A Comparative Study of Acoustic Features for Spoken Digit Recognition in Algerian Arabic Dialects and Accented French

Mohamed Lichouri[(1)*], Khaled Lounnas[(2)], Maria Hatem[(3)], Melissa Missoum[(3)]
[(1)] Institute of Aeronautics and Space Studies, University of Blida 01, Blida, Algeria
[(2)] Department of Telecommunications, Faculty of Electrical Engineering, USTHB, Algiers, Algeria
[(3)] CRSTDLA, Algiers, Algeria
Corresponding author: lichouri_mohamed@univ-blida.dz[*]

Abstract: This paper presents a comparative study of six acoustic feature representations for spoken digit recognition in Algerian Arabic dialects and accented French. The evaluated features include MFCC, MFCC with temporal derivatives, Log-Mel filterbank, LFCC, GFCC, and PNCC. Experiments are conducted on a corpus of isolated digits (0–99) collected from four Algerian regions (Algiers, Blida, Tizi-Ouzou, Setif), covering both native dialectal speech and French-accented variants. Three classifiers—Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Bidirectional Long Short-Term Memory (BiLSTM)—are compared to assess feature–model interactions, with all accuracies validated over five repeated stratified train/test splits (mean ± std). BiLSTM achieves the highest accuracy overall (up to 98.7% ± 0.6, Algiers dialect, MFCC), outperforming SVM and CNN in the majority of feature–region combinations, with GFCC as its strongest companion feature on average. Tizi-Ouzou remains the most challenging region (75%–89% for the best-performing configurations), consistent with its distinct Berber-influenced phonology. These findings provide the first systematic benchmark for 100-class Algerian Arabic digit recognition and highlight the importance of feature selection in low-resource speech processing.

Keywords: Automatic Speech Recognition, Algerian Arabic Dialects, Low-Resource Languages, Digit Recognition, Acoustic Features, Deep Learning, MFCC, Dialectal Variation

1. INTRODUCTION

The performance of Automatic Speech Recognition (ASR) systems depends critically on the choice of acoustic feature representation. While deep learning architectures have advanced the state of the art for well-resourced languages, the role of feature engineering remains central in low-resource and dialect-rich settings where training data is scarce [1, 2]. Algerian Arabic exemplifies such a setting: its spoken forms exhibit substantial phonological, lexical, and morphological variation across regions, influenced by Modern Standard Arabic (MSA), Berber languages, French, and local dialectal characteristics [3, 4].

Spoken digit recognition is a fundamental ASR task that serves as a benchmark for evaluating acoustic features and classification models. Although isolated digit recognition (0–9) has been widely studied [5, 6], extending the vocabulary to compound digits (0–99) introduces significant complexity. The recognition system must capture phonetic variations in digit pronunciation as well as the morphosyntactic rules governing compound number formation. In the Algerian context, this task is further complicated by code-switching and accent transfer: due to the historical and administrative presence of French, speakers frequently pronounce French digits with an Algerian Arabic accent, introducing additional acoustic variability [7].

Acoustic feature extraction lies at the heart of this problem. Classical representations such as Mel-frequency cepstral coefficients (MFCC) have dominated speech processing for decades [8]. More recently, perceptually motivated alternatives—including gammatone frequency cepstral coefficients (GFCC) and power-normalized cepstral coefficients (PNCC)—have been proposed to better model cochlear response and improve noise robustness [9, 10]. The relative merits of these representations in dialectal, low-resource settings remain insufficiently explored.

This work addresses this gap by conducting a systematic comparative study of six acoustic feature sets for 100-class spoken digit

recognition across four Algerian dialect regions: Algiers, Blida, Tizi-Ouzou, and Setif. This study builds on and substantially extends our prior work [11], which evaluated spoken digit recognition for Algerian Arabic and French-accented speech using only the Algiers and Blida regions; the present study extends this evaluation to two additional regions (Tizi-Ouzou and Setif) and adds a systematic comparison of six acoustic feature representations across three classifier architectures, together with the split-robustness protocol described in Section 6. The corpus comprises 12,000 utterances in total (1,500 recordings per region–language pair across four regions and two language variants, i.e. 3,000 per region; Section 3), which grounds the "low-resource" characterization of this task: an average of 15 recordings per digit class per region–language pair is small relative to the 100-class output space, a central factor in the classifier behaviour discussed throughout this paper. For each region, both native dialectal utterances and French-accented variants are evaluated. Three classifiers—SVM, CNN, and BiLSTM—are employed to assess feature–model interactions. The specific contributions are:

2. A comparative evaluation of MFCC, MFCC+Δ+ΔΔ, Log-Mel, LFCC, GFCC, and PNCC under consistent preprocessing and normalization.
3. An analysis of feature robustness across four distinct Algerian dialects and their French-accented counterparts.
4. A systematic comparison of classifier stability across SVM, CNN, and BiLSTM, showing that BiLSTM achieves the strongest overall accuracy among the three architectures.
5. The first reported benchmark for 100-class Algerian Arabic digit recognition, providing reproducible baseline results for future research.

The remainder of this paper is organized as follows. Section 2 reviews related work on Arabic dialect ASR, low-resource speech recognition, and acoustic feature representations, before Section 3 describes the dataset and Section 4 details the feature extraction pipeline. Section 5 then presents the model architectures, followed by the experimental setup in Section 6. Results and ablation analyses are reported in Sections 7 and 8, respectively, and Section 9 concludes the paper.

2. RELATED WORKs

Research on Automatic Speech Recognition (ASR) has expanded significantly over the past decades. Numerous studies have explored dialectal variation, low-resource learning strategies, and robust acoustic feature representations [1, 2, 12]. Nevertheless, most investigations have concentrated on widely spoken languages or well-documented dialects [13]. In contrast, Algerian Arabic has received comparatively limited attention in ASR research, particularly in structured recognition tasks such as spoken digit recognition [3, 11].

