Int. J. Advance Soft Compu. Appl, Vol. 18, No. 2, July 2026
Print ISSN: 2710-1274, Online ISSN: 2074-8523

ArDSS: An Arabic Down Syndrome Speech Corpus

Fatimah Alzahrani[1] and Maha Alamri[2]

[1]Faculty of Computing and Information, Al-Baha University, Al-Baha 65779, Saudi Arabia
e-mail: 446050384@stu.bu.edu.sa
[2]Faculty of Computing and Information, Al-Baha University, Al-Baha 65779, Saudi Arabia
e-mail: m.alamri@bu.edu.sa

Abstract

Down Syndrome (DS) is associated with characteristic speech impairments, yet Arabic resources remain limited. This paper introduces ArDSS, an annotated Arabic speech corpus of DS speech for acoustic analysis and for supporting future Automatic Speech Recognition (ASR) research. Recordings were collected from 13 Arabic-speaking children with DS (6 males, 7 females; ages 5–17) in Saudi Arabia using a clinically guided set of 100 phonetically diverse isolated words, captured in 3–4 sessions per participant under controlled conditions. Following quality control and segmentation, ArDSS comprises 5,509 utterances (16 kHz, 16-bit WAV) totalling 14.6 hours. Each utterance is linked to XML metadata and attempt-level annotations, and a standardised pipeline extracts fundamental frequency, formants (F1–F3), and voice-quality measures (harmonics-to-noise ratio, jitter, shimmer) after feature-validity screening, yielding 3,279 productions with complete acoustic profiles. Corpus analyses indicate substantial inter-speaker variability and recurrent Arabic-relevant pronunciation patterns, including liquid substitution, emphatic neutralisation, cluster reduction, and affricate simplification. ArDSS provides a foundational resource for Arabic clinical phonetics and supports future research on DS-aware ASR and assistive applications.

Keywords: *Down Syndrome, Arabic, Speech Corpus, Acoustic Analysis, Voice Quality, Formant Frequencies.*

1 Introduction

Down Syndrome (DS) is a genetic disorder that occurs in about one out of every 800 live births globally [1]. Individuals with DS usually present with impaired speech and language development, including delayed oral–linguistic abilities, decreased intelligibility, and atypical (both developmental and non-developmental) phonological patterns [2].

These difficulties may severely interfere with communication, thereby emphasising the need for intervention strategies that enhance performance in all areas of communication (academic, social, and vocational). Prior studies have also shown that individuals with DS display a number of anatomical and physiological differences from typically developing

Received 20 March 2026; Accepted 12 June 2026

children in their speech production, including poor articulation and reduced vowel production with centralized low vowels as well as an increase in variability of the vowels produced [3][4].

Arabic presents many unique challenges for conducting speech research as a result of its complex phonological structure (including the presence of emphatic consonants), highly rich morphology, and dialectal variation, which varies across phonology, morphology, and syntax [5]. The phonetic and articulatory attributes of Arabic may interact differently with the speech patterns of individuals with DS than those of more-studied languages.

Utilising speech recognition technologies with artificial intelligence (AI) has the potential to greatly enhance access to communication for individuals with speech impairments [6, 8]. Automatic Speech Recognition (ASR) can facilitate learning and communication by providing aligned text representations for spoken language, thereby improving access to education and meeting different learning styles [7]. However, achieving good performance generally requires large speech corpora; in contrast, the available disordered speech data remains small, more difficult to obtain compared to typical speech corpora [8, 9].

Although there has been significant progress in disordered-speech corpora for English and other Western languages [10], Arabic language disorder resources remain very few [11]. The lack of corpora will negatively impact the ability to design inclusive ASR systems designed to address variation among different types of speakers, including speakers with speech impairments and atypical speaking patterns [12,13]. Therefore, creating a speech corpus for Arabic speakers with DS is critical in addressing this need and supporting communication, diagnosis, and intervention for Arabic speaking individuals with DS.

Thus, the purpose of this paper is to create and describe a speech corpus of Arabic-speaking individuals with DS. This corpus provides a foundational resource for speech analysis, modelling, and automatic recognition of disordered Arabic speech. The potential use of this resource is in the areas of clinical assessment/diagnosis, assistive devices, and educational programmes for special needs children.

The remainder of this paper is organised as follows. Section 2 reviews the speech characteristics associated with individuals with DS, the phonological complexity of Arabic, and existing corpora of speech disorders. Section 3 describes the methodology, including ethical approval, participant recruitment, vocabulary design, recording procedures, acoustic analysis, and metadata annotation. Section 4 presents the corpus overview, acoustic characteristics, inter-speaker variability, and observed pronunciation error patterns. Section 5 interprets the findings in relation to clinical, linguistic, and technological implications, and addresses limitations and future directions. Finally, Section 6 summarises the main contributions.

2 Related Work

2.1 Overview of Arabic characteristics

Arabic presents significant computational and linguistic challenges, including morphological complexity [14,15] and dialectal varieties that add another layer of complexity [16].

Arabic uses a root-and-pattern morphological system [5]. Compared to English, Arabic has been reported to show approximately 2.5 times higher vocabulary growth and about 10 times higher out-of-vocabulary (OOV) rates when using a fixed-size lexicon [5]. The morphological complexity of Arabic creates additional computational challenges for speech and language processing systems [15].

Individuals with DS demonstrate difficulties with the structural aspects of language, including morphology and syntax [17]. Phonologically, Arabic includes emphatic consonants characterised by secondary pharyngeal articulation [18]. Recent research in Arabic-speaking individuals with DS has revealed a general consonant accuracy of only 50%, with place substitutions being the most frequent error pattern, and consonant cluster reduction being the most common mismatch of word structure [12]. Mashaqba et al. [17] mentioned that individuals with DS show difficulties with the structural aspects of Arabic morphology, with singular forms produced more accurately (83.3%), followed by plural forms (27%) and dual forms showing the greatest difficulty (10.3%).

2.2 Down Syndrome speech characteristics

Individuals with DS demonstrate consistent patterns of speech that differentiate it from all other forms of speech produced by typically developing individuals and individuals with other developmental disorders [2]. The process of speech development for individuals with DS is clearly delayed compared to typically developing infants, and first words are often produced at an average age of approximately 21 months versus 14 months for typically developing children [19]. Across all ages, and even into the adulthood of development, speech intelligibility is also very low for individuals with DS; studies of adults have reported scores between 41% and 75% for intelligibility [20]; compared to typically developing peers, who generally achieve substantially higher intelligibility.

The temporal characteristics of speech in individuals with DS have been found to be significantly disturbed. Some studies have reported mixed results regarding speech rate; however, many studies have documented substantial interruption to the rhythm and prosodic structure of speech [2]. Timing deficiencies have important implications for phonological systems like Arabic language because they classify long and short vowels as basic units of phonology [5].

Individuals with DS generally have breathy and hoarse voice qualities. Acoustic studies have reported some variable findings on this, including increased frequency perturbations (jitter) and amplitude perturbations (shimmer) but results are variable between studies, thus it is complex to evaluate only an individual's voice patterns due to several factors such as abnormal vocal fold vibration, nasalization and vocal tract resonances interacting to affect voice quality [2]. Research has also explored the fundamental frequency (F0), where early descriptions suggested a lower pitch, although the acoustic evidence remains mixed, studies looking at acoustic data are still inconsistent in terms of whether or not there are significant differences from individuals who develop typically [2].

In addition to phonatory measurements, the use of segmental acoustic analysis can also provide information regarding articulatory stability. Due to greater centralized low vowels and greater variability with low vowels for individuals with DS than for typically developing controls, vowel formant dispersion has been shown to be a particularly sensitive measure for distinguishing between groups [4].

Issues with consonant articulation have been frequently raised, including reports of substitution and deletion processes; however, fricatives, clusters and liquids have been noted as particularly challenging to produce [2]. Articulatory placement and movements have also been shown to differ from non-DS speakers. For example, there are fewer articulatory places of constriction when producing vowels. The average surface area of the tongue body movement has been approximated to be 30 % smaller in individuals with DS compared to control speakers [21].

The prosodic characteristics represent a particularly challenging aspect of speech in individuals with DS; all available studies demonstrate difficulties in the perception, imitation and spontaneous production of prosodic features [2]. Acoustic and perceptual studies show that prosodic differences affect the intonation, accentuation, rhythm, and speech rate; each of these differences has a significant impact on communicative effectiveness [22].

Compared with typically developing control subjects, individuals with DS have been reported to have higher values of fundamental frequency; there is, however, increased variability in temporal parameters such as vowel duration [22]. Disfluency problems (stuttering and cluttering) are also frequently reported in individuals with DS and may further disrupt both the rhythm and the prosodic control of speech [22].

The Arabic language uses prosody to differentiate between words. One of the ways the Arabic language differentiates between words is using stress and vowel length [5, 15]. Vowel length is phonemically contrastive in Arabic. This means that short and long vowels have separate phonetic categories [5]. Plain and emphatic consonant pairs, including pharyngeal sounds, also contribute to the unique sound of Arabic and require different ways of making these sounds [5].

2.3 Speech corpora for individuals with speech disorders

Current research has increasingly focused on collecting speech samples from people with communication disorders, in order to create a more representative view of how atypical speech is produced acoustically and linguistically [9,23]. However, there is a large discrepancy in the resources available for use in this field. Most currently available disordered-speech databases are based on English-language recordings and predominantly contain samples of acquired motor speech disorders, particularly dysarthria [6,25]. This leaves a lack of resources representing disordered speech in non-English languages and developmental speech disorders [12,24].

Table 1 summarises key speech corpora that focus on individuals with speech disorders, including the development of data collection efforts and the specific focus of each corpus.

