NORTHWESTERN UNIVERSITY

Multilevel Contributions to Low Level Multisensory Integration Processes

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Psychology

By

L. Jacob Zweig

EVANSTON, ILLINOIS

June 2017

Abstract

Traditionally, research on perception and sensory systems has considered the senses as independent and modular functions that only converge after sufficient processing in unisensory areas. Recently, however, that view has been called into question with a number of demonstrations of multisensory interactions that may occur as early as primary cortex. In the following studies, I demonstrate evidence for early and remarkably content-specific multisensory interactions. First, I demonstrate that auditory influences on visual flicker detection which critically depend on stimulus frequency, eccentricity, and temporal correspondence. Next, I extend my investigation of early multisensory integration processes by exploring how visual processing of speech includes rapid and highly specific exchanges of information. Finally, I demonstrate a novel multivariate decoding technique using deep recurrent convolutional neural networks that allows further insight into the dynamic relationship between neurophysiological processes and perception. Taken together, this research represents a radical departure from the traditional view of multisensory perception by highlighting the strikingly early nature of multisensory interactions as well as novel techniques to further explore them.

Acknowledgements

I would like to thank my supervisors, Marcia Grabowecky and Satoru Suzuki for their time, support, and invaluable training. Their ability to balance wild creativity with rigorous scientific method is something that I hope to continue to emulate throughout my own career. I thank my committee member, Steve Franconeri, for his insights and for being an exceptional model of how to communicate science effectively.

To my labmates, Alexandra List, John Plass, Sasha Sherman, Julia Mossbridge, Emmanuel Guzman, Laura Ortega, Stacey Huntington, and Melisa Menceloglu, I thank you for being incredibly helpful in providing mentorship, ideas, amazing friendship, and sometimes even therapy. To my office-mate, David Brang, thank you for challenging me in many ways and igniting the creative spark which led me to pursue these research ideas and methods. I thank my cohort, Alex, Alissa, Amanda, Annie, Brock, Christian, Christina, JD, Kelly, Mia, and Vida, for making graduate school so much fun.

I thank my parents, Robin and Howard, and my sister, Sara, who have always pushed me to pursue greater challenges and provide unconditional love and support. Finally, to my wonderful wife, Angelica, without you I would not be the person I am today. Your dedication and curiosity is a constant source of inspiration, and your bad jokes a constant source of entertainment. I truly appreciate your unwaivering love and encouragement.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Sensory systems have traditionally been considered independent modular functions. Multisensory processing has therefore been thought of as a hierarchical process that occurs only after sufficient processing in unimodal areas. For example, Ghazanfar and Schroeder (2006) describe a large body of research demonstrating multisensory interactions in higher-order association cortex such as STS, IPS, and areas in the frontal lobe. Recent evidence, however, has challenged this view by demonstrating early multisensory interactions within cortical areas that have traditionally been considered unimodal.

Critically, before identifying areas that exhibit multisensory interactions, it is necessary to have a complete definition of what signifies such interactions. In several earlier studies, activity in unimodal cortex during multisensory tasks was often ignored or misattributed. Current studies define multisensory interactions as a modulation or elicitation of activity in one modality by a separate sensory modality. Kayser and Logothetis (2007) outline three basic principles of multisensory interactions: (1) interactions are subject to spatial constraints such that responses are greater when stimuli are in the same location across the modalities, (2) interactions are subject to temporal constraints such that responses are greater when stimuli occur in close temporal proximity, and (3) multisensory interactions abide by the principle of inverse effectiveness such that response

enhancement is greater when one modality provides little information alone. In the context of this framework, a variety of recent findings have suggested that multisensory interactions may happen early within unimodal cortices.

Behaviorally, multisensory interactions are often characterized by speeded reaction times. For example, Molholm et al. (2002) demonstrated in a reaction time task that responses to multisensory stimuli were faster than the responses to unimodal stimuli, and they exceeded the upper limit specified by Miller's race model (Miller, 1982). By exceeding the race model, multisensory interactions demonstrate non-linear effects such that response times are faster than the summation of unimodal responses. Multisensory interactions may also be characterized by perceptual changes, such as during the sound-induced-flash illusion, in which presenting two beeps in close temporal proximity to a flash can induce the perception of a second illusory flash (Mishra et al., 2010, 2007; Shams et al., 2000, 2002). Similarly, delivering sub threshold TMS to peripheral visual cortex with simultaneous auditory stimulation can induce phosphene perception when stimulation is both spatially and temporally congruent (Bolognini et al., 2010). While these results suggest strong multisensory influences at sites of early sensory processing, behavioral responses are limited in that they do not allow us to precisely understand the sites and timing of such multisensory effects.

