Lip2Seq: Sequence-to-Sequence Lip Reading via Phoneme Prediction and Mel Spectrogram Reconstruction

UP904749

Abstract—Despite considerable advances in lip-reading research using pre-trained audio-visual encoders and fine-tuned decoders, there remains a lack of extensive evaluation of potential shortcomings inherent to phoneme-centric methodologies. This paper aims to bridge this gap by scrutinizing different lip-reading video representations and their accompanying classifiers to ascertain the most effective approach. Furthermore, we extend our exploration to the simultaneous prediction of acoustic speech features and phoneme classes from video frames, driven by encouraging results obtained from the same audio-visual encoders. To ensure robust validation, we evaluate these tasks on two novel datasets specifically designed to represent a simplified version of real-world lip-reading scenarios.

Index Terms—Lip phoneme recognition, phoneme confusion, lip reading, speech reconstruction, multi-task learning

1 Introduction

Unlike studio recordings, speech in naturalistic settings is rarely free of noise: outside recordings suffer corruption from natural and man-made interference such as wind and traffic noise [1]. Whereas speech recorded indoors suffers corruption from mechanical noise and overlapping speech from non-target speakers [2].

This motivates the need for recognising speech in the presence of distorted or absent auditory information. One solution is lip reading, often referred to as visual speech recognition, which has gained significant traction in the speech recognition and speech synthesis domains [3]. Lip reading is the technique of understanding speech by visually interpreting the movements of the lips, face, and tongue.

There is active research into speech synthesis and recognition with only visual information. Prior synthesis methods relied only on visual information which holds incomplete information about speech [4], and has been demonstrated to have much lower performance compared to visual speech recognition [5][6][7][8].

However, recent advances such as Audio-Visual Hidden Unit BERT (AV-HuBERT) [9] provide a self-supervised representation learning framework utilising both audio and visual speech related information, with strong lip reading performance (achieving a 26.9% WER on LRS3). This paper provides a pre-trained encoder model which computes an audio-visual embedding of videos which can be used for vision-only speech related downstream tasks, such as speech recognition and speech synthesis [10]. Speech recognition in the AV-HuBERT paper is performed using either phoneme prediction with Connectionist Temporal Classification (CTC) [11], or a sequence-to-sequence (S2S) approach using sub-word units. The paper found that the S2S model provided superior performance, but the CTC approach had better performance with smaller datasets.

Other approaches [12] have recently achieved a phoneme accuracy rate of 70% which led to an 18% lower word-error rate compared with state-of-the-art lip-reading approaches, which provides compelling evidence to explore this approach further. Phoneme classes are also used as an input for many speech synthesis systems [13], [14], [15] which shows how phoneme classification can be used across lip reading tasks.

Despite the advantages, phoneme classification for lip reading suffers from the issue of mapping visemes (visual equivalent of phonemes) to phonemes, as multiple phonemes may appear the same on the lips (homophenous) [16].

This phoneme classification confusion poses a significant challenge to lip reading technologies, as the misclassification of phonemes can lead to inaccurate interpretation of visual cues. This could subsequently decrease the overall accuracy and efficiency of speech recognition or synthesis systems, impacting their real-world applicability.

Prior research has shown that discretised SSL (self-supervised learning) speech features from the HuBERT [17] model encode mostly phonemic information and less about speaker and noise characteristics, however, the multi-modal AV-HuBERT [9] approach captures linguistic and phonetic information from the lip movement and audio streams into its latent representation. This suggests that it might also be possible to predict the mel-spectrogram features from this audio-visual latent space.

To investigate the pitfalls of current phoneme classification systems and whether the auxiliary task of mel-spectrogram prediction from AV-HuBERT embeddings is possible, this paper proposes a novel sequence-to-sequence lip reading model called Lip2Seq, with a focus on the phoneme classification component. Dlib [18] facial landmarks and embeddings from the AV-HuBERT model are explored with a variety of classifiers to determine the best method for phoneme classification. The auxiliary mel-spectrogram prediction is achieved by adding a projection layer from the neural network to also predict the acoustic features.