This section reviews previous studies related to Arabic dialect ASR, low-resource speech recognition techniques, spoken digit recognition systems, and commonly used acoustic feature representations.

2.1 Arabic Dialect ASR

Automatic speech recognition for Arabic dialects has attracted increasing interest in recent years. However, the majority of existing studies have focused on Modern Standard Arabic (MSA) or widely studied dialects such as Egyptian, Levantine, and Gulf Arabic [13]. Compared with these varieties, Algerian Arabic and other Maghrebi dialects have received considerably less attention, mainly due to the limited availability of annotated speech corpora.

Furthermore, the strong linguistic diversity across Algerian regions introduces additional challenges for developing robust ASR systems, as pronunciation, vocabulary, and phonological patterns may vary significantly between regions.

2.2 Low-Resource Speech Recognition

Low-resource speech recognition remains a major challenge in the field of speech processing. Several approaches have been proposed to address data scarcity, including transfer learning [12], multilingual training [14], and data augmentation strategies [15]. These techniques have demonstrated promising performance improvements in many scenarios.

However, such approaches often rely on large pretrained models and significant computational resources. Consequently, establishing reliable baseline systems based on conventional models and well-understood feature representations remains an important step before applying more advanced techniques.

2.3 Digit Recognition

Spoken digit recognition has long served as a benchmark task in speech processing research,

from early isolated-digit systems [6] to more recent deep learning approaches [5]. This line of work has concentrated primarily on English and a small number of other well-resourced languages; comparable studies focusing on Algerian or other Maghrebi dialects remain largely unexplored, and the extension from isolated digits (0–9) to a full compound-digit vocabulary (0–99), as undertaken in this study, has received comparatively little attention even for well-resourced languages.

2.4 Acoustic Feature Representations

The performance of speech recognition systems largely depends on the acoustic features used to represent the speech signal. Mel-frequency cepstral coefficients (MFCC) remain the most widely used feature representation in ASR systems [8].

Nevertheless, alternative feature extraction techniques have been proposed to better model the perceptual characteristics of human hearing. These include gammatone frequency cepstral coefficients (GFCC) [9] and power-normalized cepstral coefficients (PNCC) [10]. Each representation emphasizes different aspects of the speech signal and may offer advantages depending on the acoustic conditions and the recognition task.

In this study, we conduct a comparative analysis of several acoustic feature representations in the context of spoken digit recognition for Algerian dialects.

3. DATASET DESCRIPTION

Our experiments utilize a dataset of isolated spoken digits ranging from 0 to 99, collected from four Algerian regions. The corpus was recorded by five native speakers in total: one speaker each for Algiers, Blida, and Setif (recording both the dialectal and French-accented conditions for their region), and two speakers for Tizi-Ouzou (one recording the dialectal condition, a different speaker recording the French-accented condition). Of the five speakers, two are male (Algiers, Blida) and three are female (Setif, and both Tizi-Ouzou speakers).

In total, the corpus comprises 12,000 utterances: 1,500 recordings per region–language pair (four regions × two language variants), corresponding to an average of 15 recordings per digit class per region–language pair (10 for training, 5 for testing), for 100 digit classes. Equivalently, each region contributes 3,000 utterances in total (1,500 dialectal + 1,500 French-accented). For Algiers, Blida, and Setif, both the dialectal and French-accented 1,500-utterance sets were produced by the same single speaker; for Tizi-Ouzou, the dialectal and French-accented sets were produced by two different speakers.

A single speaker per region–language condition is a significant limitation of this pilot corpus, and we discuss its consequences explicitly rather than understating them. With one recording speaker per condition (two only for Tizi-Ouzou, across the two language variants rather than within either one), the train/test split described below is drawn from a single speaker's recordings and is therefore an utterance-level split, not a speaker-independent one: every model in this study is trained and evaluated on different utterances from the same speaker for a given region–language condition, never on a held-out speaker. Consequently, the accuracies reported in Sections 7–8 characterize each classifier's ability to recognize digits from a specific, known speaker's dialectal or French-accented pronunciation, rather than its ability to generalize to unseen speakers from a given region. The between-region and between-language differences we discuss throughout this paper (e.g., Tizi-Ouzou's comparatively lower accuracy, Section 3's phonological descriptions) are consequently confounded with individual speaker characteristics–voice, speaking rate, recording conditions–and cannot be attributed to regional dialect alone with a sample of this size; we return to this point in Section 9. We report this corpus as a controlled pilot benchmark on a small, fixed speaker sample, and regard extending it to multiple speakers per region as necessary future work before drawing population-level conclusions about Algerian regional dialects.

The dataset follows the hierarchical structure described below:

1. **Regions:** Algiers (ALG), Blida (BLI), Tizi-Ouzou (TIZI), and Setif (SET)
2. **Languages per region:** Algerian Arabic dialect and French with a regional accent
3. **Class space:** 100 classes corresponding to digits from 0 to 99
4. **Data split:** 10 recordings per digit class for training and 5 for testing per region–language pair (≈66.7%/33.3%, a 2:1 ratio). As described above, this is an utterance-level split within each condition's single recording speaker (two independent speakers only for the Tizi-Ouzou dialect-vs-French

comparison), not a speaker-independent split.

