Saudi ASWAT: A Large-Scale Corpus of Spontaneous Saudi Arabic Speech

Abdullah I. Alharbi, Afrah A. Altamimi, Muneera Alhoshan, Amal Almazrua
Halah Munif Alharbi, Bayan M. Almuqhim, Hawra Aljasim
Abdulrahman Alosaimy, Yahya A. Asiri, Abdullah Alfaifi
King Salman Global Academy for Arabic Language
Riyadh, Saudi Arabia
{aialharbi, a.altamimi, malhoshan,aalmazrua hmuneef, balmuqhim, haljasim,aalosaimy
yasiri, aalfaifi}@ksaa.gov.sa

Abstract

Spontaneous Arabic speech is scarce in current corpora, and it is not well represented. This poses a limitation invisibility of spontaneous Arabic to automatic speech recognition (ASR), speaker diarization, and sociolinguistic research. The Saudi ASWAT project fills a major gap by creating the first nationwide corpus of natural Saudi speech, where data has been recorded and transcribed under a systematic methodology and ecologically valid conditions. The corpus aims to collect 2,500 hours of natural conversations from a diverse range of participants. These has been selected from five major Saudi regional varieties, Najdi (Central), Eastern, Hijazi (Western), Northern, and Southern, covering more than fifty five local varieties. Speech has been recorded by trained fieldworkers using participants own devices to reflect real-life variation. The annotated data incorporate a variety of speaker demographics, regional vocabularies which differ from the standard lexicon, and structured metadata. TF–IDF profiling shows regional differences in a range of performing words. Data also represent balanced age and gender sampling to support studies of intergenerational and sociophonetic variation. Saudi ASWAT provides the most linguistically diverse resources of Saudi Arabia to date. Additionally, it establishes an ethical governed framework for Arabic speech data creation to enable advances in both computational modeling and linguistic research.

Keywords: Arabic speech corpus, Saudi dialects, spontaneous speech;

1. Introduction

Arabic is characterized by diglossia (Ferguson, 1959), a sociolinguistic phenomenon in which two distinct linguistic varieties are used in parallel within Arabic-speaking communities: Modern Standard Arabic (MSA) and regional dialects. This structural duality presents challenges not only for daily communication but also for computational applications.

Much of the progress in automatic speech recognition (ASR) and spoken language understanding has been fueled by the availability of large-scale corpora like LibriSpeech, Switchboard, and CommonVoice. Yet Arabic remains notably limited in spontaneous speech resources, especially when compared to English and other widely studied languages.

Beyond the scarcity of data lies a more fundamental linguistic issue. Arabic is not only diglossic but also highly diverse, with significant dialectal variations across and within countries (Prochazka, 1988). In only Saudi Arabia, spoken Arabic is typically categorized into five broad regional varieties, Najdi (Central), Hijazi (Western), Eastern, Northern, and Southern, each encompassing multiple local speech forms. Based on initial linguistic mapping conducted by our team, we estimate that these regions have about fifty five distinct local varieties. Differing significantly in phonology, morphology, lexicon, syntax, pragmatics, semantics, and prosody. Thus, this diversity highlights the importance of building a speech corpus to reflect Saudi variations for linguistic analysis and language technology development.

Existing Arabic speech corpora, however, more likely reflect MSA or speech from broadcast sources such as scripted television, interviews and news media (Alkanhal et al., 2023; Alharbi et al., 2024). Whereas these sources offer clean acoustic conditions and formal speech structures, they may fail to capture the informal, unscripted dynamics of everyday interaction. As a result, ASR model trained on such data often struggle with dialectal variations, spontaneous disfluencies, informal vocabulary and code-switching. To Address this gap, it requires a corpus that captures the linguistic realities of Saudi Arabia in natural settings, with both demographic and regional breadth.

