ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746699

INTEGRATING MULTIPLE ASR SYSTEMS INTO NLP BACKEND WITH ATTENTION FUSION

Takatomo Kano[1], *Atsunori Ogawa*[1], *Marc Delcroix*[1], *and Shinji Watanabe*[2]

[1]NTT Corporation, Japan
[2]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA

ABSTRACT

Spoken language processing (SLP) systems such as speech summarization and translation can be achieved by cascade models. It combines an automatic speech recognition (ASR) frontend and a natural language processing (NLP) backend including machine translation (MT) or text summarization (TS). With this cascade approach, we can exploit large non-paired datasets to independently train state-of-the-art models for each module. However, ASR errors directly affect the performance of the NLP backend in the cascade approach. In this paper, we reduce the impact of ASR errors on the NLP backend by combining transcriptions from various ASR systems. Recognizer output voting error reduction (ROVER) is a widely used technique for system combination. Although ROVER improves ASR performance, the combination process is not optimized for backend tasks. We propose a system combination that resembles ROVER using attention fusion to achieve the alignment and the combination of multiple ASR hypotheses. This allows the combination process to be optimized for the backend NLP task without changing the ASR frontend. Our proposed technique is general and can be applied to various SLP tasks. We confirm its effectiveness on both speech summarization and translation experiments.

Index Terms— Speech summarization, Speech translation, Automatic speech recognition, Attention fusion, ROVER.

1. INTRODUCTION

Many spoken language processing (SLP) systems, such as speech summarization and speech translation are achieved by cascade models. It combines two main sub-modules: an automatic speech recognition (ASR) frontend and a natural language processing (NLP) backend. This cascade connection has two benefits. First, we can lerverage state-of-the-art models for each individual sub-module. Second, it does not require large paired data composed of speech data and the associated target text to train the SLP system. However, the NLP module cannot avoid ASR error propagation that affect the SLP system's performance [1].

Many SLP studies mitigate the influence of ASR error propagation by extending and tuning the NLP backend. For example, speech translation studies [2–6] mitigate ASR errors by considering multiple ASR hypotheses and auxiliary information at the input of the NLP backend. For speech summarization, we recently proposed an attention fusion mechanism to fuse multiple ASR hypotheses into Bidirectional Encoder Representations from Transformers (BERT) [7]. Thus, the NLP backend reduces the affection of ASR errors by integrating the multiple ASR hypotheses with considering the naturalness of sentence. However, these studies only consider the multiple hypotheses extracted from the n-best list of a single ASR system. Consequently, there is little variation among the hypotheses. Moreover, these approaches require the sharing of a common vocabulary between the ASR decoder and the NLP backend encoder, to prevent building optimal ASR or NLP backend sub-modules.

Many ASR studies improve the ASR performance by combining hypotheses from multiple ASR systems [8–11]. These ASR systems combination approaches can estimate and recover the ASR errors by comparing multiple ASR hypotheses. In this paper, we focus on the most commonly used approach recognizer output voting error reduction (ROVER) [10]. It consists of alignment and voting steps to combine ASR hypotheses. ROVER is widely used because it usually improves the ASR performance [12–16]. It does not require a special setting for ASR frontend and NLP backend systems since it can operate on word sequences. However, ROVER has some limitations. First, it does not consider word similarities and naturalness of a sentence during the alignment and voting processes. Moreover, since it does not have trainable parameters, it cannot be tuned for a specific SLP task or focus on important words. For SLP systems, good ASR transcription does not always lead to good NLP performance [3, 17]. For example, an ideal ASR system would transcribe all words even fillers and stutters. However, for NLP, fillers and stutters are unnecessary, and the importance of each word is not equal. The named entities and conjunctions are essential in a summarization task. Consequently, ROVER may not always provide an optimal input for the NLP backend since it handles all words equally.

