Speech Sense Disambiguation:
Tackling Homophone Ambiguity in End-to-End Speech Translation

Tengfei Yu[1] **Xuebo Liu**[1*] **Liang Ding**[2] **Kehai Chen**[1] **Dacheng Tao**[3] **Min Zhang**[1]
[1]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
[2]The University of Sydney [3]Nanyang Technological University
tengfeiyu@stu.hit.edu.cn, {liuxuebo,chenkehai,zhangmin2021}@hit.edu.cn
liangding.liam@gmail.com, dacheng.tao@ntu.edu.sg

Abstract

End-to-end speech translation (ST) presents notable disambiguation challenges as it necessitates simultaneous cross-modal and cross-lingual transformations. While word sense disambiguation is an extensively investigated topic in textual machine translation, the exploration of disambiguation strategies for ST models remains limited. Addressing this gap, this paper introduces the concept of speech sense disambiguation (SSD), specifically emphasizing homophones - words pronounced identically but with different meanings. To facilitate this, we first create a comprehensive homophone dictionary and an annotated dataset rich with homophone information established based on speech-text alignment. Building on this unique dictionary, we introduce AmbigST, an innovative homophone-aware contrastive learning approach that integrates a homophone-aware masking strategy. Our experiments on different MuST-C and CoVoST ST benchmarks demonstrate that AmbigST sets new performance standards. Specifically, it achieves SOTA results on BLEU scores for English to German, Spanish, and French ST tasks, underlining its effectiveness in reducing speech sense ambiguity. Data, code and scripts are freely available at https://github.com/ytf-philp/AmbigST.

1 Introduction

Speech translation (ST) translates acoustic speech from one language to another and sees a steady increase in its application (Duong et al., 2016; Bérard et al., 2016; Anastasopoulos and Chiang, 2018; Ansari et al., 2020; Li et al., 2021b; Bentivogli et al., 2021). Particularly, recent research involves the development of unified end-to-end ST models, leveraging the joint pre-training of speech and text (Wang et al., 2020b; Dong et al., 2021a,b; Inaguma et al., 2021; Tang et al., 2022; Zhang et al., 2022b). This approach effectively mitigates the challenges posed by the limited availability of ST data.

Despite advances in end-to-end ST, these systems struggle with managing concurrent cross-modal and cross-lingual conversions, suffering from ambiguity in both acoustics and semantics. This challenge mirrors word sense disambiguation (WSD) in textual machine translation (Rios Gonzales et al., 2017; Luan et al., 2020; Campolungo et al., 2022). Specifically within ST, a paramount obstacle is the disambiguation of homophones - words possess identical phonetic properties but hold distinct semantic interpretations.

Traditionally, numerous ST approaches were reliant on cascade models, where homophone ambiguities were addressed either at the speech recognition stage (Ghosh et al., 2016; Zheng et al., 2020) or the machine translation stage (Xue et al., 2020; Qin et al., 2021; Liu et al., 2018). This delineated approach is in contrast to the integrated philosophy of end-to-end ST. Although several end-to-end ST studies (Zhang et al., 2021; Bang et al., 2022; Zhang et al., 2023a) try to incorporate context as a mechanism to enhance homophonic clues, such indirect methods restrict optimal disambiguation to the scope of understanding context. Consequently, semantic ambiguities persist as a notable source of errors in modern ST models.

To address this pressing concern, our study focuses on the issue of homophones, framing it as a disambiguation task that we term Speech Sense Disambiguation (SSD). Our objective is to explore the question: *how can we efficiently leverage SSD in ST models?* As an initial step, we develop comprehensive homophone dictionaries for the English (En), French (Fr), German (De), and Spanish (Es) languages and generate six language-pair annotated speech translation datasets through speech-text alignments. To bolster the model's capacity for recognizing and processing ambiguous words, we introduce *AmbigST*, a novel homophone-aware

[*] Corresponding Author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8020–8035
August 11-16, 2024 ©2024 Association for Computational Linguistics

contrastive learning methodology. This approach intertwines a homophone-aware masking strategy with contrastive learning, operating at the levels of individual tokens, entire sentences, and the ST model. Through this, the model robustly discerns homophone details adjacent to ambiguous tokens by consistently directing the extraction of context-sensitive representations from speech, thus effectively addressing semantic ambiguity.

We integrate AmbigST within a robustly pre-trained model (Zhang et al., 2022b) to evaluate its efficacy across various datasets. Our evaluation encompasses En-{De, Es, Fr} ST tasks within the MuST-C dataset, as well as {De, Es, Fr}-En ST tasks within the CoVoST datasets. AmbigST consistently outperforms the strong baseline across all metrics. Specifically, this substantial improvement not only sets a new performance standard, surpassing the current state-of-the-art, but also makes significant strides in addressing the pervasive problem of speech sense ambiguity in ST tasks. The **main contributions** of this paper are:

- We focus on addressing the intricate problem of semantic disambiguation in ST and define a task called SSD, which specifically aims to disambiguate homophones where the underlying surface form is different.
- We construct comprehensive homophone dictionaries for {En, Fr, De, Es}, and generate annotated datasets for six ST language-pairs encompassing a total of 1M instances.
- We propose a novel AmbigST method with the homophone-aware masking strategy and multi-level contrastive learning methodology, for conducting effective SSD in ST models.
- Further analysis shows AmbigST boosts the translation of low-frequency words and shorter sentences, which often encounter more pronounced ambiguity issues.

2 Related Work

End-to-End ST Addressing the inherent challenges of error propagation and high latency in cascaded speech translation systems, Bérard et al. (2016); Duong et al. (2016) underscore the viability of end-to-end ST models, eliminating the need for intermediary transcription. This methodology gains significant attention recently (Vila et al., 2018; Salesky et al., 2018, 2019; Gangi et al., 2019; Di Gangi et al., 2019b; Bahar et al., 2019; Inaguma et al., 2020). However, due to the high cost associated with data collection, ST data is often scarce. A multitude of studies focus on enhancing model performance under the constraint of limited data. For instance, existing works employ multitask learning (Le et al., 2020; Vydana et al., 2021; Ye et al., 2021) and leverage monolingual data (Deng et al., 2023) to share knowledge across different tasks, implement curriculum learning to enhance the robustness of the ST model (Kano et al., 2017; Wang et al., 2020b), and combine self-supervised learning with semi-supervised learning for speech translation (Wang et al., 2021; Bapna et al., 2022). To bridge the modality gap, Han et al. (2021); Huang et al. (2021); Xu et al. (2021); Yu et al. (2023) introduce further encoding of acoustic states, which adapt more aptly to the decoder. Most recently, the unified pretraining of speech and text emerges as a dominant paradigm (Zheng et al., 2021; Xu et al., 2021; Zhang et al., 2022b). In this paper, we introduce a novel approach, AmbigST, explicitly designed to address ambiguity challenges in the unified end-to-end ST model effectively.

