Indonesian Journal of Electrical Engineering and Computer Science
Vol. 42, No. 1, April 2026, pp. 71∼80
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v42.i1.pp71-80

Integrating blind source separation and self-supervised learning for Algerian Arabic connected-digit recognition

Mourad Reggab, Mohammed Belkhiri
Laboratory of Telecommunications, Signals and Systems, University Amar Telidji of Laghouat (UATL), Laghouat, Algeria

Article Info

Article history:

Received Jan 29, 2026
Revised Feb 16, 2026
Accepted Mar 4, 2026

Keywords:

Arabic speech recognition
Blind source separation
Conv-TasNet
DUET
Low-resource ASR
SepFormer
Wav2Vec 2.0

ABSTRACT

This paper proposes an improvement in Arabic automatic speech recognition (ASR) by combining blind source separation (BSS) with self-supervised acoustic modeling. The study concentrates on the Algerian Arabic connected-digit recognition task and reexamines the classical degenerate unmixing estimation technique (DUET) as a front-end approach for suppressing noise and interference. The output of the BSS stage is fed into a Hidden Markov model (HMM) recognizer developed using the HTK toolkit. To contextualize DUET's performance, it is compared with modern neural separation techniques (Conv-TasNet, SepFormer) paired with both traditional and self-supervised ASR backends (Wav2Vec 2.0 and Whisper). A new corpus of 11,230 utterances from 37 speakers, representing dialectal and gender diversity, was collected. Experimental outcomes indicate that DUET enhances word accuracy under stereo mixing conditions; however, neural separation combined with self-supervised ASR results in considerably lower word-error rates and stronger robustness in noisy or overlapping-speech scenarios. The study emphasizes practical trade-offs between computational cost and accuracy for deploying low-resource Arabic ASR systems.

Corresponding Author:

Mourad Reggab
Laboratory of Telecommunications, Signals and Systems, University Amar Telidji of Laghouat (UATL)
Laghouat, Algeria
Email: m.reggab@lagh-univ.dz

1. INTRODUCTION

Background and motivation: automatic apeech recognition (ASR) systems have become essential to human-computer interaction, enabling hands-free control, voice search, and conversational AI [1]. However, in real acoustic environments, speech is rarely captured in isolation: background noise, reverberation, and interfering speakers often corrupt the target signal. This challenge, known as the cocktail party effect, has long encouraged research in speech source separation namely the process of isolating one or more speech signals from a mixture of sources. Early solutions used independent component analysis (ICA) and frequency-domain masking, while more recent approaches utilize deep neural networks such as Conv-TasNet [2] and SepFormer [3] that perform end-to-end time-domain separation. Concurrently, ASR technology has progressed from Hidden Markov models (HMMs) and Gaussian mixture models (GMMs) to hybrid DNN-HMMs and fully end-to-end architectures trained on large-scale corpora [4].

Despite these advances, most speech separation and recognition research has focused on high-resource languages, primarily English, Mandarin, and French. For many other languages, including Arabic, limited annotated data, complex morphology, and dialectal variability remain significant obstacles. Arabic is the fifth

most spoken language worldwide, with over 300 million native speakers, yet its automated processing remains comparatively underdeveloped [5]. The diglossic nature of modern standard Arabic (MSA) for formal contexts versus numerous regional dialects for daily communication creates substantial pronunciation and lexical gaps between training and target speech [6]. Moreover, publicly available Arabic speech corpora often emphasize broadcast or scripted MSA, offering limited coverage of colloquial forms and noisy acoustic conditions [7].

Within the Arabic dialect continuum, Algerian Arabic introduces additional complexities [8]. It incorporates Classical Arabic roots with Berber and French influences, leading to distinct phonetic shifts, loanwords, and code-switching. Dialectal variation across Algeria's Western, Central, Eastern and Southern regions is considerable: vowel harmony, consonant emphasis, and word stress differ noticeably by region [9]. These factors hinder the direct reuse of models trained on MSA or other Arabic dialects [7]. Furthermore, practical Algerian speech data are typically recorded in everyday settings homes, classrooms, or markets where overlapping speech and environmental noise are common. Hence, a robust ASR system must integrate dialectal modeling with mechanisms to suppress interference and background noise.