Validation of the proposed approach is performed with

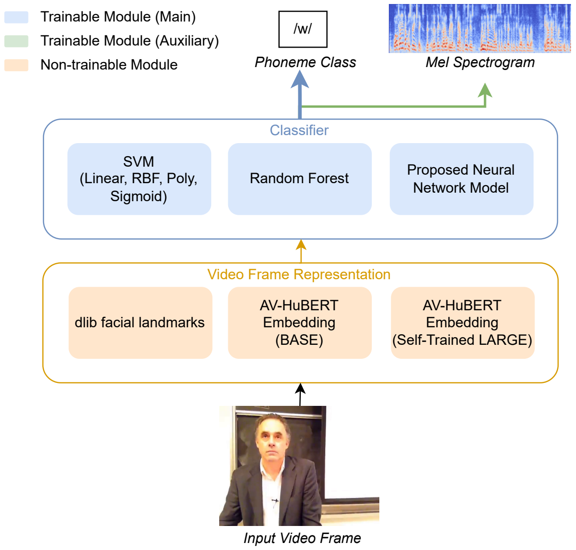
Trainable Module (Main)
Trainable Module (Auxiliary)
Non-trainable Module
/w/
Phoneme Class
Mel Spectrogram
Classifier
SVM (Linear, RBF, Poly, Sigmoid)
Random Forest
Proposed Neural Network Model
Video Frame Representation
dlib facial landmarks
AV-HuBERT Embedding (BASE)
AV-HuBERT Embedding (Self-Trained LARGE)
Input Video Frame

Fig. 1: Overview of the proposed multi-task model framework.

two custom datasets which are created to match the criteria of a single speaker talking with their mouth clearly visible, with these criteria explained further within the paper.

Thorough experimentation shows that the proposed method achieves strong phoneme classification results and provides an in-depth phoneme confusion analysis for the best performing, along with model analysis of the predicted mel-spectrogram features.

2 Method

Let $X = \{x1, ..., xT\} \in R^{TxLxC}$ be dlib predicted facial landmarks or $X = \{x1, ..., xT\} \in R^{TxD}$ a AV-HuBERT BASE or Self-Trained LARGE embedding, $Y = \{y1, ..., yT\} \in R^{TxP}$ be target phoneme classes for each video frame, and $A = \{a1, ..., aT\} \in R^{TxB}$ be the acoustic feature of ground-truth speech represented as a mel-spectrogram. T indicates the frame lengths (which are trimmed to the minimum of X, Y and A to resolve rounding errors), L and C are the number of facial landmarks and number of co-ordinates per landmark, respectively, D is the embedding dimension size for either of the AV-HuBERT models, P is the number of possible phoneme classes and B is the number of mel-spectrogram bins.

The main task is to translate the input video feature X into the phoneme class Y and for the auxiliary learning task, to simultaneously predict the mel-spectrogram features A. Training the model on both tasks simultaneously can risk performance degradation. To mitigate this, we apply a weight to both terms of the loss function for stabilisation. To this end, the joint loss function for the neural network with the auxiliary task is shown in Eq 1.

λ_{main} controls the phoneme classification loss and λ_{aux} controls the reconstruction loss. λ_{main} is set to 1.0 and λ_{aux} is set to 0.01 by default, respectively.

The proposed multi-task model framework is illustrated in Figure 1.

Refer to Appendix A for phoneme generation details.

$$\begin{aligned} \mathrm{L}_{Lip2Seq} = \lambda_{main} CrossEntropyLoss(\hat{Y},\ Y) \\ +\lambda_{aux} MSE(\hat{A},\ A) \end{aligned} \tag{1}$$

3 Experiments

3.1 Datasets

The following dataset creation criteria are used for this paper:

- Single speaker (Acoustic representations might contain different characteristics of speakers, tones, accents, etc, which makes it hard to infer mel-spectrogram features across different settings)
- Speakers mouth is constantly visible (as the lip movements are required to infer the phoneme and acoustic speech features)

Two novel datasets are constructed for this paper from public YouTube videos. The first video was chosen as it had a short duration and served to validate the proposed Lip2Seq model. The second dataset contains approximately 16.63 times more data (frame-wise) than the first. A reduced version of the ARPABET [19] phoneme dictionary is adapted for the phoneme class targets. Full details of both dataset sources and statistics are provided in Appendix B.

3.1.1 Jordan Peterson Shorts Video (Shorts)

This video lasts 51.38 seconds and runs at 30FPS, for a total of 1,233 frames. It contains 143 contiguous words and spaces. The video contains 540 phonemes in total.

3.1.2 Jordan Peterson Lecture Video (Lecture)

This video lasts 11 minutes and 23.72 seconds and runs at 30FPS, for a total of 20,532 frames. It contains 2,165 contiguous words and spaces. The video contains 7187 phonemes in total.

3.2 Implementation Details

Full source code can be found at: [1]

3.2.1 Video Frame Features

Two types of visual features are explored, raw facial landmarks inferred from the dlib library [18], which produces 68 facial landmarks with an x and y value each, where these features are included as a baseline. Also, the visual feature embeddings pre-computed from the AV-HuBERT [9] BASE and Self-Trained LARGE models, with 768 and 1024 dimensions, respectively, are used as they have known strong performance for phoneme classification based on the original paper [9].

