



The role of audio-visual integration in the time course of phonetic encoding in self-supervised speech models

Yi Wang¹, Oli Danyi Liu¹, Peter Bell¹

¹University of Edinburgh, United Kingdom

Wang.Yi@ed.ac.uk, Oli.Liu@ed.ac.uk, Peter.Bell@ed.ac.uk

Abstract

Human speech perception is multimodal. In natural speech, lip movements can precede corresponding voicing by a non-negligible gap of 100-300 ms, especially for specific consonants, affecting the time course of neural phonetic encoding in human listeners. However, it remains unexplored whether self-supervised learning models, which have been used to simulate audio-visual integration in humans, can capture this asynchronicity between audio and visual cues. We compared AV-HuBERT, an audio-visual model, with audio-only HuBERT, by using linear classifiers to track their phonetic decodability over time. We found that phoneme information becomes available in AV-HuBERT embeddings only about 20 ms before HuBERT, likely due to AV-HuBERT's lower temporal resolution and feature concatenation process. It suggests AV-HuBERT does not adequately capture the temporal dynamics of multimodal speech perception, limiting its suitability for modeling the multimodal speech perception process.

Index Terms: audio-visual speech processing, multimodal speech perception, self-supervised learning representation

1. Introduction

Recent years have seen a rising trend in using artificial neural networks (ANN) to model sensory processes in humans [1]. For speech processing, it has been shown that the neural activity in the auditory cortex of human listeners correlates with representations in a recurrent neural network model trained for speech recognition, with correspondence between levels of language processing and different layers within the model [2]. Apart from encoding hierarchy, their findings also revealed striking resemblance between the RNN model and the human brain in terms of the temporal order of linguistic representations. This comparison between model representations and brain activities could potentially help us identify the brain area and the time course in which different levels of language processing unfolds.

While various characteristics of speech processing have been examined in ANN models in comparison to human listeners, most existing work has focused on models with only audio input. In reality, human speech perception process is fundamentally multimodal, involving audio-visual integration, as illustrated by the McGurk effect [3], in which contradictory audio and visual cues cause a shift in human phoneme perception. While previous work on multisensory integration has studied how simple motion and spatial cues are integrated [4], the integration mechanism of higher-order semantically complex information, such as audio-visual cues in speech, remains unclear.

Some prior work [5, 6] have used audio-visual speech recognition models as a tool for investigating audio-visual speech perception. In contrast to [2], they aimed at answering the developmental questions of brain speech processing,

and trained ASR models under specific conditions to see if such conditions can trigger the development of corresponding characteristics in ANN models. For example, audio-visual ANNs trained by unsupervised learning proposed by Ngiam et al. [5] did exhibit the McGurk effect. Furthermore, a subsequent study [6] reported that the McGurk effect could emerge without specific training on the latest audio-visual self-supervised learning (SSL) speech representation model, Audio-visual HuBERT (AV-HuBERT) [7]. They thus suggested that audio-visual self-supervised learning (SSL) models could be useful for simulating how humans perform audio-visual integration.

A crucial factor in audio-visual integration during speech perception, overlooked in [6], is the earlier onset of the visual signal relative to the corresponding audio signal. Statistical analyses indicate that in natural speech production, mouth movements precede the resulting vocalization by approximately 100 to 300 milliseconds in the bilabial and labial consonants they tested [8]. This lead time, known as the "time-to-voice," reflects a fundamental characteristic of audio-visual speech timing. Humans have adapted to such asynchronicity, as evidenced by the reduction in amplitude and latency of auditory event-related potentials when speech audio is paired with visual stimuli [9, 10, 11]. However, it is unexplored to what extent incorporating visual cues accelerates the decodability of phonemes in ANN representations. If an ANN model exhibits little to no difference in the time course of phonetic encoding, the validity of using the model to simulate temporally sensitive aspects of speech perception would be significantly compromised.