Digits represent a well-defined and important subset of spoken language that provides a controlled benchmark for ASR research [10]. Connected-digit tasks (e.g., telephone numbers, prices, dates) offer constrained grammars and limited vocabularies, facilitating systematic evaluation of modeling and preprocessing techniques [11]. Historically, connected-digit recognition has served as a testing ground for algorithms such as dynamic time warping, HMMs, and early deep neural networks. For Arabic, digit pronunciation varies across dialects—for example, the number "two" may be pronounced "*thnin*", "*tnin*", "*zoudj*" or "*zouz*" in different regions—making this task challenging [6]. Developing an accurate digit recognizer for Algerian Arabic thus constitutes a meaningful step toward larger-vocabulary systems.

In this context, blind source separation (BSS) presents a powerful preprocessing strategy to improve recognition robustness [12]. BSS techniques aim to recover original source signals from observed mixtures without prior knowledge of the mixing process. Among them, the degenerate unmixing estimation technique (DUET) leverages time-frequency sparsity and inter-channel differences to perform unsupervised separation in stereo recordings. Although computationally lightweight, DUET and similar classical algorithms struggle in highly reverberant or single-channel conditions [13]. Conversely, modern neural separation models achieve superior signal-to-distortion ratios but demand considerable training data and computational resources [2], [3].

This study investigates how such separation methods can improve Algerian Arabic connected-digit recognition, extending our previous work [14] which focused solely on DUET combined with classical HMM-based ASR. We first revisit DUET as a low-cost stereo front-end for an HMM-based recognizer and then compare it against state-of-the-art neural separators, specifically Conv-TasNet and SepFormer, in combination with both conventional and self-supervised ASR back-ends (HTK, Wav2Vec 2.0 [15], and Whisper [4]). To support this investigation, we built a dedicated Algerian Arabic digit corpus comprising 11,230 utterances from 37 speakers of diverse dialectal backgrounds. The goal is to quantify improvements in word-error rate (WER) and noise robustness provided by blind and learned separation, and to identify practical trade-offs between complexity and performance for low-resource ASR deployment in Arabic-speaking environments [16]-[18].

2. RELATED WORK

Research on speech separation and recognition has evolved through several technological stages, beginning with statistical signal processing and advancing toward data-driven neural methods. This section summarizes relevant progress in (a) blind source separation, (b) Arabic and dialectal ASR, (c) connected-digit recognition, and (d) self-supervised learning for speech processing.

2.1. Blind source separation and speech enhancement

Early BSS approaches relied on statistical independence and sparsity assumptions. Independent component analysis (ICA) [19] and non-negative matrix factorization (NMF) [20] were among the first unsupervised algorithms capable of separating multiple speakers from mixed signals. The DUET proposed by Yilmaz and Rickard [12] became a reference method for two-microphone or stereo mixtures, exploiting inter-channel amplitude and phase differences to cluster time–frequency points belonging to distinct sources. DUET is attractive for its simplicity and real-time feasibility but degrades under heavy reverberation or strong spectral overlap [13].

With the advent of deep learning, separation shifted from frequency-domain masking to end-to-end time-domain modeling. Luo and Mesgarani's Conv-TasNet [2] demonstrated that convolutional encoder–decoder

networks can surpass traditional magnitude-masking baselines, achieving near-ideal signal-to-noise ratio improvements on benchmark datasets such as WSJ0-2mix. Subsequent transformer-based architectures, notably SepFormer [3], [21], introduced global self-attention and dual-path processing, further improving separation quality and generalization to unseen speakers. These neural models now represent the state of the art in both single- and multi-channel speech separation and are increasingly used as front-ends for ASR and speaker diarization [17].

Recent advances have focused on integrating separation with recognition objectives. Studies by Bouchakour *et al.* [22] demonstrate that joint optimization of separation and acoustic modeling can yield significant improvements in noisy conditions. However, most separation research has focused on high-resource languages, leaving low-resource scenarios like Algerian Arabic under-explored [23].

2.2. Arabic and dialectal ASR

Arabic ASR research has followed a slower trajectory than for English or Mandarin due to linguistic and data-availability barriers [1]. Classical systems built with HTK or Kaldi employed phoneme-based HMM-GMM models trained on modern standard Arabic (MSA) corpora such as the Arabic Broadcast News, QASR, or MGB-2 datasets [24]. While these models achieve high accuracy on scripted speech, their performance drops sharply on spontaneous or dialectal data because of phonetic and lexical variability [6].