Table 1. Comparison of Speech Corpora for Individuals with Speech Disorders.

Corpus	*Year*	*Lang*	*Hour*	*Recording*	*Speaker*	*Disorder*	*Age*	*Ava.*
Rondal-DS	1978	EN	21	21	21	DS	Child	Public
UASpeech	2008	EN	N/A	14,535	19	CP	Adult	Restricted
TORGO	2012	EN	N/A	16,394	15	CP, ALS	Adult	Public
Project Euphonia	2021	EN	>1,300	1M+	>1,000	Disordered	Adult	Restricted
PRAUTOCAL	2021	SP	~3.8	2,151	50	DS	Adult	Restricted
EasyCall	2021	IT	N/A	21,386	31	Dysarthria	Adult	Public
Impairments Arabic	2022	AR	<1	770	38	Speech impairment	Child	Restricted
SANACS	2024	SL	15+	N/A	37	Autism	Child	Restricted
MSA-SDC	2024	AR	~59	61,294	40	Articulation	Adult	Restricted
Arabic Voice (DS)	2025	AR	<1	N/A	27	DS	Child	Public

Note: DS = Down Syndrome; CP = Cerebral Palsy; ALS = Amyotrophic Lateral Sclerosis. Hours denote total duration; ~ indicates approximate value. N/A = not specified in original publication.

As part of earlier foundational work, the Rondal-DS Corpus [26] was established as part of the CHILDES database. The Rondal-DS Corpus provides transcribed recordings of

interactions between 21 English-speaking individuals with DS and their mothers, along with a matched group of 21 typically developing children and their mothers. It consists of approximately 484,000 words and is a well-regarded reference point in linguistic research. In the domain of acquired motor speech disorders, the UASpeech corpus [27] provides a reference for dysarthric speech, featuring 14,535 isolated-word recordings from 19 speakers with Cerebral Palsy (CP). Similarly, the TORGO corpus [28] offers a rich multimodal resource from 15 speakers (8 with cerebral palsy or amyotrophic lateral sclerosis and 7 age- and sex-matched controls), which contains aligned acoustic and articulatory data (including 3D measurements), supplemented by detailed clinical assessments.

More recently, efforts have diversified. The EasyCall corpus [29] was developed for Italian, comprising 21,386 speech commands from 31 individuals with dysarthria and 24 healthy controls, with the severity of dysarthria assessed using the Therapy Outcome Measure (TOM). The Slovak Autistic and Non-Autistic Child Speech Corpus [30] provides a valuable resource for studying developmental communication differences, containing more than 15 hours of task-oriented speech from 37 children with autism spectrum disorder and 30 neurotypical controls (ages 6–13 years), with clinical assessments including autism diagnostic measures, IQ testing, and theory of mind evaluations. For DS specifically, the PRAUTOCAL corpus [31] introduced an innovative approach by collecting the speech of 50 Spanish-speaking individuals with DS (age range 13–42 years) through a therapeutic video game designed to analyse prosodic and linguistic features.

In Arabic, Shareef et al. [32] developed an Arabic speech corpus containing 770 speech samples from 38 children with impairments (ages 7–11 years), collected from schools for children with special needs in Mosul, Iraq, focusing on incorrectly pronounced Arabic letters and numerals. The Modern Standard Arabic Speech Disorders Corpus (MSA-SDC), introduced in 2024, contains recordings from 40 adult Jordanian speakers (20 male and 20 female) with articulation disorders, including distortion and/or substitution patterns [11]. A more targeted contribution is the Arabic Voice Recognition (Down Syndrome Children) corpus [33], a publicly available collection of isolated Arabic letters, words, and numbers from 37 children (27 with DS). Although the creation of this corpus is a significant development in inclusive speech recognition, its scope is limited to single-word recordings and it lacks comprehensive participant information. This leaves an important gap in structured clinical and linguistic data. Thus, there is a strong need for an Arabic speech corpus of individuals with DS that includes comprehensive clinical evaluations.

3 Arabic Down Syndrome Speech Corpus (ArDSS)

To create an Arabic speech corpus from individuals with DS, a structured approach was utilised. The initial steps were to obtain ethical approval, define the inclusion and exclusion criteria, and determine how speech samples would be recorded. After that, appropriate speech material was selected, and informed consent was obtained from participants or their guardians before data collection. The following subsections detail the key processes for designing, recording and annotating the proposed Arabic DS speech corpus.

3.1 Ethical approval

The researchers obtained ethical approval to conduct this study from Al-Baha University's Standing Committee for Scientific Research Ethics (Approval No. 46117515, dated 20/10/1446 H; corresponding to 2025-04-18 CE). Prior to data collection, legal guardians of participating children were provided with detailed information about the purpose of the

study, how their child's data would be collected, the potential benefits and risks of participation, and how their data would be kept confidential. All recorded audio files were assigned unique identification codes (S1–S13) to ensure participant confidentiality, and no identifying information was included in the associated metadata files.

3.2 Participants' information

Through working together, three specialised centres in Saudi Arabia that serving individuals with special needs recruited participants to this study. These centres provide comprehensive services to children and adult clients with various types of disabilities including DS and retain complete medical and developmental records for every client. To be eligible to participate in this study, each child had to meet five criteria: (1) a confirmed clinical diagnosis of DS (trisomy 21) as documented by the medical records of the center in which they were recruited; (2) an age of five years or older and less than 18 years; (3) Arabic as their primary language; (4) an ability to produce single words when requested; and (5) an absence of a hearing impairment severe enough to prevent them from following verbal instructions.

A total of 13 children (6 males, 7 females) met the inclusion criteria and completed the full data collection protocol. Demographic and clinical characteristics are summarised in Table2.

Table 2. Participant demographics, recording duration, and distribution of recordings.

ID	*Age*	*Gender*	*Recordings (%)*	*IQ*	*Speech Level*	*Duration (min)*
S1	5	F	11.69	42	Average	63
S2	5	M	1.58	48	Average	40
S3	6	M	3.45	52	Average	115
S4	7	F	8.82	64	Good	60
S5	7	M	2.00	45	Good	32
S6	8	F	3.45	49	Good	49
S7	9	F	4.54	45	Excellent	28
S8	9	F	3.27	49	Average	30
S9	10	M	9.08	38	Poor	74
S10	11	M	1.62	55	Excellent	169
S11	12	M	17.10	40	Poor	104
S12	16	F	15.25	40	Poor	71
S13	17	F	18.15	50	Excellent	39
Mean ±SD	9.38± 3.82	6M/7F	100.00	47.46± 7.10	—	874

Note: Recordings (%) denotes the relative contribution of each participant to the total number of recordings (n = 5,509). Speech levels were assessed by licensed speech-language pathologists.

The ages of the participants ranged from 5 to 17 years (mean = 9.38 ± 3.82 years), and the IQ scores ranged from 38 to 64 (mean = 47.46 ± 7.10), assessed by clinical psychologists using standardised intelligence tests. Speech intelligibility was evaluated by licensed speech-language pathologists (SLPs) and classified as poor, average, good, or excellent. Audiological screening confirmed normal or corrected hearing for all participants (pure-

tone thresholds ≤ 25 dB HL at 250–8000 Hz). All participants received ongoing SLP services at their centres (1–3 sessions per week).

3.3 Vocabulary selection

The selection of target vocabulary was guided by the principles of clinical relevance, linguistic representativeness, and developmental appropriateness. In collaboration with Arabic-speaking speech-language pathologists in participating centres, 100 words were identified that were: (i) commonly used in therapeutic contexts; (ii) representative of Arabic phonological patterns; (iii) appropriate for the developmental levels of the participating speakers; and (iv) suitable for single-word production tasks.

The final vocabulary includes a wide range of semantic categories, as shown in Table 3. This selection ensures comprehensive phonological coverage and is useful for testing and intervention. The vocabulary includes 6 numbers (6.0%), 5 colours (5.0%), 7 family terms (7.0%), 4 verbs (4.0%), 13 animals (13.0%), 12 food items (12.0%), 27 household objects (27.0%), 5 emotions (5.0%), 5 body parts (5.0%), and 16 other terms (16.0%).

Table 3. Vocabulary distribution by semantic category.

Category	*Words*	*Percentage*	*Examples (Arabic)*
Numbers	6	6.0%	واحد، اثنين، ثلاثة
Colours	5	5.0%	احمر، اخضر، ازرق
Family Terms	7	7.0%	اخ، اخت، بابا
Verbs	4	4.0%	اشرب، اكل، نام
Animals	13	13.0%	اسد، بطة، سمك
Food Items	12	12.0%	تفاح، خبز، موز
Household	27	27.0%	باب، سرير، قلم
Emotions	5	5.0%	تعبان، فرحان، خايف
Body Parts	5	5.0%	عين، يد، راس
Other	16	16.0%	خلاص، شمس، صاروخ
Total	100	100%	—

A review of the phonological properties of the chosen vocabulary reveals a broad range of Arabic sounds, comprising all three short and long vowels, as well as all 28 Arabic consonants, including emphatic and pharyngeal consonants. The syllables targeted in the study varied from one to four, with a predominance of two-syllable words in the sample; this reflects typical patterns of early lexical acquisition. The full range of Arabic syllable structures is represented in the stimulus materials (e.g., CV, CVC, CVCC), thus providing a representative measure of the participants' phonotactic abilities.

