TEMPORAL CONSTRAINTS ON HUMAN AND ARTIFICIAL MULTI-SENSORY SPEECH RECOGNITION

CHRISTOPHER PERLETTE
Bachelor of Science, University of Lethbridge, 2019

A thesis submitted
in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

NEUROSCIENCE

Department of Neuroscience
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

TEMPORAL CONSTRAINTS ON HUMAN AND ARTIFICIAL MULTI-SENSORY SPEECH RECOGNITION

CHRISTOPHER PERLETTE

Date of Defence: January 17, 2023

Dr. M. Tata	Associate Professor	Ph.D.
Dr. J. Zhang	Associate Professor	Ph.D.
Thesis Co-Supervisor		
Dr. A. Gruber	Associate Professor	Ph.D.
Thesis Examination Committee Member		
Dr. C. Gonzalez	Professor	Ph.D.
Thesis Examination Committee Member		
Dr. D. Euston	Associate Professor	Ph.D.
Chair, Thesis Examination Committee		

Dedication

This thesis is dedicated to my family for supporting me during this process.

To the late William Slemko for his wisdom and guidance.

Abstract

Audio Visual Speech Recognition (AVSR) is the process of perceiving and understanding speech using audio and visual information. Combining visual information with auditory stimuli has been shown to improve AVSR performance when compared to purely auditory speech recognition when the task is performed in adverse conditions with large amounts of distracting noise. This work examines the relationship of auditory and visual speech information and the effect audio-visual temporary desynchronization has on AVSR performance. Using a whole report task, we show that (1) consistent with prior similar work, performance declines asymmetrically depending on the direction and quantity of a temporal lag, and (2) a common, modern architecture for computational AVSR does not show this asymmetry indicating a fundamental difference in biological and computational AVSR methods.

Acknowledgments

I would like to express my appreciation to:

My supervisors Dr. Matthew Tata and Dr. John Zhang.

My committee members Dr. Claudia Gonzalez and Dr. Aaron Gruber

My lab mates Lukas Grasse and Sylvain Boutros.

My family and friends.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Communication is at the core of human civilization. Without mechanisms to exchange complex information between people, society would not have developed to the stage it is at today. The primary method for this communication is speech, be it talking in-person, speaking remotely, or watching an audio-visual recording. The process of perceiving and understanding human speech appears simple at first. Someone talks while we listen and interpret the sounds they make to form words and sentences, and then derive meaning from that processed information. Despite how simple it may seem on the surface, there are many discrete processes that allow the successful perception and recognition of speech, with each process effectively working to solve a particular challenge associated with speech recognition. Speech recognition, as used in this thesis, refers to the process of perceiving speech information and transforming it into a phonetic representation. Examples of speech recognition challenges include the well-known "cocktail party problem" [5] in which a listener attempts to perceive speech in an environment with many background talkers, and the related segmentation problem [28] in which we split continuous speech into individual words even when those words are acoustically continuous. This thesis explores one of the brain's most interesting solutions to these problems: the use of vision to improve the perception of speech.

1.1 Visual Speech Perception

Visual speech perception, sometimes referred to as "lip reading", is the ability of an individual to use the visual information provided by mouth movements to understand speech. Although lip reading by hearing-impaired individuals is perhaps the most compelling instance of visual speech recognition, it has been shown that normal-hearing listeners also use visual speech to aid in speech recognition [15, 42, 54]. Thus, people are able to use visual information in conjunction with auditory information to better solve the problems described above. The fusion of these two sensory modalities in the service of speech perception is called Audio-Visual Speech Recognition (AVSR). How visual information affects overall speech perception/recognition is demonstrated by the "McGurk effect" [38] where when two conflicting auditory and visual syllables are observed, a third "fused" syllable is perceived.

1.2 Comparison of Human and Artificial Methods

Humans have evolved a range of mechanisms for sensation and perception of complex stimuli, and we have also developed a variety of artificial methods that attempt to emulate human perception in machines. These range from early perceptron networks that take in a feature vector and then output a binary classification [47] to modern machine learning and artificial neural networks (ANNs). Modern ANNs can perform a wide variety of tasks based on their structure and training, as demonstrated by Deep Speech [17], which achieved near human-like performance on speech-to-text transcription tasks, and various ANN image classifiers trained on the ImageNet dataset [46, 33]. Because state-of-the-art neural networks exhibit human-like performance in some perception tasks, understanding how these networks process information can have significant ramifications for neuroscience. It remains unclear whether the architectures of such networks have converged on computational solutions that are common with the human brain, or whether they represent computationally novel ways to solve the same problems. Some recent work suggests that biological and artificial networks exhibit some similarities. For example, the technique of Representational Similarity Analysis (RSA) allows comparison of neural signals (electroencephalography, magnetoencephalography, functional magnetic resonance imaging, etc.) to artificial models designed to perform the same tasks [31, 29, 30, 23, 18]. RSA is applied such that the activation patterns of biological and artificial neurons are compared to identify how different systems represent various inputs. In this way we can gain a better understanding of a wide variety of neural processes by comparing similarities and differences in biological and artificial processes. Conversely, we can also design ANNs to replicate known neural processes and use RSA to confirm if the ANN behaves like its biological counterpart [48, 30]. Given the rapid advances in artificial neural networks for solving perception problems, it has become of particular importance to investigate in what ways such networks share commonalities with human perception. Exploring the perception of humans and computers can provide insight to the relationship of these two modalities of speech as well as highlight differences in how humans and computers perform the same tasks, particularly in adverse conditions.

1.3 Our Contributions to the Field

When multi-modal stimuli that is normally consistently related in time becomes temporally asynchronous, our ability to perceive it is affected. In the case of audio-visual speech, changes in the temporal relationship of the corresponding audio and visual stimuli have been shown to have an effect on our ability to perceive word-like sounds, as well as the presentation of certain perception illusions [40, 55, 36, 7]. Despite there being work on how the perception of audio-visual illusions and speech-like sounds are affected by audio-visual temporal asynchrony, there is no work on how asynchronous stimuli affect AVSR in people or computational methods. The goals of this thesis are therefore (1) to determine how temporal lags of audio and visual information affect people's ability to understand natural speech, and (2) determine if these temporal lags affect computer programs in the same way. It is our hope that these insights will lead to better understanding of predictive coding mechanisms in general as well as aiding in the development of better multi-sensory processes and architectures for computer programs for speech perception.

1.4 Thesis Structure

Chapter 1 introduces the concepts and ideas explored by this thesis. Chapter 2 outlines prior work performed in the space of human AVSR. Chapter 3 presents a brief history of computational methods in AVSR and describes the computer program tested in the experiments presented in this thesis. Chapter 4 describes the experimental methodologies as well as presents the results of the experiments, then interprets the results and discusses the implications of our findings. Finally, chapter 5 summarizes the work performed in the completion of this thesis.

Chapter 2

The Problem of Human Speech Perception and the Role of Vision

2.1 The Cocktail Party Problem

Contending with complex auditory environments is one of the main challenges in speech perception. The cocktail party problem describes the task of attempting to perceive a single speaker's voice in an acoustic environment composed of competing talkers and sounds. An array of mechanisms exist in the human auditory brain, and artificial computational methods have been developed, that attempt to solve this problem. These include (1) spatial hearing, in which the listener uses a soundscape to separate and isolate speakers [27], (2) selective attention, when once we have identified a desired signal we are able to modulate the sensory gain on certain features of that signal [21], (3) and gaining information by binding visual cues with acoustic speech perception, which is the subject of this thesis [8, 19, 2, 9, 54].

2.2 The Segmentation Problem

One of the core problems associated with speech recognition is word segmentation. When people speak they have a steady cadence of about 4-6 syllables per second [43]. Notably, the phonemes that make up a syllable are nested within this acoustic envelope, creating structures characterized by amplitude and frequency modulations that are reflected in the waveform envelope of speech. Despite syllables being perceptually easy to distinguish from each other in continuous speech, syllables and the words they comprise, often do not have acoustic boundaries between them. Instead, temporally adjacent syllables are often co-articulated such that their temporal boundaries merge [58, 10], resulting in a continuous stream of syllables that needs to be segmented into individual words. Solving this segmentation problem is made more difficult when linguistic boundaries need to be also differentiated from an uncorrelated acoustic background. Yet, this is a difficult task that is performed every day by people when observing vocal speech [28].

2.3 The McGurk Effect

The relationship between auditory and visual speech sounds has been investigated in several prior studies. One of the first studies to examine this relationship was McGurk et al. [38]. The McGurk effect is an audio-visual speech perception illusion where when the video of one syllable (e.g. "Ba") is played while being dubbed with a different syllable (e.g. "Ga"), the observer will perceive a different syllable (e.g. "Da"). The syllable perceived depends on the two syllables presented, as well as which syllable is auditory and which is visual. In addition to this, not all audio-visual syllable combinations will result in a fused syllable when perceived, indicating there is some degree of specificity in terms of how visual speech affects speech perception. It is important as well to note that when such McGurk stimulus is presented and the participant does not report the fused sound, they are much more likely to report the auditory component of the stimuli than the visual component. This lends credence to the assertion that AVSR is primarily an auditory process as seen in Giordano et al. [13] where they conclude that the weighting of visual speech is modulated by the quality of the auditory signal, where as the quality of the audio signal degrades reliance on the visual signal increases.

Of particular interest are follow up studies that have examined how temporal misalignment of auditory and visual speech information affects how often the McGurk effect is reported in experiments. This work shows that the McGurk effect requires that video be temporally aligned or precede the audio[40]. When audio leads video, the rate at which listeners tend to perceive the illusion drops rapidly with increasing lag. However when video leads audio there is a more gradual drop in the rate at which fused syllable is reported [40, 55]. This asymmetric dependency of audiovisual speech on the relative timing of audio and video is perhaps not surprising because visual information in speech is known to precede auditory information by up to several hundred milliseconds [49]. Importantly, this is not due purely to the dramatically different speeds of light and sound, but rather reflects the fact that lip movements give rise to subsequent acoustic events in speech production, and these lip movements occur prior to the auditory onset of produced speech.

2.4 Predictive Coding for Audio-Visual Speech

Although the enhancement of speech perception by visual speech is well-known, it remains unclear how the brain actually uses visual information in the service of speech perception. One promising mechanistic theory is *predictive coding*. Predictive coding theory proposes that the brain maintains a model of a given environment and continuously updates that model based on our interactions with the sensory world [11, 20]. Internal models make predictions about our environment and these predictions are used, in conjunction with sensory evidence, to arrive at successful perception of incoming stimuli. These predictive coding mechanisms use error between predictions and sensory evidence as a means to update the model. This theory of brain function explicitly describes two flows of information, bottom-up and top-down, representing afferent sensory input and prediction signals, respectively. By computing the mismatch between these signals, the brain can then adjust how it models the sensory environment to minimize the prediction error [44].

While predictive coding as a mechanistic framework is promising, there are details that are unresolved. For example, in the case of audiovisual speech, there are two ways that visual information could be utilized. Visual information might be used to make a forward prediction about upcoming speech features that will appear later in the bottom-up signal (in the next few hundred milliseconds) [49]. This view of predictive coding imagines that the two streams of information need not be perfectly coupled to perform AVSR. Some prior work investigating the McGurk Effect supports the notion that visual information is used asynchronously to provide forward predictions for the auditory system. This is discussed further in section 2.6. A conceptually alternative mechanism is that visual information might be encoded, cached, and then fused with later auditory information after some fixed lag to account for the typical delays between auditory and visual signals in audiovisual speech [49]. In this view, visual information is not used for prediction but rather simply brought forward in time to constrain later auditory evidence. In either case, there is a pressing question about the temporal dynamics of the mechanism. Does the brain fuse auditory and visual information with a strict temporal offset, or are forward predictions derived from vision and then used without strict timing constraints to guide auditory perception?

These two possible modes of temporal interaction between visual and auditory speech features lead to two different predictions with respect to the temporal structure of audiovisual speech. If lip movements are used to forward predict upcoming speech features and/or their timing, then the mechanism should be somewhat tolerant of the predicting modality (vision) leading the predicted modality (hearing), at least within some range of temporal lags. However, any lag in the other direction, that is, audio leading video, should immediately disrupt the advantage of audiovisual speech, because the auditory brain cannot use predictions based on visual features that have not yet happened. By contrast, If lip movements are fused with strict time-locking to their corresponding speech features, we would expect the system to be highly sensitive to disruptions in the temporal relationship of auditory and visual stimuli. That is, temporal asynchrony (i.e. audio leading video or video leading audio) should thus disrupt the benefit of audiovisual speech, and the direction of asynchrony should not matter: there should be symmetry of forward-shifted and backward-shifted stimuli.

2.5 Human Audio Visual Speech Recognition

2.5.1 Perception of Visual Speech

Sensory perception is the first component of speech recognition. Visual speech is a key sensory modality used in speech recognition. By viewing the externally visible motor movements of the vocal tract as someone speaks, we can infer what is being said and even mentally synthesize speech sounds [4]. It is important to note that not all motor movements that produce speech are visible externally, and that different speech features can involve nearly indistinguishable lip movements. Thus visual speech is not unambiguous. Furthermore, the time difference between when the visible movements occur and when a given speech feature is produced is not consistent across phonemes. For example, for the sound "pa" there is an approximately 300-400ms auditory lag on visual speech, but for "ka" the lag is approximately 200-300ms [49]. For this reason, visual speech alone does not always contain enough information to perfectly predict speech, and is usually used in conjunction with auditory speech. The impact visual speech can have on overall speech perception can be seen in the previously described McGurk Effect.