To better understand multisensory interactions, techniques including neuroanatomical tracing, electrophysiology, and imaging can provide additional insight. Ultimately, it is necessary to use a combination of techniques to account for the limits of each individually. For example, while fMRI has excellent spatial resolution, it lacks the temporal resolution to precisely classify the time course of activity. And, while EEG provides excellent temporal resolution, it lacks detailed spatial resolution of fMRI. Electrocorticography (ECoG) provides both excellent spatial (3-5 mm) and temporal (1 ms) resolution, but is limited in that patients may present with deficits stemming from intractable epilepsy and electrodes are placed based on clinical needs and therefore may not lie within the immediate area of interest. Therefore, a thorough investigation of multisensory

interactions must incorporate information from multiple techniques and consider the strengths and limitations of each.

Mounting evidence supports the hypothesis that direct connectivity between auditory and visual cortices is thought to underlie many of these multisensory effects (e.g., Clavagnier et al., 2004; Falchier et al., 2002; Ghazanfar and Schroeder, 2006; Rockland and Ojima, 2003). Schroeder and Foxe (2005) describe connectivity between auditory and visual areas as early as A1 to the periphery of V1. Although these connections are sparse, the authors speculate they may underlie some multisensory interactions, particularly by modulating activity through excitatory or inhibitory feedback. Indeed, in the following studies, we provide additional evidence for early multisensory interactions that occur only in the periphery and may rely on these neural pathways.

Using fMRI and PET, several studies have demonstrated modulation of activity in unimodal cortex from multisensory stimulation. Noesselt et al. (2007) investigated the differences between auditory and visual cortices when streams of information were temporally correspondent. Using directed-information-transfer, a statistical technique to identify the flow of information with the brain, they observed increased information flow from STS to both A1 and V1 during temporal correspondence, leading them to speculate that the observed effects may have resulted from back propagation from STS to unimodal cortices. Martuzzi et al. (2007) attempted to further clarify the sites and timing of such multisensory interactions by analyzing the dynamics of the BOLD response. They demonstrated cross-sensory responsiveness to unimodal stimulation (e.g., primary auditory cortex was responsive to visual stimuli). In addition, they performed latency analysis to look at the time course of activation and discovered that multisensory stimulation led to reduced latency in unimodal responses, consistent with the previously discussed behavioral facilitation.

Recent evidence for a potential top-down influence on these early multisensory interactions was provided by a study in which auditory activation in visual cortex was observed during unimodal stimulation, however only in subjects who had previously been exposed to a bimodal condition (Zangenehpour and Zatorre, 2010). Using functional connectivity analyses, the authors demonstrate

correlation in activity between A1 and V1 suggesting that the observed activity may result from direct connectivity. The dependence on previous exposure suggests a top-down influence such that auditory stimulation may serve as preparatory information only when expected. Interestingly, several studies in the visual cognition literature have demonstrated a reduced influence of top-down effects on areas early in the visual ventral stream. The modulatory effects described by Zangenehpour and Zatorre (2010) may suggest multisensory features provide enhanced top-down feedback, allowing activity in earlier sensory cortices to be modulated.

EEG and MEG have also been used to investigate the timing of multisensory interactions, as they have the advantage of providing extremely fine temporal resolution. Molholm et al. (2002) found early audiovisual interactions corresponding to the timing of the C1 visual component, suggesting even initial stages of visual processing within unimodal cortex might be modulated by multisensory stimulation. Using MEG, Shams et al. (2005) demonstrated modulation of visual cortex at extremely short latencies during perception of the sound-induced-flash illusion. In a follow up study, Mishra et al. (2007) used EEG to further investigate the timing of modulatory effects in response to the sound-induced-flash illusion. They found that perception of the illusion was associated with early modulation of both auditory and visual cortex. Additionally, they reported changes in gamma power in visual cortex associated with perception of the illusion.

While traditional views of sensory processing held that multisensory interactions only occurred in higher-level association cortex, a number of recent findings have challenged that view. Across a variety of techniques, accumulating evidence points to interactions within unimodal cortex at even the earliest stages of processing. These interactions may occur based on feedback from higher levels or even direct connectivity between cortices. In the following studies, I provide additional evidence for early multisensory interactions with converging evidence pointing towards reliance upon direct neural pathways between auditory and visual corticies. Additionally, while existing evidence points to such early multisensory interactions as purely modulatory, using novel techniques for trial-by-

trial decoding of information, I provide evidence of highly specific exchanges of multisensory information within early sensory cortex.