3.4 Recording procedures

Recordings were collected in quiet rooms within specialized centres to reduce background noise. All recordings were captured using Audacity software (version 3.7.1) and saved as 16 kHz, 16-bit WAV files to provide sufficient quality to conduct acoustic analyses of the speech across all children who participated in the study while being within a manageable file size for all participants. The presentation of the target words included combined visual and auditory prompts; a tablet was used to display large, clear Arabic text representing the target word with a corresponding pictorial representation to support the child's understanding. The researcher presented an auditory model of the target word by clearly

pronouncing each word. After the child heard the auditory model, the researcher asked the child to repeat the word. If the child did not provide a response within 10 seconds after the researcher presented an auditory model, then the researcher presented an auditory model again up to three times before the researcher moved on to the next target word.

In order to capture natural variability in speech production for each target word, participants were asked to produce each word several times (5–10 attempts), with breaks provided as needed to prevent fatigue. Individual sessions lasted approximately 35–40 minutes, and each participant generally completed 3–4 sessions to accommodate differences in attention span and reduce fatigue effects. Quality was monitored throughout the recording sessions; sessions were paused if the participant showed signs of fatigue, distress, or loss of attention. Frequent positive reinforcement was also provided to help maintain motivation. The total recording time for all participants was 874 minutes (14.6 hours). Figure 1 illustrates the complete corpus creation workflow, from participant recruitment to final corpus preparation.

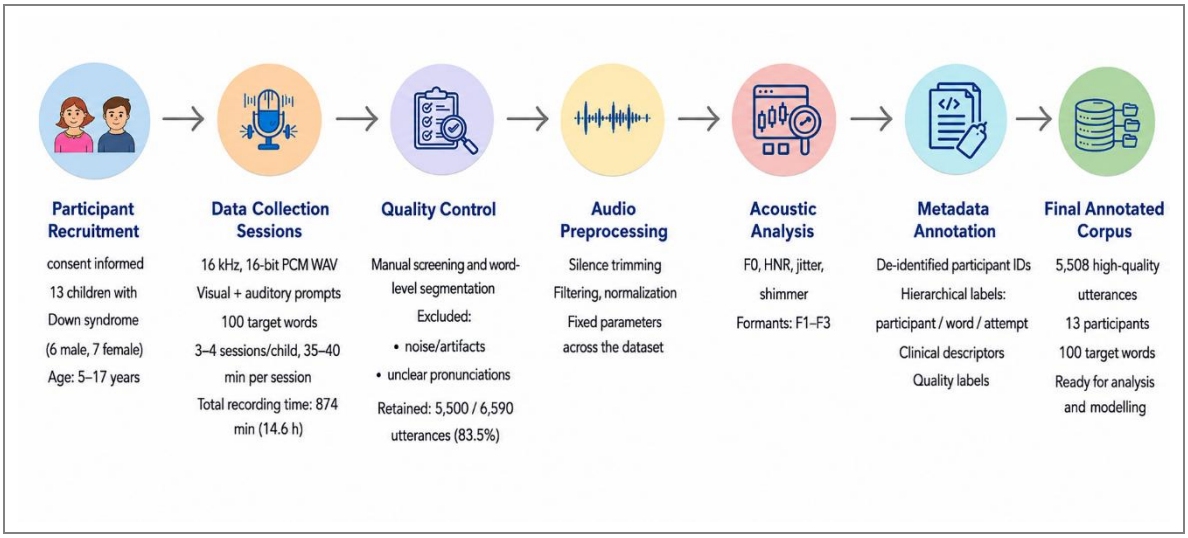

Fig. 1. ArDSS creation workflow.

3.5 Data processing and acoustic analysis

The methods for signal processing and acoustic analysis of disordered child speech all had a focus towards optimising the consistency, reproducibility, and interpretability of clinical outcomes. These methods were selected based on previously published procedures for examining voice quality in disordered speech [2].

Once the data were collected, all continuous recordings were segmented into individual word productions using Audacity. The original timing characteristics of speech, such as pauses and hesitations, were deliberately retained because they may have clinical significance. Each individual production was inspected manually to ensure that the audio quality was acceptable and that the recording corresponded to the intended target word. Any production with excessive background noise, microphone artefacts, or unclear articulation was removed from the dataset. The resulting dataset consisted of 5,509 acceptable word productions from an initial pool of approximately 5,890 recordings (93.5% retention rate). The segmented files were organised systematically by target word, and each audio file was labelled using anonymised speaker identification codes to ensure the confidentiality of the participants.

Before acoustic analysis, all audio files were subjected to a standardised preprocessing pipeline implemented using custom Python scripts with the librosa library and SciPy. The complete preprocessing and feature extraction pipeline is illustrated in Fig. 2.

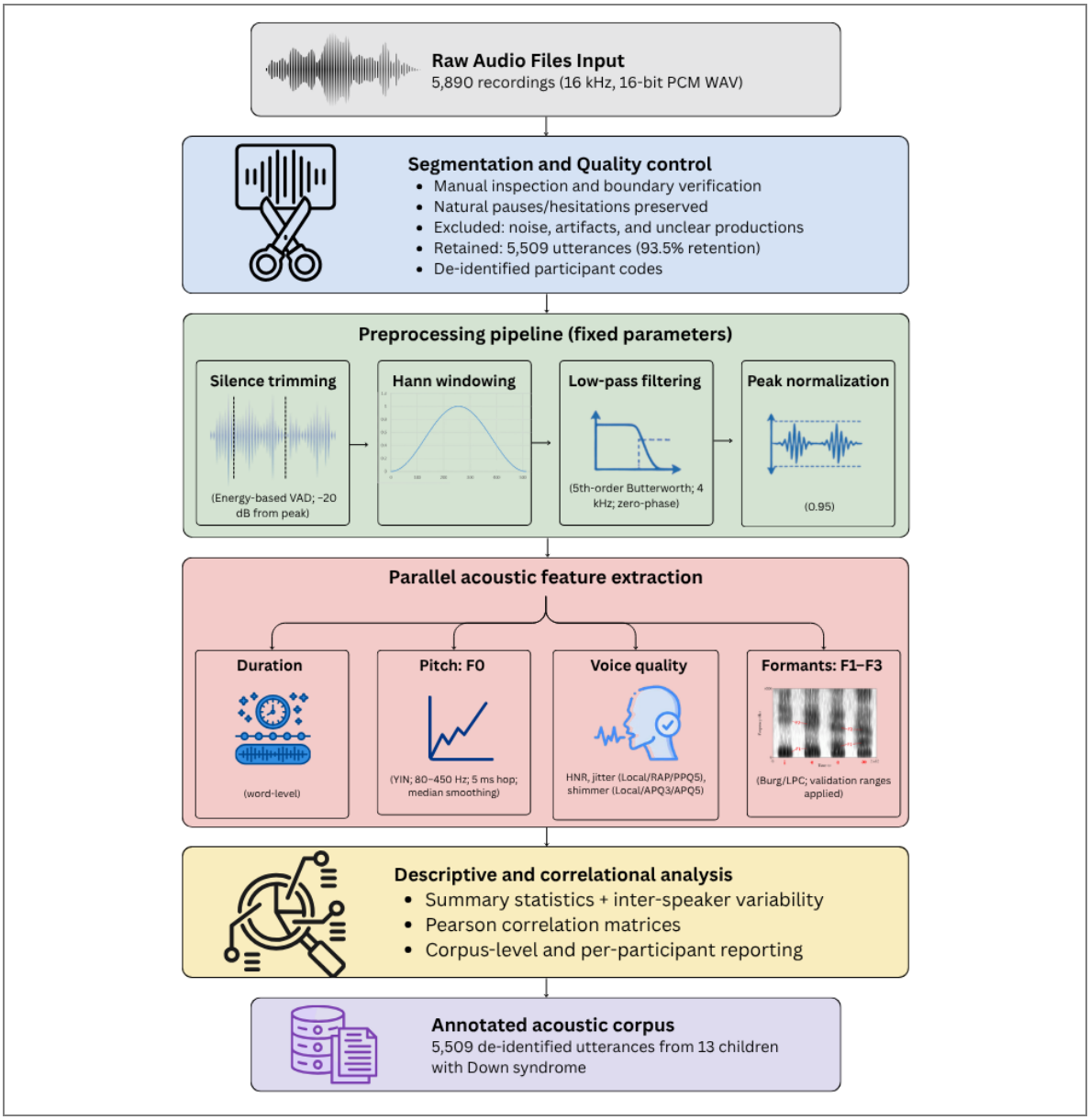

Fig. 2. Standardised audio preprocessing and acoustic feature extraction pipeline for the ArDSS corpus.

The preprocessing steps included: (1) loading audio signals at a fixed sampling rate of 16 kHz; (2) automatic trimming of leading and trailing silence using energy-based voice activity detection with a threshold set at 20 dB below peak amplitude; (3) application of a Hann (Hanning) window to reduce spectral leakage; (4) fifth-order Butterworth low-pass filtering with a cutoff frequency of 4000 Hz using zero-phase filtering (scipy.signal.filtfilt) to attenuate high-frequency noise while preserving speech-relevant information; and (5) peak amplitude normalisation to 0.95 to ensure consistent signal levels while avoiding clipping artefacts. To ensure consistency and comparability throughout the corpus, all preprocessing parameters remained constant across all recordings. Acoustic analysis was performed with Parselmouth [34], a Python interface to Praat [35], in combination with

librosa and custom feature extraction algorithms. For each preprocessed recording, the following acoustic parameters were extracted. Specifically, the fundamental frequency (F0) of each recording was estimated using librosa's YIN algorithm [36], an autocorrelation-based pitch-tracking method designed to reduce estimation errors. The pitch tracking was configured using a predefined frequency range appropriate for child speech to ensure robust pitch estimation; a hop length of 5 ms was used to capture fine-grained pitch variations. The resulting F0 contours were then smoothed using median filtering with an adaptive kernel size of up to 5 samples to reduce octave errors and spurious pitch estimates. Only valid F0 values were retained after excluding non-finite and clearly spurious estimates. From these contours, the mean, median, standard deviation, minimum, maximum, range, and coefficient of variation (CV = SD/Mean × 100%) were calculated for each recording. The voiced ratio was also computed as the proportion of voiced frames, providing an additional indicator of phonation continuity.