Semantic Disambiguation Ambiguous words in translation pose a challenge as the model needs to determine the appropriate meaning in the given context (Rodd et al., 2002). Studies have been dedicated to addressing the intricate issue of WSD within the scope of textual machine translation (Navigli, 2009; Rios Gonzales et al., 2017; Luan et al., 2020). A parallel challenge within ST involves the handling of homophones. The prevalent approaches predominantly employ cascade models, wherein homophone ambiguities are addressed at the stage of speech recognition (Ghosh et al., 2016; Zheng et al., 2020), or machine translation (Li et al., 2018; Liu et al., 2018; Xue et al., 2020; Qin et al., 2021). In end-to-end models, the research centers on extracting contextual information. Strategies such as incorporating an additional context encoder (Bang et al., 2022), utilizing context in the output (Hussein et al., 2023), integrating document information (Zhang et al., 2021), and optimizing the connectionist temporal classification loss (Zhang et al., 2023a) are explored. Unlike these studies, our research emphasizes preserving the inherent model architecture and eliminating the necessity for supplementary modifications.

Contrastive Learning Our methodology is inspired by the recent advancements in contrastive

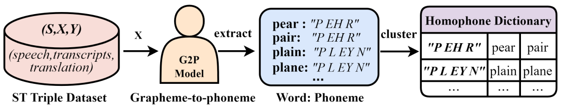

Figure 1: Homophone dictionary construction process.

Speech:	*Ted_23_22.wav:24162314305:109163*
Transcript:	*I fell asleep on the plane at night ...*
Translation:	*Ich bin nachts im Flugzeug eingeschlafen ...*
Homophone word:	*plane (plain), night (knight)*
Homophone index:	*5,7*

Table 1: An instance from the annotated dataset.

representation learning. In the NLP area, contrastive frameworks are employed for tasks such as pretraining (Gao et al., 2021; Zhong et al., 2023), sentence representation learning (Fang et al., 2020; Shen et al., 2020; Wu et al., 2022b; Yan et al., 2021), retrieval (Rao et al., 2023; Lei et al., 2023), and machine translation (Pan et al., 2021; Zhang et al., 2022a; Wu et al., 2022a). More recently, contrastive learning is also applied to cross-modal topics in speech translation (Dong et al., 2019; Zhou et al., 2020; Li et al., 2021a; Ouyang et al., 2022). Unlike prior research focusing on the contrast between speech and text representations, our method centers on the speech representation itself.

3 Speech Sense Disambiguation

Typically, a speech translation corpus includes triples of *speech-transcription-translation*, denoted as $\mathcal{D} = (\mathcal{S}, \mathcal{X}, \mathcal{Y})$. We leverage transcription $\mathcal{X}$ as the basis to explain the notion of speech sense ambiguity. Consequently, we curate a dedicated dataset that embodies homophone information, thereby providing valuable resources to tackle the issue of ambiguity in ST tasks.

Problem Definition Homophones, for example, "I" versus "eye" and "would" versus "wood", are prevalent linguistic phenomena (Zhang et al., 2021; Chung et al., 2022). SSD's process entails discerning the specific meaning of a homophone, which shares its pronunciation with other words but carries a distinct meaning. Given the auditory context $\mathbf{a}$, the goal of SSD is to identify the correct speech sense $\hat{e}$ from a defined set of potential senses $\mathcal{E}$ for a given speech sequence $\mathbf{s}$. This process can be mathematically formulated as:

$$\hat{e} = \arg\max_{e_i \in \mathcal{E}} P(e_i | \mathbf{s}, \mathbf{a}), \tag{1}$$

where e_i denotes the predicted speech sense of the sequence $\mathbf{s}$ given the auditory context $\mathbf{a}$. SSD plays a crucial role due to the varying translations associated with different speech senses. To conduct SSD in ST, we propose to construct homophone dictionaries and then use them to annotate ST datasets.

Annotating the Dataset Figure 1 provides the process of constructing a homophone dictionary. We use transcriptions $\mathcal{X}$ to construct a homophone dictionary by identifying sets of words that possess identical phonemes, facilitated by the utilization of the Montreal Forced Aligner (McAuliffe et al., 2017), an open-source tool engineered for the accurate alignment of speech with its corresponding orthographic transcription. The homophone dictionary consists of sets of words that share the same pronunciation. For instance, the phonemes of *"P EH R"* would include terms such as *pear, pair*. Similarly, the phonemes of *"P L EY N"* would encompass terms like *plain, plane*. The annotated dataset, organized in quintuples, highlights ambiguous words and their positions within sentences. A detailed example is presented in Table 1. We annotated the datasets from the MuST-C dataset (Di Gangi et al., 2019a) covering translations from En-{De, Es, Fr} and CoVoST dataset (Wang et al., 2020a) covering translations from {De, Es, Fr}-En. To ensure easy and automatic implementations, we design our process to be as streamlined as possible. The annotation process can also be easily applied to other datasets. We provide the statistics of our constructed homophone dictionary and annotated data in Appendix A.1.

4 AmbigST

4.1 Model Architecture

Inspired by the latest advancements in end-to-end ST (Zhang et al., 2022b; Fang et al., 2022), our work restructures the foundational model into three distinct components: the speech encoder, the translation encoder, and the translation decoder. The speech encoder first compresses speech representations into hidden states. These hidden states subsequently serve as inputs for the translation encoder, yielding enriched semantic information derived from the condensed speech data. The translation decoder generates the result based on the output of the translation encoder. Furthermore, our model incorporates pre-trained parameters from a unified speech-text pre-training methodology (Zhang et al.,