2.5.2 Language Models

A language model is a structure used to predict what the next feature in a segment of speech will be, based on what came previously in that string. A language model considers factors such as sentence structure, grammar, and common phrases to predict likely results of what is being and will be said. Additionally, speech patterns and lexical clues specific to an individual speaker can be utilized if a listener is familiar with a speaker. For example, "recognize speech" and "wreck a nice beach" sound very similar phonetically and look similar in terms of the speech-producing motor movements, but conversational context would generally provide clues as to which of those two phrases is more likely to have been said. Furthermore, the word "speech" is more likely than "beach" following the word "recognize", particularly in the present context. When speaking to another person perceived to be a competent speaker of a language, people speak with a cadence that tends not to place clear acoustic space between individual words, instead spacing syllables evenly in uninterrupted speech. This is to say that if someone was to look at the waveform of speech, individual syllables would be distinct, but words would not. For this reason the complex task of grouping syllables into words, and words into phrases, is achieved by applying a language model [37].

2.5.3 Multi-Modal Speech Recognition

Auditory and visual speech in conjunction with a language model provides the human brain with sufficient information to perceive speech, even in complex auditory environments. It has also been shown that our minds dynamically modulate how much they rely on visual or auditory information based on the perceived reliability of auditory information. When there is a high signal-to-noise ratio (SNR) for auditory information (when the target auditory information isn't being significantly obscured by other signals), auditory information is primarily used to perceive speech, but when the auditory SNR decreases, there is a stronger reliance on visual information. Additionally, during low auditory SNR there is more activity in regions of the brain that are thought to have the unique role of enhancing the quality of auditory information, the premotor and superior frontal cortex [13]. These regions are associated with voluntary movement and memory, and it has been shown that when observing speech, these regions activate indicating that people mentally replicate the speech production process when performing speech recognition [53]. This demonstrates two effects: One, the human brain is able to recognize low SNR situations, and two, it is able to automatically modulate which stimuli are used in SR based on the reliability of the presented stimuli. This means that as the quality of auditory stimuli degrades, the brain will begin to rely more on visual stimuli for AVSR.

2.6 The Temporal Structure of Speech and its Effect on AVSR

When exploring the temporal dynamics of multi-sensory stimuli we must consider differences in the stimuli and its transmission from production to detection. First we must consider modal temporal resolution: audio contains several orders of magnitude more time steps than video and conversely each time step in video is more data dense than audio. This can be seen in the two modalities are modeled in computational representations, specifically that audio has tens of thousands of time steps per second, though only one value per time step and video has tens of time steps per second, but often millions of values per time step [45]. Next we must consider transmission speed of stimuli: light moves much faster than sound. We also know that while there is a range of intervals between the beginning of the motivating motor movements and production of auditory speech, the motivating movements will always precede the auditory onset [49]. When considering the travel speed of the two modalities along with the temporal relationship of visual and auditory onset, we can conclude that visual speech will always precede auditory speech in a natural sensory environment.

In addition to the previously discussed experiments in temporal misalignment on McGurk stimuli, experiments examining the intelligibility of speech-like sounds and identifying synchronous and asynchronous speech have been performed. These experiments found that visual information can lead auditory information up to a few hundred milliseconds without having a significant impact on recognition performance [36] and people are more likely to identify stimuli as synchronous when video leads audio [7] than when audio leads video. Massaro et.al. tested participant's ability to identify the speaker of a specific speech-like sound and Conrey et al. tested peoples ability to identify temporally asynchronous speech. Massaro et al. found that modulation of correct speaker identification was consistent with larger temporal windows when video lead audio rather than audio leading video. Conrey et al. found that when video lead audio, participants were much more likely to identify audio-visual speech as synchronous than when audio lead video. This indicates that in simple speech sounds, visual information likely acts as a forward prediction rather than a "guess". Despite many experiments being performed on McGurk stimulus and speech-like sounds, we are not aware of any literature regarding temporal lags and natural speech in a longer format such as phrases or sentences.

Prior work on audio-visual perception of simple speech sounds in humans has shown (1) that visual speech can have a significant impact on overall perception, and (2) there appears to be an asymmetry in how the temporal alignment of the modal streams affect perception. Despite relatively extensive work on McGurk stimuli and simple speech sounds, there appears to be no work on natural speech. For this reason, this work will explore how temporal misalignment of natural speech affects our AVSR capabilities.

Chapter 3

The Problem of Artificial Perception of Audio-Visual Speech

3.1 Artificial Methods for Speech Recognition

Artificial methods for Automatic Speech Recognition (ASR) have advanced rapidly in recent years. As discussed in Chapter 1, speech recognition is an exceptionally complex task that involves converting auditory and sometimes visual information into words. This process requires the ability to parse and interpret large amounts of sparse data to compress and convert it into another form. For this reason ANNs are utilized for the task. An ANN is a computer program that is designed to emulate the way it is believed biological brains process information [41].

A question regarding artificial AVSR neural networks follows: if ANNs are designed to emulate human cognition, then we would expect them to perform AVSR tasks in a similar way to humans. This can be tested by comparing the AVSR performance patterns of the two groups.

3.1.1 Artificial Neural Networks

The general structure of an ANN consists of a series of connected layers that are broadly classified as the input layer, the first layer where information is fed into the network, the output layer, the last layer where processed data is reported by the network, and the hidden layers, any layer between the input and output layers. Each layer is composed of units known as neurons, and the depth of a network is measured in the number of layers in the network. A key feature of ANNs is they can learn to perform various tasks through a process known as training.

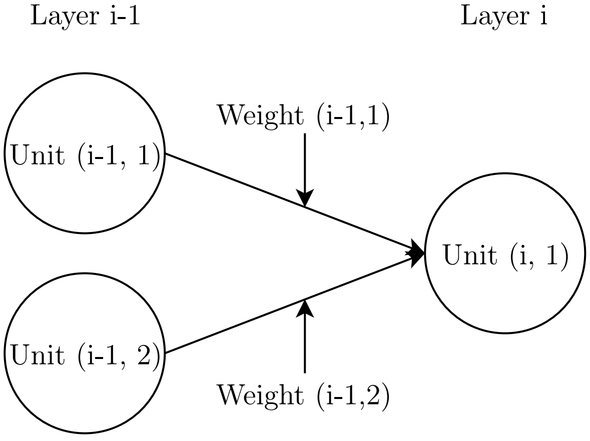

Figure 3.1: A Basic ANN Structure

How the layers and neurons in an ANN are composed defines the network and heavily influences how effective it is at successfully performing a task. The input and output layers will have a number of neurons equal to the number of inputs and outputs respectively, and hidden layers contain a number of neurons defined by the network designer. Each neuron is connected to all the neurons in the immediately adjacent layers, but not their own (with some exceptions, see Section 3.1.3), and these connections are each assigned a unique modifier known as a weight. This will be expanded on later. Neurons apply a mathematical operation to their input, called an activation function, that determines the output of that unit. The input for all neurons and their respective activation functions (other than those in the input layer as those are provided to the network) is determined by calculation in equation 3.1, with a visual diagram in figure 3.1. In this equation i is the layer index, j is the neuron index, N is the number neurons in layer i, O is the output of a neuron, and W is the weight of the connection between neurons $I_{i,j}$ and $O_{i-1,j}$, and B is the bias of the neuron. Bias is simply a value that is added to the input of a neuron regardless of the previous layers output. This allows models to shift an activation function to be more positive or negative overall.

$$I_{i,j} = B_{i,j} + \sum_{j=1}^{N_{i-1}} (O_{i-1,j} W_{i-1,j}) \tag{3.1}$$

The process of calculating the final output of a network is known as forward propagation where the equation above is applied to each neuron in a layer starting at the first hidden layer and moving through the network. After the input of a neuron is determined, the activation function chosen by the designer is applied and that final value is used as the output of that neuron for the following layers input calculations. An example of an activation function is the sigmoid function that can be seen in equation 3.2. In this equation $O_{i,j}$ is the output of neuron, $I_{i,j}$ is the input to the neuron, and e is Euler's number. This is just one of many commonly used activation functions, with the key takeaway being that the activation function takes the value calculated in equation 3.1 and uses it as an input to calculate the output of the neuron. At the end of this process, when the outputs for the neurons of the output layer are calculated, this output is returned by the network.

$$O_{i,j} = \frac{1}{1 + e^{-I_{i,j}}} \tag{3.2}$$

In training an ANN "learns" to perform a task in a similar way to how biological brains learn according to the predictive coding theory [14]. Inputs are processed by the network using forward propagation to create a prediction, and that prediction is compared to a known value. The difference between the known value and the ANN output (often called "loss" or "cost") of the network is stored and later used in adjusting the value of the weights in the ANN such that the cost is minimized. This process is known as gradient descent and can be seen in figure 3.2.

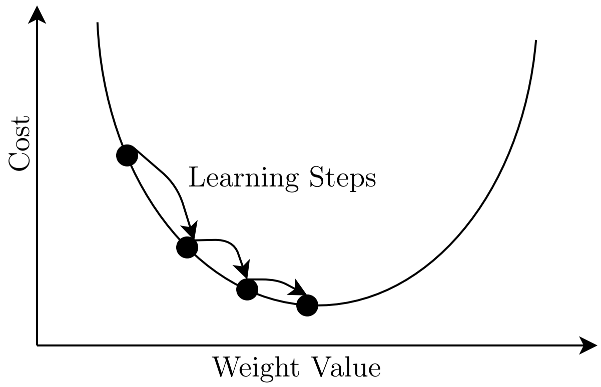

Figure 3.2: Gradient Descent Process

3.1.2 The Application of Artificial Neural Networks in Automatic Speech Recognition

Given the importance of speech perception in communication, it has been a pressing problem in human-computer and human-robot interaction to develop effective speech recognition systems, resulting in modern machine learning algorithms and a variety of speech-specific ANNs [1, 26, 52, 51].

3.1.3 Recurrent Neural Networks and Long Short-Term Memory

ANNs began to have success in ASR tasks with the development and application of Long Short-Term Memory (LSTM) units. To understand an LSTM, you must first understand the principles of Recurrent Neural Networks (RNNs). RNNs are a form of ANN where the output of a neuron also functions as the input to other neurons in the same layer on the next input, as shown in figure 3.3. RNNs excel in the analysis of time series data which makes them a good candidate for ASR, however, RNNs struggle with long term preservation of past information (generally when in excess of 1000 time steps) without greatly exaggerating or diminishing that information [39, 35]. This is an effect often seen in systems that feed into themselves. LSTMs were developed to address this short coming. LSTMs add a gate to the recurrent portion of conventional RNNs. The LSTM holds a value that is fed back into itself and other units in the same layer, and the gate can be thought of as its own unit in the ANN.

Where the LSTM unit learns how it connects to itself and other units, the gate learns when to modify the data stored in the LSTM and how much to modify it by, as shown in 3.4. This allows LSTMs to be less affected by irrelevant or erratic inputs and preserve pertinent information [22]. This made LSTMs ideal for application in ASR due to the long-term nature of speech signals, as well as the cluttered nature of audio data.

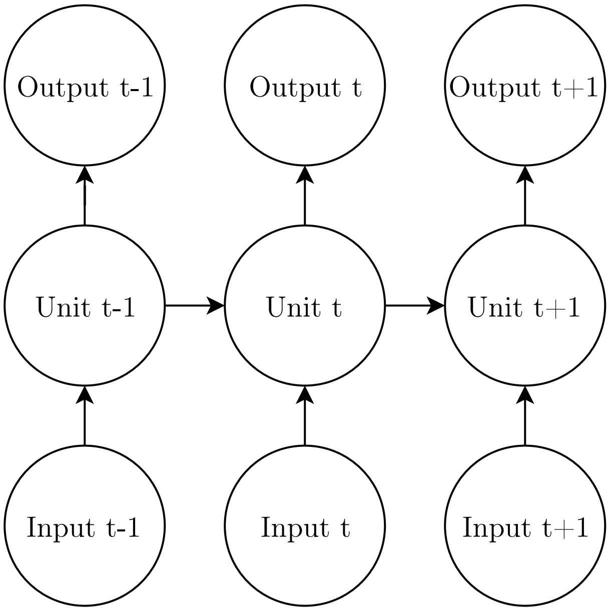

Figure 3.3: Recurrent Neural Network Structure

3.1.4 Sequence-to-Sequence Classification

Auditory speech is an extremely sparse form of data. On average people speak in the range of 10-15 phonemes per second, based on a combination of syllables per second and phonemes per syllable [43], and audio data that is being classified is often recorded at or down sampled to 16000Hz. This means even in faster end of average, there are more than 1000 auditory samples per phoneme. In the visual domain the sample rate is much lower, commonly 25-30Hz for machine learning inputs, however, each sample is often composed of hundreds or, more often in AVSR, thousands of pixels. On top of this, speech is irregular. People often stop speaking mid-sentence

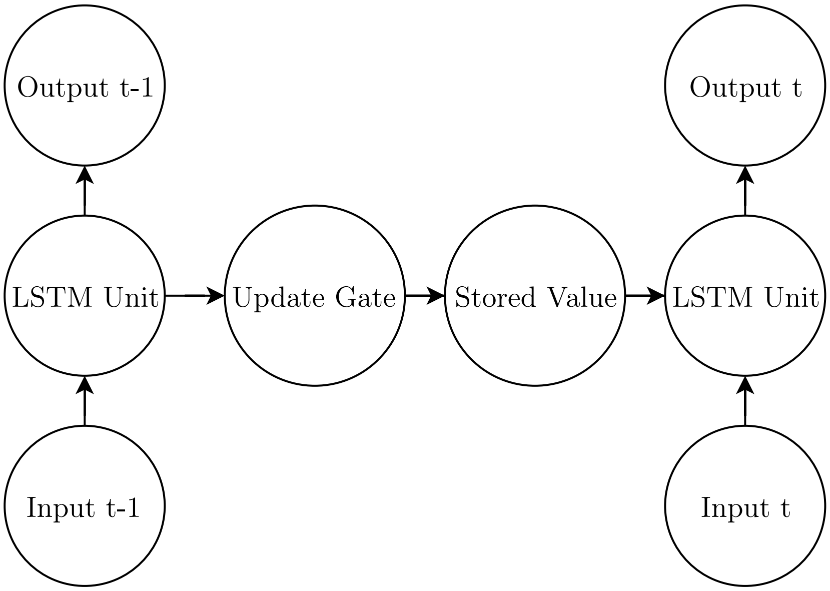

Figure 3.4: Long Short Term Memory Structure

to either breath or to think about what they will say next. In ASR the final output of an ANN is generally either plain text or phoneme representations of speech. These are very dense forms of data, and in terms of the number of data points they are much shorter. The result of this is one of the most difficult problems to solve in machine learning, sequence-to-sequence classification. The complexity of this problem requires the application of specialized ANN architectures one of the most popular being connectionist temporal classification, which is discussed later in this chapter.

