INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

An Efficient and Noise-Robust Audiovisual Encoder for Audiovisual Speech Recognition

Zhengyang Li, Chenwei Liang, Timo Lohrenz, Marvin Sach, Björn Möller, Tim Fingscheidt

Technische Universität Braunschweig
Institute for Communications Technology
Schleinitzstr. 22, 38106 Braunschweig, Germany
{zhengyang.li, c.liang, t.lohrenz, m.sach, bjoern.moeller, t.fingscheidt}@tu-braunschweig.de

Abstract

Boosted by self-supervised learning (SSL) on large amounts of unlabeled data, computationally demanding transformer-based audiovisual ASR (AV-ASR) achieves state-of-the-art performance. In this work, we are the first to propose teacher-student model distillation for an *efficient* and *noise-robust* AV encoder for AV-ASR. First, we compare two options for the teacher, a non-task-specific and a task-specific one. Second, we investigate the design and the components in the student neural network. Third, we explore loss function choices during distillation. By distillation with a simplified loss function, the final efficient conformer-based student has 69% fewer parameters and 23% less computational power than the teacher, but excels the baseline student with a WER of 4.6% (11.4%) in clean condition, and with 20.2% (35.7%) in 0dB babble noise. On average over noise types in 0dB SNR, our proposed student even achieves more than 50% relative WER reduction compared to the baseline student.

Index Terms: audiovisual speech recognition, efficient and robust networks, teacher/student model distillation, transformer, conformer

1. Introduction

Audiovisual speech recognition (AV-ASR) utilizes the movements of the speaker's lips and mouth region as compensation for acoustic information to recognize the spoken utterances. Compared to pure acoustic ASR, AV-ASR has shown its superior performance in acoustically noisy environments [1, 2, 3]. The robustness of AV-ASR enables its application in smartphones or cars in noisy or multi-talker environments.

AV-ASR models usually comprise an AV encoder and an autoregressive decoder to predict the output token sequence. The AV encoder extracts the audio and video features by an audiovisual frontend (red block of Fig. 1) and models the temporal dependencies of these two modalities by the sequence modeling architecture (encoder blocks with dark green background in Fig. 1). The sequence modeling architecture of an AV encoder has experienced a paradigm shift from recurrent neural networks [4, 5] to all-attention-based transformers [6], building upon the success of transformers in neural machine translation [7] and acoustic ASR tasks [8, 9, 10]. A conformer [11], a variant of the transformer designed specifically for ASR, has also been employed in AV-ASR [2] to improve performance. For a better initialization of AV encoders, pre-training approaches have been investigated. Afouras et al. [6] pre-trained a visual CNN frontend in the encoder on the non-public MV-LRS dataset [12]. Ma et al. [2] pre-trained the entire encoder with an isolated word classification task on the Lip Reading in the Wild (LRW) dataset [13]. Recently, Shi et al. [14] applied self-supervised learning (SSL) to pre-train the AV encoder on a large amount of unlabeled audiovisual data. The resulting audiovisual hidden unit BERT (AV-HuBERT) model achieves state-of-the-art performance on the Lip Reading Sentences 3 (LRS3) audiovisual speech recognition task [15].

