SLIDESPEECH: A LARGE SCALE SLIDE-ENRICHED AUDIO-VISUAL CORPUS

Haoxu Wang[1,3], *Fan Yu*[3], *Xian Shi*[3], *Yuezhang Wang*[3], *Shiliang Zhang*[3✉], *Ming Li*[1,2]

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Duke Kunshan University, Kunshan, China
[3]Speech Lab of DAMO Academy, Alibaba Group, China

ABSTRACT

Multi-Modal automatic speech recognition (ASR) techniques aim to leverage additional modalities to improve the performance of speech recognition systems. While existing approaches primarily focus on video or contextual information, the utilization of extra supplementary textual information has been overlooked. Recognizing the abundance of online conference videos with slides, which provide rich domain-specific information in the form of text and images, we release SlideSpeech, a large-scale audio-visual corpus enriched with slides. The corpus contains 1,705 videos, 1,000+ hours, with 473 hours of high-quality transcribed speech. Moreover, the corpus contains a significant amount of real-time synchronized slides. In this work, we present the pipeline for constructing the corpus and propose baseline methods for utilizing text information in the visual slide context. Through the application of keyword extraction and contextual ASR methods in the benchmark system, we demonstrate the potential of improving speech recognition performance by incorporating textual information from supplementary video slides.

Index Terms— audio visual speech recognition, corpus, slides

1. INTRODUCTION

Recently, there has been a significant advancement in the development of automatic speech recognition (ASR) technology. Traditional methods based on Hidden Markov Models (HMM)[1, 2] have been replaced by deep learning based techniques such as Connectionist Temporal Classification (CTC)[3, 4], Attention-based Encoder-Decoder (AED)[5, 6, 7], and Neural Transducer[8, 9].

However, the development of a robust and generalized ASR model remains a challenge. Recognition performance tends to degrade in far-field or specific domain scenarios. To address this issue, scholars have explored the integration of multi-modal information such as lip movements[10, 11, 12], open domain information[13, 14], and contextual biasing list[15, 16], etc. This multi-modal information gives complementary lip movements, semantic images, and the context to improve the ASR results.

In the past, there are many multi-modal video datasets, some of which focus on facial and lip movements, such as LRS2[10], LRS3[17], VoxCeleb2[18], and CN-Celeb-AV[19]. Others focus on open domains, such as HowTo100[20], How2[21], and VisSpeech[13]. However, in addition to facial expressions and semantic images, there is multi-modal textual information in the video that is closely related to the current speech of the speaker. Few articles consider utilizing textual information in videos to improve ASR results. In online conference sharing scenarios and online education scenarios where there are screen recordings or live broadcasts, there are often slides that are synchronized with the speaker's speech. These slides contain much rich textual information that has not been fully utilized. If only ASR technology is used, it may lead to the wrong recognition of proprietary entities in the current slide. In the field of audio processing assisted by slides, some early papers use slides to build language models[22, 23] or do some down-stream tasks[24, 25], and others use complete static slides to extract rare words and improve results using a contextual bias ASR model[26].

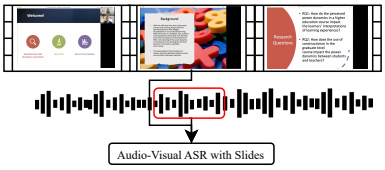


Fig. 1. Example of our SlideSpeech using the Audio-Visual ASR Benchmark.

Previous works mainly focus on using static slides, with little consideration given to the high correlation between each short speech segment and the current slide. As shown in Fig. 1, we propose to use real-time synchronized slide and speech stream for multi-modal ASR, as a "what you see is what you get" approach to improve the recognition of proprietary terms, and to avoid focusing solely on speech without paying attention to the textual information in the video slides.

We investigate past database containing slides, and find only few database, such as AMI[27] containing slide information. However, there are no time-synchronized recordings with slides sharing, and only few recordings containing slide images. To further study the improvement of ASR performance using real-time textual streams in multi-modal scenarios, we release SlideSpeech, a large-scale multi-modal audio-visual corpus with a significant amount of real-time synchronized slides. The key features of SlideSpeech include:

- Abundance of slides in videos. These can be used for text-enhanced multi-modal ASR to correct the wrong recognized proprietary terms.
- Sizeable. 1,705 videos with a total duration of 1,000+ hours, and 473 hours of transcribed speech with a confidence level above 95%.
- Diverse. A diverse range of domain categories, with 22 classes.
- Easy to expand or target specific scenarios. The slides in the videos also include relevant images and some facial information. This corpus can be applied for automatic subtitle generation in online education scenarios.

