FCGCL: Fine- and Coarse-Granularity Contrastive Learning for Speech Translation

Hao Zhang, Nianwen Si, Yaqi Chen, Zhen Li, Tong Niu, Xukui Yang*, and Dan Qu*
University of Information Engineering, Zhengzhou, China
(haozhang012, snw1608, chyaqi163, gzyangxk)@163.com
(li zhenjojo, jerry newton, qudanqudan)@sina.com

Abstract

It is notoriously difficult to implement end-to-end speech translation (E2E-ST) model because of the task complexity and data scarcity. Existing techniques often attempt to carry out ***implicit knowledge transfer*** from machine translation (MT) to ST model by imposing various constraints. However, in this transfer scenario, a significant problem is that the performance of the MT will drop significantly and the final transfer effect is also restricted. In this article, we recommend Fine and Coarse Granularity Contrastive Learning (FCGCL), which conduct ***explicit knowledge transfer*** from MT to ST model. Specially, we ensure through multi granularity contrastive learning that inputs with similar semantic between different modalities are encoded closely in the shared semantic space while inputs with different semantics are kept apart. Experiments on the MuST-C datasets on all 8 languages and further analysis show that our method can effectively improve the E2E-ST performance and achieves an average BLEU of 29.0 [1].

1 Introduction

End-to-end (E2E) speech-to-text translation (ST) (Bérard et al., 2018; Sperber et al., 2019; Liu et al., 2019; Dong et al., 2021; Liu et al., 2020b; Du et al., 2021; Fang et al., 2022; Ye et al., 2021; Xu et al., 2021; Han et al., 2021) is the task of translating a source-language audio directly to a foreign language text without any intermediate outputs. Different from the traditional cascade method (Ney, 1999; Mathias and Byrne, 2006; Sperber et al., 2017; Lam et al., 2021; Bahar et al., 2021; Dalmia et al., 2021) which decomposes ST into two sub-tasks -automatic speech recognition (ASR) for transcription and machine translation (MT) for translation, E2E-ST jointly handles them in a single neural network. This endows E2E-ST with special advantages, such as lower latency, less error propagation, and fewer parameters.

However, E2E-ST is a cross-modal translation task, and it is non-trivial to train such a model well. Speech is more complicated and finer granularity than text, making it more difficult to extract representations containing rich semantic information. Such modal gap between speech and text results in the performance of ST model is usually inferior to the corresponding MT model (Liu et al., 2020b). To overcome this problem, existing techniques often attempt to carry out ***implicit knowledge transfer*** by imposing various constraints (Liu et al., 2020b; Du et al., 2021; Fang et al., 2022; Han et al., 2021). In this transfer scenario, a significant problem is that the performance of the MT as a constraint will drop significantly and the final transfer effect is also restricted.

In this paper, we propose a cross-modal multi granularity contrastive learning method to make ***explicit knowledge transfer*** from MT to ST model. The embedding from a ST model encoder can be regarded as frame-level representation of fine granularity. Correspondingly, the mean vector of embeddings is termed as the coarse granularity sentence-level representation. The same as ST model, there are two types of representations with different granularities obtaining from the MT model encoder. We perform fine and coarse granularity contrastive learning (FCGCL) on both sentence-level and frame-level to provide comprehensive guidance for the extraction of speech representations to bridge the modal gap. The MT model is pretrained and we frozen its parameters to avoid the performance drop.

Moreover, another problem needs to be solved is the representation degeneration (Gao et al., 2019; Wang et al., 2019; Ethayarajh, 2019) suffered by the pretrained MT model, which is also called anisotropy. In this paper we take a simple whiten-

* Corresponding author.
[1] https://github.com/zhhao/fcgcl

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3048–3059

ing approach (Su et al., 2021; Huang et al., 2021) to alleviate representation degeneration problem by transforming the representation into a standard normal distribution, which satisfies isotropy. We conduct experiments on the MuST-C benchmark on all 8 language pairs. The experiment results and detailed analysis verify the effectiveness of our proposed method.

2 Method

In this section we first analysis the basic problem formulation of E2E-ST and introduce the overall framework of FCGCL. Then the coarse granularity contrastive learning and the whitening operation to alleviate the representation degeneration problem are stated in Section 2.3. Section 2.4 details the fine granularity contrastive learning and the maximum similarity method which is used to find the corresponding text token for each speech frame in an unsupervised manner. Finally, we describe knowledge distillation in Section 2.5.

2.1 Problem Formulation

The speech translation corpus usually contains *speech-transcription-translation* triples, denoted as $\mathcal{D} = \{(x^{(n)}, y^{(n)}, z^{(n)})\}_{n=1}^{N}$, where x represents audio, y is the translation in target language and z is the corresponding transcription in the source language. E2E-ST strives to generate translated sequences y without generating intermediate transcription z, and the standard training objective is to optimize the maximum likelihood estimation loss on the training set:

$$\mathcal{L}_{ST}(\theta) = -\sum_{(x,y)\in\mathcal{D}} \log p(y|x) \tag{1}$$

2.2 Model Architecture

The overall structure of FCGCL is shown in Figure 1, including an E2E-ST model, a pretrained MT model, and a contrastive learning module. The encoder and decoder in ST and MT are consistent with the original Transformer (Vaswani et al., 2017). Following the previous works (Fang et al., 2022; Ye et al., 2021; Han et al., 2021), we use a pretrained Wav2vec2.0 (Baevski et al., 2020) without finetune to extract speech representations.