In this paper, we investigate whether the incorporation of visual signals affects the time course of phonetic encoding in a self-supervised learning (SSL) speech model, as it does with human listeners. To this end, we compare an audio-visual model, AV-HuBERT, with its audio-only counterpart, HuBERT, by examining the window of phonetic decodability [12, 13]. Using a set of linear classifiers, we obtain a fine-grained view of when information about a phone becomes available in the model before its onset in the acoustics, with a temporal resolution up to the frame rate of the representations. In this way, we model how much of the temporal dynamics of visual and audio signals in speech can be encoded by the SSL representations.

We find that while visual cues can precede acoustic cues by hundreds of milliseconds, they advance phonetic decodability in AV-HuBERT over HuBERT only by approximately 20 ms. This limited effect is likely due to AV-HuBERT's down-sampling on the acoustic features by stacking every four frames of acoustic features to align with each visual feature frame. This process advances the audio features by an average of approximately 20 ms from their original timing. Furthermore, by manually introducing greater asynchronicity between visual and acoustic signals when depicting decodability curves, we find that AV-

HuBERT’s phonetic encoding is more strongly influenced by acoustic than visual signals. Overall, our findings suggest that greater caution is needed when using AV-HuBERT to model temporally sensitive aspects of multimodal speech perception.

2. Related works and method

Our analysis relies on using pre-trained SSL models as feature extractors, and training phoneme classifiers to identify the decodable window of a phone.

2.1. HuBERT and Audio-visual HuBERT

We used an audio-only and an audio-visual SSL model as the feature extractors for the corresponding modality. We chose the AV-HuBERT model [7] and its audio-only counterpart HuBERT [14], as AV-HuBERT has been the most widely used SSL audio-visual model. Both models were pre-trained by the masked prediction of hidden units method [14]. For abbreviation, HuBERT and AV-HuBERT will hereinafter be addressed as “HuBERTs” if being described together.

AV-HuBERT shares a similar masked prediction pre-training paradigm as HuBERT, but with the following modifications: 1) involving cross-modal transfer learning in the first pre-training iteration by using the pseudo labels extracted from acoustic features (MFCCs) as the initial pre-training target; 2) adding a deeper feature extractor ResNet-18 [15] to encode visual inputs; 3) applying early fusion at feature level by concatenation before feeding the audio and visual features into the transformer structure. More specifically, to align the audio log Mel-filterbank energy feature at 100 Hz to the visual ResNet-18 feature at 25 frames per second, every four audio feature frames are stacked and concatenated with a visual feature. The amount and domain of data being used for pre-training AV-HuBERT and HuBERT are also different.

Our analysis in Section 4 conducts a normalization step on the decodability curve to mitigate the impact from such design differences, and will focus on revealing the characteristic of AV-HuBERT more by comparing the peak time of its decodability curve to the phone onset time.

In the feature extraction steps of all experiments in this work, we used the publicly available HuBERT BASE¹ pre-trained on the 960-hour subset of LibriSpeech, and the AV-HuBERT BASE model² pre-trained on the 433-hour subset of LRS3, both model consisting of 12 transformer blocks.

2.2. Phonetic decodability window

Previous works [13, 16] decoded phone identity from brain recordings within a context window of 200-400 ms centered at the phone onset. The analysis was further applied in [12] on audio-only SSL speech representations [17] to observe encoded temporal dynamics.

Taking a context window with duration T centered at phone onset $[-\frac{T}{2}, \frac{T}{2}]$, it covers $2w + 1 = \frac{T}{\text{sample.rate}}$ frames of speech features $f_{-w:w}$. Given a set of data consists of N phone-level speech sequences $s_{1:N}$ cut from utterances and labelled by phone identity $p_{1:N}$ obtained from forced-alignment, we use $f_{i,j}$ to represent feature in phone-level sequence s_i at time step j relative to the center. All frames that share the same time step j relative to the phone onset $F^j = f_{1:N,j}$ can be paired with phone labels $p_{1:N}$ to train a Logistic Regression classifier C^j on phone classification. When testing the classifier C^j on a batch of evaluation speech feature frames at time step j ,

¹<https://github.com/facebookresearch/fairseq/tree/main/examples/hubert>

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

i.e. $f_{1:N,j}^{eval}$, the average classification accuracy Acc_j reflects the ability of decoding the phone identity from the SSL model feature j steps preceding / following the phone onset. The curve drawn from predictive accuracy score sequence $Acc_{-w}:Acc_w$ versus the relative time steps in the context window $-w : w$ reflect how the temporal dynamics of speech input is captured by the SSL model features.

