Journal of Applied Research and Technology

www.jart.icat.unam.mx

Journal of Applied Research and Technology 22 (2024) 451-470

Original

Cross-modal learning representation using new margin combination for speech recognition task

D. Karim[a]* • M. Abdelkarim[b]

[a]*Research Unite of Analyse and Processing of Electrical and Energetic Systems, Faculty of Sciences of Tunis, University of Tunis El Manar, El Manar-Tunis 2092, Tunisia*
[b]*Reseach Laboratory in Algebra, Numbers Theory and Intelligent Systems, Faculty of Sciences of Monastir, 90 Mohamed V Street, 1002 Monastir, Tunisia*

12 22 2023; accepted 04 08 2024
Available 06 30 2024

Abstract: Cross-modal retrieval aims to elucidate the fusion of information, mimic human learning, and advance the field. The main challenge in cross-modal matching is to build a shared subspace reflecting semantic proximity. Previous works fail to capture asymmetric relevance by adopting symmetric similarity computations. To overcome these shortcomings, an efficient approach called quaternion representation learning (QRL) is introduced for better cross-modal matching. Thus, a better representation of the shared semantics is offered by virtue of its richer representation capacity of the quaternionic space and its strong expressive power.

Transfer learning is a crucial aspect in this context. By leveraging pre-trained models, the knowledge gained from one task or domain can be effectively transferred to another, allowing for improved performance and generalization. In this study, transfer learning is employed to enhance the cross-modal retrieval system. Specifically, a pre-trained ResNet-512 model is utilized in conjunction with the proposed total margin (TM) loss function, which combines the QRL approach with the novel adaptive mean margin (AMM) methodology.

The TM loss function, coupled with the pre-trained ResNet-512 model, is evaluated on the Audio-Visual Arabic Speech Database (AVAS) and the Arabic Visual Speech Database (AVSD), along with other audio-visual datasets. Experimental results demonstrate the effectiveness of the TM loss function in consistently improving performance on both databases. The recall scores (R@k) and mean average precision (mAP) values achieved on the AVAS Database are as follows: R@1: 42.1±0.7, R@2: 70.2±0.1, R@5: 78.5±1.0, and mAP: 53.0±1.1. Similarly, on the AVSD Database, the results are R@1: 41.7±0.3, R@2: 69.2±1.1, R@5: 78.0±0.3, and mAP: 52.7±0.5.

By incorporating transfer learning and the TM loss function into the cross-modal retrieval framework, this study demonstrates the potential for improving clustering efficiency and enhancing visual and speech understanding. The combination of pre-trained models and the TM loss function offers a promising avenue for advancing cross-modal matching techniques and achieving state-of-the-art performance.

Keywords: ™, cross-modal representation, quaternion approach, AVAS, AVSD

*Corresponding author.
E-mail address: dabbabikarim@hotmail.com (D. Karim).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

1. Introduction

Recently, automatic speech recognition (ASR) systems have achieved significant improvements, reaching human parity (Xiong et al., 2016) or even surpassing humans on many clean speech benchmarks (Tüske et al., 2020; Nguyen et al., 2020) thanks to advances in supervised neural models (Amodei et al., 2016). However, these systems still face challenges when exposed to environmental conditions, particularly when voice recordings are corrupted by noise (Afouras et al., 2018a), which can significantly degrade their performance. In recent years, extensive research has been conducted on noise robustness (Kinoshita et al., 2020) to enhance the efficiency of ASR systems in various scenarios.

To address the limitations of traditional ASR systems, a promising research direction is the integration of noise-invariant lip motion information, which combines speaker audio and video streams. This integration, known as audio-visual speech recognition (AVSR), has the potential to improve performance across a wide range of applications and bring audio-visual speech recognition (AVSR) systems closer to human speech perception (Koguchi et al., 2018; MacDonald, 2018).

While the study of AVSR emerged in the previous era (Barbier et al., 2021), which was characterized by the development of simple model architectures and the use of small-scale datasets, the rapid development of model architectures (Xu et al., 2020) and large-scale data collection (Afouras et al., 2018b) has propelled AVSR systems to new levels of performance. However, these architectures are often data-hungry, relying on substantial amounts of labelled training data. This fully supervised nature of existing AVSR models (Zhang et al., 2021), (Anwar et al., 2023; Fenghour et al., 2021) poses a challenge due to the need for expensive labelled data. Consequently, the application of modern AVSR systems may not always be feasible, especially in scenarios where limited labelled data is available, such as for approximately 700 million spoken languages (McCarty & Coronel-Molina, 2016).

In this study, our research contributes to the actual literature on AVSR by introducing the audio-visual HuBERT pre-training model (AV-HuBERT) (Shi et al., 2022a). In contrast to previous approaches that heavily rely on fully supervised AVSR models and expensive labelled data, such as end-to-end sentence-level lipreading (LipNet) (Zhang et al., 2021), Deep End-to-End Lip Reading (Fenghour et al., 2021), and audio-visual speech recognition (AVSpeech) (Anwar et al., 2023) using a deep recurrent neural network, our approach leverages transfer learning and self-supervised learning techniques to enhance the robustness of the ASR system.

Transfer learning plays a crucial role in our strategy by leveraging knowledge acquired from pre-training on large-scale datasets. By utilizing a pre-trained model, specifically the ResNet-512 model, we can transfer the learned representations and knowledge to our AV-HuBERT model. This transfer of knowledge allows us to initialize our model with meaningful weights, enabling it to capture high-level features and generalize well to the AVSR task. Transfer learning helps to mitigate the data-hungry nature of AVSR models and improves their performance, even when labelled data is limited.

Furthermore, our research incorporates the QRL approach, which contributes to the effectiveness of transfer learning in the AVSR framework. By adopting QRL, we aim to harness its potent ability to express strong and rich representations, making it suitable for learning versatile representations for heterogeneous data. In contrast to conventional inner product-based methods, we explore the use of the Hamilton product in QRL to compute similarity, facilitating the efficient capture of asymmetric relevance in audio-visual speech data (Shi et al., 2022b). This integration of QRL strengthens the representation capacity of the AV-HuBERT model and enhances its ability to capture the nuanced relationships between lip movements and associated sounds.

Additionally, we introduce the adaptive mean margin (AMM) contrastive learning methodology, which is combined with the quaternion loss function, resulting in the total margin (TM) loss. This combination further improves the clustering efficiency of our proposed AV-HuBERT model. By optimizing the TM loss function, we generate a better distribution of features in the high-dimensional space, leading to enhanced visual and speech understanding.

By integrating transfer learning, QRL, and the AMM methodology into our AVSR framework, we address the limitations of actual AVSR models that heavily rely on substantial amounts of labelled data. Our approach reduces the dependency on labelled data, making AVSR systems more accessible and applicable, particularly in scenarios where limited labelled data is available or for under-resourced languages.

In the subsequent sections of this article, we will present related studies, followed by an overview of the methodology employed in our investigation (Section 2). This will be followed by a description of the materials used and the presentation of the results from our experiments (Section 3). Finally, in the last section, we will provide the main conclusions and summaries, emphasizing the significance of our contributions to the field of AVSR, particularly through the incorporation of transfer learning and advanced methodologies.

2. Related works

Self-supervised learning is an unsupervised visual representation learning technique that allows for obtaining features without manual labelling. As a result, the performance gap compared to supervised pre-training in speech recognition (Artetxe et al., 2018) and computer vision (Chen et al., 2020), has been quickly closed. In computer vision, many recent state-of-the-art methods rely on the instance discrimination task, which treats each dataset image (or "instance") and its transformations as distinct classes (Dosovitskiy et al., 2014). This task produces representations that can discriminate between different images while being invariant to image transformations. Recent self-supervised approaches that utilize instance discrimination depend on a combination of two components: (i) contrastive loss (Hadsell et al., 2006) and (ii) a set of image transformations. The contrastive loss compares features directly, eliminating the need for instance classes, while image transformations capture invariances encoded in the features. Both components are crucial for improving the quality of the resulting networks (Kheddar et al., 2023; Djeffal et al., 2023), enhancing performance on transformations and the objective function. The contrastive loss principle is based on comparing pairs of image representations by maximizing the distance between representations of different images and minimizing the distance between representations of transformations of the same image.

In speech recognition, the complex nature of speech, which contains intricate features, motivates the use of contrastive approaches. Unlike computer vision, contrastive models for speech recognition learn representations by differentiating a target sample (positive) from distractor samples (negatives) given an anchor representation (Mohamed et al., 2022). The objective is to maximize the similarity in the latent space between the anchor and the positive samples while minimizing the similarity between the anchor and the negative samples. The negative samples, in this case, refer to data points or instances that are considered dissimilar or unrelated to the anchor, and they are represented as embeddings or feature vectors.

One contrastive model for speech recognition is Contrastive Predictive Coding (CPC) (Oord et al., 2018). The latter is based on the noise-contrastive estimation (InfoNCE) loss, which maximizes the similarity between a localized representation and a contextualized representation by measuring their mutual information. CPC has achieved the best classification accuracy for Question Type

Classification (TREC) (Li & Roth, 2002), reaching a value of 96.8% in tests on five common natural language processing (NLP) benchmarks. Another approach, called wav2vec (Schneider et al., 2019), uses a quantization module instead of positive and negative samples to obtain a discrete representation. This approach aims to avoid finding negative samples in the same category as the positives. Wav2vec achieved a Word Error Rate (WER) of 2.43% on the nov92 test set (Anoop & Ramakrishnan, 2019) and outperformed deep speech 2 (Amodei et al., 2016), the best-character-based system in the literature, while using two orders of magnitude less labelled training data (Li & Roth, 2002). An extension of wav2vec, called wav2vec-C (Sadhu et al., 2021), incorporates a consistency term in the loss to construct input features using learned quantized representations as a vector-quantized variational autoencoder (VQ-VAE) (Razavi et al., 2019). This model achieved a 1.4% relative WER reduction compared to the baseline and a 0.7% reduction compared to wav2vec 2.0 after fine-tuning with recurrent neural network transducers (RNN-T). The achievement was demonstrated on a self-supervised task using noisy far-field real-world data and 1k hours of data for supervised ASR training. VQ-VAE, an extension of variational autoencoder (VAE), has been shown to successfully combine self-supervised learning and discrete latent space for modeling spoken languages (Polyak et al., 2021). Furthermore, combining wav2vec with HuBERT, known as wav2vec-BERT, has further improved the results (Chung et al., 2021). In Baevski et al. (2019), the authors used the vq-wav2vec approach to learn discrete representations via quantization. This approach improved the state of the art on the WSJ and TIMIT benchmarks by leveraging BERT pre-training. Bidirectional CPC also demonstrated good downstream performance on the LS corpus and a diverse speech corpus from multiple sources (Kawakami et al., 2020). A modified CPC method was proposed in (Riviere et al., 2020), where it was pre-trained on 360 hours of unlabeled Librispeech data. The main conclusion of this work is that unsupervised pre-training can be on par with supervised pre-training when sufficient data is available. The speech SimCLR approach suggested in Jiang et al. (2020) offers a new self-supervised objective for learning speech representation by applying augmentation on raw speech and its spectrogram. This approach has achieved competitive results in speech recognition and speech emotion recognition.