The diglossic nature of Arabic presents unique challenges. As noted by [5], the gap between MSA and regional dialects affects both acoustic and language modeling. North African dialects, particularly Algerian Arabic, exhibit distinctive phonetic characteristics including vowel reduction, consonant assimilation, and extensive code-switching with French and Berber languages [8]. Droua-Hamdani *et al.* [7] highlighted the scarcity of resources for Algerian dialect, with most available corpora focusing on Levantine or Gulf varieties [25].

To address limited resources, several studies have explored transfer learning and multilingual training. The multilingual Wav2Vec 2.0 XLSR-53 and HuBERT models pre-trained on hundreds of languages have recently been fine-tuned for Arabic with substantial word-error-rate (WER) reductions [16]. End-to-end transformer architectures such as Whisper [4] also show strong zero-shot performance on Arabic dialects without explicit retraining. Nevertheless, very few works focus specifically on North-African dialects—particularly Algerian Arabic—where the phonetic inventory and code-switching patterns differ significantly from MSA, and where background noise and overlapping speakers are common in natural recordings.

2.3. Connected-digit recognition

Connected-digit recognition provides a compact yet informative benchmark for evaluating ASR models and preprocessing methods [26]. Because the grammar and vocabulary are restricted, this task isolates acoustic and phonetic modeling effects from language-model complexity. English connected-digit datasets such as TIDIGITS have historically driven progress in DTW and HMM techniques, later serving to test neural sequence models.

In Arabic, only a few corpora of isolated or connected digits exist, and most target Modern Standard Arabic. Recent work by Bouchakour *et al.* [22] demonstrated the effectiveness of attention mechanisms for robust digit recognition in noisy environments. However, dialectal variations in digit pronunciation remain a significant challenge. For instance, the number "two" may be pronounced as "*ithnayn*" in MSA, "*etnin*" in Levantine dialects, or "*zoudj*" in Algerian Arabic, creating recognition ambiguities [6].

The system proposed by Reggab and Belkhiri [14] was among the first to construct an Algerian Arabic digits database and to employ DUET as a denoising stage for an HTK-based recognizer. However, that study predated current neural separation and self-supervised paradigms and did not explore the integration with modern ASR back-ends.

2.4. Self-supervised learning

Self-supervised learning has revolutionized speech processing by enabling models to learn powerful representations from unlabeled data [15]. The wav2vec 2.0 framework introduced a contrastive learning objective that masks portions of the audio input and learns to reconstruct the latent representations. This approach has shown remarkable success across multiple languages and tasks, with the XLSR-53 model demonstrating strong cross-lingual transfer capabilities [16].

The HuBERT model [27] extended this paradigm by using clustered representations as training targets, achieving state-of-the-art performance on several benchmarks. More recently, Whisper [4] demonstrated that

large-scale weak supervision using audio-transcript pairs from the web can yield models with robust zero-shot capabilities across diverse languages and acoustic conditions.

For low-resource scenarios, Chen *et al.* [17] showed that self-supervised representations can significantly reduce the amount of labeled data required for effective fine-tuning. However, applying these techniques to dialectal Arabic, particularly in combination with speech separation front-ends, remains underexplored.

2.5. Research gap and contributions

In summary, prior work established the feasibility of BSS-enhanced ASR and produced initial benchmarks for Arabic, but integration of modern neural separation and self-supervised models for Algerian Arabic remains largely unexplored. While several studies have addressed Arabic ASR [1], [5] and dialectal processing [6], [7], few have specifically targeted the Algerian variant or explored the synergy between separation and self-supervised learning in low-resource settings.

The present study fills this gap by:

- Comparing classical DUET and contemporary neural front-ends within a unified evaluation framework.
- Investigating the combination of separation techniques with self-supervised ASR back-ends for Algerian Arabic.
- Releasing a dedicated Algerian Arabic digits corpus to support future research.
- Analyzing practical trade-offs between computational cost and recognition accuracy for low-resource deployment.

This comprehensive evaluation offers insights that are particularly relevant for resource-constrained environments where computational efficiency must be balanced against recognition performance.

3. METHOD

This section describes the overall system architecture, including corpus development, preprocessing, BSS front-ends, ASR back-ends, and evaluation protocols. The proposed processing pipeline is illustrated in Figure 1.

