Informatics and Control Problems 46 Issue 1 (2026)
journal homepage: www.icp.az

Pichilti: A monolingual Azerbaijani distilled Whisper model

Mirakram Aghalarov, Mahammad Mehdi, Javidan Zeynalov, Sabuhi Aghayev

Baku Higher Oil School, Baku, Azerbaijan

ARTICLE INFO

Article history:
Received 24.10.2025
Received in revised form 14.11.2025
Accepted 20.11.2025
Available online 20.03.2026

Keywords:
Knowledge Distillation
Speech to Text
Voice Processing
Low Resource Languages

ABSTRACT

The recent saturation in the development of Speech-to-Text (STT) models has been disrupted by the release of multilingual models trained using zero-shot learning. While these models offer powerful and robust capabilities for voice processing in noisy environments, their multilingual nature leads to increased GPU utilization. Moreover, smaller models exhibit poor performance on low-resource languages such as Azerbaijani. This paper introduces a methodology and a large-scale voice dataset designed for training STT models in Azerbaijani. Over 500 hours of speech data have been collected, and knowledge distillation techniques have been applied at various levels. As a result, the distilled Whisper variant (Pichilti-base) outperforms Whisper-large v3 in Azerbaijani for voice recognition tasks. Additionally, specific post-processing methods have been implemented to mitigate hallucination effects in silent recordings.

1. Introduction

The study of voice processing dates back to the 20th century, when researchers began analyzing specific patterns in audio recordings. Given the time-intensive nature of manual voice analysis, identifying these patterns became crucial for automating the process. The transition to frequency-domain analysis and the classification of voices based on different objectives significantly improved accuracy. The increasing maturity of deep learning further automated feature extraction, enabling more efficient preliminary analysis of speech data.

Several key applications have emerged in the field of voice processing:

- Speech-to-Text (STT): Converts spoken language into text, facilitating tasks such as context analysis and voice-based assistants.
- Voice Classification: Categorizes voice inputs for applications such as smart home automation.
- Speaker Identification: Embeds speaker characteristics to recognize individuals based on pre-recorded voice data.

Many applications exist in this domain; however, language dependency remains a significant bottleneck. Some tasks, such as speaker identification, are language-agnostic, as voice frequency patterns can generate person-specific embeddings. In contrast, Speech-to-Text (STT) systems often struggle with multilingual adaptability, as they require extensive language-specific data. This raises questions about the level of digitization for a given language. Consequently, low-resource languages suffer from limited and lower-quality datasets, leading to less accurate results.

*Corresponding author
E-mail addresses: mirakram.agalarov@bhos.edu.az (M.Sh. Aghalarov), mahammad.mehdi.std@bhos.edu.az (M.G. Mehdi), dzhavidan.zeinalov.std@bhos.edu.az (J.E. Zeynalov), sabuhi.aghayev.std@bhos.edu.az (S. Aghayev)

www.icp.az/2026/1-07.pdf https://doi.org/10.54381/icp.2026.1.07

Leading companies in the machine learning market have introduced innovative solutions to reduce dependence on language-specific voice datasets, enabling the development of multilingual models. OpenAI's Whisper [1] is an open-source multilingual model trained on various languages, including some low-resource ones like Azerbaijani. OpenAI has released multiple versions of Whisper with varying model sizes, where the largest model consistently demonstrates the highest accuracy. However, smaller variants struggle to match the performance of monolingual models, which themselves often underperform in low-resource languages.

This raises a critical concern for real-world deployment. Transformer-based architectures are highly compute-intensive, leading to full utilization of GPU cores for each request. This trend creates a challenging trade-off, as the two key aspects of the solution—accuracy and efficiency—diverge significantly, making it difficult to achieve both simultaneously.

The collection and annotation of speech data are resource-intensive processes, particularly for low-resource languages. Gathering data in diverse real-world scenarios, such as noisy environments and multi-speaker settings, presents additional challenges, as the level of digitization for a given language significantly impacts data availability. Moreover, many state-of-the-art models rely on transformer architectures, which are inherently data-hungry. While hybrid approaches that combine Hidden Markov Models (HMMs) [2] with deep learning have been explored, they generally fail to achieve the accuracy of fully deep learning-based architectures. Consequently, the development of automatic speech recognition (ASR) systems for low-resource languages requires alternative techniques that can function effectively with limited labeled data.