We extracted HuBERT and AV-HuBERT features from speech within the context window of each phone, and drew the predictive decoding curve based on each type of features in a 1200 ms time window, which covers 30 frames of AV-HuBERT feature sequence or 60 frames of HuBERT feature sequence, resulting in 30 / 60 classifiers trained to draw an audio / audio-visual HuBERT decodability curve.

3. Task description

3.1. Dataset

We analyzed the phonetic decodability window on two datasets separately. The phone boundaries and labels for both datasets were obtained through the Montreal forced aligner on English arpabet dictionary [18, 19].

Grid: The GRID dataset [20] includes speech videos collected in a controlled environment, with 34 speaker sitting at the same spot relative to the camera, front-faced video and noise-free audio being recorded simultaneously. It consists of each speaker pronouncing 1000 sentences with a fixed six-word structure: command(4) color(4) preposition(4) letter(25) digit(10) adverb(4), with the number in parentheses indicating the vocabulary size at the corresponding position.

LRS3: The LRS3 dataset [21] includes TED and TEDx videos, along with subtitles and word boundaries, and has three sets: pre-train, train-val and test. In calculating the phonetic decodability in context window, we adopted LRS3 train-val split, consisting of 30 hours of speech from 4004 speakers.

Considering that utterances in Grid has little flexibility in its structure and limited vocabulary, each frame of SSL features may capture context information that covers neighbouring phones, allowing the classifier to distinguish phone identity from context phones with long time away from the target phone.

3.2. HuBERT models feature extraction settings

We conducted three-fold cross-validation on both datasets. Then in actual experiments, we randomly selected $\frac{1}{15}$ samples from both the training and the validation split of the data, further split them into 20 batches during LR training and inference, and averaged the resulting scores across the batches and folds but on each time step separately. Data sampling is due to the limitation of sklearn LR package .

4. Experiments

4.1. Full context window

We initially experimented with HuBERTs features extracted given full utterance as inputs, which potentially allows them to encode information from other phones in the utterance. Layer 9 of HuBERT / AV-HuBERT is selected for feature extraction, because it has relatively high correlation to phonetic information among layers [14, 7]. Apart from drawing decodability curves averaging samples from all phoneme categories, we also drew from only plosive stops, due to previous work [11] suggested more prominent time-to-voice on such consonants.

Fig. 1 (a) and (b) present phonetic decodability curves, which show prediction accuracy against time relative to the

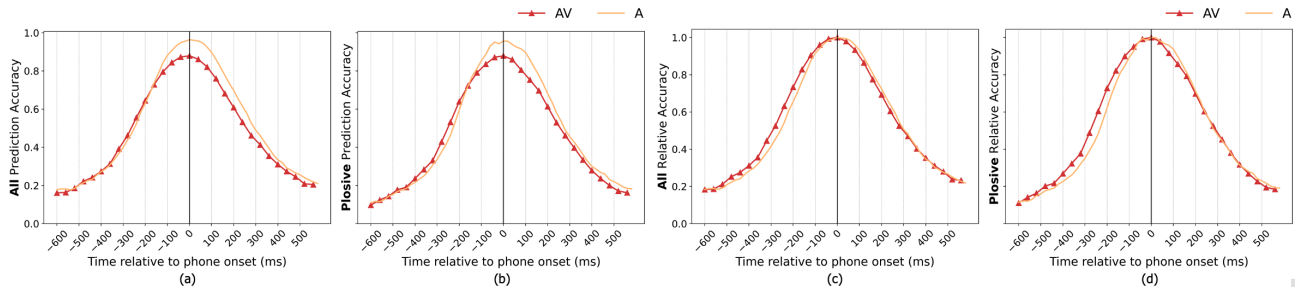


Figure 1: *Phonetic decodability curves on Grid data cross-validation experiments, representing (a) absolute prediction accuracy on all phonemes, (b) absolute prediction accuracy on plosive stop consonants, (c) relative accuracy on all phonemes, (d) relative accuracy on plosive stop consonants, with x axis labels representing time (ms) relative to phone onset (0). Plain orange line denotes HuBERT and triangular-marked red lines denotes AV-HuBERT.*