[1] https://github.com/MiscellaneousStuff/comp-vis-avhubert

3.2.2 Dataset Preprocessing

For all datasets, the audio with a sample rate of 16kHz is converted into a mel-spectrogram using a hop length matching the input video frame temporal resolution (either 33ms for 30fps or 41ms for 24 fps), with excess items in the mel-spectrogram trimmed to match the length of the visual features, and a window size of 64ms. The mel-spectrograms have 80 channels, which matches many SOTA vocoders [20]. This allows seamless audio generation in future work. Refer to Appendix A for phoneme pre-processing details.

3.2.3 Model Details

For the phoneme classification task, multiple classifiers are explored, namely multiple SVM [21] variants (Linear, Polynomial, Radial Basis Function, Sigmoid) due to its known strong classification performance, Random Forest [22] (due to its data efficiency and robustness to raw data inputs) and the proposed custom PyTorch-based [23] neural network model. The architecture of the neural network model is a simple feedforward neural network where the input layer is the video frame features (dlib or AV-HUBERT embedding), this is projected to a hidden layer with a unit count of 256 or 512 (512 units are only used for the Self-Trained AV-HuBERT LARGE embedding) with a ReLU activation function, and then a final softmax layer is applied for the phoneme class prediction. For the auxiliary mel-spectrogram prediction task, another layer is projected from the hidden layer with an output dimension of 80 (number of mel channels per frame).

3.2.4 Training Details

A fixed random seed of 42 is used for all experiments to ensure reproducibility. All models are initially assessed by overfitting to the entire dataset. This process, while avoided in the final model, can indicate the viability of a model by showing whether it can potentially attain a 100% or near 100% phoneme accuracy rate, and whether it can accurately reproduce the mel-spectrogram for that task, based on MSE loss and visual inspection.

For the neural network training, the initial learning rate is set to $1e^{-4}$, a batch size of 1 and the AdamW optimiser [24] is used. A StepLR optimiser is used with a step size of 2500 and a gamma of 0.1, and the epoch count is 5,000. A dropout value of 0.5 is used during training for regularisation [25]. These hyperparameters were found by empirically testing which epoch saw the training loss continued to decrease but the test loss began to plateau or increase.

The random forest model uses the default number of estimators (100), and all of the SVM models use the default sklearn settings.

All training runs use the same shuffled dataset, per dataset, and use k-fold cross-validation with 5 splits, where the mean score of the 5 splits is recorded in the results.

3.2.5 Evaluation Metrics

For the phoneme classifiers, the classification accuracy is used to evaluate all of the models. For the best performing model for each dataset, the precision and recall are used to evaluate the models for each phoneme class. For the auxiliary mel-spectrogram synthesis, change in phoneme accuracy and the MSE Loss of the reconstructed mel-spectrogram is used evaluate the effect of the auxiliary tasks loss weighting on the multi-task model, and to determine reconstruction fidelity, respectively.

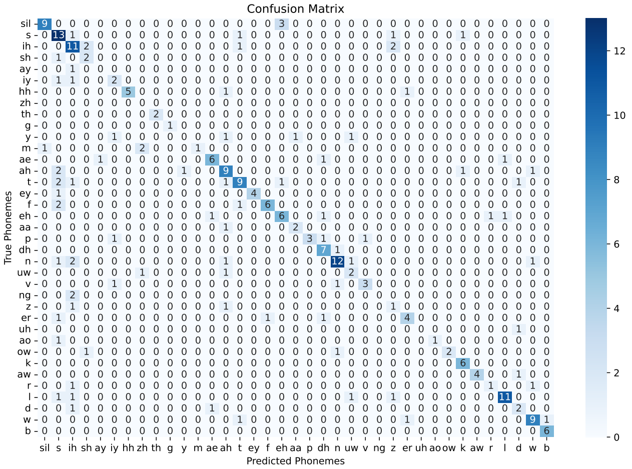

Fig. 2: Confusion Matrix for Shorts Dataset (AV-HuBERT Large and 512 dim Neural Network)

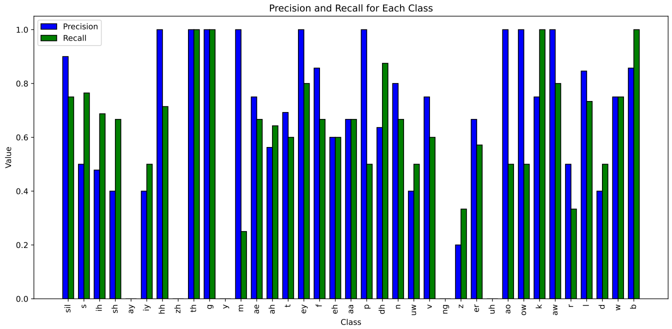

Fig. 3: Precision and Recall per Phoneme Class for Shorts Dataset (AV-HuBERT Large and 512 dim Neural Network)

3.3 Discussion

The phoneme classifiers are examined per dataset, with the shorts then lecture dataset results being presented.