While representations learned by contrastive methods have shown good viability in various downstream applications, they still face challenges when applied to speech data. One challenge involves the strategy used to define positive and negative samples, which indirectly enforces invariances on the learned representations. Another challenge is related to speech input, which lacks explicit segmentation of acoustic units. Instead of representing an entire linguistic unit, positive and negative samples only capture partial or multiple units based on the span covered by each sample. Furthermore, speech input is characterized by its smoothness and lack of natural segmentation, making it challenging to define an exact

contrastive sampling strategy that accurately assigns samples to the corresponding anchor. Self-supervised learning is an unsupervised visual representation learning technique that enables the acquisition of features without relying on manual labeling. This approach has significantly narrowed the performance gap compared to supervised pre-training in fields such as speech recognition and computer vision. In computer vision, many state-of-the-art methods utilize instance discrimination, where each image and its transformations are treated as distinct classes. By comparing features directly using contrastive loss and incorporating image transformations, these approaches generate representations that can discriminate between different images while being invariant to transformations. Contrastive models for speech recognition, on the other hand, differentiate between target and distractor samples using an anchor representation. The objective is to maximize the similarity between the anchor and positive samples while minimizing the similarity with negative samples.

One popular contrastive model for speech recognition is Contrastive Predictive Coding (CPC), which maximizes the similarity between localized and contextualized representations by measuring their mutual information. CPC has achieved impressive results in tasks such as question type classification (TREC). Another approach called wav2vec uses a quantization module to obtain discrete representations and has outperformed previous character-based systems in speech recognition, such as Connectionist Temporal Classification (CTC) (Liu et al., 2018), sequence-to-sequence learning with neural networks (Sutskever et al., 2014), and deep speech (Amodei et al., 2016). wav2vec-C incorporates a consistency term in the loss to construct input features using learned quantized representations as a vector-quantized variational autoencoder (VQ-VAE). This approach has demonstrated improved performance compared to the baseline. VQ-VAE itself has successfully combined self-supervised learning and discrete latent space for modelling spoken languages. Combining wav2vec with HuBERT, known as w2v-BERT, has further enhanced the results. Other methods such as vq-wav2vec, bidirectional CPC, and speech SimCLR have also shown promise in speech recognition tasks.

While contrastive learning has proven effective in various downstream applications, it faces challenges when applied to speech data. One challenge is defining positive and negative samples to enforce invariances in the learned representations. Speech input lacks explicit segmentation of acoustic units, making it difficult to represent entire linguistic units accurately. Additionally, the smoothness and lack of natural segmentation in speech data pose challenges in defining an appropriate contrastive sampling strategy.

On the contrary, leveraging multiple modalities can prove beneficial across various settings, where each modality can offer complementary information about the others. Historically, supervised multimodal methods have been integrated for decades for tasks such as person identification (Aleksic & Katsaggelos, 2006) and ASR audio-visual tasks (Potamianos et al., 2003). However, current trends favor unsupervised multimodal techniques in ASR and related fields. Both techniques are deemed useful for mitigating the effects of noise, like supervised methods, as noise tends to be independent and uncorrelated across different modalities. Additionally, combining speech data with image or video signals can enhance the learning of representations, thereby encoding more semantic information. Furthermore, supplementary information can be gleaned from contextual signals, albeit this may occasionally corrupt speech content.

Early computational approaches to multimodal language learning, inspired by human language acquisition, focused on the integration of visual cues (Mercado III et al., 2014). These approaches can be classified into two main categories: Intrinsic and extrinsic (Djeffal et al., 2023). Intrinsic approaches encompass modalities produced by the speech source, such as images or videos of the speaker's articulatory flesh point (Narayanan et al., 2011), lip movement (Shi et al., 2022b), face (Chung & Zisserman, 2017), or simultaneous magnetic resonance imaging (MRI) scans (Narayanan et al., 2011). Learning multiple intrinsic modalities aims to enhance robustness to noise, given the uncorrelated relationship acoustic noise has with other modalities. This type of representation learning, termed multi-view learning, involves approaches such as AV-HuBERT (Hadsell et al., 2006), audio-visual extensions of masked prediction methods (Hadsell et al., 2006; Shi et al., 2022b), and multi-view contrastive losses (Wang et al., 2015).

Conversely, extrinsic modalities, though not produced by the same source, can provide context for each other. A common example is the spoken caption paired with its image. Extrinsic approaches include learning a neural representation model for each modality using multimodal contrastive loss, where the same representation is assigned for paired examples while keeping unpaired ones different across modalities (Peng & Harwath, 2022a). Training with a masked prediction loss (Chan et al., 2022) or a masked margin SoftMax loss (Sanabria et al., 2021) presents alternative options for extrinsic modalities. Typically, evaluation in this case entails cross-modal retrieval, although modalities can be utilized for other downstream tasks, such as SUPERB and zero speech benchmark tasks (Peng & Harwath, 2022b). Analyses of different models have revealed that despite the overarching learning goal of matching speech to corresponding images (or other contextual modalities), these models typically learn

multiple levels of linguistic representations, from shallow to deep layers of models (Harwath et al., 2019). They are also capable of learning word-like units (Wang & Hasegawa-Johnson, 2020) and can be explored for multilingual research, with the visual signal acting as an "interlingua" (Harwath et al., 2018). In certain contexts, even with some textual supervision (i.e., transcribed speech), visual anchoring consistently improves representation learning (Pasad et al., 2019).

Moreover, there is a growing interest in learning joint representations of speech and text using paired and unpaired data. The SLAM approach (Bapna et al., 2021) exemplifies such endeavors for speech and text representation, employing two separate pre-trained encoders followed by a multimodal encoder to construct joint representations. The entire model is trained using a multi-task loss comprising two supervised and two self-supervised tasks.

However, exploring multimodal methods presents key challenges, including the limited supply of multimodal data compared to single-modal data and the specificity of multimodal data, often drawn from fields such as visual scene descriptions. Additionally, the extent to which learned speech representations apply to speech fields not necessarily describing or located within visual scenes remains unclear and warrants further study (Hadsell et al., 2006).

These challenges underscore the substantial semantic gap between asymmetric relevance and heterogeneous data. The awareness of asymmetric relevance in similarity computation, compounded by the complexity of intermodal pairing, calls for careful consideration when designing mechanisms to evaluate semantic similarity between cross-modal data. While existing approaches for computing semantic similarity between cross-modal data, such as vector space models, Kernel methods, and deep learning models, rely on symmetric metrics like Euclidean distance and inner product (e.g., cosine function) for similarity computation, these methods struggle to model asymmetric relevance due to their commutativity (Wei et al., 2021). Hence, the dominant cross-modal matching approaches lack mechanisms to capture asymmetric relevance. Recent developments, such as Polysemic Visual Semantic Embedding (PVSE) with multi-head attention (Song & Soleymani, 2019) and Probabilistic Cross-Modal Embedding (PCME) (Chun et al., 2021), offer promising avenues for computing diverse representations. However, these models, though advanced, remain complex and less practical. Furthermore, the application of probabilistic integration may lead to uncontrolled performance outcomes. The inherent limitation of learned representations in real space, which may overlook the complex semantic interaction between different modalities, further underscores the need for a more expressive representation space to capture asymmetric relevance.

To address these challenges, QRL has emerged as a viable solution to explicitly model asymmetric relevance in quaternion space, driven by complex-valued representation. This approach has garnered significant interest among researchers in various deep learning domains (Tu et al., 2020). QRL introduces a quaternion space composed of hypercomplex values with three imaginary components to represent features.

3. Materials and methods

3.1. Methods

To summarize, the proposed cross-modal speech recognition model, as depicted in Figure 1, consists of four main components: pre-processing, encoder representation, cross-modal quaternion, and decoder mechanism.

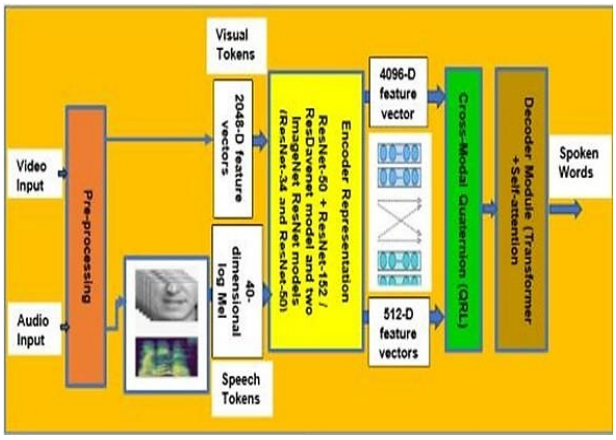

Figure 1. An overview of the cross-modal speech recognition system.

In the pre-processing stage, the video and voice inputs are transformed into visual and voice tokens. To accomplish this, image and video encoders are employed, including a ResNet-50 model pre-trained on M-MiT with TSM and a ResNet-152 model pre-trained on ImageNet. These encoders generate 2048-D feature vectors for each input video. Additionally, a 40-dimensional log Mel spectrogram is calculated as the speech representation input, which is then processed by the ResDavenet model and two ImageNet ResNet models (ResNet-34 and ResNet-50).