2.3 Coarse Granularity Contrastive Learning

Given N speech-transcription pairs $\{(x^{(n)}, z^{(n)})\}_{n=1}^{N}$, the encoded representations are denoted as $\{(h_x^{(n)}, h_z^{(n)})\}_{n=1}^{N}$, where $h_x^{(n)} \in T_x \times d$ and $h_z^{(n)} \in T_z \times d$ are the speech and text representations of the n-th sample from ST encoder and MT encoder, respectively. T_x and T_z are the lengths of the speech frame and text token sequences, respectively. We average the encoded representation over the time dimension to get the coarse granularity representations. The text representation is further whitened to relieve the representation degeneration problem.

Whitening Affected by word frequency, the word representations space finally learned by the MT model encoder is squeezed into a cone, which is anisotropy. The sentence embedding - as average of context embeddings from last encoder layer - suffer from the same issues, thus the sentence embedding space is semantically non-smoothing and poorly defined in some areas (Li et al., 2020a). This will lead to the phenomenon that some samples are not similar to the anchor sample, but the similarity calculated by metrics such cosine similarity is relatively large. We address the problem by a simple linear transformation called whitening. The whitening method (Su et al., 2021; Huang et al., 2021) will transform the sentence embeddings into the standard normal distribution, which satisfies isotropy. The isotropic sentence embedding space ensures that the cosine similarity can correctly measure sample similarity.

Contrastive loss In order to ensure the consistency of negative sample representation, only text samples are regarded as negative samples and other speech samples in batch are not considered as negative samples. We denote the representation to compute the contrastive loss as $\{(s_x^{(n)}, s_z^{(n)})\}_{n=1}^{N}$, where $s_x^{(i)} = AveragePooling(h_x^{(i)})$, $s_z^{(i)} = Whitening(AveragePooling(h_z^{(i)}))$, $s_x^{(i)} \in 1 \times d$, $s_z^{(i)} \in 1 \times d$. In order to decouple the relationship between the number of negative samples and the batch size, we set up a we set up a First-In-First-Out (FIFO) queue (He et al., 2020) to store negative samples in previous mini-batch. The contrastive loss for coarse granularity is as follows:

$$\mathcal{L}_{coarse} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{sim(s_x^i \cdot s_z^i/\tau)}}{\sum_{j=1}^{N} e^{sim(s_x^i \cdot s_z^j/\tau)} + \sum_{k=1}^{K} e^{sim(s_x^i \cdot s_z^k/\tau)}} \tag{2}$$

where $sim(\cdot)$ is the cosine similarity function, τ is the temperature and K is the number of negative samples in queue.

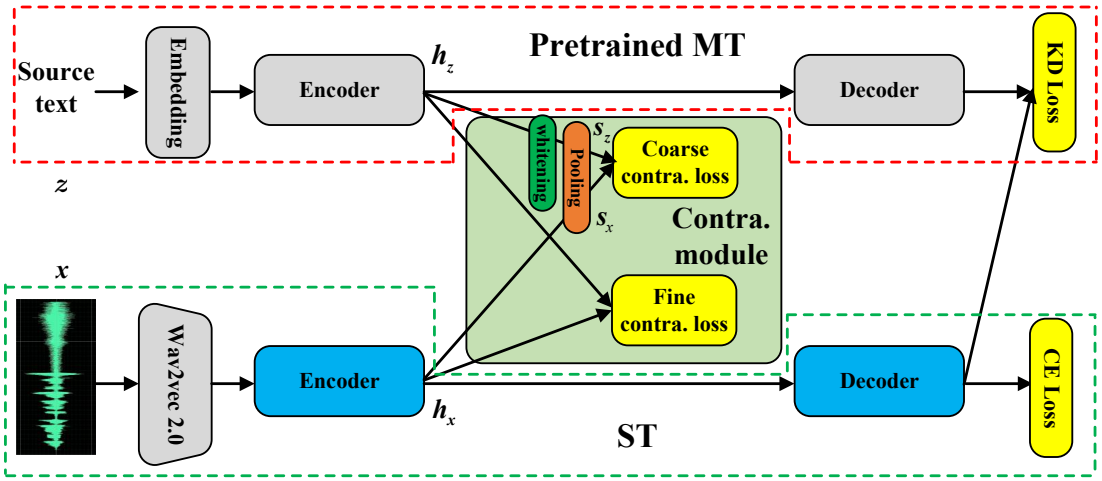

Figure 1: Overview of our proposed method. The grey modules in the figure indicate that the parameters are no longer updated during training..