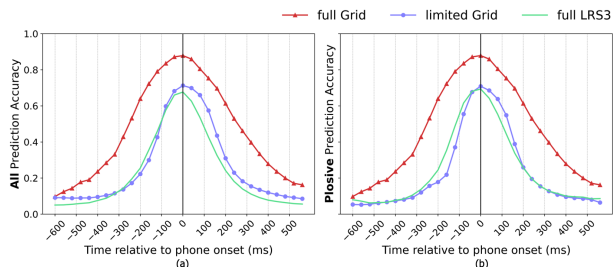


Figure 2: *Phonetic decodability curves from audio-visual features in the cross-validation experiments, full-context on Grid (full Grid, triangular-marked red lines); limited-context window of 200ms on Grid (limited Grid, circle-marked blue lines) or full-context on LRS3 (full LRS3, plain green lines), plotted on all phonemes (a), or on plosive stop consonants (b).*

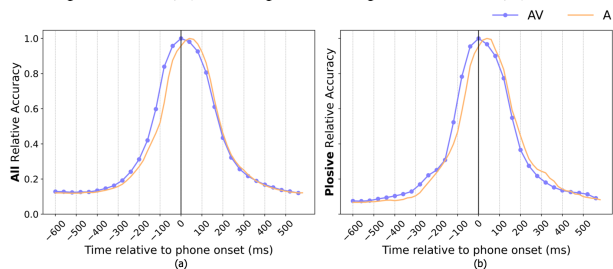


Figure 3: *Relative prediction accuracy curves on Grid data based on limited-context window, (a) all phoneme and (b) plosive stops. Circle-marked blue lines for AV-HuBERT. Plain orange lines for HuBERT*

phone onset, for all phonemes (Fig. 1a) and for plosive stop consonants (Fig. 1b), respectively, plotted with data from Grid. HuBERT curves are mildly shifted towards right than centered at the phone onset time, compared to the AV-HuBERT curves (Fig. 1a & 1b), but there is no significant sign of AV-HuBERT decodability window preceding that of HuBERT. Then we normalized each absolute accuracy curve by its maximum value, as shown in Fig. 1 (c) and (d), trying to mitigate the interference on inspecting the timing of the window from the curves' amplitude and focus only on the decodable timing. Considering the relative decodability curves (Fig. 1c & 1d), AV-HuBERT features become decodable approximately 20 ms earlier than HuBERT, which is too small to be an evidence of visual cues accelerating the decodability of phonemes.

Meanwhile, the normalized AV-HuBERT and HuBERT curves (Fig. 1c & 1d) show similar shape and width of decodability window, making further comparisons easier.

4.2. Limited context window

We further attempted to reduce the impacts from contextual phonetic information as mentioned in Section 3.1, by depicting decodability curve on LRS3 dataset or still on Grid data but with limited context (LC) provided to SSL models. A context window of 200 ms, centered on the corresponding time step of each feature frame, is enforced, with speech signals outside this window replaced by zero padding. Fig. 2 presents the resulting decodability windows drawn from absolute prediction accuracy on Grid with limited context (circle-marked blue lines) compared to that on full-context Grid data (triangular-marked red lines) and full-context LRS3 data (plain green lines). From comparing the Grid LC window curves to the Grid full context curve, we observed that the approach of limiting contexts leads to a reduced performance and narrower decodability windows, making it approximate the size and shape of the LRS3 curve. Taking the LRS3 curve as a reference, we regard applying context window at 200 ms as an effective way to reduce impacts from overwhelming phonetic contexts. Meanwhile, the LRS3 curves precede the Grid LC curves, indicating influences from data domain on the decodability curves. Potential reasons include the asynchronicity between audio and video recordings, the precision of forced-alignment results, etc. So we will consistently use Grid data with limited contexts in following experiments, considering that LRS3 consists of more noisy data.

