



Efficient Noise-Robust Hybrid Audiovisual Encoder with Joint Distillation and Pruning for Audiovisual Speech Recognition

Zhengyang Li¹, Pascal Reichert¹, Thomas Graave¹, Patrick Blumenberg¹, Tim Fingscheidt¹

¹Institute for Communications Technology, Technische Universität Braunschweig,
38106 Braunschweig, Germany

{zhengyang.li, pascal.reichert, thomas.graave, p.blumenberg, t.fingscheidt}@tu-bs.de

Abstract

Powered by self-supervised learning (SSL) on vast amounts of unlabeled data, a computationally intensive audiovisual encoder—a hybrid architecture combining ResNet and transformer in series—achieves state-of-the-art performance in audiovisual speech recognition (AV-ASR). In this work, we are the first to apply joint distillation and pruning (DP) with a teacher-student model for an efficient and noise-robust audiovisual encoder. First, we compress the transformer of the AV encoder. Second, we extend joint DP to both the ResNet and transformer of the hybrid AV encoder. In addition, we provide analyses on the teacher and the final student, respectively. With a similar number of parameters, our proposed student outperforms the previous state-of-the-art in clean condition (word error rate of 3.1% vs. 4.6%) and across all noisy conditions, while at the same time reducing computational complexity by 31.8%. Our code is at [GitHub](https://github.com/ifnspaml)¹.

Index Terms: audiovisual speech recognition, pruning, distillation, efficient and robust networks

1. Introduction

Since the psychological finding that speech perception is inherently multimodal [1, 2], researchers have been developing audiovisual speech recognition (AV-ASR) systems, which recognize the spoken utterance by incorporating speakers’ lip movements to complement the speech modality [3, 4, 5, 6, 7]. Compared to ASR based on acoustics, AV-ASR has demonstrated superior performance in noisy and multi-speaker conditions [8, 6, 5]. The noise robustness of AV-ASR systems facilitates their deployment in smart home devices [9] and automobiles [10].

Recent AV-ASR systems usually comprise an encoder (light green background color in Fig. 1) and an autoregressive decoder [6, 8, 11]. The encoder is generally a hybrid architecture containing two serially connected components: The first is an audiovisual frontend (red blocks in Fig. 1) responsible for feature extraction, where convolutional neural networks (CNNs) [4, 12, 8, 7] are widely used to process video inputs due to their parameter efficiency in vision tasks compared to vision transformers [13, 14]. The second component consists of serially stacked encoder blocks (dark green blocks in Fig. 1) for modeling temporal dependencies, which has undergone a paradigm shift as in natural language processing (NLP) from CNNs [15] and recurrent neural networks (RNNs) [16] to transformers [17, 6, 18, 19]. Hybrid AV encoders built by CNNs and transformers are usually pre-trained either separately [4] or jointly [6, 18] to achieve better initialization for fine-tuning on AV-ASR tasks. The recent audiovisual hidden unit BERT (AV-HuBERT) [18] is a

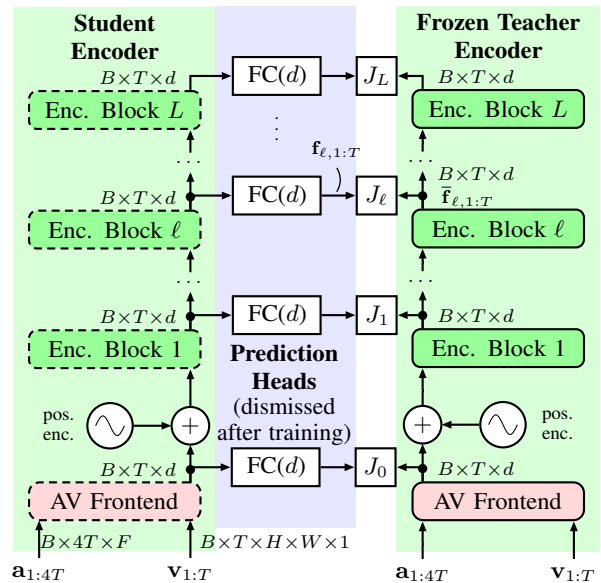


Figure 1: *Teacher-student architecture proposed in this work. The student is initialized by the teacher. The dashed blocks in the student have prunable modules shown in Table 1 in bold.*

hybrid AV encoder pre-trained jointly by self-supervised learning (SSL) on 1,759 hours of unlabeled audiovisual data, resulting in an AV encoder with 103M (AV-HuBERT base) or 325M (AV-HuBERT large) parameters. AV-HuBERT achieves state-of-the-art performance combined with transformer decoders on AV-ASR tasks [8, 20, 21]. However, such computationally heavy systems fail to meet the energy and memory efficiency requirements of edge devices.

