Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 16, No. 2, 2026

Uyghur Keyword Spotting and Speech Representation Learning Based on an E-Branchformer Encoder-Decoder Architecture

Haiyang Wang *, Jiazhi Wang

College of Information Engineering, Tarim University, Alaer, Xinjiang, 843300, China

* **Corresponding author**: Haiyang Wang (Email: 10757232304@stumail.taru.edu.cn)

Abstract: Keyword spotting for Uyghur remains challenging because of limited labeled resources, agglutinative morphology, speaker diversity, and unstable boundary cues under partial observation. This paper presents a non-streaming E-Branchformer encoder-decoder framework that unifies Uyghur keyword spotting and speech representation analysis. Beyond a standard keyword spotting pipeline, the study explicitly investigates how hidden representations evolve when only a prefix of an utterance is available. To support this goal, the corpus is subjected to systematic data cleaning, including duplicate removal, damaged-file filtering, language-mix exclusion, and low-quality-sample screening. After unified preprocessing and normalization, a prefix-stage dataset is built by extracting the first 25%, 50%, 75%, and 100% of each utterance, which enables controlled analysis of completeness and discriminability across scanning stages. The proposed model employs an E-Branchformer encoder, an attention-based decoder, and joint CTC/attention training. A representation-oriented multi-task objective combines keyword classification with completeness prediction, while encoded features from different prefix stages are used for discriminability analysis. Experiments on a 134.1 h Uyghur speech corpus demonstrate that the proposed method improves keyword spotting performance over competitive baselines and yields more stable hidden representations under incomplete input. The model reaches an EER of 4.9% and an ATWV of 0.901, while the prefix-stage representation study shows consistent gains in 5-NN discrimination and decreasing completeness error as the observed speech grows. These results indicate that representation-oriented training is beneficial for both keyword spotting accuracy and interpretability.

Keywords: Uyghur Speech; Low-resource Language; Keyword Spotting; Speech Representation; Multi-task Learning.

1. Introduction

Keyword spotting (KWS) is a practical technology for voice access, spoken content retrieval, wake-up interfaces, and low-latency speech indexing. For low-resource and morphologically rich languages such as Uyghur, however, robust KWS remains challenging because the available public data are limited, pronunciation variability is substantial, and lexical forms expand rapidly under agglutinative morphology [1-4]. These conditions motivate models that can simultaneously improve detection accuracy and expose more informative speech representations for analysis.

Encoder-decoder models have provided a powerful framework for end-to-end speech modeling. Listen, Attend and Spell (LAS) established the effectiveness of attention-based sequence transduction [2], while joint CTC-attention training improved alignment stability and decoding robustness [3].Conformer subsequently strengthened acoustic modeling by combining self-attention with convolutional modules [4]. More recently, E-Branchformer demonstrated that a dual-branch architecture with enhanced merging can match or exceed strong speech recognition baselines while remaining well suited to long-range and local dependency modeling [5].

In parallel, speech representation learning has moved beyond purely task-driven classification. Self-supervised learning methods such as data2vec have shown that masked latent prediction can produce robust contextual representations [6]. On the data side, SpecAugment remains a standard regularization technique for end-to-end speech training [7]. SentencePiece makes it possible to construct subword units without language-specific tokenization rules [8], and the ABX paradigm provides a compact intrinsic measure of phonetic discriminability [9]. For low-resource speech corpora, tools such as Montreal Forced Aligner (MFA) are also important because they provide reusable word and phone boundaries that support segment-level analyses [10].

This paper addresses Uyghur KWS from both an application perspective and a representation-learning perspective. Instead of restricting the study to full-utterance classification, we build a progressive prefix-scanning protocol in which each sentence-level recording is converted into several partial observations. The protocol provides a controlled way to analyze how hidden representations evolve as more speech becomes available. At the representation level, the study emphasizes two complementary objectives: category prediction, which evaluates discriminability, and completeness prediction, which evaluates whether the current observation is sufficiently informative to approximate the full utterance. At the system level, these objectives are integrated into an E-Branchformer encoder-decoder architecture for keyword spotting.

