IEICE TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2023EDP7249

Publicized:2024/05/24

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Neural End-to-end Speech Translation Leveraged by ASR Posterior Distribution

Yuka KO[†a)], **Katsuhito SUDOH**[†], ***Nonmembers***, **Sakriani SAKTI**[†], ***and*** **Satoshi NAKAMURA**[†,††], ***Members***

SUMMARY End-to-end speech translation (ST) directly renders source language speech to the target language without intermediate automatic speech recognition (ASR) output as in a cascade approach. End-to-end ST avoids error propagation from intermediate ASR results. Although recent attempts have applied multi-task learning using an auxiliary task of ASR to improve ST performance, they use cross-entropy loss to one-hot references in the ASR task, and the trained ST models do not consider possible ASR confusion. In this study, we propose a novel multi-task learning framework for end-to-end STs leveraged by ASR-based loss against posterior distributions obtained using a pre-trained ASR model called ASR posterior-based loss (ASR-PBL). The ASR-PBL method, which enables a ST model to reflect possible ASR confusion among competing hypotheses with similar pronunciations, can be applied to one of the strong multi-task ST baseline models with Hybrid CTC/Attention ASR task loss. In our experiments on the Fisher Spanish-to-English corpus, the proposed method demonstrated better BLEU results than the baseline that used standard CE loss.

key words: *end-to-end speech translation, spoken language translation, multi-task learning, knowledge distillation*

1. Introduction

Speech translation (ST), which translates a source language speech to a target language text, has been improved largely by recent advances in deep neural network-based methods for speech and language processing. This work focuses on such neural methods for ST. A simple approach to ST is to cascade automatic speech recognition (ASR) and machine translation (MT) [1]–[3]. However, a crucial flaw in this approach is the error propagation from ASR to MT, which makes an ST system sensitive to ASR errors. A major solution is to use many ASR hypotheses in the form of N-best lists and lattices [4].

Recent ST studies have attempted an end-to-end approach that directly translates a source language speech to the target language [5]–[7]. Since it does not use any intermediate ASR results, it is thus free from ASR error propagation. However, end-to-end ST's translation performance is usually worse than a cascade ST due to the lack of ST data. Multi-task learning [8]–[11] is a promising approach to fill the gap between cascade and end-to-end STs since it is leveraged by an ASR subtask while training the end-to-end ST. This approach introduces an additional output layer to obtain ASR results from hidden vectors in an end-to-end ST model that is trained using an ASR-oriented loss function against reference transcriptions.

Such ASR loss is usually calculated as softmax cross entropy (CE) against a one-hot distribution given by the unique reference tokens at every step of the ASR prediction. This approach prompts the ASR layer to predict a single hypothesis that dominates the output distribution and avoids confusion in the outputs. However, since our objective here is to obtain correct translations, we do not need to obtain correct one-hot ASR predictions without confusion.

Osamura et al. [12] focused on this issue and proposed a robust cascade ST that takes ASR word posterior distributions as input to its MT module to consider ASR confusion in the translation step. Bahar et al. [13] proposed tight integration of cascade ST by passing the ASR posterior probabilities to MT during training. Dalmia et al. [14] trained a cascade ST model by passing to the MT subnet the intermediate hidden representation given by the ASR decoder. However, most of these previous studies focused on cascaded frameworks that addressed ASR error propagation by considering other hypotheses in ST integration using ASR posterior probabilities or intermediate hidden representations.

Inspired by them, we take a further step and propose a novel multi-task learning framework for end-to-end ST leveraged by ASR-based loss against posterior distributions obtained using a pre-trained ASR model called ASR posterior-based loss: ASR-PBL. In this method, the posterior distribution is given by another pre-trained ASR model. Then the ST model trained with ASR-PBL mimics the outputs of the pre-trained model in the ASR subtask.

From another perspective, Chuang et al. [15] proposed using such distributional loss in multi-task learning for end-to-end STs. They incorporated semantic similarity in ASR-based loss; we focus on ASR confusion and errors caused by pronunciation variations, unclear utterances, noise, etc., all of which frequently appear in practice.

There are several methods using knowledge distillation (KD) [16], [17] for ST models. Liu et al. [18] used sequence-level KD [17] in which an ST model is trained against pseudo-reference translations given by a pre-trained MT model. Gaido et al. [19] extended this approach by a token-level KD to fine-tune an ST model, which is trained to mimic the posterior distribution given by a pre-trained MT model at every step of predicting the ST output. Their motivation was to use the knowledge of MT models created

[†]Nara Institute of Science and Technology

[††]Satoshi Nakamura conducted the research at Nara Institute of Science and Technology. He is now also affiliated with School of Data Science, Chinese University of Hong Kong, Shenzhen.

a) E-mail: ko.yuka.kp2@is.naist.jp

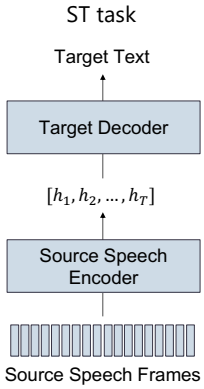


Fig. 1 End-to-end ST by single-task learning

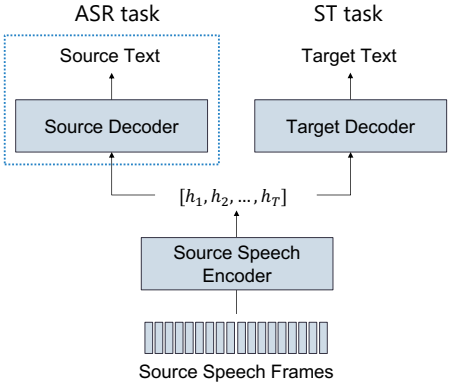


Fig. 2 End-to-end ST by multi-task learning

with large amounts of data to leverage the performance of the ST model that was created with only small amounts of ST data. On the contrary, our work uses a token-level KD in an ASR subtask of multi-task learning to include ASR confusion in the training of end-to-end ST. Instead of focusing on KD with a large amount of data like previous works, we focus on improving the model's performance by giving ASR confusion with only existing ST data. This method can be applied to a robust multi-task ST baseline model with Hybrid CTC/Attention ASR task loss. In our experiments, we applied our proposed ASR-PBL to the Hybrid CTC/Attention ASR task loss structure. Based on the Hybrid CTC/Attention ASR task loss, our experimental results on the Fisher Spanish-to-English data show that our proposed method obtained better BLEU scores than the baseline with the standard CE loss.

