using professional recording equipment, including a high-quality mono microphone and Audacity, to ensure clean, spontaneous, and natural speech, free of background noise and minimal acoustic artifacts.

The corpus comprises speech from 47 speakers, 28 female and 19 male, aged 9 to 59 and 20 to 63, respectively. The total recorded speech amounts to approximately 2.5 hours, with female speakers contributing approximately 1.5 hours and male speakers approximately 1 hour. The speech covers a range of topics relevant to community and individual issues in the region.

Technically, the audio was sampled at 16 kHz with 16-bit resolution in mono, adhering to standard digital speech processing practices to balance quality and data size efficiently. The paper details the corpus design, metadata on speakers (e.g., age, gender, and accent), and post-processing steps, including segmentation, annotation, and error correction.

The study situates this corpus in the broader landscape of Arabic linguistic resources. It contrasts the SH_ArabicIraqiAccent corpus with other Arabic resources, including Modern Standard Arabic corpora, the MADAR dialect corpus, and parallel corpora such as PADIC. The authors highlight the importance of dialectal corpora for accurate Arabic-language understanding and model training, as Modern Standard Arabic alone does not capture regional spoken varieties.

The paper references several prior Arabic and Iraqi speech and text corpora and discusses their scope, limitations, and applicability. The SH_ArabicIraqiAccent corpus complements these by focusing on spontaneous spoken Arabic from a relatively underrepresented dialectal subset.

For future works, the researchers suggest: Build corpora for each Iraqi language and each Iraqi accent, such as Kurdish and Turkish, and distinct accents in the south and west; analyze the speech of these corpora, and then compare these analyses with our analysis; compare the speech of our corpus with the corpus of accents in other Arabic-speaking countries such as Egypt and Syria; enlarge our corpus to contain other ages and persons who are not available in our corpus.

Acknowledgments The authors thank the Electrical Engineering Department, College of Engineering, Mustansiriyah University, and the Electronic Technologies Department, Technical Instructor Training Institute, Middle Technical University, Baghdad, Iraq.

References

Abbas, N. F., Qasim, T. A., & Jasim, H. A.-S. (2023). Request Constructions in Classical Arabic versus Modern Arabic. *Journal of Ethnic and Cultural Studies*, *10*(5), 1-15. https://doi.org/https://doi.org/10.29333/ejecs/1598

Al-Haff, K., Jarrar, M., Hammouda, T., & Zaraket, F. (2022). Curras+ baladi: Towards a levantine corpus. Proceedings of the Thirteenth Language Resources and Evaluation Conference, DOI: https://doi.org/10.48550/arXiv.2205.09692,

Al-Thubaity, A. O. (2015). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, *49*(3), 721-751. https://doi.org/https://doi.org/10.1007/s10579-014-9284-1

Alotaibi, H. M. (2016). AEPC: Designing an Arabic/English parallel corpus. *Research in Corpus Linguistics*, 1-7. https://doi.org/https://ricl.aelinco.es/index.php/ricl/article/view/36/22

Alrayzah, A., Alsolami, F., & Saleh, M. (2024). AraFast: Developing and evaluating a comprehensive modern standard arabic corpus for enhanced natural language processing. *Applied Sciences*, *14*(12), 5294. https://doi.org/https://doi.org/10.3390/app14125294

Alyafeai, Z., Masoud, M., Ghaleb, M., & Al-shaibani, M. S. (2022). Masader: Metadata sourcing for arabic text and speech data resources. Proceedings of the Thirteenth Language Resources and Evaluation Conference, DOI: https://doi.org/10.48550/arXiv.2110.06744,

Bordonaba-Plou, D., & Jreis-Navarro, L. M. (2023). Linguistic Injustice in Multilingual Technologies: The TenTen Corpus Family as a Case Study. In *Multilingual digital humanities* (pp. 129-144). Routledge. https://doi.org/https://doi.org/10.4324/9781003393696

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., & Erdmann, A. (2018). The MADAR Arabic dialect corpus and lexicon. Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), URL: https://aclanthology.org/L18-1535.pdf,

Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H., Al-Natsheh, H. T., Bouamor, H., Bouzoubaa, K., Cavalli-Sforza, V., El-Beltagy, S. R., & El-Hajj, W. (2021). A panoramic survey of natural language processing in the Arab world. *Communications of the ACM*, *64*(4), 72-81. https://doi.org/https://doi.org/10.48550/arXiv.2011.12631

El-Khair, I. A. (2016). Abu el-khair corpus: A modern standard arabic corpus. *International Journal of Recent Trends in Engineering & Research (IJRTER)*, *2*(11), 5-13. https://www.researchgate.net/publication/310321022_Abu_El-Khair_Corpus_A_Modern_Standard_Arabic_Corpus

Hasan, S. I., & Kadhim, H. M. (2024,). *SH_ArabicIraqiAccent Corpus*. https://github.com/SaraEsHassan/SH_ArabicIraqiAccent

Hasan, S. I., & Kadhim, H. M. (2025). Statistical analysis of F0 for Iraqi Arabic accent corpus. IET Conference Proceedings CP959, DOI: https://doi.org/10.1049/icp.2025.4386,

Ibrahim, H. S., Abdou, S. M., & Gheith, M. (2015). Sentiment analysis for modern standard arabic and colloquial. *arXiv preprint arXiv:1505.03105*. https://doi.org/https://doi.org/10.48550/arXiv.1505.03105

Jarrar, M., Zaraket, F. A., Hammouda, T., Alavi, D. M., & Wählisch, M. (2023). Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic dialect corpora with morphological annotations. 2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA), DOI: https://doi.org/10.1109/AICCSA59173.2023.10479250,

Khalifa, S., Habash, N., Abdulrahim, D., & Hassan, S. (2016). A large scale corpus of Gulf Arabic. *arXiv preprint arXiv:1609.02960*. https://doi.org/https://doi.org/10.48550/arXiv.1609.02960

Madisetti, V. (2018). *Video, speech, and audio signal processing and associated standards*. CRC Press. https://doi.org/https://doi.org/10.1201/9781315219820

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., & Smaili, K. (2015). Machine translation experiments on PADIC: A parallel Arabic dialect corpus. Proceedings of the 29th Pacific Asia conference on language, information and computation, URL: https://aclanthology.org/Y15-1004.pdf,

Meyer, C. F., & Nelson, G. (2020). Data collection. *The handbook of English linguistics*, 81-101. https://doi.org/https://doi.org/10.1002/9781119540618.ch6

Rabiner, L. R., & Schafer, R. W. (2007). Introduction to digital speech processing. *Foundations and Trends® in Signal Processing*, *1*(1–2), 1-194. https://doi.org/https://doi.org/10.1561/2000000001

Sachdeva, P., Barreto, R., Bacon, G., Sahn, A., Von Vacano, C., & Kennedy, C. (2022). The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022, URL: https://aclanthology.org/2022.nlperspectives-1.11.pdf,

Sherif, M. A., & Ngonga Ngomo, A.-C. (2015). Semantic quran: A multilingual resource for natural-language processing. *Semantic Web*, *6*(4), 339-345. https://doi.org/https://doi.org/10.3233/SW-140137

Zaidan, O. F., & Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, *40*(1), 171-202. https://doi.org/https://doi.org/10.1162/COLI_a_00169