A Multimodal Simultaneous Interpretation Prototype : Who Said What

Xiaolin Wang xiaolin.wang@nict.go.jp
Masao Utiyama mutiyama@nict.go.jp
Eiichiro Sumita eiichiro.sumita@nict.go.jp
Advanced Translation Research and Development Promotion Center
National Institute of Information and Communications Technology, Japan

Abstract

"Who said what" is essential for human users to understand video streams that have more than one speaker, but conventional simultaneous interpretation (SI) systems merely present "what was said" in the form of subtitles. Because translations unavoidably have delays and errors, users often find it difficult to trace the subtitles back to speakers. Therefore, we propose a multimodal SI system that explicitly presents users "who said what" – translation annotated with the textual tags and face icons of speakers.

Speaker recognition requires heavy computation which poses a big challenge to implementing our proposed system especially given the real-time characteristics of SI. We integrate multimodal speaker recognition with online sentence-based SI to meet this challenge as follows. First, our system employs automated speech recognition and online sentence segmenter to segment video streams into video clips each of which contains one sentence. Next, our system recognizes the speaker in each video clip using active speaker detection, voice embeddings and face embeddings; in the meantime, it translates the sentence into target language in the meantime. In the end, our system presents users both the translation and the speaker of each sentence.

Our method has two major merits. First, speaker recognition is performed per video clip, so GPUs can have enough input data to produce large batches and run efficiently. Second, speaker recognition is synchronized with machine translation, so no extra latency is introduced. As a result, our demo system is capable of interpreting video streams in real-time on a single desktop equipped with two Quadro RTX 4000 GPUs.

In addition, we full respect the privacy of users. Our system aims at distinguishing different speakers appearing in a video stream rather than figuring out the real name or identity of speakers in the physical world. As a side merit, our system requires no prior knowledge of speakers.

1 Introduction

Automated simultaneous interpretation (SI) is promising for facilitating real-time cross-lingual communication. Video streams have become a most popular form of communication nowadays because of the blossom of smart phones, video sites, chat apps and so on. Therefore, there is an increasing trend towards applying SI to video streams.

Figure 1: User Interface of Multimodal Simultaneous Interpretation

Spoken language such as conversations, discussions and debates are common in video streams where two or more speakers are involved. Working out "who said what" is a natural path for people to understand these video streams. However, when applying conventional SI, users are merely presented "what was said", that is, the translation of the transcripts from speech recognition. Therefore, users must guess who the speaker of the source utterance of each translation was. Given the following facts between the source utterances and translations,

- uncertain delays in the timeline;
- mismatch in length due to different languages;
- speech recognition errors;
- translation errors;
- speaking too fast;
- ...

users often become exhausted or even desperate in working out the speaker of each translation.

To address this problem, this paper presents a novel multimodal SI system that presents users "who said what" from video streams. As illustrated by Figure 1, in the graphic user interface of our system, each record consists of the translation of an utterance which indicates "what was said"; a speaker tag, e.g., $spk\ 0$, $spk\ 1$, and a face icon which indicates "who said". Users will be able to understand the video streams easily through reading these records.

The main challenges for us to build the proposed multimodal SI system are

1. How to recognize speakers from video streams?
2. How to maintain low latency for interpretation?

To solve the first challenge, we develop a speaker predictor (SP). SP creatively adapts a multimodal speaker recognition approach that combines voice embedding (Li et al., 2017), face

I work well under pressure
wonderful
and what would you say are some of your weaknesses
one of my biggest weaknesses is asking for help when I need it .
I 'd like to do better at that
I appreciate your honesty mister wang
what can you tell me about some of your goals over the next few years
my primary goal is to gain more work experience
so a position like this would help me meet that goal
I 'd also like to learn more about the different aspects of banking
I think those goals are very smart

(a) Plain Translations

spk0 : I work well under pressure
spk1 : wonderful
spk1 : and what would you say are some of your weaknesses
spk0 : one of my biggest weaknesses is asking for help when I need it .
spk0 : I 'd like to do better at that
spk1 : I appreciate your honesty mister wang
spk1 : what can you tell me about some of your goals over the next few years
spk0 : my primary goal is to gain more work experience
spk0 : so a position like this would help me meet that goal
spk0 : I 'd also like to learn more about the different aspects of banking
spk1 : I think those goals are very smart

