Virkkunen, Anja; Huang, Guangpu; Grosz, Tamas; Kurimo, Mikko

INVESTIGATING THE CLUSTERS DISCOVERED BY PRE-TRAINED AV-HUBERT

Published in:
2024 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2024 - Proceedings

DOI:
10.1109/ICASSP48485.2024.10447434

Published: 01/01/2024

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

INVESTIGATING THE CLUSTERS DISCOVERED BY PRE-TRAINED AV-HUBERT

Anja Virkkunen *Marek Sarvaš** *Guangpu Huang* *Tamas Grosz* *Mikko Kurimo*

Aalto University
Department of Information and Communications Engineering
firstname.lastname@aalto.fi

ABSTRACT

Self-supervised models, such as HuBERT and its audio-visual version AV-HuBERT, have demonstrated excellent performance on various tasks. The main factor for their success is the pre-training procedure, which requires only raw data without human transcription. During the self-supervised pre-training phase, HuBERT is trained to discover latent clusters in the training data, but these clusters are discarded, and only the last hidden layer is used by the conventional finetuning step. We investigate what latent information the AV-HuBERT model managed to uncover via its clusters and can we use them directly for speech recognition. To achieve this, we consider the sequence of cluster ids as a 'language' developed by the AV-HuBERT and attempt to translate it to English text via small LSTM-based models. These translation models enable us to investigate the relations between the clusters and the English alphabet, shedding light on groups of latent clusters specialized to recognise specific phonetic groups. Our results demonstrate that using the pre-trained system as a quantizer, we are able to compress the video to as low as 275 bit/sec while maintaining acceptable speech recognition accuracy. Furthermore, compared to the conventional finetuning step, our solution has considerably lower computational cost.

***Index Terms*—** ASR, audiovisual, AV-HuBERT, SSL, machine translation

1. INTRODUCTION

In recent years, we saw the rise of self-supervised learning (SSL), which produces pre-trained models optimized on large quantities of unannotated data. Their popularity is owed to the fact that after the unsupervised pre-training phase, they require only a limited amount of supervised data in the finetuning phase to achieve state-of-the-art performance on various tasks. The success of SSL models hints that the pre-training phase enables them to learn a general concept about their input modality, while in finetuning, they can be quickly specialized to a given task.

Audio-visual speech recognition combines lip reading with regular audio-based speech recognition. In the first experiments in the field, the two modalities, audio and video, were combined either using early, middle or late fusion [1, 2]. More recently, the encoder-decoder paradigm in neural networks has made middle fusion approaches easier to implement. A common approach is to use separate CNN encoders for both inputs and then fuse encoder outputs in an attention-based decoder [3]. The latest development in the speech technology field has been the rise of large self-supervised and pre-trained models, like wav2vec 2.0 [4], Whisper [5], WavLM [6] and Hidden-Unit BERT (HuBERT) [7] in the audio-only domain, and in image processing Vision Transformers (ViT) [8] are gradually replacing Residual Networks [9]. For the audio-visual domain, the most prominent example of such a model is the audio-visual HuBERT [10].

While SSL models showcase excellent results on specific tasks, it is also noted that the finetuning process often leads to Catasrofic Forgetting (CF) [11]. The loss of a considerable amount of general knowledge during the finetuning step could lead to performance issues. An additional problem with the finetuning procedure is that it completely discards the latent clusters discovered during pre-training, only focusing on the outputs of the Transformer layers, possibly leading to severe information loss. While most approaches ignore the pre-trained clusters discovered by SSL models, in a recent work [12] demonstrated that the codebook indexes of wav2vec 2.0 could be valuable assets to distil its knowledge into a smaller model. Inspired by this observation, we aim to investigate how the clusters of AV-HuBERT can be used for ASR without losing the model's generalization. Our primary aim is to investigate the clusters of the pre-trained model, thus improving AV-HuBERT's transparency.

To achieve our goal, we freeze the pre-trained model and use it as an Encoder. As a second step, we train low-resource decoders, which serve as a tool to interpret the clusters via their attention maps. This choice enables us to conduct a thorough investigation of the latent clusters and the capabilities of the pre-trained models via Decoders trained to translate the sequence of cluster IDs produced by AV-HuBERT to English.

