Language Adaptation Wake Word Spotting via Latent Space from Pre-trained Speech Models

Shifu Xiong*, Hengshun Zhou†, Kai Shen†, Shi Cheng†
Hang Chen*, Genshun Wan*, Kewei Li*, Jun Du*, Lirong Dai*
* University of Science and Technology of China
† iFlytek Research, Hefei, Anhui, P. R. China
{sfxiong, zhhs, kaishen, chengshi, ch199703, gswan, keweili12}@mail.ustc.edu.cn, {jundu, lrdai}@ustc.edu.cn

***Abstract*—This paper presents an approach for multilingual Wake Word Spotting (WWS), ingeniously fusing pre-trained extensive speech models with tailored hidden units dedicated to WWS. Initially, the Whisper encoder functions as a spatial cornerstone, seamlessly integrating with a lightweight ShuffleNet-based encoder, followed by a shared decoder tailored for monolingual WWS. Then, to refine the encoder's capability, k-means clustering is leveraged to extract aligned targets from the latent space of the pre-trained speech model, thereby bolstering monolingual performance. Finally, linguistic priors are adaptively incorporated into the proposed framework, facilitating effective multilingual WWS. Experimental evaluations in Spanish and Arabic demonstrate the proficiency of the proposed approach in enhancing the performance of WWS.**

***Index Terms*—wake word spotting, latent space, multilingual, language adaptation.**

I. Introduction

Large-scale model advancements, notably in self-supervised learning (SSL) [1]–[4] and semi-supervised learning [5], [6], have set new benchmarks performance and accuracy of neural network models. These models, fueled by extensive datasets, demonstrate remarkable versatility and prowess across diverse downstream tasks. For example, HuBERT [3] achieves substantial WER reduction through offline clustering, aligning targets for BERT-like prediction. Microsoft's WavLM [4] excels in various speech processing tasks. In 2022, OpenAI's Whisper [7], trained on 680k hours of weakly supervised data, approaches human-level robustness and accuracy in English speech recognition and showcases impressive generalization across benchmarks, even in zero-shot scenarios.

These pre-trained models, though effective feature extractors, tend to be substantial in size. This poses a deployment obstacle for resource-constrained devices, especially in wake word spotting (WWS), a specialized keyword spotting (KWS) technique that identifies preset wake words to activate speech-enabled devices [8]–[10]. Thus, prioritizing the development of compact, low-latency WWS systems is crucial. Researchers have focused on crafting compact network architectures specifically for WWS to mitigate the challenges of the model size [11]–[17]. These meticulously designed networks have achieved remarkable outcomes. Simultaneously, researchers are exploring avenues to simplify pre-trained networks [18], [19]. Additionally, neural network pruning strategies [20], [21] and diverse loss functions [22], [23] have been studied to streamline KWS models. Recently, to harness the power of pre-trained models, researchers have integrated knowledge distillation techniques into self-supervised speech representation learning (S3RL) frameworks [8], [24]–[26]. Most researches focus on pre-training customized WWS teacher models, incurring significant development and maintenance costs. Although automatic speech recognition (ASR) offers a promising latent space for exploration, directly leveraging encoder information from ASR may be redundant for KWS, as activations can occur from identical pronunciations regardless of text content. In this context, K-means clustering shows promise, and further research in this area is worthwhile.

Traditional monolingual WWS approaches struggle to scale to multilingual scenarios due to development costs and insufficient end-to-end resources. Multilingual WWS performance is further hindered by linguistic differences and limited data. Multilingual fusion attempts often suffer struggles, prompting researchers to seek innovative strategies for transcending single-language models' limitations and enabling seamless multilingual transitions [27]. The authors introduced two locale-conditioned universal models leveraging locale feature concatenation. While manipulating the encoder's output, further investigation into multilingual input prior to encoding remains crucial.