The Saudi ASWAT project stands for (**Arabic Speech With Annotated Transcripts**) which also means “voices” in arabic (أصوات), has been developed to meet this demand. It is the first large-scale nationwide corpus of spontaneous Saudi Arabic speech, It has been collected under ecologically valid conditions and designed to specifically reflect the country’s rich linguistic and demographic diver-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1595–1602
11–16 May 2026.

sity. Saudi ASWAT prioritize natural conversation, diverse recording contexts and reproducible protocols to ensure both authenticity and methodological reliability. The resulting corpus supports varied research applications, from speech recognition and diarization to dialectometry and sociolinguistics. More broadly, it contributes to national efforts promoting Arabic through data-driven technologies as well as standardized linguistic data practices.

Saudi ASWAT corpus comprise 2,500 hours of spontaneous speech, all of which has been recorded, transcribed and enriched with corresponding metadata. Subsets are currently undergoing iterative quality review to refine annotation consistency and ensure alignment accuracy across records. A sample is available to qualified researchers upon request [1]. To expand accessibility and usability, the dedicated online platform provide tools[2] for linguistic search, visualization, and dialect-aware NLP analysis.

This paper propose several key contributions:

- **Scale:** 2,500 hours of records for spontaneous Arabic speech, outperforming comparable corpora in Arabic and other languages in size and diversity.
- **Dialectal coverage:** Speech data spans not only five major Saudi dialect zones but also covering 55 distinct local varieties and is balanced by gender and age to reflect sociolinguistic variations.
- **Natural conditions:** Recordings has been made by participants using their own devices in everyday environments to ensure the spontaneity and acoustic variability of real-world interaction.
- **Linguistic authenticity:** Data combines disfluencies, regional and different features that are often absent from scripted or broadcast corpora.
- **Structured metadata and standardized annotation:** Each file is linked to speaker level and clip level metadata. Transcriptions follow an adapted systematic schema, ensuring consistency and cross-dialect compatibility.

The remainder of this paper is structured as follows: Section 2 review related work on Arabic speech corpora and spontaneous datasets. Section 3 outlines Saudi ASWAT design and recruitment protocol. Section 4 describes the data collection and annotation pipeline. Section 5 present summary statistics and baseline benchmarks and Section 7 concludes with future direction.

[1]For research access requests, please contact: arai@ksaa.gov.sa

[2]https://falak.ksaa.gov.sa/corpora

2. Related Work

Spontaneous speech corpora have been foundational in advancing ASR, speaker diarization, and linguistic modeling in many widely spoken languages (Furui et al., 2005; Bang et al., 2020; Lima et al., 2025). For English, the Switchboard corpus (260 hours) (Godfrey et al., 1997) and the Fisher English corpus (2,000 hours) (Cieri et al., 2004) are two of the earlier resources. These datasets capture spontaneous telephone conversations between diverse speakers and have enabled robust ASR models and discourse-level analyses. The design of these corpora—structured conversations between unfamiliar speakers with topic prompts—ensures both lexical richness and natural dialogue flow.

In other languages, comparable resources have achieved similar impact. For Mandarin Chinese, the HKUST corpus (200 hours) (Liu and Fung, 2005) and AISHELL (Bu et al., 2017) have become standard benchmarks, capturing natural conversational speech and public lectures. Japanese benefits from the Corpus of Spontaneous Japanese (CSJ), which offers over 600 hours of naturally occurring talks, lectures, and academic presentations (Maekawa, 2003). Spanish is represented by Fisher Spanish (163 hours) (Post and Garofolo, 2010), while corpora such as Buckeye (American English) (Pitt et al., 2002) and Verbmobil (German, Japanese, English) (Wahlster, 2000) focus on sociolinguistic and task-based interaction, respectively. These corpora are united by their emphasis on ecological validity: capturing real speech in context, with attention to dialect, speaker diversity, and spontaneous interaction. These resources are instrumental because large-scale pretraining and fine-tuning in ASR now rely heavily on high-quality data with rich variation. The success of multilingual models like wav2vec 2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2023) in languages such as English and Mandarin owes much to the availability of such corpora. Without them, even powerful self-supervised learning approaches struggle to generalize across dialects and speaking styles.