In this work, we make the following key contributions:

1. Empirical investigation of the relation between the clusters formed by the AV-HuBERT models and ASR units to uncover subgroups dedicated to recognizing the same content.
2. Propose a new way of utilizing pre-trained AV-HuBERT for audiovisual ASR with low computational cost.

2. METHODS

2.1. AV-HuBERT

AV-HuBERT developed by Shi et al. [10] is the audio-visual counterpart to audio-based HuBERT model [7]. HuBERT models are a family of self-supervised models for learning speech representations. The speech representations are learned through an iterative two-step process of feature clustering and masked prediction. In [7], Hsu et al. show how these learned speech representations can be applied to the

*The author performed the work as an Erasmus student from the Brno University of Technology

We are grateful for the Academy of Finland project funding number 345790 in ICT 2023 programme's project "Understanding speech and scene with ears and eyes". The computational resources were provided by Aalto ScienceIT. This work was supported by the Academy of Finland Flagship programme: Finnish Center for Artificial Intelligence FCAI

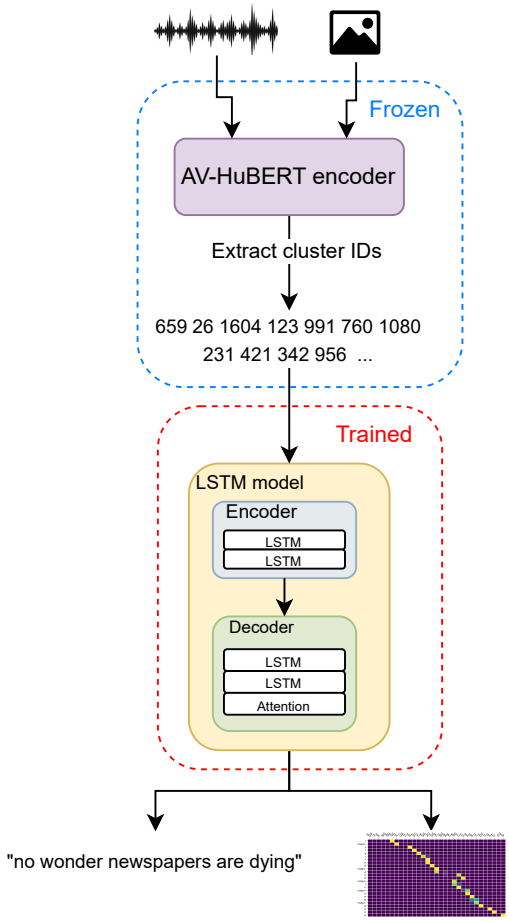

Fig. 1. Proposed audio-visual ASR pipeline.

task of automatic speech recognition, and other works have applied pre-trained AV-HuBERTs to for example speaker verification [13] and speech-to-speech translation [14].

The model architecture of AV-HuBERT can be split into four steps. First, each input modality is fed through their respective encoder, which for visual input is a ResNet and for audio input and a single feed-forward linear projection layer. Second step concatenates these two feature vectors into one. During model training, modality dropout is also applied at this step to prevent the model from relying too much on the audio input. In the third step, the concatenated feature vector is passed through a transformer encoder to produce contextualized audio-visual representations. The last step is to predict the cluster IDs using the audio-visual representation.

The initial target cluster IDs are created by applying k-means clustering on the audio features. The model is trained on these targets for one iteration, after which new k-means clustering is done using features extracted from the model. In [10], this two-step pre-training process was repeated five times.

2.2. Mapping Latent Clusters to Characters

After the pre-training step, AV-HuBERT models can be finetuned by connecting a new output layer to the last Transformer layer and using the CTC loss [15]. This approach, however, discards the clustering layer, optimized in the per-training steps and incurs a considerable computational cost to optimize a large number of parameters inside the BERT model. As an alternative, we propose using the cluster output of the original model, and keeping its weights frozen.

Our solution reduces computational cost, by avoiding the optimization of AV-HuBERT, only using it as a fixed encoder. Once the predicted cluster ids are extracted for the training data, we connect a simple LSTM-based translation model, to find the connection between the latent groups and the English alphabet and some other symbols (i.e. word separator and apostrophe). The proposed system is illustrated in Fig. 1.