(b) Translations annotated with speakers

Figure 2: Comparison of Readability of Translations with and without Speaker Annotations

embedding (Schroff et al., 2015) and active speaker detection (Roth et al., 2019). To solve the second efficiency challenge, we synchronize all the heavy multimodal computation with the sentence-based interpretation (Wang et al., 2019) so that neural networks can run efficiently on GPUs through processing large batches as input.

The rest of this paper is organized as follows. First, Section 2 describes the user interface of our system. Then, Section 3 presents the architectures of sentence-based multimodal SI and multimodal speaker recognition. Next, Section 4 describes the implementation of each module in detail. After that, Section 5 briefly analyzes the real-time capability of our system. Furthermore, Section 6 compares our system with related works on multimodal machine translation. Finally, Section 7 concludes this paper with a description on future works.

2 User Interface

The user interface of our system is split into two parts as illustrated by Figure 1. The input video stream is displayed in the left part of the window, below which translations are presented in the conventional way as subtitles. The main output of our SI system is displayed in the right part of the window which consists of speaker tags, face icons and translations.

Speaker Tags are texts that follow the pattern of $spk\ n$, such as $spk\ 0$ and $spk\ 1$, where n is a number assigned to each physical speaker based on the order of his or her appearance in video streams. Because these tags are plain texts, in addition to being shown in the graphic user interface, they can also be used to generate textual transcripts for review or post-editing. Speaker annotations can greatly improve the readability of the transcripts as illustrated by Figure 2.

Face Icons are the face shots of the speakers when they are saying the corresponding utterances. These face icons have two merits. First, they allow users to double check whether

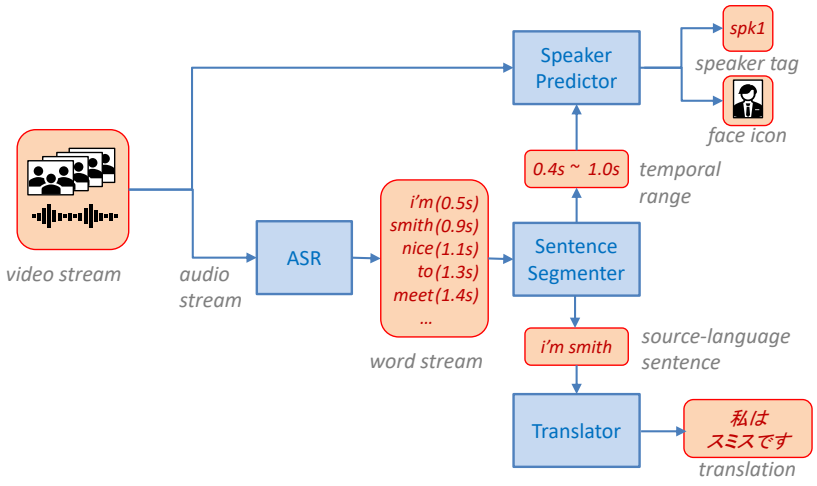

Figure 3: System Architecture

the speaker tags are correct or not. Second, they allow users to predict the sentiments of the utterances from the facial expressions of the speakers.

Translations are translated sentences. Our system performs sentence-based simultaneous interpretation. Instead of waiting for a speaker to finish speaking, it detects the event that the speaker has finished a sentence (Wang et al., 2016b). Then our system translates that sentence and displays it to users.

3 System Architecture

Our system consists of an automatic speech recognition (ASR) engine, a sentence segmenter, a speaker predictor and a translator as illustrated by Figure 3). Blue rectangles represent modules and red rounded rectangles represent data examples. Arrows show the direction of data flow.

The system takes a video stream as input, and generates the speaker tags, face icons and translations in an online manner. The system accomplishes this multimodal simultaneous interpretation task in four steps as follows.