Moving on to the encoder representation stage, independent feature extraction methods are applied to represent each modality. The video encoders produce 4096-D feature vectors, while the speech model generates 512-D feature vectors.

In the cross-modal quaternion stage, the model explores correlation modeling to learn common representations from the multi-modal inputs. This approach aims to capture relationships and correlations between speech and visual information, thereby enhancing the understanding of cross-

modal interactions. To achieve this, the model incorporates the QRL method, leveraging the concept of quaternionic space. The QRL method explicitly addresses non-symmetric correlations and cross-modal correspondence, considering the inherent asymmetry in the relationships between modalities. By doing so, the model learns representations for separate modalities and effectively exploits the unique characteristics of cross-modal interactions, resulting in improved performance for speech and visual language tasks.

Finally, the decoder module employs a transformer architecture to decode the spoken words. This transformer architecture allows for word generation through a mechanism of self-attention and cross-attention between vision and language.

3.1.1. Pre-processing

To reduce noise and prepare for further processing, the input data is pre-processed. This stage consists of converting the video and voice inputs into visual and voice tokens. There are differences between the two modalities, so pre-processing will make a differentiation.

3.1.2. Encoder representation

This second period consists in representing each modality independently by exploiting feature extraction methods. The visual and speech inputs are gathered by the encoder stage, and then intermediate states are generated to encode the semantic content. Following the embeddings, the most common methodologies for constructing an encoder are to use LSTM, convolution, and other techniques to encode speech sequences. Regarding speech representation, word embeddings, positional embeddings, and segment embeddings are introduced into the BERT encoder. Additionally, a series of features, such as the image representation, are aligned with a speech representation. In this scenario, the patch, grid, and region functionality are taken out of the visual domain.

Pre-training models in visual language ensure the combination of feature extraction and fusion with pre-training tasks. These parts address diverse challenges, such as quantifying the speech and image, and passing them to the model for training, managing representational interaction challenges, and creating pre-training tasks to help models learn alignment information. Pre-training on large-scale data can learn semantic correlation through separate modalities, addressing the problem of hard access to expensive manual annotations. There are two basic pre-training choices when it comes to merging encoders and dual encoders to regroup information into paired data. The single encoder enhances BERT input, while the dual encoders effectuating co-/cross-BERT. Studies have proven that the single-stream design directly drives self-attention across two modalities, neglecting intra-modality interactions. Therefore, dual-stream architecture has been advocated and adopted by several researchers to characterize cross-modal interactions. In contrast to single-stream architectures, dual-stream architectures explore a cross-modal mechanism for modeling two unidirectional cross-attention sublayers. Sub-layers are characteristically composed of a cross-attention layer. They transferred information and harmonized semantics. In this event, the parameters are commonly distributed between two sub-layers and the contextualized embedding information is learned by separate transformers.

Some researchers expand the integration of segments from various sources to specify the input elements, i.e., VL-BERT and VisualBERT. Dual-stream models encompass ViLBERT (Zhang et al., 2021), UNIMO (Quan et al., 2021), CLIP (Dong et al., 2022), ViLLA (Anwaar et al., 2021), LXMERT (Wu et al., 2021), Lightning Dot (Bi et al., 2022), ALIGN (Li et al., 2019), WenLan1.0 (Wang et al., 2023), COTS (Gupta et al., 2021) ALBEF (Liu et al., 2017), and ERNIE-ViL (Liu et al., 2019). In ViLBERT, the co-transformer manages a two-stream interaction. Additionally, the construction has been revealed for interactivity, especially considering the speech context when interpreting the image. Moreover, LXMERT is identical to ViLBERT in the pre-training model. UNIMO produces innovative ideas, which consider both single mode and multiple modes to create a fusion of features. As for ViLLA, it uses adversarial training in the pre-training and fine-tuning phases. Adversarial training can strengthen the model generalizes better, allowing performance at the fine-tuning stage. ALBEF exhibits two categories, producing strong single-peak and multi-peak representations with improved retrieval and rationale ability. Lightning Dot's study suggests converting expensive attentional mechanisms into three types of learning goals.

3.1.2.1. Feature representation

Visual representation: The QRL method can be precisely explored as a plug-in module without the need for additional layers to learn the representation. Consequently, the input features of the different modalities are identical to the baseline. For an image V, it is represented as in (Liu et al., 2019) by a set of salient region features $V = \{w_1, \dots, w_m\}, w_k \epsilon\, R^{d_w}$, where regions detection is performed using a faster-CNN pre-trained on visual genomes with bottom-up attention, then they are fed to pre-trained ResNet-101 to extract the different features. Next, a d_w-dimensional feature with an fc layer in real space is obtained such that:

$$w_k = W.\big(CNN(V)\big) + b \tag{1}$$

where CNN(.) ensures the extraction of regional features in the boundary boxes, and W and b represent the learned parameters for the f (c) layer.

Two image and video encoders were explored to represent the input videos: a ResNet-50 model (Flamant et al., 2021) Temporal Shift Module (TSM) pre-trained on M-MiT (Trabelsi et al., 2017) and a ResNet-152 model (Dong et al., 2022) pre-trained on ImageNet (Yang et al., 2022). Each encoder produces a 2048-D feature vector after applying the max-Pooling operation on the temporal dimension (8 frames for the TSM (~3 ips) and 3 frames for the image model (1 ips)). The two 2048-D vectors were concatenated and fed into a multi-layer perceptron (MLP) projection head to achieve the final visual representation of the 4096-D. Indeed, the structural diagram of the ResNet-20 and ResNet-152 video encoders is given in Figure 2.

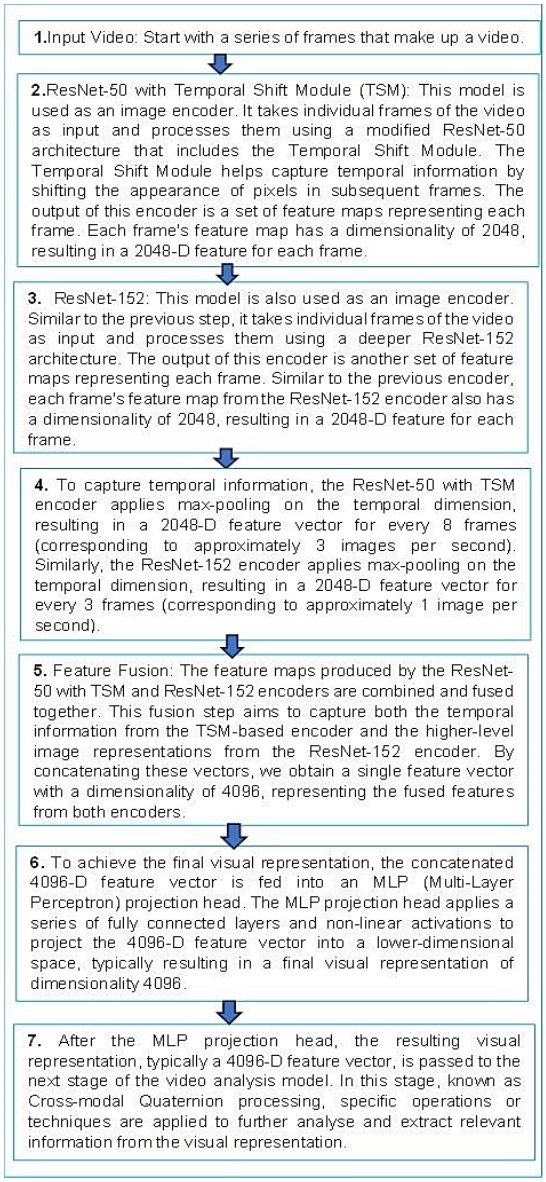

Figure 2. Structural diagram of the ResNet-50 and ResNet-152 video encoders.

Speech representation: Another set T of speech features such as T={e_1,…,e_n },e_j ∈ R^(d_e) is used to denote a caption. Each word is first represented as a single vector and then integrated into d_e dimensional characteristic with a bidirectional GRU. The characteristic of the jth word is averaged by the GRU hidden states back and forth at the jth step.

Also, the models were trained with raw speech sequences instead of the corresponding transcription. For each sequence, every 10 seconds of speech was sampled to be used for training, and then a 40-dimensional log Mel spectrogram was calculated to serve as input to the speech model. These models are ResDavenet (Wu et al., 2021) and two ImageNet ResNet models (Dong et al., 2022) (ResNet-34, ResNet-50) where the first convolutional layer has been modified to take the 1-channel input for the ResNet models so that the spectrograms can be processed. Additionally, the wav2vec model (Guo et al., 2021) was also implied in our experiments, which takes the input in the form of raw waveform. The speech sequences are first inputted into the pre-trained wav2vec model, which generates 512-D vectors per 210ms. Then, they are fed into a learnable ResStack, generated from ResDavenet, with the aim of learning speech sequence representations.

3.1.3. Cross-modal quaternion

Many works have been consecrated to the representation through the modeling of multimodal interactions. For learning common representations, correlation modeling has been explored based on multimodal representations. In fact, cross-modal interaction sustains further interactions between the two different modalities to enhance speech and visual language tasks. However, the degree of cross-modal information merging differentiates between cross-modal mechanisms, such as self-attention.

The quaternion neural factorization machine (Chen et al., 2021) was explored as a solution to model the complex interaction of features. Complex-valued networks also allowed interpretations of physical meanings with well-limited constraints (Li et al., 2019). Undeniably, many works (Wang, Wang et al., 2021) have made great strides, whether exploring only the phase information of complex-valued vectors or designing a diversity of complicated structures, involving novel activation, and consecrated neural units for their completely hyper/complex valued networks to ameliorate representational capacity. Nevertheless, most cross-modal matching approaches rely on real-valued networks, and thus these devoted strategies may restrict generalization. To overcome these limitations, the QRL method explores only the concept of quaternionic space and the mathematical characteristic of the Hamilton product, i.e., all calculations are performed in real space. Therefore, this characteristic allows us to simply hybridize the QRL method with the present cross-

modal matching methods. Contrary to common works, the use of quaternionic space allows us to learn the representations of separate modalities and explicitly takes up those non-symmetric correlations with the asymmetry inherent in the cross-modal correspondence.