2. End-to-end Speech Translation

2.1 Single-task End-to-end Speech Translation

An end-to-end ST model consists of a source language speech encoder and a target language text decoder. Let $\mathbf{X} = x_1, ..., x_T$ be a sequence of source speech feature vectors and let $\mathbf{Y} = y_1, ..., y_N$ be a sequence of target language tokens, where T and N are the lengths of sequences $\mathbf{X}$ and $\mathbf{Y}$. Typically, each x_i $(1 \leq i \leq T)$ corresponds to a speech frame (e.g., 20 milliseconds), and each y_i corresponds to a subword token. Here i-th output token y_i depends on input $\mathbf{X}$ and the outputs up to the previous step, $y_1, ..., y_{i-1}$, and the posterior probability of v chosen from target language vocabulary V as y_i is denoted as:

$$P_{\mathrm{ST}}(y_i = v) = p(v|\mathbf{X}, y_{1:i-1}). \tag{1}$$

The ST model is trained using the following loss function defined by CE against reference y_i represented by a one-hot distribution:

$$\mathcal{L}_{\mathrm{ST}} = -\sum_{i=1}^{N}\sum_{v\in V} q(y_i, v) \ln P_{\mathrm{ST}}(y_i = v), \tag{2}$$

where $q(y_i, v)$ is an indicator function that returns 1 if $y_i = v$ and otherwise 0.

Recent studies on neural sequence-to-sequence models usually apply label smoothing [20], [21] to avoid overfitting, which distributes the probability mass onto the elements in V other than the ground truth reference. When v is a token in vocabulary V, the reference probability of vector $q(y_i, v)$ is represented as:

$$q(y_i, v) = \begin{cases} 1-\epsilon & \text{if } y_i = v, \\ \epsilon/(V-1) & \text{otherwise} \end{cases} \tag{3}$$

where label smoothing weight is ϵ (generally set to 0.1). When $\epsilon = 0$, it denotes CE loss without label smoothing. Fig. 1 diagrams a the single-task ST. The speech encoder converts input $\mathbf{X}$ into a sequence of hidden vectors $\mathbf{H} = h_1, ..., h_T$, and the target language text decoder predicts $\mathbf{Y}$ as the translation result.

2.2 Multi-task End-to-end Speech Translation

In contrast to a single-task end-to-end ST explicit ASR module, the multi-task approach uses an additional decoder to transcribe the input speech from the hidden vectors given by the speech encoder as an ASR subtask (Fig. 2). The loss function in it is denoted as follows using ASR posterior probability P_{ASR}, defined similarly to P_{ST}:

$$\mathcal{L}_{\mathrm{ASR}} = -\sum_{i=1}^{N}\sum_{v\in V} q(y_i, v) \ln P_{\mathrm{ASR}}(y_i = v). \tag{4}$$

The overall loss function is given by a weighted sum of $\mathcal{L}_{\mathrm{ST}}$ and $\mathcal{L}_{\mathrm{ASR}}$ using a hyperparameter λ_{ASR}:

$$\mathcal{L} = (1 - \lambda_{\mathrm{ASR}})\mathcal{L}_{\mathrm{ST}} + \lambda_{\mathrm{ASR}}\mathcal{L}_{\mathrm{ASR}}. \tag{5}$$

2.3 Hybrid CTC/Attention Loss for Multi-task Speech Translation

A Connectionist Temporal Classification (CTC) [22] model

can be optimized using the forward-backward technique, which determines the optimal alignments between speech and textual representations. CTC's key idea is the implementation of intermediate label structure, denoted as $\mathbf{\Pi} = \pi_1, ..., \pi_T$. This step allows repeated labels and the inclusion of unique blank labels (–) that represent emissions devoid of labels, implying $\pi_t \in \{1, ..., K\} \cup \{-\}$. The aim of CTC is to enhance the value of $P(\mathbf{Y}|\mathbf{X})$, which denotes the probability distribution over all possible label sequences $\Phi(\mathbf{Y}')$:

$$P(\mathbf{Y}|\mathbf{X}) = \sum_{\mathbf{\Pi} \in \Phi(\mathbf{Y}')} P(\mathbf{\Pi}|\mathbf{X}) \tag{6}$$

In this context, $\mathbf{Y}'$ is an augmented version of $\mathbf{Y}$, achieved by interspersing blank symbols between each label as well as at the start and finish. An example is transforming from (h, e, l, l, o) in $\mathbf{Y}$ to $(-, h, e, -, l, -, l, o)$ in $\mathbf{Y}'$. The probability of a label sequence, $P(\mathbf{\Pi}|\mathbf{X})$, is approximated as the outcome of independent network outputs:

$$P(\mathbf{\Pi}|\mathbf{X}) \approx \prod_{t=1}^{T} P(\pi_t|\mathbf{X}) = \prod_{t=1}^{T} q_t(\pi_t). \tag{7}$$

Here $q_t(\pi_t)$ denotes softmax activation corresponding to label π_t in output layer q at time t. The CTC loss, which needs to be minimized, is the negative log-likelihood of true token sequence $\mathbf{Y}^*$:

$$\mathcal{L}_{\text{CTC}} = -\ln P(\mathbf{Y}^*|\mathbf{X}). \tag{8}$$

Probability distribution $P(\mathbf{Y}|\mathbf{X})$ can be determined using the forward-backward method:

$$P(\mathbf{Y}|\mathbf{X}) = \sum_{u=1}^{|\mathbf{Y}'|} \frac{\alpha_t(u)\beta_t(u)}{q_t(y'_u)}, \tag{9}$$

where $\alpha_t(u)$ is the forward variable that, represents the total probability of all possible prefixes $(y'_{1:u})$ that end with the u-th label, and $\beta_t(u)$ is a backward variable of all possible suffixes $(y'_{u:U})$ that start with the u-th label. The network is trained with standard back-propagation by taking the derivative of the loss function concerning $q_t(k)$ for all labels k, including the blanks.

Since CTC does not explicitly model the relationships between labels due to its assumption of conditional independence, as shown in Eq. 7, its capacity to represent character-level linguistic details is constrained. As a result, it's typical to integrate lexicons or linguistic models, such as in the hybrid approach.