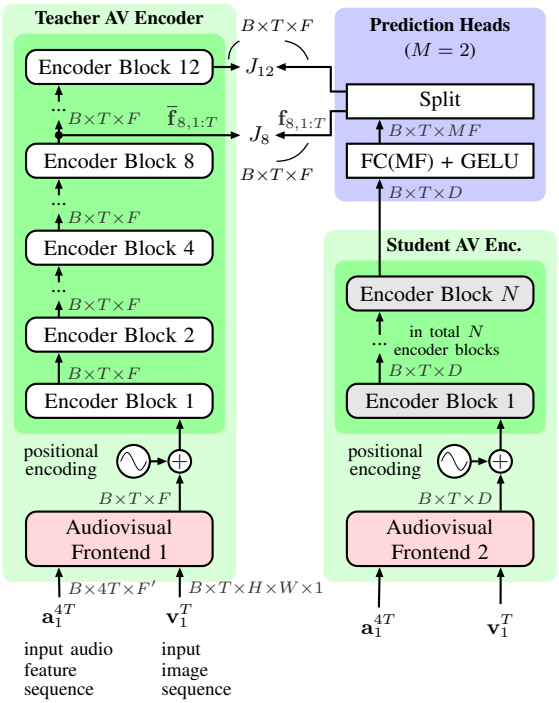

Figure 1: ***Teacher-student architecture*** *proposed in this work. The $M{=}|\mathcal{L}|{=}2$ prediction head outputs and the encoder block 8 and 12 outputs are used in (2) to compute J_ℓ, $\ell \in \mathcal{L} = \{8, 12\}$.*

Along with the performance improvement by the SSL pre-trained AV encoder, the memory requirements and computational complexity are drastically increased. Most research to compress the oversized SSL pre-trained models is conducted in language modeling [16, 17] and ASR based on acoustic input [18, 19, 20, 21, 22]. A common approach to improve the model efficiency is the model distillation with a teacher-student neural network [19, 20, 21], where a small student model learns the targets generated by the pre-trained large teacher model. The first work to distill an SSL pre-trained model in acoustic ASR is DistilHuBERT from Chang et al. [19]. Wang et al. [21] and Lee et al. [20] improved the performance of the distilled model with different student model designs. These works are evaluated in different downstream tasks on the speech processing universal performance benchmark (SUPERB) [23], where the

10.21437/Interspeech.2023-793

ASR task is assessed on the clean datasets of Librispeech [24]. A major difference between AV-ASR and ASR based on acoustics is that AV-ASR demands a visual frontend to process the video information. The `ResNet`-based visual frontends applied in prominent AV encoders [1, 2, 14] have a small size but an extremely high computation cost. Efficient visual frontends have so far only been explored by Ma et al. [25] on the LRW isolated word classification task.

For continuous audiovisual speech recognition, efficient visual frontends have not been explored. To the best of our knowledge, model distillation to obtain an efficient AV encoder has so far not been applied to AV-ASR. In addition, only a few recent works in acoustic ASR explored the robustness of the distilled student against noise [26]. Teacher/student architectures which are robust against noise are insufficiently investigated as concerns the AV encoder.

In this work, we therefore propose a teacher-student framework for efficient AV encoders and evaluate their performance on AV-ASR tasks. To improve the noise robustness of distilled AV encoders, we first generate student learning targets specific to AV-ASR tasks by utilizing a finetuned teacher. Second, we increase the depth of student models and simplify the loss terms to learn more linguistic features, which are proven more profitable for speech recognition tasks [27, 28]. Third, we build the student with advanced conformer models. To improve the efficiency of distilled AV encoders, we apply a light-weight `ShuffleNetv2`-based visual frontend to significantly reduce the computational complexity.

The paper is structured as follows: In Section 2, we introduce our proposed methods. Section 3 gives the experimental setup, training details, and then results and discussion on the LRS3 AV-ASR task. The paper is concluded in Section 4.

2. Methods

2.1. Our Investigated Teacher-Student Framework

The proposed teacher-student framework for the audiovisual (AV) encoder is shown in Fig. 1. The teacher (light green on the left) and the student (light green on the right) use the same image sequence $\mathbf{v}_1^T{=}(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_T)$ and audio feature sequence $\mathbf{a}_1^{4T}{=}(\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_{4T})$ as input to an audiovisual frontend followed by positional encoding. Note that in our case the frame rate is 25 Hz (video) and 100 Hz (audio), causing the fourfold length $4T$ of the audio feature sequence. Both, teacher and student models are transformer encoder architectures, with the latter comprising less parameters and reduced computational complexity for improved efficiency by design. The M prediction heads in the blue block project the student's output to M vectors, where each vector of length F is trained to match the learning targets from the outputs of the teacher's encoder blocks. As a *framework baseline*, we employ the student model design and distillation settings from `DistilHuBERT` [19].