3.1.3.1. Quaternion approach

Let consider τ_r a regrouped cross-modal pair training composed of N instances, denoted as $\tau_r = \{(V,T)\}^N$. We denote also $V\epsilon\ R^{d_v}$ as a visual feature for an image or video, and $T\epsilon\ R^{d_t}$ as a speech feature for the corresponding caption. The ultimate objective of the diverse mapping functions is to classify the visual manifestation V which is connected to the semantics of the query T as high as possible, and vice versa. It has been demonstrated that the representation capacity of hypercomplex space is better than that of real space (Reichert & Serre, 2013). Meanwhile, the conventional inner product lacks the ability to model asymmetric relevance well due to its symmetry. Hence, the QRL approach is formulated to learn quaternion representations for cross-modal correspondence and determine a solution to the asymmetric relevance problem.

3.1.3.2. Intra-modal similarity

In real space, a fully connected (f_c) layer is used to calculate the similarity of intra-modal data. A similar quaternion-based layer can also be explored, where Q_w∈H represents the weight parameter as real space. For an input Q_in∈H^(d_in), the formulation of the quaternion layer f_c is given as follows:

$$Q_{out} = Q_w \otimes Q_{in} \tag{2}$$

where the Hamilton product operation ⊗ is used to improve the representation of the quaternion input.

3.1.3.3. Inter-modal similarity

By far, most of the cross-modal similarity is determined based on attentional strategies (Wang, Kou et al., 2021). Hence, a quaternionic attention is defined for the cross-modal similarity. Let us consider $Q_A \in H^{d_a}$ and $Q_B \in H^{d_b}$ define the quaternions representations of two features resulting from identical or different similarities, the score function Q_A and Q_B can therefore be computed using the Hamilton product in quaternion space, such as:

$$Q_{Atten} = Q_A^T \otimes Q_B \tag{3}$$

Hamilton's product is intrinsically non-commutative and is perfect for modeling asymmetric relevance. Thus, we can use this characteristic of mathematics to compute the similarity between diversified data rather than the inner product in real space, which represents the main difference from output methods (Wei et al., 2021).

3.1.3.4. Quaternion representation learning

Even though the QRL approach can be worked in the space called quaternion, its inputs and outputs are all considered as real-valued representations. This is almost identical to similar works (Parcollet et al., 2018a; Li et al., 2019), which have the objective of building a proficient and miniaturized network for less parameter learning. Thus, a modeling of asymmetric relevance is used with the mathematical characteristics of the product of Hamilton.

Similarly, for recent works, a set of region features V is obtained such that $V = \{w_1, \dots, w_m\}$, where $w_1\epsilon\ R^{d_w}$ denotes an image V in real space in addition to a set of speech features T which is given such that $T = \{e_1, \dots, e_n\}, e_j \in R^{d_e}$.
Without involving differentiation and activation in complex space, and without introducing a specially designed new layer, the QRL technique ensures the mapping of real representations in quaternionic space by equidimensional truncated nonlinear maps, formulated as follows:

$$Q_{w_k} \coloneqq r_{w_k} + a_{w_k} * i + b_{w_k} * j + c_{w_k} * k,$$
$$\left[r_{w_k} \otimes a_{w_k} \otimes c_{w_k}\right] = w_k \tag{4}$$

where ⊗ represents the concatenation operation and $\partial_{w_k} \in H^{d_w/4}$ denotes the quaternion representations of the kth image region. This means that c_{w_k} indicates the real part of Q_{w_k}, whereas r_{w_k}, a_{w_k}, and c_{w_k} are designated for the imaginary components in Q_{w_k}, and all of them possess the same dimension of $r_w/4$. The initialization of Q_{e_j} for speech can also be done in the same way and for fair comparisons the representation of the quaternion is not extended to d_w.

3.1.3.5. Similarity calculation

Actual approaches (Diao et al., 2021; Guo et al., 2021) have conducted region-text alignments through different attentional mechanisms, most of which relied on the inner product (cosine function). Due to its symmetry, the naive attention strategies cannot pick up the asymmetric relevance, and so Hamilton's product in attention mechanisms. More precisely, the quaternion attention strategy is used for each region to address the matching words in the sentence with a weighted sum of all representations of the word quaternions. Due to the transformation of all local features into quaternion space, the local similarity $s(w_k, e_j)$ between Q_{w_k} and Q_{e_j} can be computed by Equation 2. As the Hamilton product also designates a quaternion vector, a component aware SoftMax is then attributed to smooth the four components of the quaternion $s(w_k, e_j)$ (involving one real component and three imaginary components) to generate the attention weights, determined as:

$$Q_{\alpha_{kj}} = componentSoftMax(\lambda * s(w_k, e_j)),$$

$$Q_{e_{kj}} = r_{\alpha_{kj}} r_{e_j} + a_{\alpha_{kj}} a_{e_j} * i + b_{\alpha_{kj}} b_{e_j} * j + c_{\alpha_{kj}} c_{e_j} * k \quad (5)$$

for each attention component, the parameter λ is explored to control its smoothness. Next, concatenation is used to transform the quaternion representation of context $Q_{e_{kj}}$ back into real space, defined as:

$$c_k = \sum_{j=1}^{n} \begin{matrix} c_{kj} = [r_{e_{kj}} \oplus a_{e_{kj}} \oplus b_{e_{kj}} \oplus c_{e_{kj}}], \\ c_{kj}, \end{matrix} \quad (6)$$

where c_k is the weighted context to get a caption.

3.1.3.6. Cross-modal alignment

The feature importance of each region can be determined for a given context c_k in real space using the generated features, such that:

$$R(w_k, c_k) = \frac{w_k^T c_k}{\|w_k^T\| \|c_k\|}. \quad (7)$$

Finally, exploring the modeling of asymmetric relevance in quaternionic space, the similarity between a given visual and speech pair (V, T) can be computed as:

$$S(V,T) = \frac{\sum_{k=1}^{m} R(w_k, c_k)}{m}. \quad (8)$$

Considering that the cross-modal matching problem is a two way recovery process, the commonly used triplet rank loss (Wang, Kou et al., 2021) can be used as an objective function.

$$L(V,T) = [S(V,T^{-1}) - S(V,T) + \Delta]_+ + [S(V^{-1},T) - S(V,T) + \Delta]_+ \quad (9)$$

Where $[x]_+ \equiv (x, 0)$, Δ represents a fixed margin between negative pair $(V,T^{-1}), (V^{-1},T)$ and positive pair (V,T). Moreover, T^{-1} (V^{-1}) defines the most difficult negative instance being given a query V(T).

To improve more the feature alignment by learning to discriminate between positive and negative pairs of feature embeddings, we propose to combine the loss $L(V,T)$ with that of contrastive learning called masked margin SoftMax loss (MMS) (Wei et al., 2021). This function and that of large margin cosine loss (LMCL) (Monfort et al., 2021) allow to integrate a margin in the contrastive learning framework to enhance feature discrimination among non-paired embeddings. Indeed, MMS explores a monotonically increasing margin to permit initial learning to start to converge before a huge loss change is added. However, a theoretical bound has been proposed on the maximum margin size of $1 - cos(\frac{2\pi}{N})$ for LMCL where N denotes the number of discriminated classes. To align speech with visual information, the class size can be considered unlimited because each caption is itself a slightly different representation that one wishes to distinguish by giving a maximum margin size of 1. Practically, MMS can add a margin as follows:

$$L_{VT} = -\frac{1}{B}\sum_{i=1}^{B} log\ log \left(\frac{e^{S(V_i,T_i)-M}}{e^{S(V_i,T_i)-M} + \sum_{j=1}^{B} I_{i\neq j} e^{S(V_i,T_i)}} \right) \quad (10)$$

Where the margin M evolves exponentially every 1000 training steps by a factor of 1.002 starting from an initial value of 0.001.

Table 1 presents the results obtained from the training process using the proposed margin-based contrastive learning approach. The displayed values demonstrate a consistent decrease in the loss metric as the training progresses. This decreasing trend suggests that the model is achieving improved feature alignment and discrimination throughout the training iterations.

Also, we have explored in this study the extension of the idea of margin increases in MMS to an adaptive framework called adaptive mean margin (AMM) (Monfort et al., 2021) which does not require adjustment of the initial value of the margin or the rate of growth. Thus, the loss total L_{Total} which combines the loss L_{VT} with that of $L(V,T)$ referred to as total margin (TM) is given such that:

$$L_{TM} = L_{AMM} + L(V,T).w_{AMM} \quad (11)$$

where w_{AMM} is the weighting of AMM.

Table 1. Training results.

Training step	Low value
1000	0.023
2000	0.018
3000	0.015
4000	0.012
5000	0.010

TM loss function effectively captures the asymmetric relevance between modalities by considering adaptive margins and quaternion representations. Adaptive mean margin (AMM) allows for different margins for different modalities, accommodating their varying levels of relevance. For instance, in speech-to-image cross-modal learning, the margin for speech may be larger than that for images, reflecting speech's higher importance. QRL enables the model to learn modality-specific quaternion representations that capture unique characteristics. This allows the model to better

distinguish between relevant and irrelevant modalities, even when they share similar visual features. By combining QRL and AMM, the TM loss function provides a robust framework for cross-modal learning that effectively navigates the complexity of multimodal data and captures asymmetric relevance. This leads to improved performance in tasks such as cross-modal retrieval, recognition, and generation, tailored to the specific characteristics of the cross-modal learning task.

To tailor QRL and AMM to specifically address the challenges in cross-modal learning for speech recognition, the following customization process can be followed:

1. Quaternion representation learning (QRL)

Use a quaternion-based loss function. The quaternion-based loss function, such as that of the quaternion Frobenius norm, is used to learn quaternion representations that are more discriminative for cross-modal learning. This loss function can be designed to minimize the distance between the quaternion representations of the speech and visual modalities while maximizing the distance between the quaternion representations of the speech modality and other irrelevant modalities.

Use a quaternion-based regularization term. The quaternion-based regularization term is explored to encourage the quaternion representations to be more compact and discriminative. This regularization term can be designed to penalize the quaternion representations that are to spread out or that are too like each other.