Recently, a hybrid approach using both the attention-based encoder-decoder and CTC models has been proposed for ASR performance improvement. Since no language model is included in the above CTC-based ASR framework, such a hybrid approach supports the inclusion of language model constraints in the CTC-based ASR, such as Hybrid CTC/Attention [23]. In this paper, we applied Hybrid CTC/Attention loss to the ASR task loss of a multi-task ST. The ASR task loss $\mathcal{L}_{\text{ASR}}$ function is given by a weighted sum of $\mathcal{L}_{\text{Att}}$ and $\mathcal{L}_{\text{CTC}}$ using hyperparameter λ_{CTC}:

$$\mathcal{L}_{\text{ASR}} = (1 - \lambda_{\text{CTC}})\mathcal{L}_{\text{Att}} + \lambda_{\text{CTC}}\mathcal{L}_{\text{CTC}}. \tag{10}$$

Note that we did not apply Hybrid CTC/Attention to our pre-trained ASR model for implementation simplicity.

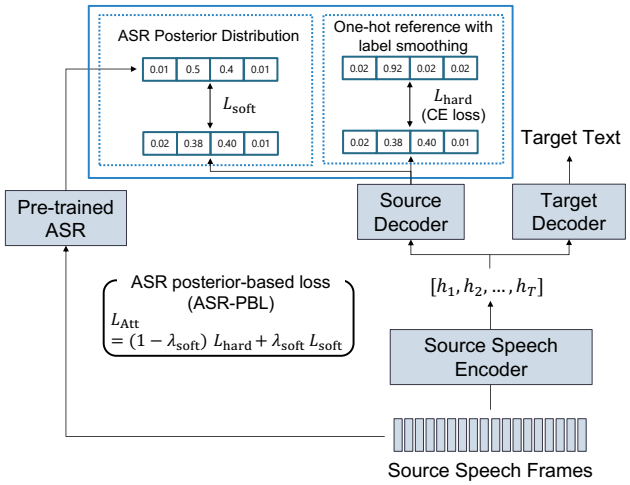


Fig. 3 Multi-task end-to-end ST with hard or soft target loss by pre-trained ASR model: We defined ASR task loss mixed with hard and soft target loss as our proposed ASR-PBL.

3. Proposed Method

We propose a method to train an ST model in a multi-task manner using a loss function based on posterior distributions given by a pre-trained ASR model instead of single ASR reference tokens from the the ground truth transcriptions. Fig. 3 illustrates a diagram of the training using the standard CE loss and ASR-PBL. ASR-PBL is defined over the ASR posterior distributions as a reference to include the ASR confusion in the hidden vectors in the ST model. This decision prompts the source decoder to mimic the ASR posterior distributions given by the pre-trained ASR model through a token-level KD. The loss, which is back-propagated into the encoder, also affects our main ST task. As a result, the ST decoder learns to handle ASR confusion in its training and should become more robust against possible ASR confusion.

The ASR posterior distributions are obtained through softmax in the output layer of the source decoder. Let $P_{\text{soft}}(i, v)$ be the posterior probability of ASR token hypothesis v at the i-th position of the ASR result. ASR-PBL $\mathcal{L}_{\text{soft}}$ is defined as:

$$\mathcal{L}_{\text{soft}} = -\sum_{i=1}^{N}\sum_{v \in V}^{V} P_{\text{soft}}(i, v) \ln P_{\text{ASR}}(y_i = v). \tag{11}$$

Note that P_{ASR} is obtained from the ASR decoder in the ST model and differs from P_{soft} obtained from pre-trained ASR model. Here subscript *soft* reflects the distributional nature of ASR-PBL, as opposed to the CE loss with a *hard* one-hot objective. We used ASR-PBL with the original CE loss in

the weighted mixture and $\mathcal{L}_{\text{Att}}$ in Eq. 10 is represented as follows:

$$\mathcal{L}_{\text{Att}} = (1 - \lambda_{\text{soft}})\mathcal{L}_{\text{hard}} + \lambda_{\text{soft}}\mathcal{L}_{\text{soft}}, \quad (12)$$

where λ_{soft} is a mixture weight. Here, $\lambda_{\text{soft}} = 0.0$ denotes that the final loss function ignores ASR-PBL and becomes equivalent to the CE loss without label smoothing.

4. Experiments

The following experiments investigated the effectiveness of the proposed method using the Fisher Spanish corpus [24], which consists of approximately 140 K pairs and 160 hours of conversational Spanish speech along with its transcriptions and corresponding English translations. It was used for training both the ASR and ST models.

4.1 Data Setup

For the experiments, we filtered out long utterances with more than 3000 frames or 400 characters and preprocessed the text part with lowercasing, punctuation normalization [24], and tokenization with `tokenizer.perl` script in the Moses toolkit[†] [25] for both Spanish and English. We also extracted 80-channel log-Mel filterbank coefficients and 3-dimensional pitch features from the speech part using Kaldi [26], resulting in 83-dimensional features per frame. These features were normalized by the mean and standard deviation for each training set. We also applied speed perturbation [27] by a factor of 3 for data augmentation.

We trained a subword unigram model using SentencePiece [28] for text tokenization, with a shared subword vocabulary having a maximum of 1000 entries.

Table 1 WERs by our pre-trained ASR model and those in ESPnet document for reference: We also put WER of soft-label 1-best WER for one epoch training.

Model	Dev	Dev2	Test	Soft labels 1-best
Data size	3.9k	3.9k	3.6k	415.8k
Decoding beam size		10		1
Our model	30.2	29.1	27.2	9.3
ESPnet [†††]	24.2	23.6	21.5	-

4.2 Model Setup

The ST and pre-trained ASR models were based on Transformer [29], implemented using ESPnet[††] [30] with a single random seed of 1.

The hyperparameter settings of the model followed the defaults of ESPnet unless otherwise noted in subsequent analyses. The Transformer model consisted of an encoder with twelve layers using 2048-dimensional vectors and a decoder with six layers using 2048-dimensional vectors. The encoder and decoder employed attention mechanisms with six heads using 256-dimensional vectors.

