Enhancing Word Discrimination and Matching in Query-by-Example Spoken term detection with Acoustic Word Embeddings

Pantid Chantangphol and **Theerat Sakdejayont** and **Tawunrat Chalothorn**
Kasikorn Labs Kasikorn Business–Technology Group, Thailand
{pantid.c,theerat.s,tawunrat.c}@kbtg.tech

Abstract

In this paper, we propose a novel approach to enhance query-by-example spoken term detection using Acoustic Word Embeddings (AWEs). Our AWEs model combines CNN and LSTM layers to capture sequential information and generate fixed-dimensional word-level embeddings. To address the challenge of distinguishing between words, we introduce a deep word discrimination loss that enhances embedding discrimination. Additionally, we employ an embedding-matching scheme based on cosine similarity computation and sliding window smoothing. Our experimental results demonstrate the effectiveness of our approach in word discrimination tasks, achieving high mean Average Precision scores and outperforming baseline models. Moreover, our embedding-matching scheme shows promising performance in query-by-example spoken term detection, opening up possibilities for advancements in audio indexing and search techniques.

Index Terms: spoken term detection, query-by-example, acoustic word embedding, word discrimination, audio retrieval

1 Introduction

The field of Spoken Term Detection (STD) (Mandal et al., 2014)—identifying specific terms within audio streams or files—has gained importance due to the widespread availability of internet media and the proliferation of smart devices. This has led to an increasing demand for proficient audio search tools and efficient voice control mechanisms. Query by Example (QbE) represents a specialized application of STD, offering advantages over traditional text-based searches by directly matching audio samples. This is especially valuable for handling unknown or out-of-vocabulary search terms.

Query by Example Spoken Term Detection (QbE-STD) has historically employed Dynamic Time Warping (DTW) in conjunction with frame-level features for keyword matching (Rodriguez-Fuentes et al., 2014; Mantena et al., 2014). Both supervised (Zhang et al., 2019) and unsupervised approaches (Chen et al., 2016; Holzenberger et al., 2018) have been examined, each with distinct advantages. While unsupervised methods primarily utilize traditional acoustic features (Vasudev et al., 2016; Wang et al., 2018), supervised techniques frequently employ neural network-derived phonetic features. The field has witnessed a paradigm shift with the introduction of Acoustic Word Embeddings (AWEs) (Ma et al., 2021; Kamper et al., 2019; Settle et al., 2017; Kamper et al., 2016; Yuan et al., 2018), which transform variable-length speech segments into fixed-dimensional vectors (Levin et al., 2013). This approach overcomes the computational limitations of traditional DTW-based methods, facilitating more efficient searching, clustering, and similarity comparisons. Neural networks, particularly Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, are now widely utilized for extracting these AWEs (Ram et al., 2018; Svec et al., 2022; Chen et al., 2015; Settle and Livescu, 2016; Chung and Glass, 2018; Naik et al., 2020; Ram et al., 2020; Lopez-Otero et al., 2019; Madhavi and Patil, 2017). Consequently, the current focus in QbE-STD research has largely shifted towards search and indexing tasks, with these deep learning frameworks playing a pivotal role in feature extraction.

The main challenge resides in mapping sequential speech information into vector space without losing sequential integrity. Our proposed method addresses this challenge through deep neural networks and introduces an additional loss function designed for enhanced word discrimination. This paper presents an architecture combining CNN layers for local feature extraction, Long Short-Term Memory (LSTM) layers for capturing temporal dependencies, and Fully Connected Layers (FC

Layers) for dimensionality reduction. An additional loss function is incorporated to improve word discrimination and optimize generalization across keywords spoken by various speakers by aligning embeddings with acoustic word centroids while maximizing inter-class and minimizing intra-class variation. Moreover, our method utilizes a cosine similarity-based query centroid matching technique, supplemented by moving average smoothing, for efficient word search in spoken utterances. Our contributions to this work are as follows:

1. Introduction of an Acoustic-to-Embedding network (A2E-Net) for generating word-level acoustic word representations.

2. Development of a Deep Word Discrimination (DWD) loss function aimed at enhancing the discrimination capabilities of acoustic word embeddings by minimizing intra-word distance and maximizing inter-word distance within each acoustic word embedding.

3. Establishment of Query Centroid Similarity Matching (QC-matching), a technique for acoustic word embedding matching that employs query centroids to facilitate QbE-based audio indexing.

The remainder of this paper is organized as follows: Section 2 details the proposed system, Section 3 discusses implementation aspects, Section 4 presents the results, and Section 6 outlines the conclusions.