1. The ASR engine receives audio signals from the video stream and convert it a word stream. The word stream consists of words and their time stamps. For example, "i'm (0.5s)" means that the word "i'm" appears at the 0.5 seconds of the video stream.

2. The sentence segmenter splits the word stream into source-language sentences. Each sentence consists of a text and a temporal range. For example, the first sentence is "i'm smith" with a temporal range of 0.4 seconds to 1.0 seconds.

3. The speaker predictor first extracts a clip from the video stream following the temporal range of each sentence, and then recognizes the active speaker, that is, the person who is speaking in the clip. The active speaker is assumed to say the corresponding sentence. Each active speaker is presented by a textual tag and a face icon.

4. The translator translates the text of each sentence into target-language, which is a standard machine translation task. For example, "i'm smith" is translated into a Japanese one.

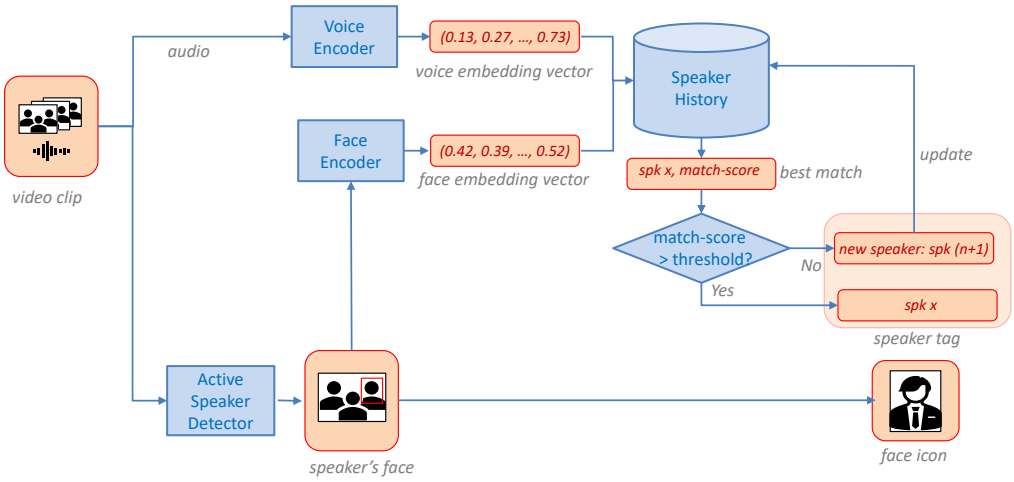

Figure 4: Speaker Predictor Architecture

3.1 Speaker Predictor

The speaker predictor is more complicated than the other three modules, which further consists of an active speaker detector, a face encoder, and a voice encoder (Figure 4). In addition, an ad-hoc database named speaker history contains the face embedding vectors and voice embedding vectors of the speakers who have appeared in the current video stream.

The speaker predictor takes a video clip as input and generates the speaker tags and face icons as output. The speaker prediction task is accomplished as follows,

1. The active speaker detector recognizes the person who is speaking in the video clip. A face icon of the active speaker is extracted to represent the prediction result.

2. The face encoder converts the face icon into a face embedding vector, noted as v_f.

3. The voice encoder converts the audio of the video clip into a voice embedding vector, noted as v_a.

4. The database of speaker history is searched for the speaker that best matches the face and voice embedding vectors, noted as spk x, with a matching score, formulated as,

$$\begin{aligned} x &= \underset{x}{argmax}\, cos(v_f, v_f^x) \\ &\quad + cos(v_a, v_a^x) \end{aligned} \tag{1}$$

   where v_f^x and v_a^x is the face and voice embedding vectors of the speaker x, cos means cosine similarity.

5. The matching score is compared with a predefined threshold. If the matching score exceeds the threshold, the current speaker will be predicted as spk x.