Baseline teacher model: As the *baseline teacher* model, the `base` configuration of the audiovisual hidden unit BERT (`AV-HuBERT` [1]) is chosen, which is pre-trained on noise-augmented and unlabeled audiovisual data from the Voxceleb2 [29] and LRS3 datasets [15]. The `base AV-HuBERT` uses a `ResNet`-based audiovisual frontend and a total of 12 transformer blocks each having frozen parameters during distillation. This baseline teacher encoder is not finetuned on any specific task, so it produces general audiovisual representations.

Baseline student model: As a *baseline student* model, in accordance with the student design employed by `DistilBERT` [16] for language modeling and `DistilHuBERT` [19] for speech tasks, we use the same audiovisual frontend and the encoder block design as the teacher model, with the exception of having a reduced number of encoder blocks ($N = 2$). The student is initialized with the audiovisual frontend and the first two encoder blocks of the teacher.

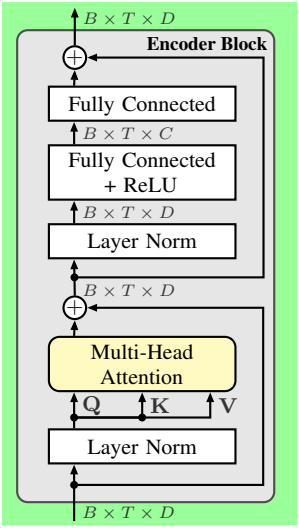

Figure 2: *Single **transformer encoder block** during training with a self multi-head attention (yellow block). The first fully connected layer projects the feature dimension from D to the expanded feature dimension $C > D$.*

Distillation: For distillation, the baseline framework uses $M = 3$ prediction heads. The learning targets are outputs from the 4th, 8th, and 12th encoder blocks of the teacher. The according loss function is

$$J = \sum_{\ell \in \mathcal{L}} J_\ell \tag{1}$$

with $\ell \in \mathcal{L} = \{4, 8, 12\}$ indexing those teacher encoder blocks that contribute to the loss. The loss term in layer ℓ is given by

$$\begin{aligned} J_\ell &= \lambda J_\ell^{\cos} + J_\ell^{L1} \\ &= -\lambda \sum_{t \in \mathcal{T}} \log \sigma \left(\cos \left(\mathbf{f}_{\ell,t}, \bar{\mathbf{f}}_{\ell,t} \right) \right) + \sum_{t \in \mathcal{T}} \frac{1}{D} \| \mathbf{f}_{\ell,t} - \bar{\mathbf{f}}_{\ell,t} \|_1 \end{aligned} \tag{2}$$

and consists of a cosine similarity $-J_\ell^{\cos}$ and an $L1$ loss J_ℓ^{L1} with some hyperparameter λ. The F-dimensional feature vector produced by the teacher's ℓ-th encoder block at time step $t \in \mathcal{T}{=}(1, 2, ..., T)$ is denoted as $\bar{\mathbf{f}}_{\ell,t}$, while the entire sequence is $\bar{\mathbf{f}}_{\ell,1:T} = (\bar{\mathbf{f}}_{\ell,1}, \bar{\mathbf{f}}_{\ell,2}, ..., \bar{\mathbf{f}}_{\ell,T})$. The matching F-dimensional feature vector at time step t processed by the student model and the prediction heads is $\mathbf{f}_{\ell,t}$ taken from sequence $\mathbf{f}_{\ell,1:T}$. Minimizing the loss (1) during training aims at maximizing the cosine similarity and reducing the $L1$ loss simultaneously.

2.2. Proposed Teacher and Distillation Loss

Task-specific teacher: Models pre-trained by SSL are capable of extracting general representations that are suitable for a wide range of downstream tasks. `AV-HuBERT` models also demonstrated their utility in various audiovisual tasks such as AV-ASR [1, 14], automatic lip-reading [30, 31], speaker verification [32], and audiovisual speech enhancement [33]. In recent works on model distillation [19, 20, 21, 26], SSL pre-trained models are used directly as teacher, but the choice of the teacher is insufficiently investigated. Here, striving to obtain *audiovisual* representations from the teacher that are more specific for the