Voice quality measures were extracted using Parselmouth's interface to Praat's voice quality analysis functions. Harmonics-to-noise ratio (HNR) was calculated using Praat's cross-correlation method with a time step of 10 ms and 4.5 periods per analysis window, providing a measure of the vibration quality of the periodic vocal fold [35]. Jitter (frequency perturbation) and shimmer (amplitude perturbation) were quantified using Praat's periodicity analysis functions [37].

To accommodate the variable voice quality characteristics of speech in individuals with DS, an adaptive silence-threshold strategy was employed by systematically testing threshold values of 0.03, 0.05, 0.10, and 0.15 and selecting the lowest threshold that yielded valid measurements.

Local jitter (cycle-to-cycle variability), relative average perturbation (RAP), and five-point period perturbation quotient (PPQ5) were computed as jitter measures. Similarly, local shimmer, three-point amplitude perturbation quotient (APQ3), and five-point amplitude perturbation quotient (APQ5) were extracted as shimmer measures. All Praat-based perturbation analyses were conducted using consistent pitch settings across recordings, and the perturbation measures were expressed as percentages. Given the variability in voice quality among individuals with DS, perturbation measures were interpreted descriptively within the context of the extracted dataset. These measures were included to characterise variability in voice quality rather than to impose strict exclusion criteria.

Praat's Burg method was applied to estimate the F1, F2, and F3 formant frequencies. Based on established child speech formant frequency ranges, and the criterion of having at least three valid estimates per recording, formant values were considered valid. Formant frequencies were analysed to describe vowel-space properties [38].

Descriptive statistics (mean, standard deviation, median, minimum, maximum, skewness, and kurtosis) were computed for each acoustic parameter for the complete corpus and separately for each participant using pandas and NumPy. Inter-speaker variability was assessed using coefficients of variation and range statistics. Pearson correlation coefficients were calculated to examine associations among the acoustic features in the corpus; this approach is relevant for corpus development and acoustic characterization studies [2,38].

3.6 ArDSS annotation

Comprehensive metadata were recorded for each utterance using a structured XML schema organised hierarchically across three levels.

At the participant level, each recording contains demographic and clinical information, including anonymised participant ID (tag: *id*), age (tag: *age*), gender (tag: *gender*), IQ score (tag: *IQ*), and speech intelligibility rating (tag: *clinicalSeverity*). These were independently assessed by licensed speech-language pathologists with experience working with individuals with DS. At the word level, the entries include the target word in Arabic orthography (tag: *word*) and the transcription of the observed pronunciation (tag: *observedPronunciation*). At the attempt level, given the inherent variability in speech of individuals with DS, each of the multiple production attempts (5–10 per target word) was annotated with: (1) attempt number (tag: *attempt*), (2) audio quality rating (Good/Excellent) (tag: *audioQuality*), (3) Pronunciation-related annotations include pronunciation level (pronunciationLevel), categorized as Poor, Average, Good, Excellent, or Fluent based on phonemic precision and overall intelligibility, and notable features (notableFeatures), which capture clinically relevant characteristics such as hesitations, sound substitutions, and specific articulatory errors. These pronunciation quality annotations were manually performed by two Arabic native speakers.

To assess the reliability of these annotations, inter-annotator agreement (IAA) was calculated through Cohen's Kappa [39], which is a widely accepted method for determining the degree of agreement between two annotators. This approach is commonly used to ensure annotation reliability across language research [40,41,42] including children's oral reading error detection [43], and conversational speech labelling [44], and emotional state recognition in children with DS [45].

Therefore, the degree of agreement for the pronunciation quality annotations was measured using Cohen’s Kappa, resulting in an agreement score of 88% for pronunciationLevel, indicating almost perfect agreement, and resulting in an agreement score of 77% for notableFeatures, indicating substantial agreement. This high level of consistency between annotators enhances the credibility and reliability of the annotated corpus.

Fig. 3 provides an illustrative example of the hierarchical annotation structure for two target words from participant S3, demonstrating how multiple production attempts capture natural variability in speech patterns characteristic of individuals with DS.

4 Acoustic Analysis Results

The acoustic analyses were conducted on productions for which the relevant acoustic measurements could be validly extracted after preprocessing and feature-validity checks. The main analysed subset included 3,279 productions, although a small number of F1 estimates were unavailable under the formant-validity constraints. The results therefore summarize the distribution of fundamental frequency (F0), voice quality measures, and formant-related features, and examine inter-speaker variability and relationships among key variables. Prior to analysis, all recordings underwent careful manual segmentation and inspection. Only speech segments corresponding to the target lexical items were retained, while non-speech intervals, examiner speech, background noise, and non-target utterances were excluded. The completed corpus consists of 100 phonologically diverse vocabulary items, chosen to reflect a variety of consonantal and vocalic structures of Arabic. The individual speaker contribution percentage was anywhere from 1.58% to 18.15% of the total recordings, highlighting variability in speaker contributions, but ensuring adequate representation for each speaker.

4.1 Descriptive statistics of acoustic features

As summarised in Table 4, descriptive statistics at the corpus level for the extracted acoustic features were obtained. Word duration showed substantial variability (Mean = 0.67 s, SD = 0.27 s), ranging from 0.19 s to 2.78 s. This variability is consistent with differences in lexical length, articulation rate, and speaker-specific production patterns. In addition, the voiced ratio was relatively high (Median = 1.00), suggesting that the analysed productions predominantly contained voiced segments suitable for periodicity-based measures.

```
<recording>
  <speaker>
    <id>S3</id>
    <age>6</age>
    <gender>Male</gender>
    <IQ>52</IQ>
    <clinicalSeverity>Average</clinicalSeverity>
  </speaker>
  <targetWord>
    <word>كتاب</word>
    <observedPronunciation>كتاب</observedPronunciation>
    <attempts>
      <attempt number="1">
        <audioQuality>Good</audioQuality>
        <pronunciationLevel>Average</pronunciationLevel>
        <notableFeatures>Slight hesitation, correct pronunciation</notableFeatures>
      </attempt>
      <attempt number="2">
        <audioQuality>Excellent</audioQuality>
        <pronunciationLevel>Good</pronunciationLevel>
        <notableFeatures>Fluent, correct pronunciation</notableFeatures>
      </attempt>
    </attempts>
  </targetWord>
  <targetWord>
    <word>أخضر</word>
    <observedPronunciation>خدل</observedPronunciation>
    <attempts>
      <attempt number="1">
        <audioQuality>Good</audioQuality>
        <pronunciationLevel>Poor</pronunciationLevel>
        <notableFeatures> ض → د substitution, simplification of "خ" </notableFeatures>
      </attempt>
    </attempts>
  </targetWord>
</recording>
```

Fig. 3. Example of structured metadata annotation for participant S3 showing hierarchical organization at participant, word, and attempt levels.

Table 4. Descriptive statistics of key acoustic features after feature-validity screening (N = 3,279).

Feature	*N*	*Mean*	*SD*	*Median*	*Min*	*Max*
Duration (s)	3,279	0.67	0.27	0.61	0.19	2.78
Mean F0 (Hz)	3,279	267.01	48.16	259.48	143.11	482.90
F0 SD (Hz)	3,279	39.31	28.79	33.67	0.33	159.63
F0 CV (%)	3,279	15.31	11.90	12.55	0.15	65.14
Voiced ratio	3,279	0.98	0.07	1.00	0.13	1.00
HNR (dB)	3,279	8.97	4.33	8.83	−3.71	21.90
Jitter (local, %)	3,279	2.00	2.03	1.40	0.04	15.89
Shimmer (local, %)	3,279	3.83	0.89	3.82	1.58	6.36

F1 (Hz)	3,272	563.26	103.38	559.35	291.72	871.71
F2 (Hz)	3,279	1,744.78	309.49	1,723.68	857.29	2,841.23
F3 (Hz)	3,279	2,838.72	294.72	2,860.56	1,703.00	3,610.43

Note: Formant extraction yielded a small number of missing values for F1 (N = 3,272), reflecting instances where reliable estimates were not available under the formant validity constraints.

4.2 Fundamental frequency (F0) characteristics

The mean F0 across all analysed productions was 267.01 Hz (SD = 48.16 Hz), ranging from 143.11 to 482.90 Hz, as shown in Figure 4. Most mean F0 values clustered around the midrange, with a right tail reflecting higher-pitched productions. Pitch variability within utterances (F0 SD) averaged 39.31 Hz, with observed values ranging from 0.33 to 159.63 Hz (see Table 4). The F0 CV averaged 15.31%, indicating moderate pitch variability relative to mean F0. The high median voiced ratio (1.00) indicates that the F0 measures were derived from predominantly voiced speech segments.