2.4 Fine Granularity Contrastive Learning

The above sentence-level contrastive loss is sufficient for classification tasks. However, for the sequence-to-sequence generative tasks like ST, the semantic of each small unit must be accurate which means the representation of each frame learned from sentence level contrastive learning may be sub-optimal. Thus, we further recommend fine granularity contrastive learning to find the optimal token-level representation for the decoder.

The *j*-th frame representation of the *i*-th speech sample is denoted as $h^{(i)}_{x,j}$. For now, let us assume that we can easily find the positive sample $pos^{(i)}_{x,j}$ for the anchor sample $h^{(i)}_{x,j}$. The setting of negative samples is the same as that in Section 2.3. Then the fine granularity contrastive loss is defined as:

$$\mathcal{L}_{fine} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{T_x}\log\frac{e^{sim(h^{(i)}_{x,j}\cdot pos^{(i)}_{x,j}/\tau)}}{e^{sim(h^{(i)}_{x,j}\cdot pos^{(i)}_{x,j}/\tau)}+\sum_{k=1,k\neq i}^{K} e^{sim(h^{(i)}_{x,j}\cdot s^{k}_{z}/\tau)}} \tag{3}$$

where K is the number of negative samples, T_x is the length of frame-level speech and N is the number of speech.

Obtaining positive samples Given speech representation $h^{(i)}_x$ and the corresponding text representation $h^{(i)}_z$, we calculate the cosine similarity matrix $\Delta \in {}^{T_x \times T_z}$, in which each element means the cosine similarity between corresponding speech frame and text token. The matching rule is that each feature vector in speech is matched to the most similar feature vector in transcription. The positive sample selection process can be formulated as:

$$pos^{(i)}_{x,j} = \{h^{(i)}_{z,m} | \arg\max_{m} sim(h^{(i)}_{x,j}\cdot h^{(i)}_{z,m}), m = 1, 2, \cdots T_z\} \tag{4}$$

where $h^{(i)}_{x,j}$ is the *j-th* frame representation of the *i-th* speech, $h^{(i)}_{z,m}$ is the *m-th* token representation of the corresponding transcription. The entire matching process can be quickly calculated through matrix with a small amount of calculation. Because of the unsupervised paradigm, no additional alignment model is required. In addition, whitening is not used when calculating the similarity matrix. This is because the transform matrix cannot be estimated for some input by SVD in some case where the length of the token sequence is limited, which will cause computational instability.

2.5 Knowledge Distillation

In addition, in order to provide further supervision signals and give multi-level guidance for the training of the ST model, word-level knowledge distillation (KD) (Liu et al., 2019) is used to further transfer the knowledge from the MT to the ST model. The KD loss is defined as:

$$\mathcal{L}_{KD} = -\sum_{(x,z)\in\mathcal{D}}\sum_{t=1}^{N}\sum_{k=1}^{|V|} Q(y_t = k|y_{<t}, x; \theta_{ST}) \times \log P(y_t = k|y_{<t}, x; \theta_{MT}) \tag{5}$$

where $Q(y_t = k|y_{<t}, x; \theta_{ST})$ and $P(y_t = k|y_{<t}, x; \theta_{MT})$ are the decoder output distributions

of the teacher and the student model, respectively. x, y, z are speech, translation and transcription, respectively. V is the shared vocabulary between ST and MT.

2.6 Training

Pretrain A pretrained MT model is used to initialize the whole ST model. During training, the parameters of MT are frozen to avoid performance drop. Our preliminary experiment shows that joint training does not improve the performance of the MT model, but has a negative impact. In addition, the parameters of Wav2vec 2.0 are also frozen to facilitate quick experiments.

The overall training objective is the weighted sum of all previous losses:

$$\mathcal{L} = \alpha\mathcal{L}_{ST} + \beta\mathcal{L}_{coarse} + \gamma\mathcal{L}_{fine} + \eta\mathcal{L}_{KD} \quad (6)$$

where $\alpha, \beta, \gamma, \eta$ are hyper-parameters to adjust the weight of each loss. During inference, only ST model is preserved, and all other modules are discarded.

3 Experiment

3.1 Dataset and Processing

MuST-C MuST-C (Di Gangi et al., 2019a) is a multilingual dataset based on English TED talks, including English speech, English transcription and the translation in 8 language direction: German (De), French (Fr), Russian (Ru), Spanish (Es), Italian (It), Romanian (Ro), Portuguese (Pt), and Dutch (NL). It is one of the largest training data for speech translation. We select the model according to its performance on the validation set and use tst-COMMON set to test.

External MT Datasets The MT model is trained separately and has the same structure as the ST model, which allows us to use parallel sentence pairs in the external MT datasets in addition to the transcription-translation pairs in the ST corpus.

Table 1 lists the statistics of all the datasets included.

Processing For speech input, we use the original 16-bit 16kHz mono-channel audio waveform. We tokenize and truecase all texts via Moses[2]. Punctuation is kept, but split from words, and then normalized. In each language direction, we apply BPE (Sennrich et al., 2015) on the combination of source and target text to obtain shared sub-word units, and the vocabulary size is set to 8K.