Table 1: *WER (%) on the LRS3 test set. Models are evaluated with* ***clean*** *speech, at an SNR of* ***0dB babble*** *noise, and at an SNR of* ***0dB second interfering talker (talker)****. The student has N encoder blocks as depicted in Fig. 1. Parameter $D < F$ is the feature dimension and $C > D$ is the expanded feature dimension as shown in Fig. 2. Best student results are in* ***bold*** *font, second best are* <u>underlined</u>*.*

Method	Teacher	Student						Distillation		#params in the student	#FLOPs/ frame	WER (%)		
	Task-specific	Init. by teacher	Visual frontend (FE)	Transformer/Conformer				M	Learning target			0 dB babble	0 dB talker	clean
				Type	N	D	C							
`base AV-HuBERT` [1] teacher			`ResNet18`	Transf.	12	768	3072	3	4,8,12	103M	875M	6.3	4.0	1.9
① Baseline student	✗	✓	`ResNet18`	Transf.	2	768	3072	3	4,8,12	32M	734M	35.7	41.4	11.4
② + Task-specific teacher	✓	✓	`ResNet18`	Transf.	2	768	3072	3	4,8,12	32M	734M	33.7	37.7	10.6
③ - Student initialization	✓	✗	`ResNet18`	Transf.	2	768	3072	3	4,8,12	32M	734M	25.5	27.2	7.1
④ + Thin and deep student	✓	✗	`ResNet18`	Transf.	6	384	3200	3	4,8,12	31M	678M	<u>20.4</u>	<u>19.0</u>	5.0
⑤ + Simplified loss	✓	✗	`ResNet18`	Transf.	6	384	3200	2	8,12	31M	678M	<u>20.4</u>	<u>19.0</u>	**4.6**
⑥ + Conformer student	✓	✗	`ResNet18`	Conf.	6	384	1536	2	8,12	32M	674M	**20.2**	**17.5**	**4.6**
⑦ + Light-weight visual FE	✓	✗	`ShuffleNetv2`	Conf.	6	384	1536	2	8,12	22M	107M	22.9	21.1	<u>4.8</u>

AV-ASR task, we finetune the `AV-HuBERT` model on labeled training data from the LRS3 AV-ASR task and use the finetuned AV encoder as the *task-specific* teacher in our approach.

Simplified loss: In the baseline framework, the targets for the student are generated by the set of teacher encoder block layers $\mathcal{L} = \{4, 8, 12\}$. The linguistic features in the later layers of the encoder have more influence on the ASR performance. This motivates us to simplify the loss function by learning the targets only from layers $\mathcal{L} = \{8, 12\}$, see Fig. 1.

2.3. Proposed Student

Student initialization: An appropriate model initialization can often result in improved performance. A recent study about the interpretability of transformer-based ASR [28] found that the early transformer layers tend to extract acoustic features, which is beneficial for speaker identification tasks, while later layers are responsible for extracting phonetic information, which is more important for ASR. For this reason, the weights of the teacher's early layers may not be optimal starting weights for the student. Accordingly, we train the student model from scratch during distillation and examine the impact of student initialization.

Thin and deep student: The trade-off between width and depth in deep learning has been widely discussed [34, 35]. In recent transformer-based acoustic ASR, a deeper neural network has been demonstrated to be effective [20, 36]. Our chosen distillation baseline may not have sufficient depth to effectively capture various patterns and model the interactions between video and audio modalities, as the student only uses $N = 2$ transformer encoder blocks. In this work, we propose to reduce the feature dimension from F to $D < F$ and to adjust the feature dimension $C > D$ (depicted in Fig. 2) to build a thinner but deeper student.

Conformer student: The conformer [11] adds a convolutional block after the multi-head self-attention (MHSA) employed in the encoder blocks. The conformer focuses more on local information and has been used effectively for ASR [11] and also AV-ASR [2] tasks. Compared to the student model in the baseline framework, we replace the transformer encoder blocks with conformer layers to improve the AV-ASR performance.