Compression of SSL-pretrained models has been explored in NLP [22, 23, 24], ASR [25, 26, 27], and AV-ASR [7]. Knowledge distillation by a teacher-student architecture is one of the most commonly employed methods [22, 23, 25, 26]. In ASR tasks, increasing the depth and reducing the feature dimensions of transformers lead to deeper and thinner student models, which outperform shallower students [28]. This improvement is attributed to more layers (network depth) extracting more linguistic information which is beneficial for speech recognition [29, 11]. However, the student architecture must be manually designed and fixed during distillation. Structured pruning is another widely used compression method in SSL-pretrained models [24, 27, 30], allowing an encoder to learn an optimal architecture. However, only transformers are typically pruned [24]. Peng et al. [27, 30] utilized joint distillation and pruning to ASR based on acoustics. In AV-ASR, the visual fron-

¹<https://github.com/ifnspaml/Efficient-AV-ASR>

tend, often based on ResNet for video processing, has a small amount of parameters but usually accounts for the majority of computational complexity in an AV encoder. Li et al. [7] applied teacher-student distillation for an AV encoder and achieved state-of-the-art performance in the respective model size regime. However, replacing the pre-trained ResNet-based visual frontend with a small ShuffleNetv2-based frontend results in performance degradation.

In this work, we propose joint distillation and pruning (DP) within a teacher-student framework to develop efficient AV encoders and evaluate their performance on AV-ASR tasks. First, we perform an analysis of the hybrid teacher AV encoder, revealing an unbalanced distribution of computational complexity and a parameter count in the range between the ResNet-based visual frontend and the transformer blocks. Second, we apply joint DP to transformer blocks in the AV encoder. Third, we extend joint DP to both the visual frontend and the transformer blocks in the hybrid AV encoder, demonstrating that computational complexity can be significantly reduced while improving performance and robustness. In addition, we present the learned student encoder to highlight the importance of different layers.

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

2. Joint Distillation and Pruning for the AV Encoder

In Fig. 1, we apply joint distillation and pruning (DP) with our proposed teacher-student framework for the audiovisual (AV) encoder. The frozen teacher and the student take the same image sequence $\mathbf{v}_{1:T}=(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_T)$ and audio feature sequence $\mathbf{a}_{1:4T}=(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{4T})$ as input. In our case, the frame rate is 25 Hz (video) and 100 Hz (audio), causing the fourfold length $4T$ of the audio feature sequence. The teacher encoder is pre-trained and frozen. The student encoder is initialized by the teacher encoder with the same architecture and weights. The dashed blocks have prunable *modules*, e.g., a convolutional channel, a single attention head, or the intermediate units in feed-forward networks of the transformer encoder blocks. The distillation occurs layer-to-layer with up to L fully connected layers as shown in the purple background, which is explained in the following. **Model distillation:** As shown in Fig. 1, for layer-to-layer distillation, the projected intermediate feature sequence $\mathbf{f}_{\ell,1:T}=(\mathbf{f}_{\ell,t})$ with $\mathbf{f}_{\ell,t} \in \mathbb{R}^d$ from certain student encoder blocks is learned to match the learning targets from the teacher encoder blocks $\bar{\mathbf{f}}_{\ell,1:T}=(\bar{\mathbf{f}}_{\ell,t})$. The distillation loss of an utterance with the video length of T frames is $(\cdot)^\top$ denotes the vector transpose)

$$\begin{aligned} J^{\text{dist}} &= \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} (J_\ell^{\text{cs}} + \beta J_\ell^{L1}) \\ &= \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \left(- \sum_{t \in \mathcal{T}} \log \sigma \left(\frac{\mathbf{f}_{\ell,t}^\top \cdot \bar{\mathbf{f}}_{\ell,t}}{\|\mathbf{f}_{\ell,t}\| \cdot \|\bar{\mathbf{f}}_{\ell,t}\|} \right) \right. \\ &\quad \left. + \beta \sum_{t \in \mathcal{T}} \frac{1}{d} \|\mathbf{f}_{\ell,t} - \bar{\mathbf{f}}_{\ell,t}\|_1 \right) \end{aligned} \quad (1)$$

with $\ell \in \mathcal{L} = \{0, 4, 8, 12\}$ indexing those encoder blocks contributing to the loss. The loss term in block ℓ consists of a cosine similarity loss J_ℓ^{cs} and an L_1 loss J_ℓ^{L1} scaled by a hyperparameter $\beta = 0.1$ for all experiments. The sigmoid function $\sigma(\cdot)$ is applied to the cosine similarity to stabilize the training.