2 Proposed framework

In this section, we present the components of our proposed method for enhancing QbE-STD. They are as follows:

2.1 Acoustic-to-Embedding network (A2E-Net)

Our proposed AWEs model architecture aims to effectively capture and represent acoustic features at both the frame and word levels. The input comprises raw audio signals, which are divided into frames using a windowing size of 25 ms and a step size of 10 ms. To extract local acoustic features, we employ two CNN layers with 3x3 kernels and 64 filters each, followed by a max-pooling layer that reduces dimensions and extracts essential features. Two additional CNN layers with 3x3 kernels and 128 filters each extract higher-level features, followed by another max-pooling layer for further dimension reduction.

To capture temporal dependencies and sequence information, we utilize two sets of LSTM layers. The first set consists of two LSTM layers with 1024 units, followed by another set of two LSTM layers with 512 units each. These LSTM layers are crucial for modeling the sequential nature of acoustic features. Subsequently, two FC layers map the LSTM outputs to lower-dimensional spaces, reducing dimensionality and facilitating subsequent embeddings. The resulting frame-level AWEs, with a size of 256x1, are obtained from the output of the FC layer. The statistical pooling layer then aggregates the variable-length frame-level AWEs into a fixed-length representation by computing the mean and standard deviation, concatenating these, and finally mapping them to a 4096-dimensional space through a linear transformation. This fixed-length representation encapsulates both the mean and variance of the frame-level features, making it a rich and comprehensive descriptor for each word. Another FC layer maps a 4096x1 representation to a 2048-dimensional space, generating word-level AWEs. During training, the model parameters are optimized using both cross-entropy loss, a common classification loss, and an auxiliary word-discrimination loss designed to enhance embedding discrimination.

In summary, our AWEs model architecture combines CNN layers for local feature extraction, LSTM layers for capturing temporal dependencies, and FC layers for dimensionality reduction and mapping to lower-dimensional embeddings. By representing acoustic features at both the frame and word levels, our model enables the effective calculation of word-level embeddings and facilitates meaningful similarity comparisons.

2.2 Deep word discrimination Loss (DWD)

The DWD loss is introduced to address the challenge of accurate word discrimination. In such tasks, where the search content and query keyword are typically spoken by different speakers, it is crucial to ensure that the AWEs of the same spoken keyword by different speakers are identical. However, traditional embedding approaches often encode speaker-related information, which hinders precise word discrimination. To overcome this limitation, we incorporate a variability-invariant loss in the training phase.