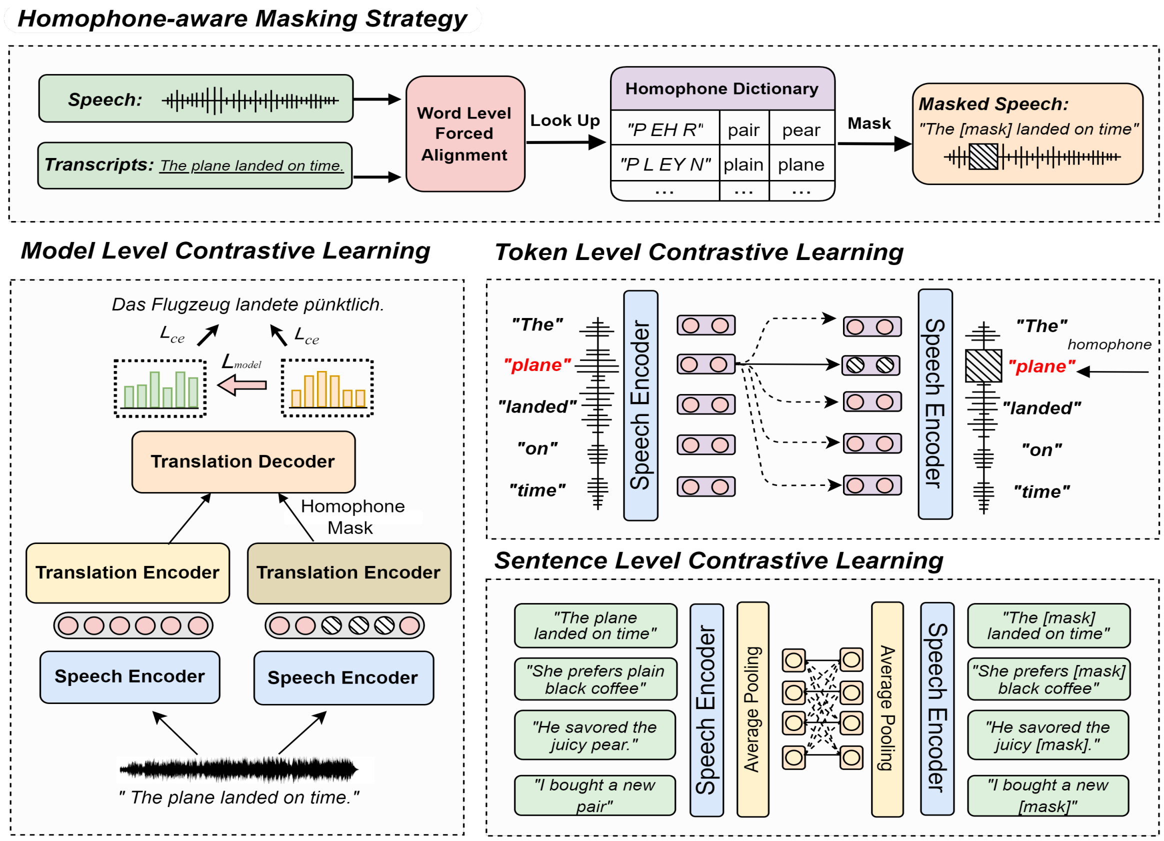
Homophone-aware Masking Strategy
Speech:
Transcripts: The plane landed on time.
Word Level Forced Alignment
Look Up
Homophone Dictionary
"P EH R" pair pear
"P L EY N" plain plane
Mask
Masked Speech:
"The [mask] landed on time"
Model Level Contrastive Learning
Das Flugzeug landete pünktlich.
L_{ce}
L_{model}
L_{ce}
Translation Decoder
Homophone Mask
Translation Encoder
Translation Encoder
Speech Encoder
Speech Encoder
" The plane landed on time."
Token Level Contrastive Learning
"The"
"plane"
"landed"
"on"
"time"
Speech Encoder
Speech Encoder
"The"
homophone
"plane"
"landed"
"on"
"time"
Sentence Level Contrastive Learning
"The plane landed on time"
"She prefers plain black coffee"
"He savored the juicy pear."
"I bought a new pair"
Speech Encoder
Average Pooling
Average Pooling
Speech Encoder
"The [mask] landed on time"
"She prefers [mask] black coffee"
"He savored the juicy [mask]."
"I bought a new [mask]"

Figure 2: Overall framework of AmbigST. The dashed/solid lines indicate negative/positive pairs. We first utilize a homophone-aware masking strategy, effectively masking the speech representation associated with ambiguous tokens. Then, we propose three-level contrastive learning methods. These methods facilitate the alignment of ambiguous words with their corresponding reference representation, generated from the original speech input.

2022b), enhancing its effectiveness in speech translation tasks. An overarching visual depiction of our proposed methodology can be found in Figure 2.

4.2 Homophone-aware Masking Strategy

As mentioned in §3, the conventional ST model encounters difficulties in handling speech sense ambiguity. Consequently, the model struggles to accurately capture the intended semantic properties of these ambiguous tokens. To alleviate the aforementioned challenge, we introduce a novel homophone-aware masking strategy leveraging the annotated dataset we have constructed. Given an speech, the speech encoder receives the original sequence $\mathbf{s}$ as input and produces its contextual representation $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_\mathrm{I}]$. We employ a word-level forced alignment technique (Fang et al., 2022) between the speech and transcriptions to pinpoint the temporal occurrence of individual words within the speech segment. Consequently, the homophone-aware masking matrix of the speech representation denoted as $\mathbf{m} = [\mathrm{m}_1, \mathrm{m}_2, ..., \mathrm{m}_\mathrm{I}]$, is generated in line with a homophone dictionary to identify the precise location of homophonic segments in the speech representation. With a certain probability p^*, the calculation of the masking matrix is:

$$\mathrm{m_i} = \begin{cases} 1 & i \in \mathbf{index}, p > p^* \\ 0 & else \end{cases}, \qquad (2)$$

where p is sampled from the uniform distribution $\mathcal{U}(0, 1)$, **index** represents the index set of the homophone representation in the speech sequence.

4.3 Homophone-aware Contrastive Learning

Token-level Learning We argue that the precise semantics of individual small homophone units are crucial to address speech sense ambiguity effectively. Our objective is to utilize homophone information to advance progress in contrastive learning, with a specific focus on intricate levels of granularity. Inspired by Su et al. (2022), we propose a token-level contrastive learning method. We utilize the identical model to generate the speech encoder output twice. In one instance, we apply a homophone-aware masking strategy to generate a masked representation, denoted as $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_\mathbf{1}, \tilde{\mathbf{h}}_\mathbf{2} ..., \tilde{\mathbf{h}}_\mathbf{I}]$. The proposed token-aware contrastive learning objective is then defined as:

$$\mathcal{L}_{\text{Token}} = -\sum_{i=1}^{I} \mathbb{I}(\tilde{\mathbf{h}}_{\mathbf{i}}) \log \frac{e^{\text{cosine}(\tilde{\mathbf{h}}_{\mathbf{i}},\mathbf{h}_{\mathbf{i}})/\tau}}{\sum_{j=1}^{I} e^{\text{cosine}(\tilde{\mathbf{h}}_{\mathbf{i}},\mathbf{h}_{\mathbf{j}})/\tau}}, \tag{3}$$

where $\mathbb{I}(\tilde{\mathbf{h}}_{\mathbf{i}})$ is an indicator function that evaluates to 1 if $\tilde{\mathbf{h}}_{\mathbf{i}}$ represents a masked homophone, and 0 otherwise. τ is a temperature hyper-parameter and cosine$(\cdot,\cdot)$ computes the cosine similarity. The underlying intuition of this approach is to encourage the generation of masked token that closely aligns with the corresponding homophone generated by the model while remaining distinguishable from other tokens within the sequence. As a result, this approach enhances the model's capacity to effectively understand the distinctive characteristics associated with homophones.