For the ST models, we applied model-averaging with the best five models among 30 training epochs according to BLEU [31] in the Fisher development (dev) (3.9k pairs) set. We applied label smoothing for the ST task with a weight of 0.1. The minibatch size was set to 64 in the number of segments with a gradient accumulation (accumgrad) of four.

The evaluation metric for ST was a 4-reference, case-insensitive BLEU [31] on Fisher development 2 (dev2) (3.9k pairs) and the test data (3.6k pairs), given by `multi-bleu.detok.perl` in Moses. Then we used a beam search size of 10.

In this experiment, we used Hybrid CTC/Attention loss in the ASR task loss to reveal the effectiveness of ASR-PBL in each setting. Except for parameter λ_{CTC} that controlled the CTC loss weight, we followed the default implementation in ESPnet for the settings of the CTC loss calculation.

4.2.1 Baseline ST models

In this study, we used a multi-task ST with standard CE loss λ_{Att} and CTC loss λ_{CTC} in Eq. 10 as the baseline. Throughout the experiment, we empirically set λ_{CTC}=0.5 in Eq. 10 and used the CE loss with label smoothing with a weight of 0.1. This implementation is based on ESPnet. We also included the values of the multi-task ST with Hybrid CTC/Attention ASR task loss reported in ESPnet[†††], named *Transformer ASR-MTL [32]*.

4.2.2 Proposed ST models

In the proposed ST models, we used ASR-PBL with CTC loss in the ASR task of multi-task ST training. After training the ASR in the setting below, our model generated ASR posterior distributions using all of the training speech data. The ST model used the ASR posterior distributions for calculating $\mathcal{L}_{\text{soft}}$ in ASR-PBL following Eq. 11. During the ST training, $\mathcal{L}_{\text{Att}}$ was calculated in Eq. 12 by a certain λ_{soft} parameter. Our proposed model was also trained with $\lambda_{\text{CTC}} = 0.5$ in Eq. 10, like the baseline CE model. $\mathcal{L}_{\text{Att}}$ in Eq. 10 was set to the proposed ASR-PBL by mixing the hard CE loss and soft loss instead of the baseline CE loss. The proposed method is named ASR-PBL.

The pre-trained ASR model was trained using pairs of Spanish speech and transcripts in the training data. The label smoothing weight was set to 0.1. The batch size was set to 64 segments with an accumgrad of 2. We chose the best model based on the output accuracy in the development set among 30 training epochs.

By inference, we used a beam search size of 10 and applied model averaging with the best model among 30 training epochs according to WER [31] in the Fisher development set. We also evaluated the ASR outputs by WER.

[†] `https://github.com/moses-smt/mosesdecoder`

[††] `https://github.com/espnet/espnet`

[†††] `https://github.com/espnet/espnet/blob/master/egs/fisher_callhome_spanish/asr1b/RESULTS.md`

Table 2 BLEU results on Fisher with hyperparameters λ_{ASR} and λ_{soft} resulting in best development score

Model				BLEU		
Task	ASR task loss	λ_{ASR}	λ_{soft}	Dev	Dev2	Test
Single-task ST	-	-	-	41.10	41.61	40.66
Multi-task ST	Transformer ASR-MTL [32]	-	-	46.64	47.64	46.45
	w/o CE (λ_{CTC}=1.0)	0.3	-	45.98	47.16	45.83
	CE (λ_{CTC}=0.5)	0.5	-	47.18	47.43	46.59
	ASR-PBL (λ_{CTC}=0.5)	0.3	0.7	47.20	**48.36**	**46.82**

Table 1 compares the performance of the pre-trained ASR model in WER with the scores in the ESPnet document[†††] for reference. Our model's performance was worse than the reference, probably due to the difference in the loss function. The reference scores were from a model trained using CTC-based loss [22]. For the output of the ASR posterior distribution of ASR-PBL, we used a greedy search. Table 1 also shows the WER of the 1-best ASR results derived from posterior of distributions for the training of the proposed method in 1 epoch. In Eq. 12, the hyperparameter values for λ_{ASR} and λ_{soft} show the best ones in the Fisher development, chosen among $\{0.3, 0.4, 0.5\}$ for λ_{ASR} and $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ for λ_{soft}.

4.3 Main Results

Table 2 shows the main BLEU results on the Fisher dev2 and the test in each multi-task learning method. The BLEU score of the re-implemented baseline ST model introducing the Hybrid CTC/Attention in the ASR task loss improved more than in the ASR task loss only with CTC or the reported Transformer ASR-MTL score in ESPnet [32]. In the ASR task loss only with CTC, BLEU was best with a small λ_{ASR} of 0.3 among $\{0.3, 0.4, 0.5\}$. In the ASR task loss only with CTC, BLEU was best with a large λ_{ASR} of 0.5 among $\{0.3, 0.4, 0.5\}$. On the other hand, for the Hybrid CTC/Attention ASR task loss, BLEU was best with a small λ_{ASR} of 0.3 among $\{0.3, 0.4, 0.5\}$. Our proposed method outperformed all other baseline methods. In our proposed method with ASR-PBL, BLEU was best with a small λ_{ASR} of 0.3 among $\{0.3, 0.4, 0.5\}$ and a large λ_{soft} of 0.7 among $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$.

4.4 Detailed Results

Following the main results in Section 4.3, we conducted the following analyses to investigate the effectiveness of our proposed method in more detail.

1. We observed of BLEU with the baseline and proposed methods across various WER ranges to analyze the influence of ASR difficulty on the ST model.
2. We applied time-masking to the input speech frames of each utterance while training the ST model with proposed method, aiming to emulate variations in the pre-trained ASR performance.
3. We examined the performance fluctuations in both the ST and ASR task outputs by varying the λ_{soft} values.

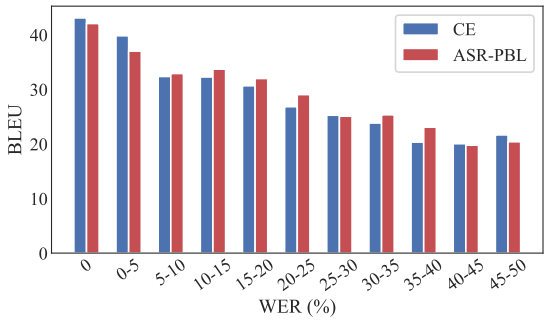


Fig. 4 BLEU comparisons in different WERs between baseline and proposed methods in Fisher test.