Joint distillation and pruning through L_0 regularization: The student encoder is parameterized by a set Θ with trainable

modules as its elements. Among these, the student encoder includes prunable modules $\theta_m \in \Theta$ indexed by $m \in \mathcal{M} = \{1, 2, \dots, M\}$, where M is the total number of prunable modules. We prune each module θ_m by applying a binary mask $z_m \in \{0, 1\}$, resulting in pruned modules $\hat{\theta}_m = z_m \cdot \theta_m$. The set of non-zero modules forms a subset $\hat{\Theta} \subseteq \Theta$. The count of non-zero elements in the set of pruned modules $\hat{\Theta}$ is represented as an L_0 term $\lambda \|\hat{\Theta}\|_0$ weighted by a hyperparameter $\lambda \in \mathbb{R}^+$, which can be added to the loss function (1) to penalize non-zero parameters and to promote a sparse model.

However, the discrete binary masks $z_m \in \{0, 1\}$ are not differentiable and cannot be optimized jointly with the student's modules Θ . To make the masks z_m differentiable, Louizos et al. proposed masks $z_m \in [0, 1]$ following a hard concrete distribution $z_m \sim \text{p}(z_m; \alpha_m)$ with a trainable parameter α_m [31]. We denote all masks as a set $\mathcal{Z} = \{z_m\}_{m=1}^M$ and all trainable parameters of the mask distributions as a set $\mathcal{A} = \{\alpha_m\}_{m=1}^M$. To control sparsity during pruning, Wang et al. [24] applied the augmented Lagrangian method, reformulating the objective function as an adversarial game. For detailed derivation, please refer to [31, 24]. Here we give the final objective function:

$$\begin{aligned} \max_{\lambda_1, \lambda_2} \min_{\Theta, \mathcal{A}} \mathbb{E}_{\mathcal{Z} \sim \text{p}(\mathcal{Z}; \mathcal{A})} & \left(\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{a}_{1:4T}, \mathbf{v}_{1:T}) \in \mathcal{D}} J^{\text{dist}} \right) \\ & + \lambda_1 \cdot (\rho(\mathcal{A}) - \tau) + \lambda_2 \cdot (\rho(\mathcal{A}) - \tau)^2, \\ & \text{with } \mathcal{Z} \sim \text{p}(\mathcal{Z}; \mathcal{A}) : z_m \sim \text{p}(z_m; \alpha_m), \forall m \in M \end{aligned}$$

where the model's sparsity during pruning is $\rho(\mathcal{A}) \in [0, 1]$ defined in [24] and the target sparsity is a hyperparameter $\tau \in [0, 1]$. The training dataset \mathcal{D} has in total $|\mathcal{D}|$ utterances. The trainable modules Θ of the student encoder and the mask distributions' parameters $\mathcal{A} = \{\alpha_m\}_{m=1}^M$ are updated by gradient descent, while the trainable Lagrange multipliers $\lambda_1, \lambda_2 \in \mathbb{R}$ are updated by gradient ascent, which will increase the training loss continuously unless the equality constraint between the current sparsity $\rho(\mathcal{A})$ and the target sparsity τ is satisfied, ensuring the desired model size [24, 30].