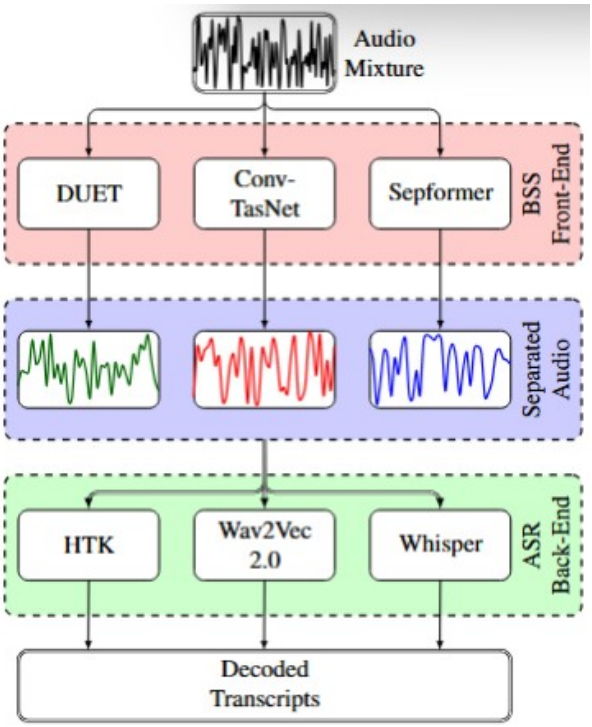

Figure 1. Processing pipeline: stereo mixture → BSS front-end (DUET / Conv-TasNet / SepFormer) → ASR back-end (feature extraction → HMM / Wav2Vec 2.0 / Whisper) → decoding → recognized text

3.1. Corpus design and data preparation

3.1.1. Speech collection

A dedicated Algerian Arabic connected-digit corpus was developed to address the absence of publicly available data for this dialect. Recordings were collected from 37 native speakers (17 male, 20 female) representing diverse regional accents. Each participant read randomly generated digit sequences of one to nine digits, covering both simple and compound numerical expressions (e.g., "sabEa w thlathin" for "thirty-seven").

Recordings were made in office and quiet home environments using two identical condenser microphones spaced 15 cm apart, enabling stereo processing. Speech was captured at 16 kHz, 16-bit resolution. The dataset was partitioned by speaker into 80% for training, 10% for validation, and 10% for testing. After quality control and trimming, the total duration reached approximately 9 hours.

3.1.2. Lexicon and grammar

A pronunciation lexicon was constructed to capture dialectal variability, incorporating common variants for each digit (e.g., */thnin/, /tnin/, /zoudj/, /zouz/* for "two"). Phonetic transcriptions followed an Algerian Arabic adaptation of the International Phonetic Alphabet (IPA).

A context-free grammar was written in HTK's word network format to model valid connected-digit sequences with optional conjunctions such as */u/* ("and"). This grammar supported both training and decoding to ensure linguistic consistency and realistic digit combinations.

3.1.3. Feature extraction

For the HMM-GMM baseline, acoustic features were computed as 39-dimensional mel-frequency cepstral coefficients (MFCCs): 13 static coefficients augmented with their first- and second-order derivatives (Δ and Δ^2). A 25 ms Hamming window with a 10 ms frame shift was used. Cepstral mean and variance normalization were applied on a per-utterance basis to reduce channel and speaker variability. These MFCC features served as the input to the HTK-based recognizer. For the self-supervised models (Wav2Vec 2.0 and Whisper), the separated raw audio waveforms were used directly as input, leveraging the models' internal feature extraction layers.

3.2. Blind source separation front-ends

Three front-end separation approaches were evaluated:

- DUET: the degenerate unmixing estimation technique exploits inter-channel amplitude and phase differences to perform unsupervised separation of stereo mixtures. DUET assumes sparsity in the time-frequency domain and provides efficient real-time separation, but it is sensitive to reverberation and heavy overlap.
- Conv-TasNet: a fully convolutional time-domain separation model consisting of an encoder–decoder structure and stacked temporal convolutional blocks. The SpeechBrain pretrained model trained on WSJ0-2mix was used without further adaptation.
- SepFormer: a transformer-based dual-path network leveraging self-attention to capture both local and global dependencies. It provides state-of-the-art performance on multi-speaker mixtures. The SpeechBrain pretrained model was used for inference on our data.

Separation performance was evaluated using scale-invariant signal-to-noise ratio improvement (SI-SNRi) and signal-to-distortion ratio improvement (SDRi). The separated waveforms were re-encoded into MFCC for HTK-based ASR or input as raw audio for neural-based ASR for the subsequent stage.