In this paper, we introduce a lightweight yet potent approach for wake word spotting, seamlessly integrating pre-trained large speech models with k-means clustering-derived hidden units. Initially, we leverage a fixed Whisper encoder followed by a WWS decoder during the training phase, ensuring a proficient decoder. Subsequently, we adopt the Whisper encoder as a spatial anchor, introducing a lightweight ShuffleNet-based encoder and a shared decoder for WWS training. Additionally, we implement auxiliary training for the lightweight encoder in the latent space, utilizing k-means clustering to assign discrete labels based on the Whisper encoder's output. This auxiliary loss is training weighted alongside the WWS loss from both branches, enhancing the training process. Furthermore, we adaptively incorporate language prior information for effective multilingual WWS. Verified on car scene corpora, the proposed method boosts performance in both Spanish and Arabic.

979-8-3315-7206-8/25/$31.00

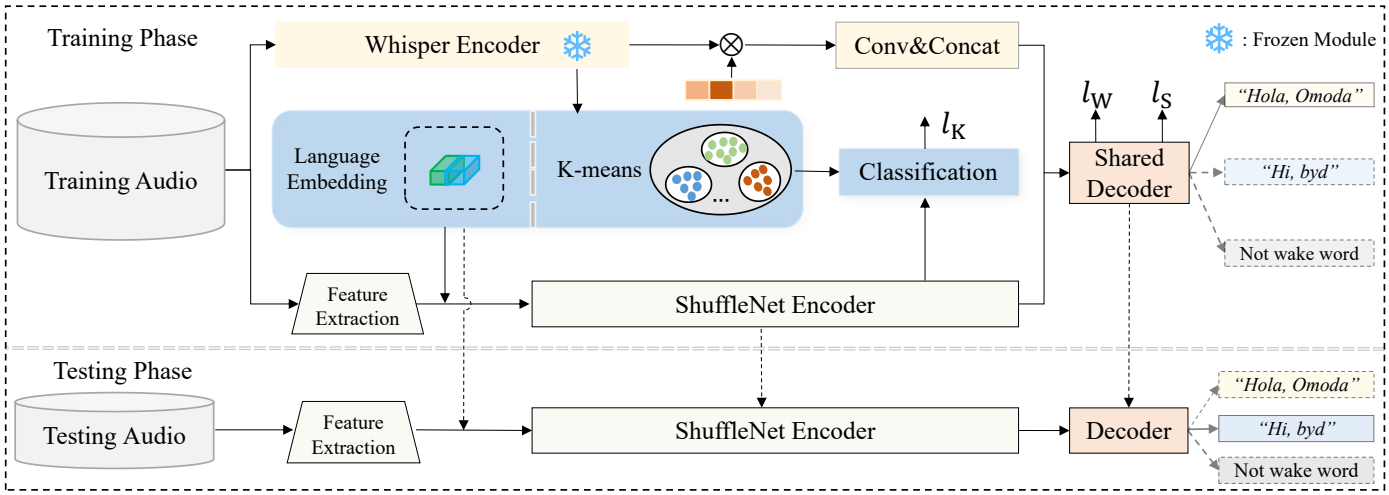


Fig. 1. The architecture of the proposed language adaptation lightweight wake word spotting system via latent space from the pre-trained speech model.

II. Proposed Frameworks

The comprehensive flowchart illustrating the proposed multilingual wake word spotting approach is presented in Fig. 1, showcasing an end-to-end design that guarantees a smooth transition from input to output.

A. WWS System Combining Whisper and Lightweight Encoder

We employ an encoder-decoder framework for end-to-end wake word spotting. Our experiments incorporate two distinct encoders: a Whisper encoder [7] and a ShuffleNet-based network [28] boasting a minimal parameter count of merely 774k, ensuring a lightweight and efficient model. For decoding, we leverage the multi-scale dilated temporal convolutional network (MDTC) [29], renowned for capturing temporal audio patterns. This setup efficiently extracts pertinent audio features, facilitating accurate wake word detection.