Structured pruning for the hybrid AV encoder: The student encoder comprises three types of prunable modules: (1) The attention heads in multi-head attention of the transformer blocks, where we apply mask vectors $\mathbf{z}_{\ell'} = (z_\mu, \dots, z_{\mu+H-1})$ with scalar elements for each head with index $h \in \mathcal{H} = \{1, \dots, H\}$ of the in total $H = 12$ attention heads per transformer encoder block. The encoder block index is $\ell' \in \mathcal{L}' = \{1, \dots, L\}$, and we have in total L encoder blocks. We have a starting index μ related to m and $z_m \in \mathcal{Z}$. (2) The expanded intermediate dimension $4d$ of the FFN layers in the transformer blocks, to which we apply mask vectors $\mathbf{z}_n = (z_\mu, \dots, z_{\mu+4d-1})$ with a starting index μ related to m (and $z_m \in \mathcal{Z}$) in the n -th transformer encoder block following [32, 27]. (3) Output channels in the ResBlocks of the AV frontend. The ResNet-18-based visual frontend consists of eight serially connected ResBlocks indexed by $j \in \{1, 2, \dots, 8\}$, which includes ResBlock-B's with projection shortcut and ResBlock-A's with identity shortcut. Each ResBlock has two activation functions indexed by $b \in \{1, 2\}$, one within the shortcut indexed by $b = 1$, and the other one outside the shortcut indexed by $b = 2$. Except the first two ResBlock-A's, we apply mask vectors $\mathbf{z}_{j,b} = (z_\mu, \dots, z_{\mu+C_{j,b}-1})$ with a starting index μ related to m (with $z_m \in \mathcal{Z}$) to the $C_{j,b}$ output channels after the activation functions. Due to the identity shortcut in ResBlock-A, the second mask vector in ResBlock-A's $\mathbf{z}_{j,2}$ has to be the same as the second mask vector in the preceding block $\mathbf{z}_{j-1,2}$ to avoid dimensional inconsistencies.

3. Evaluation and Discussion

3.1. Experimental Setup

Databases and pre-processing: We train and evaluate our models on the Lip Reading Sentences 3 (LRS3) audiovisual speech recognition task, which is the largest publicly available labeled English AV-ASR dataset with 433 h training data collected from TED and TEDx talks [33]. We follow the pre-processing pipeline for the LRS3 dataset detailed in [18]: For input audio features, we extract 26-dimensional log-filterbank outputs from raw audio sampled at 16 kHz with a 25 ms window and a frame shift of 10 ms, yielding 100 audio frames per second. For video frames sampled at 25 Hz, we convert them to grayscale and crop them to a 96×96 region of interest based on the face alignment.

Joint pruning and distillation: Based on the PyTorch-based fairseq sequence-to-sequence toolkit [34], we implement the teacher-student model to perform joint distillation and pruning (DP) for AV encoders. The joint DP procedure consists of two steps. First, the student model is initialized with the teacher’s weights and jointly distilled and pruned for 50k updates to the target sparsity τ . The learning rates are adjusted using a linear decay learning rate scheduler with 0.0002, 0.02, 0.02 peak learning rates for the model parameters and auxiliary parameters $\{\alpha_m\}_{m=1}^M, \lambda_1, \lambda_2$, respectively, and using 15k warm-up steps. Second, the student model is further distilled using the teacher from step one for 25k updates to improve the performance. The learning rate is adjusted using a linear decay learning rate scheduler with 0.0001 peak learning rate and 5k warm-up steps. The noise-augmented AV-HuBERT base² is used as the teacher model and remains frozen. The target sparsity is linearly increased to the specified value using a warm-up phase of 5k steps. The hard concrete masks are sampled for each training batch and shared among the training samples within the batch.

Fine-tuning for AV-ASR: For a fair comparison, we employ the same decoder architecture as the baseline method [8] for fine-tuning the models across all experiments. The decoder network comprises six transformer decoder blocks with 57M parameters. The outputs of the encoder-decoder architecture are subword tokens generated by SentencePiece [35] with a vocabulary size of 1000. The fine-tuning process is done using the PyTorch-based fairseq toolkit. To ensure comparability, we adhere to the fine-tuning setup outline in [7]: The entire encoder-decoder model is fine-tuned for 60k updates. The encoder is frozen for the first 48k updates. We apply the same data augmentation as in AV-HuBERT [8], where 25% of the training data is augmented with noise chosen from babble, music, natural noise, and second interfering talker condition at a signal-to-noise ratio (SNR) of 0dB. There is no speaker overlap in babble noise and second interfering talker condition among different splits.

Evaluation in noisy environments: To add noise to our speech data, we exactly follow [8]. First, *babble noise* is generated by mixing utterances from 30 different speakers from the MUSAN dataset [36], with each speaker exclusively assigned to either the training, validation, or test partition. We also evaluate our approaches under conditions of speech with *music* noise and a *second interfering talker* from the LRS3 dataset.