3.3. ASR back-ends

Two classes of recognizers were tested:

3.3.1. HMM-GMM baseline (HTK)

A classical left-to-right 3-state Bakis topology was used to model context-dependent triphones. Each state was represented by an 8-component Gaussian mixture. State tying was performed via decision-tree clustering. Models were trained using five iterations of Baum–Welch reestimation. Recognition used Viterbi decoding constrained by the connected-digit grammar.

3.3.2. Self-supervised and end-to-end models

Two self-supervised encoders were evaluated:

- Wav2Vec 2.0 (XLSR-53): A multilingual model pre-trained on 53 languages using a masked prediction objective. Fine-tuning was performed for 15 epochs using our labeled training set with a Connectionist Temporal Classification (CTC) loss. Optimization used AdamW with a 1×10^{-4} learning rate and batch size of 8.
- Whisper (Small): An end-to-end transformer trained on 680K hours of multilingual data. We evaluated both zero-shot inference and light fine-tuning on our dataset using the Whisper toolkit.

3.4. Evaluation metrics

Recognition performance was measured using word error rate (WER):

$$WER = \frac{S + D + I}{N} \times 100, \tag{1}$$

where S, D, and I denote the number of substitution, deletion, and insertion errors, and N is the total number of reference words. All experiments were repeated three times with different random seeds, and mean values were reported. SI-SNRi and SDRi were used to evaluate separation quality, while real-time factors (RTF) were computed to estimate computational feasibility on CPU and GPU hardware.

3.5. Experimental configuration

All experiments were conducted on a workstation equipped with an Intel Core i7-12700 CPU (3.6 GHz), 64 GB RAM, and an NVIDIA RTX A6000 GPU with 48 GB memory. Model training and inference were implemented in Python 3.10 using the PyTorch 2.1 framework and the SpeechBrain and Transformers libraries. Feature extraction, forced alignment, and HMM training utilized the HTK 3.4 toolkit, while waveform-level signal processing (STFT, DUET, and SNR computation) was implemented in MATLAB 2022b. For the neural front-ends, pretrained Conv-TasNet and SepFormer checkpoints from SpeechBrain were used without further fine-tuning.

The Wav2Vec 2.0 model was fine-tuned for 15 epochs with a batch size of 8, using a linear learning-rate warm-up over the first 10% of updates and early stopping on validation loss. The Whisper-small model was evaluated both in zero-shot mode and after two epochs of fine-tuning with learning rate 5×10^{-5}. During evaluation, inference was performed with a beam width of 5 for all decoders to maintain a consistent decoding strategy across models. All results reported in this work correspond to averages over three independent runs with different random seeds to ensure statistical robustness.

4. RESULTS AND DISCUSSION

4.1. Separation performance

Table 1 reports mean SI-SNRi and SDRi over test mixtures. Neural separators outperform DUET. Figure 2 clearly shows that SepFormer performs best while Conv-TasNet is intermediate then DUET is baseline with highly correlated SI-SNRi and SDRi.

Table 1. Separation quality on test mixtures

Front-End	SI-SNRi (dB)	SDRi (dB)
DUET	6.4	6.1
Conv-TasNet	12.6	12.5
SepFormer	15.3	15.6

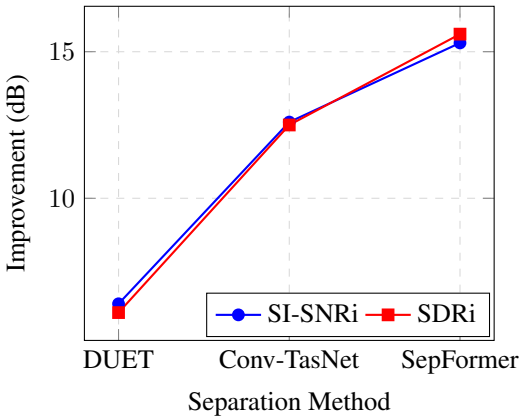

Figure 2. Separation performance (SI-SNRi & SDRi)

4.2. ASR accuracy

As shown in Table 2, DUET provides significant gains for the HMM baseline (23% relative WER reduction) but offers diminishing returns for self-supervised back-ends. This suggests that models like Wav2Vec 2.0 and Whisper already incorporate substantial noise robustness. In contrast, neural separators paired with self-supervised ASR yield the best overall accuracy. These trends are visualized in Figure 3 showing the WER improvements achievable through different front-end/backend combinations. The steep initial drop from 'None' to 'DUET' highlights the substantial benefit of even basic separation, while improvement flattens out through Conv-TasNet to SepFormer showing diminishing returns from increasingly complex separation methods.