1) WWS system based on Whisper encoder: OpenAI Whisper [7] boasts a comprehensive suite of multilingual multi-task models with pre-trained versions released openly, fostering wide adoption and further research. Shinji Watanabe et al. recently introduced Open Whisper-style Speech Model (OWSM), replicating Whisper's training with open-source tools and public data [30], [31]. Despite field advancements, Whisper retains its competitive edge, emphasizing its robustness and efficacy. We incorporate a Whisper encoder with an MDTC decoder for wake word spotting. For efficiency and performance, we utilize the tiny Whisper model with a four-layer encoder. During training, we freeze the Whisper encoder parameters, optimizing the utilization of the encoder information by introducing weights based on encoder's hidden layers and an additional convolutional layer, and concatenating four adjacent frames, as illustrated in Fig. 1.

Following the methodology presented in [32], we assign values of 1 to the n frames immediately surrounding the wake word endpoint and 0 to all others. Utilizing a hybrid LSTM-CNN architecture, we train a voice activity detection (VAD) model on matched scenario data. Our WWS networks are optimized using the max-pooling loss function, as detailed in [33], [34]. For each sample, our model outputs a probability $p(y = 1|\Theta)$ indicating the possibility of the wake word's presence, where Θ represents the model parameters. The objective is to minimize the binary cross-entropy loss between this prediction and the ground truth, formulated as below

$$l_{\mathrm{W}} = -\sum_{t\epsilon\widehat{\mathrm{F}}} \log y_t - \log y_{f_p}^{k_t^\dagger} \tag{1}$$

where $\widehat{\mathrm{F}}$ encompasses all indices for frames not corresponding to the wake word. $k_t^\dagger$ designates the positive target label for frames, indicating the wake word's presence. Within $k_t^\dagger$, f_p labels a specific frame where the posterior probability for $k_t^\dagger$ attains its maximum value.

2) WWS system based on ShuffleNet-based encoder: Given the unique demands of the application, we opt to replace the Whisper encoder with ShuffleNet-based [28] as the alternative encoder. The loss function for the WWS system, utilizing ShuffleNet-based network, is formulated as:

$$l_{\mathrm{S}} = -\sum_{t\epsilon\widehat{\mathrm{F}}} \log y_t - \log y_{f_p}^{k_t^\dagger} \tag{2}$$

To leverage semantic information in Whisper, utilizing the Whisper encoder as a spatial anchor, facilitating its integration with the ShuffleNet-based encoder. After training the Whisper-based WWS system, we added the ShuffleNet-based encoder, fostering a shared decoder architecture. This approach ensures independent processing of the same training sample by both encoders. The combined loss function for the WWS system is formulated as:

$$l_{\mathrm{WS}} = \alpha \times l_{\mathrm{W}} + \beta \times l_{\mathrm{S}} \tag{3}$$

here, α and β are coefficients that sum to 1, adjusting the relative weight of two components in the loss function. Specifically, $\alpha = 0.5$ is used in this study.

During the testing phase, we exclusively employ the ShuffleNet-based encoder, followed by the shared decoder for

decoding, maintaining a focused and efficient pipeline for accurate wake word detection.

B. WWS System Combining Hidden Units

Drawing inspiration from [35] and [36], we train a ShuffleNet-based encoder model to predict hidden-level labels generated via k-means clustering on Whisper encoder's hidden layer embeddings. We initially extract embeddings for clean, playback, and universal negative data using the Whisper encoder. These embeddings are then clustered into 4096 categories using k-means for training. The ShuffleNet-based encoder incorporates a parallel frame-level classification branch, utilizing the k-means-derived labels. The classification is optimized with cross-entropy loss (CE-loss), which can be expressed as follows:

$$l_{\mathrm{K}}^{t} = -\log p_{c_t}(\mathbf{X}_t) \tag{4}$$

$$l_{\mathrm{K}} = -\sum_{c=1}^{K} y_c \log p_c(\mathbf{X}_t) \tag{5}$$

here, $\mathbf{X}_t$ represents the d-dimensional spectral feature, and $p_i(\mathbf{X}_t)$ denotes the softmax output for the i-th dimension from the encoder's classification layer. c_t indicates the frame-level label at frame t. The overall system loss is calculated as a weighted sum of three distinct loss sets, formulated as follows:

$$l_{\mathrm{WSK}} = \alpha \times l_{\mathrm{W}} + \beta \times l_{\mathrm{S}} + \gamma \times l_{\mathrm{K}} \tag{6}$$

where γ is a weight, with a value of $\gamma = 1.0$ being employed.

C. WWS System with Language Prior Information

In [37], the authors presented a multilingual query-by-example KWS system leveraging a residual neural network, achieving promising accuracy for streaming keyword spotting. Recently, notable performance gains were reported through a local conditional approach for multilingual KWS [27]. We introduce an adaptive language hidden space method by integrating data from various languages. For wake word spotting across languages, we dynamically adjust the decoder's neuron nodes. To mitigate the high false alarm rates often seen in multilingual training, we employ a language adaptation hidden space scheme, assigning unique learnable embeddings per language and concatenating them with hand-crafted features along the channel dimension. This language-specific concatenation is then processed by the encoder, whose first convolutional layer's input channels are modified accordingly, optimizing wake word spotting tasks.

III. Experiments and Results

A. Databases and Implementation Details

We implemented the proposed approach in Spanish and Arabic, leveraging audio data gathered from a smart car setting. For Spanish, *"Hola, Omada"* is used as the wake word, while *"Hi, byd"* was employed for Arabic. We replayed clean wake word audio across varying noise levels, generating diverse signal-to-noise ratios (SNRs). Furthermore, we enriched the negative samples with a mix of noises, universal negatives (unconstrained conversations), and challenging similar-sounding words. Both languages' training corpus totaled approximately 600 hours of audio. The validation and test sets featured 48 hours of negatives respectively and 8000 Spanish/6800 Arabic positive instances.

To set the false alarm threshold (TH), we aimed for no more than three false alarms in 48 hours, prioritizing recall as our key metric. Recall measures the system's success in detecting wake words present in samples.

Using Pytorch, we trained all models with SGD optimization, setting the learning rate to 0.001. The training spanned 8 epochs on 8 V100 GPUs.

TABLE I
Test set performance for Spanish and Arabic WWS systems. [TH: false alarm threshold]

WWS systems	es		ar	
	TH	Recall (%)	TH	Recall (%)
ShuffleNet Enc	0.851	64.90	0.964	75.57
Whisper Enc	0.853	76.40	0.846	86.16
Combined Encs	0.842	77.80	0.832	89.85
+ Hidden Units	0.829	79.30	0.819	91.44

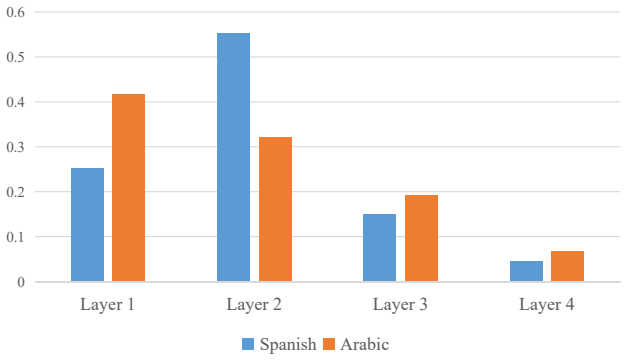


Fig. 2. Distribution of weights for the Whisper encoder with different sets of layers.