2. Adaptive mean margin (AMM)

Use an adaptive margin. The adaptive margin is exploited to dynamically adjust the margin between the positive and negative samples during training. This margin can be adjusted based on the difficulty of the training data, the progress of the training process, or the characteristics of the quaternion representations.

Use a quaternion-based distance metric. The quaternion-based distance metric, such as the quaternion dot product, is used to measure the distance between the quaternion representations of the speech and visual modalities. This distance metric can be designed to consider the unique properties of quaternion representations, such as their non-commutativity and their ability to represent rotations.

By customizing the QRL and AMM in this way, it is possible to address the challenges in cross-modal learning for speech recognition, such as the different modalities of speech and visual data, the lack of correspondence between the two modalities, and the need for discriminative and robust representations.

In fact, the proposed model maintains stability in learning from the cross-modal data while accurately capturing the variability inherent in speech and visual inputs through the following mechanisms:

- Multi-modal fusion: The model uses a multi-modal fusion strategy to combine information from the speech and visual modalities. This helps to stabilize the learning process and prevents the model from overfitting to either modality.
- Data augmentation: The model is trained on a large and diverse dataset of speech and visual data. This helps to ensure that it can generalize well to new data, even if it is noisy or contains variability.
- Regularization: The model uses a variety of regularization techniques to prevent overfitting. These techniques include dropout, weight decay, and early stopping.
- Curriculum learning: The model is trained using a curriculum learning strategy. This means that the model is first trained on easier data and then gradually exposed to more difficult data. This helps to stabilize the learning process and prevents the model from becoming confused by the more complex data.

In addition to these mechanisms, the model also uses the TM novel loss function that is designed to explicitly capture the variability inherent in speech and visual inputs. This loss function encourages the model to learn representations that are both discriminative and robust to noise and variability.

As a result of these mechanisms, the model can maintain stability in learning from the cross-modal data while accurately capturing the variability inherent in speech and visual inputs. This leads to improved performance on a variety of cross-modal speech recognition tasks. data that is collected from a variety of sources, such as YouTube videos, TV shows, and movies. This data is likely to be diverse and noisy, as it will contain a variety of different accents, speaking styles, and visual conditions.

The suggested model can use the techniques described above to learn to capture the variability inherent in this data. For example, the model can use data augmentation to learn to generalize to new and unseen data, and it can use regularization to prevent overfitting. The model can also use multi-modal fusion to learn to capture the complementary information that is available in both the speech and visual modalities.

As a result, the model can learn to accurately capture the variability inherent in speech and visual inputs while maintaining stability. This makes the model more robust to noise and more effective at learning from diverse and noisy datasets.

3.1.4. Decoder module

Following the encoding of the interaction of visual and linguistic features, the next step is to explore intermediate states to decode the spoken words at each stage. Since the

decoder module produces outputs in inference, it is most like the encoder module. There are different decoding approaches, such as convolution, LSTM, GRU, and transformer. For example, LSTM generates every word auto regressive. Additionally, the transformer first permits word generation through a mechanism of self-attention and cross-attention between vision and language. Therefore, the decoder function demurs the encoder mentioned above.

4. Results and discussions

4.1. Materials

4.1.1. Databases

Audio-Visual Arabic Speech (AVAS) database: We explored the Audio-Visual Arabic Speech (AVAS) database (Antar & Sagheer, 2013) which includes different isolated words and a few sentences for the continuous speech recognition task. Additionally, it included samples acquired under different conditions where each sample involves four lighting conditions and five head poses. The AVAS database included approximately 13,850 videos plus 10,000 static images from 50 speakers.

Arabic Visual Speech Database (AVSD): This database (Elrefaei et al., 2019) included approximately 1100 videos collected from 22 speakers for 10 daily communication words. These videos were recorded with Full HD resolution and 30 frames per second. AVSD specified for isolated Modern Standard Arabic (MSA) words where each isolated word begins and ends with a silence. These words are given as follows: "ً تفّضل " tafaddal"," shukran -شكرا -, ""marhaban-مرحبا, " وداعا-wada'an", " "taiyeb-أعتذر" , "طيّب-a'atather," "سلم -salam",- "حسنا Hassanan",-أهلا" ahlan," and آسف -"aassef."

4.2. Implementation details

We have conducted different experiments on different databases. For ResNet-152 (Wang et al., 2023) and ResNet-50 (Lin et al., 2019) models, we have followed the same configurations as in Miech et al. (2019). As regards the size of the aligned cross-modal representation in latent space and in that of quaternionic, it was set at 64 for all databases. For the fully connected layer in addition to that of the nonlinear activation, they were included in the final SoftMax classifier. The input size for this fully connected layer was set to 64 and the output size was determined by the total number of categories.

The implementation of different approaches was conducted using Nvidia Tesla V100 GPUs with 16 clips per GPU. Additionally, Pytorch was used as a framework to develop the various models and the training was done for 100 epochs using the Adam as an optimizer. For the learning rate, it was set at 1.5e-04.

4.3. Results

Table 2 presents the comprehensive results obtained from experiments conducted on the speech recognition system. The evaluation utilized various pre-trained models and a cross-modal representation method to generate feature representations for different captions. The performance of these models was assessed using R@k recall scores (k=1, 2, 5) and mean average precision (mAP) metrics on the AVAS and AVSD databases, employing different loss functions, namely noise-contrastive estimation (NCE) (Djeffal et al., 2023), masked margin Softmax loss (MMS) (Ilharco et al., 2019), Semi-hard negative mining (SHN) (Schroff et al., 2015), AMM (Monfort et al., 2021), and TM.

The results obtained with ResNet-50 and the TM loss function are highly favorable, demonstrating the effectiveness of our speech recognition model. This approach achieves a mAP of 53.0±1.1 on the AVAS database, surpassing the mAP of 52.7±0.5 obtained on the AVSD database. Consequently, it proves to be more effective in recognizing speech in noisy environments.

Furthermore, the approach achieves high recall rates at all ranks on both the AVAS and AVSD databases. This indicates that our approach can correctly identify a considerable proportion of the relevant speech segments in the database. Specifically, our approach achieves an R@1 of 42.1±0.7 on the AVAS database and an R@1 of 41.7±0.3 on the AVSD database. This signifies that our approach can correctly identify the top-ranked relevant speech segment in 42.1% of the queries on the AVAS database and in 41.7% of the queries on the AVSD database. Moreover, our proposed ResNet-152 approach combined with the TM loss function achieves an R@2 of 70.2±0.1 on the AVAS database and an R@2 of 69.2±1.1 on the AVSD database, indicating the ability to correctly identify one of the top two ranked relevant speech segments in 70.2% of the queries on the AVAS database and in 69.2% of the queries on the AVSD database. Additionally, our approach achieves an R@5 of 78.5±1.0 on the AVAS database and 78.0±0.3 on the AVSD database, signifying the ability to correctly identify one of the top five ranked relevant speech segments in 78.5% of the queries on the AVAS database and in 78.0% of the queries on the AVSD database.

Comparing the different loss functions employed in the experiments, it is evident that the proposed TM loss function consistently outperformed the others on both datasets. This indicates that the TM loss function effectively improves the feature representations, leading to better speech recognition performance. Notably, the ResNet-152 language model demonstrates exceptional capabilities in generating strong representations, particularly for the recovery task, as observed in Table 2.

These findings highlight the effectiveness of the proposed TM loss function and the ResNet-152 language model in enhancing the performance of the speech recognition system. The results strongly indicate that the inclusion of the TM loss function and the utilization of the ResNet-152 language model as a backbone can make a substantial contribution to achieving accurate and reliable speech recognition. These improvements are particularly valuable in the context of cross-modal learning representation on the AVAS and AVSD databases.

As observed in Table 2, the TM loss function consistently outperforms the other loss functions in terms of recall scores and mean average precision (mAP) on both the AVAS and AVSD databases. This suggests that the TM loss function is more effective at learning discriminative representations for speech recognition, even when using audio and visual modalities. One explanation for this is that the TM loss function explicitly penalizes the model for making mistakes on hard examples. This encourages the model to focus on learning the most difficult examples, which can lead to improved performance on the overall task, even when using multiple modalities. Another explanation is that the TM loss function encourages the model to learn more compact representations of the data. This can make the model more efficient to train and deploy, and it can also lead to improved generalization performance, even when using multiple modalities.

Table 2. Different results obtained for different Loss functions on both AVAS and AVSD databases.