This paper presents a methodology for leveraging publicly available videos to enhance the quality of automatic speech recognition (ASR) for low-resource languages like Azerbaijani without the need for manual annotations (Figure 1) This approach reduces the computational burden during deployment by enabling the use of smaller model sizes. Additionally, findings indicate that the distilled model outperforms its teacher model, benefiting from natural regularization and improved hallucination management.

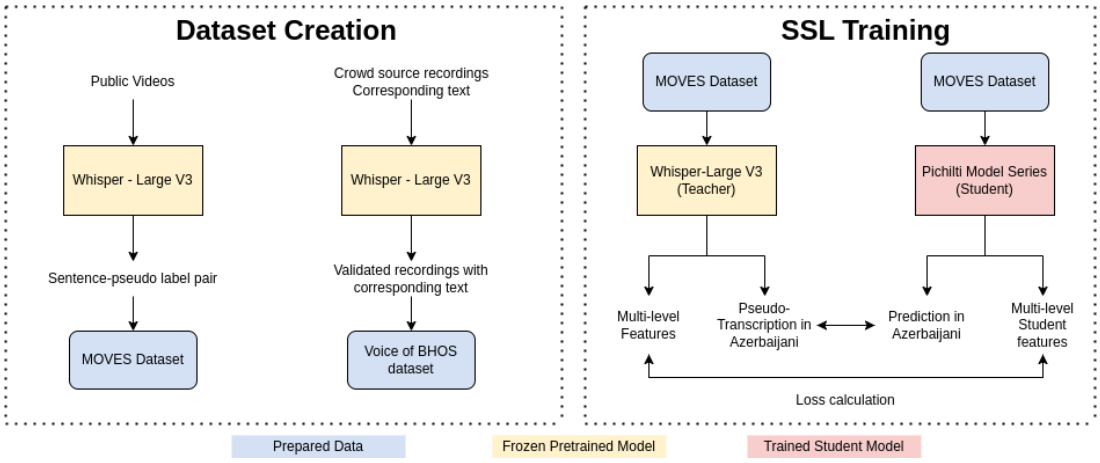

Fig. 1. 2 parts of the work done for this project. Dataset created for training and benchmarking, while the same dataset has been used for the SSL techniques

This research has been conducted in several stages:

- Public dataset analysis: Publicly available datasets were categorized into two groups—crowdsourced and professionally labeled—to assess their quality and suitability.
- Whisper model evaluation: A detailed analysis of the Whisper model was performed to understand feature distribution across different types of input data.
- Ablation study: A series of ablation experiments were conducted to evaluate the impact of training at different levels.

- Benchmark development: A dedicated benchmark dataset and evaluation framework were created to facilitate future research in this domain.

2. Literature Review

The popularity of speech recognition tasks is increasing every day as there are a lot of use cases. However, in order to conduct literature, review efficiently we divided this section into 3 subsections.

3.1 Speech Recognition Models

Selection of the model architecture plays important role in this study. Strength of the model is not only the training strategy but also the model architecture. Therefore, several surveys have made extensive studies. [3] have separated the Automatic Speech Recognition models into categories like Hybrid Models with Hidden Markov Chain, Transformer based architectures. Authors also paid attention to different training setups like Deep Transfer Learning (DTL), Federated Learning and Deep Reinforcement Learning.

Wav2Vec [4] is one of the most selected models which is based on Convolution inside transformers model and GeLU Activations. The model training contains 2 steps motivated by BERT training: Pre-training and fine-tuning. This allows to train over unlabeled large-scale data and adapt it to the given language with labelled data. Contrastive features are also used in order to develop distinct features according to the voice patterns. [5] proposes a solution to increase robustness over voice data. Authors also draw attention to disentanglement with extensive ablation study. Additional studies and methodologies have been carried out in [6], [7] by adding multimodality and increasing efficiency.