In the case of Arabic, however, the situation is markedly different. While MSA is used in formal contexts such as news broadcasts and official documents, it is not the native spoken language for any community. Instead, Arabic is characterized by a complex landscape of dialects—regionally, socially, and even individually variable—that are used in everyday communication. Despite this, most existing Arabic speech corpora focus on MSA, with few resources covering dialectal Arabic.

Some progress has been made in developing dialectal corpora. The CALLHOME Egyptian Arabic corpus (Canavan et al., 1997) and the Levan-

tine and Iraqi telephone speech corpora (Makhoul and Zawaydeh, 2005; LDC, 2006b) provide tens of hours of two-sided conversations. The Gulf Arabic CTS dataset (LDC, 2006a) and the Arabic Fisher Levantine (Makhoul and Zawaydeh, 2005) extend this to additional dialects. However, these datasets are limited in scale (typically 20–60 hours), constrained to telephone audio (8 kHz), and often lack detailed speaker or recording context metadata.

A recent effort is the Saudi ASWAT dataset introduced by (Alkanhal et al., 2023), which includes 732 hours of Arabic audio scraped from YouTube and SoundCloud. While effective for pretraining speech representation models (e.g., wav2vec, data2vec), the dataset consists mostly of monologue-style content and lacks supervised annotations such as transcriptions or speaker metadata. Furthermore, only 27% of the dataset reflects Saudi dialects, and the absence of natural dialogue limits its utility for conversational modeling. Another major development was the release of the SADA corpus, Saudi Audio Dataset for Arabic, proposed by (Alharbi et al., 2024). SADA contains 668 hours of high quality audio taken from 57 Saudi television shows across 11 genres (e.g. comedy, drama and documentaries). It was annotated through multi-stage human validation. However, spontaneous speech coverage is constrained by the nature of broadcast media. Television dialogues even when unscripted are usually shaped by production settings and editorial finishing. These factors can introduce a level of formality and acoustic homogeneity that can limit the generalization.

These gaps motivate Saudi ASWAT project to construct the largest and most diverse spontaneous Arabic speech corpus with 2,500 hours of natural conversation; not only across five major regional varieties in Saudi Arabia but also across more varieties. Unlike prior efforts, Saudi ASWAT's design ensures spontaneous, ecologically valid speech. The use of real-life devices introduces acoustic variability reflective of real-world communication settings. Another distinguishing feature of Saudi ASWAT is taht it focus on demographic and linguistic diversity.

3. Corpus Design

The Saudi ASWAT corpus targets spontaneous spoken Arabic across Saudi Arabia major dialectal and sociocultural regions. It spans five geographic zones (Central, Eastern, Western, Northern, and Southern) ensuring balanced representation by region, gender, and age. Each zone has multiple local speech variations, as can be seen in Table 1.

The entire corpus contain 2,500 hours of natural conversation collected over 10 months from a large number of participants and coordinated by 55 trained field recorders. Each one was responsible about 40 to 50 hours of speech recorded in authentic contexts such as homes, workplaces, schools and public spaces. They used different methods to interview participants, engaging them in short, spontaneous conversations about everyday interests. They were instructed to speak with the participants in the local dialect to encourage natural conversation, and were prompted to continue speaking if they paused. They also used wordless cartoon stories, especially with children, to prompt spontaneous responses. All the stories and conversations were inspired by Saudi and local culture to promote the use of more local vocabulary. This design emphasizes ecological validity and sociolinguistic diversity.