[2]https://www.statmt.org/moses/

Language (EN-)	MuST-C		External MT	
	hours	#sent	Source	#sent
Germany (DE)	408h	234K	WMT	4.6M
French (FR)	492h	280K	WMT	40.8M
Russian (RU)	489h	270K	WMT	2.5M
Spanish (ES)	504h	270K	WMT	15.2M
Italian (IT)	465h	258K	OPUS100	1.0M
Romanian (RO)	432h	240K	WMT	0.6M
Portuguese (PT)	385h	244K	OPUS100	1.0M
Dutch (NL)	442h	253K	OPUS100	1.0M

Table 1: Statistics of all datasets

3.2 Experimental setups

Model Configuration The Wav2vec 2.0 follows the large configuration in (Baevski et al., 2020), which is self-supervised pretrained on Librispeech (Panayotov et al., 2015) audio data only[3]. We use Transformer (Vaswani et al., 2017) as the backbone, including 6 encoder layers and 6 decoder layers. Each of these layers comprises of 256 hidden units, 4 attention heads, and 2048 feed-forward hidden units.

MT model Pretrain When using additional datasets to train the MT model, we first train MT model on additional MT corpora and then finetune it on the transcription-translation pairs in the MuST-C corpus to solve the domain mismatch problem. This is same with (Xu et al., 2021).

E2E-ST Training and Inference During training, we use the Adam (Kingma and Ba, 2014) optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$ and adopt the default learning schedule in ESPnet (Inaguma et al., 2020). The dropout rate and the value of label smoothing are all set to 0.1. We adopt dropdim (Zhang et al., 2022), a recently proposed structured dropout method, as a data augmentation strategy. We adopt the random mask strategy described in their paper with a mask rate of 0.05. An early stop strategy is adopted during training, that is, training is stopped if the accuracy of the model on the validation set does not increase for three consecutive epochs. The training takes about one day to converge. We set α,β,γ and η to 0.4, 1.0, 1.0, 0.6 respectively.

During inference, we average the best 5 checkpoints for evaluation. We use beam search with a beam size of 10, and the length penalty is 0.6. We report the case-sensitive SacreBLEU[4] (Post, 2018) for fair comparison with previous work.

[3]https://huggingface.co/facebook/wav2vec2-large-960h

[4]https://github.com/mjpost/sacrebleu, signature: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.0.0

Model	External Data		BLEU								
	Speech	MT	EN-DE	EN-FR	EN-RU	EN-ES	EN-IT	EN-RO	EN-PT	EN-NL	Avg.
w/o external MT data											
STAST (Liu et al., 2020b)	×	×	23.1	-	-	-	-	-	-	-	-
AFS (Zhang et al., 2020)	×	×	22.4	31.6	14.7	26.9	23.0	21.0	30.0	24.9	23.9
SATE (Xu et al., 2021)	×	×	25.2	-	-	-	-	-	-	-	-
Dual-Decoder (Le et al., 2020)	×	×	23.6	33.5	15.2	28.1	24.2	22.9	30.0	27.6	25.7
XSTNet (Ye et al., 2021)	✓	×	25.5	36.0	16.9	29.6	25.5	**25.1**	31.3	30.0	27.5
TDA (Du et al., 2021)	×	×	25.4	36.1	16.4	29.6	25.1	23.9	31.1	29.6	27.2
STEMM (Fang et al., 2022)	✓	×	25.6	36.1	17.1	30.3	25.6	24.3	31.0	30.1	27.5
FCGCL	✓	×	**25.7**	**36.5**	**17.5**	**30.4**	**26.0**	24.6	**31.4**	**30.3**	**27.8**
w/ external MT data											
SATE (Xu et al., 2021)	✓	✓	28.1	-	-	-	-	-	-	-	-
JT-S-MT (Tang et al., 2021)	×	✓	26.8	37.4	-	31.0	-	-	-	-	
XSTNet (Ye et al., 2021)	✓	✓	27.8	**38.0**	18.5	30.8	26.4	25.7	**32.4**	**31.2**	28.8
Chimera (Han et al., 2021)	✓	✓	26.3	35.6	17.4	30.6	25.0	24.0	30.2	29.2	27.3
TDA (Du et al., 2021)	✓	✓	27.1	37.4	-	-	-	-	-	-	-
STEMM (Fang et al., 2022)	✓	✓	28.7	37.4	17.8	31.0	25.8	24.5	31.7	30.5	28.4
FCGCL	✓	✓	**28.7**	37.5	**19.1**	**31.2**	**26.5**	**26.0**	32.1	31.0	**29.0**

Table 2: BLEU scores on MuST-C tst-COMMON set. "External Data" indicates whether the method uses additional data.