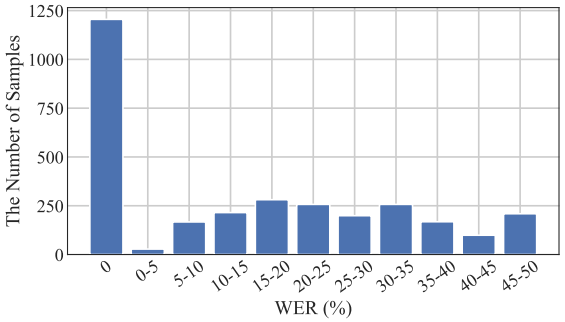


Fig. 5 Number of samples in different WERs in Fisher test.

4. We compared the output examples from both the baseline and proposed methods.

4.4.1 BLEU differences in various WER ranges in pre-trained ASR

To investigate the impact of ASR difficulty on the BLEUs, we compared the BLEUs of the baseline and proposed models in different ASR WER ranges. Fig. 4 shows the BLEUs of each ASR WER range in the baseline and proposed methods in the Fisher test. Fig. 5 shows the number of samples of each ASR WER range in the Fisher test. We calculated the WERs based on the pre-trained ASR in Section 4.2.2 and divided them into different ranges. The BLEU in each range was calculated as a corpus BLEU of SacreBLEU [33].

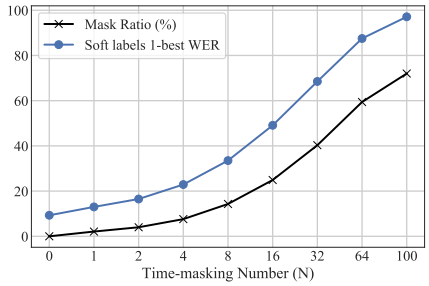


Fig. 6 Relations among N, mask ratio, and soft labels 1-best WERs: N represents maximum number of masked frames. Mask Ratio represents how many speech frames were masked.

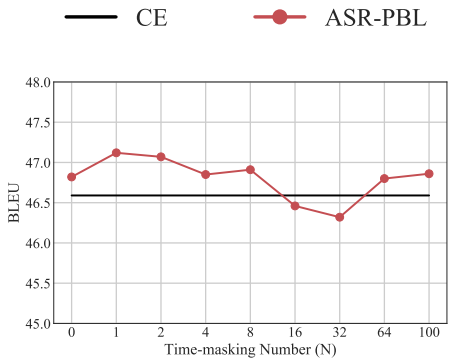


Fig. 7 Relations between each time mask number (N) and BLEU in $\lambda_{\text{soft}} = 0.7$ on Fisher test

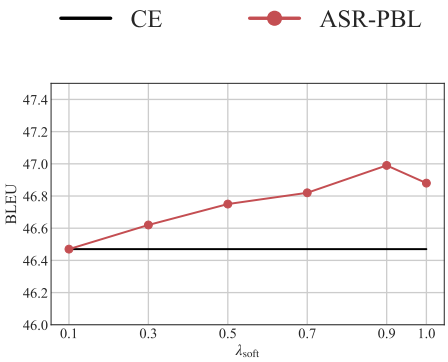


Fig. 8 BLEU results on Fisher test with different λ_{soft}

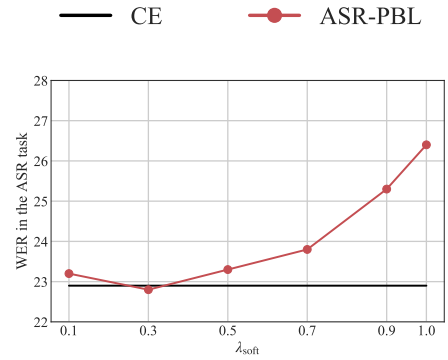


Fig. 9 WER results in ASR task on Fisher test with different λ_{soft}

In Figs. 4 and 5, we excluded the samples whose WERs exceeded 50% because they did not appear frequently. In most of the ranges from 5% to 40%, we found that the proposed method got better scores than the baseline. Fig. 4 shows that the proposed method was inferior to the baseline in the range of 0% and 0% to 5%. The number of samples was very small at the range of 0% to 5% as shown in Fig. 5. Those samples contain few ASR errors.

This result indicates that the proposed method was effective in ranges from about 5% to 40%, although no effect was observed in situations where the WER is small or large. Our result also shows the proposed method was inferior to the baseline in situations where ASR error was very small. The proposed method was effective when the ASR error occurred to some extent; while it was ineffective when the ASR error was low.

4.4.2 Performance effect of pre-trained ASR for ASR-PBL

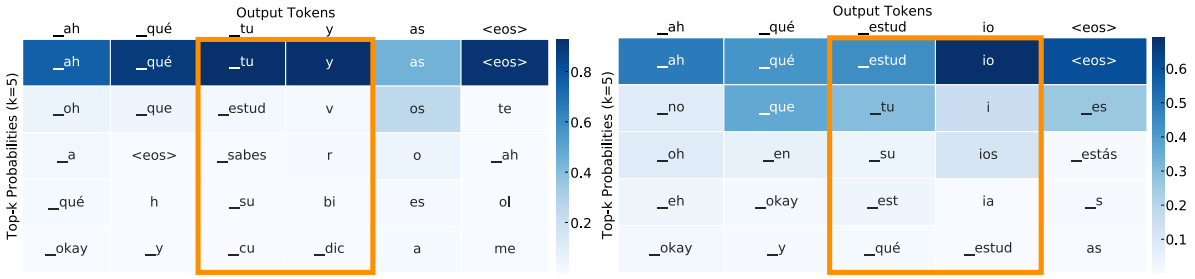
We conducted further analyses on ASR-PBL based on an intuition: *The constraints derived by the soft loss are probably affected by the performance of the pre-trained ASR model*. To simulate the differences in the ASR performance while training the models, we applied perturbation by time-masking over the input speech frames of each utterance in the ASR inference using a pre-trained ASR model to obtain posterior distributions.