3.1.5 Encoder-Decoder Model of Classification

The encoder-decoder model for ANNs is a popular method for converting complex, sparse data sources into more compressed and concise forms and is often applied to sequence-to-sequence classification problems. The basic pipeline for the model is as follows. Raw information is collected and presented to the network. That raw information is then processed and condensed into useful features that identify and differentiate the raw data. What these features are will depend on the type of raw data and what the network is designed to do. In the case of auditory speech data it could be descriptors such as pitch or timbre. These encoded features are often referred to as embedded features, or simply embeddings. These embeddings are then processed by the decoder to classify some aspect of the raw data based on the embedded features. The ability for the encoder-decoder model to take sparse, complex data and effectively classify it makes it an extremely effective model for ASR [25].

3.1.6 Multi-Headed Attention Transformer Model for Classification

The transformer model in ANNs is an evolution of the encoder-decoder model. With the broad concept of encoding raw data and decoding embeddings still employed, the transformer model makes use of structures known as attention heads. What a head attends to is dependant on the training it receives, but broadly there are 3 forms of attention. Self-attention in the encoder, self-attention in the decoder, and encoder-decoder-attention in the decoder. These act in the following ways respectfully, the input sequence attends to itself, the target sequence attends to itself, and the target sequence attends to the input sequence. Individually these "heads" break down an input and determine if a section of the input should be attended to as well as how to attend to it. By using multiple attention heads (hence the term multi-headed attention) in the same network it is possible to attend to multiple distinct attributes of the input data. Due to the nature of multi-headed attention having multiple independent functions in the same section of the input analysis, the use of a transformer model enables large scale parallel processing allowing for the potential of greatly decreased real-time processing as well as more accurate results. The transformer model is also able to process an entire input sequence simultaneously, giving them an edge in computation time over conventional RNN and LSTM based networks that must process each data point individually [6]. One disadvantage of using multi-headed attention is that unlike RNN and LSTM architectures that "perceive" each chunk of data in order sequential order, multi-headed attention "perceives" the entire input simultaneously. To address this shortcoming, positional encoding is employed. This is a relatively simple process where information for the position of a sample including its relative and absolute positions in the input are stored in a data structure and included in the input of that sample into the encoder. This results in a relatively simple and lightweight method for achieving temporal resolution and still enabling parallel computing.

3.1.7 Connectionist Temporal Classification

As explained previously, ASR requires the mapping of a long sequence of raw data on to a relatively short output. To achieve this, it is common to use Connectionist Temporal Classification (CTC) and its associated scoring function, CTC loss, for the creation and evaluation of a networks output. The output of an analysis network, regardless of the architecture (RNN, LSTM, Transformer, etc.), is a single label [14]. By taking a series of short segments from raw data and using them as the input for an analysis network, you get a series of labels. Importantly blank/null is a valid label, and this series of labels is decoded into a final output. The decoding process utilized by the CTC architecture is what makes it so effective for sequence to sequence classification. The decoding process compresses the output by considering the output in sequential order and removing repeated labels in the sequence. This makes the model particularly good at ASR when the target output is a series of phonemes as phonemes rarely repeat in a word. However, in text outputs, repeated letters are common. In these instances it can be seen why null being a valid output is important, as it can be used to indicate breaks between repeated letters. This prevents the decoding process from removing these repeated letters and can correctly identify words that contain 2 or more of the same character in a row [16].

3.2 Comparing Human and Artificial Methods

Comparing human and artificial methods for performing the same task can have a wide range of benefits. Previously we have discussed the history of ANNs in SR tasks, and briefly touched on the general history of ANNs. ANNs were initially developed to emulate the processes believed to take place in brains and neural tissue [47]. This followed the principles of biomimetics where artificial designs are based on biological systems that already perform a given task [56]. Starting with more basic perceptron networks and evolving into the structures we know today, ANNs have grown in scale and complexity, however, during this evolution there has been a decoupling of the roots of ANNs and their current directions. Advances in computing science have rapidly out paced advances in neuroscience and while ANNs were initially often specifically designed to emulate what was believed to be the human process of performing a task, this is often not the case anymore. Now, often when a network is designed to perform a human task, designers think in terms of what types of networks and units handle a specific type of data well or have been known to perform a similar task in the past rather than directly considering the biological systems that perform that task. This exposes the interesting question of potential convergent evolution. If an artificial system is initially designed to emulate a biological one, but diverges from this design philosophy and then after many iterations is trained to perform a task that the biological system can also perform, could the artificial system exhibit the same behaviours as the biological system?

A variety of studies have shown that the behavioural and empirical experimental results between artificial and human methods show similar trends, and when RSA is applied the underlying dynamics of how data is processed are often similar [30, 23]. This shows that utilizing biomimetic designs for artificial models can have benefits including the reduction of network size as well as requiring less training data and training cycles [48, 18, 24]. If we compare the behavioural traits of human and artificial

AVSR and find distinct differences between them, it is likely that the architecture of Deep AVSR, the ANN tested in this research, could be modified to better recognize speech in adverse conditions.

3.3 Deep AVSR

As one of the purposes of this research was to determine differences between the behaviour of human and artificial methods in AVSR, it was vital to select a model that was developed relatively recently and implemented modern ANN structures. For these reasons the network called Deep AVSR [50], based on the paper Deep Audio-Visual Speech Recognition from Afouras et al. [1], was chosen. This is a state of the art AVSR network that uses a Transformer Model CTC (TM-CTC) architecture for AVSR. The network separately processes the audio and visual streams initially. Both data streams feed into their own encoders that have 6 layers with each layer containing 8 attention heads to create feature vectors. These 2 feature vectors are then concatenated and go through a convolution layer to reduce their size by half for the purpose of maintaining consistency in the transformer functions. The compressed feature vector then feeds into a decoder with the same architecture as the encoders. An initial decoding is performed using a beam search that minimizes the total distance between all likely output options, then the CTC decoder is applied as described in the above section. Finally a language model is applied to the result of the CTC decoder. This converts character outputs to real words based on a combination of 1) how close the string of characters is to a given word and 2) how likely a word is to exist at that point in the output based on prior words in the output. For a broad overview of the structure and data flow of Deep AVSR, as shown in 3.5. A more in-depth description of the architecture can be found in the paper and in the repository of the implementation [50]. This implementation of the network is trained on the LRS2-BBC dataset, containing approximately 150,000 AV utterances with a

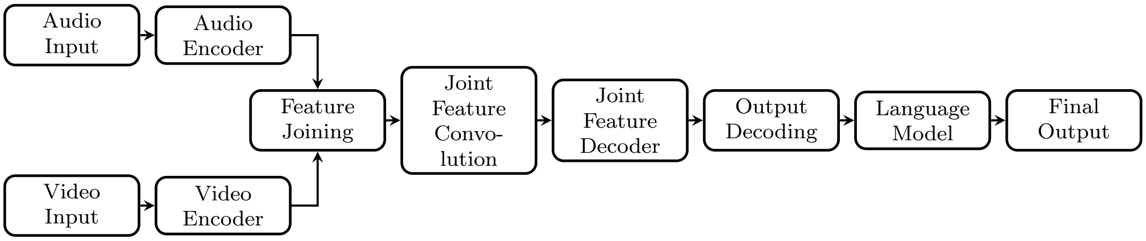

Figure 3.5: **Deep AVSR Structure**

total of approximately 2.4 million words.

Chapter 4

An Investigation of Temporal Dynamics in Human and Artificial Audio-Visual Speech Recognition

4.1 Preamble

In chapter 1 we outlined the primary goals of this research. Broadly, these were to compare and contrast human and artificial speech recognition, particularly with respect to the integration and use of *visual* speech information.

More specifically, we aimed to test two ideas that follow from a Predictive Coding view of the temporal dynamics of audio-visual integration for speech recognition. These ideas are derived from a 1) continuous (human) and 2) simultaneous (transformer model) model of the perception of time-series data where correlated events do not occur simultaneously. These ideas are 1) *Queuing*, wherein a "queue" of visual inputs are continuously compared to new auditory events in a broad range of time steps, and 2) *mapping* where visual feature "embeddings" are "mapped" onto and compared to embeddings generated from the auditory input at an interval in the future. This interval is informed by the visual embedding. In the case of *queuing*, we would expect to observe an asymmetric performance reduction where when video leads audio the drop is gradual but when video lags audio there will be a rapid drop in performance. Conversely, *mapping* should exhibit a symmetric, rapid drop in AVSR performance as a lag grows.

Importantly, these predictions are irrespective of whether the listener is a human brain or an artificial neural network. In this way we can make a novel assertion regarding any temporal dependency of AVSR in natural speech, in both people and ANNs. Using a *whole-report* task we tested the AVSR capabilities of human participants and an artificial program, then analyzed and compared the results. This chapter describes the experiment and analysis methods performed, and discusses the results of the experiments.

We predicted that the human listeners would exhibit an asymmetry in performance with respect to asynchrony between auditory and visual speech. Specifically, we predicted that, consistent with the previously described "queuing" behaviour, performance should be relatively robust to video leading audio by up to a few hundred milliseconds, whereas performance should decline rapidly and substantially when audio leads video. These predictions follow from previous related studies [40, 55] that examined the influence of asynchrony in audio-visual speech. In these studies McGurk-style stimuli (i.e. single audio syllables paired with congruent or incongruent visual lip movements) were lagged in time such that audio would either lead or lag video, thereby introducing an artificial audio-visual temporal asynchrony. The results showed that the frequency of the McGurk effect occurring (i.e. misperception rate) rapidly dropped when audio led video. This means that visual input had little effect on perception of auditory syllables if those syllables lead the video even by a small lag. Conversely, the misperception rate dropped slowly as audio lagged video by increasing amounts, demonstrating a distinct asymmetry in misperception rate depending on the direction of the temporal lag. Thus visual input effectively modulated auditory perception even when it led by many tens of milliseconds. We expected AVSR performance in a whole-report speech recognition task to modulate in the same way, with a rapid drop in performance when audio led video and a gradual drop when video led audio.

In potential contrast with human listeners, Deep AVSR does not include layers or neuron types that would be able to dynamically compensate for shifting audio-visual synchronicity if such shifting synchronicity is not present in the training dataset. Notably, the training dataset for this network primarily contained television clips, which are often temporally well aligned. For these reasons we speculated that AVSR performance emulate the previously described "mapping" behaviour, and performance would drop uniformly and rapidly in each direction as the audio-visual asynchrony increased. If this were the case, it would mean that Deep AVSR and human listeners use visual information in fundamentally different ways.

4.2 General Design

4.2.1 Construction of the Stimulus Test Set

Six fluent English speakers, 3 male and 3 female, were video recorded saying 28 unique phrases selected from the TIMIT dataset [12]. These 28 videos were then copied and programmatically edited so that the audio was shifted temporally by a series of lags (in ms, +/- 60, +/-240, +/-360, +/-480, +/-600, +/-720, and +/-840 for video leading (+) and lagging (-) audio, respectively). As audio was shifted to achieve the lags, depending on the direction of the shift there was empty audio space created at the beginning or end of a video and additional audio extending past the beginning or end of the video. The empty space was filled with silence and extra audio was truncated to keep the length of the audio and video the same. For the lagged videos an alphanumeric code is used to represent the temporal lag in the format of a letter followed by a three-digit number. The letter will be a ‘B’ or ‘I’ indicating audio "Behind" video and audio "In-front" of video respectively, and the number is the millisecond value of the lag. For example, I240 indicates that the audio track lead the video by 240ms. Additionally, there were "base" and "jumble" trials. In base trials, video was the original unmodified video, with audio and video perfectly synchronized; that is, 0 ms lag. Jumble takes the video and audio tracks from two different base clips of the same length and speaker, and combines them to create a video with incongruent auditory and visual components. The intention of the jumble trials was to provide a worst-case baseline in which the visual information could not be used at all to aid in recognition. This choice of worst-case baseline has the advantages that 1) it uses audio-visual speech as the stimuli, and 2) unlike the temporal lags, there is no consistent correlation between the audio and video that could be used to aid in the recognition of speech. Additionally, for the human trials, babble noise was added to increase the difficulty of the task. Babble noise was directly added over the audio track of the video, including the sections that were filled with silence. This extra background noise encouraged the use of visual information to aid in the recognition of the speech [13], and adjusted performance to avoid floor or ceiling effects. Babble noise mimics the temporal and spectral dynamics of human speech, though crucially does not contain linguistic information [32]. This allows it to act as a distractor or masker to make the experimental task more difficult without introducing information that could be construed as identifiable speech [32]. Noise was not added to the computational experiment as it caused a floor effect for Deep AVSR.