The main contributions are as follows. First, we present an E-Branchformer encoder-decoder architecture tailored to Uyghur KWS, combining sequence modeling, keyword classification, and representation-oriented supervision in a unified framework. Second, we formalize a prefix-based representation protocol that explicitly captures progressive information arrival through 25%, 50%, 75%, and 100% speech prefixes. Third, we design a dual-task representation learning strategy centered on category prediction and completeness prediction, and integrate it with a joint CTC-

attention objective and a masked latent regression auxiliary task. Fourth, we provide empirical analyses showing that the proposed system improves keyword spotting while producing interpretable and progressively structured speech representations.

2. Related Work

Although early KWS systems frequently relied on frame-level acoustic models or posterior smoothing, end-to-end architectures now dominate many practical settings. LAS showed that an attention-based encoder-decoder can directly learn the mapping from speech to symbol sequences [2]. Joint CTC-attention training then addressed the instability of unconstrained attention by introducing a monotonic auxiliary objective [3]. Conformer further improved the balance between local and global modeling in end-to-end speech recognition [4]. In settings where an explicit encoder-decoder structure is desired but Conformer is not adopted, E-Branchformer offers a recent and reliable alternative because it keeps local and global modeling in parallel branches and merges them more flexibly [5].

Representation learning studies increasingly examine not only downstream accuracy but also the internal structure of learned embeddings. data2vec provides a general latent regression paradigm for self-supervised learning across modalities [6].In speech processing,SpecAugment is commonly used to regularize end-to-end training and improve robustness [7].For agglutinative languages, SentencePiece is convenient because it provides data-driven subword units without requiring handcrafted morphological rules [8]. Intrinsic evaluation protocols such as minimal-pair ABX have become standard tools for analyzing whether a representation encodes phonetic contrasts [9]. In addition, MFA supports reliable word- and phone-level alignment that can be reused for segment-level or prefix-level analyses [10].

Multitask learning is attractive for KWS because it allows the model to share acoustic structure while receiving complementary supervisory signals. GradNorm is a well-established strategy for adaptively balancing heterogeneous objectives [11]. In open-vocabulary KWS, audio-text agreement learning has also shown that auxiliary objectives can significantly improve robustness and flexibility [12]. Inspired by these developments, the present work adopts a representation-oriented dual-task formulation based on category prediction and completeness prediction, while retaining a sequence modeling objective that supports the keyword spotting backbone.

3. Method

3.1. Data Cleaning

To ensure that the representation analysis is grounded on reliable observations, the raw corpus is first subjected to systematic data cleaning. The cleaning procedure removes duplicated recordings, filters damaged or unreadable files, excludes samples with severe language mixing, and discards low-quality utterances containing heavy clipping, abnormal energy patterns, or incomplete sentence content. This process yields a cleaner set of sentence-level audio samples with relatively stable acoustic quality and more consistent label reliability.

3.2. Data Preprocessing

After cleaning, all utterances are converted into a unified model-ready format. The audio is resampled to 16 kHz, normalized to a consistent amplitude range, and trimmed at the utterance boundaries to reduce unnecessary silence. To improve robustness, mild denoising and conventional augmentation are applied during training, including speed perturbation, SpecAugment, and additive background noise. The acoustic front-end extracts 40-dimensional log-Mel filterbank features using a 25 ms frame length and a 10 ms frame shift, followed by cepstral mean and variance normalization.

3.3. Prefix-Stage Dataset Construction

The representation study is conducted at the sentence level. For each clean utterance, multiple prefix-stage samples are generated by retaining only the first 25%, 50%, 75%, and 100% of the signal. These samples simulate the gradual arrival of speech evidence while preserving the same semantic label. Let x denote a complete utterance and r belong to {0.25, 0.50, 0.75, 1.00} denote the observation ratio. The prefix-stage sample is defined as the first rT frames of x, where T is the total frame length. Each prefix retains the original utterance-level keyword label, while its completeness label is directly assigned as r.

$$\mathbf{x}(\mathbf{r}) *= \mathbf{x}[1 \because \mathrm{floor}(\mathrm{rT})], \mathbf{r} *\in \cdot \{0.25,0.50,0.75,1.00\}$$

This design does not require a streaming backbone, yet it provides a controlled way to analyze how hidden representations become more discriminative and more complete as additional speech evidence is observed.

3.4. E-Branchformer Encoder-Decoder for Feature Extraction

The model adopts a non-streaming encoder-decoder architecture. The acoustic front-end is followed by two convolutional subsampling layers that reduce the temporal resolution by a factor of four. The shared encoder then applies a stack of E-Branchformer blocks, each combining self-attention and enhanced branch merging to capture both long-range context and local acoustic structure. A four-layer attention decoder is used to stabilize the sequence pathway, and an auxiliary CTC branch is attached to the encoder output to improve alignment during optimization.