Sentence-level Learning To further enhance the effectiveness of contrastive learning and identify the optimal sentence-level representation, we introduce a self-supervised approach that focuses on sentence-level contrastive learning. Our approach incorporates the application of dropout noise (Gao et al., 2021) in conjunction with our proposed homophone-aware masking strategy. Subsequently, we average both $\mathbf{H}$ and $\tilde{\mathbf{H}}$ across the temporal dimension. As a result, we obtain sentence-level representations for both the original and homophone-masked forms, denoted as $\mathbf{z}$ and $\tilde{\mathbf{z}}$. For a minibatch of K examples, the sentence-level contrastive learning objective of n-th sentence is defined as:

$$\mathcal{L}_{\text{Sentence}} = -\log \frac{e^{cosine(\mathbf{z}_{\mathbf{n}},\tilde{\mathbf{z}}_{\mathbf{n}})/\tau}}{\sum_{k=1}^{K} e^{cosine(\mathbf{z}_{\mathbf{n}},\tilde{\mathbf{z}}_{\mathbf{k}})/\tau}}. \tag{4}$$

Employing this objective enables the model to grasp sophisticated semantic nuances by comprehensively considering the wider context and interconnections inherent within a sentence.

Model-level Learning We propose a refined model-level contrastive learning framework to ensure consistent guidance in extracting context-aware representations from speech. Crucially, this framework is specifically designed to tackle the challenges presented by homophones and assist the model in robustly identifying the information present near ambiguous tokens. Unlike the traditional knowledge distillation approach (Kim and Rush, 2016) matching predictions of a single sample from two networks, we enhance model performance by leveraging the inherent knowledge of a single network through predictions of diverse samples. *i.e.*, self-knowledge distillation (Zhang et al., 2019; Wang et al., 2022). This strategy can be seen as a unique variant of contrastive learning characterized by the presence of positive examples alone. Formally, given the original and masked contextual representations obtained from the speech encoder, the model-level contrastive learning objective is defined as:

$$\mathcal{L}_{\text{MCL}} = \sum_{j=1}^{|\mathbf{y}|} \text{KL}\left(P_\theta(\mathbf{y}_j|\mathbf{y}_{<j},\mathbf{H}) \| P_\theta(\mathbf{y}_j|\mathbf{y}_{<j},\tilde{\mathbf{H}})\right), \tag{5}$$

where KL denotes the Kullback-Leibler (KL) divergence. $P_\theta(\mathbf{y}_j|\mathbf{y}_{<j},\mathbf{H})$ is the predicted probability distribution of the j-th target token given the speech representation $\mathbf{H}$ as the input of translation encoder, and $P_\theta(\mathbf{y}_j|\mathbf{y}_{<j},\tilde{\mathbf{H}})$ is that given the masked representation as input. Therefore, by incorporating ST loss $\mathcal{L}_{\text{ST}}$, the ultimate training objective $\mathcal{L}_{\text{Ambig}}$ can be stated as follows:

$$\begin{aligned}\mathcal{L}_{\text{Ambig}} = \lambda(&\mathcal{L}_{\text{ST}} + \alpha\mathcal{L}_{\text{MCL}}) \\ &+ (1-\lambda)(\mathcal{L}_{\text{Token}} + \mathcal{L}_{\text{Sentence}}),\end{aligned} \tag{6}$$

where λ, α is the coefficient weight to control $\mathcal{L}_{\text{Ambig}}$. We use them to ensure a balanced distribution across diverse tasks, which in turn enhances the proficiency of disambiguation processes.

4.4 Discussion

In summary, our proposed three-level contrastive learning approach fundamentally focuses on refining training data from multiple granularities. The token-level method enhances the representation of ambiguous words, while sentence-level methods capture more detailed contextual features, and model-level optimization directly refines the model's parameters. Importantly, there is no expansion in model parameters or training data size, and no influence on decoding speed, making it easily integrated into existing frameworks.

5 Experiments

5.1 Experimental Setups

Dataset Our experiment are conducted on the MuST-C dataset (Di Gangi et al., 2019a), a multilingual speech translation corpus featuring over 385 hours of TED Talks speech in English with corresponding manual transcriptions and translations. For our specific analysis, we select three language pairs of En-{De, Es, Fr}. The dev set is employed for validation and the `tst-COMMON` set for testing. Additionally, we utilized the CoVoST-2 ST dataset (Wang et al., 2020a), another extensive multilingual ST corpus, focusing on translations of {De, Es, Fr}-En to further assess the model's translation accuracy across different linguistic contexts.

Pre-processing For speech input, we utilize raw 16-bit, 16kHz mono-channel audio waves. From the transcripts within the training set, we construct the homophone dictionaries, then we annotate dataset through a speech-text alignment process using the Montreal Forced Aligner[1]. For each translation direction, we use a unigram SentencePiece[2] to learn a vocabulary on the text data from the ST dataset. The vocabulary is shared for source and target with a size of 10K.

Model Configuration Our model consists of three modules. For the *speech encoder*, we use Hubert (Hsu et al., 2021). It adheres to the base configuration, utilizing six transformer layers with a hidden dimension size of 512. For the *translation encoder*, we use $N_e = 6$ transformer encoder layers. For the *translation decoder*, we use $N_d = 6$ transformer decoder layers. Each of these transformer layers comprises 512 hidden units, 8 attention heads, and 3,072 feed-forward hidden units.

Training and Inference During model training, we load a speech-text pre-training model[3]. The max-token is set to 800K (50 seconds), and we drop the training samples longer than that. We apply a speech masking probability of 5% and set label smoothing to 0.3. The learning rate increases linearly to $3e - 5$ in the first 5K steps, then decays linearly to zero in total 50K steps. The masking probability p is set at 0.5, a value determined through validation results among options of 0.3, 0.5, and 0.7. The temperature parameter, τ, is set at 0.01 and selected based on validation scores of 0.01, 0.1, and 1.0. The weight parameter, α, is set to 1.0 for En-De and 0.5 for other tasks, with the parameter λ being fixed at 0.9 across all tasks. We implement our models based on fairseq[4]. Models train on 2 GPUs with an update frequency of 16.

During inference, we average the checkpoints of the last five epochs for evaluation. We employ beam search with a beam size of 10 to generate optimal results. For the computation of evaluation metrics, we utilize sacreBLEU[5] (Post, 2018) to assess translation quality (Papineni et al., 2002), APT (Miculicich Werlen and Popescu-Belis, 2017) for evaluating homophone accuracy and BLEURT score[6] (Sellam et al., 2020) for measuring the quality of the generated text. We conduct the statistical significance with paired bootstrap resampling (Koehn, 2004) for BLEU score, paired t-tests (Hsu and Lachenbruch, 2014) for homophone accuracy, and Z-tests for BLEURT score.