We adopted the time-masking module of SpecAugment [34]. We set max mask window size W to 40 and tried different max numbers of masks N among {1, 2, 4, 8, 16, 32, 64, 100}. In other words, we randomly chose the window size of a mask from 1 to 40 every time, and the number of masks was chosen randomly from 1 to N (e.g., 100) every time. The masked parts were replaced with the average feature values over the utterances and fixed throughout the training. Note that we did not apply this approach for the one-hot results used in $\mathcal{L}_{\text{hard}}$ in Eq. 12 or $\mathcal{L}_{\text{CTC}}$ in Eq. 10.

Fig. 6 shows the WER on the soft labels for training ASR-PBL. When N grew, the mask ratio and the WER of the one-hot soft labels increased. Fig. 7 shows the relation between each time mask number (N) and BLEU in $\lambda_{\text{soft}} = 0.7$ on the Fisher test set. Overall, BLEU gradually decreased with an increase in N. These results show that the performance degradation in our proposed method was suppressed by the *hard* loss in Eq. 12 and the CTC loss in Eq. 10, not only using *soft* loss as in Eq. 12. In Fig. 7, one important finding here is that the BLEU score with $N = 32$ was just about 0.5 point less than when using no mask in $N = 0$. These results suggest that the mixture of the CE and CTC losses dominated the training and suppressed the negative effects of ASR-PBL.

Table 3 Examples from Fisher test by baseline CE and proposed ASR-PBL

	ASR output	ST output
Example 1		
Reference	la música *alegre* y romántica siempre que hable del amor positivo	*joyful* music and romantic if talks about positive love
CE (Baseline)	la música ley de eh y romántica siempre que hable del amor positivo	the music law eh romantic music always talks about love positive
ASR-PBL (Proposed)	la música **alegir** eh i romántica siempre que hable del amor positivo	**joyful** music and romantic if talks about positive love
Example 2		
Reference	ah qué *estudias*	oh what do you *study*
CE (Baseline)	ah qué tuyas	ah what are you doing
ASR-PBL (Proposed)	ah qué **estudio**	oh what do you **study**



(a) CE ASR task (Baseline) (b) ASR-PBL ASR task (Proposed)

Fig. 10 Top-5 ASR task posterior probabilities: Sentences are from Example 2 in Table 3. First raw means 1-best output tokens.

4.4.3 BLEUs in the ST task and WERs in the ASR task varying λ_{soft}

As shown in Eq. 12, our proposed method is based on adjusting the contributions of both the hard and soft losses. We examined the performance by varying their weights. The line charts in Fig. 8 show the BLEU results on the Fisher test by changing λ_{soft}. The proposed method outperformed the baseline with CE in most cases. BLEU was best when $\lambda_{\text{soft}} = 0.7$ in the Fisher development, and BLEU was best when $\lambda_{\text{soft}} = 0.9$ in the Fisher test. We observed high BLEUs in high λ_{soft}.

Fig. 9 compares WER on the Fisher test obtained by the ASR decoder. We evaluated WER on the output generated from the ASR decoder to directly examine the effect on the ASR-PBL and ASR posterior distributions. In general cascade ST systems, the BLEU result would be better when the ASR model's WER is small because the error propagation is alleviated. In contrast, Figs. 8 and 9 show that a low WER in the ASR task did not result in a high BLEU in the ST task. It suggests that minimizing WER does not always increase the BLEU scores in a multi-task ST. It also supports the results in which our proposed method outperformed the baseline even if the WER in the ASR task is higher than that of the baseline.

4.4.4 Output examples

Table 3 shows the translation examples of the Fisher test. In Example 1, the baseline CE model mistakenly derived *ley* in its ASR output. As a result, the ST task also mistakenly derived *law*, which is the counterpart of *ley* while missing correct ST output *joyful*, which is the counterpart of *alegre*. On the other hand, the proposed method also chose in its 1-best ASR output a different word, *alegir*. However, its pronunciation resembles *alegre*. Finally, the proposed ST task derived appropriate output: *joyful*.

In Example 2, the baseline CE model predicted *tuyas* rather than *estudias* in the reference by its ASR decoder and missed *study* corresponding to *estudias*. The proposed method also mistakenly chose *estudio* in the ASR subtask. However, *estudio* would be a more suitable choice due to its closer pronunciation and semantic similarity, unlike the baseline's choice of *tuyas*. As a result, the proposed ST task generated an appropriate word: *study*.

We next investigated Example 2 in detail. Fig. 10 shows the top-5 ASR token-level hypotheses by the baseline CE and proposed ASR-PBL models, which are colored by their posterior probabilities. In the probability distribution of the baseline in Fig. 10(a), high probability was assigned to the top ranked token (e.g., *_tu* and *y*). This means that the ST model ignores a possible ASR hypothesis of *estudias* in the ASR task. In contrast, the posterior distribution of *_tu* is large after *_estud*, indicating that not only the posterior probability of *_estud* but also that of *_tu* is large in the ASR task posterior probabilities from the proposed method like Fig. 10(b). Although *_tu* is part of the mistaken word *tuyas* in the baseline, it also considered the possibility of that *_estud*, which is part of *estudias*, is the most likely prediction and translated the

result as *study*. This result suggests that our proposed ST model predicted the output based on the acoustic confusion among *estudio*, *tuyas* and *estudias*. These examples show that our proposed method generated correct ST outputs while considering other ASR hypotheses in the ASR subtask.

5. Conclusions

We proposed a method to train an end-to-end ST model using ASR-PBL. It basically keeps ASR confusion and improves the robustness of an ST model by knowledge distillation from a pre-trained ASR model.

Our experimental results demonstrate effectiveness against baselines with standard CE loss. Further analyses with perturbations on ASR showed that the mixed use of standard CE loss and ASR-PBL worked effectively, regardless of the performance of the pre-trained ASR model. Future work includes the application of the proposed method to simultaneous ST and in such challenging language pairs as English-Japanese.

Acknowledgments

Part of this work was supported by JST SPRING Grant Number JPMJSP2140 and JSPS KAKENHI Grant Numbers JP21H05054 and JP21H03467.

An earlier version of this work was presented in another paper [35]. In this paper, we expanded the previous work's contents and conducted an additional experiment to reveal the effectiveness of our proposed method when applied to CTC loss. We also conducted further analyses on the impact of ASR difficulty and the effectiveness of the pre-trained ASR model in our proposed method.

References

[1] H. Ney, "Speech translation: coupling of recognition and translation," Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99, Phoenix, Arizona, USA, March 15-19, 1999, pp.517–520, IEEE Computer Society, 1999.