In this paper, we propose an alternative to ROVER to integrate multiple ASR systems into an NLP backend and introduce an attention fusion mechanism within an NLP backend. Thus, our proposed method can integrate transcriptions from multiple ASR systems considering word similarities and sentence naturalness based on the hidden representation of the NLP backend. The attention fusion has trainable parameters which can be tuned for a specific task. Our proposed attention fusion is derived from our recent work [7], and resembles hierarchical attention [18, 19] proposed for multi-stream combinations [20] and audio-visual processing [21]. We treat the multiple ASR hypotheses as a multi-stream of hierarchical attention and achieve the alignment and combination processes for ASR hypotheses using an attention mechanism. Moreover, we investigate attention fusion for the combination of ASR hypotheses from different systems, i.e., unaligned hypotheses with different lengths.

We compared our proposed method with ROVER and related approaches that combine hypotheses from a single ASR system [1,5,7]. We experimentally confirmed its effectiveness on summarization and translation tasks, which are tasks that have very different characteristics. For example, the translation model needs to learn such complex mapping as English to German or English to Portuguese. However, it does not need to consider the long context of source sentence. Although a summarization model does not process multiple languages, it does consider the long context of a whole document. Our proposed method combines multiple-ASR hypotheses within the NLP backend by the attention fusion and retrains the NLP backend such as a BERT and Transformer [22] encoder for each task. Our experimental results show that our proposed method improved the summarization and translation performance on HOW2 [23] and TED talk [24, 25] datasets.

2. MULTI-ASR INTEGRATION WITH ATTENTION FUSION

2.1. System combination

Many studies achieved multi-ASR systems combination using lattice combination [8–10] or minimum Bayes risk decoding [11]. In this paper, we focus on the most commonly used approach ROVER [10]. ROVER uses two basic steps to combine multiple ASR hypotheses: alignment and combination. The alignment step finds the best word alignment of the hypotheses relative to a reference hypothesis. In the combination step, the aligned hypotheses are combined into a single-word sequence. For example, ROVER performs alignment by building a word transition network (WTN) that allows accounting for insertion, deletion, and substitutions in the hypotheses. Then, ROVER uses dynamic programming to find the optimal alignments. In the combination step, ROVER performs a voting process, i.e., for each time step, chose the word candidate by considering the frequency of occurrence.

ROVER usually operates on word tokens and does not have any learnable parameters, which makes the approach generally applicable. However, it cannot be easily used with continuous word representations that provide more informative context information. Nor can it be optimized for a given backend task because it has no learnable parameters.

To address these issues, we propose an attention-fusion-based approach for system combination. The following subsection explains the word encodings we use and how we implement the basic alignment and combination steps.

2.2. Word encodings

We perform a system combination using a word or sub-word encoding that includes context or semantic information. We use the representation provided by the first few layers of the encoder of an NLP network as sub-word encodings $E \in \mathbb{R}^{L\times D}$:

$$E^n = g(S^n), \tag{1}$$

where S^n is the n-th ASR hypothesis, $n = 1, \dots, N$, and N denotes the number of ASR systems. Since each ASR hypothesis S^n may have a different length, we pad each S^n with a special token, i.e., [MASK] [26], so that all the hypotheses have the same length L. $g(\cdot)$ represents the first few layers of the NLP backend, and D is the hidden dimension of the NLP backend encoder.

Fig. 1 compares ROVER and our proposed attention fusion. In the figure,$h(\cdot)$ denotes the upper layers of the NLP backend after $g(\cdot)$. ROVER combines ASR hypotheses before the NLP backend, while the attention fusion combines ASR hypotheses within the NLP backend.

2.3. Attention-based hypothesis alignment

We align each transcription using an attention mechanism between each hypothesis and a reference hypothesis E^r. The attention query is $E^r = [e^r_1, \dots, e^r_L]$, which is chosen to be the hypothesis from the ASR system that achieves the best performance on a validation dataset. The key and value are E^n. The attention has trainable parameters, which consist of linear transformations $W^Q \in \mathbb{R}^{D\times D'}$, $W^K \in \mathbb{R}^{D\times D'}$, $W^V \in \mathbb{R}^{D\times D'}$ for query, key, and value. D' denotes the hidden dimension of upper layers of the NLP backend; in this paper, D equals D'. The attention function generates the following aligned encoding sequence $\tilde{E}^n$:

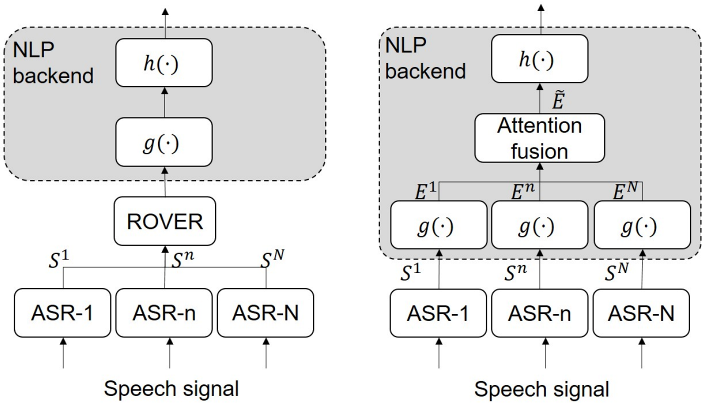

Fig. 1. Comparison of ROVER and attention fusion

$$\tilde{E}^n = \text{softmax}\left((E^r W^Q)(E^n W^K)^{\mathrm{T}}\right) E^n W^V, \tag{2}$$

where T is the transpose operation.

Compared to ROVER, the attention-based alignment step achieves soft alignment. Since the encoder captures the whole context of the input sentence through e.g. self-attention, the alignment process can exploit the forward and backward context information by performing alignment with the encoding. Then, the attention fusion can align reference and hypothesis sequences not only at the word-level but also at the phrase or sentence level.

2.4. Attention-based hypothesis combination

In the combination step, we perform attention over different ASR systems for every aligned sub-word position l in a similar way as hierarchical attention [18, 27, 28]. As introduced in Eq. (2), $\tilde{E}$ is a matrix containing the N aligned encoding sequences and $\tilde{E}_l$ the aligned encoding for the l-th sub-word position in the sequence. We can perform attention over the hypotheses to obtain a modified encoding vector e^{att}_l:

$$\alpha_l = \text{softmax}\left((e^r_l)^{\mathrm{T}} W^Q \tilde{E}_l\right), \tag{3}$$

$$e^{\text{att}}_l = \alpha_l \tilde{E}^{\mathrm{T}}_l \tag{4}$$

where $\alpha_l \in \mathbb{R}^{1\times N}$ is the attention weight over N recognition hypotheses. The difference from ROVER is that ROVER chooses a word from multiple hypotheses by voting process using frequency of occurrence, but attention fusion merges multiple hypotheses by a weighted summation using a similarity in an encoding space. Here ROVER makes hard decisions; our proposed attention-fusion performs soft decisions.

Table 1. ROUGE (R1, R2, RL) scores for different speech summarization systems: System (7) is our proposed method.

Method	TED			HOW2		
	R1	R2	RL	R1	R2	RL
(0) ASR-GT	32.1	6.2	19.0	56.5	37.8	59.3
(1) ASR w/ASR BPE	29.9	6.9	18.3	47.4	27.1	46.1
(2) ASR w/BERT BPE	28.9	6.2	17.8	45.3	26.8	45.0
(3) ROVER	29.9	5.8	19.2	48.0	27.3	47.1
(4) Retrain	31.5	5.6	20.4	47.2	27.0	45.6
(5) Confidence [1]	30.1	6.8	20.4	48.4	29.0	47.3
(6) Nbest fusion [7]	31.9	6.0	19.3	49.3	28.8	48.2
(7) System fusion	31.9	6.1	19.0	50.1	29.0	48.3

3. SPEECH SUMMARIZATION EXPERIMENT

3.1. Experimental settings

In this paper, we compare related studies [1,5,7,10] and our proposed attention fusion on speech summarization and translation tasks.We evaluated the the speech summary systems on the TEDSummary [7] and HOW2 summarization [23] tasks. We built all the SLP systems based on a cascade manner. We trained the Transformer-based ASR models using the ESPnet toolkit[1], by following the published recipes for the TEDLIUM2 [29] and HOW2 [23] tasks. We used the same ASR models for the speech summarization and translation systems and varied the BPE size of the ASR to prepare $N= 5$ ASR systems (Section 2.2) for ROVER and our proposed model (Table 2).