Regional Variety	Example Local Varieties
Najdi	Riyadh, Al-Qassim, áa'il
Northern	Al áudūd ash Shamālīyah, Al-Jawf, Tabuk
Southern	Najrān, Asir, Al-Bāáah, Jāzān
Hijazi	Makkah, Al-Madīnah
Eastern	Ash Sharqīyah

Table 1: Main Saudi Regional Varieties and their corresponding Local Varieties

3.1. Participants and Field Recorders

Speakers (participants) were selected to represent five age bands—children (5–10), teens (10–15), youth (15–20), adults (20–40) and older adults (40–60) and were long-term residents of their respective regions. All speakers (or guardians) signed written consent. Fifty five trained recorders, native to their assigned regions, managed participant selections and data capture after completing short training on ethics, equipment calibration as well as transcription standards. Each recorder collected about 40 to 50 hours of audio and also should maintain minimal intervention (<25 % of total talk time).

3.2. Recording Conditions and Content

Recordings were deliberately non studio, conducted in every day environments (homes, classrooms, cafes, offices). Moderate background noise was permitted to preserve naturalness; long silences (>10 s) were trimmed and normalized. Sessions were averaged between 10 and 20 minutes per file. Speech was entirely spontaneous for all dialogic and monologic seesions. They covered broad topical domains including as an examples education, work, culture, entertainment and daily life. This diversity encourages variation in vocabulary, register and pragmatic style. The supervisor ensures that the recording is free of any personal

information about the participants or others, and also prevents any discussion involving abuse or hate speech. All participants or their guardians also sign consent forms agreeing to participate, with the understanding that their personal information will not be shared, either in the recording or in any metadata that identifies them.

3.3. File Format and Metadata

Audio was stored as MP3 extension (44.1 kHz) or WAV extension (16 kHz) using the standardized convention as follows: <Region>-<RecorderID>-<SpeakerIDs>-<Date> (e.g., E-R13-MY3FO4-17012023.mp3). Two metadata layers attached with each file:

- **Person metadata:** demographic as well as linguistic attributes.
- **Clip metadata:** topic, setting, date and conversational type.

  Metadata are kept in CSV and JSON formats with unique identifiers linking all audio, transcripts and demographic data.

The project follows international research ethics standards and national data governance regulations. Personal identifiers are excluded and public releases will be anonymized. Each participant consented to the use of their recordings for research and educational purposes, ensuring lawful reuse and long term management under regulated access.

4. Recording and Annotation Protocol

4.1. Recording Workflow

The entire process was managed through a secure web based corpus platform. Recorders should start with registering participants, then uploaded files and tracked monthly progress throgh personal dashboard. Each recording was automatically validated for minimum duration ($\geq$60 s), maximum silence ($\leq$10 s) and volume stability ($\pm$3 dB RMS) before finally acceptance.

4.2. Transcription and Review

Transcription was done using a customized web-based transcription platform. It was built with Arabic interfaces designed to support the specific requirements of Arabic speech annotators. The tool provides right-to-left editing, time-aligned segmentation and multi-speaker management. All this within a browser-based environment that allows both online and offline operation (see Figure 1).

The transcription interface present two complementary display modes. The first is a table based view, in which each row represents a time-aligned segment of audio file. Annotators can edit transcriptions easily, assign or modify speaker labels as well as adjust text and audio alignment directly within the interface. It also has filtering options by speaker, facilitating quality control and consistency checks across multi-participant conversations. The second mode provides a continuous text display to enable annotators to review the transcript holistically and verify discourse coherence across turns.

To ensure transcriptions efficiency and precision, the tool supports a set of keyboard shortcuts for rapid navigation, playback control and the timestamp insertion. Annotators can play the audio by pressing the Enter key, also record precise temporal markers where transcription should begin or pause. The system then allow entering the corresponding text segment at that timestamp, to ensure accurate temporal alignment. An auto-save feature continuously saves progress to minimize the risk of data loss during extended transcription sessions.

Recorders first reviewed their own clips for clarity, also trimming background noise and extended pauses before transcriptions. They then can produce initial transcripts shortly after recording to maintain contextual and situational accuracy . Each file subsequently underwent linguistic review by trained annotators, who can check transcriptions accuracy, spelling consistency and meeting teh given criteria.