6. If the matching score is lower than the threshold, the current speaker will be treated as a new speaker. A new tag spk $(n$+1) will be assigned, where n is the number of appeared speakers. In addition, the new tag together with the face and voice embedding vectors will be added into the database of speaker history

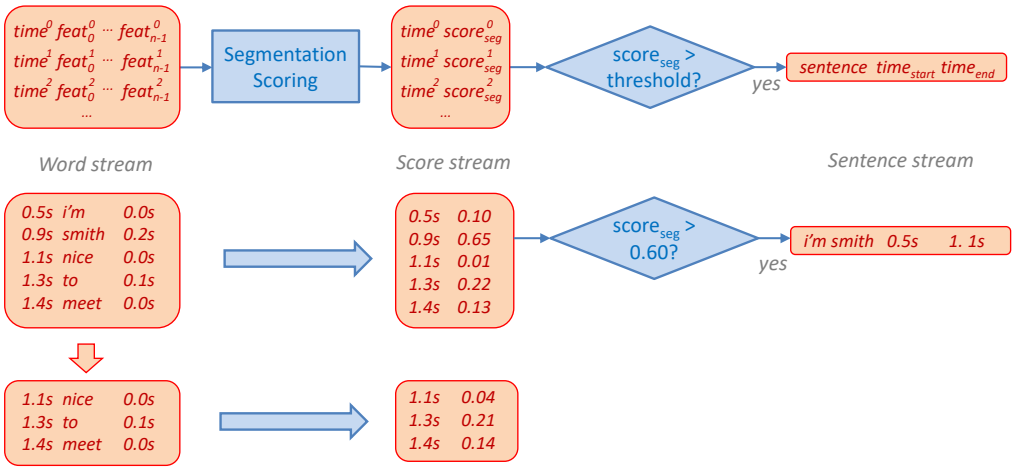

Figure 5: Senentence Segmenter

3.2 Sentence Segmenter

The sentence segmenter adopts a sequence labelling architecture (Figure 5). The input of word stream is viewed as a sequence of time stamps $time^k$ and features $feat_0^k \ldots feat_{n-1}^k$ which represent the k-th word.

The sentence segmenter calculates a segmentation confidence score as,

$$score_{seg}(k) = F_{seg}(feat_0^0, \cdots, feat_{n-1}^0, \cdots, feat_0^{k+K}, \cdots, feat_{n-1}^{k+K}) \quad (2)$$

where $score_{seg}(k)$ means the confidence of segmenting after the k-th word, F_{seg} represents a scoring model, and K is the size of right context,

Our system employs two features to represent a word as illustrated by Figure 5. One is the surface text of the word, and the other is the duration of the speaker pause after the word. The implementation of the scoring model is presented in Section 4.2.

The segmentation scores are compared with a predefined threshold. If $score_{seg}(k)$ exceeds the threshold, a segment will be produced from $time_0$ to $time_{k+1}$. After that, the segmenter module will be reset to process the remaining words that start from $time_{k+1}$.

4 System Implement

4.1 Automated Speech Recognition

The ASR module is required to be not only accurate but also low-latency due to the real-time characteristic of the simultaneous interpretation task (Wang et al., 2016b; Novitasari et al., 2019; Nguyen et al., 2020). Our current solution is the streaming convolution model proposed by Pratap et al. (2020) [1]. We are aware that state-of-the-art speech recognition models are transformers (Likhomanenko et al., 2021) and conformers (Gulati et al., 2020). They are more accurate than convolution models, but they typically operate on audio segments instead of audio streams. Adapting transformers and conformers to the input of audio streams is a trending topic (Moritz et al., 2020; Tsunoo et al., 2020; Chen et al., 2021). We are paying close attention

[1] https://github.com/flashlight/wav2letter/tree/main/recipes/streaming_convnets

to this field, and plan to upgrade our system when matured streaming transformer or conformers are available.

4.2 Sentence Segmenter

The sentence segmenter module segments word stream into sentences in an online manner (Stolcke et al., 1998; Sridhar et al., 2013; Wang et al., 2016a, 2019; Iranzo-Sánchez et al., 2020; Li et al., 2021; Wicks and Post, 2021; Gravellier et al., 2021).