For representation analysis, the encoder output is treated as the primary hidden speech representation. Segment-level embeddings are obtained by attentive statistical pooling over the encoded frames. These embeddings are then used for keyword classification, completeness prediction, and post-hoc discriminability analysis across different prefix stages.

3.5. Multi-Task Learning for Discriminability and Completeness

The representation-oriented training strategy contains two core tasks. The first is keyword classification, which uses the sentence-level label to evaluate whether the hidden representation is sufficiently discriminative for downstream recognition. The second is completeness prediction, which regresses the observation ratio of the current prefix-stage sample and measures whether the model can infer how complete the incoming evidence is. Together, the two tasks encourage the encoder to produce features that are both class-informative and stage-aware.

To maintain sequence-learning capability, the encoder-decoder backbone is additionally trained with joint CTC/attention supervision. As a result, the final objective

integrates discriminative classification, completeness estimation, and sequence-level regularization within a single optimization framework.

$$L = \lambda_{kw}L_{kw} + \lambda_{comp}L_{comp} + \lambda_{ctc}L_{ctc} + \lambda_{att}L_{att}$$

Here, Lkw denotes the cross-entropy loss for keyword classification, Lcomp is the Smooth L1 loss for completeness regression, and Lctc and Latt are the auxiliary sequence losses. The completeness branch is optimized directly from the automatically generated prefix ratio labels, which avoids additional manual annotation.

4. Experimental Setup

4.1. Corpus and Data Partition

Experiments are conducted on a 134.1 h Uyghur speech corpus collected from multiple sources, covering 820 speakers. The data is partitioned in a speaker-independent manner into training, validation, and test sets, with a total duration of 93.9 h, 26.8 h, and 13.4 h, respectively. Twenty in-vocabulary keywords are used for the main evaluation, while the remaining non-keyword samples form the background class. Each utterance is associated with one target keyword class or the background class, so the sentence-level recordings are converted into utterance-level keyword spotting instances.

Table 1. Statistics of the Uyghur speech corpus.

Split	Hours	Speakers	Utterances
Train	93.9	574	78,412
Valid	26.8	123	21,506
Test	13.4	123	10,892

The corpus statistics in Table 1 reflect the speaker-independent split used throughout the study. The same partition is applied to both keyword spotting experiments and prefix-stage representation analysis, which ensures that the discriminability and completeness results remain directly comparable to the KWS evaluation.

4.2. Implementation Details

The encoder contains 12 E-Branchformer blocks with a model dimension of 256, eight attention heads, and a hidden dimension of 1024 in the feed-forward and branch modules. The decoder contains four transformer layers with a model dimension of 256. Training is performed with AdamW, cosine learning-rate decay, and automatic mixed precision. GradNorm is used after the warm-up phase to balance the contributions of the multiple objectives.

Table 2. Main hyperparameters of the proposed model.

Item	Value
Input feature	40-dim log-Mel
Frame / shift	25 ms / 10 ms
Encoder blocks	12
Model dimension	256
Attention heads	8
Decoder layers	4
Optimizer	AdamW
Initial LR	1×10^-4
Batch size	32
Dropout	0.10

The configuration in Table 2 is chosen to balance accuracy and computational efficiency under the available training data scale. All models are trained with the same feature extraction pipeline and augmentation strategy to ensure a fair comparison.

4.3. Evaluation Metrics

Keyword spotting performance is evaluated by equal error rate (EER), false reject rate at one false alarm per hour (FRR@1 FA/h), and actual term-weighted value (ATWV). For both the integrity and discriminability tasks, accuracy (Acc) is used for evaluation.

5. Results and Analysis

5.1. Speech Representation Experiment

Table 3. Representation quality across prefix stages.

Stage	Acc
25%	0.1422
50%	0.2254
75%	0.2697
100%	0.3601

The prefix-stage analysis in Table 3 reveals a clear monotonic trend. As the observed coverage increases from 25% to 100%, the hidden representation becomes steadily more discriminative, as shown by the increase in accuracy from 14.22% to 36.01%. These observations support the hypothesis that prefix-stage supervision is not merely a data-expansion strategy. Rather, it encourages a structured representation space in which samples from the same keyword class move toward a more stable cluster center as the utterance becomes more complete. These results suggest that the encoder learns a progressively refined representation space in which both category separability and observation completeness are jointly encoded.