We then presented the LC feature generated decodability curves in Figure 3, which shows a high similarity between the curves of AV-HuBERT and HuBERT. The AV-HuBERT feature becomes decodable slightly earlier than HuBERT and reaches the most decodable point 20 ms in advance. We hypothesized that such a 20 ms time gap is not because AV-HuBERT managed to capture the earlier onset of visual signals. AV-HuBERT model downsamples audio inputs by stacking 4 frames of the audio low-level features before concatenating them with one visual feature frame, so that their temporal resolutions are consistent for joint processing. This downsampling step blurs the temporal resolution in audio signals, which we suspected causes the 20 ms gap. More specifically, suppose the probability for a phone onset to fall in each of the four frames being equal, the audio feature corresponding to exactly the phone onset time will be advanced by an expectation of 20 ms. HuBERT can also blur the temporal resolution in audio signals, but its shorter frame window length (20 ms) will lead to shorter time shifts being expected.

4.3. Asynchronicity simulation

The reasoning on AV-HuBERT's relatively decodable window advancing HuBERT by 20 ms, as mentioned in Section 4.2, re-

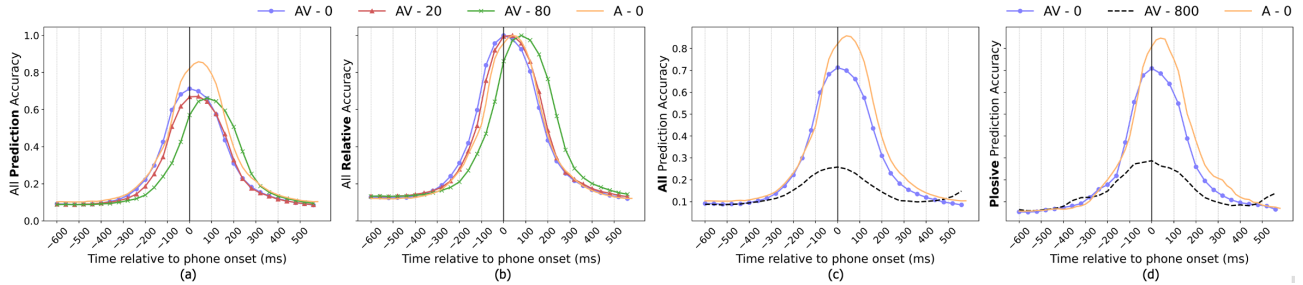


Figure 4: *Phonetic decodability curves on asynchronous audio-visual signals from Grid data: (a) absolute prediction accuracy on all phonemes (b) relative accuracy on all phonemes, (c) absolute prediction accuracy on all phonemes (d) absolute prediction accuracy on plosive stops. Subfigure (a) and (b) present AV-HuBERT curves with audio signals delayed by 0, 20 and 80 milliseconds (labelled by AV-0/20/80), respectively, and a HuBERT curve on synchronized audio-visual signals (labelled by A-0). Subfigure (c) and (d) include AV-HuBERT and HuBERT curves on synchronized audio-visual signals (labelled by AV/A-0), and an AV-HuBERT curve with audio signals delayed by 800 ms (labelled by AV-800).*