In Chapter 2, I demonstrate auditory influences on visual flicker detection which critically depend on stimulus frequency, eccentricity, and temporal correspondence. The pattern of specificity demonstrated in this study, in tandem with relevant evidence from neuroanatomical and neurophysiological studies, suggests a likely role of direct neural connectivity which underlies these effects.

In Chapter 3, I extend my investigation of early multisensory integration processes by exploring how visual processing of speech includes rapid and highly specific exchanges of information. Using deep convolutional neural networks, I demonstrate that viewing speech movements and hearing speech generate similar dynamic activity in auditory cortex, suggesting real-time content-specific crossmodal interactions in early sensory processing.

In Chapter 4, I highlight a novel technique for decoding electrophysiologic information using recurrent convolutional neural networks. I provide evidence for robust end-to-end classification using this technique without any feature preprocessing, reducing the potential for bias in the decoding pipeline.

In summary, this investigation demonstrates the early nature of multisensory binding processes which are potentially reliant upon direct neural connectivity. These results represent a dramatic departure from the traditional modular view of multisensory processing by highlighting highly specific exchanges of information at early stages of cortical perceptual processing.

Chapter 2

Amplitude-modulated sounds selectively facilitate the detection of fast peripheral flicker with 90° phase sensitivity

2.1 Abstract

Auditory-visual interactions in the real world involve sustained dynamic stimuli. While many laboratory studies focused on single isolated auditory-visual events, some investigated dynamic crossmodal interactions using amplitude-modulated (AM) sounds and visual flicker. These studies revealed slow (<4 Hz) crossmodal binding mechanisms that mediate conscious tracking of crossmodal synchrony and synchrony-based crossmodal attention capture, as well as revealed crossmodal temporal mechanisms that influence perceived number and rate of visual flashes by strongly weighting the more reliable auditory temporal processing. Do dynamic sounds also directly influence the processing of visual dynamic signals? We demonstrate that AM sounds reduce flicker-detection thresholds in the periphery (not in the fovea), preferentially for a fast flicker rate (e.g., 12 Hz) above the temporal limit for consciously tracking auditory-visual synchrony (4 Hz)

but not for a slower rate (e.g., 3 Hz), with sensitivity to 90° (but not 180°) crossmodal phase-shift. This pattern of spatiotemporal specificity in combination with the relevant knowledge from neuroanatomy and neurophysiology suggest that AM sounds boost responses of frequency-doubling visual neurons to subtle flicker, potentially through direct neural connections from auditory cortex that target the peripherally-tuned neurons in visual cortex.

2.2 Introduction

Temporal coincidence of multisensory stimuli has been shown to provide perceptual enhancements. For example, a flash presented simultaneously with a beep appears brighter than a flash presented in isolation (e.g., Bolognini et al., 2010; Stein et al., 1996). Research investigating these multisensory interactions has primarily focused on single isolated events. However, in the real world we often encounter trains of multisensory events dispersed in time and space. For example, while enjoying a drink at a street café and casually looking at people engaged in conversations, you may hear rhythmic footsteps and notice a group of tourists entering your peripheral vision. It remains unclear how rhythmic auditory and visual stimuli interact in central and peripheral visual fields. Here we examined how hearing amplitude-modulated sounds influenced the detection of visual flicker.

Previous research suggests that the perception of temporal features is strongly influenced by auditory signals. For example, perceived intervals and durations of auditory-visual events are primarily determined by auditory signals (e.g., Burr et al., 2009; Ortega et al., 2014). Relevant to the current study, the perceived rate of visual flicker is influenced by the rate of concurrent amplitude-modulated sounds (e.g., Gebhard and Mowbray, 1959; Recanzone, 2003; Wada et al., 2003). Further, a rapid succession of two auditory pulses can generate the illusory perception of two visual flashes from a single flash — the sound-induced flash illusion (e.g., Shams et al., 2000, 2002). These crossmodal effects demonstrate that auditory rhythms can influence the perceived timing, rate, and number of visual flashes, likely reflecting a greater weighting of auditory information in

making temporal decisions due to the superior temporal resolution provided by the auditory system (e.g., Alais and Burr, 2004; Ernst and Bülthoff, 2004). However, these effects do not necessarily require that auditory rhythms influence the visual processing of flicker. In the current study, we investigated whether amplitude-modulated (AM) sounds facilitated the detection of synchronized visual flicker and whether they generated flicker perception of a non-flickered image.