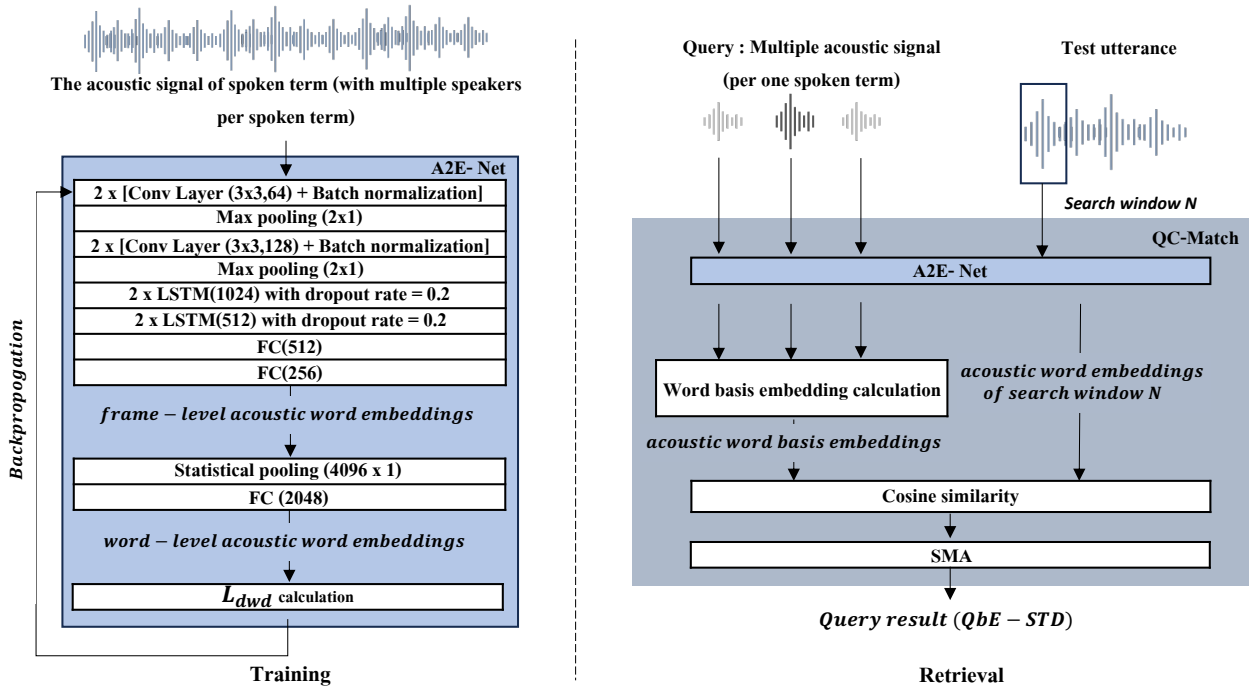

Figure 1: A2E-net with QC-matching framework for Query-by-Example Spoken term detection system

To address the inter-class and intra-class covariance in QbE-STD tasks, our additional loss function aims to maximize variation across different word classes while minimizing variation within the same class. We construct batches of size $N_{word} \times M$, where M denotes the number of acoustic signals for each word, spoken by different speakers. This is designed to capture the diversity in pronunciation, accent, and other speech characteristics unique to each speaker. Such diversity is crucial in a QbE-STD task, where the goal is to accurately identify a keyword regardless of the speaker. By incorporating acoustic signals from multiple speakers for each word, the model is trained to recognize words independently of speaker-specific characteristics. N_{word} represents the number of distinct words and indicates the size of the vocabulary in the training set. This set is generated from the alignment of acoustic signals and their transcriptions using the Montreal Forced Aligner (McAuliffe et al., 2017). After alignment, feature vectors x_{ji} are extracted from the i^{th} acoustic signal of the j^{th} word. These vectors are input into the AWEs model, comprising convolutional layers with ReLU activation and batch normalization, followed by max-pooling layers. The model also includes four LSTM layers and two dense layers with ReLU activation, culminating in the output layer that represents frame-level AWEs. Each word-level AWE emb_{ji} is normalized to enable accurate comparisons.

The main objective during training is to optimize the embedding representation of each acoustic signal. This involves aligning the embedding closely with the centroid of embeddings from the same word while ensuring a significant separation from centroids of other words. The word embedding centroid is computed by averaging the word-level embeddings, excluding the i^{th} acoustic word embedding, denoted as $[emb_{j1}, ..., emb_{jM}]$ with the M acoustic signals per word, resulting in c_j.

$$c_j = \frac{\sum_{m=1, m\neq i}^{M} emb_{jm}}{M-1} \tag{1}$$

To measure the similarity between the word-level embeddings and the centroid, we employ cosine similarity. The similarity matrix $(S_{ji,k})$ represents the scaled cosine similarities between each embedding vector emb_{ji} and all centroids c_k.

$$S_{ji,k} = \cos\left(emb_{ji}, c_k\right) \tag{2}$$

To enhance conventional contrastive loss in QbE-STD tasks, a softmax operation is applied to similarity scores, enabling a probabilistic interpretation of the similarity between embedding vectors. The loss on each embedding vector (emb_{ji}) is defined as follows:

$$L_{sm} = -S_{ji,j} + \log \sum_{k=1}^{N_{word}} \exp S_{ji,k} \quad (3)$$

where L_{sm} represents the softmax loss.

Finally, we introduce the contrastive centroid Loss (L_{cc}) to encourage embeddings of positive examples (words in the query) to be close to their respective class centers while simultaneously pushing them away from the class centers of negative examples (other words). By considering both the centrality and contrastive aspects, this loss promotes effective discrimination in QbE audio indexing.

$$L_{cc} = \sum_{j=1}^{N_{word}} \sum_{i=1}^{M} (1 - S_{ji,j}) + \max_{1<k<N_{word}, j \neq k} S_{ji,k} \quad (4)$$

The $(1 - S_{ji,j})$ targets positive pairs, measuring and minimizing their dissimilarity from the class center to enhance intra-class compactness. The second term addresses the most dissimilar negative pairs. It identifies the maximum similarity between emb_{ji} and centroids of all other classes $(k \neq j)$ The aim is to decrease the similarity of an embedding vector to centroids of different words, thus increasing inter-class variability.

