SCoPE: Submodular Combinatorial Prototype Learner for Continuous Speech Keyword Spotting

Anonymous ACL submission

Abstract

Keyword Spotting (KwS) in the continuous speech setting encapsulates localization and recognition of keywords amongst a large volume of non-keyword tokens, further exemplified by variation in speakers and the presence of rare keywords. Our paper presents a novel **S**ubmodular **Co**mbinatorial **P**rototype (SCoPE) learning framework that not only contrasts between target keywords but also ensures sufficient separation of keywords from non-keyword tokens. Additionally, our work proposes a weakly-supervised training strategy, utilizing forced alignment on phoneme-level embeddings to guide a windowing function to correctly localize keywords of interest. We evaluate our model on the popular LibriSpeech and L2-Arctic datasets under varying numbers of keywords demonstrating a class-imbalanced distribution and show that our proposed architecture consistently outperforms existing baselines by up to 1.8%.

1 Introduction

Keyword Spotting (KwS) for the Continuous Speech (CS) setting diverges from traditional trigger word recognition (Kundu et al., 2023; Vygon and Mikhaylovskiy, 2021; Seo et al., 2021) tasks in the *localization and identification* of target keywords from a continuous speech signal (Zhao et al., 2022a). With a recent surge in the democratization of digital communication in corporate meetings/ conferences, classroom teaching (online spoken tutorials), telemedicine, *etc.*, CS-KwS (Zhao et al., 2022a) has gained importance in identifying technical keywords and filtering of prohibited keywords in recorded speech summarization. Recent advancements, inspired by machine vision (Lin et al., 2017) include anchor-free detector (Zhao et al., 2022b) and speech-to-vision signal transformation (Samragh et al., 2023). These approaches have enhanced keyword spotting accuracy and computational efficiency (Bittar et al., 2024). The presence of rare keywords, large diversity in the keyword catalog, and a large volume of non-catalog keywords continue to remain challenging problems.

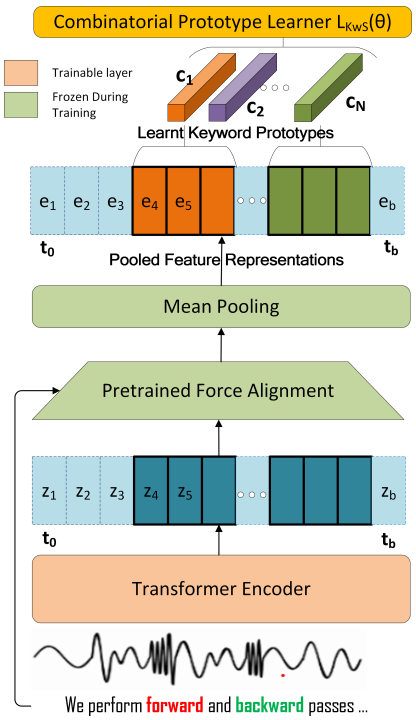

Figure 1: Overview of our proposed Continuous Speech Keyword Spotting Architecture showing the Combinatorial Prototype Learner objective.

Contemporary research in representation learning (Khosla et al., 2020; Sohn, 2016; Frosst et al., 2019; Deng et al., 2019) has made significant strides to overcome class imbalance and resilience to outliers, a predominant issue in existing approaches (Rumelhart et al., 1986). A key component of such techniques is the design of an objective function tailored to overcome specific challenges in the target task. Contrastive learning, stemming from noise contrastive estimation (Gutmann and Hyvärinen, 2010) in self-supervised learning (Chen

et al., 2020a; He et al., 2020; Chen et al., 2020b), is a notable milestone in this direction. In supervised contexts, Khosla et al. (2020); Sohn (2016); Frosst et al. (2019) explicitly focus on forming feature clusters beyond simple feature-to-centroid alignment or handling the most challenging negatives (Song et al., 2016). These methods primarily rely on pairwise similarity and have been shown in Majee et al. (2024) to not ensure clear cluster separation and resilience to imbalance which is a key requirement for the CS-KwS task.