Because large-scale supervised training corpora for the sequence labelling problem of sentence segmentation (Section 3.2) are not publicly available, we manually craft the scoring model for sentence segmentation as

$$\begin{aligned} score_{seg}(k) &= score_{RNN}(w_0, \cdots, w_{k+K}) \\ &\quad + \alpha\, pause(k) \end{aligned} \tag{3}$$

where $score_{RNN}$ is the segmentation score from the RNN-based model proposed by Wang et al. (2019) [2], $pause(k)$ is the duration of speaker pause after the word w measured in seconds, and α is a manually tuned weight.

4.3 Translator

The translator module translates one source-language sentence into one target-language sentence, which is a standard machine translation task (Brown et al., 1993; Zens et al., 2002; Chiang, 2005; Bahdanau et al., 2014). We employ our in-house machine translation system as the translator module. The machine translation system is publicly accessible through a Web API [3].

4.4 Active Speaker Detector

The active speaker detector module recognizes who is speaking in a visual scene from one or more candidates (Roth et al., 2019; Kim et al., 2021). Active speaker detection is an emerging research topic (Chakravarty et al., 2016; Chung, 2019; Zhang et al., 2021; Tao et al., 2021; Köpüklü et al., 2021; León-Alcázar et al., 2021). Our system adopts the end-to-end multimodal (video and audio) active speaker detection framework proposed by Roth et al. (2019) because of the trade-off between accuracy and efficiency. The framework first employs 3-D convolutional neural networks to convert visual and audio into embedding vectors, and then concatenates the embedding vectors to make predictions.

4.5 Face Encoder

The face encoder module converts face images into embedding vectors, the similarity of which directly corresponds to a measure of face similarity (Schroff et al., 2015). For our application, the face encoder is highly demanding on efficiency as simultaneous interpretation is a real-time task, while less demanding on accuracy as the encoder only needs to distinguish a small number of people that appear in a same video stream. Therefore, our face encoder is a middle-sized model of Resnet50 (He et al., 2016) which is trained on the dataset of labeled faces in the wild (LFW) (Huang et al., 2008) using the triplet loss proposed by Schroff et al. (2015).

4.6 Voice Encoder

The voice encoder module converts audio utterances into embedding vectors, the similarity of which corresponds to a measure of voice similarity. Voice encoder is related to the task of speaker verification which aims at verifying the identity of a person from the characteristics of

[2] https://github.com/arthurxlw/cytonNss

[3] https://mt-auto-minhon-mlt.ucri.jgn-x.jp/

Module	Workload
ASR	45.0%
Sentence Segmenter	0.2%
Translator	2.2%
Speaker Predictor	10.4%

Table 1: Workload Percentage of Each Module. The lower the better.

his or her voice (Li et al., 2017; Chung et al., 2018; Wan et al., 2018; Nagrani et al., 2020). Our voice encoder is a pretrained Resnet34 model released in (Chung et al., 2020; Heo et al., 2020)[4].

5 Real-time Capability

Our SI system employs a parallel pipeline to integrate the main modules to meet the real-time requirement. We estimate the real-time capability of each module using workload percentage, which is calculated as,

$$\frac{T_{running}}{T_{running} + T_{idle}} \times 100\%, \tag{4}$$

where $T_{running}$ and T_{idle} are running and idle durations respectively.

Table 1 shows that the workload percentages of all the modules are below 100% thus our system is fully capable of interpreting video streams in real-time.

6 Related Works

Multimodal machine translation – the task of doing machine translation with multiple data sources – is a trending topic (Specia et al., 2016; Di Gangi et al., 2019; Sanabria et al., 2018). A large volume of research effort has been dedicated to improving the translation quality through drawing information from modalities other than text (Sulubacak et al., 2019; Hirasawa et al., 2019; Lin et al., 2020; Yao and Wan, 2020; Mitzalis et al., 2021).

Our interpretation system approaches the task of multimodal machine translation from a different angle. Imaging when interpretating a video stream, the visual contents of the video stream will mainly fall into two categories,

1. the speakers;

2. the subjects of the speeches.