3.3 Main Results

Comparison with E2E Baselines Table 2 shows the comparison of our proposed method on the MuST-C dataset and the reference E2E-ST systems. We mainly compare with previous works with or without additional MT datasets. (a) Without external MT datasets. Compared with the previous best model, our proposed method gains an average improvement of 0.3 BLEU in 8 language directions. Different from the previous works whose main ideas is to implicitly bound the parameter space of ST model by treating the MT as a constraint term with the sharing mechanism, we employ contrastive learning and knowledge distillation to provide direct guidance for ***explicit knowledge transfer*** across modality and thus achieve better results. (b) With external MT datasets. Our method can further achieve a 1.2 BLEU improvement compared FCGCL without external MT data and outperform the previous state-of-the-art (SOTA) model by 0.2 BLEU. Contrastive learning is also used in Chimera (Han et al., 2021), which designed a shared semantic memory to learn the semantic information shared between modalities, but it limits the feature output lengths of the two modalities to be consistent. Our method does not have this limitation and is more efficient.

Comparison with Cascaded Baselines To further validate the effectiveness of our proposed method, we compare with several strong cascaded baseline systems, all of which are trained with additional datasets. As described in Table 3, our proposed method can outperform the cascade model and achieve better performance, showing the potential of FCGCL.

Model		BLEU	
		En-De	En-Er
Cascaded	XSTNet (Ye et al., 2021)	25.2	34.9
	STEMM (Fang et al., 2022)	27.5	-
	STAE (Liu et al., 2020b)	28.2	-
End-to-end	FCGCL	**28.7**	**37.5**

Table 3: Comparison with cascaded models on MuST-C En-De and En-Fr tst-COMMON set.

4 Analysis

4.1 Ablation Studies

Contributions of Different Parts To better evaluate the contribution of each part of our proposed method, we perform ablation studies on the MuST-C En-De datasets. The results in Table 4 show that each part of FCGCL is necessary and has a positive effect in improving model performance. It is worth noting that when dropdim is removed, the model performance has a significant drop, about 0.95 BLEU. This is mainly because dropdim in FCGCL is not only used to enhance the generalization ability of the model, but also to generate harder representations to enhance the effect of contrastive learning as a data augmentation method.

Size of Negative Sample Queue In negative example-based contrastive learning, the size of the queue directly affects the performance of the model. Figure 2 shows the experiment results. The blue and yellow curves in the figure represent the variation of model performance with the size of the

Model	BLEU
FCGCL	25.71
-coarse	25.03
-fine	25.26
-kd	25.04
-dropdim	24.76

Table 4: BLEU scores on MuST-C En-De tst-COMMON set when different parts are removed.

τ	0.06	0.08	0.10	0.12	0.14	0.16
BLEU	25.29	25.21	25.46	25.71	25.43	25.03

Table 5: BLEU scores on MuST-C En-De tst-COMMON under different temperature.

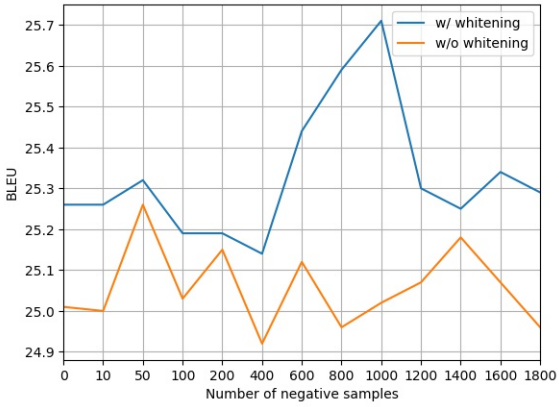

Figure 2: Model performance with different negative queue size.

negative sample queue with or without whitening, respectively.

Without whitening, the optimal queue size is 50. When whitening is used, the model performance increases with the queue size, reaching best performance at 1000. This is mainly because the whitening operation can reduce the number of false negative samples. These false negative samples are not similar to the anchor sample, but the calculated similarity is large due to the anisotropy phenomenon of the uneven distribution, which affects the calculation of the contrast loss and leads to the degradation of the model performance. The whitening operation can alleviate this anisotropy phenomenon, so the model equipped with whitening can benefit from more negative samples. However, when the queue size exceeds 1000, the model performance drops significantly. We speculate that although the whitening operation can reduce the number of false negative samples, the overall false negative samples are still very large in this case. In Section 4.2 we conduct further experiments on the effect of whitening to validate our analysis here.

Effect of Temperature Temperature τ is another important hyperparameter in contrastive learning, which controls the strength of penalties on hard negative samples. The smaller the temperature, the more the model pays attention to the hard negative samples and gives them a larger gradient to separate from the anchor sample. Table 5 shows the model performance at different temperature. We choose several temperature hyper-parameters ranging from 0.06 to 0.16. The model achieves the best performance at the temperature of 0.12.

4.2 Effect of Whitening

To verify the effect of whitening, we compute the samples pairs similarity distribution. Specifically, we randomly select a speech sample from the MuST-C En-De tst-COMMON set as the anchor sample, and 1000 transcription samples. For each sample, we average the encoder output over the time dimension to get the overall representation. Suppose the similarity distribution of 1000 samples and anchor sample (speech sample) is $[s_x^1 \cdot s_z^1, s_x^1 \cdot s_z^2, \cdots, s_x^1 \cdot s_z^j], j = 1, \cdots 1000$, where s_x^1 denotes the speech representation, s_z^j is j-th text representation. s_z^1 is the corresponding positive sample representation, and the rest are treated as negative samples. Then we normalize the similarity distribution $[s_x^1 \cdot s_z^1, s_x^1 \cdot s_z^2, \cdots, s_x^1 \cdot s_z^j] / s_x^1 \cdot s_z^1, j = 1, \cdots 1000$.