We will present the pipeline used to construct the entire SlideSpeech corpus and discuss our approach to addressing the challenges of multi-

✉Corresponding author: sly.zsl@alibaba-inc.com.

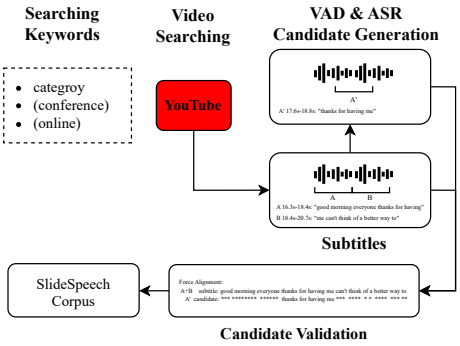


Fig. 2. Diagram of our Creation pipeline.

modal ASR in the context of slides. Taking inspiration from the work of [26], we develop a pipeline for text-based multi-modal ASR in synchronized slides and video scenarios. This pipeline incorporates various models, including text detection (TD), optical character recognition (OCR), semantic keyword extraction, and contextual bias ASR. Leveraging this pipeline, we establish a benchmark system based on SlideSpeech, enabling comparative analysis and further research.

2. CREATION PIPELINE

As shown in Fig. 2, we introduce the detailed creation pipeline of our SlideSpeech corpus, including candidate video searching, candidate audio/text segments generation, and candidate validation.

2.1. Candidate Video Searching

Our corpus primarily consists of video content from YouTube. We design retrieval keywords as specific categories + (conference) + (online), focusing on video playlists curated by uploaders or users, along with independent channel videos. Manual screening is conducted to collect slide videos, categorizing a video as a candidate if it features slides for over 50% of its duration. Such videos are included in the download list. Playlists with 10 or more candidate videos or playlists comprising at least 80% of candidates are also added to the download list. Finally, the yt-dlp[1] tool is used to download all the candidate videos.

2.2. Candidate Segments Generation

Inspired by GigaSpeech[28] and WenetSpeech[29], we use our in-house VAD and ASR systems to generate candidate transcripts for the downloaded videos. Our ASR system has shown exceptional performance on public benchmark platforms, consistently achieving accuracy rates exceeding 95% across diverse testing scenarios and various public ASR databases. The audio from all videos is segmented using VAD, and the ASR system is then utilized to generate candidate transcripts for each segment. Thus, we obtain the audio/text segments.

2.3. Candidate Validation

While downloading the videos, we also obtain the automatically generated or user-uploaded subtitle files from YouTube. These files and the transcripts generated in Sec. 2.2 are used for candidate validation. However, due to their imprecise nature, we do not utilize the timestamps from the YouTube-generated subtitles. Nevertheless, the subtitle files generally cover all the text in the subtitles.

Though the subtitles and ASR timestamps do not perfectly align, we still can use them for validation. By concatenating multiple subtitles based on their timestamps, we ensure complete coverage of each short segment with a candidate transcript generated by the ASR system. Then, we obtain corresponding text subtitles and ASR candidate transcripts. As the concatenated subtitles' time duration exceeds the segment duration, the subtitle text is longer than the candidate transcript. To align the longer subtitles and candidate transcripts, we employ the Smith-Waterman algorithm[30].

After aligning the two texts, we remove any unmatched words at the beginning and end of the subtitle text. We then computed the word error rate (WER) between the resulting subtitle text and the candidate transcript, and set $confidence = (1 - WER)$ for the current speech-text pair segment, which is used for subsequent database filtering.

3. THE SLIDESPEECH CORPUS

In this section, we present the metadata, training sets, evaluation sets and the diversity of the corpus. For downloading the current corpus, please refer to the GitHub link provided[2]. This corpus is not commercially available; it is only intended for academic research.