Overall, the tool design emphasizes efficiency, accuracy and accessibility for native Arabic transcribers. this is to ensure the resulting annotations capture the spontaneous and dialectally rich characteristics of Saudi Arabic speech with fine temporal and linguistic granularity.

The annotation guidelines adapt the CODA conventions for dialectal Arabic propsed by (Habash et al., 2012). It also was modified for Saudi phonological and lexical phenomena. Transcribers recorded dialectal forms, marked non-lexical events (e.g., laughter, hesitation) and aligned text to $\leq$10 s audio segments. After the review, the entire materials were integrated into a unified corpus structure and analyzed using internal tools.

4.3. Quality Control

Quality assurance combined the automated and manual checks. Automated validation included SNR > 20 dB, silence trimming, clipping detection and metadata completeness. Manual review covered $\approx 10\%$ % of files monthly to ensure transcriptions and labeling accuracy. Non-compliant items were returned to the recorder for correction.

The corpus design supports future multi layer annotations including disfluency, code switching,

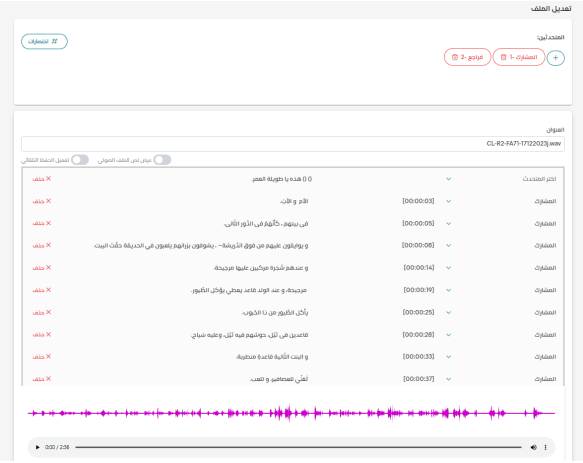

Figure 1: Saudi ASWAT online transcriptions platform showing the table based view for multi speaker annotations and time aligned editing. Each row represents a speech segment with options for alignment adjustment plus speaker labeling and text filtering.

overlap and paralinguistic cues. All corpus assets (audio, transcripts and metadata) are stored in a secure national linguistic data infrastructure with version control and persistent identifiers. Access to the corpus is provided upon request for academic and non-commercial research purposes. Approved access is granted under a research-use agreement that prohibits redistribution, commercial use, or attempts to identify participants, ensuring compliance with ethical and data protection requirements. This ensures data integrity , ethical compliance , and reproducibility for future Arabic-speech research.

4.4. Ethical Considerations

The Saudi ASWAT project follows established international research ethics standards for human-participant data collection. Participation in the recordings was entirely voluntary. All participants were informed about the purpose of the project, the type of data being collected, and the intended research use of the recordings before giving written consent.

For participants under the age of 18, written consent was obtained from a parent or legal guardian prior to participation. In addition, minors were informed in age-appropriate language about the recording activity and could refuse to participate at any time.

Participants were informed of their right to withdraw from the study at any stage. In such cases, their recordings and associated metadata can be permanently deleted from the corpus upon request.

To protect privacy, the recording protocol explicitly prohibited the collection of personally identifiable information during conversations. Personal identifiers are not stored in the released metadata. All public versions of the dataset will be anonymized and distributed under controlled access conditions.

The corpus infrastructure complies with national data governance regulations, and all data are stored in a secure institutional environment with restricted access and audit logging. Any future public releases will only include recordings that meet anonymization and consent requirements.