For the NLP backend, we built a BERT-based text summarization (TS) model using the OpenNMT toolkit[2]. We used a pre-trained BERT model provided by huggingface[3] and Transformer decoder in the same way as [28] . We used the TEDSummary [7] and HOW2 datasets to train these TS models.

We considered several baseline systems. First, **(0) ASR-GT** uses ground-truth ASR transcriptions as input to the NLP backend. **(0) ASR-GT** illustrates the upper-bound performance on speech summarization.

(1) ASR w/ASR BPE and **(2) ASR w/BERT BPE** use the ASR transcription as input to the NLP backend. System (1) uses the BPE model of ESPnet and (2) uses the BPE model of BERT. **(3) ROVER** uses ROVER output as input to the NLP backend.

(4) Retrain uses an NLP backend with ASR transcriptions. **(5) Confidence** is based on the work of Weng et al. [1] that uses fine-tuned NLP backend with a confidence score. These baselines use ASR transcriptions or transcriptions and confidence scores to fine-tune the NLP backend. However, these NLP backends cannot consider multiple ASR hypotheses.

(6) Nbest fusion uses attention fusion to combine multiple ASR hypotheses derived from a single ASR system. We create the N hypotheses as follows. First, we save the sequence of summation of output log-softmax values from the ASR and the language models for the best beam-search path. After decoding, for each step in the 1-best path, we select the N=5 tokens with the top 5 values of saved output vectors in the path [4, 5, 7].

(7) System fusion our proposed method, uses attention fusion to combine ASR hypotheses provided by multiple ASR systems.

Table 2. ASR WER for each BPE size. 30.5k* corresponds to the BPE size of the BERT model. ROVER denotes output of ROVER system for these ASR systems

	BPE size						
	best		2-nd	3-th	4-th	5-th	
	500	5k	10k	20k	30k	30.5k*	ROVER
How2	n/a	13.0	13.6	14.1	14.3	14.6	12.2
TED	8.5	n/a	8.7	9.5	10.0	10.4	8.3

Table 3. BLUE scores for different speech translation systems. System (7) is our proposed method.

	Method	TED (En-De)	HOW2 (En-Pt)
(0)	ASR-GT	27.2	55.2
(1)	ASR w/ASR BPE	24.3	44.6
(2)	ASR w/MT BPE	24.1	44.8
(3)	ROVER	24.4	45.1
(4)	Retrain	24.0	45.5
(5)	Posterior [5]	25.1	45.8
(6)	Nbest fusion [7]	25.1	45.6
(7)	System fusion	25.4	46.0

3.2. Experimental results

Table 1 shows the ROUGE scores [30] for the baseline systems ((0-6)) and our proposed system (7) with attention fusion[4].

Our proposed system (7) achieved superior ROUGE scores than the ROVER-based combination and conventional cascade-based systems. Moreover, it achieved better scores than systems (4-6) except for ROUGE-2 (R2) and ROUGE-L (RL) on the TED task. For the TED task, note that due to the data themselves and the overall better ASR performance, systems (4-6) achieved performance close to the system (0), which leaves little margin for improvement. On the other hand, on the HOW2 task, there is more room for improvement, and our proposed system (7) provides more substantial performance gains over the other baseline systems (0-6).

4. SPEECH TRANSLATION EXPERIMENT

4.1. Experimental settings

Next we built a Transformer-based machine translation (MT) [22] model using OpenNMT. For English to German translation, we used IWSLT21 dataset [25] to train the SLP model and test by the "dev" dataset. For English to Portuguese translation, we used the HOW2 dataset to train and test the SLP model.