The Deep Word Discrimination loss (L_{dwd}) is a combination of the softmax loss and the contrastive centroid Loss, as follows:

$$L_{dwd} = L_{sm} + L_{cc} \quad (5)$$

By incorporating the Deep Word Discrimination loss into the training process, our goal is to enhance the discriminative power of the embeddings, thereby facilitating accurate word discrimination in QbE search tasks.

2.3 Query centroid similarity matching (QC-matching)

Our proposed word-searching system employs an embedding-matching scheme based on cosine similarity computation with a sliding window. To initiate the process, the search content is divided into segments using a fixed-size sliding window along the time axis, forming a sequence of segments. These segments are then passed through a trained A2E-Net, resulting in a sequence of acoustic word embeddings derived from the FC layer.

To ensure consistency in segment lengths, the keyword audio is either padded or clipped to match the size of the sliding window. Subsequently, each input segment (x) is transformed into its corresponding embedding (emb_x) using deep CNN. In order to capture the representation of acoustic signals of a spoken query term, the basis embedding of the word is computed by averaging the word-level embeddings in the following manner:

$$c_b = \frac{\sum_{x=1}^{B} emb_x}{B} \quad (6)$$

where B is the number of multiple acoustic signals of a spoken query term. The basis embedding, denoted as c_b, captures the representative acoustic features of the spoken query.

By calculating the cosine similarity between the segment sequence of the search content y and the basis embedding of the spoken query (emb_x), we generate a time-dependent score sequence. To mitigate the impact of random score fluctuations, we apply a simple moving average (SMA) operation (Koul and Awasthi, 2019) to smooth the sequence. This smoothing process involves summing recent scores and dividing the sum by the number of frames involved at each point.

The resulting smoothed score sequence provides a measure of similarity between the search content and the spoken query, enabling the identification of relevant word occurrences within the search content. This embedding-matching approach, employing cosine similarity computation with a sliding window and subsequent SMA smoothing, offers an effective means of searching for specific words in spoken utterances.

3 Experimental Details

In this section, we provide the experimental details of our study, covering evaluation metrics, the dataset, baselines, data preparation, and model configuration.

3.1 Evaluation

Our evaluation of the method employs two key metrics: mean Average Precision (mAP) and Precision at 5 (P@5), same as (Ma et al., 2021). The mAP metric assesses the average precision for each word in word discrimination and search content. It is calculated by averaging precision values for all queries, providing a holistic measure of retrieval performance. Precision at k documents (P@k) evaluates the precision of the retrieval system by considering the relevance of the top k retrieved word

occurrences. By using mAP and P@5, we gain insights into the retrieval performance and precision of our word-searching system, accurately retrieving desired words from spoken utterances.

3.2 Dataset

This study explores word discrimination across Buckeye (Pitt et al., 2005) (6 hours for development and testing), Librispeech (Panayotov et al., 2015) (5.4 hours for development and clean testing), TIMIT (Garofolo et al., 1983) (4620 audio files for training, 1690 for testing), and English Command Voice corpus 12.0 (Ardila et al., 2020) (986,897 utterances for training, 16,365 for development and testing).

To evaluate word discrimination, we train an AWEs model using the English Common Voice dataset and assess discrimination using Librispeech and Buckeye. We investigate the QbE technique for spoken term detection and compare the performance of our embedding-matching method with other approaches. For embedding-matching, we use spoken queries from Librispeech and test utterances from TIMIT. We examine the effectiveness of fixed-dimensional acoustic embedding by obtaining unseen spoken queries from Librispeech and test utterances from TIMIT. Through these experiments, our aim is to gain insights into word discrimination and evaluate the effectiveness of our proposed method in unseen word search scenarios.

3.3 Baseline

Network: Due to the high performance of supervised acoustic word embedding models, as cited in (Ram and Aldarmaki, 2022) and (Sanabria et al., 2023), we evaluate our proposed AWE model in comparison with baseline models such as Wav2Vec 2.0 (W2V2) (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and XLSR-53 (Conneau et al., 2021) for word discrimination tasks. These baseline models leverage pre-trained supervised representations for constructing acoustic word embeddings (AWEs). Notably, HuBERT with mean-pooling outperforms other AWE systems employing simpler pooling strategies, as evidenced in (Sanabria et al., 2023), thus showcasing its robust performance across various AWEs. Additionally, XLSR-53 demonstrates promising performance, as reported in (Ram and Aldarmaki, 2022).