Our interpretation system focuses on the first category. It recognizes the speaker of each utterance, and then annotates the translation with the speaker, so that users can better understand the video stream despite translation latencies and errors. In contrast, the related works focus on the second category of contents so that users can get better translations.

Nevertheless, our interpretation system and the related works on multimodal machine translation are complement with each other. Integrating our system with the related works will lead to very effective interpretation systems which can generate both well-annotated and high-quality translations from video streams.

[4] https://github.com/clovaai/voxceleb_trainer

7 Conclusion

In this paper, we propose an automated multimodal simultaneous interpretation system to improve the user experience on interpreting video streams, and build an efficient implementation based on the sentence-based interpretation.

Our system has been tested on various video streams. The system works very well on some of the video streams and produces high-quality translations which are correctly annotated with the tags and face icons of speakers.

However, our system performs poorly on some video streams which have difficult speeches. When the speech is not clear enough for the ASR module to generate decent transcripts, the sentence segmenter will fail to produce sensible sentences, and then the whole system will perform poorly. Therefore, in the future, we plan to address this problem through adding more audio and visual features into the sentence segmenter and the translator to improve the robustness of our system.

Ethic

Our proposed simultaneous interpretation system fully respects users' privacy. The system is designed not to figure out the real name or identity of speaker in the physical world. Instead, speakers are only given plain tags such as $spk\,0$ and $spk\,1$ to distinguish from each other.

As a result, our simultaneous interpretation system requires no prior knowledge of speakers. It only collects the necessary information to distinguish speakers when performing interpretation tasks. The collected information will be erased when the tasks finish.

Acknowledgement

A part of this work was conducted under the commissioned research program "Research and Development of Advanced Multilingual Translation Technology" in the "R&D Project for Information and Communications Technology (JPMI00316)" of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *Proceedings of the 3rd International Conference on Learning Representations.*, pages 1–15.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational linguistics*, 19(2):263–311.

Chakravarty, P., Zegers, J., Tuytelaars, T., and Van hamme, H. (2016). Active speaker detection with audio-visual co-training. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 312–316.

Chen, X., Wu, Y., Wang, Z., Liu, S., and Li, J. (2021). Developing real-time streaming transformer transducer for speech recognition on large-scale dataset.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*, pages 263–270.

Chung, J. S. (2019). Naver at activitynet challenge 2019–task b active speaker detection (ava). *arXiv preprint arXiv:1906.10555*.

Chung, J. S., Huh, J., Mun, S., Lee, M., Heo, H. S., Choe, S., Ham, C., Jung, S., Lee, B.-J., and Han, I. (2020). In defence of metric learning for speaker recognition. In *Interspeech*.

Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., and Turchi, M. (2019). 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.

Gravellier, L., Hunter, J., Muller, P., Pellegrini, T., and Ferrané, I. (2021). Weakly supervised discourse segmentation for multiparty oral conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1381–1392.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heo, H. S., Lee, B.-J., Huh, J., and Chung, J. S. (2020). Clova baseline system for the voxceleb speaker recognition challenge 2020. *arXiv preprint arXiv:2009.14153*.

Hirasawa, T., Yamagishi, H., Matsumura, Y., and Komachi, M. (2019). Multimodal machine translation with embedding prediction. *arXiv preprint arXiv:1904.00639*.

Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

Iranzo-Sánchez, J., Pastor, A. G., Silvestre-Cerda, J. A., Baquero-Arnal, P., Saiz, J. C., and Juan, A. (2020). Direct segmentation models for streaming speech translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2599–2611.

Kim, Y. J., Heo, H.-S., Choe, S., Chung, S.-W., Kwon, Y., Lee, B.-J., Kwon, Y., and Chung, J. S. (2021). Look who's talking: Active speaker detection in the wild.

Köpüklü, O., Taseska, M., and Rigoll, G. (2021). How to design a three-stage architecture for audio-visual active speaker detection in the wild.

León-Alcázar, J., Heilbron, F. C., Thabet, A., and Ghanem, B. (2021). Maas: Multi-modal assignation for active speaker detection.

Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., and Zhu, Z. (2017). Deep speaker: an end-to-end neural speaker embedding system.