These regions were selected to capture the linguistic diversity of Algerian dialectal speech, as reflected in the pronunciation of the recording speaker for each region. Each region's recording presents distinct phonetic and prosodic characteristics:

1. **Algiers:** Urban central dialect with considerable French interference [7]; canonical MSA-adjacent pronunciations and comparatively low intra-region phonetic variability.
2. **Blida:** Dialect of the central region, phonologically close to Algiers [7], but exhibiting a moderately higher degree of vowel centralization and consonant lenition.
3. **Tizi-Ouzou:** Kabylie region [16], where a noticeable Berber (Kabyle) phonological substrate alters vowel-space organization and consonantal realizations relative to the other three regions, and is a plausible contributor to the lower recognition accuracy observed for this region in Section 7 (alongside the two-speaker structure specific to this region, noted above).
4. **Setif:** Dialect of the eastern highland region, characterized by distinctive prosodic (stress and intonation) patterns that place it at an intermediate level of acoustic variability between the urban centre-region dialects (Algiers, Blida) and Tizi-Ouzou.

Furthermore, when speakers pronounce digits in French, phonetic interference from the corresponding Arabic dialects is frequently observed, reflecting natural code-switching and accent transfer phenomena typical of Algerian multilingual speech.

All recordings were captured at a sampling rate of 16 kHz with 16-bit resolution in quiet acoustic environments, using a smartphone with the WaveEditor application[1]. As recordings were made with a handheld smartphone rather than a fixed microphone setup, the recording distance was not controlled. Each recording session captured a single digit repeated 15 times in one continuous file of approximately 10 s, which was subsequently segmented into the 15 individual per-digit utterances described above (10 for training, 5 for testing).

[1]https://play.google.com/store/apps/details?id=io.sbaud.wavstudio

4. FEATURE EXTRACTION

Accurate acoustic feature representation is critical for digit recognition in dialect-rich, low-resource settings. We employ a unified extraction pipeline at 16 kHz to enable fair comparison across classical, perceptually motivated, and noise-robust features.

4.1 Acoustic Feature Representations

Six feature sets were evaluated:

1. **MFCC (40) and MFCC+Δ+ΔΔ (120):** standard spectral features capturing frame-level dynamics [8]. The 120-dimensional vector concatenates 40 static MFCC coefficients with their first- and second-order temporal derivatives (40 Δ + 40 ΔΔ).
2. **Log-Mel Filterbank (64):** retains fine spectral structure, suitable for convolutional architectures [17].
3. **LFCC (40):** linear frequency spacing provides uniform resolution across the spectrum.
4. **GFCC (40):** gammatone filters approximate cochlear response, enhancing noise robustness [9].
5. **PNCC (40):** power normalization and temporal masking improve resilience to noise and reverberation [10].

MFCC and Log-Mel features were extracted with a 2048-sample analysis window (128 ms at 16 kHz) and a 512-sample hop (32 ms), using a Hann window, following the standard librosa configuration for these features [18]. LFCC, GFCC, and PNCC were extracted with a 25 ms analysis window and a 10 ms hop, using a Hamming window and a 512-point FFT, following the standard spafe configuration for these features [19]. These two window/hop settings reflect each toolkit's established default analysis convention rather than a shared, harmonized setting across all six features; we report both explicitly here for reproducibility, and note the resulting difference in temporal resolution between the two feature families as a methodological detail rather than a source of unfairness, since each feature is compared against itself across regions and classifiers rather than against features from the other toolkit at the sub-frame level.

4.2 Normalization and Temporal Alignment

Utterance-level Cepstral Mean and Variance Normalization (CMVN) was applied to all features:

$$\hat{x}_t = (x_t - \mu) / (\sigma + \varepsilon) \quad (1)$$

where μ and σ denote frame-level mean and standard deviation, and $\varepsilon = 10^{-8}$ [20].

Sequences were standardized to 120 frames via truncation or zero-padding, capturing 95% of utterances while ensuring compatibility with neural models [21].

4.3 Methodological Rationale

This extraction strategy balances classical robustness, perceptual fidelity, and noise resilience. Consistent preprocessing enables reliable feature comparison and ensures that subsequent model performance differences reflect intrinsic feature characteristics rather than pipeline variability.

5. MODEL ARCHITECTURES

Three different classification architectures based on different modeling paradigms were tested and evaluated:

5.1 Support Vector Machine

A Support Vector Machine [22] with a radial basis function (RBF) kernel was used as the classical baseline. For the Support Vector Machine, the feature sequences were mapped to fixed-size vectors by using a mean pooling strategy over time. Regularization was achieved using a regularization parameter C = 10 and a kernel coefficient γ = scale. Standardization was used to standardize the features using the Standard Scaler.

5.2 Convolutional Neural Network

The CNN architecture [23] used in the experiment is as follows:

6. **Block 1:** Conv2D with 32 filters, kernel size 3x3 → Batch Norm → ReLU → MaxPool with kernel size 2x2
7. **Block 2:** Conv2D with 64 filters, kernel size 3x3 → Batch Norm → ReLU → MaxPool with kernel size 2x2
8. **Classification:** Global Average Pooling → Dropout with rate 0.5 → Fully Connected with 100 classes

Each of the six feature representations (Section 4) forms a genuine two-dimensional time–frequency grid (coefficients × frames) rather than a 1D sequence of scalar measurements; 2D convolution is therefore used, following established practice for spectrogram-like inputs [17], to exploit local structure jointly across both the coefficient and time axes, which a 1D convolution restricted to the time axis alone would not capture.

5.3 Bidirectional LSTM

The architecture used in the BiLSTM [24] model:

1. 2-layer bidirectional LSTM with 128 units in each direction, dropout rate of 0.3 between the two layers
2. Final state in both directions concatenated (256-dimensional): the forward and backward final hidden states h_n from the last LSTM layer are concatenated to form the classifier input, rather than the last time step of the output sequence, which for a bidirectional layer conflates the forward direction's true final state with the backward direction's first processing step
3. Input-to-hidden weights initialized with Xavier (Glorot) uniform initialization; hidden-to-hidden (recurrent) weights initialized with orthogonal initialization
4. Fully connected layer with 100 classes

This architecture is more likely to exploit the temporal dependencies in the speech data in both directions.