		AVAS database				AVSD Database			
Spoke caption Model	Loss	R@1	R@2	R@5	mAP	R@1	R@2	R@5	mAP
ResNet-50 (Djeffal, et al., 2023)	NCE	34.5±0.1	61.1±0.5	71.6±1.0	46.7±0.1	33.5±0.1	60.6±0.1	71.2±0.5	46.4±0.1
	MMS	36.1±0.5	62.7±1.5	72.6±1.5	48.3±0.5	35.5±0.5	62.3±1.1	72.0±1.1	48.0±0.5
	SHN	34.2±0.4	60.4±0.5	70.5±0.7	46.5±0.8	33.5±0.2	60.0±0.4	69.7±0.2	46.0±0.8
	AMM	38.6±1.1	66.0±1.1	74.2±0.5	51.0±1.2	38.1±1.1	65.6±1.1	74.0±0.5	50.6±1.0
	TM	40.3±0.7	68.2±0.5	76.4±1.2	53.0±1.1	40.0±0.5	67.7±0.5	75.9±1.0	52.6±1.5
ResNet-152 (Elrefaei, et al., 2019)	NCE	36.5±0.5	63.0±0.1	73.2±1.1	46.7±0.1	36.1±0.1	62.4±0.1	73.0±1.0	46.4±0.7
	MMS	38.1±0.3	64.5±1.3	74.6±1.0	48.3±0.5	37.6±0.3	64.0±1.3	74.0±0.2	48.0±0.3
	SHN	36.2±0.1	62.8±0.9	72.5±0.7	46.5±0.8	35.3±0.1	62.6±0.9	72.0±0.1	46.5±0.8
	AMM	40.6±0.7	68.0±1.7	76.2±0.5	51.0±1.2	40.2±0.7	67.6±1.2	75.7±0.1	50.8±1.0
	TM	**42.1±0.7**	**70.2±0.1**	**78.5±1.0**	**53.0±1.1**	**41.7±0.3**	**69.2±1.1**	**78.0±0.3**	**52.7±0.5**
wav2vec (Schneider et al., 2019)	NCE	33.7±0.1	60.1±0.5	70.6±1.0	45.7±0.1	33.2±0.1	59.7±0.2	70.2±0.5	45.4±0.3
	MMS	35.1±0.5	61.7±1.5	71.6±1.5	47.3±0.5	34.8±0.5	61.2±1.3	71.0±1.2	47.0±0.8
	SHN	33.2±0.4	59.4±0.5	69.5±0.7	45.5±0.8	32.9±0.1	59.0±0.5	69.3±0.5	45.2±0.8
	AMM	37.6±1.1	65.0±1.1	73.2±0.5	50.0±1.2	37.1±1.0	64.6±1.1	72.7±0.3	49.9±1.0
	TM	39.3±0.7	67.2±0.5	75.4±1.2	52.0±1.1	38.8±0.3	66.7±0.5	74.7±1.0	51.6±0.5
ResDavenet (Harwath et al., 2018)	NCE	32.4±0.2	59.1±0.6	69.6±1.2	44.9±0.3	32.4±0.2	58.5±0.6	69.1±1.0	44.5±0.1
	MMS	34.1±0.7	61.4±1.0	70.6±1.2	46.4±0.6	34.1±0.7	60.9±1.0	70.0±1.0	46.0±0.6
	SHN	32.2±0.5	58.7±0.8	68.8±0.3	44.6±1.1	32.2±0.5	58.2±0.8	68.2±0.1	44.1±1.1
	AMM	36.6±1.5	64.0±1.7	72.2±1.0	49.0±1.0	36.6±1.5	63.8±1.7	71.7±1.0	48.6±1.0
	TM	38.3±0.5	66.2±0.7	74.4±1.5	51.0±1.3	38.3±0.5	65.9±0.7	74.0±1.2	50.6±1.3

The two Figure 3(a) and (b) show the results of an experiment that was conducted on the two AVAS and AVSD datasets to compare the performance in terms of accuracy of different pre-trained models combined with various loss functions. The results of this experiment suggest that the ResNet-152-TM model is the best performing model for both the AVAS and AVSD datasets. This model achieved the highest accuracy on both datasets, reaching 93.5% and 91.2%, respectively, and outperformed the other models by a significant margin. The second-best performing model was the ResNet-152-AMM model, which achieved an accuracy of 93.2% and 90.9%, respectively. The third best performing model was the ResNet-152-SHN model, which achieved an accuracy of 93.1% and 90.8%, respectively.

It is worth noting that the TM loss function performed well compared to other loss functions on both databases, followed by AMM and SHN. The performance order on AVAS and AVSD datasets for the different explored pre-trained models with the different combinations of loss functions places ResNet-152 first, followed by ResDavenet, ResNet-50, and wav2vec.

As regards the convergence speed of the different pre-trained models combined with various loss functions, it can be analyzed by observing the number of epochs required for the model to reach a certain level of accuracy. The Resnet-152-TM converges the fastest, followed by Resnet-152-AMM and Resnet-152-SHN. The same loss function performance order is observed with the different pre-trained models ResDavenet, ResNet-50, and wav2vec.

In terms of loss functions, the models with the lowest loss values converge faster. For this reason, the models with the TM loss function converge faster than the models with the AMM, SHN, and NCE loss functions. This is because the TM loss function is more discriminative than the other loss functions, which makes the model learn faster.

To examine the strength of our proposed model in AVAS and AVSD databases, we compare the generalization performance of the models trained on four different datasets (ViLLA as well as UNIMO (Quan et al., 2021), ALBEF (Liu et al., 2017) and Lightning Dot (Bi et al., 2022) for video/caption recovery. Each model was trained on a single data set using the approach described in section 2.1.3 and then evaluated on the test set of each other data set. This allows us to fairly compare results between test sets of varied sizes.

Each model in this assessment was trained using the ALBEF language model and the proposed TM loss function which was found to give the best results (as shown in Tables 2 and 3). In Table 3, we can see that the ALBEF model generalizes better than the other models despite the additional noise introduced by the ASR model.

4.4. Discussions

We can notice according to the various tests conducted that there is an effect due to the increase in the total margin (TM) making an increase of the difference between the similarity of true pairs and that of negative pairs. Also, we can notice that each time there is progress in training and a convergence in learning approaches, this is accompanied by an increase in the margin by increasing the distinction between positive and negative similarities by pairs. This also eliminates the need to adjust margin and growth rate which may have various optimal values for diverse similarity metrics, batch sizes, and data sets.

Additionally, the QRL approach explored with TM loss helped improve performance by better capturing asymmetric relevance. This was achieved not only on Arabic databases but also on other databases exploring pre-trained models. Also, there are other works done on Arabic databases (Antar & Sagheer, 2013; Elrefaei et al., 2019) but we are not able to compare the spoken captioning models in Table 3 here to those works because those datasets only include visual or voice captions.

To support the superiority of quaternion space over real space in the context of speech recognition, there are some studies that have been performed and provide empirical evidence. The work of (Parcollet et al., 2018b) investigated modern quaternion-valued models, including convolutional and recurrent quaternion neural networks, using the TIMIT dataset. The experiments demonstrate that the quaternion

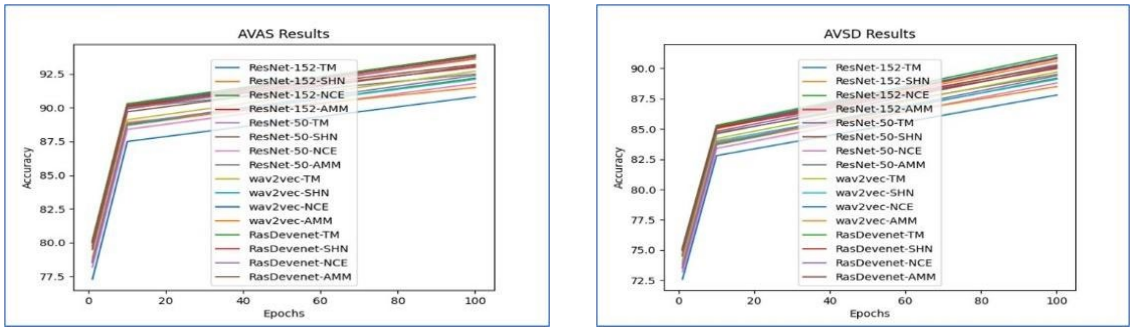

Figure 3. The results of accuracy obtained with different combinations of pre-trained models with various loss function on both AVAS and VSD databases.

neural networks (QNNs) consistently outperformed their real-valued counterparts, achieving better performance with significantly fewer learning parameters. This leads to a more efficient, compact, and expressive representation of relevant information. Another study (Qiu et al, 2020) explored the use of a quaternion long-short term memory neural network (QLSTM) for multi-channel distant speech recognition. The QLSTM, trained on concatenated multi-channel speech signals, outperformed equivalent real-valued LSTMs in this context.

Table 4 shows the results achieved with the state-of-the-art models explored with different modalities
and learning approaches for the speech recognition task. From this table, we can see that our approach, the ResNet-152 model combined with the TM loss function, achieved competitive results compared to these approaches.

It is important to note that the other state-of-the-art methods listed in Table 4 use different modalities and learning approaches. For example, the Conformer-Transducer model uses audio data and a supervised learning approach, while the multimodal dual recurrent encoder (MDRE) model uses text and audio data and a multimodal learning approach. Despite these differences, the ResNet-152 model with the TM loss function outperforms the other methods on both the AVAS and AVSD databases. This suggests that the TM loss function is a powerful approach for learning discriminative representations for speech recognition, regardless of the modality or learning approach used.

Overall, the ResNet-152 model with the TM loss function is a promising innovative approach for speech recognition. It outperforms other state-of-the-art methods in terms of accuracy and robustness to noise, and it is particularly well-suited for learning from diverse and noisy datasets, as it can capture the variability inherent in speech and visual inputs.

The ResNet-152 model with the TM loss function is also computationally efficient and easy to train, making it a practical choice for real-world applications. Additionally, it can be used for a variety of speech recognition tasks, including speaker recognition, speech emotion recognition,
and multimodal speech recognition.

Table 3. Results obtained with different cross-evaluations of video/audio recovery datasets.

	Evaluated on															
Trained on	ALBEF				ViLLA				UNIMO				Lightning Dot			
	R@1	R@2	R@5	mAP	R@1	R@2	R@5	mAP	R@1	R@2	R@5	mAP	R@1	R@2	R@5	mAP
ALBEF (Liu et al., 2017)	**44.4**	**77.2**	**84.2**	**59.1**	18.6	**54.8**	**55.5**	**30.9**	**34.1**	**65.6**	**78.4**	**48.4**	**40.2**	**69.7**	**79.7**	**53.2**
ViLLA (Anwar et al., 2023)	27.0	58.0	70.2	41.1	20.3	49.6	62.4	33.7	16.6	38.2	51.1	27.1	10.6	29.2	41.1	20.5
UNIMO (Harwath et al., 2018)	21.0	51.3	66.5	39.0	10.4	28.9	40.9	20.2	30.3	65.0	78.7	45.6	15.3	40.3	54.5	27.3
Lightning Dot	42.9	74.6	82.5	57.3	16.5	40.0	52.8	28.4	23.0	50.8	64.1	36.9	14.6	34.1	46.6	24.4

Table 4. Different results obtained with different modalities and learning approaches on different databases for speech recognition task.