Li, D., Te, I., Arivazhagan, N., Cherry, C., and Padfield, D. (2021). Sentence boundary augmentation for neural machine translation robustness. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7553–7557. IEEE.

Likhomanenko, T., Xu, Q., Pratap, V., Tomasello, P., Kahn, J., Avidov, G., Collobert, R., and Synnaeve, G. (2021). Rethinking evaluation in asr: Are our models robust enough?

Lin, H., Meng, F., Su, J., Yin, Y., Yang, Z., Ge, Y., Zhou, J., and Luo, J. (2020). Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329.

Mitzalis, F., Caglayan, O., Madhyastha, P., and Specia, L. (2021). Bertgen: Multi-task generation through bert.

Moritz, N., Hori, T., and Roux, J. L. (2020). Streaming automatic speech recognition with the transformer model.

Nagrani, A., Chung, J. S., Xie, W., and Zisserman, A. (2020). Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027.

Nguyen, T. S., Niehues, J., Cho, E., Ha, T.-L., Kilgour, K., Muller, M., Sperber, M., Stueker, S., and Waibel, A. (2020). Low latency asr for simultaneous speech translation.

Novitasari, S., Tjandra, A., Sakti, S., and Nakamura, S. (2019). Sequence-to-sequence learning via attention transfer for incremental speech recognition. *Proceedings of Interspeech*, pages 3835–3839.

Pratap, V., Xu, Q., Kahn, J., Avidov, G., Likhomanenko, T., Hannun, A., Liptchinsky, V., Synnaeve, G., and Collobert, R. (2020). Scaling up online speech recognition using convnets.

Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., and Pantofaru, C. (2019). Ava-activespeaker: An audio-visual dataset for active speaker detection.

Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., and Metze, F. (2018). How2: A large-scale dataset for multimodal language understanding.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Specia, L., Frank, S., Sima'An, K., and Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.

Sridhar, V. K. R., Chen, J., Bangalore, S., Ljolje, A., and Chengalvarayan, R. (2013). Segmentation strategies for streaming speech translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 230–238.

Stolcke, A., Shriberg, E., Bates, R. A., Ostendorf, M., Hakkani, D., Plauche, M., Tür, G., and Lu, Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proceedings of 5th International Conference on Spoken Language Processing*, pages 2247–2250.

Sulubacak, U., Caglayan, O., Grönroos, S.-A., Rouhe, A., Elliott, D., Specia, L., and Tiedemann, J. (2019). Multimodal machine translation through visuals and speech.

Tao, R., Pan, Z., Das, R. K., Qian, X., Shou, M. Z., and Li, H. (2021). Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. *Proceedings of the 29th ACM International Conference on Multimedia*.

Tsunoo, E., Kashiwagi, Y., and Watanabe, S. (2020). Streaming transformer asr with blockwise synchronous beam search.

Wan, L., Wang, Q., Papir, A., and Moreno, I. L. (2018). Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE.

Wang, X., Finch, A., Utiyama, M., and Sumita, E. (2016a). An efficient and effective online sentence segmenter for simultaneous interpretation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 139–148, Osaka, Japan. The COLING 2016 Organizing Committee.

Wang, X., Finch, A., Utiyama, M., and Sumita, E. (2016b). A prototype automatic simultaneous interpretation system. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 30–34.

Wang, X., Utiyama, M., and Sumita, E. (2019). Online sentence segmentation for simultaneous interpretation using multi-shifted recurrent neural network. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 1–11.

Wicks, R. and Post, M. (2021). A unified approach to sentence segmentation of punctuated text in many languages. 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 3995–4007, Online. Association for Computational Linguistics.

Yao, S. and Wan, X. (2020). Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350.

Zens, R., Och, F. J., and Ney, H. (2002). Phrase-based statistical machine translation. In *Advances in Artificial Intelligence*, pages 18–32. Springer.

Zhang, Y., Liang, S., Yang, S., Liu, X., Wu, Z., Shan, S., and Chen, X. (2021). Unicon: Unified context network for robust active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3964–3972.