3.3.1 Feature and Classifier Analysis

Tables 1 and 2 show the shorts and lecture dataset phoneme classification results, respectively. For both datasets, going from dlib facial landmarks to the Av-HuBERT BASE embeddings significantly improves performance (24.67% to 56.95% and 12.47% to 58.02%), with the best model being the Lip2Seq model for both datasets (63.56% and 65.65%).

However, going from the BASE to LARGE Av-HuBERT model only benefits the SVM (Linear) and Lip2Seq models (64.31% to 65.06% and 65.65% to 65.82%, respectively) for the Lecture dataset, and reduces performance for everything else. This is possibly due to the SVM (Linear) and Lip2Seq models being able to deal with the larger feature count (1024) of the LARGE Av-HuBERT embedding.

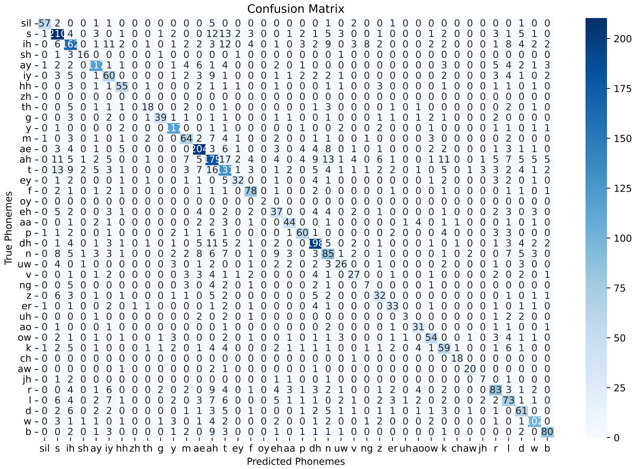

Fig. 4: Confusion Matrix for Lecture Dataset (AV-HuBERT Large and 512 dim Neural Network)

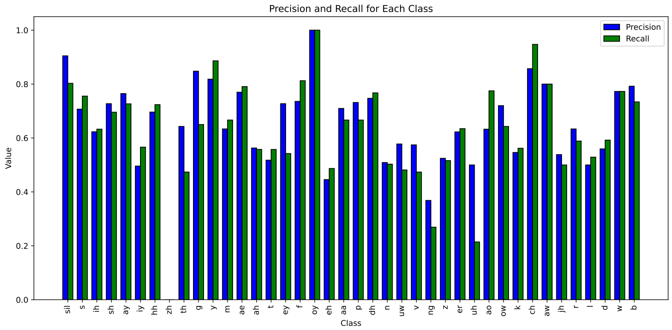

Fig. 5: Precision and Recall per Phoneme Class for Lecture Dataset (AV-HuBERT Large and 512 dim Neural Network)

The best performing model for both datasets was the Lip2Seq model with the LARGE AV-HuBERT features. Whereas for the dlib features, the Random Forest performs the best (49.91% and 29.44%), likely due to it being best suited to the raw inputs and lower dataset sizes. This is further demonstrated as it's performance drops the most out of all classifiers (-20.47%) when applied to the lecture dataset, instead of the shorts one.

3.3.2 Precision, Recall and Phoneme Confusion

The precision and recall values per phoneme class, per dataset are shown in Figures 3 and 5, for the best performing model for each dataset. The macro precision and recall (0.6314, 0.5713 for shorts and 0.6459, 0.6241 for lecture), show that the precision remained similar between datasets, but the recall improved significantly (+0.0528).

The improved recall is likely due to more examples of minority classes such as ″th″ and ″y″, which aren't present in the shorts dataset (refer to Fig 7) but is in the lecture dataset (refer to Fig 8), as the recall values for ″th″ and ″y″ both go from 0 to roughly 0.5 and 0.9, respectively.

The confusion matrices for the Lip2Seq Large AV-HuBERT features are displayed in Figures 2 and 4. Only the confusion matrix for the lecture dataset is considered, due to the multiple missing phonemes for the shorts dataset confounding the analysis. Interestingly, phonemes which would possibly be expected to be confused (plosives: ″b″, ″p″, which are phonemes which are produced by stopping the airflow using the lips, teeth or palate) are almost never confused for one another (only 1 prediction for ″p″ when it was truly ″b″, and 0 vice-versa) which makes sense given both of their high precisions (roughly 0.8 and 0.75 for ″p″ and ″b″, respectively). In fact, ″b″ is more likely to be confused as ″ah″ (9 false predictions) compared to any other phoneme.