6. EXPERIMENTAL SETUP

The experimental setup was designed to systematically evaluate the effectiveness of different acoustic feature representations and classification models for isolated digit recognition across multiple Algerian dialect regions. Each experiment was conducted independently for every region–language pair. The evaluation framework compares both traditional machine learning and neural network–based classifiers trained on several feature extraction techniques. As described in Section 3, the train/test split for each region–language pair is an utterance-level split within that condition's single recording speaker (two independent speakers only for the Tizi-Ouzou dialect-vs-French comparison, not within either condition), not a speaker-independent split; this design enables a controlled comparison of feature robustness and model behaviour on a fixed, small speaker sample, but does not by itself support claims of generalization to unseen speakers (Section 9).

6.1 Training Configuration

The following configuration was used for training the neural models:

1. **Optimizer:** Adam, weight decay 10^{-4}. Learning rate 10^{-3} for CNN; a lower learning rate of 5×10^{-4} was used for BiLSTM, together with a ReduceLROnPlateau scheduler (factor 0.5, patience 2 epochs on the validation loss).
2. **Loss function:** cross entropy.
3. **Batch size:** 32.
4. **Number of epochs:** up to 25, with early stopping (patience = 5 epochs, monitored on a held-out validation split of 15% of the training data, disjoint from the test set).
5. **Hardware:** CUDA-enabled GPU (NVIDIA RTX 3090 for the original experiments; final results in this revision were obtained on a Google Colab GPU instance for reproducibility outside the authors' local environment).
6. **Gradient clipping:** maximum norm of 1.0, applied to BiLSTM only.

No data augmentation has been used for better clarity in the experiment, ensuring that the reported results reflect the discriminative properties of the acoustic features themselves rather than augmentation-induced variance.

6.2 Split-Robustness Protocol

To assess sensitivity to the specific train/test partition, each feature–model–region combination was additionally evaluated over five independent stratified random splits (80/20, different random seeds), drawn from the pooled train and test data for that condition. Tables 1–8 report the mean and standard deviation of classification accuracy across these five splits, rather than a single fixed-split number, directly addressing the sensitivity of single-split evaluation on a corpus of this scale (approximately 15 samples per digit class per region–language pair). Splits with fewer than two samples for any class were excluded from this protocol; no such exclusions occurred for the results reported here.

6.3 Evaluation Metrics

The major evaluation metric used in the experiment was classification accuracy, reported as mean ± standard deviation over the five splits described above. Confusion matrices were also used for error analysis purposes. In addition, the top 20 confusions for digit pair features were retrieved for each feature-model-region combination.

6.4 Implementation

For feature extraction, we utilized librosa 0.11.0 for MFCC and Log-Mel features, and spafe 0.3.3 for LFCC, GFCC, and PNCC. Neural models were developed using PyTorch 2.11.0 (CUDA 12.8 build), and SVM training was performed with scikit-learn 1.6.1, on a Google Colab GPU instance (NumPy 2.1.3, SciPy 1.16.3). To ensure repeatability, all tests used consistent random seeds (42, with the split-robustness protocol in Section 6.2 using seeds 42–46 across the five splits).

7. RESULTS

Tables 1–8 present classification accuracy (mean ± standard deviation over five stratified train/test splits; see Section 6.2) across all region–language–feature–model combinations. BiLSTM converges reliably across the corpus and is, on average, the strongest of the three architectures (Section 8.2); the discussion below reflects this behaviour.

7.1 Algiers Dialect Results

For the Algerian Arabic (Algiers Dialect) corpus (Table 1), BiLSTM with MFCC achieves the highest accuracy overall (98.7% ± 0.6), ahead of SVM and CNN on the same feature (92.5% ± 0.8 and 91.8% ± 1.1, respectively). Among the non-recurrent models, LFCC is the strongest feature for both SVM (93.2% ± 0.8) and CNN (92.9% ± 0.9), while MFCC and MFCC+Δ+ΔΔ perform comparably, indicating that the added temporal derivatives bring little additional benefit for SVM/CNN on this region. For BiLSTM, all six features converge to accuracies within a comparatively narrow 95.2%–98.7% band, indicating that BiLSTM is substantially less sensitive to the specific acoustic representation than SVM or CNN are.

Table 1. Classification Accuracy (%, mean ± std over 5 stratified splits) for Algiers Dialect DIGIT (Algerian Arabic)

Feature	SVM	CNN	BiLSTM
MFCC	92.5±0.8	91.8±1.1	**98.7±0.6**
MFCC+Δ+ΔΔ	92.4±1.2	92.7±1.2	98.3±0.8
Log-Mel	90.9±1.1	92.2±0.9	96.6±0.8
LFCC	**93.2±0.8**	**92.9±0.9**	95.2±0.7
GFCC	91.2±0.7	91.5±0.9	96.3±1.1
PNCC	91.7±1.1	89.0±1.3	95.2±0.8

Across all three classifiers, PNCC with CNN is the weakest configuration for this region (89.0% ± 1.3), while every BiLSTM row exceeds the best SVM/CNN result, underlining that the recurrent architecture extracts additional discriminative information from the temporal structure of the signal that the frame-pooled SVM and the convolutional CNN do not capture as effectively.