Loss function: We compare the performance of our DWD loss with other methods, including Triplet loss (Ge et al., 2018) and Multi-Similarity Loss (MS loss) (Wang et al., 2019), demonstrating the strong and consistent performance of our DWD loss across various AWEs.

Word matching : Furthermore, we compare our proposed embedding-matching approach for QbE-STD with the baseline Cosine Distance Pattern Matching (CDP matching) method (Ma et al., 2021). This baseline method employs cosine distance computation in conjunction with a sliding window to match spoken query segments to the search content. A simple moving average is then applied to smooth the score sequence, thereby reducing random fluctuations. Additionally, a multi-template strategy is used to average values across templates, resulting in a fused embedding. We also compare our proposed approach with the best-performing model, One Softmax AWE with V-I Loss (s-AWE), as outlined in the work of (Ma et al., 2021).

3.4 Experimental setup

To evaluate the performance of A2E-Net in word discrimination tasks, reference is made to the experiment detailed in Section 4.2. We utilize six different systems for this evaluation: the proposed A2E-Net with DWD loss, softmax loss, and MS loss, as well as pre-trained W2V2, Hubert, and XLSR-53 models. The objective is to examine the efficacy of A2E-Net across different loss functions, including DWD, softmax, and MS loss. Performance comparisons are made against high-performing pre-trained models. The metric for evaluation is mAP, and word categorization employs a cosine similarity threshold of 0.5.

To assess the proposed Query-by-Example (QbE) approach for Spoken Term Detection (STD), experiments outlined in Section 4.2 are referenced. The systems examined specifically include A2E-Net with QC matching, A2E-Net with CDP matching, and s-AWE with CDP matching (Ma et al., 2021). The performance of A2E-Net is scrutinized by employing various word-matching methods and is compared against the benchmark technique of CDP matching.

To evaluate the efficacy of the proposed method in word discrimination tasks, an experiment was conducted to compare frame-level and word-level acoustic embeddings. Two variations of the A2E-Net model were employed: one with DWD loss and another with softmax loss. Detailed results and analyses can be found in Section 4.3.

To investigate the effectiveness of the proposed method in retrieving unseen words for real-world applications, an experiment was conducted as outlined in Section 4.4. The word-level A2E-Net with DWD loss was used, and the experiment focused on two categories of words: all-selected and unseen. The objective is to evaluate the ability of the system to retrieve and rank unseen words compared to a pre-selected set of words.

3.5 Data Preparation

Speech signals in all experiments were processed at a sampling rate of 16 kHz with 16-bit resolution.

To train the word discrimination model and conduct evaluations, precise word timestamps were necessary. Forced alignment techniques were employed for datasets without manual timestamps, using the MFA (McAuliffe et al., 2017) for all datasets. The evaluation focused on words with a minimum duration of 0.5 seconds.

During word discrimination training and inference, acoustic word segments were divided into 25 ms frames with a step size of 10 ms. These frames were transformed into 25-dimensional feature vectors for the acoustic word embedding model. The model generated embeddings, and cosine similarity with a threshold was used for comparison and classification.

For embedding-matching in QbE-STD, a query word with multiple acoustic words was indexed within a recording file. The word basis embedding and average duration of the query word were calculated. The recording file was segmented into segments of the average duration with a step size of 50 ms. Acoustic word embeddings were compared to the word basis embedding using cosine similarity, enabling identification and indexing based on a similarity threshold.

3.6 Model Configuration

To compare with the baseline, we conducted experiments using frame-level and word-level representations from various models. For frame-level representations, we evaluated word discrimination models with different loss functions. For word-level representations, we examined word discrimination models with the DWD loss.

For each reported model, we employed specific hyperparameter configurations, including a learning rate of 0.001, a batch size of 32, and the Adam optimizer. The output layer of the word discrimination model generated a 2048-dimensional word embedding with N_{word} nodes, representing the number of words in the training set. We implemented early stopping, and halting training if the validation loss did not improve for more than 10 epochs or started to increase for more than 3 epochs. The maximum number of epochs was set to 100. These hyperparameter settings and training strategies played a crucial role in achieving optimal model performance.

Table 1: The performance evaluation of A2E-Net in word discrimination task

Methods		mAP(%)	
Model	Loss	Librispeech	Buckeye
A2E-Net	DWD loss	**63.9**	**72.9**
	softmax loss	59.1	65.2
	Triplet loss	60.2	68.3
	MS loss	62.5	69.1
W2V2 (Baseline)		47.4	53.1
Hubert (Baseline)		58.2	64.8
XLSR-53 (Baseline)		54.7	60.1

4 Experimental result and discussion

In this section, we present the experimental results and discussion of our study, focusing on performance evaluation and comparisons across various aspects.