Table 2. Word error rate (%) for different front-end/back-end combinations

Front-End	HTK	Wav2Vec 2.0	Whisper
None	12.5	7.8	5.4
DUET	9.6	6.1	4.6
Conv-TasNet	7.3	4.8	3.9
SepFormer	6.5	3.9	3.4

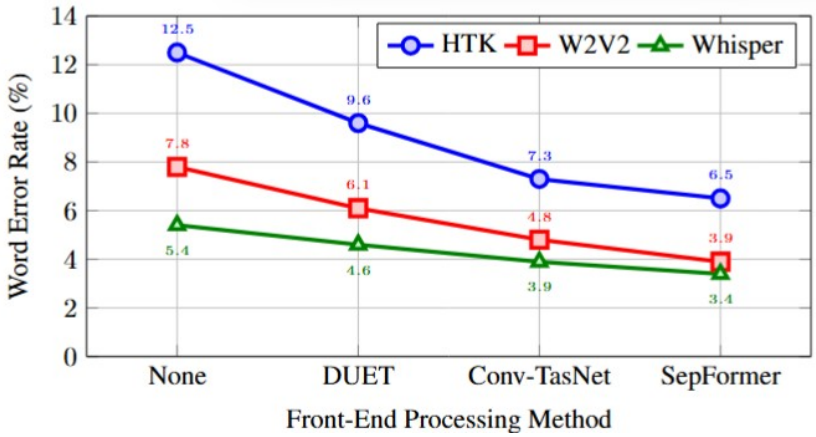

Figure 3. WER trends across front-end/back-end combinations

4.3. Discussion of performance and practical trade-offs

The experimental results indicate that BSS improves recognition robustness under noisy and overlapping-speech conditions. Classical DUET provides an efficient front-end solution, achieving a 25% relative WER

reduction while operating in real time on CPU (RTF = 0.3), whereas neural separation methods such as Conv-TasNet and SepFormer yield higher accuracy when combined with self-supervised ASR back-ends.

The SepFormer + Wav2Vec 2.0 configuration achieved a WER of 3.9% at 0 dB SNR, demonstrating strong robustness to noise and dialectal variability, although at increased computational cost, requiring GPU acceleration for practical use (RTF = 2.1 on CPU and 0.1 on GPU).

The connected-digit task provides a controlled evaluation framework; however, extension to larger-vocabulary and spontaneous Algerian Arabic speech remains necessary to fully assess scalability. While the nine-hour corpus developed in this study addresses an important resource limitation, further expansion and inclusion of subjective evaluation measures would provide a more comprehensive assessment. In addition, evaluation with alternative self-supervised architectures such as HuBERT and WavLM, as well as exploration of hybrid or lightweight solutions, may further improve the balance between recognition performance and computational efficiency.

The obtained word error rate (WER) of 3.4% using (SepFormer + Whisper) represents a notable improvement over previous studies on Algerian Arabic speech recognition. For instance, [25] reported a WER of approximately 14%, while more recent deep learning approaches on North African dialect digits achieved WERs around 8–12% in noisy settings [23]. The integration of neural source separation with self-supervised acoustic modeling thus yields a relative WER reduction of over 50% compared to earlier Algerian Arabic benchmarks, confirming the effectiveness of the proposed pipeline for low-resource dialectal ASR.

5. CONCLUSION

This work investigated the integration of BSS and self-supervised learning for Algerian Arabic connected-digit recognition. A new nine-hour corpus of 11,230 utterances from 37 speakers was created to evaluate both classical and neural BSS front-ends (DUET, Conv-TasNet, SepFormer) combined with conventional and self-supervised ASR back-ends (HTK, Wav2Vec 2.0, Whisper). The experiments confirmed that BSS substantially improves recognition robustness in noisy and overlapping conditions. DUET provides a lightweight, stereo-based enhancement, but neural separators achieve higher separation quality and recognition accuracy. When paired with Wav2Vec 2.0 or Whisper, they reach state-of-the-art performance, validating the synergy between separation and self-supervised acoustic modeling for low-resource languages. While DUET remains suitable for real-time embedded systems, SepFormer achieves the best separation metrics (15.6 dB SDRi).