The benefit of our solution is two-fold, on the one hand, it allows us to build a low-cost ASR solution without losing the generalization of AV-HuBERT; simultaneously, the LSTM translation model enables us to investigate the relation between the clusters and the ASR units. To achieve the later task, we examine the attention values of the final model to find latent groups that are relevant for predicting certain characters.

3. EXPERIMENTAL SETUP

As a first step, we aimed to recreate the finetuned models presented in the original AV-HuBERT paper [10]. By utilizing the pre-trained models, which have been trained with only LRS3 dataset [16] and the codebase[1] the authors have publicly shared, we were able to optimize three models: the *BASE* sized model after iterations 4 and 5 of pre-training and the *LARGE* one after full pre-training. Our selection was motivated by the goal of seeing how the size of the model and the final pre-training step affect the performance. We finetune each of these three pre-trained models on the 30h subset of LRS3 using the audio-visual modality.

The LRS3-TED dataset is a large English audio-visual corpus designed primarily for lip reading and audio-visual speech recognition tasks. It contains cropped facetrack videos and associated speech transcripts extracted from TED and TEDx videos available on Youtube. In total, there are 438 hours of data split between pre-train (407h), train-val (30h) and test (1h) sets.

On the low computational side, we explored two alternatives. The first one utilized AV-HuBERT as a deep cluster tool and translated its output consisting of cluster IDs to English using LSTM-based models. The second solution, serving as a comparable baseline, extracted the audiovisual embeddings from the last hidden layer of AV-HuBERT, which is typically connected to the CTC output during finetuning, and quantized it via an additional k-means step. The main motivation for including k-means in this system was to avoid memory and storage issues, the extracted hidden representations of the 30h training data required 8 and 11 GB (for *base* and *large* respectively). In contrast, storing the cluster IDs only needed $\leq$20 MB of space. With this second approach, we aimed to investigate whether the learned clusters contained any additional information compared to the high dimensional latent embeddings of the Transformer component. In all cluster-based solutions, we removed repeated cluster IDs from the sequence as preliminary experiments proved that subsequent duplicates make the translation task harder.

We implement two variations of the LSTM model using the OpenNMT toolkit [17]. The first version is the OpenNMT default LSTM model with two layers and 500 hundred hidden units. The second, larger version has 1024 hidden units. Both LSTM types are trained with the Adam optimizer using early stopping, learning rate of 0.001, dropout of 0.1 and label smoothing of 0.1. Batch size is set at 256 for the smaller LSTM and at 512 for the bigger LSTM. The

[1] https://github.com/facebookresearch/av_hubert

Table 1. Test set results for finetuned AV-HuBERTs and our proposed translation LSTMs. Each finetuned AV-HuBERT uses audio-visual input both in training and decoding. The Size column contains the number of parameters that need to be optimized during training.

Model	Size	WER (%)
Finetuned base iter 4	161M	**7.35** (±0.0)
Finetuned base iter 5	161M	7.75 (±0.20)
Finetuned large iter 5	477M	8.84 (±0.19)
base iter 4 + LSTM500	10M	23.65 (±2.24)
base iter 4 + LSTM1024	42M	24.56 (±0.0)
base iter 5 + LSTM500	11M	18.92 (±0.74)
base iter 5 + LSTM1024	43M	20.97 (±0.44)
large iter 5 + LSTM500	11M	**17.85** (±1.29)
large iter 5 + LSTM1024	43M	19.15 (±0.0)
k-means base iter 5 + LSTM500	11M	**17.55** (±0.0)
k-means base iter 5 + LSTM1024	43M	20.93 (±0.0)
k-means large iter 5 + LSTM500	11M	35.80 (±0.0)
k-means large iter 5 + LSTM1024	43M	39.77 (±0.0)

input of the models consisted of cluster IDs of the 1000 clusters used in iteration 4, while after the final (5th) iteration, it was increased to 2000 clusters as in [10].

4. RESULTS

Table 1 summarizes our experimental results. Each model type was trained five times to assess their robustness, and we report the mean performance with 95% confidence interval.