Light-weight visual frontend (FE): In the AV encoder, the visual frontend based on `ResNet` comprises much fewer parameters than the subsequent transformer, but its computational load is extremely large due to convolutional operations. We substitute the `ResNet` in the audiovisual frontend of the student with a light-weight `ShuffleNetv2` [37] to further reduce the memory and computational cost. We modify the original `ShuffleNetv2` architecture by substituting the first 2D convolutional layer by a 3D convolutional layer with kernel size $5 \times 5 \times 7$ to incorporate the additional temporal dimension, and reduce the number of output channels in the last convolutional layer from 1024 to 512.

3. Evaluation and Discussion

3.1. Experimental Setup and Training Details

Databases and pre-processing: We evaluate our models on the Lip Reading Sentences 3 (LRS3) audiovisual speech recognition task. The LRS3 dataset is the largest publicly available labeled AV-ASR dataset, which includes 433 h of labeled audiovisual training data collected from TED and TEDx talks [15]. The video frames and raw speech signal have a sample rate of 25 Hz and 16 kHz, respectively. We follow the pre-processing pipeline of the LRS3 dataset detailed in [14]. As input audio features we use 26-dimensional log-filterbank outputs, which are extracted with a 25 ms window and a frame shift of 10 ms, resulting in 100 audio frames per second. Video frames are converted to grayscale and cropped to a 96×96 region of interest based on the face alignment.

Model distillation: Based on the `PyTorch`-based `s3prl` speech toolkit [23], we implement the teacher-student model distillation for AV encoders. During model distillation, we use a batch size of 4 with an accumulate gradients of six batches to simulate the batch size $B = 24$. The teacher-student model is trained on a single `Nvidia A100` GPU for 100k updates. The learning rate is linearly increased to 0.0002 in the first 14% of updates, then linearly decreased to 0. The weight of the cosine similarity in loss term (2) is $\lambda = 1$ for all experiments.

Fine-tuning for AV-ASR: For a fair comparison, we report the same decoder architecture as the baseline method [1] to finetune the models for all experiments. The decoder network consists of six transformer decoder blocks (cf. Fig. 2 in [10]) with 57M parameters. The outputs of the encoder-decoder architecture are subword tokens generated by `SentencePiece` [38] with a vocabulary size of 1000. The finetuning process is done using the `PyTorch`-based `fairseq` toolkit. We finetune the entire encoder-decoder model for 60k updates. The learning rate is linearly increased to 0.001 for the first 30% of updates, then linearly decreased to 0. The encoder is frozen for the first 48k updates. We apply the same data augmentation as in `AV-HuBERT` [1], where 25% of the training data is augmented with 0dB noise chosen from babble, music, natural noise, and second interfering talker condition. There is no speaker overlap in babble noise and second interfering talker condition among different splits.

Evaluation in noise environments: To add noise to our speech data, we follow the exact same procedure as detailed in [1]. First, we generate *babble noise* by mixing utterances from 30 different speakers from the MUSAN dataset [39] where each speaker is used exclusively for either the training, validation, or the test partition. We also evaluate our approaches for speech with a *second interfering talker* from LRS3 data following [1] for comparability reasons. To accurately evaluate the noise robustness of models, we report the average word error rate (WER) based on

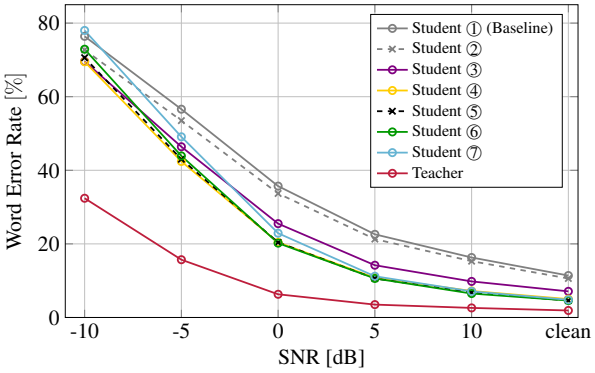

Figure 3: *WER (%) on the LRS3 test set at different SNRs in* ***babble noise****. Student identifiers ⓘ from Table 1.*