3.1. Metadata

We provide detailed metadata for the videos, including the original YouTube channel, YouTube playlist ID, domain tags, and segments. Each segment is accompanied by its corresponding timestamp, transcript, and confidence score. We also provide the scripts to download the videos. The downloaded video files are in 720p format, and the audios are converted to a 16k sample rate, single-channel, and 16-bit signed-integer format. We also offer preprocessed OCR results and extracted keywords for each segment.

3.2. Training Sets

The training set consists of 1659 videos totaling 1065.86 hours and is named as L. Moreover, a subset (S) is sampled from L, comprising 279 videos with a total duration of around 206 hours. All segments are automatically annotated with varying confidence. From S, another subset named S95 is created, containing segments with a confidence score above 95%. S95 has an effective speech duration of about 161 hours. Similarly, a subset called L95 is formed from L, resulting in an effective speech duration of about 473 hours. Annotations with confidence levels below 95% are retained in the original corpus for other academic purposes, such as self-supervised training.

3.3. Evaluation Sets

We provide development (dev) and test sets in addition to the training set. The dev set consists of 21 videos (5.07 hours), while the test set comprises 25 videos (8.75 hours). Unlike the training set, annotations in the dev and test sets are manually labeled. Additionally, after manually checking and random sampling 100 segments, we found that 94% of segments in the dev and test sets contain slides. The dev and test sets are collected from YouTube, and we ensured no overlap with the training set regarding category and ID.

3.4. Diversity

The categories in the retrieval keyword are utilized as labels for videos. Videos without a specific category or those manually identified as not belonging to any category were assigned to an "other" category. As shown in Table 1, the L set has 12 categories, the dev set has 4 categories, and the test set has 6 categories. Our corpus encompasses diverse categories, ensuring no bias towards a specific conference scenario. This makes it suitable for developing text-based multi-modal ASR techniques applicable to various scenarios.

4. BENCHMARK SYSTEM

In this section, we present the text-based multi-modal ASR pipeline from the original video to the speech transcript.

[1] https://github.com/yt-dlp/yt-dlp

[2] https://slidespeech.github.io/

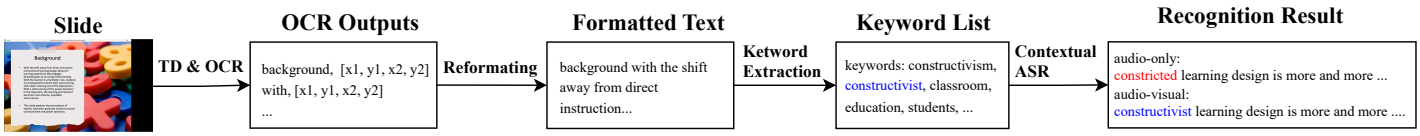


Fig. 3. Diagram of our benchmark pipeline.

Table 1. The domian distribution with video counts and duration of the SlideSpeech.

Set	Domian	Num.	Dur.(h)
L	Computer Science	435	161.0
	Musical Instruments	23	29.1
	History	310	240.3
	Agriculture	67	69.3
	Animation	192	123.7
	Music	87	42.9
	Parenting	107	97.3
	Travel	108	93.6
	Life	43	36.3
	Talent	7	7.1
	English	20	17.0
	Other	260	148.4
Dev	Pet	5	1.57
	Health	6	1.36
	Dance	4	0.82
	Medical	6	1.32
Test	Fitness	3	1.98
	Design	5	1.09
	Traffic	3	0.84
	Education	5	1.88
	Tradition Culture	5	1.24
	Child	4	1.73

4.1. Pipeline

As shown in Fig. 3, for each speech segment, we extract the middle frame image and apply TD[31] and OCR[32] models from the MMOCR toolkit[33] to extract the words in the slide. The words are then reformatted based on their coordinates to obtain formatted long text. Keyword extraction is performed using the KeyBert[34] technique, and contextual ASR is employed to enhance recognition for specific terms. Our benchmark system utilizes the contextual phrase prediction network (CPP network) ASR[35], as the contextual ASR model.