As shown in Figure 3, blue and yellow represent the normalized similarity distribution histogram with or without whitening, respectively. Before whitening, the number of samples with the normalized similarity greater than 0.2 is about 104. After whitening, the number is reduced to 34. This means that whitening operation can partly solve the representation degeneration of text, alleviating the problem that some negative samples are not similar to the anchor sample, but the calculated cosine similarity is relatively large.

5 Visualization

5.1 Visualization

Visualization of Coarse Granularity Alignment We randomly select 30 speech-transcription pairs from MuST-C En-De tst-COMMON set, and then apply T-SNE (Van der Maaten and Hinton, 2008)

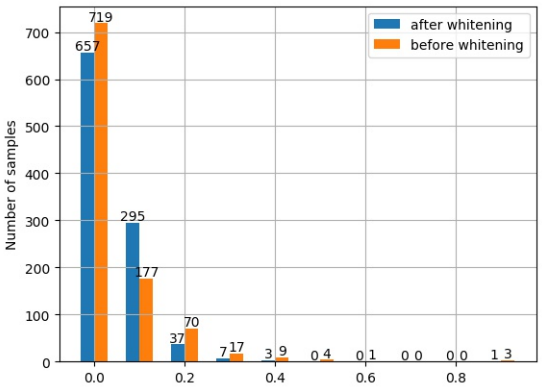

Figure 3: Histogram of normalized sample similarity distribution. The horizontal coordinate represents the similarity normalized by the maximum value. Here, the maximum value is the similarity between the anchor sample and the positive sample. The vertical coordinate represents the number of samples located in the corresponding interval.

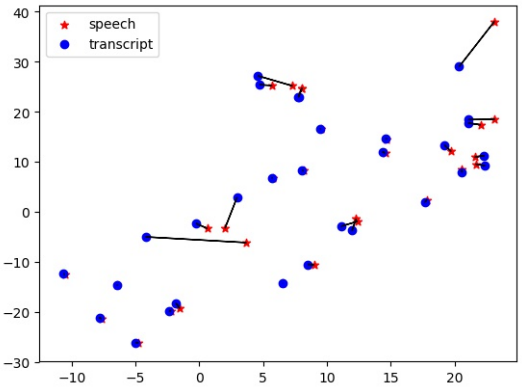

Figure 4: Visualization sentence-level representation.

to the vector representations of these samples to reduce the dimension to two. Note that these vector representations are obtained by averaging the encoder outputs over the time dimension.

The results are visualized in Figure 4. Each speech-transcription pair is connected by a solid line. It can be intuitively seen from the figure that most paired speech-transcription are projected together, and some even overlap with each other. This proves that FCGCL is capable of bridging the representation divergence of the two modalities. In addition, some speech representations in the figure are still clustered together, mainly because we compute the contrastive loss across modalities and not within the modal. Thus, the speech representations do not show good uniformity. We'll make further research about this.

Visualization of Fine Granularity Alignment In

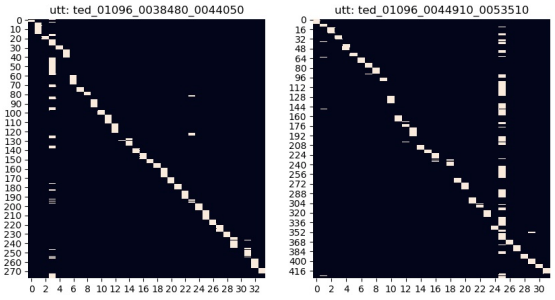

Figure 5: Visualization of similarity matrix after masking.

order to find this correspondence between speech frame and text token, we recommend the maximum similarity method. ***Is this method reasonable?*** We prove its feasibility by performing fine granularity visualization. We randomly select paired speech-transcription from MuST-C En-De tst-COMMON, then get the speech representation $h_x^{(i)} \in T_x \times d$ and text representation $h_z^{(i)} \in T_z \times d$. Then we calculate the cosine similarity matrix $\Delta \in {}^{T_x \times T_z}$, whose maximum value in each row is set to 1. The remaining values in the similarity matrix are masked with 0.

Figure 5 shows the alignment of two speech-transcription pairs. Ideally, the speech frame and its corresponding transcription token should be monotonically aligned. The overall similarity is monotonic. However, there are a small number of speech frames showing non-monotonicity. Possible reason is that there are some silence frames in speech and the model aligns them to a certain token. In addition, the text may contain the same token. This might result in wrong alignment since the model cannot determine which token the speech frame should correspond to. In general, by adopting this method, we can effectively and efficiently find the correspondence between speech frame and text token in an unsupervised manner with negligible latency overhead.