Previous research suggests that the perception of temporal features is strongly influenced by auditory signals. For example, perceived intervals and durations of auditory-visual events are primarily determined by auditory signals (e.g., Burr et al., 2009; Ortega et al., 2014). Relevant to the current study, the perceived rate of visual flicker is influenced by the rate of concurrent amplitude-modulated sounds (e.g., Gebhard and Mowbray, 1959; Recanzone, 2003; Shipley, 1964; Wada et al., 2003). Further, a rapid succession of two auditory pulses can generate the illusory perception of two visual flashes from a single flash — the sound-induced flash illusion (e.g., Shams et al., 2000, 2002). These crossmodal effects demonstrate that auditory rhythms can influence the perceived timing, rate, and number of visual flashes, likely reflecting a greater weighting of auditory information in making temporal decisions due to the superior temporal resolution provided by the auditory system (e.g., Alais and Burr, 2004; Ernst and Bülthoff, 2004). However, these effects do not necessarily require that auditory rhythms influence the visual processing of flicker. In the current study, we investigated whether amplitude-modulated (AM) sounds facilitated the detection of synchronized visual flicker and whether they generated flicker perception of a static image.

We considered several factors that may influence the effects of AM sounds on flicker detection, visual eccentricity, flicker rate, and phase. In photopic vision, most aspects of visual perception, including pattern, motion, and flicker perception, are degraded in the periphery (relative to the fovea) at least when stimulus size is held constant (see Strasburger et al. (2011), for a review). Thus, based on the principle of "inverse effectiveness" of crossmodal interactions (e.g., Stein and Stanford, 2008), concurrent auditory information might be more effective in influencing peripheral

rather than foveal visual processing. Indeed, human behavioral research has generally demonstrated stronger auditory effects on the perception of peripherally presented visual stimuli (e.g., Noesselt et al., 2005; Regan and Spekreijse, 1977; Shams et al., 2002, 2001). Furthermore, a retrograde tracing study in monkeys has shown that crossmodal neural connections from auditory cortex primarily target the peripheral representations of visual cortex (e.g., Clavagnier et al., 2004; Falchier et al., 2002; Hall and Lomber, 2008). To test the possibility that AM sounds might selectively influence flicker detection in the periphery, we compared flicker detection between the fovea and a large retinal eccentricity (56°).

It is reasonable to expect that any facilitative effects of AM sounds on visual flicker detection would depend on phase alignment. If auditory AM signals were to facilitate the processing of visual flicker signals, they should boost visual responses in phase with the flicker. We note that both auditory and visual cortices include neural populations that respond primarily to either stimulus onsets or offsets, thus responding at the stimulus modulation rate — the frequency-following neurons — and those that respond well to both stimulus onsets and offsets, thus responding at twice the stimulus modulation rate — the frequency-doubling neurons (e.g., Benucci et al., 2007; Hubel and Wiesel, 1968; Kim et al., 2011; Qin et al., 2007; Recanzone, 2000). If crossmodal interactions between AM sounds and visual flicker are mediated by the frequency-following neurons, the interactions should be abolished when AM sounds and visual flicker are 180° out-of-phase. Alternatively, if the crossmodal interactions are mediated by the frequency-doubling neurons, the interactions should be unchanged whether AM sounds and visual flicker are in-phase or 180° out-of-phase because the responses of the frequency-doubling neurons would be synchronized in either case. The responses of the frequency-doubling neurons would be out-of-phase if AM sounds and visual flicker are 90° phase-shifted. Thus, if AM sounds influence visual flicker detection through the frequency-following neurons, a 180° and 90° phase shifts should both abolish the crossmodal effect, whereas if AM sounds influence visual flicker detection through the frequency-doubling neurons, a 180° phase shift should have no effect whereas a 90° phase shift should abolish

the crossmodal effect. We included a 180° phase shift in Experiment 1, and a 90° phase shift in Experiments 2 and 3.

Are phase-specific interactions limited by the observers' ability to perceptually judge crossmodal phase alignment? Using a variety of periodically changing auditory and visual stimuli, Fujisaki and Nishida (2005) have demonstrated that perceptual judgments of auditory-visual phase alignment have an upper limit of $4\ Hz$. Thus, if the perceptual limit of auditory-visual phase judgments is reflected in a temporal limit of phase-specific auditory-visual interactions, AM sounds should not facilitate visual flicker detection beyond $4\ Hz$. Alternatively, because the auditory system is sensitive to much higher modulation rates than the visual system (e.g., O'Connor et al., 2011; Van Hateren, 1993), AM sounds might strongly facilitate the detection of rapid visual flicker regardless of whether auditory-visual phase could be perceptually discriminated. We thus used three temporal rates, 3 Hz (for which auditory-visual phase discrimination is reliable), 6 Hz (just above the temporal limit of auditory-visual phase discrimination), and 12 Hz (well above the temporal limit of auditory-visual phase discrimination).