Whisper model by OpenAI is robust model based on encoder-decoder style transformers [1]Authors specifically mentioned that in order to study strength of the method, they did not make any modification over the architecture. Model has been trained on 680k hours of unlabeled data with zero-shot learning for multilingual setup. Model supports voice to text conversion. Additionally, voice in any language can be converted to English text with given model. Model outperforms most of the other open-source models while in multi-lingual settings only largest model size can perform in compromising way for low-resource languages. Additional trainings and fine-tunings based on whisper model have been proposed in [8], [9], [10] for adding specific languages to "knowledge" of the model or strengthening the given language more.

3.2 Speech Recognition in Azerbaijani

Considering that digital maturity of the Azerbaijani language is very low, research in field of Automatic Speech recognition has not been carried out a lot unfortunately. Some works have been published but without academic paper. [11] has fine-tuned the model based on CommonVoice dataset for Azerbaijani based on whisper. We have used this model for benchmarking. There are some cloud solutions in order to perform automatic speech recognition tasks, while their details are not exposed. Several questions remain open like: Which data has been used for training, what is the model architecture, which methodology has been used. Apart from Cloud solutions, only OpenAI's Whisper [1] model can perform in Azerbaijani voice transcriptions. The main reason for this capability is zero-shot learning technique as massive corpus (680k hours of data) have been utilized in order to increase the accuracy and robustness.

3.3 Speech Recognition for Low-resource languages

Considering number of languages in all over the world, maintaining the same digital maturity of language is not possible. Therefore, in order to train the model in Azerbaijani, other training methodologies for low-resource languages could be beneficial. [9] has conducted an experiment in which Whisper model has been fine tuned for Swiss-German language. Paper illustrates high performance results while methodology can seem little bit overhead. It is due to the fact that synthetically joining 2 different sounds (sentences) even with voice identification and additional effect can create confusion as model will get more contrastive pattern between 2 sentences. It will be problem when pattern similarity is high between consecutive sentences in speech which will fail sentence checkpointing task. [7] has created specific dataset for training in Malasar which was recorded in studio with clean environment. They used this dataset for multi-stage training and also brought middle level fine-tuning with more mature intermediate language. It is also mentioned in the paper that for automatic speech recognition tasks, there should be some noises, however it should be minor. This will bring the issue of robustness in different environments.

Training strategies have been also modified in several stages. [12] has modified loss function to draw more attention for the low-resource languages while [13] developed methodology to by using meta learning and KNN sampling. These approaches show tremendous amount of improvement over vanilla whisper model while both of them used common voice dataset, and benchmarks were not shown on Azerbaijani language.

3. Datasets

In the literature review, we discussed training approaches, model architectures, and strategies for improving efficiency, such as reducing the need for manual annotation. However, most studies reference commonly used voice datasets, such as Common Voice [14]. This dataset is a large-scale multilingual corpus containing samples from a wide range of languages. Upon extensive examination of different versions of this corpus, we found that the Azerbaijani language samples remained unchanged across all patches, with only 3–4 unique speakers present. Additionally, only less than 1000 samples were available. These limitations mislead researchers in benchmarking and fail to provide a robust evaluation of model performance. Noisy, high-variability data is crucial for effective automatic speech recognition (ASR).

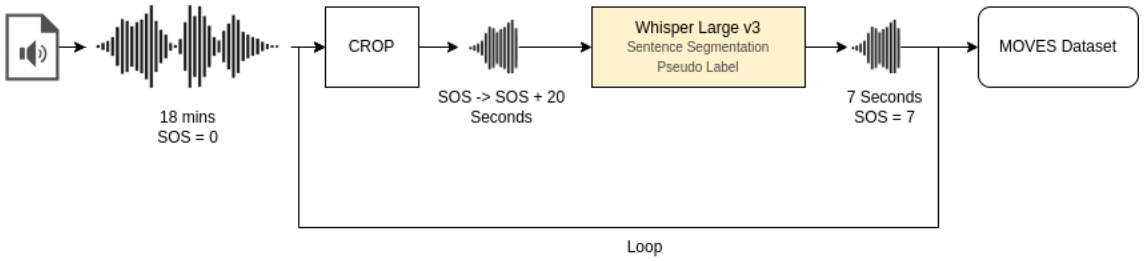

Fig. 2. Processing of 1 audio in a way that it does not lose any information. Dataset is composed of sounds which correspond to 1 sentence.