5.2 Results on En-XX ST Tasks

Table 2 shows the results on MuST-C En→De, En→Fr, and En→Es `tst-COMMON` set. Models (1) to (5) represent the existing benchmarks. Model (6) represents our implementation of the previously state-of-the-art method, SpeechUT, in which we initialize the pre-trained parameters and fully fine-tune them. One should note that all our implemented systems share the same training hyperparameters and steps.

Sub-module Results We decompose AmbigST into sub-modules, starting with Model (7) employing model-level CL for foundational comprehension. Model (8) integrates sentence-level CL into Model (7), followed by Model (9), which replaces sentence-level CL with token-level CL to examine the impact of finer-grained methods. The results reveal that the individual application of each sub-module can lead to improvements.

Overall Results Model (10), denoted as AmbigST, incorporates all three proposed CL methods. AmbigST surpasses all strong baselines across all these three tasks, achieving an average BLEU score of 35.5, establishing a new state-of-the-art performance. Notably, AmbigST achieves marked improvements (+0.9 on average) in the more reliable metric BLEURT, indicating that it addresses more semantically complex difficulties rather than surface-level ones, which tend to pose more significant challenges. Subsequently, we assess the accuracy of homophone translation using our constructed homophone dictionary. AmbigST significantly and consistently outperforms the baseline across all language directions. This suggests that AmbigST can address the pervasive problem of speech sense ambiguity in ST tasks.

[1] `https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner`

[2] `https://github.com/google/sentencepiece`

[3] `https://github.com/microsoft/SpeechT5/tree/main/SpeechUT`

[4] `https://github.com/pytorch/fairseq`

[5] `https://github.com/mjpost/sacrebleu`

[6] `https://github.com/google-research/bleurt`

ID	Model	En-De			En-Es			En-Fr		
		BLEU	HomoP	BR	BLEU	HomoP	BR	BLEU	HomoP	BR
	Existing Method									
1	Context-ST (Zhang et al., 2021)	22.9	-	-	27.3	-	-	32.5	-	-
2	STEMM (Fang et al., 2022)	28.7	-	-	31.0	-	-	37.4	-	-
3	ConST (Ye et al., 2022)	28.3	-	64.5	32.0	-	-	38.3	-	-
4	SpeechUT (Zhang et al., 2022b)	30.1	-	-	33.6	-	-	41.4	-	-
5	CMOT (Zhou et al., 2023)	29.0	-	-	32.8	-	-	39.5	-	-
	Our Implemented Method									
6	Baseline (Zhang et al., 2022b)	30.1	39.2	67.1	33.5	42.8	67.7	41.4	49.1	66.2
7	6 + Model-level CL	30.3	39.7	67.7	33.7	43.2	**68.2**†	41.7	49.6	66.5
8	7 + Sentence-level CL	30.2	40.0	67.7	33.5	43.3	68.1	41.6	49.2	66.4
9	7 + Token-level CL	30.5	39.8	67.5	33.9	43.4	68.1	41.8	49.9	66.6
10	AmbigST (6 + All)	**30.6**†	**40.2**†	**67.8**†	**34.0**†	**43.5**†	68.0	**41.9**†	**49.9**†	**67.7**†

Table 2: Case-sensitive tokenized BLEU score, homophone accuracy and BLEURT score on MuST-C `tst-COMMON` set. "+" means incorporating contrastive learning (CL) into our models. "†" indicates that the proposed method is significantly better than baseline results at a significance level ($p < 0.05$).

Model	De-En			Es-En			Fr-En		
	BLEU	HomoP	BR	BLEU	HomoP	BR	BLEU	HomoP	BR
Baseline	24.2	52.8	56.2	29.6	48.5	61.7	29.9	36.9	48.6
AmbigST	**24.8**†	**53.6**†	**56.8**†	**30.1**†	**49.3**†	**62.4**†	**30.8**†	**37.8**†	**49.2**†

Table 3: Case-sensitive tokenized BLEU score, homophone accuracy and BLEURT score on CoVoST test set.

5.3 Results on XX-En ST Tasks

Table 3 shows the results on the CoVoST test sets. We began by initializing the parameters from En→De, En→Es, En→Fr pretrained model (Zhang et al., 2022b), subsequently fine-tuning these parameters with out proposed contrastive learning methods. AmbigST outperforms established baselines in terms of all the metrics. This evidence highlights AmbigST's ability to boost model performance consistently across different languages, showcasing its language-agnostic nature.

6 Analysis

6.1 Effect of Homophone Dictionary Size

To evaluate the robustness of our approach against varying sizes of homophone dictionaries, we conducted experiments using the English homophone dictionary segmented into distinct scales {0.3, 0.5, 0.7}. We reconstituted the annotated En-De dataset to determine the impact of homophone dictionary size on AmbigST performance. As depicted in Figure 3, our method consistently enhances performance across different dataset sizes, demonstrating its efficacy and stability. As the size of the homophone dictionary increases, there is a corresponding improvement in overall performance, affirming the critical influence of dictionary size. By employing our fully constructed homophone dictionary, we achieved peak performance across all metrics, thereby confirming the precision and effectiveness of our homophone dictionary construction process.

6.2 Variant of Homophone Setups

To investigate whether the enhancements are attributed to homophone-aware masking, we execute two supplementary experiments: (1) replacing the masked representation with speech-text mixup representation (Fang et al., 2022) and (2) substituting a homophone mask with a random mask.

Mixup Representation We replace the segment of speech representation corresponding with the word embedding of the homophone. The experimental results, as depicted in Table 4, demonstrate that while the "Mixup" brings a slight improvement in model performance, "AmbigST" is superior. Compared with the mixup strategy, our approach further improves the performance on all metrics, proving that the homophone-aware mask may be a more promising strategy for addressing speech sense ambiguity.

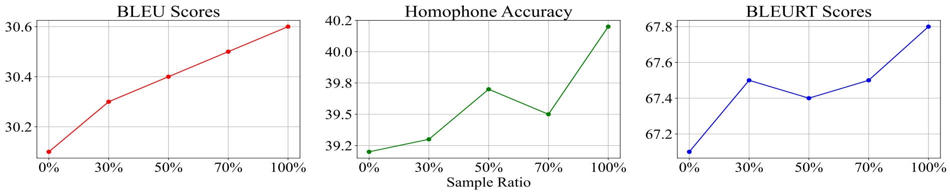

Figure 3: BLEU, homophone accuracy and BLEURT on En-De `tst-COMMON` under different size of dictionary.