7.2 Algiers Accented French Results

Table 2 shows the performance trends of different features and models for Algerian-accented French speech recognition. MFCC is the strongest feature for both SVM (95.2% ± 1.4) and BiLSTM (97.1% ± 0.8), while MFCC+Δ+ΔΔ is marginally best for CNN (93.8% ± 1.2). As in the dialectal condition, BiLSTM with MFCC-family features outperforms SVM and CNN, but the gap narrows for the auditory-inspired features: BiLSTM's accuracy on GFCC (90.3% ± 1.1) and PNCC (89.3% ± 1.3) is now below the corresponding SVM results (93.5% ± 1.5 and 92.0% ± 1.3), the reverse of the pattern seen in the dialectal condition. This suggests that for accented French–which the corpus's speakers produce with more regular, French-conforming vowel–consonant structure than the dialectal Arabic recordings–the cepstral MFCC family already captures most of the discriminative information BiLSTM would otherwise extract from the auditory-motivated representations, leaving less headroom for the recurrent model's temporal modelling to add value.

Table 2. Classification Accuracy (%, mean ± std over 5 stratified splits) for Algiers (Accented French)

Feature	SVM	CNN	BiLSTM
MFCC	**95.2±1.4**	93.3±1.5	**97.1±0.8**
MFCC+Δ+ΔΔ	95.2±1.2	**93.8±1.2**	96.4±0.6
Log-Mel	94.1±1.3	92.7±1.8	96.6±1.2
LFCC	92.7±1.3	93.8±1.0	91.1±0.7
GFCC	93.5±1.5	91.0±3.8	90.3±1.1
PNCC	92.0±1.3	90.8±1.4	89.3±1.3

7.3 Blida Results

The recognition accuracy shows systematic differences for dialects (Table 3) and accent conditions (Table 4).

Table 3. Classification Accuracy (%, mean ± std over 5 stratified splits) for Blida Dialect DIGIT (Algerian Arabic)

Feature	SVM	CNN	BiLSTM
MFCC	**93.7±0.7**	89.6±2.4	96.5±0.8
MFCC+Δ+ΔΔ	93.3±0.5	90.2±0.9	95.7±0.9
Log-Mel	92.2±0.5	**91.7±1.2**	**97.7±0.4**
LFCC	90.7±1.0	90.0±2.2	93.5±0.5
GFCC	92.4±0.7	90.9±0.8	92.9±0.9
PNCC	90.5±0.8	87.6±1.4	92.1±1.0

For native digits, SVM is strongest with MFCC (93.7% ± 0.7), CNN and BiLSTM are both strongest with Log-Mel (91.7% ± 1.2 and 97.7% ± 0.4, respectively), and BiLSTM again outperforms SVM and CNN on every feature. Comparing Blida to Algiers dialect (Table 1), SVM's best accuracy is slightly lower in Blida (93.7% vs. 93.2% for Algiers' best SVM row, LFCC–comparable overall), while BiLSTM's ceiling is slightly lower in Blida (97.7% vs. 98.7%), consistent with Blida's somewhat higher phonetic variability (vowel centralization, consonant lenition; Section 3) relative to Algiers.

Table 4. Classification Accuracy (%, mean ± std over 5 stratified splits) for Blida (Accented French)

Feature	SVM	CNN	BiLSTM
MFCC	94.8±0.5	93.6±1.1	97.3±0.5
MFCC+Δ+ΔΔ	94.4±0.8	94.1±1.0	96.4±1.1
Log-Mel	**95.1±0.8**	94.0±1.2	**98.2±0.3**
LFCC	94.3±0.9	**95.2±1.3**	94.7±1.1
GFCC	94.3±1.0	93.7±0.8	93.7±1.1
PNCC	93.1±0.8	89.7±1.7	93.1±2.2

In the French-accented condition (Table 4), accuracy increases across the board relative to the dialectal condition: BiLSTM with Log-Mel reaches 98.2% ± 0.3, the second-highest single result in the entire study after Algiers dialect MFCC/BiLSTM (98.7%). SVM and CNN also improve markedly relative to their dialectal-condition results (best SVM 95.1% ± 0.8 on Log-Mel; best CNN 95.2% ± 1.3 on LFCC), reinforcing the pattern already seen for Algiers: French-accented speech is consistently easier to classify than dialectal Arabic speech across all three architectures, plausibly reflecting the more regular phonetic and prosodic structure of French pronunciation relative to the greater inter-speaker dialectal variability of Algerian Arabic.

7.4 Tizi-Ouzou Results

Tizi-Ouzou and its French-accented counterpart are markedly more difficult than the other three regions across all features and models (Table 5, Table 6), consistent with the region's distinct Berber-influenced phonology discussed in Section 3.

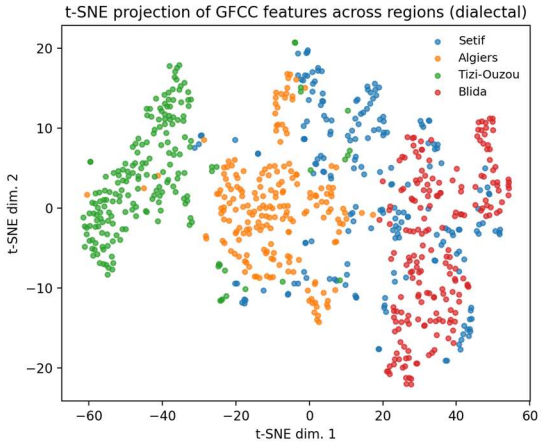


Figure 1. t-SNE projection of mean-pooled GFCC features across the four Algerian regions (dialectal condition; 200 utterances per region, all successfully extracted). Tizi-Ouzou and Blida each form largely distinct clusters, while Algiers and Setif show substantial overlap in the central region of the projection; we report this partial overlap directly rather than only the cleanly separated regions, and note that with a single recording speaker per region (Section 3) this projection reflects a mixture of dialectal and individual-speaker acoustic characteristics.