First, we looked at the *Finetuned* results and noticed that, interestingly, the smaller model yielded the best performance after the fourth iteration of pre-training. Comparing our results with the one (4.2% WER for *Finetuned large iter 5*) reported in [10], we can see considerable differences, which could be attributed to several factors like the change in the computational environment (training on single GPU instead of 32 or 64 V100 GPUs) and the fact that we did not perform any hyperparameter optimization. These differences highlight a common issue of reproducibility as it is not feasible for all to have access to such computational resources and conduct proper hyperparameter tuning.

Next, we investigated how well the pre-trained clusters encode information for speech recognition. Although the translated results are considerably worse than the finetuned ones, they are still demonstrating good ASR capabilities. Note that our proposed solution only used AV-HuBERT to infer the cluster IDs and had only 11M trainable parameters compared to the 477M parameters that need to be optimized to finetune the *large iter5* model. Overall, our LSTM-based system proved to be two times faster in the training phase and achieved a similar speed in decoding.

Similarly to the finetuned models, the smaller cluster-based models (*LTSM500*) performed better than their larger counterparts (*LSTM1024*), at the cost of lower robustness. The best model, *large iter5+LSTM500* achieved 17.85% WER on average, demonstrating that the cluster of the larger model captured more information than the base one. Additionally, comparing the *base iter4* and *iter5*, we can see that the additional pre-training step was quite beneficial, contrary to the finetuned results. Lastly, looking at the performance of models using the extra bottleneck k-means step indicates that the pre-trained cluster were able to capture valuable information that

Fig. 2. The distribution of WER (left) and CER (right) of the cluster-based model *large iter5+LSTM500* and the finetuned model *Finetuned base iter 4*. Outliers are marked with dots and dashed line shows the mean. Note that we removed the utterances with 0.0% error rate to make the figure more focused on the severity of mistakes. We also left out few extreme outliers which would have stretched the y-axis such that the plot would have been hard to read.

can be lost in the re-clustering process. Specifically, in the case of the *large* model, we observed a considerable gap between the performance of the pre-trained cluster IDs and their re-clustered counterparts, while the *base* model exhibited the opposite trend. These results suggest that even if the re-clustering step manages to capture the same information as the pre-trained ones, considerably more effort is needed to optimize the hyperparameters in order to properly exploit these new latent groups and achieve stable ASR performance.

Based on the ASR results, we selected two models (*Finetuned base iter 4* and *large iter5+LSTM500*) for further investigation. Figure 2 compares the distributions of errors made by these two models. Upon closer inspection, we found that around 32% of the utterances have equal WER and CER, and the clustering-based approach produced better CER/WER for 21.1% and 22.1% of the sentences, respectively. The large number of outliers causing high WER/CER in the case of translation-based approaches proved to be the product of hallucination (mainly phrase repetition), which is a common issue of translation models [18].

In total, ∼5% of the test data is affected by severe hallucinations, meaning that insertion errors are the dominant source of mistakes. After careful investigation, we found that the main reasons for hallucinations are common pronouns (we, they), conjunction (or) and the preposition of. After these so-called high-inflow words [19], the model tends to repeat the whole following phrase (2-4 words) several times, causing a considerable amount of insertion errors. By comparing the clustering-based solution to the fine-tuned model, we observed a 400% increase in insertion errors, in contrast to a more modest increase (approx. 100%) in deletions and substitutions. This suggests that employing techniques like [19] to mitigate the hallucinations might reduce the gap between our solution and the costly finetuned models' performance.

To further study the clusters, we inspected the attention values between clusters and character groups. The aggregated result is displayed in Figure 3. It shows that, especially in the case of consonants, many clusters put most of their attention to one category. For