4.2.2 Human Participants

The participants in the following experiments were students recruited from courses in Psychology or Neuroscience at the University of Lethbridge, Canada. The participants joined the experiments remotely via a custom-built website due to constraints of the COVID-19 pandemic. Students who participated in the experiments were compensated with course credit. Students were provided with a questionnaire prior to the start of the experiment to record their age, biological sex, handedness, if they were aware of any hearing problems, if they were using headphones or speakers to complete the study, and if their first language was English. We encouraged participants to use headphones rather than free-field speakers because headphones generally have a higher audio quality. If subjects had known hearing issues or English was not their first language they were allowed to participate in the experiment, however this was recorded to test if these participants significantly affected results. In all experiments, we observed no difference when either the ESL group, hearing impaired group, or both groups were excluded from the analysis: both the overall pattern seen in the results as well as which lags were significantly different from each other did not change. For this reason the results for both groups were included in our analysis.

The procedure adhered to the proscriptions of the Document of Helsinki and was approved by the University of Lethbridge Human Subjects Review Board. All participants gave informed consent prior to participating.

4.2.3 Computer Participant

Deep AVSR was chosen as the computational model to be tested in our experiments. This was for three main reasons: First, it had developer created pre-trained weights available for use in research that performed very well according to the results published by the developers. Second, and most importantly, it implemented a TM-CTC architecture which was shown by Afouras et al. [1] to be a highly effective architecture for performing sequence-to-sequence classification, particularly in AVSR. Additionally, as Deep AVSR uses a language model it can provide better results for the word-to-phoneme conversion tool used. Details about this tool can be found in the following section. Finally, Deep AVSR is a well-recognized solution to the AVSR problem (e.g. at the time of writing, the original 2018 Deep AVSR paper has been cited in 95 research papers). Its architecture is based on state-of-the-art approaches and it represents a good standard benchmark for such networks.

4.3 Analysis

The metric used for the analysis of the results was Phoneme Error Rate (PER). PER is a variation of Word Error Rate (WER). WER has been shown to be an effective measure of one's ability to understand speech [57] and is a standard metric for measuring speech recognition performance. We chose to use PER over WER as it allows for more granularity in our results. PER is calculated as the number of insertions (I), transformations (T), and removals (R) required to transform a response into the correct target, divided by the number of phonemes in the target (N):

$$\text{PER} = \frac{I + T + R}{N} \times 100\% \tag{4.1}$$

By decomposing words into phonemes and then comparing those versions of the responses and targets we can better determine if part of a word was correctly heard and gain a better understanding of the capabilities of the participants. For example, if a response was "word" and the target was "ward", WER would calculate a 100% error rate where PER calculates a 25% error rate.

To convert the word responses of the human participants and Deep AVSR into phonemes, a program called Phonemizer was utilized [3]. This program uses the Festival text-to-speech engine to take words and convert them to phonemes using a custom defined phoneme transcription that contains all the phonemes used in the English language.

For overall PER in the experiments, each response was analyzed individually against its target, then all PER values for a given condition in a group were averaged and standard error was calculated to get an average value and error bars.

Two variations of the PER analysis were also performed. In the first we broke PER into the individual correction operations to see if there were any differences in the required corrections between groups. In the second we compared the first and second half of responses to the respective halves in the target. This was done by taking the string of phonemes and splitting it in half, rounding up in the event of an odd number of phonemes. For example, if there was a 7-phoneme long response, the split would result in two 4-phoneme long strings with the two sharing one phoneme. As with the base PER analysis, standard error was used for error bars.

The design of the experiment allowed for direct comparison between the various conditions in the same experiment, thus t-tests were used to identify when a given lag was significantly different from its respective +/- lag as well as the jumble for that experiment. We were particularly interested in comparing "symmetric" positive and negative lags since any significant difference would indicate an asymmetrical robustness to misalignment. We thus predicted that positive lags (video leading audio) would allow significantly lower PER when compared to corresponding negative (audio leading video) lags. A Benjamini-Hochberg method with a false discover rate (FDR) of 5% was used to adjust for inflation of Type I error rate due to multiple comparisons.

Finally, the length of targets and responses (in phonemes) were averaged for each lag in each experiment. Standard error was used for error bars. This analysis was performed to provide insight into our operation results.

4.4 Experiment 1 - Human Participants

The basic paradigm for the human experiments was as follows: The participant was shown a video clip with its associated auditory track leading or lagging in time. The task was to listen to the entire sentence and then type the sentence into a text box that appeared after the video clip completed.

A practice segment consisted of 2 iterations of the basic loop, using 2 pre-determined sentences that were not included in the experiment set. Responses from the practice segment were not recorded or analyzed. After practice, participants entered the main experiment which consisted of 48 iterations of the basic loop using a data set that was generated and unique to each participant. The experiment was a full report task so participants were asked to type into a text box what they believe was said in each stimulus video. The raw responses were recorded to a database along with the speaker, clip number, and lag.

Each participant had a unique set of videos generated for them programmatically. This set contained 6 temporal lags, a base video, and a jumble video for a total of 8 videos from each of the 6 recorded speakers; thus a total of 48 videos plus the two practice videos. The set was generated such that each participant was presented with each lag being tested from every speaker, and the order of videos was shuffled. Participants were thus presented with 48 unique sentences without repetition of any sentence during the session.

Three variations of Experiment 1 were carried out, which differed in the asynchronies tested and the level of babble noise.

4.4.1 Experiment 1A

Experiment 1A had 25 participants (19 female and 6 male); 24 participants were native English speakers and 2 had known hearing impairments. The videos presented in this experiment had the base and jumble videos, along with I360, I240, I060, B060, B240, and B360 from each speaker. For this experiment the added noise track was balanced to -10dB relative to the speaker audio track.

Results

The results of Experiment 1A revealed the predicted asymmetry in AVSR performance with respect to audio leading or lagging video. This asymmetry is most evident in the PER figure 4.1 and in the t-test results found in table 4.1. We found that PER for the base, I060, B060, B240, and B360 lags was significantly different from the jumble condition whereas the I240 and I360 lags were not. Furthermore, the I/B240 and I/B360 pairs of lags were significantly different from each other: in both cases PER

Comparison Type	Offset	*p*-values		
		Experiment 1a	Experiment 1b	Experiment 1c
Same Temporal Value	I/B060	0.466	N/A	N/A
	I/B240	*3.16e-3	N/A	N/A
	I/B360	*8.69e-6	*6.01e-3	*6.17e-4
Against Jumble	I360	0.976	0.502	0.780
	I240	0.180	N/A	N/A
	I060	*2.33e-13	N/A	N/A
	Base	*4.87e-8	*2.24e-9	*1.07e-15
	B060	*1.77e-10	N/A	N/A
	B240	*2.01e-5	*7.73e-4	*1.05e-5
	B360	*8.03e-6	*1.17e-3	*5.88e-4
	B480	N/A	0.240	*8.63e-4
	B600	N/A	*2.90e-2	*3.15e-3
	B720	N/A	7.25e-2	0.239

Table 4.1: **T-Test Results for Human Experiments.** This table shows where offsets stop being significantly different from the jumble condition as well as if symmetric offsets are significantly different from each other across experiments 1a, 1b, and 1c.
* *p*-value < 0.05

was significantly worse when audio lead video than when video led audio by the same lag. This indicates that there was a lag corresponding to somewhere between I060 and I240 at which visual information ceased to be effective in aiding AVSR. However, because of the pronounced asymmetry in the data, we could not find a similar lag in the video-leading-audio direction beyond which AVSR performance was as poor as the jumble condition. At all lags we tested in Experiment 1A, when video led audio there was a significant benefit of AVSR relative to the jumble condition.

In terms of correction operations (insertions, transformations, removals), figure 4.2 shows that in experiment 1A insertions were the primary required correction, and were most modulated with overall error rate. Breaking errors down by first and second half of the sentence, we found that the first half was consistently lower than the second half as shown by figure 4.3. Finally, when examining response lengths, it appears that as PER increases, response length decreases as seen in figure 4.4. Based on the

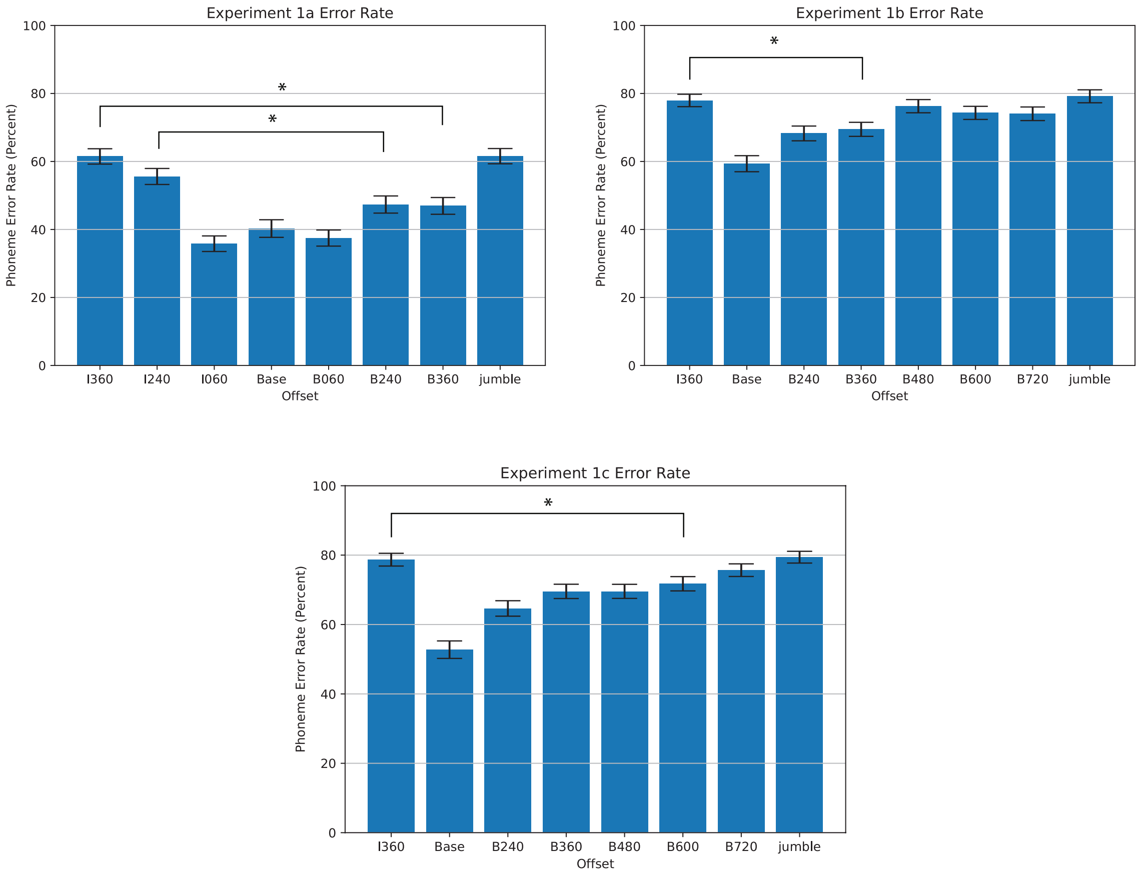

Figure 4.1: **Human Experiment PER Analysis.** For experiment 1a the asterisks show that I/B240 and I/B360 are significantly different from each other. In experiments 1b and 1c the asterisk indicates the last offset that is significantly different from I360.

correction operation results it appears that the participants will generally not respond unless they are highly confident of a response. This is backed up by the analysis of the lengths of targets and responses. Additionally, the results seen in the response section error rate are counter to what would be expected if an adaptation effect was present, though the presence of this pattern in the base and jumble conditions indicates that instead of a lack of adaptation, another effect is causing this pattern.