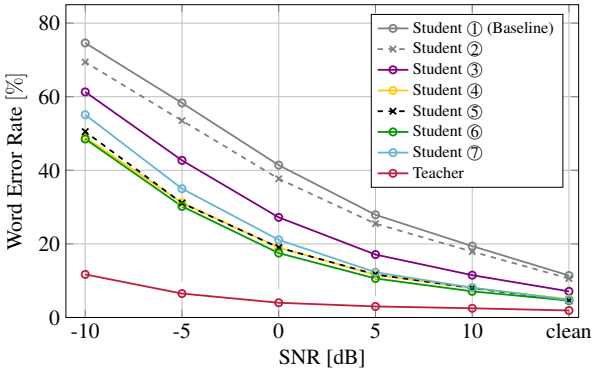

Figure 4: *WER (%) on the LRS3 test set at different SNRs in* ***second interfering talker****. Student identifiers ⓘ from Table 1.*

ten inference passes, each using different random noise snippets from the specific noise type (i.e., babble or interfering talker) and signal-to-noise ratios (SNRs).

3.2. Results and Discussion

The results of our experiments on compressing the audiovisual encoder and reducing its computational complexity are presented in Table 1. The WER is measured in three conditions: clean speech, speech mixed with 0 dB babble noise, and speech with a second interfering talker at 0 dB. Note that we follow the common practice on the LRS3 task to report performance only on test data. However, we observed that the results on the validation partition follow the same trend as on the test partition.

The teacher model (`base AV-HuBERT` [1]), reported in the first row has a total of 103M parameters and requires 875M floating point operations per frame (#FLOPs/frame). The results of the student models are displayed starting from the second row. The baseline student ① is much more efficient, however, its performance clearly deteriorates on clean speech and even more in noisy environments. We applied our methods *incrementally* while constraining the #params in the student model to approximately 32M. First, a task-specific teacher ② during distillation brings a slight improvement both on noisy and clean speech. Second, training of the student from scratch ③ and a deeper student ④ result in substantial improvements across all test conditions (e.g., -3.5% absolute WER and -2.1% absolute WER on clean speech, respectively). Third, our simplified loss ⑤ gives another -0.4% absolute WER on clean speech. The application of the conformer encoder blocks ⑥ results in our *best performing student model across all conditions and demands 69% fewer parameters and 23% less #FLOPs/frame compared to the teacher model*. It excels the baseline student with a WER of 4.6% (11.4%) in clean condition, with 20.2% (35.7%) in 0dB babble noise, and with 17.5% (41.4%) in 0dB interfering talker. *On average over noise types in 0dB SNR, our proposed student achieves more than 50% relative WER reduction (in clean even 60% relative WER reduction) compared to the baseline student.* Finally, a light-weight visual frontend ⑦ reduces the student parameters and #FLOPs/frame to 21% and 12%, respectively, when compared to the teacher model, with only a small performance loss compared to our best (in terms of WER) performing student model ⑥.

Fig. 3 depicts the WERs of all approaches presented in Table 1 assessed on different SNRs of babble noise. Expectedly, the teacher model exhibits the best WER performance across all SNR levels. Compared to the teacher, the student models show more degradation of the WER performance with decreasing SNR. The three deeper student models ④⑤⑥ have similar performance across all SNRs. In the extreme challenging -5dB and -10dB babble noise conditions, all student AV encoders exceed a WER of 40%. Above 0dB SNR, the deeper student AV encoders (④-⑦) show a clear improvement compared to the baseline student ①. Computing the average WER over all SNRs in babble noise, the student model ⑥ demonstrates with 26.5% a significantly lower WER as the baseline student ① with 36.5%.

Furthermore, the performance of the models from Table 1 is evaluated with a second interfering talker, as with different SNR, as shown in Fig. 4. The teacher model performs the best and achieves an impressive WER of 11.7% at -10dB SNR. Our best performing student ⑥ from Table 1 also has the lowest WERs across all evaluated SNRs compared to the other student models. While the performance of a purely acoustic ASR system is typically challenged in adverse conditions such as the presence of interfering talkers, our best distilled student model ⑥ achieves good WERs ranging from 4.6% to 17.5% (SNR of 0dB) in such a condition. Computing the average WER over all SNRs for an interfering talker, the student model ⑥ achieves 19.8%, whereas the baseline student ① has an almost doubled WER of 38.8%.