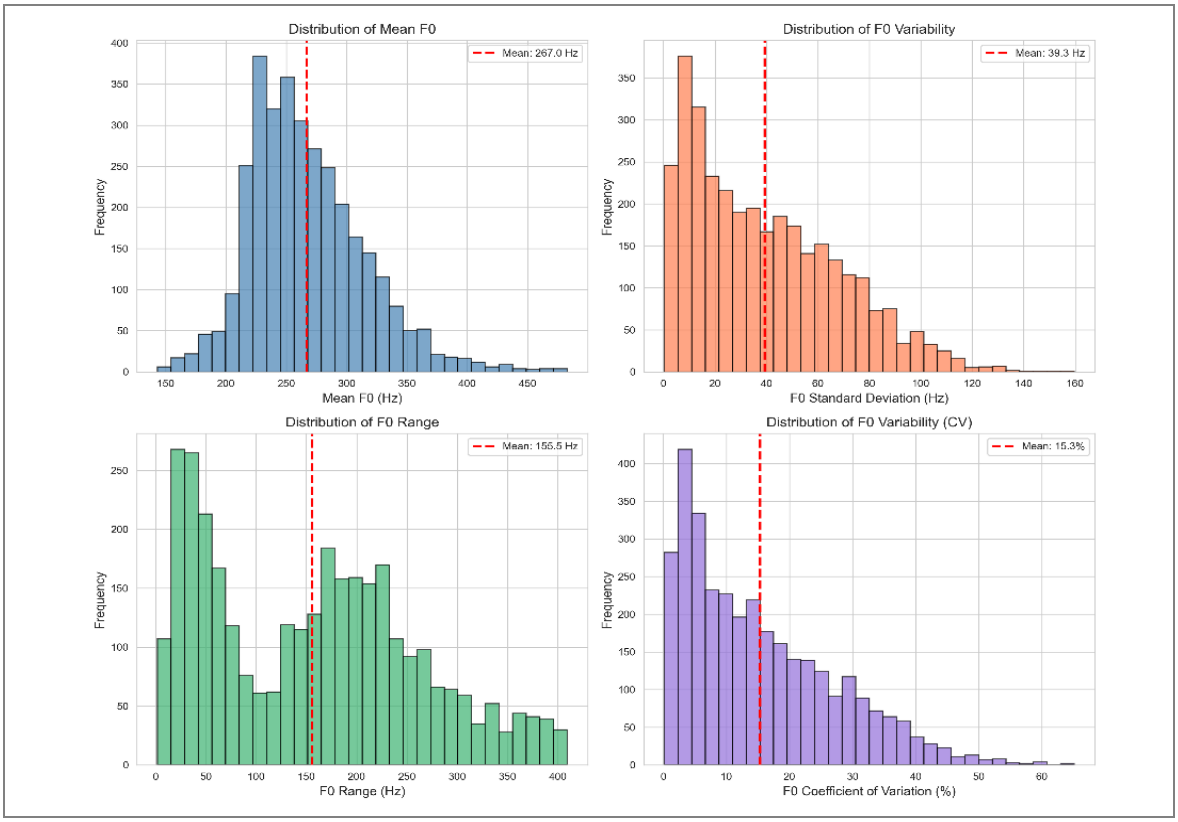

Fig. 4. Distribution of F0-related features across the analysed productions.

4.3 Voice quality measures

Voice quality was characterised using Harmonics-to-Noise Ratio (HNR) and perturbation measures (jitter and shimmer). The mean HNR was 8.97 dB (SD = 4.33 dB), with a broad range (−3.71 to 21.90 dB), indicating heterogeneous signal periodicity and noise components across productions, as illustrated in Fig. 5. Local jitter averaged 2.00% (SD = 2.03%), and local shimmer averaged 3.83% (SD = 0.89%), with distributions shown in Fig. 6. These values represent measurements that passed the predefined feature-validity

checks and therefore reflect technically reliable estimates within the adopted extraction pipeline rather than normative clinical reference ranges.

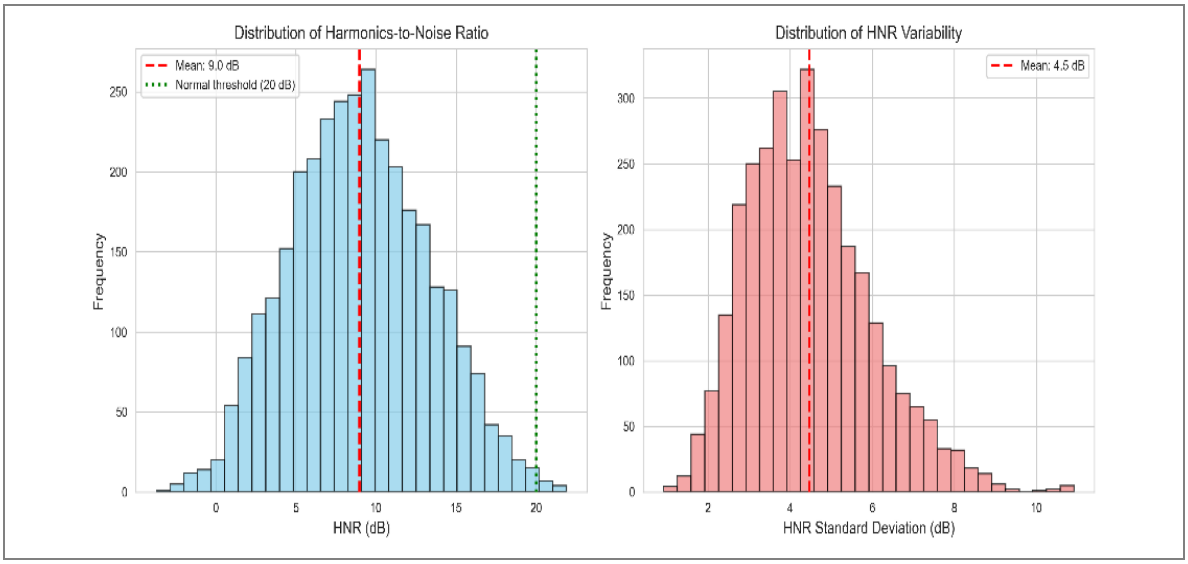

Fig. 5. Distribution of HNR features across the analysed productions.

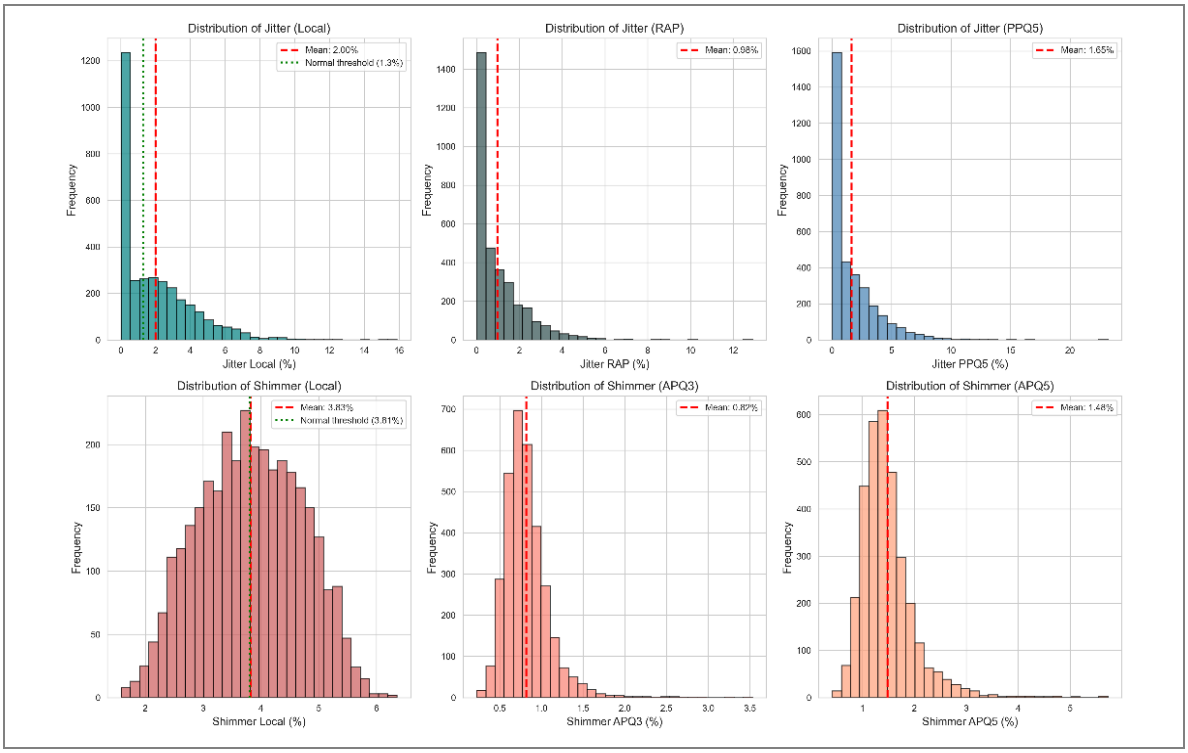

Fig. 6. Distributions of jitter and shimmer measures across the analysed productions.

4.4 Formant-related features

The average formant values across productions were F1 = 563.26 Hz, F2 = 1744.78 Hz, and F3 = 2838.72 Hz. These values reflect aggregated formant estimates derived from vowel-related segments throughout the set of recorded words. Fig. 7 visualises the distribution of the first three formants and the F1–F2 scatter, which provides an overall representation of the acoustic vowel space sampled by the corpus.