4.1 The performance evaluation of A2E-Net in word discrimination task

This study investigates the performance of various model architectures in word discrimination tasks using our proposed method. In Table 1, we compare the effectiveness of the A2E-Net model across different loss functions (DWD, softmax, Triplet, and MS) against two baseline models (W2V2 and HuBERT), employing the mAP metric for evaluation. These results contribute to the advancement of word embedding models. Specifically, the A2E-Net model with DWD loss demonstrates exceptional performance, achieving the highest mAP scores of 63.9% for Librispeech and 72.9% for Buckeye, thus outperforming both baseline models. Furthermore, the A2E-Net model employing the softmax loss function also shows competitive performance, with mAP scores of 59.1% for Librispeech and 65.2% for Buckeye. However, there remains room for further optimization. In contrast, W2V2 exhibits moderate performance, and although HuBERT outperforms W2V2, it still falls short of the mAP scores achieved by the A2E-Net

Table 2: The performance evaluation of a proposed QbE Approach for QbE-STD

Methods		mAP (%)	P@5 (%)
Model	Word matching		
A2E-Net	QC-matching	**70.22**	**80.62**
A2E-Net	CDP matching	59.1	65.2
Baseline			
s-AWE	CDP matching	59.1	65.2

models. The XLSR-53 model also demonstrates promise but requires additional tuning to match the performance of our proposed models. Overall, the A2E-Net model with the DWD loss function emerges as the most effective architecture for word discrimination tasks, highlighting the efficacy of its design and chosen loss function in achieving superior performance. This research offers valuable insights into various model architectures for word discrimination, thereby guiding future investigations in this field.

4.2 The performance evaluation of a proposed QbE Approach for QbE-STD

We conducted a comprehensive investigation to assess the effectiveness of our QbE technique for QbE-STD in Table 2, comparing it with existing approaches. We implemented two variations of our model: word embedding-based matching with a proposed loss function and pattern matching based on cosine distance with the same loss function. The evaluation was performed using the mAP metric, and the results were compared to baseline approaches. The word embedding-based model achieved high mAP scores of 70.22%, effectively detecting spoken terms. On the other hand, the pattern matching-based model showed strengths in capturing patterns but exhibited slightly lower performance. In contrast, the baseline models had lower mAP scores, indicating limitations in STD. Ultimately, the word embedding-based model emerged as the most effective, outperforming the baseline models. Our findings highlight the potential of QbE techniques and pave the way for future improvements in STD methods.

4.3 The performance evaluation of frame-level and word-level Acoustic Word Embeddings for word discrimination task

This experiment evaluates various architectures for word discrimination tasks using frame-level and word-level acoustic word embeddings. Our proposed model, which employs a specialized loss function, is compared with its softmax loss variant using the mean Average Precision (mAP) metric. Results in Figure 2 show that the specialized loss function yields high mAP scores for both frame-level and word-level embeddings, highlighting its efficacy in word discrimination. Moreover, word-level representation outperforms its frame-level counterpart, capturing discrimination patterns more effectively. The proposed model also surpasses the softmax model, validating the effectiveness of our architecture and loss function. In conclusion, we recommend using the word-level approach with our specialized loss function to improve word discrimination models, contributing to advances in speech analysis.

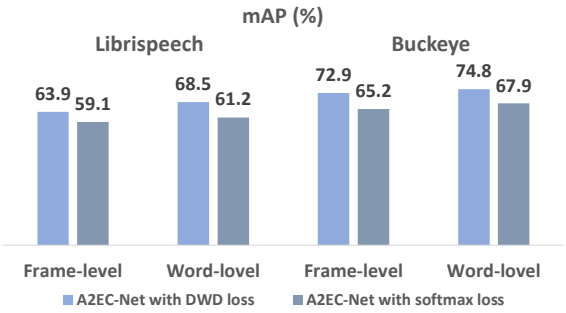

Figure 2: The performance evaluation of frame-level and word-level Acoustic Word Embedding for word discrimination task

Table 3: Performance Evaluation of AWEs for Unseen Word Retrieval

Retrieval	mAP (%)	P@5 (%)
All selected words	70.22	80.62
Unseen words	54.95	62.59