4. Conclusions

In this work, we presented a teacher-student distillation framework for developing efficient audiovisual encoder models. We investigated various optimization strategies within this framework to enhance the efficiency and noise robustness of the distilled AV encoder in audiovisual speech recognition tasks. Our best performing student model has 69% fewer parameters and 23% less computational power compared to the teacher, but excels the baseline student with a WER of 4.6% (11.4%) in clean condition, with 20.2% (35.7%) in 0dB babble noise, and with 17.5% (41.4%) in 0dB interfering talker condition. On average over noise types, this amounts to more than 50% relative WER reduction among the students at an SNR of 0dB. Our work enables the deployment of noise-robust audiovisual speech recognition systems on resource-constrained devices, and our distilled encoder models can be easily applied to other audiovisual tasks.

5. Acknowledgments

The research leading to these results has received funding from the Bundesministerium für Wirtschaft und Klimaschutz (BMWK) under funding code 01MK20011T (SPEAKER project).

6. References

[1] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, " Robust Self-Supervised Audio-Visual Speech Recognition," *arXiv:2201.02184*, Jul. 2022.

[2] P. Ma, S. Petridis, and M. Pantic, "End-To-End Audio-Visual Speech Recognition With Conformers," in *Proc. of ICASSP*, Toronto, ON, Canada, Jun. 2021, pp. 7613–7617.

[3] S. Receveur, R. Weiss, and T. Fingscheidt, "Turbo Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 846–862, May 2016.

[4] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," in *Proc. of CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 3444–3453.

[5] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-End Audiovisual Speech Recognition," in *Proc. of ICASSP*, Seoul, South Korea, Apr. 2018, pp. 6548–6552.

[6] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep Audio-Visual Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–11, Dec. 2018, (early access).

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv:1706.03762*, Dec. 2017.

[8] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," in *Proc. of ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 5884–5888.

[9] T. Lohrenz, Z. Li, and T. Fingscheidt, "Multi-Encoder Learning and Stream Fusion for Transformer-Based End-to-End Automatic Speech Recognition," in *Proc. of Interspeech*, Brno, Czech Republic, Sep. 2021, pp. 2846–2850.

[10] T. Lohrenz, P. Schwarz, Z. Li, and T. Fingscheidt, "Relaxed Attention: A Simple Method to Boost Performance of End-to-End Automatic Speech Recognition," in *Proc. of ASRU*, Cartagena, Colombia, Dec. 2021, pp. 177–184.

[11] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-Augmented Transformer for Speech Recognition," in *Proc. of Interspeech*, Shanghai, China, Oct. 2020, pp. 5036–5040.

[12] J. S. Chung and A. Zisserman, "Lip Reading in Profile," in *Proc. of BMVC*, London, UK, September 2017, pp. 155.1–155.11.

[13] J. S. v. Chung and A. Zisserman, "Lip Reading in the Wild," in *Proc. of ACCV*, Taipei, Taiwan, Nov. 2016, pp. 87–103.

[14] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction," in *Proc. of ICLR*, virtual, Apr. 2022, pp. 1–24.

[15] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition," *arXiv:1809.00496*, Oct. 2018.

[16] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," *arXiv:1910.01108*, Mar. 2019.

[17] Y. Lee, K. Jang, J. Goo, Y. Jung, and H. Kim, "TinyBERT: Distilling BERT for Natural Language Understanding," *arXiv:1909.10351*, Oct. 2020.

[18] C.-I. J. Lai, Y. Zhang, A. H. Liu, S. Chang, Y.-L. Liao, Y.-S. Chuang, K. Qian, S. Khurana, D. Cox, and J. Glass, "Parp: Prune, Adjust and Re-prune for Self-Supervised Speech Recognition," in *Proc. of NIPS*, virtual, Dec. 2021, pp. 21 256–21 272.