3.2. Results and Discussion

Analysis of the teacher model: Table 1 reports the number of parameters (M) and computational complexity (MFLOPs/frame) of each block in the AV-HuBERT base teacher encoder, which

Table 1: Number of trainable parameters and FLOPs per frame for each block of the teacher (AV-HuBERT base) model. **Bold blocks** have prunable modules. A tab indicates a sub-block. The video extractor is a modified ResNet-18. The ResBlock-A(64) appears 2 times in the ResNet-18. AV-HuBERT base contains a stack of 12 encoder blocks.

Block	Parameters (M)	FLOPs (M/frame)
AV-HuBERT base [18]	103	818.8
Audio Extractor	< 1	< 1.0
Video Extractor (ResNet-18)	12	633.2
ResBlock-A(64) ($\times 2$)	< 1	71.8
ResBlock-B(128)	< 1	55.8
ResBlock-A(128)	< 1	71.6
ResBlock-B(256)	1	66.2
ResBlock-A(256)	1	85.0
ResBlock-B(512)	4	66.2
ResBlock-A(512)	5	85.0
Other	< 1	59.8
Modality Fusion	1	2.0
Positional Encoding	5	10.0
Encoder Block ($\times 12$)	7	14.0
Multi-head attention (MHA)	2	5.0
Feed-forward network (FFN)	5	9.0

comprises 103M parameters and needs 818.8 MFLOPs/frame. An imbalance between parameter count and computational complexity in the ResNet and transformer blocks is observed: The ResNet-based video extractor has only 12% of the encoder’s parameters (12M vs. $7M \times 12$) but requires 77% of the computational complexity (633.2 MFLOPs/frame vs. $14.0 \text{ MFLOPs/frame} \times 12$). The prunable modules in the blocks marked in **bold** sum up to 97M parameters and 597.8 MFLOPs/frame (i.e., 94% and 73% of the entire teacher encoder’s parameters/complexity, respectively), which are targeted by our proposed joint distillation and pruning (DP) method.

Main joint DP results: Table 2 presents the results of the teacher model in the upper table segment and several student models are shown below. We compare the student models in terms of the model size (#Pars.), computational complexity (MFLOPs/frame), and word error rate (WER) under clean conditions and various noisy conditions at an SNR of -5dB, 0dB, and 5dB.

The upper table segment (ID=1) shows the teacher model, which achieves a 1.8% WER on the test set under clean condition. However, it consists of 103M parameters and requires 818.8 MFLOPs/frame. The second table segment (ID=2) displays Distil-AV-HuBERT [7], which is compressed solely by distillation to an efficient student model with 32M parameters. At its model size, Distil-AV-HuBERT achieves state-of-the-art (SOTA) performance so far, yielding a 4.6% WER under clean condition and 20.2% WER in 0dB babble noise on the test set. The last three table segments demonstrate our proposed student models, which are compressed by applying joint DP to various modules at different sparsity levels τ . *All of these three proposed models outperform the previous SOTA student [7] under clean and noisy conditions, even the smallest student model (ID=5) with 34.7% fewer parameters (20.9M vs. 32.0M), showing the effectiveness of joint DP for an efficient noise-robust AV encoder.*

Our best performing student model (ID=4) applies joint DP to both the ResNet and transformer in the hybrid AV encoder, surpassing the previous SOTA Distil-AV-HuBERT [7] with a 32.6% relative improvement (WER of 3.1% vs. 4.6%) under clean conditions on the test set and yields superior performance

²https://github.com/facebookresearch/av_hubert

Table 2: WER (%) on the validation and test set of LRS3 with a video frame rate of 25Hz. The results of the teacher model are presented in the first table segment. The second segment displays the results of the state-of-the-art Distill-AV-HuBERT [7]. The results with our proposed joint DP are shown in the last three segments. Models are tested in babble noise, second interfering speaker condition (2nd Speaker), and music interference at various SNRs. The best scores are in **bold** (except teacher SOTA), the second best underlined.