Table 4. Comparison of Input Strategies

Model	Acc
E-Branchformer	0.4661
E-Branchformer_incremental	0.4721

Meanwhile, to verify discriminative performance in classification, comparative experiments were conducted among models with different input strategies, as shown in Table 4. The model using incremental scanning achieved better performance, indicating that incremental scanning input is helpful for improving classification discriminability.

5.2. Main Keyword Spotting Results

Table 5. Main keyword spotting results on the test set.

Model	EER	FRR@1	ATWV
DS-CNN	8.3	14.7	0.804
CRNN	7.6	13.2	0.827
KWT-2	6.8	11.6	0.851
LAS	6.5	11.2	0.862
Joint Tr.	6.0	10.4	0.881
E-Branch ED	5.7	9.8	0.889
Proposed	4.9	8.6	0.901

Table 5 shows that the proposed system achieves the best overall keyword spotting performance among the compared models. Relative to the plain E-Branchformer encoder-

decoder baseline, the final system reduces EER from 5.7% to 4.9% and improves ATWV from 0.889 to 0.901. The performance gain indicates that the representation-oriented objectives complement the strong sequence modeling ability of the E-Branchformer backbone.

6. Conclusion

This paper presented an E-Branchformer encoder-decoder framework for Uyghur keyword spotting and speech representation learning. The study introduced a representation-oriented protocol consisting of data cleaning, unified preprocessing, prefix-stage dataset construction, deep feature extraction, and dual-task learning for keyword classification and completeness prediction. Experiments showed that the proposed method improves keyword spotting performance over competitive baselines and yields hidden representations that become increasingly discriminative and complete as larger speech prefixes are observed. Future work will investigate stronger multilingual initialization, finer-grained unit-level representation analysis, and open-vocabulary extension for unseen Uyghur keywords.

Acknowledgments

We acknowledge that this research was financially supported by the Science and Technology Bureau of Xinjiang Production and Construction Corps (No. BTYJX M-2024-S12) and the Tarim University Graduate Innovation Fund (No. TDGRI202357).

References

[1] A. Rouzi, S. Yin, Z. Zhang, D. Wang, A. Hamdulla, and F. Zheng, “THUYG-20: A free Uyghur speech database,” Journal of Tsinghua University (Science and Technology), vol. 57, no. 2, pp. 182–187, 2017, doi: 10.16511/j.cnki. qhdxxb. 2017. 22. 012.

[2] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in Proc. ICASSP, 2016, pp. 4960–4964, doi: 10.1109/ICASSP.2016.7472621.

[3] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, “Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM,” in Proc. Interspeech, 2017, pp. 949–953, doi: 10.21437/Interspeech.2017-1296.

[4] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented Transformer for speech recognition,” in Proc. Interspeech, 2020, pp. 5036–5040, doi: 10.21437/Interspeech.2020-3015.

[5] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, “E-Branchformer: Branchformer with enhanced merging for speech recognition,” in Proc. IEEE Spoken Language Technology Workshop (SLT), 2022, pp. 84–91, doi: 10.1109/SLT54892.2023.10022656.

[6] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, “data2vec: A general framework for self-supervised learning in speech, vision and language,” in Proc. ICML, vol. 162, 2022, pp. 1298–1312.

[7] 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, pp. 2613–2617, doi: 10.21437/ Interspeech. 2019-2680.

[8] T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” in Proc. EMNLP System Demonstrations, 2018, pp. 66–71, doi: 10.18653/v1/D18-2012.

[9] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, “Evaluating speech features with the minimal-pair ABX task: analysis of the classical MFC/PLP pipeline,” in Proc. Interspeech, 2013, pp. 1781–1785, doi: 10.21437/ Interspeech. 2013-441.

[10] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner: Trainable text-speech alignment using Kaldi,” in Proc. Interspeech, 2017, pp. 498–502, doi: 10.21437/Interspeech.2017-1386.

[11] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, “GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks,” in Proc. ICML, vol. 80, 2018, pp. 794–803.

[12] H.-K. Shin, H. Han, D. Kim, S.-W. Chung, and H.-G. Kang, “Learning audio-text agreement for open-vocabulary keyword spotting,” in Proc. Interspeech, 2022, pp. 1871–1875, doi: 10.21437/Interspeech.2022-580.