5. Results and Discussion

5.1. Geographic Distribution

The Saudi ASWAT corpus achieves broad national coverage, with recordings collected from thirteen regions of Saudi Arabia (Ar Riyād, Makkah al-Mukarramah, Al-Madīnah al-Munawwarah, Ash Sharqīyah, Al-Qasīm, Asīr, Tabūk, Jāzān, Najrān, Al-Jawf, Al-áudūd ash Shamāliyah, Al-Bāáah and áā'il) , As illustrated in Figure 2, the largest portions of speech taken from Asīr, Makkah al-Mukarramah, and Ar Riyād, followed by Tabūk, Al-Madīnah al-Munawwarah, and Jāzān while recordings from Al-Bāáah and áā'il remains relatively limited. Such regions will be prioritized for enrichment in the upcoming collection phases to ensure better and complete national representation.

The higher representation of Asīr and Makkah reflects are not demographic weight but the exceptional sociolinguistic complexity of these regions—characterized by intricate tribal geographies, rugged topography in addition to long standing urban rural dialect contact . Crucially, these areas have been the focus of extensive prior linguistic documentation and are among the most studied dialect zones in the Arab world (Holes, 2006). The global scholarly visibility and the relative clarity of internal dialect boundaries, established through decades of fieldwork and comparative analysis, have enabled the reliable identification and labeling of local speech varieties within the Saudi ASWAT framework.

This regional local coverage captures the main dialect zones—Najdi, Hijazi, Eastern, Northern, and Southern—and their internal sub-varieties.

5.2. TF–IDF-Based Lexical Profiling

To examine regional lexical distinctiveness , a Term Frequency-Inverse Document Frequency (TF-IDF) analysis was conducted. Each main local varieties has been treated as a separate sub-corpus. The resulting visualizations , as can be seen in Figure 3, show the most distinctive lexical items per main local variety. This emphasize words that occur frequently within a given region but rarely elsewhere.

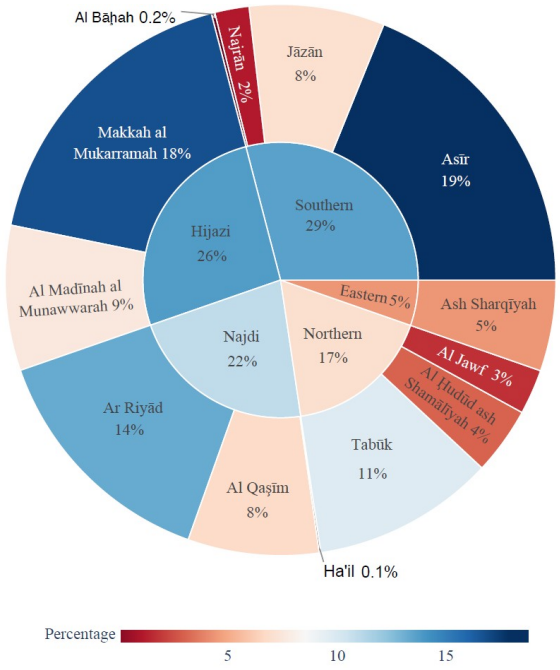

Figure 2: Regional distribution of speech recordings across thirteen regions of Saudi Arabia.

Clear regional signatures emerge from the analysis. Central regions such as Riyad and Al-Qasīm exhibit fundemental Najdi discourse markers and habitual verbs. Makkah and Madinah shows Hijazi lexical borrowings influenced by cultural contact, Ash Sharqīyah reflects Gulf-Arabic lexical features. Southern regions such as Asīr, Jāzān and Najrān present locally distinctive vocabulary and phonological reductions. Finally northern areas including Tabūk, Al-Jawf, Al-áudūd ash Shamāliyah and áā'il show features associated with the Northern continuum. It also show lexical influence from neighboring Levantine varieties. The speech of áā'il in particular demonstrates transitional characteristics between Najdi and Northern dialects which reflect both central Saudi and northern linguistic traits.

It is worth mentioning that TF-IDF weighting prove effective for isolating dialect-specific lexemes and establishing a quantitative foundation for dialect identification, clustering, and region-aware ASR model development. More studies in this direction can be potential directions for further research based on deeper data analysis.