Method	Modality	Database	Accuracy (%)	
Conformer-Transducer Models For child speech recognition (Barcovschi et al., 2023)	Audio	Child Speech Corpus	90	
Deep Learning Techniques for Speech Emotion Classification (Akinpelu & Viriri, 2022)	Audio, Text, Visual	Various speaker databases (IEMOCAP, MSP-IMPROV, etc.)	IEMOCAP	MSP-IMPROV
			75	82
Windows for Speech Emotion Recognition (Teixeira et al., 2024)	Audio	IEMOCAP, MSP-IMPROV, etc	64.1 (4 to 10 emotions)	
Multimodal Dual Recurrent Encoder (MDRE) (Kreplak López, 2020)	Text, Audio	IEMOCAP	71.8	
Multimodal Speech Emotion Recognition Using Modality Fusion (Ho et al., 2020)	Speech, Audio, Text, Images, Videos	IEMOCAP	77.58	
Improving Multimodal Speech Recognition by Data Fusion (Oneață & Cucu, 2022)	Speech, Visual	LibriSpeech clean test set and LibriSpeech other test set	word error rate (WER) 2.6% and 6.0%	

5. Conclusions and further works

The QRL method, presented in this paper, aims to explicitly model asymmetric relevance in quaternion space for cross-modal correspondence. By leveraging the combination of the adaptive mean margin (AMM) and total margin (TM) within the quaternion space, richer representations are learned, facilitating accurate modeling of asymmetric relevance. Experimental results conducted on four widely explored datasets across two cross-modal correspondence tasks demonstrate a significant improvement in performance through the TM loss function. Specifically, employing this TM loss function with ResNet-50 yielded highly favorable results, highlighting the effectiveness of our speech recognition model. For instance, our approach achieved a mAP value of 53.0±1.1 on the AVAS database and 52.7±0.5 on the AVSD database. Furthermore, our method yielded R@1, R@2, and R@5 values of 42.1±0.7, 70.2±0.1, and 78.5±1.0, respectively, on the AVAS database, and R@1, R@2, and R@5 values of 41.7±0.3, 69.2±1.1, and 78.0±0.3, respectively, on the AVSD database.

Additionally, our speech recognition model, which combines ResNet-152 with the TM loss function, has demonstrated competitive results compared to other state-of-the-art speech recognition models designed using different modalities and learning approaches.

As further work, we propose to evaluate the proposed model on other larger databases and explore its implementation using three modalities: text, speech, and visual captions. This will allow us to evaluate the model's performance on a wider range of data and explore its potential for multimodal applications. It is also important to reflect on its application in real-time applications on embedded architectures. This will involve investigating the evolution of the model's parameters and its effectiveness in this condition. By optimizing the model for real-time performance, we can explore its potential for practical applications such as real-time object recognition and speech recognition. Moreover, we plan to investigate the use of the model for cross-modal retrieval tasks. This will involve training the model on a dataset of text, speech, and visual data and evaluating its ability to retrieve relevant items from one modality given a query from another modality. This will allow us to explore the model's potential for applications such as image search and video retrieval.

Conflict of interest

The authors have no conflict of interest to declare.

Acknowledgements

The authors would like to thank the reviewers for their efforts in revising this article and share their persistent comments aimed at improving the quality of this work.

Funding

The authors received no specific funding for this work.

References

Afouras, T., Chung, J. S., & Zisserman, A. (2018a). LRS3-TED: a large-scale dataset for visual speech recognition. https://arxiv.org/abs/1809.00496

Afouras, T., Chung, J. S., & Zisserman, A. (2018b). The conversation: Deep audio-visual speech enhancement. https://arxiv.org/abs/1804.04121

Akinpelu, S., & Viriri, S. (2022). Speech Emotion Classification: A Survey of the State-of-the-Art. In *Pan-African Artificial Intelligence and Smart Systems Conference* (pp. 379-394). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-25271-6_24

Aleksic, P. S., & Katsaggelos, A. K. (2006). Audio-visual biometrics. *Proceedings of the IEEE*, *94*(11), 2025-2044. https://doi.org/10.1109/JPROC.2006.886017

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Zhu, Z. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning* (pp. 173-182). PMLR.

Anoop, C. S., & Ramakrishnan, A. G. (2019). Automatic speech recognition for Sanskrit. In *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)* (Vol. 1, pp. 1146-1151). IEEE. https://doi.org/10.1109/ICICICT46008.2019.8993283

Anwar, M., Shi, B., Goswami, V., Hsu, W. N., Pino, J., & Wang, C. (2023). Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. *arXiv preprint arXiv:2303.00628*. https://arxiv.org/abs/2303.00628

Anwaar, M. U., Labintcev, E., & Kleinsteuber, M. (2021). Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision* (pp. 1140-1149).

Antar, S., & Sagheer, A. (2013). Audio visual Arabic speech (AVAS) database for human-computer interaction applications. *The International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(9).

Artetxe, M., Labaka, G., & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. arXiv preprint arXiv:1805.06297. https://arxiv.org/abs/1805.06297

Baevski, A., Schneider, S., & Auli, M. (2019). vq-wav2vec: Self-supervised learning of discrete speech representations. https://arxiv.org/abs/1910.05453

Bapna, A., Chung, Y. A., Wu, N., Gulati, A., Jia, Y., Clark, J. H., ... & Zhang, Y. (2021). SLAM: A unified encoder for speech and language modeling via speech-text joint pre-training. https://arxiv.org/abs/2110.10329

Barbier, G., Merzouki, R., Bal, M., Baum, S. R., & Shiller, D. M. (2021). Visual feedback of the tongue influences speech adaptation to a physical modification of the oral cavity. *The Journal of the Acoustical Society of America*, *150*(2), 718-733. https://doi.org/10.1121/10.0005520

Barcovschi, A., Jain, R., & Corcoran, P. (2023). A comparative analysis between Conformer-Transducer, Whisper, and wav2vec2 for improving the child speech recognition. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (pp. 42-47). IEEE. https:///doi.org/10.1109/SpeD59241.2023.10314867

Bi, X., Shuai, C., Liu, B., Xiao, B., Li, W., & Gao, X. (2022). Privacy-preserving color image feature extraction by quaternion discrete orthogonal moments. *IEEE Transactions on Information Forensics and Security*, *17*, 1655-1668. https://doi.org/10.1109/TIFS.2022.3170268

Chan, D. M., Ghosh, S., Chakrabarty, D., & Hoffmeister, B. (2022). Multi-modal pre-training for automated speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 246-250). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9746449

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR.

Chen, T., Yin, H., Zhang, X., Huang, Z., Wang, Y., & Wang, M. (2021). Quaternion factorization machines: A lightweight solution to intricate feature interaction modeling. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 4345-4358. https://doi.org/10.1109/TNNLS.2021.3118706

Chung, J. S., & Zisserman, A. (2017). Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13* (pp. 87-103). Springer International Publishing. https://doi.org/10.1007/978-3-319-54184-6_6

Chung, Y. A., Zhang, Y., Han, W., Chiu, C. C., Qin, J., Pang, R., & Wu, Y. (2021). W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 244-250). IEEE. https://doi.org/10.1109/ASRU51503.2021.9688253

Chun, S., Oh, S. J., De Rezende, R. S., Kalantidis, Y., & Larlus, D. (2021). Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8415-8424).

Djeffal, N., Kheddar, H., Addou, D., Mazari, A. C., & Himeur, Y. (2023). Automatic Speech Recognition with BERT and CTC Transformers: A Review. In *2023 2nd International Conference on Electronics, Energy and Measurement (IC2EM)* (Vol. 1, pp. 1-8). IEEE. https://doi.org/10.1109/IC2EM59347.2023.10419784

Diao, H., Zhang, Y., Ma, L., & Lu, H. (2021). Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 2, pp. 1218-1226). https://doi.org/10.1609/aaai.v35i9.16993

Dong, X., Zhang, H., Zhu, L., Nie, L., & Liu, L. (2022). Hierarchical feature aggregation based on transformer for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(9), 6437-6447. https://doi.org/10.1109/TCSVT.2022.3164230

Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., & Brox, T. (2014). Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, *27*.

Elrefaei, L. A., Alhassan, T. Q., & Omar, S. S. (2019). An Arabic visual dataset for visual speech recognition. *Procedia Computer Science*, *163*, 400-409. https://doi.org/10.1016/j.procs.2019.12.122

Fenghour, S., Chen, D., Guo, K., Li, B., & Xiao, P. (2021). Deep learning-based automated lip-reading: A survey. *IEEE Access*, *9*, 121184-121205. https://doi.org/10.1109/ACCESS.2021.3107946

Flamant, J., Miron, S., & Brie, D. (2021). A general framework for constrained convex quaternion optimization. *IEEE Transactions on Signal Processing*, *70*, 254-267. https://doi.org/10.1109/TSP.2021.3137746

Guo, Z., Zhao, J., Jiao, L., Liu, X., & Liu, F. (2021). A universal quaternion hypergraph network for multimodal video question answering. *IEEE Transactions on Multimedia*, *25*, 38-49. https://doi.org/10.1109/TMM.2021.3120544

Gupta, D. K., Arya, D., & Gavves, E. (2021). Rotation equivariant siamese networks for tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12362-12371).

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)* (Vol. 2, pp. 1735-1742). IEEE. https://doi.org/10.1109/CVPR.2006.100

Harwath, D., Chuang, G., & Glass, J. (2018). Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4969-4973). IEEE. https://doi.org/10.1109/ICASSP.2018.8462396

Harwath, D., Hsu, W. N., & Glass, J. (2019). Learning hierarchical discrete linguistic units from visually-grounded speech. *arXiv preprint arXiv:1911.09602*. https://arxiv.org/abs/1911.09602

Ho, N. H., Yang, H. J., Kim, S. H., & Lee, G. (2020). Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access*, *8*, 61672-61686. https://doi.org/10.1109/ACCESS.2020.2984368

Ilharco, G., Zhang, Y., & Baldridge, J. (2019). Large-scale representation learning from visually grounded untranscribed speech. *arXiv preprint arXiv:1909.08782*. https://arxiv.org/abs/1909.08782v1

Jiang, D., Li, W., Cao, M., Zou, W., & Li, X. (2020). Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning. https://arxiv.org/abs/2010.13991

Kawakami, K., Wang, L., Dyer, C., Blunsom, P., & Oord, A. V. D. (2020). Learning robust and multilingual speech representations. https://arxiv.org/abs/2001.11128

Kheddar, H., Himeur, Y., Al-Maadeed, S., Amira, A., & Bensaali, F. (2023). Deep transfer learning for automatic speech recognition: Towards better generalization. *Knowledge-Based Systems*, *277*, 110851. https://doi.org/10.1016/j.knosys.2023.110851

Kinoshita, K., Ochiai, T., Delcroix, M., & Nakatani, T. (2020). Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7009-7013). IEEE. https://doi.org/10.1109/ICASSP40776.2020.9053266

Koguchi, Y., Oharada, K., Takagi, Y., Sawada, Y., Shizuki, B., & Takahashi, S. (2018). A mobile command input through vowel lip shape recognition. In *Human-Computer Interaction. Interaction Technologies: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part III 20* (pp. 297-305). Springer International Publishing. https://doi.org/10.1007/978-3-319-91250-9_2

Kreplak López, M. (2020). *Multimodal speech emotion recognition* (Master's thesis, Universitat Politècnica de Catalunya).