B. WWS System Combining Whisper Encoder and Hidden Units

Initially, we evaluated the efficacy of WWS systems equipped with diverse encoders. Table I presents exhaustive outcomes for Spanish and Arabic. "es" and "ar" denote Spanish and Arabic, respectively. "Combined Enc" signifies a cascaded shared decoder integrating Whisper and ShuffleNet-based encoders, while "+ Hidden units" denotes the additional classification of ShuffleNet-based encoder's output into hidden units, as elaborated in SectionII-B. Notably, the static Whisper encoder, which solely trained the decoder for both languages, outperforms the ShuffleNet-based encoder. Integrating both encoders with a shared decoder further boosts performance, surpassing individual encoder systems. We substituted the tiny Whisper model with the larger-v2 Whisper, acknowledging the extended training time. This substitution significantly enhanced

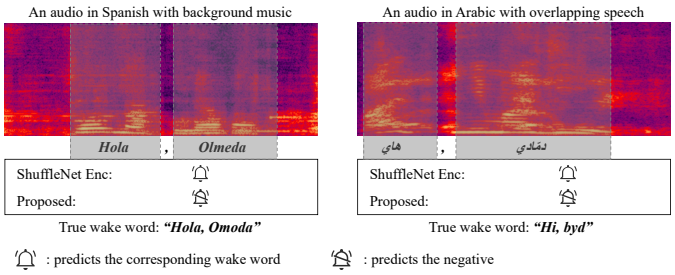


Fig. 3. Two audio exemplars, each containing pronunciations akin to the wake word elaborated upon in test set alongside their respective prediction across diverse systems.

WWS performance, hinting at potential optimizations within the larger speech model tailored for WWS.

Then we performed a comparative weight distribution analysis for the Whisper encoder across varying layer configurations, as depicted in Fig. 2, adopting a methodology akin to [24]. Notably, a distinct disparity arises between Spanish and Arabic, with Arabic weights steadily declining while Spanish weights prominently concentrate in the second layer. This disparity hints at the Whisper encoder's distinct adaptation to these languages, potentially mirroring their unique linguistic nuances.

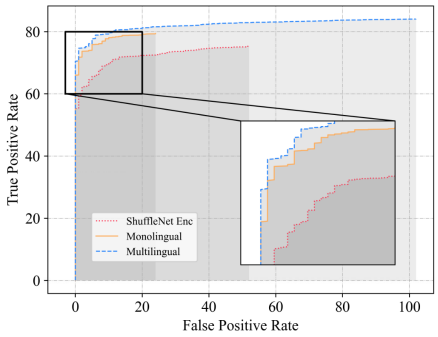
By integrating the classification of hidden units into the lightweight encoder's output, as elaborated in SectionII-B, we observed significant enhancements in system performance over the combined encoder system. The performance gains for Spanish and Arabic are summarized in the final row of Table I. To better demonstrate the advantages of the proposed method, we conducted manual auditory inspection for test samples. Fig. 3 presents two audio clips for Spanish and Arabic, respectively. In the left instance, a female voice greets someone whose name sounds similar to "*Omoda*" amidst background music. Similarly, in the right instance, a speaker pronounces a phrase as "*Hi, di ma di*" that sounds similar to "*Hi, byd*". These misclassifications observed in ShuffleNet-based WWS systems primarily from the phonetic modeling, which struggles with such minute pronunciation nuances. Conversely, Whisper excels in imparting supplementary semantic insights, thereby bolstering the system's resilience against such ambiguities, even amidst distracting background noise and sound overlaps.

TABLE II
PERFORMANCE COMPARISON OF WWS SYSTEMS. [TH: FALSE ALARM THRESHOLD]

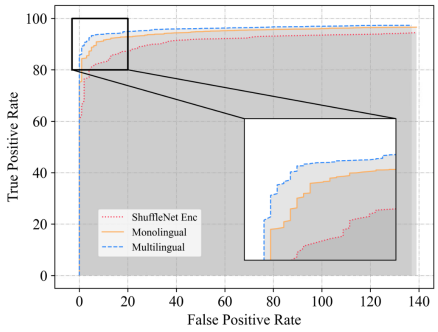
WWS systems	es		ar	
	TH	Recall (%)	TH	Recall (%)
Baseline	0.809	76.62	0.913	89.43
LE [27]	0.648	76.59	0.773	91.38
Multilingual	0.894	79.70	0.951	91.57