We built baseline systems that are identical as those in Section 3.1. We changed system (2) to **(2) ASR w/MT BPE** uses the BPE model of MT (8k sub-words for HOW2 and 32k sub-words for TED), and system (5) to **(5) Posterior** based on the work of Bahar et al. [5]. System (5) inputs a posterior vector instead of a one-hot vector to the NLP backend. The dimension of the posterior vector is the

[1] https://github.com/espnet/espnet

[2] https://github.com/OpenNMT/OpenNMT-py

[3] https://huggingface.co/transformers/

[4] Note that we confirmed the validity of our implementation of BERTSum since it achieved a similar level of performance on the HOW2 corpus [31], which reportedROUGE-1 (R1) and ROUGE-L (RL) scores of 48.3 and 44.0, although the systems cannot be directly compared because of differences in the training data and the ASR frontend.

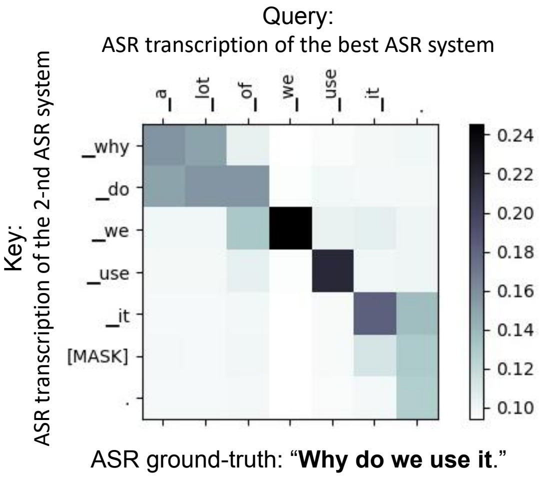

Fig. 2. Alignment result of attention fusion:
Vertical axis is an ASR transcription of the best ASR system.
Horizontal axis is an ASR transcription of the 2-nd ASR system.

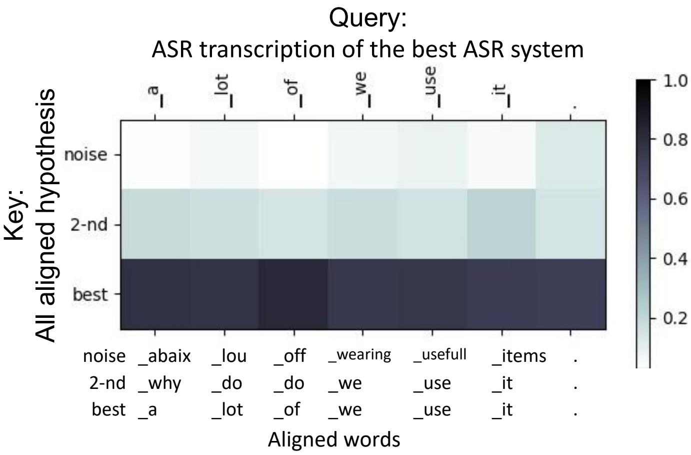

Fig. 3. Integration result of Attention fusion:
Vertical axis "best" and "2-nd" denote the aligned sequences of the best and 2-nd ASR systems, and "noise" denotes random sub-word sequence.

same as the sub-word vocabulary size, and satisfies the sum-to-one condition. The MT model combines all sub-words by a weighted sum using the posterior probability at the NLP backend embedding layer.

4.2. Experimental results

Table 3 shows the BLUE scores [32] for the baseline systems (0-6) and our proposed method with **(7) System fusion**. Our proposed system (7) outperforms ROVER-based combination and conventional cascade-based systems in the translation task.

Furthermore, the posterior method (system (5)) outperforms Nbest fusion (system (6)) on the HOW2 dataset. System (5) merges hypothesis sub-words at an NLP backend embedding layer considering the posterior probabilities. Thus, it does not consider the context of the input sequence. System (6) merges hypothesis sub-words inside the NLP backend encoder considering the naturalness and the context of the input sequence. However, this does not seem to bring performance gains for translation tasks. Our proposed system (7) achieves the best BLUE score probably because it aligns different lengths of ASR hypotheses and combines multiple hypotheses in an optimal manner for the translation.