5.3. Speaker Demographics

The demographic composition of the Saudi ASWAT corpus can be seen illustrated in Figure 4. Female speakers slightly outnumber males which providing near equal gender representations. Age groups range from early childhood up to senior adulthood, with the majority of participants belonging to the 21–40 and 41–60 year ranges which may represent the most linguistically active segment of the population for such project. The inclusion of younger speakers introduces inter-generational depth to enable comparative studies of apparent time variation and age related linguistic change.

6. Copyrights

The Language Resources and Evaluation Conference (LREC) Proceedings are published by the European Language Resources Association (ELRA). They are available online from the conference website.

ELRA's policy is to acquire copyright for all LREC contributions. In assigning your copyright, you are not forfeiting your right to use your contribution elsewhere. This you may do without seeking permission and is subject only to normal acknowledgment to the LREC proceedings. The LREC Proceedings are licensed under CC-BY-NC, the Creative Commons Attribution-Non-Commercial 4.0 International License.

7. Conclusion

The Saudi ASWAT corpus makes a major contribution to Arabic speech technology and linguistic research. Its 2,500 hours of natural Saudi Arabic, spoken in a spontaneous, colloquial style and collected under realistic conditions, tackles long-standing gaps in the availability of with ecologically valid data for dialect-rich Media Arabic and other forms. Across labs on five major dialect countries and about 55 local varieties, the corpus captures fine-grained variation. It is a resource suiting applications such as ASR and speaker diarization to dialectology and sociolinguistics.

In marked contrast to previous corpora whether dominated by prepared speech for broadcast or otherwise, the speech material collected in Saudi ASWAT is based on natural conversation. Its acoustic environments are numerous and varied, the demographic balance reflects the language community of the local variates. The inclusion of detailed metadata plus standardized and dialect-aware annotation protocols enhance the utility for both computational and descriptive linguistic tasks.

To facilitate access, a dedicated platform provides tools for search, visualization as well as linguistic data analysis. A sample of the corpus is available to qualified researchers upon request mentioned eariler and more data can be invistaged through a customized platform for computational linguistic researchers.

In future work we will concentrate on a staged quality review and regional enhancement, especially for areas now underrepresented in the data set such as Hail and Al-Bahah. The platform will

Figure 3: TF–IDF-based lexical profiling for a thirteen local varieties of Saudi Arabia which highlight regionally distinctive lexical items.

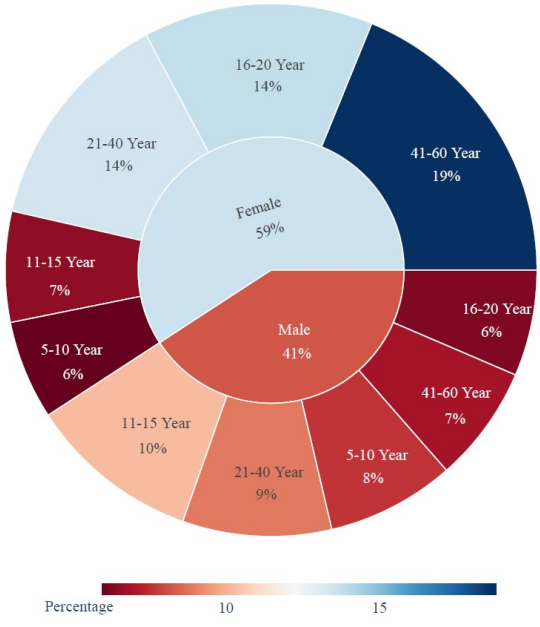

Figure 4: Distribution of participants by age and gender within the Saudi ASWAT corpus. The inner ring represents gender while the outer ring shows age-group proportions within each gender.

include more features, tools that consortium members can use, so building Saudi ASWAT into larger Arabic NLP and ASR pipelines. The corpus seeks to serve as a base for advancing Arabic language technologies and furthering our understanding of the sociolinguistic landscape in Saudi Arabia.

8. Acknowledgements

Place all acknowledgments (including those concerning research grants and funding) in a separate section at the end of the paper.