Another interesting finding is that the ″sil″ silence phoneme has a high precision and recall (0.9, 0.8), which means that the classifier is good at predicting when a person is not speaking. This is a useful property of the classifier as identifying breaks in speech is essential to accurate continuous speech recognition (i.e, being able to discern individual words and pauses).

3.3.3 Mel-Spectrogram Prediction

The results of the Lip2Seq models simultaneous mel-spectrogram reconstruction and phoneme classification are summarized in Table 3, which presents the impact of altering the λ_{aux} value across a range of weights on phoneme classification accuracy and Mean Squared Error (MSE) loss for the mel-spectrogram reconstruction. This is performed using the AV-HuBERT Large Lip2Seq model.

Interestingly, the results indicate an inverse relationship between the weightings for the mel-spectrogram reconstruction loss and phoneme accuracy. This suggests that as we put more emphasis on minimizing the mel-spectrogram reconstruction loss, the model's ability to accurately classify phonemes diminishes. One possible explanation for this trend is that the neural network finds it challenging to learn a representation over the LARGE Av-HuBERT embedding that can be used to simultaneously predict the phoneme class and reconstruct the mel-spectrogram features.

Moreover, the MSE loss values suggest that the model is most successful at reconstructing the acoustic features when the auxiliary learning task, the mel-spectrogram reconstruction, is given a small weighting of 0.01. In fact, the MSE loss seems to increase (i.e., the reconstruction becomes worse) as the weighting for the auxiliary learning task increases. This relationship demonstrates the challenges the model faces when trying to optimize for two objectives at the same time.

To further analyze these findings, Figure 6 provides a comparison between the predicted and ground truth mel-spectrograms on a validation set. This validation set is a separate 10% slice of the original dataset, which was not shuffled and was kept separate from the training and test sets. The shuffling of data during model training was deliberately avoided in order to maintain the same phoneme distribution in both training and test sets, especially for the shorts dataset.

Visual inspection of the predicted and ground truth mel-spectrograms reveals that the model has effectively learned the boundaries between predicted phonemes, as indicated by the distinct gaps between the roughly orange rectangular blocks in the mel-spectrograms. Nonetheless, the model

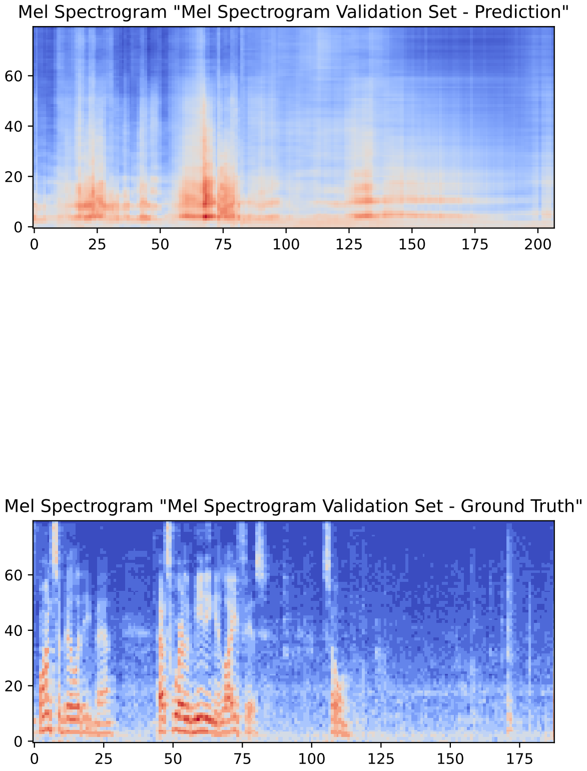

Fig. 6: $\lambda_{aux} = 0.01$ Predicted and Ground Truth Mel-Spectrogram Snippets (AV-HuBERT Large and 512 dim Neural Network)

seems to struggle with capturing the finer details of the phonemes themselves.

There are two primary possible explanations for this limitation. Firstly, it could be the case that the LARGE Av-HuBERT embedding does not contain sufficient information to discern pitch waveforms for the predicted phoneme. This possibility is supported by findings from the original HuBERT paper [17], which indicated potential limitations in the information capacity of the Av-HuBERT embeddings.

Alternatively, the challenge could lie in the architecture of our model. Our current model uses a linear projection for predicting the mel-spectrograms, which might not be adequate for capturing the complexity of the acoustic features. Future work could explore sequential architectures, such as Transformer [26] or LSTM [27] architectures, or additional training strategies to better capture these features and further improve the model's performance in both phoneme classification and mel-spectrogram reconstruction.