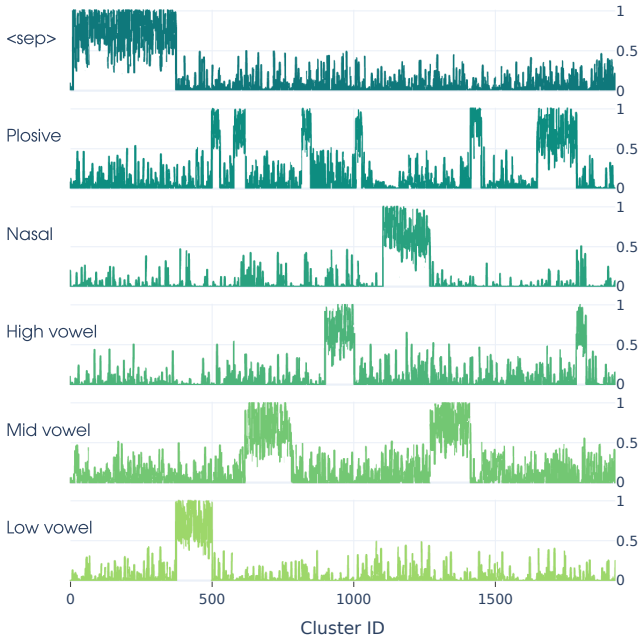

Fig. 3. A closer look at how attention is distributed across cluster IDs for <sep> token, plosives (*b,d,g,k,p,t*), nasals (*m,n*), high vowels (*i,u*), mid vowels (*e,o*) and low vowels (*a*).

vowels and silence (word separator) small amounts of attention activation are distributed throughout all the cluster IDs, although there are clear groups of clusters with higher attention as well.

Overall, we can see that most of the clusters are specialized to recognize specific groups, while a small portion captures some other information relevant for more than one character category. Interestingly, almost 400 clusters were used for predicting the word separator unit, approximately 500 proved to be devoted to recognizing vowels, and less than 1000 units were mapped to consonants. In Figure 4, a more detailed, per-character attention mapping is shown. We observed that the model reserved most clusters to predict vowel *E*, and it had considerable overlap with those used for outputting the letter *O*. *M* and *T* required the most latent units on the consonant side, while rare letters like *Q*, *X*, and *Z* attended to only few clusters.

Additionally, AV-HuBERT reserved one cluster to mark the beginning and several clusters to signal the end of the recording without any explicit instruction. Moreover, we noticed that ∼50 clusters were present in the AV-HuBERT outputs but were mostly ignored by the translation model, suggesting that they encoded some additional audiovisual information irrelevant for speech recognition. We should note that not all clusters were utilized. In total, there were 10 unused ones, most likely dedicated to some other events not occurring in the supervised data. The investigation of these ignored units is out of the scope of this paper but remains an important future task.

5. CONCLUSIONS

Self-supervised pre-training procedures have gained popularity in recent years and produced several state-of-the-art solutions for multiple tasks, including audio-visual speech recognition. Their main appeal is that only a limited amount of annotated data is required for their finetuning. While they have several advantages over previous solutions, one considerable downside of SSL models is their

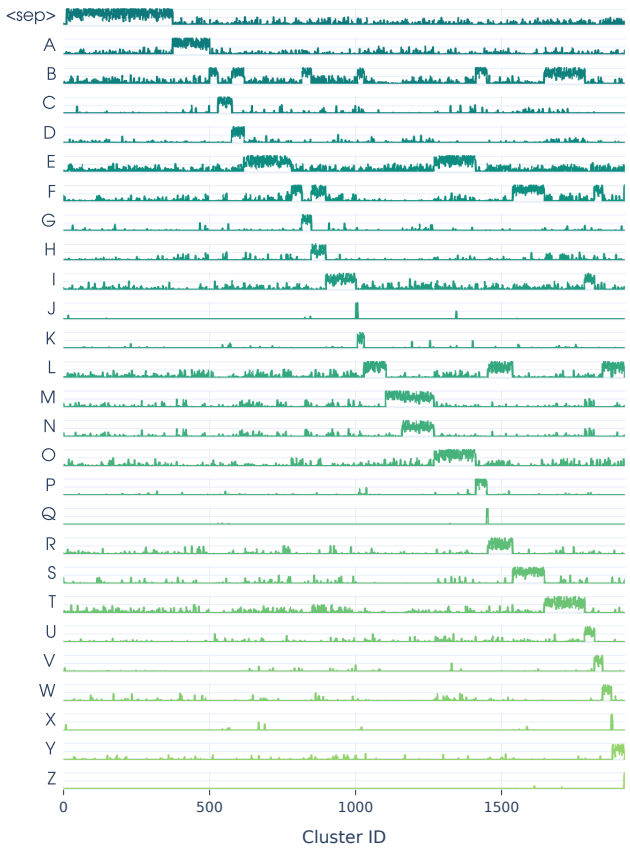

Fig. 4. Character-level attention values for each cluster ID. Attention values range between 0 and 1. The cluster IDs have been reordered by attention to make it easier to see how many clusters are dedicated to each character.

considerable size and, consequently, computational cost. This work focuses on how AV-HuBERT could be efficiently used without large clusters of GPUs by utilizing its latent clusters discovered during pre-training.