In this work, we present SCOPE, a *Continuous Speech Keyword Spotting* framework which aims to learn discriminative representations from continuous speech signals to improve performance on downstream keyword spotting tasks. Our framework depicted in Figure 1 adopts a pretrained Wav2Vec2.0 (Baevski et al., 2020) architecture to extract speech embeddings that are further mean-pooled with the help of a forced aligner (Kürzinger et al., 2020) during training. In a continuous speech setting, a large volume of non-important keywords co-exist alongside the target keywords adding to the complexity of the KwS algorithm. To address this, word-level embeddings belonging to the target keyword class are grouped together in the feature space while being separated from others through a novel combinatorial objective (refer Section 2.3). In addition to contrasting between target keywords, our objective explicitly contrasts between target and non-target keywords to encourage the underlying model to better attend to rare target keywords.

Our contributions include: **(1)** A new Continuous Speech Keyword Spotting framework (Section 2.2) *addressing the challenges of rare keywords and code-switch*. **(2)** Introduction of a *Combinatorial Prototype Learner* (CPL) objective (Section 2.3) that explicitly contrasts non-target keywords with target keyword embeddings in addition to ensuring learning of discriminative keyword-level features. **(3)** Improvements in performance on the LibriSpeech (Panayotov et al., 2015) dataset with varying number of keywords, where our model *outperforms existing baseline by significant margins* (upto 1.8%) while being resilient to varying degrees of imbalance (Section 3).

2 Method

2.1 Problem Definition

Representation learning for continuous speech starts with learning parameters θ of a feature extractor $E(s_i, \theta)$, $\forall i \in |\mathcal{T}|$ tasked with learning robust token-level representations for the keyword vocabulary $\mathcal{T}$ in our catalog C. The output of E is a set of embeddings, which are pooled, $f_t = \text{Avg.Pool}(E(\mathcal{T}; \theta))$ to retrieve word-level embeddings guided by weakly-supervised labels from a forced-alignment model. A learning objective $L_{KwS}(\theta)$ guides the feature extractor E to learn discriminative class/keyword prototypes amidst a *large volume of non-catalog keywords* $\mathcal{T}'$ and presence of *rare keywords*. Following Majee et al. (2024), we introduce a combinatorial viewpoint where the dataset $\mathcal{T}$ is represented as a collection of sets, $\mathcal{T} = \{A_1, A_2, \cdots, A_{|C|}\}$, where each A_i, $i \in [1, C]$ represents a keyword in $\mathcal{T}$. Each batch of inputs $\{s_i, c_i\}_{i=1}^{|\mathcal{T}|}$ is considered as the ground-set $\mathcal{V}$ and the loss $L_{KwS} = \sum_{k=1}^{|C|} f(A_k; \theta)$ is modelled as the sum over the total submodular information (Fujishige, 2005) contained in a set A_k, where f is the submodular function. Note that the non-catalog words $\mathcal{T}'$ are not a part of the keyword catalog C.

2.2 The SCOPE Framework

Unlike isolated KwS tasks (Warden, 2018), real-world settings demonstrate (1) **longtail imbalance** leading to inter-class bias, (2) large **diversity in speakers** and, (3) the presence of **phonetically similar words**, *e.g.*, *weak* and *week* which lead to confusions.

Previous literature in KwS indicates a strong correlation between model performance and quality of utterance-level feature representations learnt by $E(S_i, \theta)$. Learning of robust representations is further motivated by the choice of objective function $L(\theta)$. Contrastive learners like (Khosla et al., 2020; Chen et al., 2020a) demonstrate learning of robust feature vectors by modeling feature similarity, forming compact as well as well-separated keyword-specific feature clusters. To address the challenges in real-world KwS tasks, we propose a **S**ubmodular **Co**mbinatorial **P**rototyp**e** Learner for Continuous Speech Keyword Spotting (SCOPE) framework that learns keyword prototypes $C = [c_1, c_2 ... c_N]$. Figure 1 illustrates the proposed SCOPE architecture.

We adapt the encoder from Wav2Vec (Baevski et al., 2020) for keyword-level feature representation learning. The training is performed in a supervised manner as the word-level segments of an utterance are computed at once using a forced

alignment algorithm (Kürzinger et al., 2020). The resultant word-level segments serve as soft labels. During training, we mean-pool the aligned feature segments to obtain keyword-level embeddings. We stack these feature embeddings over the batch dimension. We train the model to align the features of these word-level embeddings. We use cross-entropy as our baseline loss function. We further adopted contrastive learners like SupCon (Khosla et al., 2020) to ensure that the learned keyword embeddings are invariant to variations in context and speaker identity.