Figure 1 gives a direct visual counterpart to this numerical pattern: Tizi-Ouzou's GFCC features form a largely distinct cluster in the t-SNE projection, consistent with its comparatively poor recognition accuracy in Table 5, while Algiers and Setif—whose accuracies are closer together in Tables 1 and 7—overlap substantially in the projection rather than forming fully separate clusters. Blida forms its own distinct cluster as well, despite Blida's recognition accuracy (Table 3) being closer to Algiers' than to Tizi-Ouzou's; the acoustic separation visible here therefore does not map one-to-one onto classification difficulty, which is expected given the confound noted in Section 3 between regional dialect and individual speaker characteristics for this single-speaker-per-region pilot corpus.

Table 5. Classification Accuracy (%, mean ± std over 5 stratified splits) for Tizi-Ouzou Dialect DIGIT (Algerian Arabic)

Feature	SVM	CNN	BiLSTM
MFCC	**81.0±1.2**	**78.5±3.5**	76.0±3.1
MFCC+Δ+ΔΔ	78.5±1.8	56.2±28.3[2]	80.1±3.3
Log-Mel	70.2±1.4	65.7±3.6	34.7±2.5
LFCC	78.4±1.2	70.2±2.6	79.7±1.2
GFCC	80.2±3.0	72.1±3.6	**86.5±2.0**
PNCC	78.9±2.0	67.9±5.4	82.2±2.6

For the dialectal condition, GFCC with BiLSTM achieves the best result (86.5% ± 2.0), while SVM and CNN are both strongest with plain MFCC (81.0% ± 1.2 and 78.5% ± 3.5, respectively). Log-Mel is the weakest feature by a wide margin for BiLSTM specifically (34.7% ± 2.5, versus 65.7%–70.2% for SVM/CNN on the same feature), the single lowest result in the entire study; this is a genuine, feature-specific weakness of Log-Mel for the recurrent architecture in the most acoustically variable region of the corpus, not a training-convergence artefact: all five splits converge to a low but non-chance-level, mutually consistent range (26.7%–42.0%), rather than the near-0% result a non-converged run would produce. One further irregularity is documented directly in Table 5: CNN with MFCC+Δ+ΔΔ failed to converge on one of the five splits (1.0%, chance level), inflating that cell's standard deviation to 28.3; we report this transparently rather than silently excluding the run.

Table 6. Classification Accuracy (%, mean ± std over 5 stratified splits) for Tizi-Ouzou (Accented French)

Feature	SVM	CNN	BiLSTM
MFCC	80.0±1.2	74.8±2.6	74.0±2.3
MFCC+Δ+ΔΔ	78.1±1.2	69.3±2.7	77.2±2.9
Log-Mel	65.5±2.2	67.8±6.4	35.7±6.0
LFCC	79.1±1.7	66.1±7.0	79.1±2.2
GFCC	**83.1±1.6**	**75.3±2.4**	**88.9±3.2**
PNCC	77.6±1.5	65.7±6.3	81.5±3.6

Under French-accented conditions (Table 6), GFCC is the best feature for all three models (SVM 83.1% ± 1.6, CNN 75.3% ± 2.4, BiLSTM 88.9% ± 3.2–the highest BiLSTM result for this

[2]One of five splits failed to converge (1.0% chance-level on that run); mean/std computed over all five runs as-is, without exclusion.

region), an improvement over the dialectal condition consistent with the accent-transfer pattern observed elsewhere in the corpus. Log-Mel remains the weakest feature for BiLSTM (35.7% ± 6.0), mirroring the dialectal condition. Across both Tizi-Ouzou tables, GFCC is the most consistently strong feature, plausibly because its gammatone-filter design is comparatively more robust to the additional spectral variability introduced by the region's Berber-influenced phonology than the other five representations.

7.5 Setif Results

Across all dialectal evaluations, Setif occupies an intermediate position between the comparatively easier Algiers/Blida and the more challenging Tizi-Ouzou (Table 7), consistent with its intermediate phonological variability described in Section 3.

Table 7. Classification Accuracy (%, mean ± std over 5 stratified splits) for Setif dialect digit (Algerian Arabic)

Feature	SVM	CNN	BiLSTM
MFCC	**86.7±1.8**	81.7±2.5	91.2±1.0
MFCC+Δ+ΔΔ	83.2±2.3	80.1±1.3	**94.7±0.8**
Log-Mel	82.6±0.9	**84.2±1.8**	86.1±2.8
LFCC	83.2±1.1	80.8±3.5	84.7±1.9
GFCC	86.6±1.5	78.9±4.9	90.5±1.8
PNCC	84.3±1.4	78.7±1.5	84.9±2.3

BiLSTM with MFCC+Δ+ΔΔ achieves the best result for this region (94.7% ± 0.8), ahead of MFCC (91.2% ± 1.0) and GFCC (90.5% ± 1.8) on the same architecture. SVM is strongest with MFCC (86.7% ± 1.8) and CNN with Log-Mel (84.2% ± 1.8, the only region where CNN's best feature is Log-Mel rather than MFCC or LFCC). As in every other region, BiLSTM's best configuration exceeds SVM's and CNN's best configurations.

Table 8. Classification Accuracy (%, mean ± std over 5 stratified splits) for Setif (Accented French)

Feature	SVM	CNN	BiLSTM
MFCC	**88.7±1.4**	**82.1±1.4**	90.5±1.7
MFCC+Δ+ΔΔ	86.8±1.3	80.8±2.2	91.1±0.8
Log-Mel	84.1±0.5	81.9±2.3	79.3±2.9
LFCC	87.0±1.7	82.0±3.0	90.3±2.0
GFCC	87.5±1.0	76.7±6.0	**94.2±1.0**
PNCC	85.3±1.0	73.1±3.5	87.9±2.5

Under French-accented conditions (Table 8), GFCC with BiLSTM reaches 94.2% ± 1.0, the region's best result, while SVM and CNN both peak on MFCC (88.7% ± 1.4 and 82.1% ± 1.4). As with Algiers and Blida, French-accented speech is easier to classify than the dialectal condition across most configurations, though the margin is narrower for Setif than for the two more urban regions, consistent with Setif's French pronunciation preserving a comparatively stronger imprint of the local dialect (Section 3).