To address these issues, we developed two datasets: Massive Open Voices from Every Source (MOVES) for training and Voice of University for benchmarking.

- MOVES Dataset: This dataset was collected from various YouTube channels, with careful consideration given to licensing and real-world applicability. Over 500 hours of audio data were processed using Whisper Large V2 for sentence segmentation. To mitigate

hallucinations, a specialized segmentation method was implemented. Initial crowdsourcing efforts indicated that nearly all audio clips were shorter than 20 seconds, with each clip typically corresponding to a single sentence. Consequently, 20-second audio frames were extracted from the beginning of each recording to ensure consistency and facilitate data cleaning, particularly for automatic speech recognition (ASR) tasks. After segmenting each 20-second clip into sentences, the endpoint of the first sentence was recorded, and subsequent sentences were discarded due to the increased risk of hallucinations when processing the second and third sentences. The next 20-second frame was then extracted starting from the noted millisecond, ensuring complete sentence capture. Through this segmentation process, 250,000 data samples, along with their pseudo-labels, were collected (Figure 2). Data is available openly at https://huggingface.co/datasets/BHOSAI/MOVES_Azerbaijani

- Voice of University Dataset: This dataset was collected using a crowdsourcing approach, with participants from Baku Higher Oil School University. It consists of voice recordings from university students and is shared privately to ensure fair benchmarking. Out of 16,000 sentence-voice pairs, 14,000 were automatically validated using Whisper and Levenshtein distance metrics (see Equation 1). A subset of 2,000 high-quality recordings was selected and categorized into five groups: male speech, female speech, noisy speech, silence, and music. This categorization enables a comprehensive evaluation of model robustness. To support benchmarking by other researchers, we developed the Open Speech Leaderboard, a platform similar to the Open LLM Leaderboard, hosted on Hugging Face. The Voice of University data that support the benchmarks of this study are not publicly available due to fairness criteria in benchmarking. However, they are available from the corresponding author on reasonable request.

Levenshtein distance formula is given below:

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1 \end{cases} & otherwise \end{cases} \quad (1)$$

- a,b – the two strings under comparison.
- i – length of the prefix of string a a (i.e., first i i characters).
- j – length of the prefix of string b b (i.e., first j j characters).
- lev a,b (i,j) – the minimum number of edit operations needed to transform the prefix a[1..i] into b[1..j].

4. Methodology

The creation of the dataset allowed us to carry out self-supervised learning in offline manner. This saved us a GPU resource that we did not need to run inference of whisper large and target model to be trained in the same GPU at the same time for several iterations. However, with different experimental setup, we had to run inference of whisper large at the same time due to multi-level fine-tuning. In order to go to more details, we can refer to figure in which we show 2 different experiments.

Both experiments only modify the decoder part of the model (Figure 3). For substantiation, we created several versions of the same sentence in windy, noisy and silent environments. Additionally, we also created different synthetic versions of the same sentence with different genders. Then, we checked the distribution of output of the encoder to see the effect of the environment in feature extraction level. Our experiments showed that environment does not affect the features of the same sentence and voice owner while there was slight deviation in different voice characteristics. This

proves the audio-environment disentanglement characteristics and robustness of the encoder. Therefore, in order not to lead the catastrophic forgetting, we froze the encoder and let the fine-tuning to adapt only decoder part.