C. WWS System Combining Language Prior Information

Expanding upon prior research, we refined the WWS by integrating language prior knowledge, as elaborated in Section II-C. The outcomes are concisely presented in Table II. "Baseline" indicates a mixed-language training approach without language-specific cues. Analyzing Table II, it's evident that omitting language information leads to a substantial performance drop compared to single-language WWS. Our comparison with [27] highlights that leveraging language priors achieves optimal results, emphasizing the crucial role of linguistic priors in crafting robust, efficient multilingual wake word spotting systems. Fig. 4 illustrates the true positive rate against the false positive rate, showcasing the performance of various WWS systems on a per-language basis, where 'Monolingual' and 'Multilingual' stand for combined encoders of monolingual and multilingual system. This comparison not only underlines the discriminative capabilities of each system but also serves as evidence for the exceptional effectiveness and robust stability of our novelly proposed method.



(a) ROC curves for Spanish.



(b) ROC curves for Arabic.

Fig. 4. Comparison of the receiver operating characteristic (ROC) curves for Spanish and Arabic WWS systems.

IV. CONCLUSION

This study presents a streamlined approach for multilingual wake word spotting, efficiently fusing pre-trained speech models with hidden units. The method leverages a lightweight encoder anchored by the Whisper encoder, followed by a shared decoder. Auxiliary k-means training assigns categorical labels to encoder outputs, leveraging prior language knowledge to amplify performance. Experimental results on the multilingual WWS task demonstrate consistent improvements in detection accuracy over the mixed-language baseline, while

also yielding a more compact model size suitable for deployment in resource-constrained environments. Future endeavors will integrate the Whisper language identification module and expand our scope to encompass a wider array of languages.

V. Acknowledgments

This research was partially supported by grants from the National Natural Science Foundation of China (Grant U23A20315) and the National Key Research and Development Program of China (Grant No.2022YFB4500600).

References

[1] Q. Yang, Q. Liu, and H. Li, "Deep residual spiking neural network for keyword spotting in low-resource settings," in *INTERSPEECH*, 2022, pp. 3023–3027.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[3] 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 Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[4] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[5] Y. Zhang, J. Qin, D. S. Park, *et al.*, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2020, arXiv:2010.10504. [Online]. Available: https://arxiv.org/pdf/2010.10504.

[6] Y. Zhang, D. S. Park, W. Han, *et al.*, "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1519–1532, 2022.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning (ICML)*, 2023, pp. 28 492–28 518.

[8] C. Gao, Y. Gu, F. Caliva, and Y. Liu, "Self-supervised speech representation learning for keyword-spotting with light-weight transformers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[9] G.-P. Yang, Y. Gu, S. Macha, Q. Tang, and Y. Liu, "On-device constrained self-supervised learning for keyword spotting via quantization aware pre-training and fine-tuning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 951–10 955.

[10] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4087–4091.

[11] Z. Song, Q. Liu, Q. Yang, and H. Li, "Knowledge distillation for in-memory keyword spotting model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4128–4132.

[12] S. Lv, X. Wang, S. Sun, L. Ma, and L. Xie, "Dccrn-kws: An audio bias based model for noise robust small-footprint keyword spotting," in *INTERSPEECH*, 2023, pp. 929–933.

[13] Z. Akhtar, M. O. Khursheed, D. Du, and Y. Liu, "Small-footprint slimmable networks for keyword spotting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] D. Ng, Y. Xiao, J. Q. Yip, *et al.*, "Small footprint multi-channel network for keyword spotting with centroid based awareness," in *INTERSPEECH*, 2023, pp. 296–300.