ID	Model	Compressed Module	Sparsity \uparrow τ (%)	# Pars. \downarrow (M)	FLOPs \downarrow (M/frame)	Clean \downarrow		SNR (dB)	Babble \downarrow		2nd Speaker \downarrow		Music \downarrow	
						val	test		val	test	val	test	val	test
1	AV-HuBERT base [18] (teacher, SOTA)	-	0	103.0	818.8	6.0	1.8	-5	21.9	16.7	11.0	6.0	11.9	6.9
								0	12.4	6.4	8.9	4.0	8.6	3.9
								5	8.5	3.5	7.6	2.9	7.3	2.9
2	Distil-AV-HuBERT [7] (so-far SOTA for <40M)	Transformer	69	32.0	674.0	-	4.6	-5	-	-	-	-	-	-
								0	-	20.2	-	17.5	-	-
								5	-	-	-	-	-	-
3	Joint DP (ours)	Transformer	65	35.9	682.0	<u>8.6</u>	<u>3.8</u>	-5	<u>41.8</u>	<u>33.5</u>	<u>27.9</u>	<u>22.9</u>	<u>23.4</u>	<u>18.1</u>
								0	<u>23.8</u>	<u>15.3</u>	<u>18.5</u>	<u>14.1</u>	<u>16.2</u>	<u>9.4</u>
								5	<u>14.7</u>	<u>8.4</u>	<u>13.9</u>	<u>9.0</u>	<u>12.2</u>	<u>6.4</u>
4	Joint DP (ours)	ResNet +	<u>70</u>	<u>31.0</u>	<u>459.6</u>	8.1	3.1	-5	41.2	32.4	25.4	19.5	23.1	16.2
		Transformer						0	22.8	14.4	17.3	12.0	15.5	8.6
		5						13.9	6.9	12.9	7.6	11.9	5.8	
5	Joint DP (ours)	ResNet +	80	20.9	391.4	9.2	4.2	-5	46.6	38.0	32.3	27.7	27.4	20.3
		Transformer						0	26.5	17.6	21.9	16.3	18.3	11.1
		5						16.0	9.5	15.8	10.2	12.8	7.7	

across all noisy conditions, while simultaneously reducing the computational complexity by 31.8%.

It is worth to note that our best performing student with compressed ResNet and transformer (ID=4) exceeds our student model with only compressed transformer (ID=3) in clean condition (WER of 3.1% vs. 3.8%) and across all noise conditions even with fewer parameters (31.0M vs. 35.9M) and fewer computations (459.6M vs. 682.0 MFLOPs/frame). This reveals the necessity to compress both the ResNet and transformer in the hybrid AV encoder for efficient noise-robust AV-ASR.

Pruned architecture of student models: Figure 2 illustrates the pruned architecture of our proposed students (ID=3,4,5) of Table 2. The upper subfigure displays the number of channels in the ResBlocks j with activation function index b (referred to $j.b$) of the visual frontend, where the gray student (ID=3) has the same number of convolutional channels as the teacher, while the blue (ID=4) and red (ID=5) students have less channels due to joint DP. Accordingly, these two students saved parameters and reduced computational complexity in the visual frontend. The center subfigure presents the number of attention heads, while the bottom subfigure shows the number of intermediate units of the FFN in transformer blocks. For reference, *each transformer encoder block* in the teacher model has 12 attention heads and an FFN intermediate dimension of 3,072. After joint DP, all three student models end up having much less attention heads and much smaller intermediate FFN dimensions, with a rising tendency towards later encoder blocks (see, e.g., $\ell'=8, 11$, or 12). These late encoder blocks are responsible for extracting linguistic information, which is beneficial for ASR tasks [29, 37]. Compared to the gray student (ID=3, 35.9M parameters), the best performing blue student (ID=4, 31.9M parameters) has more attention heads and more intermediate units particularly in later blocks, achieving a more efficient and noise-robust architecture by applying joint DP to both ResNet and transformer.

4. Conclusions

In this work, we presented a teacher-student framework to compress a hybrid audiovisual (AV) encoder for automatic speech recognition (ASR) composed of a ResNet-based AV frontend and transformer blocks. Building on our analysis of the hybrid