Model	Dlib (Phone Acc)	Av-HuBERT BASE (Phone Acc)	Av-HuBERT LARGE (Phone Acc)
SVM (Linear)	13.14%	63.15%	62.09%
SVM (Poly)	22.80%	48.05%	45.78%
SVM (RBF)	15.50%	56.81%	53.73%
SVM (Sigmoid)	9.00%	55.76%	54.22%
Random Forest	**49.91**%	54.38%	48.86%
Lip2Seq	37.46%	**63.56%**	**65.58%**
Mean	24.64%	56.95%	55.04%

TABLE 1: Feature-wise Comparison of Each Classifier for Shorts Dataset

Model	Dlib (Phone Acc)	Av-HuBERT BASE (Phone Acc)	Av-HuBERT LARGE (Phone Acc)
SVM (Linear)	8.16%	64.31%	65.06%
SVM (Poly)	8.23%	56.21%	56.69%
SVM (RBF)	8.03%	60.90%	57.33%
SVM (Sigmoid)	7.43%	45.84%	44.16%
Random Forest	**29.44%**	55.53%	52.79%
Lip2Seq	13.51%	**65.65%**	**65.82%**
Mean	12.47%	58.07%	56.98%

TABLE 2: Feature-wise Comparison of Each Classifier for Lecture Dataset

λ_{aux}	Phoneme Accuracy	MSE Loss
0.00	65.82%	N/A
0.01	63.75%	7.5695
0.10	63.67%	8.0597
0.50	62.94%	8.7936

TABLE 3: Comparison of λ_{aux} Weighting on Phoneme Accuracy and Mel-Spectrogramm Reconstruction Loss. (AV-HuBERT Large and 512 dim Neural Network)

4 Conclusion

In this paper, a sequence-to-sequence model is proposed which linearly projects features from a pre-trained audio-visual encoder model for phoneme classification, with another projection used for acoustic feature prediction. The experimental results prove that the approach performs well for phoneme classification but the mel-spectrogram synthesis performed poorly and may require a different architecture or other approach.

Acknowledgments

I would like to extend my gratitude to Mr. Dinghuang Zhang, my lecturer, for taking the time to entertain my many project ideas throughout the semester, useful critique of the ideas I presented and the encouragement throughout.

References

[1] J. P. Bello, C. Silva, O. Nov, *et al.*, *Sonyc: A system for the monitoring, analysis and mitigation of urban noise pollution*, 2018. arXiv: 1805.00889 [cs.SD].

[2] S. Watanabe, M. Mandel, J. Barker, *et al.*, *Chime-6 challenge:tackling multispeaker speech recognition for unsegmented recordings*, 2020. arXiv: 2004.09249 [cs.SD].

[3] A. Fernandez-Lopez and F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning," *Image and Vision Computing*, vol. 78, pp. 53–72, 2018, ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2018.07.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885618301276.

[4] M. Kim, J. Hong, S. J. Park, and Y. M. Ro, "Cromm-vsr: Cross-modal memory augmented visual speech recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 4342–4355, 2022. DOI: 10.1109/TMM.2021.3115626.

[5] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, *Video-driven speech reconstruction using generative adversarial networks*, 2019. arXiv: 1906.06301 [eess.AS].

[6] D. Michelsanti, O. Slizovskaia, G. Haro, E. Gómez, Z.-H. Tan, and J. Jensen, *Vocoder-based speech synthesis from silent videos*, 2020. arXiv: 2004.02541 [eess.AS].

[7] M. Kim, J. Hong, and Y. M. Ro, *Lip to speech synthesis with visual context attentional gan*, 2022. arXiv: 2204.01726 [cs.CV].

[8] R. Mira, K. Vougioukas, P. Ma, S. Petridis, B. W. Schuller, and M. Pantic, "End-to-end video-to-speech synthesis using generative adversarial networks," *IEEE Transactions on Cybernetics*, vol. 53, no. 6, pp. 3454–3466, Jun. 2023. DOI: 10.1109/tcyb.2022.3162495. [Online]. Available: https://doi.org/10.1109%2Ftcyb.2022.3162495.

[9] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, *Learning audio-visual speech representation by masked multimodal cluster prediction*, 2022. arXiv: 2201.02184 [eess.AS].

[10] 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*, 2022. arXiv: 2212.11377 [eess.AS].

[11] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 369–376, ISBN: 1595933832. DOI: 10.1145/1143844.1143891. [Online]. Available: https://doi.org/10.1145/1143844.1143891.