2.3 Combinatorial Prototype Learner

The challenges outlined in Section 2.1 motivate a learning framework to discern discriminative features for each keyword, with clearly established decision boundaries between catalog keywords. For continuous speech, the learner must handle numerous non-catalog keywords $\mathcal{T}'$ and catalog keywords, with some catalog words being rare or technical, creating a class imbalance. Inspired by representation learning paradigms (Khosla et al., 2020; Chen et al., 2020a; Majee et al., 2024), we propose a novel combinatorial objective $L_{KwS}(\theta)$ to address these challenges in the continuous speech setting. CPL amplifies the separation between target keywords and non-catalog words, distinguishing itself from other metric/contrastive learning objectives (Khosla et al., 2020; Chen et al., 2020a; Deng et al., 2019). This is achieved by modeling $L_{KwS}(\theta)$ as the minimization of Total Submodular Information over sets $A_k \in \mathcal{T}$ (Theorem 1).

Theorem 1 *If* $f(A) = -\sum_{i,j\in A} S_{ij}(\theta) + \sum_{i\in A} \frac{1}{|A|} \log(\sum_{j\in V} exp(S_{ij}(\theta)))$ *represents a submodular function, we can define an objective* L_{KwS} *as shown in Eqn 1 as the Total Submodular Information* $L_{KwS} = \sum_{k=1}^{|C|} \frac{1}{N_f(A_k)} f(A_k;\theta)$, *where* $N_f(A_k) = |A_k|$ *is the normalization constant. Note that the ground set* $\mathcal{V} = \mathcal{T} \cup \mathcal{T}'$.

$$L_{KwS}(\theta) = \sum_{k=1}^{|C|} \frac{-1}{|A_k|} \sum_{i,j\in A_k} S_{ij}(\theta) + \sum_{i\in A} \frac{1}{|A_k|} \log\left[\sum_{j\in\mathcal{T}} \exp(S_{ij}(\theta)) + \sum_{j\in\mathcal{T}'} \exp(S_{ij}(\theta))\right] \quad (1)$$

We provide proofs of CPL loss submodularity in Appendix A.1. CPL loss decomposes into two terms: minimizing $\sum_{i,j\in A_k} S_{ij}$ ensures intra-cluster compactness, while the second term ensures inter-keyword separation, introducing embeddings from non-catalog keywords $\mathcal{T}'$ only in the second term to maximize separation. Since $f(A_k;\theta)$ is submodular, L_{KwS} models class imbalance and is resilient to outliers (Majee et al., 2024).

Inference. After training the model, we extract speech features and store keyword prototypes by mean-pooling the features of all keyword occurrences in the training set. For each keyword, we compute the centroid by averaging these features, resulting in a centroid vector for each keyword prototype. During evaluation, we use a sliding window approach on the test utterances. For each test utterance, we slide a window of size w and stride s. Within each window, we mean-pool the features to obtain a pooled feature vector and calculate the cosine similarity between this vector and each keyword centroid. We select the keyword with the highest cosine similarity score and apply a threshold to this score to ensure confident keyword predictions.

3 Experimental Setup and Results

3.1 Datasets

LibriSpeech-100 includes 100 hours of read English speech that have been resampled at a frequency of 16 kHz and is diverse in its speakers. This dataset is specifically designed for Automatic Speech Recognition (ASR), but we adapt it to KwS as in Zhao et al. (2022a). We use Librispeech's train-clean-100 and test-clean splits. We also show results on the L2-Arctic dataset comprising recordings from 24 speakers (2 male, 2 female for each language) representing one of six native languages: Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese. Approximately 1,600 keywords were selected, with each keyword occurring once for all 24 speakers. We create a held-out set of 3400 Vietnamese-accented utterances.

3.2 Training

Wav2vec2 Large Finetuned Model (315M parameters) is used as our base model (Baevski et al., 2020). For cross-entropy, we added a projection layer after the Wav2vec2 feature extractor of size C+1, plus one denotes an extra class to represent non-catalog words. For contrastive learners, we added a projection network as described in (Khosla et al., 2020) to compress embeddings down from

Table 1: Results on LibriSpeech continuous speech dataset.