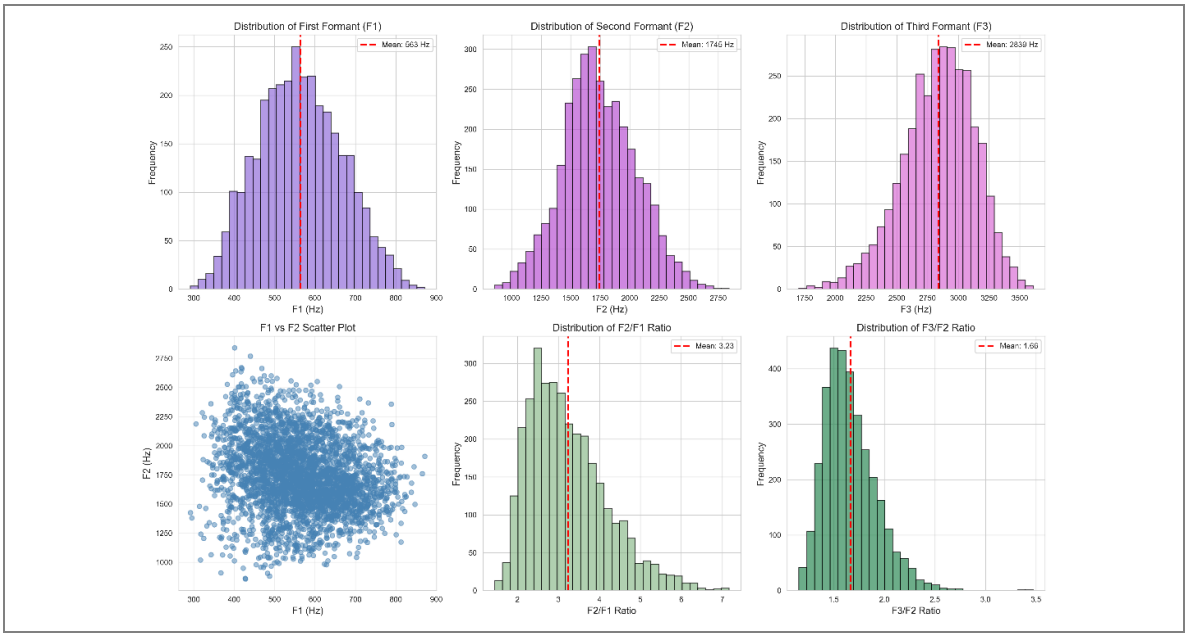

Fig. 7. Distributions of formant frequencies (F1, F2, F3) and an F1–F2 scatter plot representing the overall vowel space sampled by the recorded word set.

4.5 Inter-speaker variability

Substantial inter-speaker variability was observed among the 13 participants. The number of utterances analysed per speaker was imbalanced, ranging from 53 (S6) to 828 (S12), reflecting typical constraints in paediatric data collection and variation in task engagement between sessions. Table 5 shows the acoustic profiles for each speaker. The mean F0 ranged from 224.3 Hz (S13) to 327.8 Hz (S8), and the mean HNR ranged from 4.25 dB (S7) to 12.28 dB (S9). Fig. 8 further visualises this variability across key measures.

Table 5. Speaker-level summary of key acoustic features (N indicates the number of analysed utterances per speaker).

Speaker	*N*	*Mean F0*	*F0 SD*	*HNR*	*Jitter*	*Shimmer*	*F1 (Hz)*	*F2 (Hz)*
S1	209	226.7	36.3	5.88	3.59	3.69	577	1719
S2	55	228.8	34.7	5.79	2.30	2.57	523	2068
S3	149	258.2	31.0	11.07	1.04	2.96	511	1318
S4	355	259.3	42.9	7.25	2.93	3.89	578	1486
S5	62	256.7	60.3	7.25	2.87	3.83	597	1869
S6	53	307.1	40.6	5.32	2.26	2.59	568	1616
S7	114	269.8	34.4	4.25	2.08	3.56	602	1775
S8	179	327.8	64.0	5.48	1.92	2.72	592	1766
S9	429	280.3	37.2	12.28	2.18	4.34	525	1942
S10	60	322.5	50.5	8.40	2.56	4.73	670	1775
S11	318	316.6	45.9	6.42	2.18	3.34	604	1858
S12	828	263.4	29.4	9.76	1.60	3.96	595	1680
S13	468	224.3	11.5	11.90	1.11	4.40	478	1887

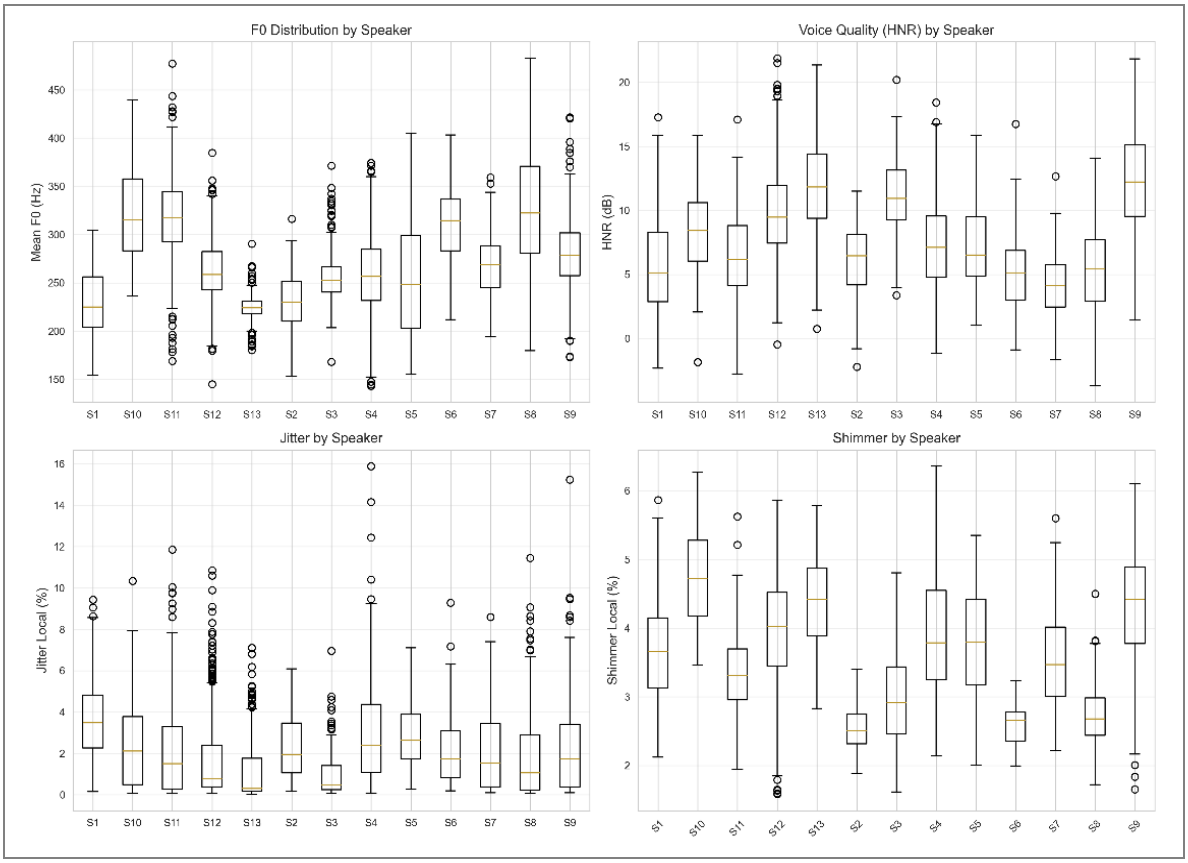

Fig. 8. Inter-speaker variability shown through speaker-level distributions for representative features (e.g., Mean F0, HNR, jitter, shimmer).

4.6 Correlation analysis

Pearson correlation analysis was conducted to examine the relationships among key acoustic features (Fig. 9). Strong positive correlations were observed among different jitter measures (e.g., local jitter and RAP; $r = 0.95$, $p < 0.001$), reflecting that these metrics capture closely related aspects of cycle-to-cycle periodicity variation. HNR showed moderate-to-strong negative correlations with jitter measures (e.g., HNR and local jitter; $r = -0.45$, $p < 0.001$), indicating that productions with higher noise components tended to exhibit increased periodicity instability. The mean F0 exhibited a weak negative correlation with HNR ($r = -0.06$, $p < 0.001$), suggesting only a small association between pitch level and HNR in this corpus. Given the large sample size, these results should be interpreted primarily in terms of effect sizes and practical relevance in addition to statistical significance.