Variant	BLEU	HomoP	BR
Baseline	30.1	39.2	67.1
AmbigST	**30.6**	**40.2**	**67.8**
Representation (Default: masked speech)			
Mixup	30.2	39.1	67.4
Mask Strategy (Default: homophone)			
Random	30.2	39.9	67.4

Table 4: BLEU, homophone accuracy and BLEURT on En-De `tst-COMMON` under different setups.

Strategy	1-gram	2-gram	3-gram	4-gram	All
Baseline	61.8	36.0	23.4	15.7	30.1
Random Mask	61.9	36.2	23.6	15.8	30.2
Homophone Mask	**62.0**	**36.5**	**23.9**	**16.1**	**30.6**

Table 5: N-gram BLEU score on En-De `tst-COMMON` under different mask strategy.

Extra Task	BLEU	HomoP	BR
Baseline	30.1	39.2	67.1
+ T2T Task	29.9	39.7	67.3
+ U2T Task	29.8	39.5	67.5
+ CoLaCTC Task	30.0	39.5	67.5
AmbigST	**30.6**	**40.2**	**67.8**

Table 6: BLEU, homophone accuracy and BLEURT on En-De `tst-COMMON` under different auxiliary tasks.

Model	Noun	Verb	Adj.	Adv.	Conj.	Pron.	Det.	All
	165	39	22	13	10	5	5	259
Baseline	35.4	23.5	**36.0**	50.9	49.6	48.3	41.1	39.2
AmbigST	**36.1**	**24.5**	35.5	**52.3**	**51.0**	**50.0**	**41.5**	**40.2**

Table 7: Homophone accuracy according to different parts of speech on the En-De `tst-COMMON`. We calculate the number of types. Most of the homophone words appearing are nouns.

Random Strategy We randomly mask the speech representation with a 10% probability and conduct a three-level contrastive learning method. As shown in Table 4, our investigation reveals that although random masking alleviates issues with homophones to some extent, there is no significant improvement in BLEU and BLEURT scores. This effect may stem from a random mask similar to the masking strategy of BERT (Kenton and Toutanova, 2019), which allows other unmasked words to enhance their ability to capture context information and, therefore, can potentially affect the translation accuracy of homophones. However, these words are not central in a sentence. As shown in Table 5, we further calculate the n-gram score in the BLEU metric to compare the homophone mask and random mask strategies. While random masking may achieve comparable 1-gram scores, but it fails to address the core challenging words within the sentences, notably in the 2-gram, 3-gram, and 4-gram evaluations. Through our proposed homophone masking strategy, the model can more effectively capture the contextual and nuanced aspects of the language, resulting in more precise translation.

6.3 Auxiliary Textual Translation Task

We conducted a comprehensive comparison with methods that indirectly address homophone ambiguity. Specifically, we evaluated our approach against models utilizing cross-modal multi-task learning techniques, including text-to-text (T2T) (Ouyang et al., 2022), unit-to-text (U2T) (Kim et al., 2023; Zhang et al., 2023b), and coarse labeling for CTC (CoLaCTC) tasks (Zhang et al., 2023a). As shown in Table 6, the inclusion of auxiliary T2T and U2T tasks resulted in slight improvements in homophone accuracy and BLEURT score, albeit with a marginal reduction in BLEU score. This observation can be attributed to the highly optimized fusion of speech and text achieved through robust unified pretraining. Despite the assertion by Zhang et al. (2023a) that the CoLaCTC loss encourages the encoder to consider contextual clues, our direct strategy exhibits clear advantages. AmbigST effectively harnesses the intrinsic homophone properties within the ST model, thereby substantiating its efficacy in mitigating speech sense ambiguity.

Model	Non-homophone		Homophone		All	
	ACC	Δ	ACC	Δ	ACC	Δ
Baseline	44.5	-	39.2	-	40.3	-
AmbigST	**44.6**	**0.1**	**40.2**	**1.0**	**41.0**	**0.7**

Table 8: Translation accuracy of randomly selected words, homophones, and all words in the En-De `tst-COMMON`.

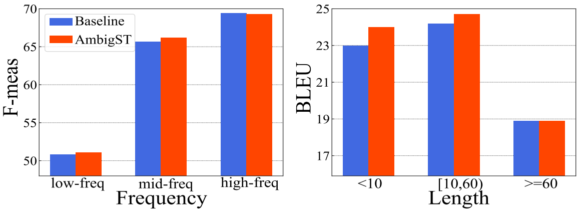

Figure 4: Word frequency and sentence length analysis.

6.4 Effect of Homophone Type

To examine the impact of homophone type, we categorized the ambiguous words that appeared in the En-De tst-COMMON set according to their part of speech. Subsequently, we analyze the count of ambiguous words in each category according to part of speech and assess their homophone accuracy respectively. As illustrated in Table 7, our approach proficiently disambiguates homophones across nearly all parts of speech except adjectives. AmbigST demonstrates an average enhancement of 1.7 in the homophone accuracy of conjunctions and pronouns. Furthermore, our analysis reveals that most homophones in our dataset comprise nouns (165 words) and verbs (39 words), predominantly utilized in daily discourse. Our method achieves substantial improvements in these domains, registering an average increase of 0.8 in the accuracy of homophone translation. This confirms that our method is effectively tuned to the requirements of daily oral communication.

6.5 Homophone vs. Non-homophone

To concisely illustrate that AmbigST can address difficulties posed by homophones, we conducted a comparative analysis using the En-De tst-COMMON set. Firstly, we selected 200 non-homophones at random and determined their translation accuracy across the test set, averaging results from five distinct data sets to ensure reliability. Secondly, we assessed translation accuracy across all word occurrences in the test set. Then, we compared the results of non-homophones, homophones, and all words between the baseline and AmbigST. Our results, as detailed in Table 8, starkly highlight the increased challenge homophones present to baseline systems. The baseline model, indeed, has a lower translation accuracy on homophones compared to other words. While AmbigST does not significantly improve translation accuracy for non-homophones, it substantially enhances the accuracy for homophones, thereby improving overall translation accuracy. This enhancement substantially increases the overall translation accuracy, demonstrating AmbigST's efficacy in effectively mitigating issues related to speech ambiguity.

6.6 Analyses of Word Frequency and Sentence Length

To elucidate how AmbigST improves translation quality, we employ compare-mt (Neubig et al., 2019) for a comparative analysis with a baseline model, focusing on *word frequency* and *sentence length*. Our findings, illustrated in Figure 4, reveal that AmbigST demonstrates enhanced resilience to low (less than 10) and mid-frequency (10-1,000) words, which often include a high incidence of homophones, underscoring our method's effectiveness on SSD. Moreover, AmbigST favors shorter sentences, particularly those under 60 words. Shorter sentences might lack explicit contextual clues; however, our method enables the model to grasp fine-grained information more effectively. Consequently, it achieves stronger translation performance, even when the text contains limited contextual information.