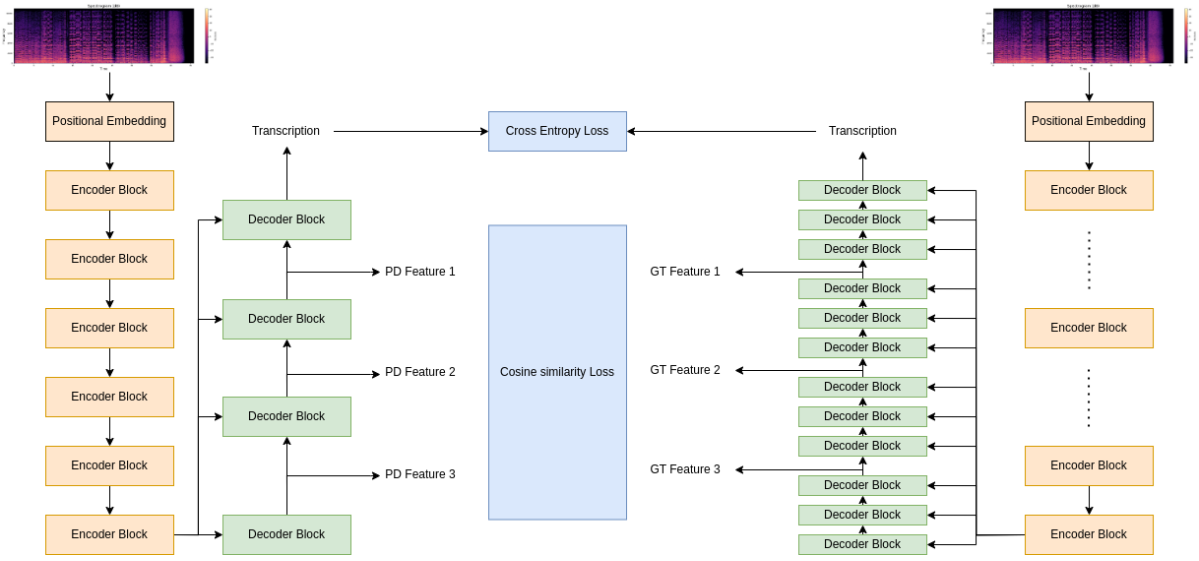

Fig. 3. Training pipeline for the base student model and large teacher model. In order to balance the number of feature outputs, group of decoder blocks have been created. Model on the left side is Teacher and right side is Student

- The first experiment is single-level fine-tuning of the whisper model. Here we take the voice as an input and pseudo-label from whisper large as our target (ground truth) then fine-tune it. It is knowledge distillation technique.
- The second experiment is called multi-level fine tuning. Here, we also make knowledge distillation in target output. However, in this experiment, we also make distillation in output of each decoder groups. So, we are calculating additional losses for the latent space in order to train the model.

5. Experimental Setup

We carried out those experiments for different model configurations and hyperparameters. Considering the strength of Knowledge distillation from the paper [15], if we make knowledge distillation in a way that teacher is large whisper and student is smaller whisper, we can bring the accuracy of the student much closer to teacher. If the teacher and student models are in the same size, e.g., both of them are large, there will be regularization effect and student model can slightly outperform the teacher model. Therefore, we made our experiments for different student models as can be seen in Table 1.

Table 1
Table of Student model sizes

Student name	*Size in Millions*
Tiny	39
Base	74
Small	244
Medium	769
Large	1550

For the experiments, we used 2 RTX4090 in custom distributed way. Trainings have been carried out with given configurations: Batch size - 16, learning rate - 0.0005, No need for gradient accumulation, weight decay - 0.01.

6. Results

For the first experiment with single-level fine-tuning in Table 2, we used offline self-supervised learning. In the first round of experiments in this setup, we noticed fluctuations regardless of the learning rate and other hyperparameters. The reason was hallucinations in pseudo-labels when there is no speech in given voice. Therefore, we used regular expressions in order to find those anomalies and remove the text and keep clean data for training.

Multi-level fine-tuning could not outperform the single level. As it is seen from the table, results are much closer to the large whisper model.

Table 2

Results of the experiment based on single-level training

Model Name	*CER %*	*WER %*
Whisper-tiny	94.5	100
Whisper-small	88.2	100
Whisper-base	86.2	76.5
Whisper-medium	52.8	64.1
Whisper-large-v3	12.3	28.2
Whisper large Azerbaijani	11.8	22.1
Pichilti-tiny (ours)	23.3	30.8
Pichilti-small (ours)	19.1	28.9
Pichilti-base (ours)	15.0	24.7
Pichilti-medium (ours)	10.8	18.8
Pichilti-large (ours)	7.4	14.4

Lack of the performance in Table 3 is due to strict guidance of the large model in also feature level. We analyzed this issue with ablation study given in Table 4. So, when we force the student less to imitate the teacher model, then the performance increases as the teacher model makes more mistakes.