4.2. Contextual Bias ASR System

We utilize Contextualized CTC/AED as our contextual ASR model, as depicted in Fig. 4. The model consists of a context encoder, a multi-head cross-attention based biasing layer, and a CPP network integrated with the traditional encoder. The contextual phrases are transformed into token sequences using byte-pair encoding (BPE). These sequences are then passed through a bidirectional LSTM, extracting the last time step's embedding ($h_i^{CE} \in R^d$) as the embedding for each contextual phrase i. Additionally, we include a special token <nobias>in the contextual biasing list before BPE encoding. This allows the model to handle cases where words from the bias list are absent in the speech, ensuring inference capability even without contextual phrases.

The biasing layer is a multi-head cross-attention mechanism that allows the speech embedding to capture contextual information from the contextual phrase embedding, modeling the relationship between speech and contextual phrases. By using the speech embedding (h^E) as the query and the contextual phrase embedding (h^{CE}) as the key and value, we obtain the contextual representation (c^E). The final contextual-enhanced speech embedding (h^{CA})

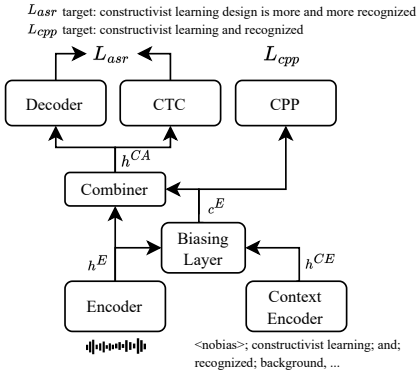


Fig. 4. Diagram of Contextualized CTC/AED Model.

is obtained by fusing c^E and h^E using the combiner, defined as $h^{CA} = FeedForward([LayerNorm(h^E), LayerNorm(c^E)])$. Additionally, the CPP network predicts contextual phrases present in the current speech. It consists of a linear projection layer and a shared CTC linear output layer with the main network. The CPP labels are generated by removing words from the original speech transcript that do not appear in the current contextual bias list, resulting in a CPP target text. During training, the context representation (c^E) is used to compute the CTC loss between c^E and the CPP target text, serving as an auxiliary loss for the ASR model. This module is not required during testing.

4.3. Training Bias List Generation

In the original CPP paper[35], a simulation approach randomly selects three phrases containing 1-3 words from the speech transcript. These phrases are combined with distractors and <nobias>to form the contextual biasing list for training. Alternatively, we can extract semantically related keywords from the actual OCR text to form a biasing list. These methods can be used independently or combined to train an effective contextual ASR network.

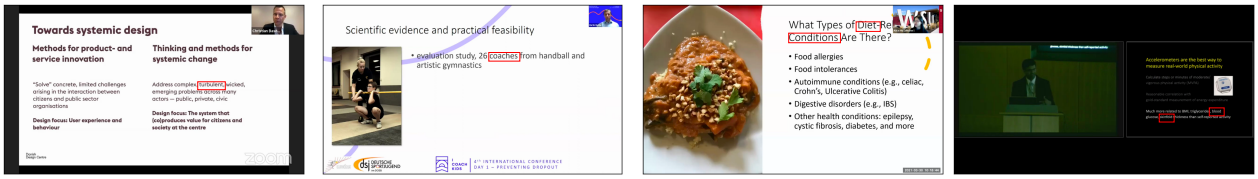
5. EXPERIMENTS AND DISCUSSION

5.1. Experments Setup

During data preparation, we generate 80-dimensional FBank features with a 25ms window and a 10ms frame shift. SpecAugment[36] is applied with frequency and time masks (F=30, T=40). Our baseline model is a conformer-based end-to-end model[37], an improved version of the transformer model that captures both long-distance dependencies and local information in speech. It consists of 12 conformer encoder blocks ($d^{ff} = 2048, H = 4, d^{att} = 256, CNN_{kernel} = 15$) and 6 transformer decoder blocks ($d^{ff} = 2048, H = 4$). We utilize a set of 5k BPE tokens generated by the SentencePiece tokenizer and an added special token <nobias>. The objective function combines CTC and attention objectives with a logarithmic linear combination using a weight parameter $\lambda = 0.3$. Label smoothing is applied to the attention objective. Training is performed on four 16GB memory V100 RTX GPUs, with a maximum trainable epoch of 70 and a mini-batch size of 22 million acoustic feature bins. We use the Adam optimizer and apply the Noam learning rate scheduler with 15k warmup steps and a learning rate of 0.002. The final model is obtained by averaging the top 10 best checkpoints.