4.4 The performance evaluation of AWEs for Unseen Word Retrieval

This experiment evaluates the effectiveness of AWEs in retrieving unseen words through QC-matching, utilizing A2E-Net and the DWD loss. We measure the system's performance in identifying and retrieving unseen words compared to randomly selected words, using the mAP metric. The results presented in Table 3 advance search techniques for speech data, offering valuable insights into the effectiveness of AWEs in Unseen Word Search Retrieval. By analyzing the strengths and weaknesses of each architecture in word discrimination tasks, the proposed model with AWEs demon-

strates impressive performance in distinguishing both seen and unseen words across languages. It achieves high mAP and P@5 scores, although there is still room for improvement in discriminating unseen words. These findings highlight the effectiveness of AWEs for word discrimination and emphasize the benefits of leveraging multilingual models. Overall, they contribute to the advancement of search techniques for STD, providing valuable insights for future research in this domain.

5 Discussion

5.1 Performance Insights

Acoustic Word Embeddings (AWEs) have played a pivotal role in advancing the field by offering a computationally efficient approach to spoken term detection. Our A2E-Net model with DWD loss function outperformed baseline models like W2V2 and HuBERT, achieving mAP scores of 63.9% on Librispeech and 72.9% on Buckeye. These scores underline the architectural efficiency and the efficacy of the DWD loss function. On both frame-level and word-level tasks, our specialized loss function improves word discrimination, thereby enhancing the versatility of the model across different granularities. AWEs were also effective in retrieving unseen words, thereby advancing search techniques for speech data. Our QbE technique surpassed existing baseline models with a high mAP score of 70.22%, underscoring the efficacy of word embedding-based models in spoken term detection.

5.2 Computation Time

One notable advantage of A2E-Net is its computational efficiency. Traditional methods (e.g. DTW) suffer from high computational complexity, especially with long sequences. A2E-Net generates AWEs that represent variable-length segments as fixed-dimensional vectors, significantly reducing computation time for search and similarity comparisons. While the training phase is resource-intensive due to the depth of the model, real-time deployment remains efficient. The specialized loss function adds minimal computational overhead, making model scalable for real-time applications.

5.3 Theoretical and Practical Implications

The research findings have important theoretical ramifications for the academic community in machine learning, acoustic modeling, and natural language processing. On the practical side, the reduced computational complexity and time efficiencies hold promise for applications in information retrieval, speech indexing, and automated customer service.

5.4 Limitations and Future Work

Despite encouraging results, limitations exist. The proposed loss functions, though superior to traditional ones, require broader linguistic testing. Additional evaluation against a more diverse set of baseline models could enrich our findings. The current A2E-Net model excels in distinguishing seen words but falls short in discriminating unseen words. Future work could focus on developing adaptive methods to enhance this specific performance aspect. The generalizability of the model across various languages, dialects, or noisy environments, as well as its practical effectiveness in real-world, real-time applications, remains to be tested. Moreover, subsequent studies could expand the A2E-Net model to include more languages, particularly those with limited resources, to increase its applicability in linguistically diverse contexts. Therefore, upcoming research could focus on overcoming these limitations and further refining the performance of the model across multiple domains.

6 Conclusion

The presented research substantially advances the understanding and development of Query-by-Example Spoken Term Detection (QbE-STD) techniques, acoustic word embeddings (AWEs), and their integration with deep learning architectures. Our study introduces an innovative approach to enhance QbE-STD through the use of AWEs. The A2E model overcomes the limitations of traditional methods by converting variable-length speech segments into fixed-dimensional vectors, thereby facilitating quicker and more efficient search operations. Experimental results confirm the model's effectiveness in word discrimination tasks, underscoring its potential for innovations in audio indexing and search techniques. The incorporation of the DWD loss function further augments the discriminative power of the embeddings. Our contributions not only advance the field of QbE-STD but also set the stage for improved audio search tools and voice-controlled applications. Particularly, the A2E-Net model with DWD loss function exhibits superior performance, offering promising avenues for future research in speech technology.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

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*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Guoguo Chen, Carolina Parada, and Tara N. Sainath. 2015. Query-by-example keyword spotting using long short-term memory networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5236–5240.

Hongjie Chen, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li. 2016. Unsupervised bottleneck features for low-resource query-by-example spoken term detection. In *17th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 923–927. ISCA.

Yu-An Chung and James R. Glass. 2018. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. In *19th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 811–815. ISCA.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *22nd Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2426–2430. ISCA.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. 1983. Timit acoustic-phonetic continuous speech corpus.

Weifeng Ge, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2018. Deep metric learning with hierarchical triplet loss. In *Computer Vision – ECCV 2018*, pages 272–288, Cham. Springer International Publishing.