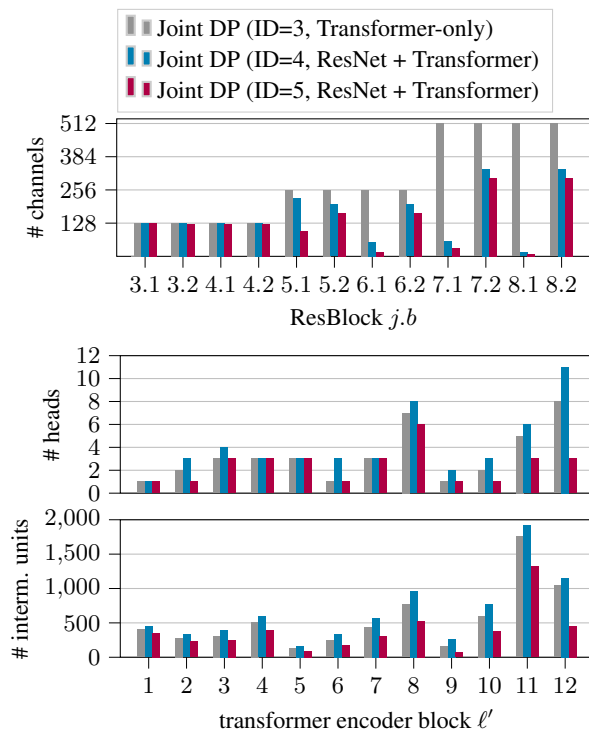


Figure 2: Pruned architectures of our three proposed students compressed by joint DP, chosen from ID=3,4,5 in Table 2.

teacher AV encoder, we carefully design joint distillation and pruning (DP) methods. At a similar model size, compared to the previous state-of-the-art (distillation-only) model, our best performing DP model achieves a 32.6% relative improvement in clean conditions (WER=3.1% vs. 4.6%) and yields superior performance across all noisy conditions, while simultaneously reducing the computational complexity by 31.8%. By applying joint DP to both ResNet-based visual frontend and transformer in the hybrid AV encoder, we thereby claim a new state of the art in the regime $\approx 20 \dots 35$ M parameter AV encoders.

5. Acknowledgments

The research leading to these results has received funding from the Bundesministerium für Bildung und Forschung (BMBF) under funding code 03VP10991 (BesserLesen project).