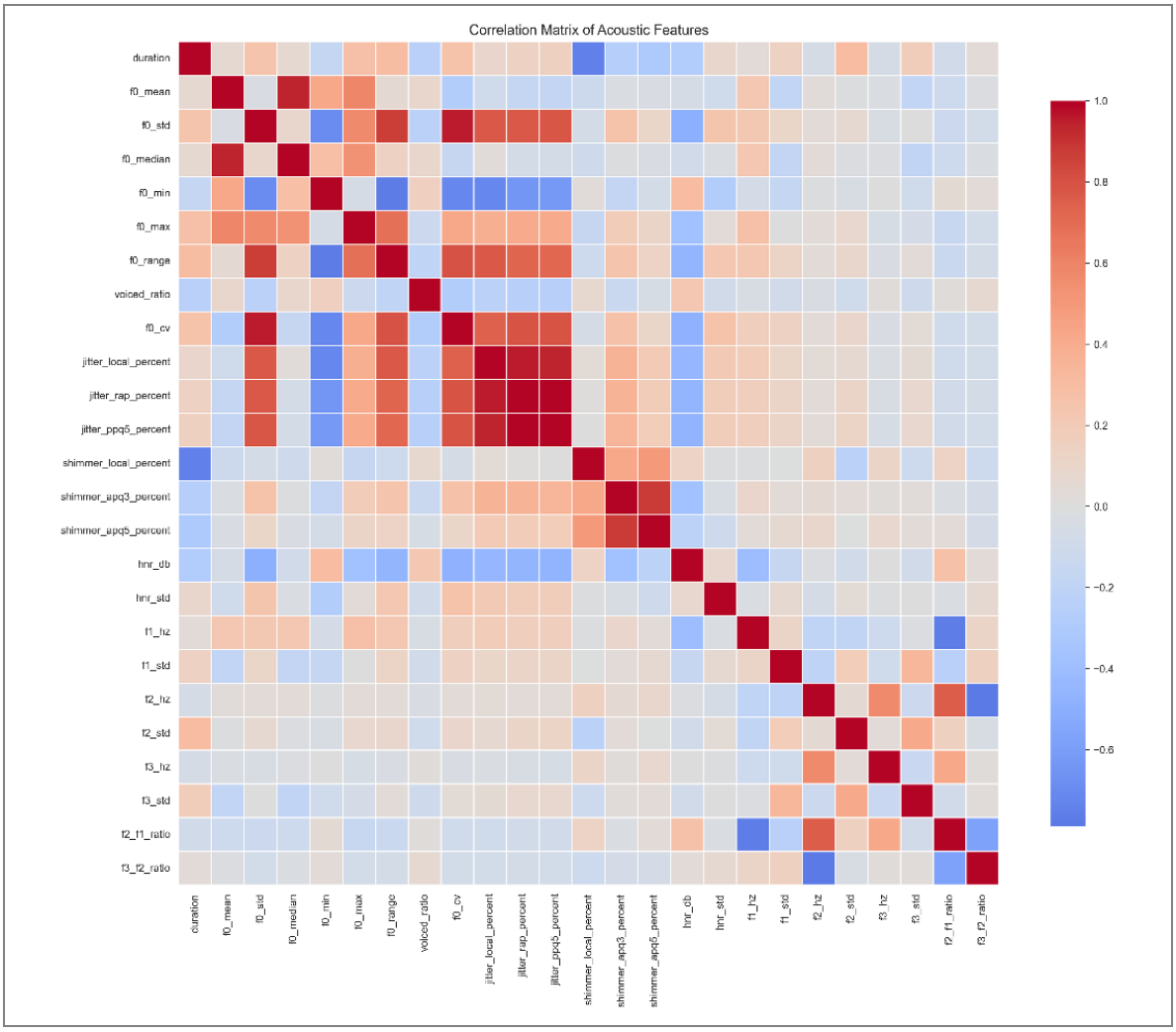

Fig. 9. Correlation heatmap of key acoustic features. Colour intensity reflects the magnitude of Pearson's r.

4.7 Pronunciation error patterns

The pronunciation quality annotations were further examined to identify recurring error patterns in the recorded productions, drawing on phonological-process categories reported in DS speech research and Arabic clinical phonology [2,19,46,47]. Table 6 summarises the most frequent patterns observed in the annotated data.

Table 6. Quantified pronunciation error patterns in ArDSS (N = 13)

Error Category	*Examples (target → observed)*	*Frequency (%)*
Liquid Substitution	أحمل → أحمر ; كبيل → كبير خلوف → خروف ; لاس → راس	92%
Affricate Simplification	زمل → جمل ; دزاز/زاز → دجاج زبل/زيل → جبل	87%
Emphatic Neutralisation	بيد/بيت → بيض ; اسفل → أصفر بتة → بطة	82%
Cluster Reduction	خدل/خضر → أخضر ازق/ازي → أزرق	78%

Weak Syllable Deletion (word-shape simplification)	تسفا → مستشفى ; ابعة/باعة → أربعة	74%
Guttural Deletion/Substitution	بز → خبز اليب → حليب ; اخة → أخ	69%

Note: Frequency represents the percentage of participants exhibiting the pattern in at least one production.

5 Discussion

The development of speech corpora for disordered populations remains limited considerably behind that of typical speech resources. While there are many large-scale resources for normal speech, such as Common Voice, which has approximately 2,500 hours of speech in 29 different languages [48], LibriTTS has 585 hours in English [49], and The People's Speech has over 30,000 hours of speech in English [50]. However, disordered-speech corpora remain far more limited in size and language coverage. In the broader landscape of disordered-speech resources, widely used databases are largely English-speaking and tend to focus on acquired motor speech disorders, particularly dysarthria [6,24]. Large-scale initiatives such as Project Euphonia have demonstrated the value of extensive English disordered-speech collections for improving inclusive speech technologies, reporting a corpus exceeding 1,300 hours and more than one million utterances from more than 1,000 speakers across several adult speech-disorder groups [8]. In parallel, infrastructure efforts such as CLARIN provide structured access to corpora for speech with communication disorders, further illustrating the growing emphasis on resource development [23]. In contrast to these resources, the ArDSS corpus is smaller in scale but contributes a clinically annotated dataset for a less well-represented language and a developmental speech disorder. It combines utterance-level acoustic measurements with structured pronunciation error annotation, offering unique value for analysing DS speech and supporting research and technology development in Arabic.

5.1 Acoustic profiles and inter-speaker variability

Across the analysed subset, timing and acoustic measures showed wide dispersion, reflecting the heterogeneity typical of paediatric clinical speech. The pitch-related measures varied substantially between speakers and productions, consistent with the broad developmental range of the cohort. The high voiced ratio in the analysed subset supports that the F0 estimates were generally derived from predominantly voiced segments, lending confidence to the pitch summaries, while the degree of variation observed across speakers suggests that age alone is unlikely to account for the observed acoustic patterns.

Voice quality measures also exhibited marked variability. The HNR ranged widely across productions, indicating heterogeneous harmonic-to-noise characteristics. The perturbation measures (jitter and shimmer) showed dispersion as well; importantly, these measures were computed only for productions that met the predefined validity constraints of the extraction pipeline. Within the correlation analysis, HNR was negatively associated with perturbation measures, supporting the interpretation that increased noise components tended to occur alongside greater periodicity instability within this corpus. Likewise, the weak association between average F0 and HNR indicates that pitch level alone provides limited insight into harmonic-to-noise characteristics in this population.

Formant-related measures report corpus-wide acoustic characteristics of vowel-related sounds across the collected word list. As the recorded productions were isolated words

rather than controlled sustained vowels, the reported distributions should be interpreted with caution as aggregated estimates across diverse phonetic contexts. Nonetheless, the wide range of formant distributions indicates considerable variability in vowel-related acoustic production across speakers and phonetic environments. These findings also illustrate the need to examine the validity of extracted acoustic characteristics in disordered child speech, as voice irregularities and coarticulation can affect automated tracking.

5.2 Arabic pronunciation error patterns

Phonological analysis showed several repeatable error patterns, broadly in line with what has been reported in the literature and relevant to clinical interpretation. The most common pattern was liquid substitution (92%), especially for /r/, which was often produced as /l/ or /j/. This may reflect the difficulty of sounds that depend on precise tongue timing and fine motor control in individuals with DS. Affricate simplification was also common (87%). Here, affricates were produced without their full affricate quality; for example, /dʒ/ was sometimes realised as a fricative-like or stop-like sound.

Arabic-specific contrasts were affected as well. Emphatic neutralisation (82%) involved producing emphatic consonants as their plain counterparts. Guttural deletion or substitution (69%) pointed to difficulty with posterior constrictions. Word-shape simplification (74%) appeared mainly in longer targets, where the extra phonological length seemed to increase sequencing and planning demands. Overall, the pattern suggests that the children's errors were shaped by both Arabic phonological contrasts and the motor-planning constraints associated with DS.

5.3 Technological implications

The heterogeneity observed in the speech of Arabic-speaking children with DS is evidenced by the distinct acoustic profiles observed across participants. Therefore, assessment and treatment should be individualized and should include the phonatory, articulatory, phonological, and timing-related aspects of the speech production process. The ArDSS corpus can serve as an initial reference point for hypothesis-driven clinical research involving Arabic-speaking children with DS, giving clinicians a more contextually appropriate way to interpret their speech than relying only on norms derived from typical speech.

From a speech-technology perspective, the ArDSS corpus provides useful data for the creation and validation of speech-system algorithms capable of processing atypical child speech. The systematic nature of the observed pronunciation patterns can motivate modelling strategies that incorporate pronunciation variants and speaker-aware adaptation. The corpus may also support the development of computer-assisted systems and pronunciation support tools that focus on empirically frequent error types, complementing services with data-driven practice opportunities.

6 Conclusion

This paper introduces the ArDSS corpus, comprising 5,509 acceptable-quality isolated-word recordings (WAV, 16-bit, 16 kHz) from 13 Arabic-speaking individuals with DS (6 males, 7 females; ages 5–17) in Saudi Arabia, yielding 14.6 hours of raw speech data. The corpus combines XML metadata, phonological error annotation, and utterance-level acoustic features extracted after feature-validity screening. To our knowledge, ArDSS

represents the first annotated Arabic DS speech corpus, documenting substantial inter-speaker variability and pronunciation patterns tied to Arabic phonology.

The development of such corpora is inherently challenging due to the limited availability of participants and the complexity of collecting disordered speech data. This is reflected in the relatively small size of the dataset and its focus on isolated-word recordings from Saudi Arabic speakers. Despite these limitations, the ArDSS corpus provides a valuable resource for analysing pronunciation patterns and speech variability in Arabic DS speech and offers a foundation for advancing research in inclusive speech technologies. The corpus is available for research purposes upon request to the corresponding author.

Future work will focus on extending the corpus to include a larger and more diverse participant population, as well as incorporating connected and conversational speech to better capture real-world communication and dialectal variation.

ACKNOWLEDGEMENTS

The authors would like to thank the participating children and their families for their time and cooperation. We also acknowledge the support of the specialized centres and speech-language pathologists who facilitated participant recruitment and data collection.

References

[1] Chappa, S. (2024). An overview of down syndrome. *International Journal of Current Innovations in Advanced Research, 7*(2), 28–35.

[2] Kent, R. D., & Vorperian, H. K. (2013). Speech impairment in Down syndrome: A review. *Journal of Speech, Language, and Hearing Research, 56*(1), 178–210.