Dataset	Loss function	Precision	Recall	F1 Score
LibriSpeech Top 20	CE	93.6	95.2	94.3
	CPL	**94.6**	**97.4**	**95.8**
LibriSpeech 500	CE	92.0	92.6	92.2
	CPL	**92.0**	**95.1**	**93.5**
LibriSpeech 1500	CE	91.8	88.7	89.1
	CPL	**92.0**	**90.0**	**90.9**
L2-Arctic	CE	**98.8**	93.8	95.6
	CPL	98.4	**96.3**	**96.6**

	Precision	Recall	F1
CE	**0.988**	0.938	0.956
SupCon	0.941	0.932	0.921
CPL	0.98	**0.963**	**0.966**

Table 2: Comparing Objective Functions for KwS task

	Dataset		AP@5
Wav2Vec2	LibriSpeech20	CE	0.690
		CPL	0.692
	LibriSpeech500	CE	0.580
		CPL	0.587
	LibriSpeech1500	CE	0.364
		CPL	0.376

Table 3: Localization of keywords using AP@5.

1024 dim to 256 dim. All the models are trained on one NVIDIA A100-SXM4-80GB for 20k steps with a batch size of 30.

3.3 Evaluation Metrics

We frame the KwS task as a word detection problem. To evaluate performance, we use precision, recall, and F1 score metrics. Additionally, we aim to detect the precise position of each keyword within the utterance. This is achieved using a sliding window module, which processes the utterance with a fixed window size and stride. Consecutive predictions of the same keyword indicate its location and duration (time frames) in the utterance. For evaluation, we use the 1D Average Precision (AP) metric from (Zhao et al., 2022a). We compute the 1D Intersection over Union (IoU) for each detected keyword against the ground truth. We slide a window across the input data, apply mean pooling, and calculate cosine similarity between the pooled features and stored keyword centroids. We calculate the start and end of the time frame using window size and stride. A predicted event is a true positive (TP) if it's similarity score exceeds a threshold and the overlap of frames with the ground-truth event of the same class is more than 5%. Ground-truth events not matched by predictions are false negatives (FNs), while unmatched predictions are false positives (FPs).

3.4 Results

Table 1 shows precision, recall, F1 scores on catalogs of varying sizes containing words of different frequencies. We use catalogs of top-20 most frequent words (as in (Zhao et al., 2022a)), 500 keywords occurring between 50 to 100 times and 1500 rare keywords appearing only 5 to 10 times in the training data.

Our proposed CPL achieves an improvement of more than 1% over the CE approach across all catalog sizes. The most significant gain of 1.8% is observed in the tail word catalog of size 1500, demonstrating that our method is robust to class imbalance.

Table 2 shows precision, recall, F1 scores comparing CPL with SupCon and CE losses. We observe that CPL shows improvements of 1% and 4.5% compared to CE and SupCon, respectively. We report AP@5 in Table 3 to compare against (Zhao et al., 2022a) Libri-Top20 result. (We note our numbers are poorer than the reported numbers in (Zhao et al., 2022a) as we trained our models on 100 hrs of Librispeech training data while they trained on Librispeech 960 hours.)

4 Conclusion

We introduce the SCOPE framework for keyword spotting in continuous speech. By ensuring clear differentiation between target keywords and non-keyword tokens and employing a weakly-supervised training strategy with forced-alignment, our approach enhances keyword detection and localization. In continuous speech, the learner must effectively handle numerous non-catalog and catalog keywords, some of which are rare resulting in a class imbalance. To address this challenge, we propose a Combinatorial Prototype Learner, a novel objective function aimed at optimizing feature representations for continuous speech contexts.

5 Limitations

A limitation of our method is the reliance on a sliding window for inference, which can sometimes incorrectly predict a keyword when it appears as a subword, suffix, or prefix. Nonetheless, our framework provides a robust and effective solution for real-world keyword spotting applications.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Alexandre Bittar, Paul Dixon, Mohammad Samragh, Kumari Nishu, and Devang Naik. 2024. Improving vision-inspired keyword spotting using dynamic module skipping in streaming conformer encoder. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10386–10390. IEEE.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. *Intl. Conf. on Machine Learning (ICML)*.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

Nicholas Frosst, Nicolas Papernot, and Geoffrey E. Hinton. 2019. Analyzing and improving representations with the soft nearest neighbor loss. In *International Conference on Machine Learning*.