4.4.2 Experiment 1B

Experiment 1A revealed the predicted asymmetry with respect to lag, but failed to show the extent of the asymmetry. This was due to the fact that listeners were able to

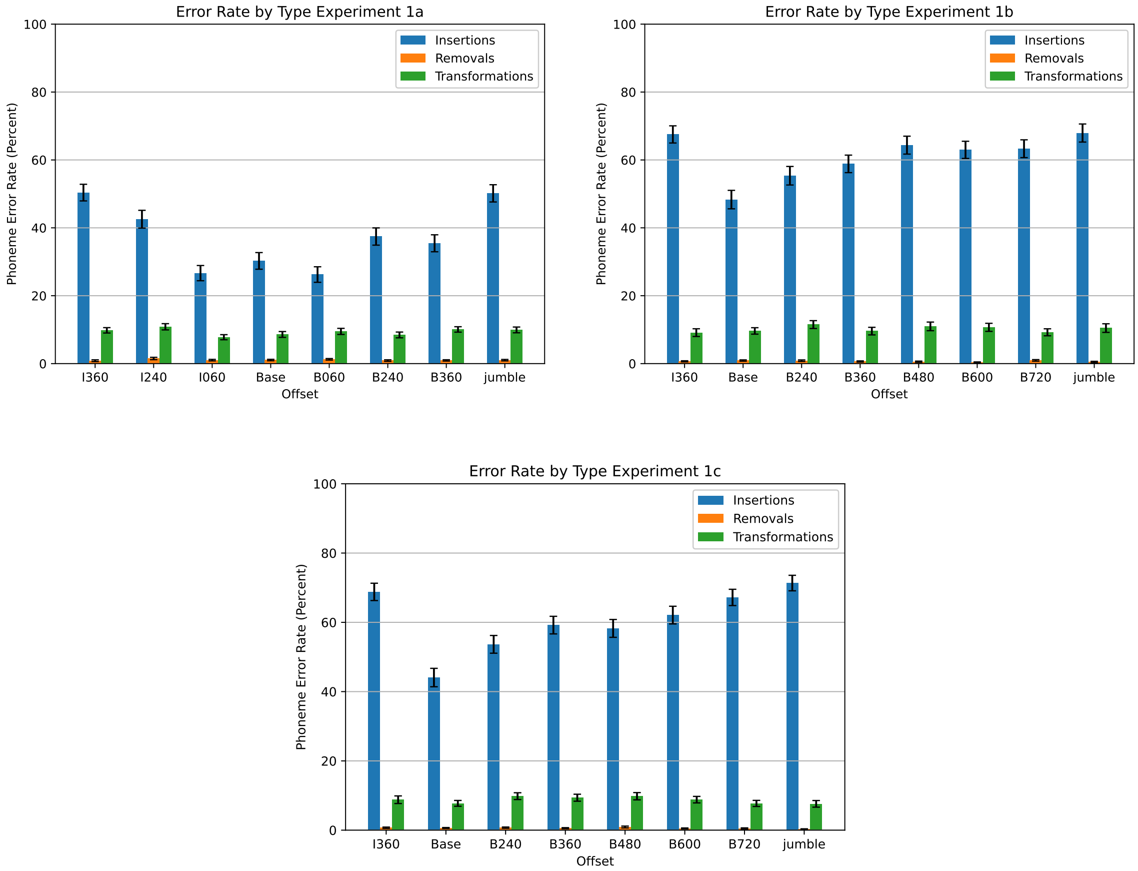

Figure 4.2: **Experiment 1 PER Analysis by Correction Type.** These graphs indicate that insertions are the primary correction type required for human responses regardless of lag direction or quantity.

derive an AVSR advantage at every positive lag tested. Experiment 1B was designed to find a lag beyond which even video-leading-audio conferred no AVSR benefit. This experiment had 19 participants (13 female, 6 male); 14 were native English speakers and none had known hearing impairments. For this experiment we tested the base and jumble conditions along with I360 as an audio-leading baseline. We tested video-leading conditions of B240, B360, B480, B600, B720. In this experiment the noise was balanced at -3dB relative to the speaker audio track in an attempt to achieve better stratification in the results.

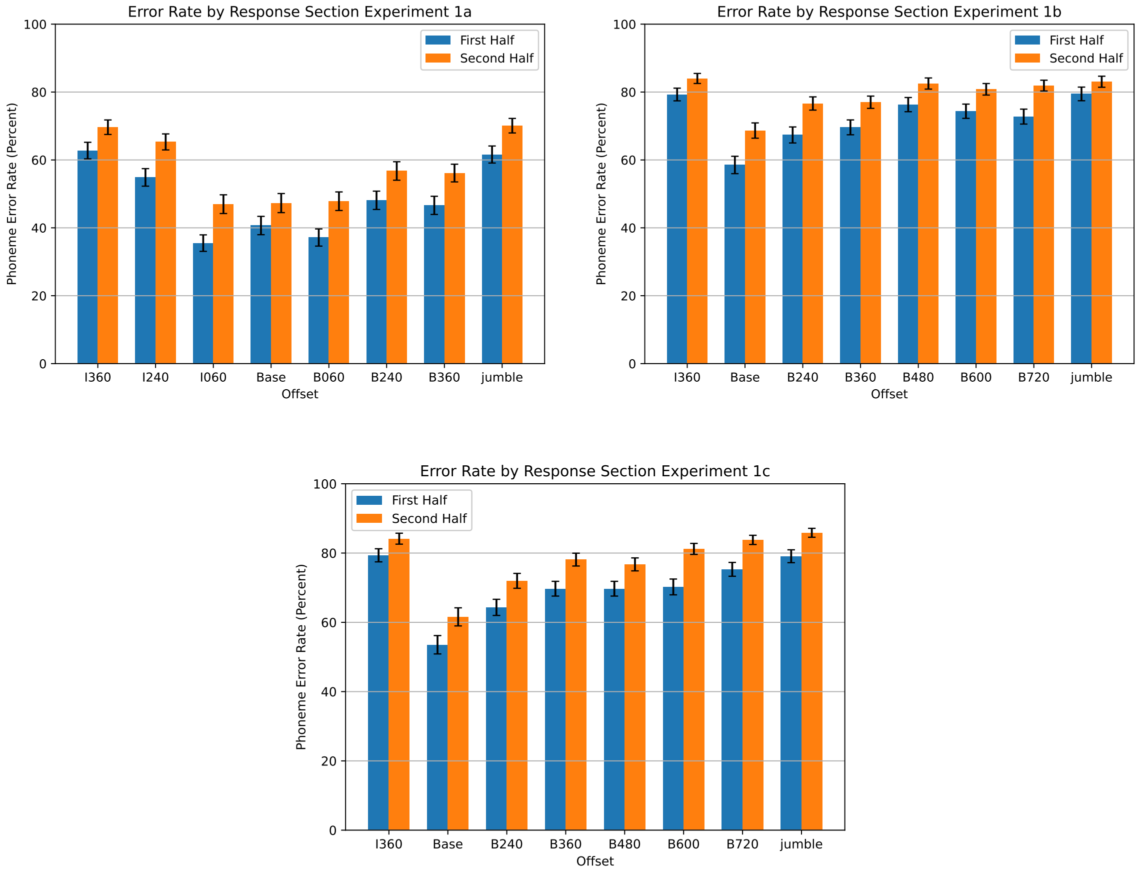

Figure 4.3: **Experiment 1 PER Analysis by Response Section.** These graphs show that contrary to what would be expected with a predictive coding model, PER in the second half of human responses is higher than that of the first half.

Results

Graphs for overall PER in experiment 1B can be found in figure 4.1 and t-test results can be found in table 4.1. Experiment 1B also replicated the asymmetry found in Experiment 1A and revealed the positive extent. We found that the base, B240, B360, and B600 lags all allowed significant AVSR benefit relative to the jumble condition, whereas the I360, B480, and B720 lags did not. Additionally, the I/B360 lags were significantly different from each other. In terms of correction operations, figure 4.2 shows that in experiment 1B insertions were the primary required correction, and were most modulated with overall error rate. This was consistent with experiment 1A. For first and second half error rates we found that the first half of the sentence

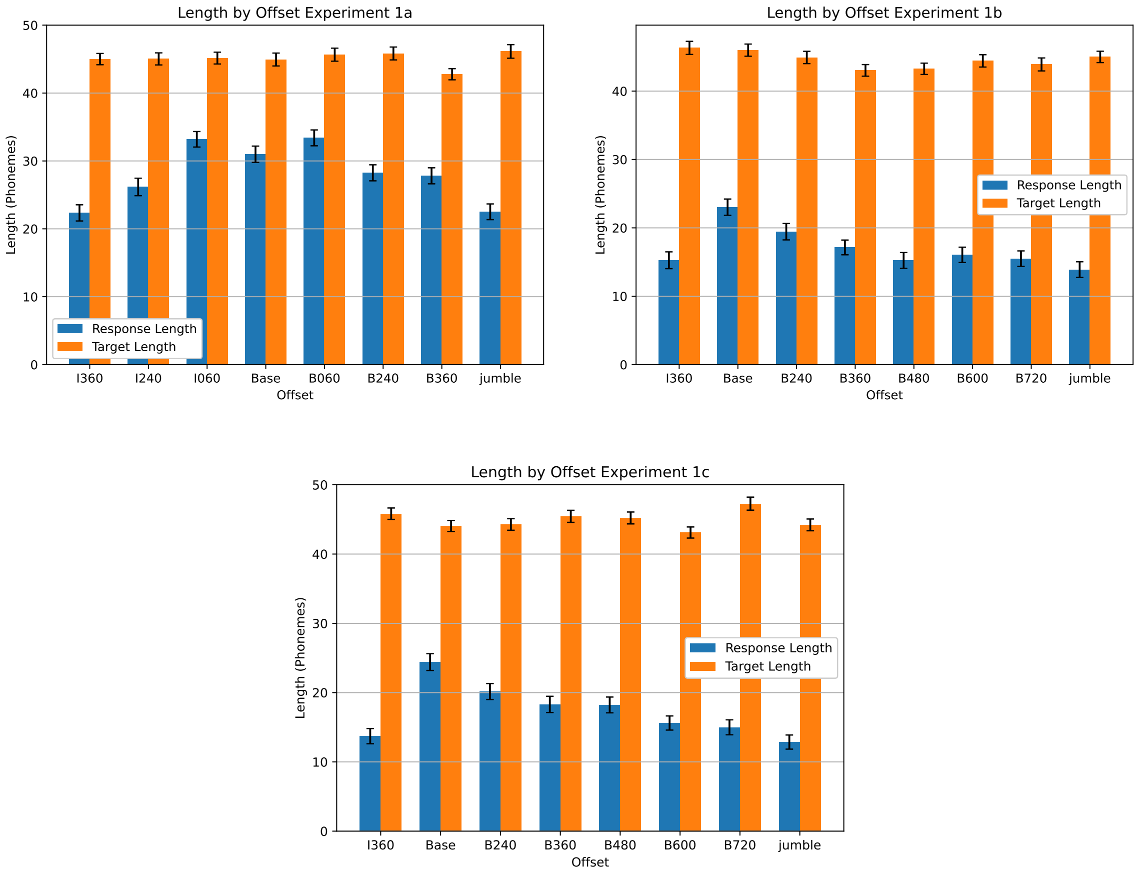

Figure 4.4: **Experiment 1 Length Analysis.** These graphs show that the length of responses is 1) much shorter than that of the target on average, and 2) inversely related to overall PER.

was consistently lower than the second half as shown by figure 4.3, which was the same as experiment 1A. Finally, when examining response lengths, we saw the same relationship between overall error rate and average response length as seen in figure 4.4 seen in experiment 1A.

4.4.3 Experiment 1C

Experiment 1B replicated the asymmetry shown in Experiment 1A and showed that it extended out to approximately 500 ms of lag, when video led audio. However the substantial increase in noise floor of Experiment 1B resulted in much higher overall PER relative to Experiment 1A and appeared to cause a flooring effect in the results. For this reason we repeated Experiment 1B but with less noise added to the speech signal. Experiment 1C had 29 participants (22 female, 7 male); 24 were native English speakers and none had known hearing impairments. As in Experiment 1B, we tested the base and jumble along with I360, B240, B360, B480, B600, B720. In this experiment the noise was balanced at -7dB relative to the speaker audio track in an attempt to resolve the flooring effect seen in Experiment 1B.

Results

Experiment 1C replicated Experiments 1A and 1B. Graphs for overall PER in experiment 1C can be found in figure 4.1 and t-test results can be found in table 4.1. The results of experiment 1C found that the base, B240, B360, B480, and B600 lags were significantly different from the jumble condition and the I360, and B720 lags were not. Additionally, the I/B360 lags were significantly different from each other. With the B480 and B600 both being significantly different from the jumble in this experiment, there is likely a similar boundary between the B600 and B720 lags as the one seen between I060 and I240 in experiment 1A. In terms of correction operations, figure 4.2 shows that in experiment 1C insertions were the primary required correction, and modulates most with overall error rate. This was consistent with the previous experiments. For first and second half error rates we found that first half was consistently lower than second half as shown by figure 4.3, which was consistent with the previous experiments. Finally, when examining response lengths, we saw the same relationship between overall error rate and average response length seen in experiments 1A and 1B as shown by figure 4.4.

4.5 Experiment 2 - Deep AVSR

Experiment 1 demonstrated an asymmetrical dependence of AVSR on positive versus negative lag between audio and video. This result reveals a critical aspect of how humans perceive and fuse visual information in the service of auditory speech recognition. Experiment 2 sought to test whether an effective deep neural network trained to recognize audio-visual speech exhibits a similar asymmetric dependency on temporal synchronization.

For this experiment a pipeline was created in which the network was initialized with pre-trained weights provided by the developer, then each video was processed by the network as per the procedure described in section 3.3. The raw responses were then recorded to a text file along with the speaker, clip number, and lag. It is important to note that the artificial model is not capable of learning the dataset as it was not in a training mode. Therefore it was possible to test every lag of every clip from every speaker. We did not train the network ourselves as 1) the pre-trained weights are made available for research purposes, and 2) the training would have taken approximately 2 years on the hardware we had available at the time.

4.5.1 Data Set

We tested the lags +/-060, +/-240, +/-360, +/-480, +/-600, +/-720, and +/-840 as well as the base and jumble videos. Additionally, the program was forced to use audio and visual information so we used clips that did not contain the babble noise present in the human experiments as the inclusion of this noise showed a flooring effect in experiments.