6. References

- [1] H. McGurk and J. MacDonald, “Hearing Lips and Seeing Voices,” *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.
- [2] L. D. Rosenblum, “Speech Perception as a Multimodal Phenomenon,” *Current Directions in Psychological Science*, vol. 17, no. 6, pp. 405–409, 2008.
- [3] T. Drugman, M. Gurban, and J.-P. Thiran, “Relevant Feature Selection for Audio-Visual Speech Recognition,” in *Proc. of MMSP*, Chania, Greece, Oct. 2007, pp. 179–182.
- [4] 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).
- [5] 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.
- [6] 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.
- [7] Z. Li, C. Liang, T. Lohrenz, M. Sach, B. Möller, and T. Fingscheidt, “An Efficient and Noise-Robust Audiovisual Encoder for Audiovisual Speech Recognition,” in *Proc. of Interspeech*, Dublin, Ireland, Aug. 2023, pp. 1583–1587.
- [8] B. Shi, W.-N. Hsu, and A. Mohamed, “Robust Self-Supervised Audio-Visual Speech Recognition,” in *Proc. of Interspeech*, Incheon, Korea, Sep. 2022, pp. 2118–2122.
- [9] H. Chen, H. Zhou, J. Du, C.-H. Lee, J. Chen, S. Watanabe, S. M. Siniscalchi, O. Scharenborg, D.-Y. Liu, B.-C. Yin *et al.*, “The First Multimodal Information Based Speech Processing (MISP) Challenge: Data, Tasks, Baselines and Results,” in *Proc. of ICASSP*, Singapore, May 2022, pp. 9266–9270.
- [10] H. Wang, P. Guo, P. Zhou, and L. Xie, “MLCA-AVSR: Multi-Layer Cross Attention Fusion based Audio-Visual Speech Recognition,” in *Proc. of ICASSP*, Seoul, South Korea, Apr. 2024, pp. 8150–8154.
- [11] Z. Li, T. Graave, J. Liu, T. Lohrenz, S. Kunzmann, and T. Fingscheidt, “Parameter-Efficient Cross-Language Transfer Learning for a Language-Modular Audiovisual Speech Recognition,” in *Proc. of ASRU*, Taipei, Taiwan, Dec. 2023, pp. 1–8.
- [12] 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.
- [13] D. Serdyuk, O. Braga, and O. Siohan, “Audio-Visual Speech Recognition Is Worth 32x32x8 Voxels,” in *Proc. of ASRU*, Cartagena, Colombia, Dec. 2021, pp. 796–802.
- [14] —, “Transformer-Based Video Front-Ends for Audio-Visual Speech Recognition for Single and Multi-Person Video,” in *Proc. of Interspeech*, Incheon, South Korea, Sep. 2022, pp. 2833–2837.
- [15] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, “Lip-Reading With Densely Connected Temporal Convolutional Networks,” in *Proc. of IEEE WACV*, virtual, Jan. 2021, pp. 2857–2866.
- [16] 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.
- [17] S. Petridis *et al.*, “Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture,” in *Proc. of SLT*, Athens, Greece, Dec. 2018, pp. 513–520.
- [18] 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.
- [19] 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.
- [20] A. Rouditchenko, Y. Gong, S. Thomas, L. Karlinsky, H. Kuehne, R. Feris, and J. Glass, “Whisper-Flamingo: Integrating Visual Features into Whisper for Audio-Visual Speech Recognition and Translation,” in *Proc. of Interspeech*, Sep. 2024, pp. 2420–2424.
- [21] U. Cappellazzo, M. Kim, H. Chen, P. Ma, S. Petridis, D. Falavigna, A. Brutti, and M. Pantic, “Large Language Models Are Strong Audio-Visual Speech Recognition Learners,” *arXiv:2409.12319*, pp. 1–5, Sep. 2024.
- [22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter,” *arXiv:1910.01108*, pp. 1–5, Mar. 2019.
- [23] Y. Lee, K. Jang, J. Goo, Y. Jung, and H. Kim, “TinyBERT: Distilling BERT for Natural Language Understanding,” in *Findings of the ACL-EMNLP*, virtual, Nov. 2020, pp. 4163–4174.
- [24] Z. Wang, J. Wohlwend, and T. Lei, “Structured Pruning of Large Language Models,” in *Proc. of EMNLP*, virtual, Nov. 2020, pp. 6151–6162.
- [25] 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.
- [26] 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,” in *Proc. of Interspeech*, Incheon, South Korea, Sep. 2022, pp. 1686–1690.
- [27] Y. Peng, K. Kim, F. Wu, P. Sridhar, and S. Watanabe, “Structured Pruning of Self-Supervised Pre-Trained Models for Speech Recognition and Understanding,” in *Proc. of ICASSP*, Rhodes, Greece, Jun. 2023, pp. 1–5.
- [28] Y. Lee, K. Jang, J. Goo, Y. Jung, and H. Kim, “FitHuBERT: Going Thinner and Deeper for Knowledge Distillation of Speech Self-Supervised Models,” in *Proc. of Interspeech*, Incheon, South Korea, Sep. 2022, pp. 3588–3592.
- [29] A. Pasad, B. Shi, and K. Livescu, “Comparative Layer-Wise Analysis of Self-Supervised Speech Models,” in *Proc. of ICASSP*, Rhodes, Greece, Jun. 2023, pp. 1–5.
- [30] Y. Peng, Y. Sudo, S. Muhammad, and S. Watanabe, “DPHuBERT: Joint Distillation and Pruning of Self-Supervised Speech Models,” in *Proc. of Interspeech*, Dublin, Ireland, Aug. 2023, pp. 62–66.
- [31] C. Louizos, M. Welling, and D. P. Kingma, “Learning Sparse Neural Networks through L_0 Regularization,” in *Proc. of ICLR*, Vancouver, Canada, Apr. 2018, pp. 1–13.
- [32] M. Xia, Z. Zhong, and D. Chen, “Structured Pruning Learns Compact and Accurate Models,” in *Proc. of ACL*, Dublin, Ireland, May 2022, pp. 1513–1528.
- [33] T. Afouras, J. S. Chung, and A. Zisserman, “LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition,” *arXiv:1809.00496*, pp. 1–2, Oct. 2018.
- [34] 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.
- [35] T. Kudo and J. Richardson, “SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing,” in *Proc. of EMNLP*, Brussels, Belgium, Nov. 2018, pp. 66–71.
- [36] D. Snyder, G. Chen, and D. Povey, “MUSAN: A Music, Speech, and Noise Corpus,” *arXiv:1510.08488v1*, pp. 1–4, Oct. 2015.
- [37] Z. Li, P. Blumenberg, J. Liu, T. Graave, T. Lohrenz, S. Kunzmann, and T. Fingscheidt, “Interleaved Audio/Audiovisual Transfer Learning for AV-ASR in Low-Resourced Languages,” in *Proc. of Interspeech*, Kos, Greece, Sep. 2024, pp. 2524–2528.