4.3. Visualization

For a further analysis, we plot the alignment result of our attention fusion (Section 2.3) in Fig. 2. The vertical axis is a query E^r (ASR transcription of the best ASR system) and the horizontal axis is a key E^n (the ASR transcription of the 2-nd ASR system (Table 2)) of Section 2.3. From this result, the attention fusion has an ability to consider alignment at the phrase level. The phrase "a lot of" is aligned to the phrase "why do." Then the attention map shows that identical words in the sequences are aligned to each others, and the corresponding phrase is correctly aligned.

Next, we plot the combination results of the attention fusion (Section 2.4) in Fig. 3. "best" denotes the aligned sequence of the best ASR system (which here is not the most accurate hypothesis), "2-nd" denotes an aligned sequence of the 2-nd ASR system, and "noise" denotes a random sub-word sequence that we artificially added to confirm the attention fusion behavior. From this result, we confirm that the attention fusion does not focus on the noise sequence; it can ignore that sequence. Second, in this case, the "2-nd" sequence is the same as ASR ground-truth, and the attention fusion combines the "best" and the "2-nd" sequences.

Table 4. Translation examples of each ASR hypothesis.

Reference	Porque usamos isso ?
System (4) w/best	Muito disso usamos isso .
System (4) w/2-nd	Porque usamos isso ?
System (4) w/noise	Baixo de usar itens úteis .
System (7) w/all	Porque usamos isso ?

Finally, Table 4 shows the translation results.**Reference** is a ground-truth translation.**System (4) w/best**, **System (4) w/2-nd**, **System (4) w/noise** denote the **(4) Retrain** translation results obtained with each hypothesis. **System (7) w/all** denotes our proposed **(7) System fusion** translation result with combining all hypotheses shown in Fig. 3. The result shows that our proposed method correctly translates source to the target language and avoids the ASR error propagation.

5. CONCLUSION

We proposed attention fusion to integrate multiple ASR systems into the NLP backend. It mitigates the impact of ASR error propagation on speech translation and summarization by aligning and combining different lengths of ASR hypotheses inside the NLP backend. Future works will investigate a tighter interconnection of the ASR frontend and the NLP backend to improves SLP system's performance, exploit such speech-specific information as intonation, and create richer and more informative translations and summaries.

6. REFERENCES

[1] Shi-Yan Weng, Tien-Hong Lo, and Berlin Chen, "An effective contextual language modeling framework for speech summarization with augmented features," in *EUSIPCO*, 2020, pp. 316–320.

[2] Nicola Bertoldi, Richard Zens, and Marcello Federico, "Speech translation by confusion network decoding," in *ICASSP*, 2007, pp. 1297–1300.

[3] Masaya Ohgushi, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura, "An empirical comparison of joint optimization techniques for speech translation," in *INTERSPEECH*, 2013, pp. 2619–2623.

[4] Kaho Osamura, Takatomo Kano, Sakti Sakriani, Katsuhito Sudoh, and Satoshi Nakamura, "Using spoken word posterior features in neural machine translation," in *IWSLT*, 2018.

[5] Parnia Bahar, Tobias Bieschke, Ralf Schlüter, and Hermann Ney, "Tight integrated end-to-end training for cascaded speech translation," in *SLT*, 2021, pp. 950–957.

[6] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel, "Neural lattice-to-sequence models for uncertain inputs," in *IWSLT*, 2017, pp. 1380–1389.

[7] Takatomo Kano, Atsunori Ogawa, Marc Delcroix, and Shinji Watanabe, "Attention-based multi-hypothesis fusion for speech summarization," in *ASRU*, 2021, p. (to appear).

[8] Gunnar Evermann and Woodland Philip, "Posterior probability decoding, confidence estimation and system combination," in *Speech Transcription Workshop*, 2000, vol. 27, pp. 78–81.