6 Related Works

End-to-end ST Benefiting from the development of deep learning in recent years, more and more researchers have begun to focus on the end-to-end learning paradigm (Bérard et al., 2016; Duong et al., 2016). However, due to data scarcity, it is difficult to train an E2E-ST model well. Researchers have begun to explore various solutions, ranging from efficient network architecture design (Karita et al., 2019; Di Gangi et al., 2019b; Sung et al.,

2019) to incorporating additional training signals, including multi-task learning (Weiss et al., 2017; Liu et al., 2020a), sub-module pre-training (Stoian et al., 2020; Wang et al., 2020b), knowledge distillation (Liu et al., 2019; Gaido et al., 2020), meta-learning (Indurthi et al., 2019), data augmentation (Kocabiyikoglu et al., 2018; Jia et al., 2019; Pino et al., 2019), attention transfer (Sperber et al., 2019; Wang et al., 2020a). Another recognized problem is that the encoder of the ST model is overburdened. To address this problem, some studies decouple the ST encoder into an acoustic encoder and a semantic encoder to better extract information from the input (Dong et al., 2021; Liu et al., 2020b; Xu et al., 2021). Other studies use various methods to impose constraints on the E2E-ST model for knowledge transfer between modalities. However, most of these methods make implicit knowledge transfer, while in this work we use contrastive learning combined with knowledge distillation to make explicit knowledge transfer for E2E-ST to bridge the modal gap and improve the performance.

Contrastive learning Contrastive learning has recently emerged as a powerful method for learning representations from unlabeled data. Models based on contrastive learning have achieved outstanding performance in the domains of computer vision (He et al., 2020; Chen et al., 2020), speech (Wang and Oord, 2021; Zhang et al., 2021; Xiao et al., 2021), and NLP (Gao et al., 2021; Yan et al., 2021; Ye et al., 2022). Some studies extend contrastive learning to the multimodal domain (Radford et al., 2021; Wu et al., 2022; Alayrac et al., 2020). Inspired by these works, we use cross-modal contrastive learning to ensure that inputs with similar semantic between different modalities are encoded closely in the shared feature space while inputs with different semantics is kept apart.

Fine Granularity Contrastive learning Contrastive learning usually obtains the overall representation of the input through pooling, and then performs loss calculations. For classification tasks, the overall representation is sufficient. However, the representation obtained this way is sub-optimal for generative tasks or dense prediction tasks. Some studies perform fine granularity contrastive learning to learn finer input representations (Wang et al., 2021; Zeng et al., 2021; Wang and Karout, 2021). To ensure the effectiveness and efficiency of computation, we adapt the method of (Wang et al., 2021) to perform cross-modal matching of speech frames and text tokens.

Representation Degeneration Affected by word frequency, the embedding space learned in language modeling or neural machine translation is squeezed into a narrow cone, showing an uneven distribution of anisotropy which is referred as representation degeneration (Gao et al., 2019; Wang et al., 2019; Ethayarajh, 2019). Many strategies are recommended to mitigate this problem: regularization-based methods (Gao et al., 2019; Wang et al., 2019), flow-based methods (Li et al., 2020b), whitening methods (Su et al., 2021; Huang et al., 2021), and methods that utilize noise as the negative samples (Zhou et al., 2022; Wu et al., 2021). The whitening-based approach we use, which was first applied to the BERT, directly processes the trained representation without retraining the model.

7 Conclusion

In this paper, we recommend FCGCL, which combines contrastive learning and knowledge distillation for *explicit knowledge transfer* across modalities. In addition, we recommend whitening to solve the representation degeneration problem of text representation in MT model. Experiments on the MuST-C dataset on all 8 languages demonstrate the effectiveness of our method.

Limitations

Although our method exhibits the desired effect, it still suffers from certain limitations. First, we adopt cross-modal contrastive learning to guide the training of the ST model, but the problem of representation degeneration in NLP can seriously affect the calculation of contrastive loss. In this paper, we use the whitening operation to deal with this problem, but when the number of negative samples is too large (over 1000), the model performance drops significantly. This shows that the processing capacity of the whitening operation is also limited, and more effective methods need to be explored in the future. Second, we freeze the parameters of the MT model to avoid quality degradation of text representations. When calculating the contrastive loss, in order to ensure the consistency of the representation, the rest of the speech representations in batch are not taken as negative samples, which causes some speech representations to cluster together and do not show good uniformity. In the future, we will consider adding the contrastive loss

within the speech modal to improve the uniformity of the speech representation distribution.

Acknowledgements

We thank anonymous reviewers and program chairs for their valuable and insightful feedback. This work was supported by the National Natural Science Foundation of China under Grants 61673395 and 62171470.

References

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. *Proceedings of Advances in Neural Information Processing Systems*, 33:25–37.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proceedings of Advances in Neural Information Processing Systems*, 33:12449–12460.

Parnia Bahar, Tobias Bieschke, Ralf Schlüter, and Hermann Ney. 2021. Tight integrated end-to-end training for cascaded speech translation. In *Proceedings of IEEE Spoken Language Technology Workshop*, pages 950–957. IEEE.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6224–6228. IEEE.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of International conference on machine learning*, pages 1597–1607. PMLR.

Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, and Shinji Watanabe. 2021. Searchable hidden intermediates for end-to-end models of decomposable sequence tasks. *arXiv preprint arXiv:2105.00573*.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. Must-c: a multilingual speech translation corpus. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2012–2017. Association for Computational Linguistics.

Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2019b. Adapting transformer to end-to-end spoken language translation. In *Proceedings of INTERSPEECH*, pages 1133–1137. International Speech Communication Association (ISCA).

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12749–12759.

Yichao Du, Zhirui Zhang, Weizhi Wang, Boxing Chen, Jun Xie, and Tong Xu. 2021. Regularizing end-to-end speech translation with triangular decomposition agreement. In *arXiv preprint arXiv:2112.10991*.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. Stemm: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 7050–7062.

Marco Gaido, Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2020. End-to-end speech-translation with knowledge distillation: Fbk@ iwslt2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 80–88.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 2214–2225.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. Whiteningbert: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online. Association for Computational Linguistics.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2019. Data efficient direct speech-to-text translation with modality agnostic meta-learning. *arXiv preprint arXiv:1911.04283*.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184. IEEE.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on transformer vs rnn in speech applications. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation. *arXiv preprint arXiv:1802.03142*.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2021. Cascaded models with cyclic feedback for direct speech translation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7508–7512. IEEE.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. *arXiv preprint arXiv:2011.00747*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020a. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020b. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. *arXiv preprint arXiv:1904.08075*.

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020a. Synchronous speech recognition and speech-to-text translation with interactive decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8417–8424.

Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2020b. Bridging the modality gap for speech-to-text translation. *arXiv preprint arXiv:2010.14920*.

Lambert Mathias and William Byrne. 2006. Statistical phrase-based speech translation. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, volume 1, pages I–I. IEEE.

Hermann Ney. 1999. Speech translation: Coupling of recognition and translation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210.

Juan Pino, Liezl Puzon, Jiatao Gu, Xutai Ma, Arya D McCarthy, and Deepak Gopinath. 2019. Harnessing indirect training data for end-to-end automatic speech translation: Tricks of the trade. In *Proceedings of the 16th International Conference on Spoken Language Translation*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of International Conference on Machine Learning*, pages 8748–8763. PMLR.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2017. Neural lattice-to-sequence models for uncertain inputs. *arXiv preprint arXiv:1704.00559*.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019. Attention-passing models for robust and data-efficient end-to-end speech translation. *Transactions of the Association for Computational Linguistics*, 7:313–325.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing asr pretraining for low-resource speech-to-text translation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7909–7913. IEEE.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Tzu-Wei Sung, Jun-You Liu, Hung-yi Lee, and Lin-shan Lee. 2019. Towards end-to-end speech-to-text translation with two-pass decoding. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7175–7179. IEEE.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021. Improving speech translation by understanding and learning from the auxiliary text translation task. *arXiv preprint arXiv:2107.05782*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Proceedings of Advances in neural information processing systems*, 30.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020a. Bridging the gap between pretraining and fine-tuning for end-to-end speech translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9161–9168.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020b. Curriculum pre-training for end-to-end speech translation. *arXiv preprint arXiv:2004.10093*.

Duo Wang and Salah Karout. 2021. Fine-grained multi-modal self-supervised learning. *arXiv preprint arXiv:2112.12182*.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. 2019. Improving neural language generation with spectrum control. In *Proceedings of International Conference on Learning Representations*.

Luyu Wang and Aaron van den Oord. 2021. Multi-format contrastive learning of audio representations. *arXiv preprint arXiv:2103.06508*.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. 2021. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2clip: Learning robust audio representations from clip. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4563–4567. IEEE.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2021. Smoothed contrastive learning for unsupervised sentence embedding. *arXiv preprint arXiv:2109.04321*.

Alex Xiao, Christian Fuegen, and Abdelrahman Mohamed. 2021. Contrastive semi-supervised learning for asr. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3870–3874. IEEE.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2021. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2619–2630.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. *arXiv preprint arXiv:2104.10380*.

Tong Zhang1 Wei Ye, Baosong Yang, Long Zhang, Xingzhang Ren, Dayiheng Liu, Jinan Sun, Shikun Zhang, Haibo Zhang, and Wen Zhao. 2022. Frequency-aware contrastive learning for neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. 2020. Adaptive feature selection for end-to-end speech translation. *arXiv preprint arXiv:2010.08518*.

Hao Zhang, Dan Qu, Keji Shao, and Xukui Yang. 2022. Dropdim: A regularization method for transformer networks. *IEEE Signal Processing Letters*, 29:474–478.

Zi-Qiang Zhang, Yan Song, Ming-Hui Wu, Xin Fang, and Li-Rong Dai. 2021. Xlst: Cross-lingual self-training to learn multilingual representation for low resource speech recognition. *arXiv preprint arXiv:2103.08207*.

Kun Zhou, Beichen Zhang, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130.