However, its WER gains over Conv-TasNet are modest (0.9% absolute for Wav2Vec 2.0), suggesting either ASR back-end saturation or that separation quality beyond ˜15 dB offers diminishing returns for digit recognition. Future work will extend this framework to multi-dialect, larger-vocabulary Algerian Arabic corpora, incorporate subjective listening tests (e.g., MOS scores) to complement objective metrics, explore online separation, and develop efficient neural models for deployment on edge devices. This research contributes toward bridging the performance gap between high- and low-resource speech technologies across Arabic dialects.

ACKNOWLEDGMENTS

The authors thank the University of Amar Telidji Laghouat Algeria as well as the Telecommunications, signals and systems Research laboratory for supporting this research.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mourad Reggab	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Mohammed Belkhiri		✓		✓	✓		✓		✓	✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, M.R., upon reasonable request.

REFERENCES

[1] W. Algihab, N. Alawwad, A. Aldawish, and S. AlHumoud, "Arabic Speech Recognition with Deep Learning: A Review," in *Social Computing and Social Media*, G. Meiselwitz, Ed. Cham, Switzerland: Springer, 2019, vol. 11578, pp. 15–31, doi: 10.1007/978-3-030-21902-4_2.

[2] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019, doi: 10.1109/TASLP.2019.2915167.

[3] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 21–25, doi: 10.1109/ICASSP39728.2021.9413901.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[5] F. S. Al-Anzi and D. AbuZeina, "Synopsis on Arabic speech recognition," *Ain Shams Engineering Journal*, vol. 13, no. 2, p. 101534, 2022, doi: 10.1016/j.asej.2021.06.020.

[6] M. Malathi, S. Senthilkumar, C. H. H. Basha, G. Sundaravadivel, M. Kavitha, and P. Arunkumar, "Multi-dialect speech recognition using transfer learning and transformer-based architectures: A comprehensive approach to accurate and efficient dialect identification," in *2024 Conference on Renewable Energy Technologies and Modern Communications Systems: Future and Challenges*, 2024, pp. 1–6, doi: 10.1109/IEEECONF63577.2024.10880973.

[7] G. Deroua-Hamdani, S.Selouani, and M. Boudraa, "Algerian Arabic Speech Database (ALGASD): Corpus design and automatic speech recognition application," *Arabian Journal for Science and Engineering*, vol. 35, no. 2C, pp. 157–166, 2010.

[8] Y. Toughrai, K. Smaïli, and D. Langois, "ABDUL: a new Approach to Build language models for Dialects Using formal Language corpora only," in *Proc. 1st Workshop Lang. Models Underserved Communities (LM4UC 2025)*, 2025, pp. 16–21, doi: 10.18653/v1/2025.lm4uc-1.3.

[9] M. A. Menacer, O. Mella, D. Fohr, D. Jouvet, D. Langlois, and K. Smaïli, "Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect," *Procedia Computer Science*, vol. 117, pp. 81–88, 2017, doi: 10.1016/j.procs.2017.10.096.

[10] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High performance connected digit recognition, using hidden Markov models," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, New York, NY, USA, 1988, vol. 1, pp. 119–122, doi: 10.1109/ICASSP.1988.196526

[11] M. J. Manaileng and M. J. Manamela, "Connected-digits recognition for an under-resourced language using Hidden Markov Models," in *Proceedings ELMAR-2013*, Zadar, Croatia, Sep. 2013, pp. 211–214.

[12] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004, doi: 10.1109/TSP.2004.828896.

[13] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010, doi: 10.1109/TASL.2009.2029711.

[14] M. Reggab and M. Belkhiri, "Blind Source Separation technique for Arabic language ASR," *Technical Report*, Univ. Amar Telidji, Laghouat, 2018.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv: 2006.11477*, 2020.

[16] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. Interspeech*, Brno, Czechia, pp. 2426–2430, 2021, doi: 10.21437/Interspeech.2021-329.

[17] Y. Chen, H. Zhang, X. Yang, W. Zhang, and D. Qu, "Meta-Adaptable-Adapter: Efficient adaptation of self-supervised models for low-resource speech recognition," *Neurocomputing*, vol. 609, no. 1, p. 128493, 2024, doi: 10.1016/j.neucom.2024.128493.