4.5.2 Results

The results of Experiment 2 reveal that Deep AVSR performs quite unlike human listeners with respect to audio-visual synchronization. Whereas human listeners exhibit a pronounced robustness to video leading audio, Deep AVSR can make use of visual information only when it was tightly synchronized with audio. Furthermore the asymmetric relationship between error rate and lag that was characteristic of human listeners was not evident in the Deep AVSR results. Graphs for the overall PER in

Comparison Type	Offset	p-values
Same Temporal Value	I/B060	0.826
	I/B240	0.841
	I/B360	0.987
	I/B480	0.772
	I/B600	0.838
	I/B720	0.418
	I/B840	0.352
Against Jumble	I840	0.259
	I720	0.254
	I600	0.170
	I480	0.150
	I360	0.165
	I240	0.156
	I060	*2.02e-5
	Base	*1.11e-6
	B060	*4.80e-6
	B240	0.121
	B360	0.179
	B480	9.66e-2
	B600	0.113
	B720	5.71e-2
	B840	†4.20e-2

Table 4.2: **T-Test Results for Experiment 2.** These results indicate that there is a symmetric and small window in which Deep AVSR is able to use visual speech to aid in AVSR.
* p-value < 0.05
† p-value < 0.05, but the Benjamini-Hochberg procedure (FDR=5%) determined this to be a false discovery

Deep AVSR's responses can be found in figure 4.5 with p-scores for the t-tests in table 4.2. These results show that only the I/B060 and base conditions were significantly different from the jumble condition, and that none of the lagged pairs were significantly different from their I/B counterpart. This demonstrates a lack of the performance asymmetry seen in experiment 1. Regarding the analysis of operations, as seen in figure 4.6, transformations were the primary source of error in addition to being the source of error that modulated most with overall error rate. When comparing error rates in the first and second half of responses, we found that error rate was

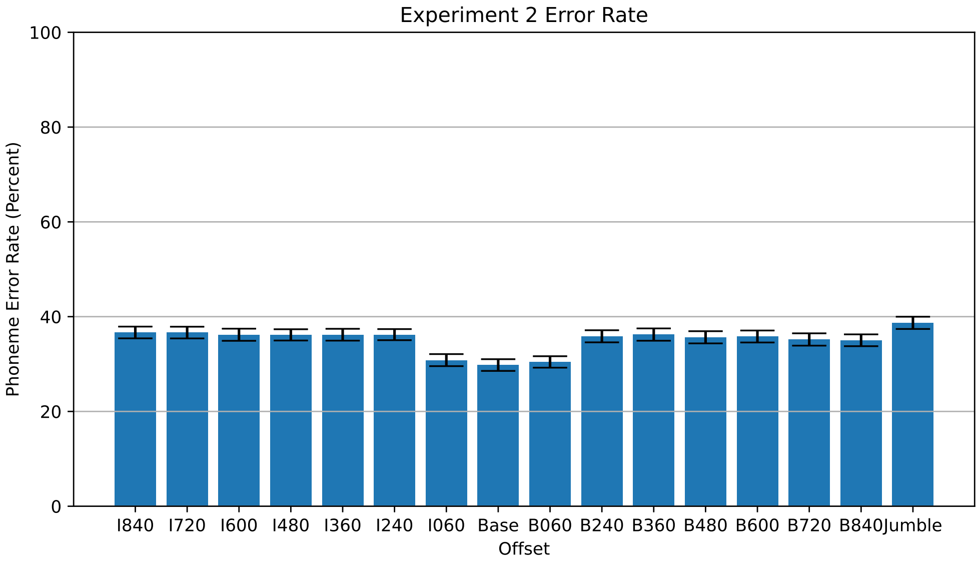

Figure 4.5: **Computer Experiment PER Analysis.** These results show that no paired I/B lags are significantly different from each other, demonstrating symmetric performance across various lags.

consistently lower in the first half, though not significantly so as seen in figure 4.7. When looking at response and target lengths in figure 4.8, we found that responses were consistently shorter than targets, though not to the extent seen in the human trials, and these lengths modulated slightly with overall error rate.

4.6 Discussion

4.6.1 Summary of Key Findings

Our research found that in the human results there was a distinct, temporally asymmetric pattern of performance. Specifically, the pattern was consistent with the "queuing" interpretation of predictive coding - suggesting that visual speech acts as a dynamic, forward predictor of upcoming auditory speech. Figures 4.1 and 4.5 show that human listeners exhibited a surprising tolerance to audio lag (up to several hundred milliseconds) but were substantially less tolerant of video lag. By contrast, we found that Deep AVSR exhibited a rapid drop in performance in both temporal directions as the value of the lag grew, consistent with the "mapping" interpretation of

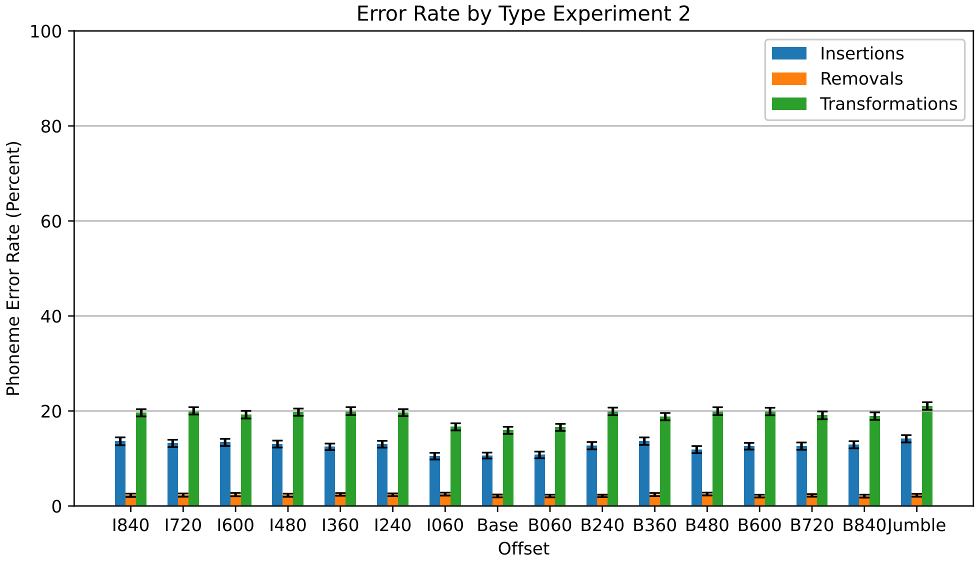

Figure 4.6: **Experiment 2 PER Analysis by Operation.** This graph shows the PER broken into correction types. We can see that transformations are the most common correction, and that it is what modulates with overall PER.

predictive coding. This pattern in the Deep AVSR data is consistent with a mechanism in which multi-sensory evidence is fused with a tightly controlled time relationship across the modalities. In this mechanism, current visual speech is used to make predictions about future auditory speech without any flexibility in when that auditory evidence should appear in the sequence. Thus Deep AVSR, at least in its current architecture, uses different mechanisms to fuse visual and auditory speech information as compared to human listeners.

We also saw a distinct difference between human participants and Deep AVSR with respect to the kinds of errors that were made. In figures 4.2 and 4.6 it is shown that the human participants and Deep AVSR consistently had different patterns in terms of the required correction types, and which correction type modulated most with overall PER. In the human participants we saw transformations and removals stay relatively consistent across different lags, whereas a decrease in the rate of required insertions accounted for decreases in overall PER. This contrasts with Deep AVSR for which the insertions and removals stayed relatively consistent while transformations

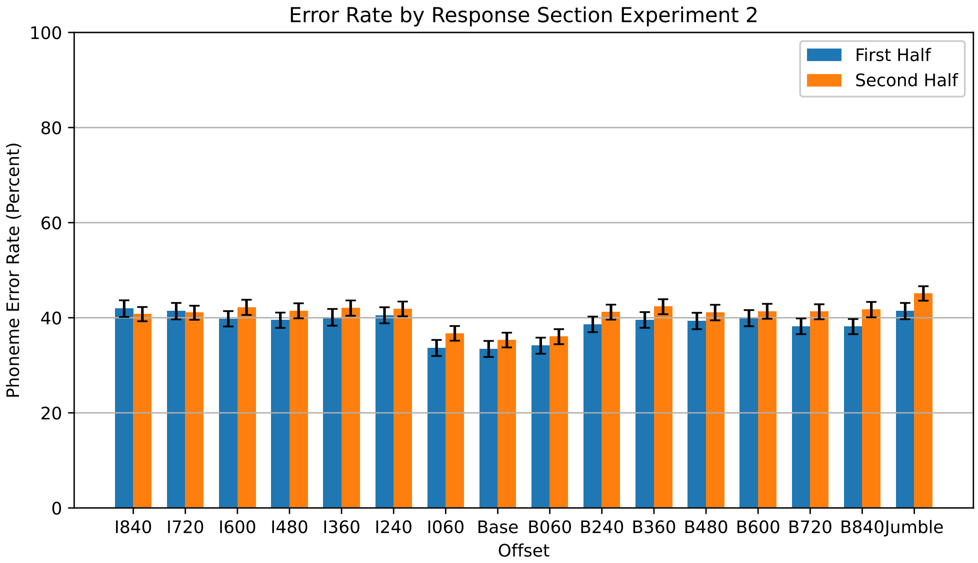

Figure 4.7: **Experiment 2 PER Analysis by Response Section.** This graph shows the error rate in the first and second half of Deep AVSR's responses. These results contradict what we would expect to see based on the predictive coding model of AVSR.

were modulated with lag. This behaviour makes sense when one considers response length across lags: human responses were consistently shorter than the targets. By contrast, for Deep AVSR the lengths of responses were more similar to that of the targets as shown in figures 4.4 and 4.8. Thus when human listeners had difficulty with the task, they tended to simply not report words, whereas Deep AVSR tended to guess incorrectly.

4.6.2 Interpretation of Key Findings

Overview

This research has shown that when auditory and visual components of speech are temporally lagged, human listeners demonstrate an asymmetric reduction in AVSR performance with respect to the direction of lag. This result is consistent with prior work, which showed that the inclusion of visual speech with auditory speech was more likely to aid AVSR performance when visual speech was aligned with or temporally

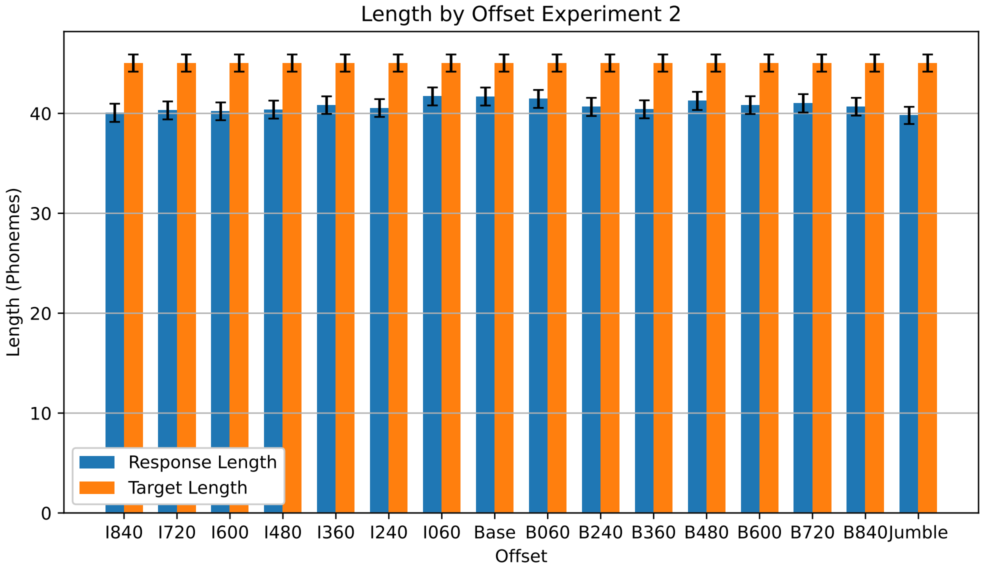

Figure 4.8: **Experiment 2 Length Analysis.** This graph shows that response lengths were slightly shorter on average than target lengths.

led auditory speech than when it lagged auditory speech [40, 36, 55, 7]. For example, the McGurk effect exhibits a similar asymmetry over similar range of asynchronies [40]. In the range of -120 to +300, the McGurk effect was consistently present (50% or greater rate of misperceived syllables) and it was particularly prominent (60% or greater misperception rate) in the range +/-000 to +180. This indicates a requirement that visual information needs to be aligned with or lead auditory information for it to have a significant effect on the human perception of AV speech [40]. Our work is consistent with these results and reinforces the assertion that visual speech perception in humans acts as a forward predictor for upcoming auditory information.