[2] F. Casacuberta, M. Federico, H. Ney, and E. Vidal, "Recent efforts in spoken language translation," IEEE Signal Process. Mag., vol.25, no.3, pp.80–88, 2008.

[3] G. Kumar, M. Post, D. Povey, and S. Khudanpur, "Some insights from translating conversational telephone speech," IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014, pp.3231–3235, IEEE, 2014.

[4] M. Sperber, G. Neubig, N. Pham, and A. Waibel, "Self-Attentional Models for Lattice Inputs," Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, ed. A. Korhonen, D.R. Traum, and L. Màrquez, pp.1185–1197, Association for Computational Linguistics, 2019.

[5] A. Berard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," CoRR, vol.abs/1612.01744, 2016.

[6] R.J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017, ed. F. Lacerda, pp.2625–2629, ISCA, 2017.

[7] J. Niehues, R. Cattoni, S. Stüker, M. Negri, M. Turchi, T. Ha, E. Salesky, R. Sanabria, L. Barrault, L. Specia, and M. Federico, "The IWSLT 2019 evaluation campaign," Proceedings of the 16th International Conference on Spoken Language Translation, IWSLT 2019, Hong Kong, November 2-3, 2019, ed. J. Niehues, R. Cattoni, S. Stüker, M. Negri, M. Turchi, T. Ha, E. Salesky, R. Sanabria, L. Barrault, L. Specia, and M. Federico, Association for Computational Linguistics, 2019.

[8] A. Anastasopoulos and D. Chiang, "Tied Multitask Learning for Neural Speech Translation," Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), ed. M.A. Walker, H. Ji, and A. Stent, pp.82–91, Association for Computational Linguistics, 2018.

[9] Y. Jia, R.J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model," Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019, ed. G. Kubin and Z. Kacic, pp.1123–1127, ISCA, 2019.

[10] T. Kano, S. Sakti, and S. Nakamura, "End-to-End Speech Translation With Transcoding by Multi-Task Learning for Distant Language Pairs," IEEE ACM Trans. Audio Speech Lang. Process., vol.28, pp.1342–1355, 2020.

[11] T. Kano, S. Sakti, and S. Nakamura, "Transformer-Based Direct Speech-To-Speech Translation with Transcoder," 2021 IEEE Spoken Language Technology Workshop (SLT), pp.958–965, 2021.

[12] K. Osamura, T. Kano, S. Sakti, K. Sudoh, and S. Nakamura, "Using Spoken Word Posterior Features in Neural Machine Translation," Proceedings of the 15th International Workshop on Spoken Language Translation, 181-188, Oct. 2018, vol.21, p.22, 2018.

[13] P. Bahar, T. Bieschke, R. Schlüter, and H. Ney, "Tight integrated end-to-end training for cascaded speech translation," IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021, pp.950–957, IEEE, 2021.

[14] S. Dalmia, B. Yan, V. Raunak, F. Metze, and S. Watanabe, "Searchable hidden intermediates for end-to-end models of decomposable sequence tasks," Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, ed. K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, pp.1882–1896, Association for Computational Linguistics, 2021.

[15] S. Chuang, T. Sung, A.H. Liu, and H. Lee, "Worse WER, but Better BLEU? Leveraging Word Embedding as Intermediate in Multitask End-to-End Speech Translation," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, ed. D. Jurafsky, J. Chai, N. Schluter, and J.R. Tetreault, pp.5998–6003, Association for Computational Linguistics, 2020.

[16] G.E. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," CoRR, vol.abs/1503.02531, 2015.

[17] Y. Kim and A.M. Rush, "Sequence-level knowledge distillation," Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, ed. J. Su, X. Carreras, and K. Duh, pp.1317–1327, The Association for Computational Linguistics, 2016.

[18] Y. Liu, H. Xiong, J. Zhang, Z. He, H. Wu, H. Wang, and C. Zong, "End-to-End Speech Translation with Knowledge Distillation," Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019, ed. G. Kubin and Z. Kacic, pp.1128–1132, ISCA, 2019.

[19] M. Gaido, M.A.D. Gangi, M. Negri, and M. Turchi, "End-to-End Speech-Translation with Knowledge Distillation:

FBK@IWSLT2020," Proceedings of the 17th International Conference on Spoken Language Translation, IWSLT 2020, Online, July 9 - 10, 2020, ed. M. Federico, A. Waibel, K. Knight, S. Nakamura, H. Ney, J. Niehues, S. Stüker, D. Wu, J. Mariani, and F. Yvon, pp.80–88, Association for Computational Linguistics, 2020.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp.2818–2826, IEEE Computer Society, 2016.

[21] R. Müller, S. Kornblith, and G.E. Hinton, "When does label smoothing help?," Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, ed. H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox, and R. Garnett, pp.4696–4705, 2019.

[22] A. Graves, S. Fernández, F.J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006, ed. W.W. Cohen and A.W. Moore, ACM International Conference Proceeding Series, vol.148, pp.369–376, ACM, 2006.

[23] S. Watanabe, T. Hori, S. Kim, J.R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," IEEE Journal of Selected Topics in Signal Processing, vol.11, no.8, pp.1240–1253, 2017.

[24] M. Post, G. Kumar, A. Lopez, D.G. Karakos, C. Callison-Burch, and S. Khudanpur, "Improved speech-to-text translation with the fisher and callhome spanish-english speech translation corpus," Proceedings of the 10th International Workshop on Spoken Language Translation: Papers, Heidelberg, Germany, December 5-6, 2013, 2013.

[25] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic, ed. J.A. Carroll, A. van den Bosch, and A. Zaenen, The Association for Computational Linguistics, 2007.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The Kaldi speech recognition toolkit," IEEE 2011 workshop on automatic speech recognition and understanding, IEEE Signal Processing Society, 2011.

[27] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015, pp.3586–3589, ISCA, 2015.

[28] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp.66–71, 2018.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," Advances in neural information processing systems, pp.5998–6008, 2017.

[30] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.E.Y. Soplin, J. Heymann, M. Wiesner, N. Chen, *et al.*, "ESPnet: End-to-End Speech Processing Toolkit," Proc. Interspeech 2018, pp.2207–2211, 2018.