lies on a hypothesis, that the timing of AV-HuBERT’s decodability curve is dominated by its audio input. We then manipulated audio signals for AV-HuBERT, delaying it by time t_d compared to the visual inputs and phoneme labels. Fig. 4 presents the AV-HuBERT decodability curves using manipulated audio-visual inputs with $t_d \in \{20, 80\}$ ms, compared with the original curve. The peak times relative to phone onset measured on AV-HuBERT curves with audio signal delayed by $t_d = 0, 20, 80$ are 0, 40, 80 ms, respectively (Fig. 4(b), the circle-marked blue, the triangular-marked red, and the cross-marked green lines). Such delays in peak are aligned with the delay time durations in audio signals, considering a delay of 20 ms may be rounded by the 40-millisecond time resolution of AV-HuBERT. The same phenomenon is observed on the plosive stop curves, whose plots are omitted here. The curve with 20 ms audio delay almost overlaps with that of audio-only HuBERT in Fig. 4(b), (triangular-marked red line vs. orange plain line), supporting our reasoning on AV-HuBERT features’ temporal information being dominated by the audio modality while its temporal resolution being perturbed due to downsampling.

Another reason may explain why the decodability curve delays are aligned with the audio signal delays. SSL models may have learned to be robust to asynchronous audio-visual inputs by backing-off on the more informative audio modality. To rule out such possibility, we applied an extreme delay $t_d = 800$ ms on audio signals, and compared its curve with the original AV-HuBERT curves, as in Fig. 4 (c)(d). 800 ms of asynchronicity disables the 200-millisecond-long LC window to encode most of the relevant audio signals and ruins models’ potential backing-off mechanism, so the decoding primarily relies on the visual input encoding. For both all / plosive stop phonemes, the black dotted curves in Fig. 4, which represent the 800 ms asynchronicity, are centered at 0 ms. There is a low bump-up in the plosive curve at around -350 ms, but having too low accuracy to be addressed as ‘decodable’. If AV-HuBERT provides adequate encoding in the visual modality based on the actual temporal dynamics of visual signals, we expect the peak to locate prominently earlier than the acoustic phone onset time. We observed that AV-HuBERT provided a temporally skewed visual encoding that follows the temporal dynamics of audio signals. This is likely because the initialization of AV-HuBERT’s pseudo labels, based on MFCC features, forced the encoders of both audio and visual modalities to learn acoustic-based targets.

4.4. decodability window from different layers

All experiments conducted in Section 4.1 to Section 4.3 used features from the 9th layer of HuBERTs models. We further

Table 1: *The time of decodability curves’ peak relative to phone onset, depicted on representations from different layers of HuBERT/AV-HuBERT models, for all phonemes or for plosive stops. L3/6/9/12 denotes the layer 3/6/9/12 in the corresponding SSL models. 800 Async denotes curves drawn with audio inputs delayed by 800 ms*

No.	Model Type	Phoneme Category	L3	L6	L9	L12
1	HuBERT	all	20	20	20	20
2		plosive	20	20	20	0
3	AV-HuBERT	all	0	0	0	0
4		plosive	0	0	0	0
5	AV-HuBERT	all	0	0	0	0
6	(800ms Async)	plosive	0	0	0	0

validated that the lack of encoding visual speech temporal dynamics is not only limited to layer 9 features. We regard the time with the highest accuracy on a decodable curve as its peak. The timing of decodability curve’s peak drawn on HuBERT/AV-HuBERT layer 3/6/9/12 features are listed in Table 1. Peak time number 5 and 6 were generated from AV-HuBERT features using asynchronous inputs with 800 ms lags in audio, approximating the encoding of visual speech when audio is absent. Tab. 1 showed the peak time from other layers throughout HuBERT/AV-HuBERT models approximately align with those from layer 9 described in previous subsections. It confirms that other layers in AV-HuBERT can hardly capture the earlier onset of visual cues either.

5. Conclusion

In this paper, we investigated how audio-visual SSL models capture the natural asynchronicity between audio and visual cues in speech, by using a sequence of linear classifier to decode phone identity from each frame of AV-HuBERT/HuBERT model features in a context window centered at phone onset. We found that AV-HuBERT’s encoding of speech temporal dynamics is dominated by its audio input. The time course of its visual encoding is overwhelmingly aligned to its audio encoding, leading to the natural asynchronicity in speech being overlooked. We believe this is due to the cross-modal transfer learning in AV-HuBERT pre-training initialization. We conclude that great caution is needed if using AV-HuBERT to model multimodal speech perception. Furthermore, all transformer models used in temporal sensitive modelling of human speech and language processing should be used with extensive caution, as the temporal correlation in their features can be manipulated by the design of their training targets.