Deep AVSR does not show this same asymmetry, instead showing a temporally symmetric drop in performance to similar that of the jumble condition at any tested lag greater than 60ms. Thus there is a fundamental difference in the human and artificial methods of performing AVSR. We speculate that this is a result of two main factors, 1) the network architecture, and 2) the training data. First, the TM-CTC architecture does not contain recurrent, convolutional, or LSTM units that function in the temporal domain. Instead, it relies on positional encoding to allow for temporal resolution in the analysis of stimuli. While this method is effective in the trials performed by the network developers and allows for parallel computing to increase real-time efficiency, it may not allow for information in the auditory and visual domain to be shifted relatively in time to affect the output at a broad range of time steps. Rather, by enforcing a static representation of cross-modal temporal dynamics learned from the BBC-LRS2 training data it mismatches encodings in time when faced with asynchronies it did not encounter in training. It has been shown that the true temporal relationship of real-world AV speech is inconsistent. When examining initial motivating mouth movements and auditory onset, depending on the phoneme and speaker, the delay between mouth movement and corresponding acoustic phoneme is in a range from 100 to 400ms [49]. The implication of this is that in the TM-CTC architecture, the visual onset of each mouth movement will be applied to the relative position in the future where the auditory onset is expected to happen, and in this way, visual information can predict, or at least inform, auditory information. As there is a speaker dependency in this lag, Deep AVSR should learn some range of asynchrony for each phoneme, however, it has not been trained with other natural lags. As a result, any significant lag in the stimuli will greatly reduce the usefulness of the visual information and it could even begin acting as a distractor - with mismatching frames of evidence being merged across modalities. For these reasons, the architecture of Deep AVSR makes it more efficient in processing speech data when assumptions of a fixed asymmetry are met, but it also "perceives" these data in a fundamentally different way that humans. Another important note is that the dataset used to train the network (BBC-LRS2) consisted of television clips. These clips will have very closely aligned data, devoid of the natural lags that listeners will often encounter in realistic audio-visual stimuli (see section 4.6.3). This is to say that there is a clear distinction in how Deep AVSR and human listeners tend to receive audio-visual information, particularly during learning. These are likely two prominent factors in the different performance patterns seen in the human and Deep AVSR PER results.

Error in Perception

When perceiving speech, whether by human listeners or by artificial means, there is always the potential for errors. Broadly speaking we observed three ways for error to be introduced into the responses of our subjects as described by the error rate metric. The observed error rate tracks these by the correction required (insertion, transformation, removal), which we take to reflect something about the underlying (error) process in the brain (or computer). Identifying the error rather than the correction provides insight into the root cause of errors in perception. These errors are replacing a phoneme that was said with a different one that was not spoken, creating a brand new phoneme, and not perceiving something that was said. I refer to these sources of error as replacement (corrected by transformation/substitution), creation (corrected by removal), and omission (corrected by insertion) respectively. Differentiating between replacement and creation errors is difficult as it is highly dependant on context, though both are caused by errors in the synthesis of information, so we group substitution and creation into generative error as both involve generating information that does not exist in the stimuli. Errors of omission are errors where produced speech is not perceived. This gives us two broad categories of error to examine, generative and omissive.

The assertion that there is a fundamental difference in how humans and Deep AVSR perform speech recognition is supported by the analysis of error types as it shows that the type of errors made by human participants and Deep AVSR are distinctly different. Humans primarily made omissive errors, whereas Deep AVSR primarily made generative errors. When examining the average response lengths at different lags we see why this is the case. As shown in table 4.3, human participants had shorter responses overall when compared to both the targets and Deep AVSR results. Further more, in figure 4.4, we see that higher error rate lags and noise conditions result in shorter average response lengths relative to less adverse experimental conditions in the human participants. By contrast, Deep AVSR consistently had a near identical average response length and variance with the targets. Notably the length of response does not modulate with error rate in any significant way. This explains the difference in correction type. Additionally, it shows that Deep AVSR is consistently able to solve the segmentation problem, finding close to the right number of phonemes despite often struggling with correct selection. Conversely, human participants appear to struggle to perceive the presented speech in the first place, however when they do so successfully, they appear to be much more likely to select the correct word, and by extension phonemes. It is important to note that in this experiment humans have the unique ability to simply give up and not input a response, or input a partial response if they are uncertain. In the experimental instructions they were asked to do their best, but if they didn't know an answer they could enter an empty response. Being a computer program, Deep AVSR will always respond with its best guess. This could partially explain the difference in error types in the human and computer participants, though further research would be required to confirm this.

Analysis of First- vs Second-half of Sentences: Real-time Adaptation to Audio-Visual Asynchrony and Bayesian Prediction

An interesting possible explanation of the difference in performance between humans and Deep AVSR might be that humans are able to adapt in near real-time to lags between audio and video. To consider this possibility we also analyzed error rates split between the first and second half of responses (see section 4.3 for the analysis approach). We predicted that real-time adaptation to a particular audio-visual lag

Experiment	Category	Average Length	Standard Deviation
Experiment 1a	Target	45.06	0.91
	Response	28.08	1.19
Experiment 1b	Target	44.60	0.89
	Response	16.95	1.15
Experiment 1c	Target	44.93	0.85
	Response	17.28	1.11
Experiment 2	Target	45.04	0.86
	Response	40.74	0.90

Table 4.3: **Response and Target Lengths and Variance for all Experiments.** These results demonstrate the differences between the lengths and variance of the human and Deep AVSR responses as compared to each other and the targets. The human participants had shorter and more varied response lengths as compared to Deep AVSR.

should result in greater error rates in the first half relative to the second half of each sentence. However, this was not the case as shown in figures 4.3 and 4.7. Instead both humans and Deep AVSR tended to make more errors related to the second half of sentences relative to the first half.

This result also stands in contrast to a predictive coding view of AVSR. Broadly, if a Bayesian mechanism exists that adjusts an adapting linguistic model to make predictions about upcoming speech sounds, one would expect it to perform better as evidence accumulates over time. Predictive coding should work better for the second half of a sentence, since the predictive mechanisms should benefit from the constraints provided by the first few words in a sentence. The fact that we see precisely the opposite pattern in our data suggests that a simple Bayesian mechanism is not a sufficiently good model of AVSR. We speculate that the performance difference seen is a result of either reduced attention in the second half of the task or sentences being more complex and difficult to predict in their second halves, though more research is required to identify the root cause of this effect.

Presentation of Predictive Coding

In the introduction we discussed that there were two ways that visual information might be used in the service of auditory speech recognition: one in which vision is used to predict upcoming audio and another in which visual information is encoded, cached, and fused with audio with an appropriate lag. Our data do not speak directly to these alternative mechanisms, but they do impose a key requirement for any future model of audio-visual speech: regardless of the mechanism, it is remarkably flexible with respect to timing provided that video leads audio, and quite inflexible when audio leads video. We speculate that it is a predictive mechanism that is likely to be more tolerant of asynchrony.

This is interesting in that it shows that how predictive coding presents is not strictly dependant on the presented task. More research is required to determine what precisely created the behavioural difference seen in the artificial and human participants, however it is likely a combination of how Deep AVSR was designed, specifically how it handles time-series stimuli, as well as the training data, which would have little to no temporal lag in AV alignment, are major contributing factors. It may seem obvious that AV training stimuli would or even should be temporally aligned, though when you consider differences in the nature of audio and visual stimuli, perhaps introducing a temporal lag into it may be a good idea as we learn using stimuli that has a lag as discussed in section 4.6.3.

4.6.3 The Temporal Nature of Audio-Visual Stimuli

Previously we have discussed how the motor movements (visual stimuli) in speech occur prior to auditory onset. We have yet to discuss differences in the speed at which light and sound travel. The speed of light is effectively infinite on the spatial scale over which a conversation can reasonably occur. By contrast, the speed of sound is a much slower $343m/s$. This means that all observed speech has an AV lag introduced by how the two forms of stimuli travel to an observer. This inherent lag between video and audio is, however, quite small relative to the set of lags tested in the present study. If someone speaks to you from 10 meters away, there is just over 29ms of lag between the visual and auditory stimuli. The LRS2 dataset used to train Deep AVSR uses stimuli taken from television programs. In the case of a newscast or a narration, microphones are generally placed within a metre of the talker, and in the case of a conventional show, there is often a microphone directly above the talker just out of frame. This means there is little to no lag in the expected visual and auditory onsets, though in the real world, conversations can happen over relatively large distances, on the order of tens of meters depending on context. This is to say that people learn to perceive audio-visual speech with variable lags that are roughly similar to the smallest lag (60ms, or approximately 20 meters) used in our stimuli.

The relatively long lags tested in the present study, especially in Experiment 1B and 1C present an interesting question regarding audio-visual perception. The 360ms lag is equivalent to listening to someone from over 100 metres away. At this distance, the entire face of the talker would span approximately 1/10th of a degree of visual angle and the speech would be barely distinguishable above a noise floor. Thus, we rarely encounter meaningful speech stimuli at the lags tested in the present study, yet participants were at least somewhat tolerant of lags in the hundreds of milliseconds. One explanation is that speech is not the only form of audio visual stimuli since people regularly observe events with associated sounds at long distances, albeit not usually intelligible speech. Therefore it is reasonable to think that we may have learned a general-purpose dynamic temporal relationship of visual and auditory events not specific to speech, and tolerance to lag in audio-visual speech is simply a consequence of tolerance to lag in other kinds of sound events. Similarly, there are no naturally occurring scenarios in which an auditory stimulus will arrive earlier than its associated visual event.

4.6.4 Limitations of the Findings

Accuracy of Metrics and the Types of Errors Generated by Humans and Deep AVSR

It is important to note that the metric used in the analysis of these results will have an artificially high error rate in the human participants, but not in Deep AVSR. This is because, for the human participants, spelling errors could contribute towards the overall PER if it changes how the phoneme conversion engine analyzed them. In the case of Deep AVSR, the language model directly selects words based on a scoring method, meaning that while it may not select the correct word, the selected word will always be spelled correctly, and thus there is no chance of accidental phoneme corruption. Notably, both groups are still able to make grammatical errors, though grammatical errors will usually not result in a phonetic difference, so there was likely minimal effect for this.

Prior Work in Adaptation for Audio-Visual Asynchrony

We performed the analysis of comparing first- and second-half split responses with the aim of testing for an adaptation effect in AVSR. Namely, we expected performance would improve in the second half of a response if a participant was able to adapt to lag in near-real-time. We concluded that the results do not support such an effect, though we also state that these results also do not support the complete absence of this effect. It has been shown by Lennert et al. that adaptation effects in audio-visual perception appeared over multiple consecutive trials [34], however, their participants could not adapt within individual trials. It should be noted that in the Lennert et al. study, a simple visual dot and auditory beep was used as the stimulus making adaptation in a single trial virtually impossible, however it shows that humans find it difficult to compensate for audio-visual lags with a small sample size of the lag. Nevertheless, the sentences presented to listeners on each trial were on a relatively short scale, and listeners encountering temporal lags at random on successive trials. These factors suggest that listeners were unable to make rapid adaptations to audio-visual lag, at least within the parameters of our study, and there is the possibility of long-term adaptation being possible.

Error Rate Discrepancies in Deep AVSR

One major concern with our results is the difference in performance when comparing our tests to that of the Deep AVSR developers. According to the repository for Deep AVSR [50], they were able to achieve a WER of 6.8% (the Deep AVSR developers did not use PER for training or testing) where in our trials the WER was 56.5% under the same conditions (no temporal lag or background noise). Previously we have discussed the testing and training data sets for Deep AVSR, that being LRS2-BBC, though in our experiments we used recordings of the TIMIT dataset. In addition to differences in general speaker accent, the LRS2 dataset is a more accurate representation of natural English speech as compared to TIMIT, which was designed to be a stringent test of phoneme combinations and speech patterns in English, with sentences often having complex sentence structure and a broad vocabulary. Additionally, the language model used was also trained on the LRS2-BBC dataset. As a language model utilizes linguistic structure to make predictions and it was trained on a much simpler dataset, it would likely have difficulty analyzing the generally more complex linguistic structures seen in TIMIT. It should also be noted that the testing data set was recorded at a much higher quality than the network can take as an input. For this reason, the data set had to be edited and exported multiple times to reduce the size of the video frame and reduce the audio sample rate. Despite our attempts to limit data compression, the quality of the audio and video was relatively low compared to what was shown to the human participants. Although these factors together explain why the performance of Deep AVSR was substantially worse on our test than reported buy the network developers, it nevertheless performed above chance. The goal was not to compare Deep AVSR to humans in terms of absolute performance, but rather to consider whether Deep AVSR would exhibit the same pattern of sensitivity to audio-visual lag. We conclude that, despite its relative overall poor performance, the evidence clearly suggests that Deep AVSR would not exhibit human-like tolerance to positive lags under any circumstances short of retraining it on datasets with longer lags explicitly included.

Audio Only Condition

The final notable limitation in this research is that there was no comparison with an audio-only condition. This was considered as an alternative to the jumble condition to fill the role of the worst case scenario. Ultimately we chose to use the jumble condition as we wanted to keep the experiment as an AVSR task. It could be beneficial for future studies to include an audio-only condition along side the jumble condition to test if incongruent visual speech acts as a distractor or is simply ignored in AVSR.

4.6.5 Future Directions

Finding a more effective method for testing for an adaptation effect should be a priority for future studies. Presenting consecutive trials with the same temporal lag and then comparing PER between consecutive trials would allow for the longer-term lag adaptation seen in Lennert et al. [34] to occur. This would allow for a more definitive conclusion regarding the presence of such a compensating effect in AVSR.

In addition to modifying the experiment to test for an adaptation effect, it is imperative to also test different ANN architectures. In the background section we discussed the history of AVSR in artificial models, and one of the previous methods involved the use of recurrent or LSTM units rather than the current TM-CTC method. Making use of recurrent or LSTM units might allow for the forward propagation of visual information in a broad temporal range.

Additionally, including stimuli with slightly lagged audio-visual streams would replicate the natural processes for learning AVSR. Developing an ANN that has structures to allow for broad temporal resolution and including training stimuli with temporal lags could result in the creation of an AVSR network that replicates the behavioural phenomena seen in human AVSR with asynchronous audio-visual streams, and potentially create a more robust ANN for AVSR. In its current form, Deep AVSR would be a poor choice for realistic free-field applications such as robotics, in which speech sometimes must propagate over long distances. Deep AVSR is not tolerant of such lags, but could be altered and trained otherwise.