9. Bibliographical References

Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, Kareem Darwish, Abdullah Alrajeh, et al. 2024. Sada: Saudi audio dataset for arabic. https://www.kaggle.com/datasets/sdaiancai/sada2022.

Lamya Alkanhal, Abeer Alessa, Elaf Almahmoud, and Rana Alaqil. 2023. Aswat: Arabic audio dataset for automatic speech recognition using speech-representation learning. In *Proceedings of the First Arabic Natural Language Processing Conference*, pages 120–127. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Jeong-Uk Bang, Seung Yun, Seunghi Kim, Mu-Yeol Choi, Min-Kyu Lee, Yeojeong Kim, Dong-Hyun Kim, Jun Park, Youngjik Lee, and Sang-Hun Kim. 2020. Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences*, 10:6936.

He Bu, Jun Du, Xu Na, Bengu Wu, and Hui Zheng. 2017. Aishell-1: An open-source mandarin speech corpus. https://openslr.org/33.

Alexandra Canavan, David Graff, and George Zipperlen. 1997. Callhome egyptian arabic speech. https://catalog.ldc.upenn.edu/LDC97S45. LDC97S45.

Christopher Cieri, David Miller, and Kevin Walker. 2004. Fisher english training speech part 1. https://catalog.ldc.upenn.edu/LDC2004S13. LDC2004S13.

Charles A. Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

S. Furui, Masanobu Nakamura, Tomohisa Ichiba, and K. Iwano. 2005. Analysis and recognition of spontaneous speech using corpus of spontaneous japanese. *Speech Commun.*, 47:208–219.

John Godfrey, Edward Holliman, and Jane McDaniel. 1997. Switchboard-1 release 2. https://catalog.ldc.upenn.edu/LDC97S62. LDC97S62.

Nizar Habash, Owen Rambow, and George Kiraz. 2012. Conventional orthography for dialectal arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1251–1258.

Clive Holes. 2006. The arabic dialects of arabia. *Proceedings of the Seminar for Arabian Studies*, 36:25–34.

LDC. 2006a. Gulf arabic conversational telephone speech. https://catalog.ldc.upenn.edu/LDC2006S48. LDC2006S48.

LDC. 2006b. Iraqi arabic conversational telephone speech. https://catalog.ldc.upenn.edu/LDC2006S29. LDC2006S29.

Rodrigo Lima, Sidney E. Leal, Arnaldo Candido Junior, and Sandra M. Aluísio. 2025. A large dataset of spontaneous speech with the accent spoken in são paulo for automatic speech recognition evaluation. In *Intelligent Systems: 34th Brazilian Conference, BRACIS 2024, Belém Do Pará, Brazil, November 17–21, 2024, Proceedings, Part I*, page 33–47, Berlin, Heidelberg. Springer-Verlag.

Deyi Liu and Pascale Fung. 2005. Hkust mandarin telephone speech. https://catalog.ldc.upenn.edu/LDC2005S15. LDC2005S15.

Kikuo Maekawa. 2003. Corpus of spontaneous japanese (csj). https://clrd.ninjal.ac.jp/csj/en/.

John Makhoul and Bassam Zawaydeh. 2005. Bbn/aub darpa babylon levantine arabic speech and transcripts. https://catalog.ldc.upenn.edu/LDC2005S08. LDC2005S08.

Mark A. Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. 2002. The buckeye corpus of conversational speech. https://buckeyecorpus.osu.edu/.

Matt Post and John Garofolo. 2010. Fisher spanish training speech part 1. https://catalog.ldc.upenn.edu/LDC2010S01. LDC2010S01.

Thomas Prochazka. 1988. *Saudi Arabian Dialects*, 1 edition. Routledge.

Alec Radford, Jong Wook Kim, Christopher Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2306.16217*.

Wolfgang Wahlster. 2000. Verbmobil speech databases. https://www.phonetik.uni-muenchen.de/Bas/BasVM.html.

10. Language Resource References