[9] Björn Hoffmeister, Dustin Hillard, Stefan Hahn, Ralf Schlüter, Mari Ostendorf, and Hermann Ney, "Cross-site and intra-site ASR system combination: Comparisons on lattice and 1-best methods," in *ICASSP*, 2007, pp. 1145–1148.

[10] Jonathan Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings.*, 1997, pp. 347–354.

[11] Haihua Xu, Daniel Povey, Lidia Mangu, and Jie Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Comput. Speech Lang.*, vol. 25, no. 4, pp. 802–828, 2011.

[12] Olivier Siohan, Bhuvana Ramabhadran, and Brian Kingsbury, "Constructing ensembles of asr systems using randomized decision trees," in *ICASSP*, 2005, vol. 1, pp. 197–200.

[13] Venkata Ramana Rao Gadde, Andreas Stolcke, Dimitra Vergyri, Jing Zheng, M. Kemal Sönmez, and Anand Venkataraman, "Building an ASR system for noisy environments: Sri's 2001 SPINE evaluation system," in *INTERSPEECH*, John H. L. Hansen and Bryan L. Pellom, Eds., 2002.

[14] Yulia Tsvetkov, Florian Metze, and Chris Dyer, "Augmenting translation models with simulated acoustic confusions for improved spoken language translation," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, 2014, pp. 616–625.

[15] Arne Mauser, Richard Zens, Evgeny Matusov, Sasa Hasan, and Hermann Ney, "The RWTH statistical machine translation system for the IWSLT 2006 evaluation," in *IWSLT*, 2006, pp. 103–110.

[16] Christian Fügen, Alex Waibel, and Muntsin Kolss, "Simultaneous translation of lectures and speeches," *Mach. Transl.*, pp. 209–252, 2007.

[17] Mark Hopkins and Jonathan May, "Tuning as ranking," in *ACL*, 2011, pp. 1352–1362.

[18] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy, "Hierarchical attention networks for document classification," in *NAACL-HLT*, 2016, pp. 1480–1489.

[19] Jindrich Libovický and Jindrich Helcl, "Attention strategies for multi-source sequence-to-sequence learning," in *ACL*, 2017, pp. 196–202.

[20] Xiaofei Wang, Ruizhi Li, Sri Harish Mallidi, Takaaki Hori, Shinji Watanabe, and Hynek Hermansky, "Stream attention-based multi-array end-to-end speech recognition," in *ICASSP*, 2019, pp. 7105–7109.

[21] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, and Kazuhiko Sumi, "Attention-based multimodal fusion for video description," in *ICCV*, 2017, pp. 4203–4212.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[23] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze, "How2: A large-scale dataset for multimodal language understanding," *CoRR*, vol. abs/1811.00347, 2018.

[24] Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura, "Transformer-based direct speech-to-speech translation with transcoder," in *SLT*, 2021, pp. 958–965.

[25] Antonios Anastasopoulos, Ondrej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Miguel Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander H. Waibel, Changhan Wang, and Matthew Wiesner, "FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN," in *IWSLT*, 2021, pp. 1–29.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[27] Potsawee Manakul, Mark J. F. Gales, and Linlin Wang, "Abstractive spoken document summarization using hierarchical model with multi-stage attention diversity optimization," in *INTERSPEECH*, 2020, pp. 4248–4252.

[28] Tzu-En Liu, Shih-Hung Liu, and Berlin Chen, "A hierarchical neural summarization framework for spoken documents," in *ICASSP*, 2019, pp. 7185–7189.

[29] Anthony Rousseau, Paul Deléglise, and Yannick Estève, "TED-LIUM: an automatic speech recognition dedicated corpus," in *LREC*, 2012, pp. 125–129.

[30] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *ACL*, 2004, pp. 74–81.

[31] Yang Liu and Mirella Lapata, "Text summarization with pretrained encoders," in *EMNLP-IJCNLP*, 2019, pp. 3728–3738.

[32] Krzysztof Wolk and Krzysztof Marasek, "Enhanced bilingual evaluation understudy," *CoRR*, vol. abs/1509.09088, 2015.