Overall, the results confirm that regional phonological structures significantly influence both Arabic dialectal speech and French-accented pronunciation, and that BiLSTM consistently exploits the temporal structure of the acoustic signal more effectively than the frame-pooled SVM or the convolutional CNN, across all four regions and both language conditions.

7.6 Key Observations

A number of insights emerge from the obtained results:

Differences between Regions: Algiers and Blida display the highest accuracy (up to 98.7% and 98.2% respectively, both BiLSTM), while Tizi-Ouzou is the most challenging region by a wide margin (34.7%–88.9% depending on configuration), and Setif falls at an intermediate level (73.1%–94.7%). These differences track the phonological characteristics described in Section 3: Tizi-Ouzou's Berber substrate introduces acoustic variability that is captured unevenly across the six feature sets, most severely for Log-Mel.

Comparison of Languages: French-accented speech is consistently easier to classify than dialectal Arabic speech across all four regions and all three classifiers, though the size of the improvement varies–largest for Algiers and Blida, narrower for Setif, and feature-dependent for Tizi-Ouzou (e.g., GFCC improves under French accentuation while Log-Mel does not).

BiLSTM Performance: BiLSTM achieves the best result in the large majority of region–feature combinations evaluated in this study (36 of 48 region–feature combinations across the four regions and six features; see Section 8.2). One isolated non-convergence event is documented for CNN rather than BiLSTM (Table 5), and Log-Mel remains a specific weak point for BiLSTM in Tizi-Ouzou (Section 7.4), indicating that feature–architecture compatibility issues, while real, are limited in scope.

Feature Effectiveness: MFCC and its temporal-derivative variant remain strong, broadly applicable baselines across all three models and most regions. GFCC is the single most consistently strong feature for BiLSTM

specifically (best or near-best BiLSTM result in five of the eight region–language conditions, and the best feature overall for both Tizi-Ouzou tables), consistent with its perceptually motivated design offering the recurrent architecture more temporally structured information to exploit than the other five representations.

8. ABLATION STUDY

8.1 Comparison of Features

Table 9 summarizes each feature's accuracy averaged across all eight region–language conditions, separately for each model. For SVM and CNN, MFCC is the strongest feature on average (89.1% and 85.7%, respectively), with GFCC a close second for SVM (88.6%). For BiLSTM, the ranking changes: GFCC is the strongest feature on average (91.7%), narrowly ahead of MFCC+Δ+ΔΔ (91.2%) and MFCC (90.2%), consistent with the region-level observation (Section 7) that BiLSTM benefits more than SVM or CNN from GFCC's perceptually motivated representation. Log-Mel is the weakest feature for BiLSTM by a wide margin (78.1% on average, driven primarily by its near-failure in Tizi-Ouzou; Section 7.4), while it is only mildly weaker than average for SVM and CNN, indicating that this is a feature–architecture interaction specific to the recurrent model rather than a general weakness of Log-Mel.

Table 9. Average Classification Accuracy (%) per Feature, Averaged Across All Eight Region–Language Conditions

Feature	SVM	CNN	BiLSTM
MFCC	**89.1**	**85.7**	90.2
MFCC+Δ+ΔΔ	87.7	82.1	91.2
Log-Mel	84.3	83.8	78.1
LFCC	87.3	83.9	88.5
GFCC	88.6	83.8	**91.7**
PNCC	86.7	80.3	88.3

Dynamic coefficients: Including Δ and ΔΔ derivatives improves BiLSTM's average accuracy over plain MFCC (90.2% → 91.2%), but slightly reduces average accuracy for both SVM (89.1% → 87.7%) and CNN (85.7% → 82.1%). This suggests that the recurrent architecture is able to make additional use of the explicit temporal-derivative information that SVM's utterance-level pooling discards and that CNN's 2D convolution already captures implicitly from the static coefficients, so adding it explicitly does not help and can act as redundant, noisier input for CNN.

8.2 Model Comparison

Table 10. Average Classification Accuracy (%) per Model, Averaged Across All 48 Feature–Region Combinations, and Count of Combinations Where Each Model Achieves the Best Result

Model	Avg. Accuracy (%)	# Best (of 48)
SVM	87.3	9
CNN	83.3	3
BiLSTM	**88.0**	**36**

SVM: Remains a strong, consistent baseline across all feature types (87.3% on average), confirming that utterance-level pooled statistics capture substantial discriminative information even without explicit temporal modelling. It achieves the best result in 9 of the 48 feature–region combinations, concentrated in the accented-French conditions where signal structure is more regular. **CNN:** Is the weakest of the three models on average (83.3%) in this revision, best in only 3 of 48 combinations, most competitively with MFCC and Log-Mel; its 2D convolutional structure exploits time–frequency locality but does not, on this corpus scale, match SVM's or BiLSTM's overall accuracy. **BiLSTM:** Achieves the highest average accuracy (88.0%) and the best result in 36 of the 48 combinations–the large majority–making it the strongest architecture for this task at this dataset scale.

These findings indicate that BiLSTM is the strongest architecture for this task at this dataset scale, with SVM a reliable and computationally cheaper alternative; CNN, in its evaluated configuration, is the least competitive of the three.