[12] R. El-Bialy, D. Chen, S. Fenghour, *et al.*, "Developing phoneme-based lip-reading sentences system for silent speech recognition," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 129–138, 2023. DOI: https://doi.org/10.1049/cit2.12131. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/cit2.12131. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cit2.12131.

[13] H. Kim, S. Kim, and S. Yoon, *Guided-tts: A diffusion model for text-to-speech via classifier guidance*, 2022. arXiv: 2111.11755 [cs.SD].

[14] Y. Ren, C. Hu, X. Tan, *et al.*, *Fastspeech 2: Fast and high-quality end-to-end text to speech*, 2022. arXiv: 2006.04558 [eess.AS].

[15] K. Shen, Z. Ju, X. Tan, *et al.*, *Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers*, 2023. arXiv: 2304.09116 [eess.AS].

[16] H. L. Bear and R. Harvey, "Phoneme-to-viseme mappings: The good, the bad, and the ugly," *Speech Communication*, vol. 95, pp. 40–67, 2017, ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2017.07.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639317300286.

[17] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 3451–3460, Oct. 2021, ISSN: 2329-9290. DOI: 10.1109/TASLP.2021.3122291. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291.

[18] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[19] D. Jurafsky, A. Bell, E. Fosler-Lussier, C. Girand, and W. Raymond, "Reduction of english function words in switchboard," vol. 7, Nov. 1998. DOI: 10.21437/ICSLP.1998-801.

[20] J. Kong, J. Kim, and J. Bae, *Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis*, 2020. arXiv: 2010.05646 [cs.SD].

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[22] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, IEEE, vol. 1, 1995, pp. 278–282.

[23] A. Paszke, S. Gross, F. Massa, *et al.*, *Pytorch: An imperative style, high-performance deep learning library*, 2019. arXiv: 1912.01703 [cs.LG].

[24] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, 2019. arXiv: 1711.05101 [cs.LG].

[25] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014. [Online]. Available: http://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf.

[26] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] S. Tomar, "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[29] jianfch, *Stable-ts: Asr with reliable word-level timestamps using openai's whisper*, 2023. [Online]. Available: https://github.com/jianfch/stable-ts.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022. arXiv: 2212.04356 [eess.AS].

[31] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," Aug. 2017, pp. 498–502. DOI: 10.21437/Interspeech.2017-1386.

Appendix A
Phoneme Dictionary, Generation and Preprocessing

A.1 Phoneme Dictionary

The phoneme dictionary used for this paper is the ARPABET [19] dictionary because the version of the Montreal Forced Aligner used to generate the phonemes was trained on the ARPABET phoneme dictionary, as is commonly used in speech recognition. However, the original version of the ARPABET dictionary resulted in a dictionary of 62 different phoneme classes which was excessive and many of them had overlapping roles. Therefore the phoneme dictionary used was reduced from the original generated phonemes down to 40 phoneme classes.

For the convenience of the reader, the mapping of phonemes from the original ARPABET predictions to the IPA phonetic alphabet is provided in Table 4.

A.2 Phoneme Generation

The phoneme class targets for each dataset are generated using the following procedure: the MP3 file of the source video is extracted using ffmpeg [28], then a modified version [29] of the large OpenAI Whisper model [30] is used to generate a word-level transcription of the audio and then a script is used to reproduce the word-level transcript into the TextGrid format expected by the Montreal Forced Aligner (MFA) [31].

At this stage, MFA was often failing to compute the phoneme predictions on the generated TextGrid file as any

Phoneme Category	ARPABET Phoneme	IPA Phoneme
Nasals	m	m
	n	n
	ng	ŋ
Plosives	p	p
	b	b
	t	t
	d	d
	k	k
	g	ɡ
Fricatives	s	s
	z	z
	f	f
	v	v
	th	θ or ð
	sh	ʃ
	zh	ʒ
	hh	h
Affricates	ch	tʃ
	jh	dʒ
Approximants	r	r
	l	l
	w	w
	y	j
Vowels	ih	ɪ
	ey	eɪ
	ae	ae
	aa	ɑ
	ow	oʊ
	uw	u
	ah	υ or ə
	er	ɜʿ or ɜr
	uh	ʊ
	ao	ɔ
	oy	ɔɪ
	ay	aɪ
	aw	aʊ
	eh	ɛ
	iy	i
Liquids	l	l
	r	r
Glides	w	w
	y	j
Voiced	z	z
	v	v
	zh	ʒ
	dh	ð
	m	m
	n	n
	ng	ŋ
	r	r
	l	l
	w	w
	y	j
	b	b
	d	d
	g	ɡ
	jh	dʒ
Voiceless	s	s
	f	f
	th	θ
	sh	ʃ
	hh	h
	p	p
	t	t
	k	k
	ch	tʃ