[18] O. H. Anidjar, R. Marbel, and R. Yozevitch, "Whisper Turns Stronger: Augmenting Wav2Vec 2.0 for Superior ASR in Low-Resource Languages," *arXiv preprint arXiv: 2501.00425*, 2024.

[19] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4–5, pp. 411–430, 2000, doi: 10.1016/S0893-6080(00)00026-5.
[20] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF," *APSIPA Transactions on Signal and Information Processing*, vol. 8, pp. 1–14, 2019, doi: 10.1017/ATSIP.2019.5.
[21] S. Ui-Hyeop, L. Sangyoun, K. Taehan, and P. Hyung-Min, "Separate and Reconstruct: Asymmetric Encoder-Decoder for Speech Separation," *arXiv preprint arXiv: 2406.05983*, 2024.
[22] L. Bouchakour, K. Lounnas, and M. Debyeche, "Enhancing Robustness of Arabic Speech Recognition in Noisy Environments Using Advanced Feature Extraction and Denoising Techniques Based on Deep Learning Models," *Circuits, Systems, and Signal Processing* , 2025, doi: 10.1007/s00034-025-03418-w.
[23] K. Lounnas, M. Abbas, M. Lichouri, M. Hamidi, H. Satori, and H. Teffahi, "Enhancement of spoken digits recognition for under-resourced languages: case of Algerian and Moroccan dialects," *International Journal of Speech Technology* , vol. 25, no. 2, pp. 443–455, 2022, doi: 10.1007/s10772-022-09971-y.
[24] H. Mubarak, A. Hussein, S. A. Chowdhury, and A. Ali, "QASR: QCRI Aljazeera Speech Resource – A large scale annotated Arabic speech corpus," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguist. 11th Int. Joint Conf. Nat. Lang. Process. (ACL-IJCNLP)*, 2021, vol. 1, pp. 2274–2285. doi: 10.18653/v1/2021.acl-long.177.
[25] A. R. Ali, "Multi-Dialect Arabic Speech Recognition," in *2020 International Joint Conference on Neural Networks (IJCNN))*, Glasgow, UK, 2020, pp. 1–7. doi: 10.1109/ijcnn48605.2020.9206658.
[26] R. Ashifur, Md. M. Kabir, M. F. Mridha, M. Alatiyyah, H. F. Alhasson, and S. S. Alharbi, "Arabic Speech Recognition: Advancement and Challenges," *IEEE Access*, vol. 12, pp. 39689–39716, 2024, doi: 10.1109/ACCESS.2024.3376237.
[27] W. N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, no. 11, pp. 3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

BIOGRAPHIES OF AUTHORS

Mourad Reggab is an Associate Professor at the University of Laghouat, Algeria. An academic with over two decades of experience in higher education, he earned his degree in Electronics Engineering from the University of Boumerdès in 2001, specializing in communication systems. He later completed a Magister degree in Electronic Systems' Engineering with a focus on Automatic Speech Recognition. His primary research areas are signal processing, speech recognition, blind source separation, and artificial intelligence. He is a member of the Telecommunications, Signals, and Systems Research Laboratory at Amar Telidji University. His specific research interests include statistical signal processing, automatic speech recognition, blind source separation, image processing, and artificial intelligence. He can be contacted via email at: m.reggab@lagh-univ.dz.

Mohammed Belkhiri Has received the Engineer degree in Electrical Engineering and Electronics from the Institute of Electrical and Electronics Engineering, University of Boumerdès, Algeria, in 2000. He then earned the Magister degree in Robotics and Automatic Control in 2002 and the Ph.D. degree in Automatic Control in 2008, both from the École Nationale Polytechnique, Algiers, Algeria. From 2003 to 2008, he served as an Assistant Professor at the University of Laghouat, Algeria. He was promoted to Associate Professor in 2008, a position he held until 2016. Since 2011, he has been affiliated with the Telecommunications, Signals, and Systems Research Laboratory at the University Amar Telidji, Laghouat, Algeria. Since 2016, he has been a Full Professor with the Electrical Engineering Department at the University of Laghouat. His research focuses on nonlinear, adaptive, and intelligent neural network control of electromechanical systems, with applications in power conversion, robotics, and autonomous systems. He can be contacted at email: m.belkheiri@lagh-univ.dz.