7 Conclusion

To mitigate speech sense ambiguity in ST models, this paper represents a pioneering effort that involves the construction of comprehensive homophone dictionaries for En, Fr, De, Es and six language pair-annotated datasets that include pertinent homophone information. In the wake of these advancements, we introduce *AmbigST*, a novel homophone-aware contrastive learning approach that incorporates a homophone-aware masking strategy operating at the token, sentence, and model levels. Experimental results on the MuST-C and CoVoST benchmark demonstrate that our proposed AmbigST approach can leverage speech sense disambiguation efficiently.

Limitations

While the proposed AmbigST can effectively alleviate speech sense ambiguity, it still has the following limitations:

- **Word Sense Ambiguity Not Address** Our research primarily tackles homophones to explore speech sense ambiguity in end-to-end ST systems. However, we do not delve into "word sense" in this research. Future efforts will aim to investigate word sense ambiguity through our three-level contrastive method to validate the effectiveness of our method.

- **Challenges in Low-Resource Languages** While our methodology is intended to be language-agnostic, its efficacy may diminish in low-resource environments where transcription quality is suboptimal. Such conditions may impede the training of precise acoustic models with tools such as MFA, leading to potential inaccuracies in both the homophone dictionary and the annotated dataset, and thereby constraining model performance.

- **System Performance Variability** Due to regional differences and dialect variations, there is a diverse understanding of ambiguity. This necessitates training acoustic models on speech tailored to specific phonetic preferences to construct a dictionary of ambiguous words. For convenience, this work uses a standard pronunciation acoustic model from MFA, which may disadvantage those with varying pronunciation preferences.

- **Performance in Specific Contexts** As detailed in Section 6.6, the efficacy of our methodology appears diminished when applied to long sentences. This observation may stem from the intrinsic abundance of contextual cues within such sentences, which inherently diminishes the necessity and impact of our approach.

Ethics Statement

Our research rigorously adheres to ethical considerations in line with the ACL Ethics Policy. We employ the publicly accessible MuST-C dataset and the pre-trained SpeechUT model, both sanctioned for research use. Our study aims to address homophone ambiguity in end-to-end speech translation, for which we develop a homophone dictionary and a detailed homophone-annotated dataset. In accordance with the original license, we release the annotated dataset under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY NC ND 4.0). For reproducibility, we offer both datasets and code to fellow researchers upon request, ensuring the objective and precise presentation of our results.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62206076 and 62276077), Guangdong Basic and Applied Basic Research Foundation (Grant Nos. 2024A1515011491 and 2024A1515011205), Shenzhen Science and Technology Program (Grant Nos. ZDSYS20230626091203008, KJZD20231023094700001 and RCBS20221008093121053), and Shenzhen College Stability Support Plan (Grant Nos. GXWD20220811173340003, GXWD20220817123150002 and GXWD20220811170358002). Xuebo Liu was sponsored by CCF-Tencent Rhino-Bird Open Research Fund. We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions.

References

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana. Association for Computational Linguistics.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019. A comparative study on end-to-end speech to text translation. In *Proc. of ASRU*, pages 792–799. IEEE.

Jeong-Uk Bang, Min-Kyu Lee, Seung Yun, and Sang-Hun Kim. 2022. Improving end-to-end speech translation model with bert-based contextual information. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6227–6231. IEEE.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mslam: Massively multilingual joint pre-training for speech and text. *ArXiv preprint*, abs/2202.01374.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.

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. In *NIPS workshop on End-to-end Learning for Speech and Audio Processing*.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.

HoLam Chung, Junan Li, Pengfei Liu, Wai-Kim Leung, Xixin Wu, and Helen Meng. 2022. Improving rare words recognition through homophone extension and unified writing for low-resource cantonese speech recognition. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 26–30. IEEE.

Hexuan Deng, Liang Ding, Xuebo Liu, Meishan Zhang, Dacheng Tao, and Min Zhang. 2023. Improving simultaneous machine translation with monolingual data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11, pages 12728–12736.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Mattia Antonino Di Gangi, Matteo Negri, Roldano Cattoni, Roberto Dessi, and Marco Turchi. 2019b. Enhancing transformer for end-to-end speech-to-text translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 21–31, Dublin, Ireland. European Association for Machine Translation.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021a. Consecutive decoding for speech-to-text translation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12738–12748. AAAI Press.

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021b. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12749–12759. AAAI Press.

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, San Diego, California. Association for Computational Linguistics.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *ArXiv preprint*, abs/2005.12766.

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 (Volume 1: Long Papers)*, pages 7050–7062, Dublin, Ireland. Association for Computational Linguistics.

Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. 2019. Adapting transformer to end-to-end spoken language translation. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1133–1137. ISCA.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Papri Ghosh, Tejbanta Singh Chingtham, and Mrinal Kanti Ghose. 2016. Slhar: A supervised learning approach for homophone ambiguity reduction from speech recognition system. In *2016 Second international conference on research in computational intelligence and communication networks (ICRCICN)*, pages 12–16. IEEE.

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 2021*, pages 2214–2225, Online. Association for Computational Linguistics.

Henry Hsu and Peter A Lachenbruch. 2014. Paired t test. *Wiley StatsRef: statistics reference online*.

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

Wuwei Huang, Dexin Wang, and Deyi Xiong. 2021. AdaST: Dynamically adapting encoder states in the decoder for end-to-end speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2539–2545, Online. Association for Computational Linguistics.

Amir Hussein, Brian Yan, Antonios Anastasopoulos, Shinji Watanabe, and Sanjeev Khudanpur. 2023. Enhancing end-to-end conversational speech translation through target language context utilization. *arXiv preprint arXiv:2309.15686*.

Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe. 2021. Source and target bidirectional knowledge distillation for end-to-end speech translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1872–1881, Online. Association for Computational Linguistics.

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.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2017. Structured-based curriculum learning for end-to-end english-japanese speech translation. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2630–2634. ISCA.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Minsu Kim, Jeongsoo Choi, Dahun Kim, and Yong Man Ro. 2023. Many-to-many spoken language translation via unified speech and text representation learning with unit-to-unit translation. *arXiv preprint arXiv:2308.01831*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

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. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3520–3533, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. 2023. Unsupervised dense retrieval with relevance-aware contrastive pretraining. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10932–10940, Toronto, Canada. Association for Computational Linguistics.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021a. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, Online. Association for Computational Linguistics.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021b. Multilingual speech translation from efficient finetuning of pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 827–838, Online. Association for Computational Linguistics.

Xiang Li, Haiyang Xue, Wei Chen, Yang Liu, Yang Feng, and Qun Liu. 2018. Improving the robustness of speech translation. *arXiv preprint arXiv:1811.00728*.

Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2018. Robust neural machine translation with joint textual and phonetic embedding. *arXiv preprint arXiv:1810.06729*.

Yixing Luan, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. 2020. Improving word sense disambiguation with translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4055–4065, Online. Association for Computational Linguistics.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 498–502. ISCA.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, and Afra Alishahi. 2023. Homophone disambiguation reveals patterns of context mixing in speech transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8249–8260, Singapore. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Siqi Ouyang, Rong Ye, and Lei Li. 2022. Waco: Word-aligned contrastive learning for speech translation. *ArXiv preprint*, abs/2212.09359.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Wenjie Qin, Xiang Li, Yuhui Sun, Deyi Xiong, Jianwei Cui, and Bin Wang. 2021. Modeling homophone noise for robust neural machine translation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7533–7537. IEEE.

Jun Rao, Liang Ding, Shuhan Qi, Meng Fang, Yang Liu, Li Shen, and Dacheng Tao. 2023. Dynamic contrastive distillation for image-text retrieval. *IEEE Transactions on Multimedia*.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Jennifer Rodd, Gareth Gaskell, and William Marslen-Wilson. 2002. Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of memory and language*, 46(2):245–266.

Elizabeth Salesky, Susanne Burger, Jan Niehues, and Alex Waibel. 2018. Towards fluent translations from disfluent speech. In *Proc. of SLT*, pages 921–926. IEEE.

Elizabeth Salesky, Matthias Sperber, and Alexander Waibel. 2019. Fluent translations from disfluent speech in end-to-end speech translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2786–2792, Minneapolis, Minnesota. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *ArXiv preprint*, abs/2009.13818.

Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. 2022. TaCL: Improving BERT pre-training with token-aware contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2497–2507, Seattle, United States. Association for Computational Linguistics.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. 2022. Unified speech-text pre-training for speech translation and recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1488–1499, Dublin, Ireland. Association for Computational Linguistics.

Laura Cross Vila, Carlos Escolano, José AR Fonollosa, and Marta R Costa-jussà. 2018. End-to-end speech translation with the transformer. In *IberSPEECH*, pages 60–63.

Hari Krishna Vydana, Martin Karafiát, Katerina Zmolikova, Lukáš Burget, and Honza Černockỳ. 2021. Jointly trained transformers models for spoken language translation. In *Proc. of ICASSP*, pages 7513–7517. IEEE.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. European Language Resources Association.

Changhan Wang, Anne Wu, Juan Pino, Alexei Baevski, Michael Auli, and Alexis Conneau. 2021. Large-scale self- and semi-supervised learning for speech translation. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2242–2246. ISCA.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020b. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3728–3738, Online. Association for Computational Linguistics.

Zhijun Wang, Xuebo Liu, and Min Zhang. 2022. Breaking the representation bottleneck of chinese characters: Neural machine translation with stroke sequence modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6473–6484.

Di Wu, Liang Ding, Shuo Yang, and Mingyang Li. 2022a. Mirroralign: A super lightweight unsupervised word alignment model via cross-lingual contrastive learning. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 83–91.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022b. ESim-CSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3898–3907, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 2619–2630, Online. Association for Computational Linguistics.

Haiyang Xue, Yang Feng, Shuhao Gu, and Wei Chen. 2020. Robust neural machine translation with asr errors. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 15–23.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2267–2271. ISCA.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5099–5113, Seattle, United States. Association for Computational Linguistics.

Tengfei Yu, Liang Ding, Xuebo Liu, Kehai Chen, Meishan Zhang, Dacheng Tao, and Min Zhang. 2023. Promptst: Abstract prompt learning for end-to-end speech translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10140–10154.

Biao Zhang, Barry Haddow, and Rico Sennrich. 2023a. Efficient CTC regularization via coarse labels for end-to-end speech translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2264–2276, Dubrovnik, Croatia. Association for Computational Linguistics.

Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. 2021. Beyond sentence-level end-to-end speech translation: Context helps. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2566–2578.

Dong Zhang, Rong Ye, Tom Ko, Mingxuan Wang, and Yaqian Zhou. 2023b. Dub: Discrete unit back-translation for speech translation. *arXiv preprint arXiv:2305.11411*.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3712–3721. IEEE.

Tong Zhang, Wei Ye, Baosong Yang, Long Zhang, Xingzhang Ren, Dayiheng Liu, Jinan Sun, Shikun Zhang, Haibo Zhang, and Wen Zhao. 2022a. Frequency-aware contrastive learning for neural machine translation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11712–11720. AAAI Press.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. 2022b. SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1663–1676, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12736–12746. PMLR.

Yi Zheng, Xianjie Yang, and Xuyong Dang. 2020. Homophone-based label smoothing in end-to-end automatic speech recognition. *arXiv preprint arXiv:2004.03437*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. E2s2: Encoding-enhanced sequence-to-sequence pretraining for language understanding and generation. *IEEE Transactions on Knowledge and Data Engineering*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. Cmot: Cross-modal mixup via optimal transport for speech translation. *ArXiv preprint*, abs/2305.14635.

A Appendix

A.1 Statistic of Homophone Dictionary and Annotated Dataset

Source	Low	Mid	High	Total
En	808	535	77	1,420
De	408	301	42	751
Fr	15,077	5,525	36	20,638
Es	914	521	16	1,451

Table 9: Statistics of the number of words in homophone dictionary. For English homophones, we calculate the words appear in the en-de dataset.

Dataset	En-De	En-Es	En-Fr	Total
MuST-C	189,890	221,974	226,912	638,776
Dataset	**De-En**	**Es-En**	**Fr-En**	**Total**
CoVoST	118,844	71,317	184,544	374,205

Table 10: Statistics of the number of items in the dataset.

We develop homophone dictionaries for English, German, Spanish and French, categorizing homophones into three levels based on word frequency appearing in training dataset: low-frequency (less than 10), mid-frequency (10-1000), and high-frequency (more than 1000). The results are detailed in Table 9. Our analysis reveals a predominance of low-frequency words among homophones, with a particularly notable concentration in the French language. This is attributed to the characteristic of many French words ending in “s” that are not pronounced. This finding is consistent with the research presented by Mohebbi et al. (2023), which underscores the widespread occurrence of homophones in French.

We employed the constructed homophone dictionary to systematically annotate the MuST-C and CoVoST datasets, which cover En-{De, Fr, Es} and {De, En, Fr}-En language pairs, respectively. This annotation process included the word-level alignment between speech and its corresponding source transcription text, alongside the identification and indexing of homophones. The quantity of all annotated datasets is detailed in Table 10, with a total of 639k instances annotated in the MuST-C dataset and 374k in the CoVoST dataset.