Chapter 5

Conclusion

Humans have incredible capabilities for perceiving audio-visual speech in adverse conditions. With the rise of artificial ANNs, automatic speech recognition is becoming a viable, everyday technology that can be leveraged by even basic computers. Given the prominence of webcams, AVSR is a widely accessible option, though determining the optimal mechanisms and architectures becomes difficult when considering the speed at which these technologies evolve.

5.1 Summary of Work and Findings

By using PER as an evaluation metric we were able to perform experiments to identify and compare behavioural patterns present in both human participants and the ANN chosen for these experiments, Deep AVSR. We were able to demonstrate that behavioural patterns regarding the temporal integration of audio-visual stimuli shown by people in the perception of asynchronous speech-like stimuli are also present in the recognition of natural speech. Conversely, this behavioural pattern was not demonstrated by Deep AVSR. Human participants demonstrated "queuing" and Deep AVSR showed a "mapping" pattern. This highlights a distinct difference in how humans observers and Deep AVSR perceive audio-visual speech information. It is unclear if this is a result of the training data, ANN architecture, or a combination of these and other factors, though we can say with confidence that Deep AVSR with the weights provided by the developers is not optimal for performing AVSR in natural environments that might include long lags between video and audio. We were unable to identify a near-term adaptation effect in our results which is counter to a purely Bayesian interpretation of predictive coding in AVSR.

5.2 Future Directions

This work has led to several ideas regarding future work. First, it is our hope that the results found in this work can be used to improve the performance and efficiency of future ANN designs. This would be achieved by implementing structures into an ANN that allow for dynamic temporal resolution as well as training on asynchronous audio-visual data. Second, we would like to perform similar experiments that include audio-only and jumble conditions to assess if visual speech can act as a distractor. Finally, we would like to modify the procedure to test if consecutive presentations of the same asynchrony allow for the perceptual compensation of said asynchrony.

Ultimately this work has identified a previously unknown effect, that human participants and Deep AVSR demonstrate separate and distinct patterns in AVSR performance when presented with asynchronous AV speech conditions, and paved the groundwork for several future studies.

Bibliography

[1] Triantafyllos Afouras et al. “Deep Audio-visual Speech Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018), pp. 1–1. DOI: 10.1109/TPAMI.2018.2889052.

[2] Vikranth Rao Bejjanki et al. “Cue Integration in Categorical Tasks: Insights from Audio-Visual Speech Perception”. In: *PLOS ONE* 6.5 (May 2011), pp. 1–12. DOI: 10.1371/journal.pone.0019812. URL: https://doi.org/10.1371/journal.pone.0019812.

[3] Mathieu Bernard and Hadrien Titeux. “Phonemizer: Text to Phones Transcription for Multiple Languages in Python”. In: *Journal of Open Source Software* 6.68 (2021), p. 3958. DOI: 10.21105/joss.03958. URL: https://doi.org/10.21105/joss.03958.

[4] Mathieu Bourguignon et al. “Lip-Reading Enables the Brain to Synthesize Auditory Features of Unknown Silent Speech”. In: *Journal of Neuroscience* 40.5 (2020), pp. 1053–1065. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.1101-19.2019. eprint: https://www.jneurosci.org/content/40/5/1053.full.pdf. URL: https://www.jneurosci.org/content/40/5/1053.

[5] E. Colin Cherry. “Some Experiments on the Recognition of Speech, with One and with Two Ears”. In: *The Journal of the Acoustical Society of America* 25.5 (1953), pp. 975–979. DOI: 10.1121/1.1907229. eprint: https://doi.org/10.1121/1.1907229. URL: https://doi.org/10.1121/1.1907229.

[6] Chung-Cheng Chiu et al. “State-of-the-art speech recognition with sequence-to-sequence models”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 4774–4778.

[7] Brianna L Conrey and David B Pisoni. “Audiovisual asynchrony detection for speech and nonspeech signals”. In: *AVSP 2003-International Conference on Audio-Visual Speech Processing*. 2003.

[8] Ariel Ephrat et al. “Looking to listen at the cocktail party”. In: *ACM Transactions on Graphics* 37.4 (Aug. 2018), pp. 1–11. ISSN: 1557-7368. DOI: 10.1145/3197517.3201357. URL: http://dx.doi.org/10.1145/3197517.3201357.

[9] Norman P. Erber. “Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli”. In: *Journal of Speech and Hearing Research* 12.2 (1969), pp. 423–425. DOI: 10.1044/jshr.1202.423. eprint: https://pubs.asha.org/doi/pdf/10.1044/jshr.1202.423. URL: https://pubs.asha.org/doi/abs/10.1044/jshr.1202.423.

[10] Janet Fletcher. "The prosody of speech: Timing and rhythm". In: *The handbook of phonetic sciences* (2010), pp. 521–602.

[11] Karl Friston. "A theory of cortical responses". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1456 (2005), pp. 815–836. DOI: 10.1098/rstb.2005.1622. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2005.1622. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2005.1622.

[12] John Garofolo et al. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Version V1. 1993. DOI: 11272.1/AB2/SWVENO. URL: https://hdl.handle.net/11272.1/AB2/SWVENO.

[13] Bruno L Giordano et al. "Contributions of local speech encoding and functional connectivity to audio-visual speech perception". In: *eLife* 6 (June 2017). Ed. by Charles E Schroeder, e24763. ISSN: 2050-084X. DOI: 10.7554/eLife.24763. URL: https://doi.org/10.7554/eLife.24763.

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[15] Michael S. Gordon and Suzanne Allen. "Audiovisual Speech in Older and Younger Adults: Integrating a Distorted Visual Signal With Speech in Noise". In: *Experimental Aging Research* 35.2 (2009). PMID: 19280447, pp. 202–219. DOI: 10.1080/03610730902720398. eprint: https://doi.org/10.1080/03610730902720398. URL: https://doi.org/10.1080/03610730902720398.

[16] Alex Graves et al. "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 369–376. ISBN: 1595933832. DOI: 10.1145/1143844.1143891. URL: https://doi.org/10.1145/1143844.1143891.

[17] Awni Hannun et al. "Deep speech: Scaling up end-to-end speech recognition". In: *arXiv preprint arXiv:1412.5567* (2014).

[18] Saeedeh Hashemnia et al. "Human EEG and Recurrent Neural Networks Exhibit Common Temporal Dynamics During Speech Recognition". In: *Frontiers in Systems Neuroscience* 15 (2021), p. 617605.

[19] Simon Haykin and Zhe Chen. "The Cocktail Party Problem". In: *Neural Computation* 17.9 (Sept. 2005), pp. 1875–1902. ISSN: 0899-7667. DOI: 10.1162/0899766054322964. eprint: https://direct.mit.edu/neco/article-pdf/17/9/1875/816375/0899766054322964.pdf. URL: https://doi.org/10.1162/0899766054322964.

[20] David J Heeger. "Theory of cortical function". In: *Proceedings of the National Academy of Sciences* 114.8 (2017), pp. 1773–1782.

[21] Steven A Hillyard, Edward K Vogel, and Steven J Luck. "Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 353.1373 (1998), pp. 1257–1270.

[22] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: `10.1162/neco.1997.9.8.1735`.

[23] Tomoyasu Horikawa and Yukiyasu Kamitani. "Generic decoding of seen and imagined objects using hierarchical visual features". In: *Nature communications* 8.1 (2017), pp. 1–15.

[24] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[25] Hirofumi Inaguma, Masato Mimura, and Tatsuya Kawahara. "Enhancing monotonic multihead attention for streaming asr". In: *arXiv preprint arXiv:2005.09394* (2020).

[26] Jacob Kahn et al. "Libri-light: A benchmark for asr with limited or no supervision". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 7669–7673.

[27] Gerald Kidd et al. "Stimulus factors influencing spatial release from speech-on-speech masking". In: *The Journal of the Acoustical Society of America* 128 (Oct. 2010), pp. 1965–78. DOI: `10.1121/1.3478781`.

[28] Jon Kleinberg, Christos Papadimitriou, and Prabhakar Raghavan. "Segmentation problems". In: *Journal of the ACM (JACM)* 51.2 (2004), pp. 263–280.

[29] Simon Kornblith et al. "Similarity of neural network representations revisited". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3519–3529.

[30] Nikolaus Kriegeskorte and Pamela K Douglas. "Cognitive computational neuroscience". In: *Nature neuroscience* 21.9 (2018), pp. 1148–1160.

[31] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. "Representational similarity analysis-connecting the branches of systems neuroscience". In: *Frontiers in systems neuroscience* (2008), p. 4.

[32] Nitish Krishnamurthy and John HL Hansen. "Babble noise: modeling, analysis, and applications". In: *IEEE transactions on audio, speech, and language processing* 17.7 (2009), pp. 1394–1407.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[34] Therese Lennert, Soheila Samiee, and Sylvain Baillet. "Coupled oscillations enable rapid temporal recalibration to audiovisual asynchrony". In: *Communications Biology* 4 (May 2021), p. 559. DOI: `10.1038/s42003-021-02087-0`.

[35] Esther Levin. “A recurrent neural network: Limitations and training”. In: *Neural Networks* 3.6 (1990), pp. 641–650.

[36] Dominic W Massaro, Michael M Cohen, and Paula MT Smeele. “Perception of asynchronous and conflicting visual and auditory speech”. In: *The Journal of the Acoustical Society of America* 100.3 (1996), pp. 1777–1786.

[37] Sven L Mattys, Laurence White, and James F Melhorn. “Integration of multiple speech segmentation cues: a hierarchical framework.” In: *Journal of Experimental Psychology: General* 134.4 (2005), p. 477.

[38] Harry McGurk and John MacDonald. “Hearing lips and seeing voices”. In: *Nature* 264.5588 (1976), pp. 746–748.

[39] Larry R Medsker and LC Jain. “Recurrent neural networks”. In: *Design and Applications* 5 (2001), pp. 64–67.

[40] K. Munhall et al. “Temporal constraints on the McGurk effect”. In: *Perception and Psychophysics* 58.3 (1996), pp. 351–362.

[41] Costas Neocleous and Christos Schizas. “Artificial Neural Network Learning: A Comparative Review”. In: *Methods and Applications of Artificial Intelligence*. Ed. by Ioannis P. Vlahavas and Constantine D. Spyropoulos. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 300–313. ISBN: 978-3-540-46014-5.

[42] Jonathan E Peelle and Mitchell S Sommers. “Prediction and constraint in audiovisual speech perception”. In: *Cortex* 68 (2015), pp. 169–181.

[43] David Poeppel and M. Assaneo. “Speech rhythms and their neural foundations”. In: *Nature Reviews Neuroscience* 21 (May 2020), pp. 1–13. DOI: 10.1038/s41583-020-0304-4.

[44] Rajesh PN Rao and Dana H Ballard. “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects”. In: *Nature neuroscience* 2.1 (1999), pp. 79–87.

[45] Gregg H Recanzone. “Interactions of auditory and visual stimuli in space and time”. In: *Hearing research* 258.1-2 (2009), pp. 89–99.

[46] Benjamin Recht et al. “Do imagenet classifiers generalize to imagenet?” In: *International Conference on Machine Learning*. PMLR. 2019, pp. 5389–5400.

[47] Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386.

[48] Martin Schrimpf et al. “Brain-score: Which artificial neural network for object recognition is most brain-like?” In: *BioRxiv* (2020), p. 407007.

[49] Jean-Luc Schwartz and Christophe Savariaux. “No, There Is No 150 ms Lead of Visual Speech on Auditory Speech, but a Range of Audiovisual Asynchronies Varying from Small Audio Lead to Large Audio Lag”. In: *PLOS Computational Biology* 10.7 (July 2014), pp. 1–10. DOI: 10.1371/journal.pcbi.1003743. URL: https://doi.org/10.1371/journal.pcbi.1003743.

[50] Smeet Shah. *Deep AVSR.* URL: `https://github.com/lordmartian/deep_avsr`. (accessed: 23.03.2021).

[51] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. "Robust Self-Supervised Audio-Visual Speech Recognition". In: *arXiv preprint arXiv:2201.01763* (2022).

[52] Brendan Shillingford et al. "Large-scale visual speech recognition". In: *arXiv preprint arXiv:1807.05162* (2018).

[53] Jeremy I Skipper, Howard C Nusbaum, and Steven L Small. "Listening to talking faces: motor cortical activation during speech perception". In: *Neuroimage* 25.1 (2005), pp. 76–89.

[54] William H Sumby and Irwin Pollack. "Visual contribution to speech intelligibility in noise". In: *The journal of the acoustical society of america* 26.2 (1954), pp. 212–215.

[55] Virginie Van Wassenhove, Ken W Grant, and David Poeppel. "Temporal window of integration in auditory-visual speech perception". In: *Neuropsychologia* 45.3 (2007), pp. 598–607.

[56] Julian FV Vincent et al. "Biomimetics: its practice and theory". In: *Journal of the Royal Society Interface* 3.9 (2006), pp. 471–482.

[57] Ye-Yi Wang, Alex Acero, and Ciprian Chelba. "Is word error rate a good indicator for spoken language understanding accuracy". In: *2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721)*. IEEE. 2003, pp. 577–582.

[58] Brigitte Zellner. "Pauses and the temporal structure of speech". In: *Zellner, B.(1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition.(pp. 41-62). Chichester: John Wiley.* John Wiley, 1994, pp. 41–62.