Table 2. *Performance (%) of the baseline model and the contextual asr benchmark. w/o K represents the case without Keyword, w K represents the case with Keyword. OCR, LR, Keyword columns denote the performance is calculated according to the OCR, LR or Keyword biasing list. U/B/R refers to the U-WER/B-WER/Recall metrics, respectively.*

Model	Train	Dev				Test			
		WER	OCR(U/B/R)	LR(U/B/R)	Keyword(U/B/R)	WER	OCR(U/B/R)	LR(U/B/R)	Keyword(U/B/R)
Baseline	S95	21.05	21.54/19.04/82.64	18.75/100/0.00	20.29/31.27/68.76	21.22	21.97/18.14/83.69	19.17/100/0.00	20.83/26.60/73.51
Contextual w/o K	S95	21.06	21.56/19.01/82.63	18.98/92.63/7.83	20.29/31.37/68.73	21.25	21.97/18.27/83.51	19.39/92.52/8.00	20.83/26.96/73.17
Contextual w K	S95	20.80	21.48/18.00/83.64	18.83/88.56/11.81	20.22/28.61/71.48	20.95	21.85/17.24/84.51	19.21/87.76/12.86	20.73/24.05/76.10
Baseline	L95	13.09	13.70/10.58/90.47	11.75/100/0.00	12.87/16.13/83.90	12.89	13.70/9.59/91.45	11.78/100/0.00	12.90/12.70/87.43
Contextual w/o K	L95	12.91	13.50/10.46/90.66	11.65/93.87/6.27	12.70/15.67/84.36	12.64	13.46/9.28/91.80	11.63/91.93/8.55	12.64/12.63/87.54
Contextual w K	L95	12.64	13.46/9.25/91.85	11.57/81.89/18.25	12.66/12.39/87.64	12.38	13.42/8.13/92.91	11.53/78.87/21.81	12.60/9.32/90.86



GT: so to address more complex and turbulent and wicked emerging problems
A: so to address more complex and triple and and wicked emerging problems
AV: so to address more complex and turbulent and wicked emerging problems

GT: the coaches planned the training sessions during this period independent
A: decoders planned the training sessions during this period independent
AV: we coaches planned the training sessions during this period independent

GT: um so what types of diet related conditions are there that affect ...
A: so what types of direlated visions are there that affect ...
AV: so what types of diet related conditions are there that affect ...

GT: a blood glucose skinfold thickness then the actually measured ...
A: of like glucose in both thickness than the actually measured ...
AV: of blood glucose in full thickness then the actually measured ...

Fig. 5. Qualitative results on the SlideSpeech dataset. We show the ground truth (GT), the recognition of the audio-only baseline (A) and benchmark system (AV). Note how the keywords of the slides help correct the recognition. Errors in the recognition compared to the GT are highlighted in red and the corrected words are highlighted in blue. The words are still wrong but in the slides are highlighted in orange.

In the contextualized version, the context encoder consists of a 1-layer bidirectional LSTM (BLSTM) with a dimensionality of 256 and a linear layer. The biasing layer is a multi-head attention (MHA) layer with an embedding size of 256 and 4 attention heads. The context prediction network consists of two linear layers. The first linear layer has a 256-dimensional input and output with a tanh activation, while the second linear layer projects the input onto the vocabulary size and shares parameters with the CTC linear layer. We initially train the contextual model using the simulated biasing list and then finetune it on a combination of simulated and keyword biasing lists with a 0.5 mix ratio. We initially only train the contextual part and freeze the others from the pretrained baseline model and unfreeze all when finetune it on the mixture of the simulated and keyword biasing lists. The experiments were conducted using the ESPnet end-to-end speech processing toolkit[38].

5.2. Results

We evaluate the results using WER, biased word error rate (B-WER), unbiased word error rate (U-WER), and the recall of words in both the biasing list and transcript. U-WER is calculated explicitly for words not in the biasing list, while B-WER is computed for words in the biasing list. In cases of insertion errors, if the inserted phrase is from the biasing list, it contributes to the B-WER calculation; otherwise, it contributes to the U-WER calculation.