Satoru Fujishige. 2005. *Submodular functions and optimization*. Elsevier.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*.

Arnav Kundu, Mohammad Samragh, Minsik Cho, Priyanka Padmanabhan, and Devang Naik. 2023. Heimdal: Highly efficient method for detection and localization of wake-words. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. Ctc-segmentation of large corpora for german end-to-end speech recognition. In *International Conference on Speech and Computer*, pages 267–278. Springer.

T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. 2017. Focal Loss For Dense Object Detection. In *ICCV*, pages 2999–3007.

Anay Majee, Suraj Nandkishor Kothawade, Krishnateja Killamsetty, and Rishabh K Iyer. 2024. SCore: Submodular combinatorial representation learning. In *Forty-first International Conference on Machine Learning*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Mohammad Samragh, Arnav Kundu, Ting-Yao Hu, Aman Chadha, Ashish Srivastava, Minsik Cho, Oncel Tuzel, and Devang Naik. 2023. I see what you hear: a vision-inspired method to localize words. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Deokjin Seo, Heung-Seon Oh, and Yuchul Jung. 2021. Wav2kws: Transfer learning from speech representations for keyword spotting. *IEEE Access*, 9:80682–80691.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Inf. Processing Systems*.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Computer Vision and Pattern Recognition (CVPR)*.

Roman Vygon and Nikolay Mikhaylovskiy. 2021. Learning efficient representations for keyword spotting with triplet loss. In *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23*, pages 773–785. Springer.

P. Warden. 2018. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *ArXiv e-prints*.

Zhiyuan Zhao, Chuanxin Tang, C. Y. Yao, and Chong Luo. 2022a. An anchor-free detector for continuous speech keyword spotting. In *Interspeech*.

Zhiyuan Zhao, Chuanxin Tang, Chengdong Yao, and Chong Luo. 2022b. An anchor-free detector for continuous speech keyword spotting. *arXiv preprint arXiv:2208.04622*.

A Appendix

A.1 Proof of Submodularity of CPL Loss

The combinatorial formulation of CPL loss as in Equation 2 can be defined as a sum over the set-function $L_{KwS}(\theta)$ as described in Theorem 1 of the main paper.

$$f(A_k;\theta) = \underbrace{-\sum_{i,j\in A_k} S_{ij}(\theta)}_{\text{Term 1}} + \underbrace{\sum_{i\in A_k} \log\Bigg(\sum_{j\in\mathcal{T}} exp(S_{ij}(\theta))}_{\text{Term 2(a)}} + \underbrace{\sum_{j\in\mathcal{T}'} exp(S_{ij}(\theta))\Bigg)}_{\text{Term 2(b)}} \tag{2}$$

The Term 1 of CPL loss is a negative sum over similarities of set A_k and is thus **submodular**. The Term 2(a) and 2(b) in eq. 2 can be combined together based on the assumption in Theorem 1 which states that $\mathcal{V} = \mathcal{T} \cup \mathcal{T}'$. We show this in eq. 3.

$$f(A_k;\theta) = \underbrace{-\sum_{i,j\in A_k} S_{ij}(\theta)}_{\text{Term 1}} + \underbrace{\sum_{i\in A_k} \log\Bigg(\sum_{j\in\mathcal{V}} exp(S_{ij}(\theta))\Bigg)}_{\text{Term 2}} \tag{3}$$

The Term 2 computed in eq. 3 depicts a sum of the exponent of similarities ($\sum_{j\in\mathcal{V}} exp(S_{ij}(\theta)) - 1$) which is a modular term as the sum is computed over the complete ground set $\mathcal{V}$. The logarithm over this term constituting the complete inter-class term represents a concave over modular function which is **submodular** in nature. Thus, the underlying function $f(A_k;\theta)$ for the CPL loss $L(\theta)$ represented in Theorem 1 is **submodular** in nature.