[15] J. Wang, M. Xu, J. Hou, *et al.*, "Wekws: A production first small-footprint end-to-end keyword spotting toolkit," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[16] Y. M. Saidutta, R. S. Srinivasa, C.-H. Lee, C. Yang, Y. Shen, and H. Jin, "To wake-up or not to wake-up: Reducing keyword false alarm by successive refinement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[17] K. Ding, M. Zong, J. Li, and B. Li, "Letr: A lightweight and efficient transformer for keyword spotting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7987–7991.

[18] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7087–7091.

[19] R. Wang, Q. Bai, J. Ao, *et al.*, "Lighthubert: Lightweight and configurable speech representation learning with once-for-all hidden-unit bert," in *INTERSPEECH*, 2022, pp. 1686–1690.

[20] Y. Yang, K. Zhang, Z. Wu, and H. Meng, "Keyword-specific acoustic model pruning for open-vocabulary keyword spotting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1–5.

[21] H. Zhou, J. Du, C.-H. H. Yang, S. Xiong, and C.-H. Lee, "A study of designing compact audio-visual wake word spotting system based on iterative fine-tuning in neural network pruning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7572–7576.

[22] B. Labrador, G. Zhao, I. L. Moreno, A. S. Scarpati, L. Fowl, and Q. Wang, "Exploring sequence-to-sequence

transformer-transducer models for keyword spotting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[23] B. Liu, S. Nie, Y. Zhang, S. Liang, Z. Yang, and W. Liu, "Loss and double-edge-triggered detector for robust small-footprint keyword spotting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6361–6365.

[24] G.-P. Yang, Y. Gu, Q. Tang, D. Du, and Y. Liu, "On-device constrained self-supervised speech representation learning for keyword spotting via knowledge distillation," in *INTERSPEECH*, 2023, pp. 1623–1627.

[25] F. Cui, L. Guo, Q. Wang, P. Gao, and Y. Wang, "Exploring representation learning for small-footprint keyword spotting," in *INTERSPEECH*, 2022, pp. 3258–3262.

[26] H. S. Bovbjerg and Z.-H. Tan, "Improving label-deficient keyword spotting through self-supervised pretraining," in *ICASSPW*, 2023, pp. 1–5.

[27] P. Zhu, H. J. Park, A. Park, A. S. Scarpati, and I. L. Moreno, "Locale encoding for scalable multilingual keyword spotting models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[28] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6848–6856.

[29] J. Hou, L. Zhang, Y. Fu, *et al.*, "The npu system for the 2020 personalized voice trigger challenge," 2021, arXiv:2102.13552. [Online]. Available: https://arxiv.org/pdf/2102.13552.

[30] Y. Peng, J. Tian, B. Yan, *et al.*, "Reproducing whisper-style training using an open-source toolkit and publicly available data," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2023, pp. 1–8.

[31] Y. Peng, J. Tian, W. Chen, *et al.*, "Owsm v3.1: Better and faster open whisper-style speech models based on e-branchformer," 2024, arXiv:2401.16658. [Online]. Available: https://arxiv.org/pdf/2401.16658.

[32] H. Lim, Y. Kim, K. Yeom, *et al.*, "Lightweight feature encoder for wake-up word detection based on self-supervised speech representation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[33] M. Sun, A. Raju, G. Tucker, *et al.*, "Max-pooling loss training of long short-term memory networks for small-footprint keyword spotting," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 474–480.

[34] H. jin Park, P. Violette, and N. A. Subrahmanya, "Learning to detect keyword parts and whole by smoothed max pooling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7899–7903.

[35] R. Alvarez and H. Park, "End-to-end streaming keyword spotting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6336–6340.

[36] S. Gandhi, P. von Platen, and A. M. Rush, "Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling," 2023, arXiv:2311.00430. [Online]. Available: https://arxiv.org/pdf/2311.00430.

[37] P. M. Reuter, C. Rollwage, and B. T. Meyer, "Multilingual query-by-example keyword spotting with metric learning and phoneme-to-embedding mapping," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.