[31] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, "BLEU: a method for automatic evaluation of machine translation," Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp.311–318, 2002.

[32] H. Inaguma, S. Kiyono, K. Duh, S. Karita, N. Yalta, T. Hayashi, and S. Watanabe, "ESPnet-ST: All-in-One Speech Translation Toolkit," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020, ed. A. Celikyilmaz and T. Wen, pp.302–311, Association for Computational Linguistics, 2020.

[33] M. Post, "A call for clarity in reporting BLEU scores," Proceedings of the Third Conference on Machine Translation: Research Papers, Belgium, Brussels, pp.186–191, Association for Computational Linguistics, Oct. 2018.

[34] D.S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E.D. Cubuk, and Q.V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019, ed. G. Kubin and Z. Kacic, pp.2613–2617, ISCA, 2019.

[35] Y. Ko, K. Sudoh, S. Sakti, and S. Nakamura, "ASR Posterior-based Loss for Multi-task End-to-end Speech Translation," Proc. Interspeech 2021, pp.2272–2276, 2021.

Yuka Ko Yuka Ko received a B.S. degree in information engineering from Osaka Prefecture University in 2020 and a M.Eng. degree in 2022 from the Nara Institute of Science and Technology (NAIST), Japan. She is now a Ph.D. Student in the Augmented Human Communication Laboratory, NAIST. She is a recipient of the JSPS Research Fellowship for Young Scientist DC2. Her research interests include speech translation, machine translation, and spoken language processing.

Katsuhito Sudoh Katsuhito Sudoh is a professor at Nara Women's University. He received a bachelor's degree in engineering in 2000 and master's and Ph.D. degrees in informatics in 2002 and 2015 from Kyoto University. He was in NTT Communication Science Laboratories from 2002 to 2017 and Nara Institute of Science and Technology from 2017 to 2024. He currently works on spoken language processing and natural language processing. He is a member of ACL, ISCA, ANLP, ASJ, IPSJ, and JSAI.

Sakriani Sakti Sakriani Sakti (Member, IEEE) received the B.E. degree in informatics (cum laude) from the Bandung Institute of Technology, Bandung, Indonesia, in 1999, and the M.Sc. and the Ph.D. degrees from the University of Ulm, Ulm, Germany, in 2002 and 2008, respectively. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003 and 2009, she was a Researcher with ATR SLC Labs, Japan, and from 2006 to 2011, she was an Expert Researcher with NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her studies from 2005 to 2008 with Dialog Systems Group, University of Ulm. She was actively involved in collaboration activities, such as Asian Pacific Telecommunity Project from 2003 to 2007, A-STAR, and USTAR from 2006 to 2011. From 2009 to 2011, she was a Visiting Professor with the Computer Science Department, University of Indonesia (UI), Indonesia. From 2011 to 2017, she was an Assistant Professor with the Augmented Human Communication Laboratory, NAIST, Japan. She was a Visiting Scientific Researcher of INRIA Paris-Rocquencourt, France, from 2015 to 2016, under JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation. From 2018 to 2021, she was a Research Associate Professor with NAIST and a Research Scientist with RIKEN, Center for Advanced Intelligent Project (AIP), Japan. From 2021 to 2024, she was an Associate Professor with JAIST, Adjunct Associate Professor with NAIST, Visiting Research Scientist with RIKEN AIP, and Adjunct Professor with the University of Indonesia. She is currently a full Professor at NAIST, Adjunct Professor with JAIST, Visiting Research Scientist with RIKEN AIP, and Adjunct Professor with the University of Indonesia. Her research interests include statistical pattern recognition, graphical modeling framework, deep learning, multilingual speech recognition and synthesis, spoken language translation, affective dialog system, and cognitive-communication. In 2000, she was the recipient of the DAAD Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm. She is a Member of JNS, SFN, ASJ, ISCA, and IEICE. She is also a Committee Member of IEEE SLTC From 2021 to 2023 and an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING from 2020 to 2023. Furthermore, she is the Chair of ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL) and a Board Member of Spoken Language Technologies for Under-Resourced Languages (SLTU).

Satoshi Nakamura Dr Nakamura is Emeritus and Specially Appointed Prof. at the Nara Institute of Science and Technology (NAIST), Japan, Full Prof. at the Chinese University of Hong Kong, Shenzhen (CUHKSZ), and Honorary Prof. at the Karlsruhe Institute of Technology (KIT), Germany. He received his B.S. from Kyoto Institute of Technology in 1981 and his Ph.D. from Kyoto University in 1992. He was an associate professor at NAIST from 1994 to 2000, department head and director of the ATR Spoken Language Communication Research Laboratories from 2000 to 2008, and vice president of ATR from 2007 to 2008. He was Director General of the Keihanna Research Laboratories and Executive Director of the Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology (NICT), Japan, from 2009 to 2010, and Project Leader of the Tourism Information Analytics Team at the Center for Advanced Intelligence Project of RIKEN Institute from 2017 to 2021. He was director of the Augmented Human Communication Laboratory and full professor at NAIST from 2011 to 2024. He is currently a full professor at CUHKSZ. His interests include modelling and systems of speech-to-speech translation, speech recognition and spoken language processing. He is one of the leaders in speech-to-speech translation research and has been served in various worldwide speech-to-speech translation research projects, including C-STAR, IWSLT, A-STAR, and ISCA SIG on Spoken Language Translation. He was awarded the LREC Antonio Zampolli Prize in 2012. He has also received the Yamashita Research Award, the Kiyasu Award from the Information Processing Society of Japan, the Telecom System Award, the AAMT Nagao Award, the Docomo Mobile Science Award in 2007, the ASJ Award for Distinguished Achievements in Acoustics, the Commendation for Science and Technology from the Minister of Education, Science and Technology, and the Commendation for Science and Technology from the Minister of Internal Affairs and Communications. He was an elected member of the Board of Directors of the International Speech Communication Association, ISCA in 2011-2018, a member of the Editorial Board of the IEEE Signal Processing Magazine in 2012-2014, and a member of the IEEE SPS Speech and Language Technical Committee in 2013-2015. He is an IPSJ Fellow, an ATR Fellow, an IEEE Fellow, and an ISCA Fellow. He is a senior member of the Institute of Electronics, Information and Communication Engineers (IEICE).