Our study demonstrates a more efficient way of using pre-trained models for creating AV-ASR systems compared to the standard finetuning approach. In contrast to conventions, we employ AV-HuBERT as a frozen encoder to cluster the data and employ a small LSTM model to transform the IDs into text. Although the costly finetuning procedure achieves higher accuracy overall, our analysis revealed that the proposed solution, which contains an order of magnitude fewer parameters and is considerably faster, still manages to outperform it in many cases. These observations suggest that the latent clusters should also be utilized by the finetuning procedure as they often contain complementary information.

In addition to the speech recognition experiments, we also conducted a thorough analysis of the clusters. We have uncovered the relation between English characters and the latent clusters using the attention values of the translation model, and identified groups of clusters relevant for recognizing various phonetic groups. Lastly, we also discovered that many clusters are dedicated to other non-ASR tasks, but determining their exact function remains an important part of our future work.

6. REFERENCES

[1] Xiaoyu Song, Hong Chen, Qing Wang, Yunqiang Chen, Mengxiao Tian, and Hui Tang, "A review of audio-visual fusion with machine learning," *Journal of Physics: Conference Series*, vol. 1237, no. 2, pp. 022144, jun 2019.

[2] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun, "Attention bottlenecks for multimodal fusion," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, Eds. 2021, vol. 34, pp. 14200–14213, Curran Associates, Inc.

[3] Linlin Xia, Gang Chen, Xun Xu, Jiashuo Cui, and Yiping Gao, "Audiovisual speech recognition: A review and forecast," *International Journal of Advanced Robotic Systems*, vol. 17, no. 6, pp. 1729881420976082, 2020.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[5] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, Eds. 23–29 Jul 2023, vol. 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518, PMLR.

[6] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Micheal Zeng, and Furu Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[7] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *CoRR*, vol. abs/2106.07447, 2021.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. 2021, OpenReview.net.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[10] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *International Conference on Learning Representations*, 2022.

[11] Vinay Venkatesh Ramasesh, Ethan Dyer, and Maithra Raghu, "Anatomy of catastrophic forgetting: Hidden representations and task semantics," in *International Conference on Learning Representations*, 2021.

[12] Liyong Guo, Xiaoyu Yang, Quandong Wang, Yuxiang Kong, Zengwei Yao, Fan Cui, Fangjun Kuang, Wei Kang, Long Lin, Mingshuang Luo, Piotr Żelasko, and Daniel Povey, "Predicting multi-codebook vector quantization indexes for knowledge distillation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[13] Bowen Shi, Abdelrahman Mohamed, and Wei-Ning Hsu, "Learning Lip-Based Audio-Visual Speaker Embeddings with AV-HuBERT," in *Proc. Interspeech 2022*, 2022, pp. 4785–4789.

[14] Rongjie Huang, Huadai Liu, Xize Cheng, Yi Ren, Linjun Li, Zhenhui Ye, Jinzheng He, Lichao Zhang, Jinglin Liu, Xiang Yin, and Zhou Zhao, "AV-TranSpeech: Audio-visual robust speech-to-speech translation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, July 2023, pp. 8590–8604, Association for Computational Linguistics.

[15] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA, 2006, ICML '06, p. 369–376, Association for Computing Machinery.

[16] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *CoRR*, vol. abs/1809.00496, 2018.

[17] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush, "OpenNMT: Open-source toolkit for neural machine translation," in *Proceedings of ACL 2017, System Demonstrations*, Vancouver, Canada, jul 2017, pp. 67–72, Association for Computational Linguistics.

[18] David Dale, Elena Voita, Loïc Barrault, and Marta R. Costa-jussà, "Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, Eds. 2023, pp. 36–50, Association for Computational Linguistics.

[19] Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi, "A theoretical analysis of the repetition problem in text generation," in *AAAI Conference on Artificial Intelligence*, 2020.