TABLE 4: ARPABET to IPA Phoneme Mapping

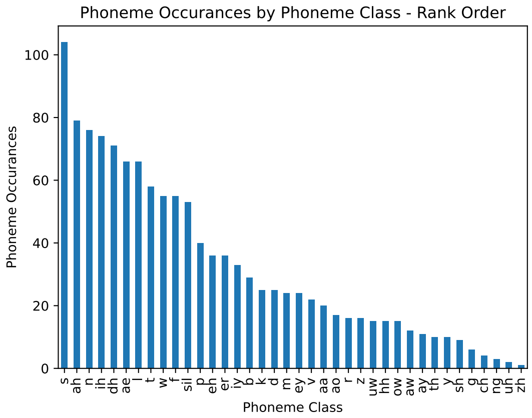

Fig. 7: Phoneme Distribution of Shorts Video

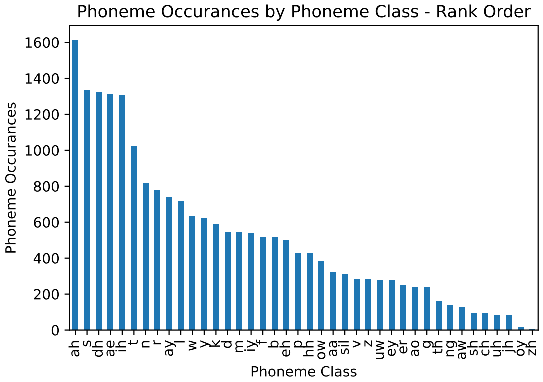

Fig. 8: Phoneme Distribution of Lecture Video

two overlapping word-level predictions would cause it to fail, therefore the TextGrid transcription was manually modified to fix these errors. The iterative process of fixing word-level timestamps and MFA validating the TextGrid file was not illustrated. Finally, the predicted phonemes are returned from MFA as a Phoneme TextGrid file containing the timestamps for each word and phoneme for that dataset. The entire procedure, along with required manual intervention and file types at each stage are illustrated in Figure 9.

A.3 Phoneme Preprocessing

The phoneme class targets are generated by getting word-aligned transcripts for the audio files using the Whisper ASR model [30], and then passing them to the Montreal Forced Aligner [31] to get the predicted phonemes with 10ms temporal resolution. These phoneme classes are interpolated to match the temporal resolution of the mel-spectrogram features and video frame features.

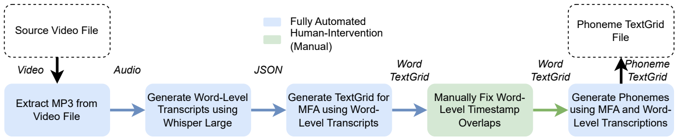
Source Video File
Fully Automated
Human-Intervention (Manual)
Phoneme TextGrid File
Video
Audio
JSON
Word TextGrid
Word TextGrid
Phoneme TextGrid
Extract MP3 from Video File
Generate Word-Level Transcripts using Whisper Large
Generate TextGrid for MFA using Word-Level Transcripts
Manually Fix Word-Level Timestamp Overlaps
Generate Phonemes using MFA and Word-Level Transcriptions

Fig. 9: Phoneme Generation Procedure

Appendix B
Dataset Sources

B.1 Jordan Peterson Shorts Video Details

Figure 7 shows the phoneme class distribution for the shorts dataset and how it is characterised by a heavy tail, with a small number of phoneme classes (s, ah, n, ih, dh, etc.) having at least 80 phoneme instances or above, whereas (sh, g, ch, ng, uh and zh all have under 10). The silence ”sil” phoneme class occurs 4.29% out of the phoneme instances.

- **Title:** The False Appeal of Communism
- **URL:** https://www.youtube.com/watch?v=wsDmwoOrpR8
- **FPS:** 23.976024
- **Width, Height:** 406x720
- **Duration:** 51 seconds

B.2 Jordan Peterson Lecture Video Details

Figure 8 shows the phoneme class distribution for the lecture dataset and how it is characterised by a heavy tail, with a small number of phoneme classes (ah, s, dh, ae, etc.) all being represented over a 1,000 times out of the 20,532 phoneme instances, and many of the classes at the opposite end of the distribution (ng, aw, sh, ch, etc.) barely occuring 100 times. The silence ”sil” phoneme class occurs 1.51% out of the phoneme instances.

- **Title:** Jordan Peterson on the meaning of life for men. MUST WATCH
- **URL:** https://www.youtube.com/watch?v=NX2ep5fCJZ8
- **FPS:** 29.970030
- **Width, Height:** 1280x720
- **Duration:** 11 minutes and 24 seconds