The biasing list used during inference and the reference biasing list used during calculating metrics can be different. We use a biasing list of 50 keywords during inference. Table 2 presents the results for various reference biasing lists. The OCR biasing list includes all the words on the slide, representing the upper limit of word assistance information that the current slide can provide. The Low-Recall (LR) biasing list comprises words present on the slide but not correctly recognized by the baseline audio-only model, representing the upper limit of utilizing slide information for correction. Lastly, the Keyword biasing list consists of the 50 keywords used during inference, closely related to the semantic information of the slide. The results of the Keyword biasing list indicate the contextual ASR model's performance.

Our baseline audio-only model trained on S95/L95 achieves 21.05/13.09 and 21.22/12.89 WER on the Dev and Test Set, respectively. On the OCR biasing list, it achieves 19.04/10.58 B-WER with a recall of 82.64%/90.47%, and 18.14/9.59 B-WER with a recall of 83.69%/91.45%. As mentioned earlier, the LR metrics represent the potential to correct recognition errors using slide information. The Keywords biasing list contains the words that carry semantic information related to the slide. The baseline trained on S95/L95 achieved 31.27/16.13 with a recall of 68.76%/83.90% on the Dev set, and 26.60/12.70 B-WER with a recall of 73.51%/87.43% on the Test set.

Compared to the baseline, the contextual ASR without the Keyword biasing list shows comparable results in WER. Using the 50 Keyword biasing list during inference, the WER of contextual ASR trained on S95/L95 improves to 20.80/12.64 and 20.95/12.38 on the Dev and Test sets. Moreover, improvements are observed in OCR, LR, and Keyword cases. In the case of Keyword, the B-WER of the model trained on S95/L95 improves to 28.61/12.39 and 24.05/9.32 on the Dev and Test sets, with a recall improvement to 71.48%/87.64% and 76.10%/90.86%. These results demonstrate the potential for enhancing specialized terms using additional slide information on this corpus. It should be noted that our main contributions include providing an open-source corpus and offering baseline methods for this corpus. The OCR metric is used to compare the performance of using slide information, and the Keyword metric is used to compare the performance of the contextual ASR model.

Qualitative Results: Further improvement results are presented in Fig. 5. Leveraging textual information from slides can effectively enhance the recognition performance of ASR systems, especially for recognizing specialized and proper nouns. The dataset used in this study and the provided baseline system serves as an essential research reference for text-based audio-visual ASR.

6. CONCLUSION

In this work, we release SlideSpeech, which is a large-scale audio-visual corpus enriched with slides. The corpus contains a significant amount of real-time synchronized slides. We introduce the creation pipeline and the details of the corpus. We also present a benchmark system for this slide-enriched corpus. The experiment results demonstrate the potential for enhancing specialized terms using additional slide information on this corpus.

7. REFERENCES

[1] Mark Gales, Steve Young, et al., "The Application of Hidden Markov Models in Speech Recognition," *Foundations and Trends® in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[2] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech*, 2016, pp. 2751–2755.

[3] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. ICML*, 2006, pp. 369–376.

[4] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki 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.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," in *Proc. NIPS 2017*, 2017, vol. 30.

[6] Linhao Dong, Shuang Xu, and Bo Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," in *Proc. ICASSP*, 2018, pp. 5884–5888.

[7] Niko Moritz, Takaaki Hori, and Jonathan Le, "Streaming Automatic Speech Recognition with the Transformer Model," in *Proc. ICASSP*, 2020, pp. 6074–6078.

[8] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar, "Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss," in *Proc. ICASSP*, 2020, pp. 7829–7833.

[9] Fangjun Kuang, Liyong Guo, Wei Kang, Long Lin, Mingshuang Luo, Zengwei Yao, and Daniel Povey, "Pruned RNN-T for fast, memory-efficient ASR training," in *Proc. Interspeech*, 2022, pp. 2068–2072.

[10] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep Audio-Visual Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8717–8727, 2022.

[11] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-To-End Audio-Visual Speech Recognition with Conformers," in *Proc. ICASSP*, 2021, pp. 7613–7617.

[12] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed, "Robust Self-Supervised Audio-Visual Speech Recognition," in *Proc. Interspeech*, 2022, pp. 2118–2122.