[19] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "DistilHuBERT: Speech Representation Learning by Layer-Wise Distillation of Hidden-Unit BERT," in *Proc. of ICASSP*, Singapore, May 2022, pp. 7087–7091.

[20] Y. Lee, K. Jang, J. Goo, Y. Jung, and H. Kim, "FitHuBERT: Going Thinner and Deeper for Knowledge Distillation of Speech Self-Supervised Learning," *arXiv:2207.00555*, Jul. 2022.

[21] R. Wang, Q. Bai, J. Ao, L. Zhou, Z. Xiong, Z. Wei, Y. Zhang, T. Ko, and H. Li, "LightHuBERT: Lightweight and Configurable Speech Representation Learning with Once-for-All Hidden-Unit BERT," *arXiv:2203.15610*, Jun. 2022.

[22] T. Ashihara, T. Moriya, K. Matsuura, and T. Tanaka, "Deep Versus Wide: An Analysis of Student Architectures for Task-Agnostic Knowledge Distillation of Self-Supervised Speech Models," in *Proc. of Interspeech*, Incheon, Korea, Sep. 2022, pp. 411–415.

[23] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "SUPERB: Speech Processing Universal Performance Benchmark," in *Proc. of Interspeech*, Brno, Czech Republic, Sep. 2021, pp. 1194–1198.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," in *Proc. of ICASSP*, South Brisbane, QLD, Australia, Apr. 2015, pp. 5206–5210.

[25] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards Practical Lipreading With Distilled and Efficient Models," in *Proc. of ICASSP*, Toronto, ON, Canada, Jun. 2021, pp. 7608–7612.

[26] K.-P. Huang, Y.-K. Fu, T.-Y. Hsu, F. R. Gutierrez, F.-L. Wang, L.-H. Tseng, Y. Zhang, and H.-Y. Lee, "Improving Generalizability of Distilled Self-Supervised Speech Processing Models under Distorted Settings," in *Proc. of SLT*, Doha, Qatar, Jan. 2023, pp. 1112–1119.

[27] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-Wise Analysis of a Self-Supervised Speech Representation Model," in *Proc. of ASRU*, Cartagena, Colombia, Dec. 2021, pp. 914–921.

[28] K. Shim, J. Choi, and W. Sung, "Understanding the Role of Self Attention for Efficient Speech Recognition," in *Proc. of ICLR*, virtual, Apr. 2022, pp. 1–19.

[29] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. of Interspeech*, Hyderabad, India, Sep. 2018, pp. 1086–1090.

[30] T. Lohrenz, B. Möller, Z. Li, and T. Fingscheidt, "Relaxed Attention for Transformer Models," *arXiv:2209.09735*, Sep. 2022.

[31] Z. Li, T. Lohrenz, M. Dunkelberg, and T. Fingscheidt, "Transformer-Based Lip-Reading with Regularized Dropout and Relaxed Attention," in *Proc. of SLT*, Doha, Qatar, Jan. 2023, pp. 723–730.

[32] B. Shi, A. Mohamed, and W.-N. Hsu, "Learning Lip-Based Audio-Visual Speaker Embeddings with AV-HuBERT," in *Proc. of Interspeech*, Incheon, Korea, Sep. 2022, pp. 4785–4789.

[33] W.-N. Hsu, T. Remez, B. Shi, J. Donley, and Y. Adi, "ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement ," *arXiv:2212.11377*, Dec. 2022.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[35] A. Goyal, A. Bochkovskiy, J. Deng, and V. Koltun, "Non-Deep Networks," *arXiv:2110.07641*, Oct. 2021.

[36] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Müller, S. Stüker, and A. Waibel, "Very Deep Self-Attention Networks for End-to-End Speech Recognition," in *Proc. of Interspeech*, Graz, Austria, Sep. 2019, pp. 66–70.

[37] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," in *Proc. of ECCV*, Munich, Germany, Sep. 2018, pp. 116–131.

[38] T. Kudo and J. Richardson, "SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing," *arXiv:1808.06226*, Aug. 2018.

[39] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.0848v1*, pp. 1–4, Oct. 2015.