Li, Q., Wang, B., & Melucci, M. (2019). CNM: An interpretable complex-valued network for matching. *arXiv preprint arXiv:1904.05298*.
https://arxiv.org/abs/1904.05298

Li, X., & Roth, D. (2002). Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Lin, J., Gan, C., & Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7083-7093).

Liu, C., Mao, Z., Liu, A. A., Zhang, T., Wang, B., & Zhang, Y. (2019). Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 3-11).
https://doi.org/10.1145/3343031.3350869

Liu, H., Jin, S., & Zhang, C. (2018). Connectionist temporal classification with maximum entropy regularization. *Advances in Neural Information Processing Systems*, *31*.

Liu, Y., Guo, Y., Bakker, E. M., & Lew, M. S. (2017). Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE international conference on computer vision* (pp. 4107-4116).

MacDonald, J. (2018). Hearing lips and seeing voices: the origins and development of the 'McGurk Effect'and reflections on audio–visual speech perception over the last 40 years. *Multisensory Research*, *31*(1-2), 7-18.
https://doi.org/10.1163/22134808-00002548

Mercado III, E., Mantell, J. T., & Pfordresher, P. Q. (2014). Imitating sounds: A cognitive approach to understanding vocal imitation. *Comparative Cognition & Behavior Reviews*, *9*.

McCarty, T. L., & Coronel-Molina, S. M. (Eds.). (2016). *Indigenous language revitalization in the Americas*. New York: Routledge.

Miech, A., Zhukov, D., Alayrac, J. B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2630-2640).

Mohamed, A., Lee, H. Y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., ... & Watanabe, S. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1179-1210.
https://doi.org/10.1109/JSTSP.2022.3207050

Monfort, M., Jin, S., Liu, A., Harwath, D., Feris, R., Glass, J., & Oliva, A. (2021). Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14871-14881).

Narayanan, S., Bresch, E., Ghosh, P. K., Goldstein, L., Katsamanis, A., Kim, Y., ... & Zhu, Y. (2011). A multimodal real-time MRI articulatory corpus for speech research. In *Twelfth Annual Conference of the International Speech Communication Association*.

Nguyen, T. S., Stüker, S., & Waibel, A. (2020). Super-human performance in online low-latency recognition of conversational speech. *arXiv preprint arXiv:2010.03449*.
https://doi.org/10.48550/arXiv.2010.03449

neaţă, D., & Cucu, H. (2022). Improving multimodal speech recognition by data augmentation and speech representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4579-4588).

Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
https://doi.org/10.48550/arXiv.1807.03748

Parcollet, T., Ravanelli, M., Morchid, M., Linarès, G., Trabelsi, C., De Mori, R., & Bengio, Y. (2018a). Quaternion recurrent neural networks. *arXiv preprint arXiv:1806.04418*.
https://doi.org/10.48550/arXiv.1806.04418

Parcollet, T., Ravanelli, M., Morchid, M., Linarès, G., & De Mori, R. (2018b). Speech recognition with quaternion neural networks. *arXiv preprint arXiv:1811.09678*.
https://doi.org/10.48550/arXiv.1811.09678

Pasad, A., Shi, B., Kamper, H., & Livescu, K. (2019). On the contributions of visual and textual supervision in low-resource semantic speech retrieval. *arXiv preprint arXiv:1904.10947*.
https://doi.org/10.48550/arXiv.1904.10947

Peng, P., & Harwath, D. (2022a). Self-supervised representation learning for speech using visual grounding and masked language modeling. *arXiv preprint arXiv:2202.03543*.
https://doi.org/10.48550/arXiv.2202.03543

Peng, P., & Harwath, D. (2022b). Fast-slow transformer for visually grounding speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7727-7731). IEEE.
https://doi.org/10.1109/ICASSP43922.2022.9747103

Polyak, A., Adi, Y., Copet, J., Kharitonov, E., Lakhotia, K., Hsu, W. N., ... & Dupoux, E. (2021). Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*.
https://doi.org/10.48550/arXiv.2104.00355

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, *91*(9), 1306-1326.
https://doi.org/10.1109/JPROC.2003.817150

Quan, Y., Chen, Y., Shao, Y., Teng, H., Xu, Y., & Ji, H. (2021). Image denoising using complex-valued deep CNN. *Pattern Recognition*, *111*, 107639.
https://doi.org/10.1016/j.patcog.2020.107639

Razavi, A., Van den Oord, A., & Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, *32*.

Reichert, D. P., & Serre, T. (2013). Neuronal synchrony in complex-valued deep networks. *arXiv preprint arXiv:1312.6115*.
https://doi.org/10.48550/arXiv.1312.6115

Riviere, M., Joulin, A., Mazaré, P. E., & Dupoux, E. (2020). Unsupervised pretraining transfers well across languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7414-7418). IEEE.
https://10.1109/ICASSP40776.2020.9054548

Sadhu, S., He, D., Huang, C. W., Mallidi, S. H., Wu, M., Rastrow, A., ... & Maas, R. (2021). Wav2vec-c: A self-supervised model for speech representation learning. *arXiv preprint arXiv:2103.08393*.
https://doi.org/10.48550/arXiv.2103.08393

Sanabria, R., Waters, A., & Baldridge, J. (2021). Talk, don't write: A study of direct speech-based image retrieval. *arXiv preprint arXiv:2104.01894*.
https://doi.org/10.48550/arXiv.2104.01894

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.
https://doi.org/10.48550/arXiv.1904.05862

Shi, B., Hsu, W. N., Lakhotia, K., & Mohamed, A. (2022a). Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*.
https://doi.org/10.48550/arXiv.2201.02184

Shi, B., Hsu, W. N., & Mohamed, A. (2022b). Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*.
https://doi.org/10.48550/arXiv.2201.01763

Song, Y., & Soleymani, M. (2019). Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1979-1988).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

Teixeira, F. L., Soares, S. P., Abreu, J. L., Oliveira, P. M., & Teixeira, J. P. (2024). Comparative Analysis of Windows for Speech Emotion Recognition Using CNN. In *International Conference on Optimization, Learning Algorithms and Applications* (pp. 233-248). Springer, Cham.
https://doi.org/10.1007/978-3-031-53025-8_17

Trabelsi, C., Bilaniuk, O., Zhang, Y., Serdyuk, D., Subramanian, S., Santos, J. F., ... & Pal, C. J. (2017). Deep complex networks. *arXiv preprint arXiv:1705.09792*.
https://doi.org/10.48550/arXiv.1705.09792

Tu, Y., Lin, Y., Hou, C., & Mao, S. (2020). Complex-valued networks for automatic modulation classification. *IEEE Transactions on Vehicular Technology*, *69*(9), 10085-10089.
https://doi.org/10.1109/TVT.2020.3005707

Tüske, Z., Saon, G., Audhkhasi, K., & Kingsbury, B. (2020). Single headed attention based sequence-to-sequence model for state-of-the-art results on switchboard. *arXiv preprint arXiv:2001.07263*.
https://doi.org/10.48550/arXiv.2001.07263

Wang, L., & Hasegawa-Johnson, M. (2020). A DNN-HMM-DNN hybrid model for discovering word-like units from spoken captions and image regions. In *Interspeech*.

Wang, N., Wang, Z., Xu, X., Shen, F., Yang, Y., & Shen, H. T. (2021). Attention-based relation reasoning network for video-text retrieval. In *2021 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.
https://doi.org/10.1109/ICME51207.2021.9428215

Wang, Y., Kou, K. I., Zou, C., & Tang, Y. Y. (2021). Robust sparse representation in quaternion space. *IEEE Transactions on Image Processing*, *30*, 3637-3649. https://doi.org/10.1109/TIP.2021.3064193

Wang, W., Arora, R., Livescu, K., & Bilmes, J. A. (2015). Unsupervised learning of acoustic features via deep canonical correlation analysis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4590-4594). IEEE. https://doi.org/10.1109/ICASSP.2015.7178840

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., ... & Qiao, Y. (2023). Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14408-14419).

Wei, J., Yang, Y., Xu, X., Zhu, X., & Shen, H. T. (2021). Universal weighting metric learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 6534-6545. https://doi.org/10.1109/TPAMI.2021.3088863

Wu, J., Wu, C., Lu, J., Wang, L., & Cui, X. (2021). Region reinforcement network with topic constraint for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(1), 388-397. https://doi.org/10.1109/TCSVT.2021.3060713

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... & Zweig, G. (2016). Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*. https://doi.org/10.48550/arXiv.1610.05256

Xu, B., Lu, C., Guo, Y., & Wang, J. (2020). Discriminative multi-modality speech recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition* (pp. 14433-14442).

Yang, S., Li, Q., Li, W., Li, X., & Liu, A. A. (2022). Dual-level representation enhancement on characteristic and context for image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(11), 8037-8050. https://doi.org/10.1109/TCSVT.2022.3182426

Zhang, A., Tay, Y., Zhang, S., Chan, A., Luu, A. T., Hui, S. C., & Fu, J. (2021). Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with $1/n $ parameters. *arXiv preprint arXiv:2102.08597*. https://doi.org/10.48550/arXiv.2102.08597

Zhang, T., He, L., Li, X., & Feng, G. (2021). Efficient end-to-end sentence-level lipreading with temporal convolutional networks. *Applied Sciences*, *11*(15), 6975. https://doi.org/10.3390/app11156975