Nils Holzenberger, Mingxing Du, Julien Karadayi, Rachid Riad, and Emmanuel Dupoux. 2018. Learning word embeddings: Unsupervised methods for fixed-size representations of variable-length speech segments. In *19th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2683–2687. ISCA.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Herman Kamper, Aristotelis Anastassiou, and Karen Livescu. 2019. Semantic query-by-example speech search using visual grounding. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7120–7124.

Herman Kamper, Weiran Wang, and Karen Livescu. 2016. Deep convolutional acoustic word embeddings using word-pair side information. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4950–4954.

Sumit Koul and Amit Kumar Awasthi. 2019. An introduction to moving average and its importance. *Journal of emerging technologies and innovative research*.

Keith Levin, Katharine Henry, Aren Jansen, and Karen Livescu. 2013. Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 410–415.

Paula Lopez-Otero, Javier Parapar, and Alvaro Barreiro. 2019. Efficient query-by-example spoken document retrieval combining phone multigram representation and dynamic time warping. *Information Processing and Management*, 56(1):43–60.

Murong Ma, Haiwei Wu, Xuyang Wang, Lin Yang, Junjie Wang, and Ming Li. 2021. Acoustic word embedding system for code-switching query-by-example spoken term detection. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5.

Maulik C. Madhavi and Hemant A. Patil. 2017. Partial matching and search space reduction for qbe-std. *Computer Speech and Language*, 45:58–82.

Anupam Mandal, K. R. Prasanna Kumar, and Pabitra Mitra. 2014. Recent developments in spoken term detection: a survey. *International Journal of Speech Technology*, 17(2):183–198.

Gautam Mantena, Sivanand Achanta, and Kishore Prahallad. 2014. Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(5):946–955.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 498–502. ISCA.

Prajyot Naik, Manisha Naik Gaonkar, Veena Thenkanidiyoor, and A. D. Dileep. 2020. Kernel based matching and a novel training approach for cnn-based qbe-std. In *2020 International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5.

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)*, pages 5206–5210.

Mark A. Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Dhananjay Ram, Lesly Miculicich, and Hervé Bourlard. 2018. CNN based query by example spoken term detection. In *19th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 92–96. ISCA.

Dhananjay Ram, Lesly Miculicich, and Hervé Bourlard. 2020. Neural network based end-to-end query by example spoken term detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1416–1427.

Sreepratha Ram and Hanan Aldarmaki. 2022. Supervised acoustic embeddings and their transferability across languages. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 212–218, Trento, Italy. Association for Computational Linguistics.

Luis J. Rodriguez-Fuentes, Amparo Varona, Mikel Penagarikano, Germán Bordel, and Mireia Diez. 2014. High-performance query-by-example spoken term detection on the sws 2013 evaluation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7819–7823.

Ramon Sanabria, Hao Tang, and Sharon Goldwater. 2023. Analyzing acoustic word embeddings from pre-trained self-supervised speech models. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Shane Settle, Keith D. Levin, Herman Kamper, and Karen Livescu. 2017. Query-by-example search with discriminative neural acoustic word embeddings. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2874–2878. ISCA.

Shane Settle and Karen Livescu. 2016. Discriminative acoustic word embeddings: Tecurrent neural network-based approaches. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 503–510.

Jan Svec, Jan Lehecka, and Lubos Smídl. 2022. Deep LSTM spoken term detection using wav2vec 2.0 recognizer. In *23rd Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1886–1890. ISCA.

Drisya Vasudev, Suryakanth V. Vasudev, K. K. Anish Babu, and K. S. Riyas. 2016. Combined mfcc-fbcc features for unsupervised query-by-example spoken term detection. In *Intelligent Systems Technologies and Applications*, pages 511–519. Springer International Publishing.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5017–5025.

Yu-Hsuan Wang, Hung-Yi Lee, and Lin-Shan Lee. 2018. Segmental audio word2vec: Representing utterances as sequences of vectors with applications in spoken term detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6269–6273.

Yougen Yuan, Cheung-Chi Leung, Lei Xie, Hongjie Chen, Bin Ma, and Haizhou Li. 2018. Learning acoustic word embeddings with temporal context for query-by-example speech search. In *19th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 97–101. ISCA.

Kun Zhang, Zhiyong Wu, Jia Jia, Helen M. Meng, and Binheng Song. 2019. Query-by-example spoken term detection using attentive pooling networks. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1267–1272. IEEE.