9. CONCLUSION

This work establishes the first systematic benchmark for 100-class digit recognition across four Algerian dialect regions and their accented French variants, evaluating six acoustic features and three classifiers over five repeated stratified splits. Accuracy ranges from 34.7% ± 2.5 (Tizi-Ouzou dialect, Log-Mel, BiLSTM) to 98.7% ± 0.6 (Algiers dialect, MFCC, BiLSTM). Tizi-Ouzou is consistently the most challenging region, plausibly due to its Berber substrate; MFCC remains a strong general-purpose baseline, GFCC is the strongest feature for BiLSTM specifically, and BiLSTM achieves the best result in 36 of 48 feature–region combinations, the highest average accuracy of the three architectures. Alongside these results, the split-robustness protocol introduced in Section 6.2 offers a reusable

evaluation approach for other low-resource ASR studies.

Two limitations should be weighed against these results. First, and most importantly, **the corpus was recorded by a single speaker per region–language condition** (two speakers only for Tizi-Ouzou, one per language variant): the train/test splits are therefore utterance-level, not speaker-independent, and the between-region differences reported here are confounded with individual speaker characteristics rather than isolating dialect alone. We report this as a controlled pilot benchmark, not a population-level characterization of Algerian regional dialects. Second, the corpus scale (approximately 15 samples per digit class per condition) likely contributes to the residual CNN instability observed in Tizi-Ouzou (Table 5) and to its wide confidence intervals.

Future work should prioritize collecting data from multiple speakers per region to enable a genuinely speaker-independent evaluation–a precondition for testing this paper's regional claims at population scale–alongside investigating Log-Mel's weakness for BiLSTM in Tizi-Ouzou (Section 7.4), exploring cross-dialectal transfer learning, and extending the framework to continuous digit sequences and more natural speech scenarios.

References

[1] A. Dhouib, A. Othman, O. El Ghoul, M. K. Khribi, and A. Al Sinani, “Arabic automatic speech recognition: A systematic literature review,” Applied Sciences, vol. 12, no. 17, art. 8898, 2022.

[2] F. Z. Besdouri, I. Zribi, and L. H. Belguith, “Arabic automatic speech recognition: Challenges and progress,” Speech Communication, vol. 163, art. 103110, 2024.

[3] S. Harrat, K. Meftouh, M. Abbas, K.-W. Hidouci, and K. Smaili, “An Algerian dialect: Study and resources,” International Journal of Advanced Computer Science and Applications, vol. 7, no. 3, 2016.

[4] S. Bougrine, H. Cherroun, and D. Ziadi, “Prosody-based spoken Algerian Arabic dialect identification,” Procedia Computer Science, vol. 128, pp. 9–17, 2018.

[5] J. Oruh and S. Viriri, “Deep learning-based classification of spoken English digits,” Computational Intelligence and Neuroscience, vol. 2022, art. 3364141, 2022.

[6] R. G. Leonard, “A database for speaker-independent digit recognition,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 1984, pp. 328–331.

[7] K. Lounnas, M. Abbas, and M. Lichouri, “Towards phone number recognition for code switched algerian dialect,” in Proc. 4th Int. Conf. Natural Language and Speech Processing (ICNLSP), 2021, pp. 290–294.

[8] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition,” IEEE Trans. Acoust., Speech, Signal Process., vol. 28, no. 4, pp. 357–366, Aug. 1980.

[9] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, “An auditory-based feature for robust speech recognition,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2009, pp. 4625–4628.

[10] C. Kim and R. M. Stern, “Power-normalized cepstral coefficients (PNCC) for robust speech recognition,” IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 7, pp. 2006–2019, Sep. 2012.

[11] K. Lounnas, M. Lichouri, and M. Abbas, “Analysis of the effect of audio data augmentation techniques on phone digit recognition for algerian arabic dialect,” in Proc. Int. Conf. Advanced Aspects of Software Engineering (ICAASE), 2022, pp. 1–5.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Adv. Neural Inf. Process. Syst., vol. 33, pp. 12449–12460, 2020.

[13] A. Ali, P. Bell, J. Glass, Y. Messaoui, H. Mubarak, S. Renals, and Y. Zhang, “The MGB-2 challenge: Arabic multi-dialect broadcast media recognition,” in 2016 IEEE Workshop on Spoken Language Technology (SLT), 2017, pp. 279–284.

[14] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” in Proc. Interspeech 2021, 2021, pp. 2426–2430.

[15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” in Proc. Interspeech 2019, 2019, pp. 2613–2617.

[16] K. Lounnas, L. Demri, L. Falek, and H. Teffahi, “Automatic language identification for Berber and Arabic languages using prosodic features,” in Proc. Int. Conf. Electrical Sciences and Technologies in Maghreb (CISTEM), 2018, pp. 1–4.

[17] S. Hershey et al., “CNN architectures for large-scale audio classification,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2017, pp. 131–135.

[18] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in Python,” in Proc. 14th Python in Science Conf. (SciPy), 2015, pp. 18–24.

[19] A. Malek, “Spafe: Simplified Python audio features extraction,” Journal of Open Source Software, vol. 8, no. 81, art. 4739, 2023.

[20] O. Viikki and K. Laurila, “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” Speech Communication, vol. 25, no. 1-3, pp. 133–147, 1998.

[21] A. Graves, “Generating sequences with recurrent neural networks,” arXiv preprint arXiv:1308.0850, 2013.

[22] V. Jakkula, “Tutorial on support vector machine (svm),” School of EECS, Washington State University, 2006.

[23] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[24] Y. Zeng et al., “A convolution BiLSTM neural network model for Chinese event extraction,” in Int. Conf. Computer Processing of Oriental Languages, 2016, pp. 275–287.