6. Acknowledgements

This work was supported by the UKRI Centre for Doctoral Training (CDT) in Natural Language Processing, funded by the UKRI grant EP/S022481/1 and the University of Edinburgh. We would like to thank Biao Zeng from University of South Wales and Hao Tang, Sharon Goldwater from ILCC, University of Edinburgh for useful discussion. We would like to thank Ondrej Klejch for emergency technical support on forced alignment, and Scott Ye for suggestions on plotting.

7. References

- [1] N. Kanwisher, M. Khosla, and K. Dobs, "Using artificial neural networks to ask 'why' questions of minds and brains," *Trends in Neurosciences*, vol. 46, no. 3, pp. 240–254, 2023.
- [2] M. Keshishian, G. Mischler, S. Thomas, B. Kingsbury, S. Bickel, A. D. Mehta, and N. Mesgarani, "Parallel hierarchical encoding of linguistic representations in the human auditory cortex and recurrent automatic speech recognition systems," *bioRxiv*, pp. 2025–01, 2025.
- [3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.
- [4] B. E. Stein and T. R. Stanford, "Multisensory integration: current issues from the perspective of the single neuron," *Nature reviews neuroscience*, vol. 9, no. 4, pp. 255–266, 2008.
- [5] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng *et al.*, "Multimodal deep learning," in *ICML*, vol. 11, 2011, pp. 689–696.
- [6] L. Grasse and M. S. Tata, "On the role of noise in audiovisual integration: Evidence from artificial neural networks that exhibit the McGurk effect," *arXiv preprint arXiv:2411.05715*, 2024.
- [7] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.
- [8] C. Chandrasekaran, A. Trubanova, S. Stillitano, A. Caplier, and A. A. Ghazanfar, "The natural statistics of audiovisual speech," *PLoS computational biology*, vol. 5, no. 7, p. e1000436, 2009.
- [9] J. Besle, A. Fort, C. Delpuech, and M.-H. Giard, "Bimodal speech: early suppressive visual effects in human auditory cortex," *European journal of Neuroscience*, vol. 20, no. 8, pp. 2225–2234, 2004.
- [10] V. Van Wassenhove, K. W. Grant, and D. Poeppel, "Visual speech speeds up the neural processing of auditory speech," *Proceedings of the National Academy of Sciences*, vol. 102, no. 4, pp. 1181–1186, 2005.
- [11] B. Zeng, K. Evans, M. Carne, L. Game, and E. Persson¹, "Asynchronicity between visual and auditory information in audiovisual speech: Evidence from four types of consonant-words/b/,t/,k/and/g," in *Proc. AVSEC 2024*, 2024, pp. 43–46.
- [12] O. D. Liu, H. Tang, N. Feldman, and S. Goldwater, "A predictive learning model can simulate temporal dynamics and context effects found in neural representations of continuous speech," *arXiv preprint arXiv:2405.08237*, 2024.
- [13] L. Gwilliams, J.-R. King, A. Marantz, and D. Poeppel, "Neural dynamics of phoneme sequences reveal position-invariant code for content and order," *Nature communications*, vol. 13, no. 1, p. 6606, 2022.
- [14] 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 Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [16] B. Khalighinejad, G. C. da Silva, and N. Mesgarani, "Dynamic encoding of acoustic features in neural responses to continuous speech," *Journal of Neuroscience*, vol. 37, no. 8, pp. 2176–2185, 2017.
- [17] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
- [18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Interspeech*, vol. 2017, 2017, pp. 498–502.
- [19] K. Gorman, J. Howell, and M. Wagner, "Prosodylab-aligner: A tool for forced alignment of laboratory speech," *Canadian Acoustics*, vol. 39, no. 3, pp. 192–193, 2011.
- [20] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.
- [21] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.