[3] Martin, G. E., Klusek, J., Estigarribia, B., & Roberts, J. E. (2009). Language characteristics of individuals with Down syndrome. *Topics in Language Disorders, 29*(2), 112–132.

[4] Carl, M., Kent, R. D., Levy, E. S., & Whalen, D. H. (2020). Vowel acoustics and speech intelligibility in young adults with Down syndrome. *Journal of Speech, Language, and Hearing Research, 63*(3), 674–687.

[5] Habash, N. Y. (2010). *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

[6] Tobin, J., Nelson, P., MacDonald, B., Heywood, R., Cave, R., Seaver, K., Desjardins, A., Jiang, P.-P., & Green, J. R. (2024). Automatic speech recognition of conversational speech in individuals with disordered speech. *Journal of Speech, Language, and Hearing Research, 67*(11), 4176–4185.

[7] Wald, M., & Bain, K. (2008). Universal access to communication and learning: The role of automatic speech recognition. *Universal Access in the Information Society, 6*(4), 435–447.

[8] MacDonald, R. L., Jiang, P.-P., Cattiau, J., Heywood, R., Cave, R., Seaver, K., Ladewig, M., Tobin, J., Brenner, M. P., Nelson, P. C., Green, J. R., & Tomanek, K. (2021). Disordered speech data collection: Lessons learned at 1 million utterances from Project Euphonia. In *Proceedings of Interspeech 2021* (pp. 4833–4837). ISCA.

[9] Green, J. R., MacDonald, R. L., Jiang, P.-P., Cattiau, J., Heywood, R., Cave, R., Seaver, K., Ladewig, M. A., Tobin, J., Brenner, M. P., Nelson, P. C., & Tomanek, K. (2021). Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. In *Proceedings of Interspeech 2021* (pp. 4778–4782). ISCA.

[10] Martin, A., MacDonald, R. L., Jiang, P.-P., Ladewig, M., Cattiau, J., Heywood, R., Cave, R., Tobin, J., Nelson, P. C., & Tomanek, K. (2025). Project Euphonia: Advancing inclusive speech recognition through expanded data collection and evaluation. *Frontiers in Language Sciences, 4*, 1569448.

[11] Alqudah, A. A. M., Alshraideh, M. A. M., Abushariah, M. A. M., & Sharieh, A. A. S. (2024). Modern Standard Arabic speech disorders corpus for digital speech processing applications. *International Journal of Speech Technology, 27*(1), 157–170.

[12] Ayyad, H., AlBustan, S., & Ayyad, F. (2021). Phonological development in school-aged Kuwaiti Arabic children with Down syndrome: A pilot study. *Journal of Communication Disorders, 93*, 106128.

[13] Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2024). Towards inclusive automatic speech recognition. *Computer Speech & Language, 84*, 101567.

[14] Abdou, S. M., & Moussa, A. M. (2018). Arabic speech recognition: Challenges and state of the art. In N. El Gayar & C. Y. Suen (Eds.), *Computational Linguistics, Speech and Image Processing for Arabic Language* (pp. 1–27). World Scientific.

[15] Alsayadi, H. A., Abdelhamid, A. A., Hegazy, I., Alotaibi, B., & Fayed, Z. T. (2022). Deep investigation of the recent advances in dialectal Arabic speech recognition. *IEEE Access, 10*, 57063–57079.

[16] Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

[17] Mashaqba, B., Abu Sa'aleek, H., Huneety, A., & Al-Shboul, S. (2020). Grammatical number inflection in Arabic-speaking children and young adults with Down syndrome. *South African Journal of Communication Disorders, 67*(1), e1–e7.

[18] Watson, J. C. E. (2002). *The phonology and morphology of Arabic*. Oxford University Press.

[19] Stoel-Gammon, C. (2001). Down syndrome phonology: Developmental patterns and intervention strategies. *Down Syndrome Research and Practice, 7*(3), 93–100.

[20] Bunton, K., Leddy, M., & Miller, J. (2007). Phonetic intelligibility testing in adults with Down syndrome. *Down Syndrome Research and Practice, 12*(1), 1–4.

[21] Bunton, K., & Leddy, M. (2011). An evaluation of articulatory working space area in vowel production of adults with Down syndrome. *Clinical Linguistics & Phonetics, 25*(4), 321–334.

[22] Corrales-Astorgano, M., Escudero-Mancebo, D., & González-Ferreras, C. (2018). Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome. *Speech Communication, 99*, 90–100.

[23] CLARIN ERIC. (2024). *Corpora of disordered speech*. Retrieved from https://www.clarin.eu/resource-families/corpora-disordered-speech

[24] Wan, Y., Sun, M., Kang, X., Li, J., Guo, P., Gao, M., & Wang, S.-J. (2024). CDSD: Chinese Dysarthria Speech Database. In *Proceedings of Interspeech 2024* (pp. 4109–4113). ISCA.

[25] Rowe, H. P., Gutz, S. E., Maffei, M. F., Tomanek, K., & Green, J. R. (2022). Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective. *Frontiers in Computer Science, 4*, 770210.

[26] Rondal, J. A. (1978). Maternal speech to normal and Down's syndrome children matched for mean length of utterance. In C. E. Meyers (Ed.), *Quality of life in severely and profoundly mentally retarded people: Research foundations for improvement*. Washington, DC: American Association on Mental Deficiency.

[27] Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., & Frame, S. (2008). Dysarthric speech database for universal access research. In *Proceedings of Interspeech 2008* (pp. 1741–1744). ISCA.

[28] Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation, 46*(4), 523–541.

[29] Turrisi, R., Braccia, A., Emanuele, M., Giulietti, S., Pugliatti, M., Sensi, M., Fadiga, L., & Badino, L. (2021). EasyCall corpus: A dysarthric speech dataset. In *Proceedings of Interspeech 2021* (pp. 41–45). ISCA.

[30] Kruyt, J., Sabo, R., Polónyiová, K., Ostatníková, D., & Beňuš, Š. (2024). The Slovak autistic and non-autistic child speech corpus: Task-oriented child-adult interactions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 16094–16099). European Language Resources Association.

[31] Escudero-Mancebo, D., Corrales-Astorgano, M., Aguilar-Cuevas, L., González-Ferreras, C., & Cardeñoso-Payo, V. (2022). PRAUTOCAL corpus: A corpus for the study of Down syndrome prosodic aspects. *Language Resources and Evaluation, 56*(1), 191–224.

[32] Shareef, S. R., & Al-Irhayim, Y. F. (2022). Towards developing impairments Arabic speech dataset using deep learning. *Indonesian Journal of Electrical Engineering and Computer Science, 25*(3), 1400–1405.

[33] Alhalabi, A., Abd Albaki, W., & Abu Saad, H. (2025). Arabic voice recognition (Down syndrome children) [Dataset]. Kaggle.

[34] Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics, 71*, 1–15.

[35] Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

[36] de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America, 111*(4), 1917–1930.

[37] Maryn, Y., Corthals, P., De Bodt, M., Van Cauwenberge, P., & Deliyski, D. (2009). Perturbation measures of voice: A comparative study between Multi-Dimensional Voice Program and Praat. *Folia Phoniatrica et Logopaedica, 61*(4), 217–226.

[38] Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research, 50*(6), 1510–1545.

[39] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

[40] Sakib, F. A., Karim, A. R., Khan, S. H., & Rahman, M. M. (2023). Intent detection and slot filling for home assistants: Dataset and analysis for Bangla and Sylheti. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)* (pp. 48–55). Association for Computational Linguistics.

[41] Kolesnyk, A. S., & Khairova, N. F. (2022). Justification for the use of Cohen's kappa statistic in experimental studies of NLP and text mining. *Cybernetics and Systems Analysis, 58*(2), 280–288.

[42] Liqreina, H. A., Jarrar, M., Khalilia, M., El-Shangiti, A. O., & Abdul-Mageed, M. (2023). Arabic fine-grained entity recognition. In *Proceedings of ArabicNLP 2023* (pp. 310–323). Association for Computational Linguistics.

[43] Cleuren, L., Duchateau, J., Ghesquière, P., & Van hamme, H. (2008). Children's oral reading corpus (CHOREC): Description and assessment of annotator agreement. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)* (pp. 998–1005). European Language Resources Association.

[44] Shirai, K., & Fukuoka, T. (2018). JAIST annotated corpus of free conversation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 741–748). European Language Resources Association.

[45] Lyakso, E., Frolova, O., Matveev, A., Matveev, Y., Grigorev, A., Makhnytkina, O., & Ruban, N. (2022). Recognition of the emotional state of children with Down syndrome by video, audio and text modalities: Human and automatic. In *Speech and Computer: 24th International Conference, SPECOM 2022* (pp. 438–450). Springer International Publishing.

[46] Ayyad, H. S., & Bernhardt, B. M. (2022). When liquids and fricatives outrank stops: A Kuwaiti Arabic-speaking child with Down syndrome and protracted phonological development. *Clinical Linguistics & Phonetics, 36*(7), 670–682.

[47] Alzyoudi, M., Albustanji, Y., Khan, A., & Almazroui, K. M. R. (2022). Phonological process in Arabic-speaking children with Down syndrome: A psycholinguistic investigation. *Africa Education Review, 19*(1), 1–14.

[48] Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)* (pp. 4218–4222). European Language Resources Association.

[49] Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., & Wu, Y. (2019). LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In *Proceedings of Interspeech 2019* (pp. 1526–1530). ISCA.

[50] Galvez, D., Diamos, G., Ciro, J., Cerón, J. F., Achorn, K., Gopi, A., Kanter, D., Lam, M., Mazumder, M., & Reddi, V. J. (2021). The People's Speech: A large-scale diverse English speech recognition dataset for commercial usage. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS 2021)*. NeurIPS.