[13] Valentin Gabeur, Paul Hongsuck Seo, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid, "AVATAR: Unconstrained Audiovisual Speech Recognition," in *Proc. Interspeech*, 2022, pp. 2818–2822.

[14] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid, "AVFormer: Injecting Vision into Frozen Speech Models for Zero-Shot AV-ASR," in *Proc. CVPR*, 2023, pp. 22922–22931.

[15] Golan Pundak, Tara N. Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao, "Deep Context: End-to-end Contextual Speech Recognition," in *Proc. SLT*, 2018, pp. 418–425.

[16] Feng-Ju Chang, Jing Liu, Martin Radfar, Athanasios Mouchtaris, Maurizio Omologo, Ariya Rastrow, and Siegfried Kunzmann, "Context-Aware Transformer Transducer for Speech Recognition," in *Proc. ASRU*, 2021, pp. 503–510.

[17] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[18] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[19] Lantian Li, Xiaolou Li, Haoyu Jiang, Chen Chen, Ruihai Hou, and Dong Wang, "CN-Celeb-AV: A Multi-Genre Audio-Visual Dataset for Person Recognition," in *Proc. Interspeech*, 2023, pp. 2118–2122.

[20] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic, "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips," in *Proc. ICCV*, 2019, pp. 2630–2640.

[21] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze, "How2: A Large-scale Dataset for Multimodal Language Understanding," in *Proc. NIPS*, 2018.

[22] João Miranda, João Paulo Neto, and Alan W Black, "Improving ASR by integrating lecture audio and slides," in *Proc. ICASSP*, 2013, pp. 8131–8135.

[23] Yuya Akita, Yizheng Tong, and Tatsuya Kawahara, "Language model adaptation for academic lectures using character recognition result of presentation slides," in *Proc. ICASSP*, 2015, pp. 5431–5435.

[24] Lin-shan Lee, James Glass, Hung-yi Lee, and Chun-an Chan, "Spoken Content Retrieval—Beyond Cascading Speech Recognition with Text Retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1389–1420, 2015.

[25] James Glass, Timothy J. Hazen, Scott Cyphers, Igor Malioutov, David Huynh, and Regina Barzilay, "Recent progress in the MIT spoken lecture processing project," in *Proc. Interspeech 2007*, 2007, pp. 2553–2556.

[26] Guangzhi Sun, Chao Zhang, and Phil Woodland, "Tree-constrained Pointer Generator with Graph Neural Network Encodings for Contextual Speech Recognition," in *Proc. Interspeech*, 2022, pp. 2043–2047.

[27] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., "The AMI meeting corpus: A pre-announcement," in *Proc. MLMI*, 2005, pp. 28–39.

[28] Guoguo Chen, Shuzhou Chai, and et al, "GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio," in *Proc. Interspeech*, 2021, pp. 3670–3674.

[29] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng, "WENETSPEECH: A 10000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition," in *Proc. ICASSP*, 2022, pp. 6182–6186.

[30] William R Pearson, "Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms," *Genomics*, vol. 11, no. 3, pp. 635–650, 1991.

[31] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai, "Real-Time Scene Text Detection with Differentiable Binarization and Adaptive Scale Fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[32] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang, "Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition," in *Proc. CVPR*, 2021, pp. 7098–7107.

[33] Zhanghui Kuang, Hongbin Sun, Zhizhong Li, Xiaoyu Yue, Tsui Hin Lin, Jianyong Chen, Huaqiang Wei, Yiqin Zhu, Tong Gao, Wenwei Zhang, Kai Chen, Wayne Zhang, and Dahua Lin, "MMOCR: A Comprehensive Toolbox for Text Detection, Recognition and Understanding," in *Proc. MM*, 2021, p. 3791–3794.

[34] Maarten Grootendorst, "KeyBERT: Minimal keyword extraction with BERT.," `https://github.com/MaartenGr/KeyBERT`, 2020.

[35] Kaixun Huang, Ao Zhang, Zhanheng Yang, Pengcheng Guo, Bingshen Mu, Tianyi Xu, and Lei Xie, "Contextualized End-to-End Speech Recognition with Contextual Phrase Prediction Network," in *Proc. Interspeech*, 2023, pp. 4933–4937.

[36] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[37] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[38] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.