Table 3

Results of the experiment based on multi-level training

Model Name	*CER %*	*WER %*
Whisper-tiny	94.5	100
Whisper-small	88.2	100
Whisper-base	86.2	76.5
Whisper-medium	52.8	64.1
Whisper-large-v3	12.3	28.2
Whisper large Azerbajiani	11.8	22.1
Pichilti-tiny (ours)	32.7	51.6
Pichilti-small (ours)	29.6	40.0
Pichilti-base (ours)	28.1	38.9

Table 4

Ablation study to understand the effect of the multi-level training with different number of stages on base student and large teacher model

Model Architecture	*CER %*	*WER %*
4 Decoder blocks	28.1	38.9
3 Decoder blocks	24.5	36.1
2 Decoder blocks	22.1	32.2
Last Layer	15.0	24.7

7. Conclusion

This paper highlights the necessity of self-supervised learning and its role in mitigating the challenges associated with data availability and computational requirements through experimental validation. Despite extensive research in self-supervised learning (SSL), it is preferable not to impose specific output constraints at different stages, as seen in multi-level training. Instead, single-level training allows the model to exert a greater regularization effect on the decoder blocks and improves error correction in large models through generalization, as supported by knowledge distillation research. As a result of this study, we have released the Pichilti model under an open-source license, providing a benchmarking framework for automatic speech recognition (ASR) models compatible with the Transformers library. The quantitative results demonstrate that leveraging unlabeled data with a greater variety of noise enhances model accuracy when employing the proposed methodology.

References

[1] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision (2022). https://arxiv.org/abs/2212.04356

[2] E. Trentin, M. Gori, A survey of hybrid ann/hmm models for automatic speech recognition, Neurocomputing. 37 No.1 (2001) pp.91-126. https://doi.org/10.1016/S0925-2312(00)00308-8

[3] H. Kheddar, M. Hemis, Y. Himeur, Automatic speech recognition using advanced deep learning approaches: a survey. Information Fusion 109, 102422 (Sep 2024). https://doi.org/10.1016/j.inffus.2024.102422

[4] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations (2020). https://arxiv.org/abs/2006.11477

[5] A. Gulati, J. Qin, C.C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-augmented transformer for speech recognition (2020). https://arxiv.org/abs/2005.08100

[6] X. Pan, P. Chen, Y. Gong, H. Zhou, X. Wang, Z. Lin, Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition (2022). https://arxiv.org/abs/2203.07996

[7] L.G. Pillai, K. Manohar, B.K. Raju, E. Sherly, Multistage fine-tuning strategies for automatic speech recognition in low-resource languages (2024). https://arxiv.org/abs/2411.04573

[8] S. Rijal, S. Adhikari, M. Dahal, M. Awale, V. Ojha, Whisper finetuning on nepali language (2024). https://arxiv.org/abs/2411.12587

[9] V. Timmel, C. Paonessa, R. Kakooee, M. Vogel, D. Perruchoud, Fine-tuning whisper on low-resource languages for real-world applications (2024). https://arxiv.org/abs/2412.15726

[10] M. Qian, S. Tang, R. Ma, K.M. Knill, M.J. Gales, Learn and don't forget: Adding a new language to asr foundation models (Sep 2024). https://doi.org/10.21437/interspeech.2024-1045

[11] Drishti Sharma: whisper-large-v2-azerbaijani (revision 9b5d30c) (2025). https://doi.org/10.57967/hf/3960

[12] A. Piñeiro-Martín, C. García-Mateo, L. Docio-Fernandez, M.D.C. López-Pérez, G. Rehm, Weighted cross-entropy for low-resource languages in multilin-gual speech recognition (Sep 2024). https://doi.org/10.21437/interspeech.2024-734

[13] M.H. Hsu, K.P. Huang, H. Yi Lee, Meta-whisper: Speech-based meta-icl for asr on low-resource languages (2024). https://arxiv.org/abs/2409.10429

[14] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F.M. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus (2020). https://arxiv.org/abs/1912.06670

[15] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations (2020). https://arxiv.org/abs/2002.05709