In summary, we investigated how AM sounds influenced visual flicker detection as a function of visual eccentricity, flicker rate, auditory-visual phase, and the perceptual judgment of auditory-visual alignment. Anatomical evidence of auditory-visual neural connectivity in mammals (e.g., Cappe and Barone, 2005; Clavagnier et al., 2004; Falchier et al., 2002; Rockland and Ojima, 2003) and the principle of inverse effectiveness predicted that the crossmodal effect should target peripheral vision. The potential involvement of the frequency-following neurons predicted that the crossmodal effect should be abolished by both a 90° and a 180° auditory-visual phase shifts, whereas the potential involvement of the frequency-doubling neurons predicted that the crossmodal effect should be abolished by a 90° (but not by a 180°) auditory-visual phase shift. Finally, whereas the potential involvement of the mechanisms that allow perceptual judgments of auditory-visual alignment predicted that the crossmodal effect would have the temporal limit of $4\ Hz$, the potential involvement of a crossmodal mechanism that takes advantage of the superior temporal resolution

of the auditory system to compensate for the relatively low temporal resolution of the visual system predicted that the crossmodal effect would be especially effective for the detection of fast flicker.

2.3 Experiment 1

2.3.1 Method

Participants

Twelve individuals from Northwestern University gave informed consent to participate in this experiment. Five were trained psychophysical observers and seven were undergraduate students who participated for partial course credit. They all had normal or corrected-to-normal vision and normal hearing, and were tested individually. The five psychophysically trained observers were tested with 3 Hz, 6 Hz, and 12 Hz auditory-visual modulations, but because the data from 3 Hz and 6 Hz modulations little differed, we only included 3 Hz and 12 Hz modulations for the seven undergraduate participants, and the presented analyses were conducted only for 3 Hz and 12 Hz modulations. The undergraduate participants were tested in two sessions (about a week apart) to increase the reliability of the data by obtaining two flicker-detection thresholds for each sound condition (see below). Two of those participants did not return for the second session and another participant used incorrect response keys throughout the first session, so that only one threshold value per condition was obtained from these three undergraduate participants.

Stimuli and Procedures

Participants were tested in a dimly lit room with their heads stabilized using a chin rest. Visual stimuli were presented with three 6 mm red LEDs (each subtending 1.15° of visual angle) mounted on a large black cardboard surface (103° horizontal by 81° vertical) positioned 30 cm from the participant. The LEDs were placed at the center (the foveal condition) as well as at 56° to

the left and right of the center (the peripheral condition). Auditory stimuli were square-wave amplitude-modulated tones (2500 Hz) presented through a piezoelectric speaker positioned directly below each LED (4° visual angle center-to-center). Presentations of auditory and visual stimuli were controlled by an Arduino Uno microcontroller providing sub-millisecond timing accuracy. The timing of auditory and visual signals was verified using a photoresistor circuit and digital oscilloscope to measure the onset of auditory and visual signals. A Macbook Pro computer running MATLAB with PsychToolbox extensions (Brainard, 1997; Pelli, 1997) was used to control the experiment and to record participants' responses.

Participants performed a two-interval forced choice task. On each trial, they were presented with two stimulus intervals, and decided which interval contained visual flicker and responded with a button press while maintaining central eye fixation. The flicker and no-flicker intervals were each 1.5-s long separated by a 1-s inter-stimulus interval (ISI). Flicker consisted of square-wave luminance modulation at 3 Hz or 12 Hz (6 Hz was also included for the five psychophysically trained observers).

There were four sound conditions (Figure 2.1). On an *AM-in-phase* trial, an AM sound in-phase with visual flicker was presented during the flicker interval, whereas an unmodulated control sound was presented during the no-flicker interval (Figure 2.1, top row). On an *AM-180°-phase-shifted* trial, an AM sound 180° phase-shifted from the visual flicker was presented during the flicker interval, whereas an unmodulated control sound was presented during the no-flicker interval (Figure 2.1, second row). To prevent participants from being biased toward choosing the AM-sound-present intervals as the flicker intervals irrespective of the actual perception of flicker, we included *AM-catch trials* that were equal in number to the AM-in-phase and AM-out-of-phase trials combined. On an AM-catch trial, an AM sound, either the one that would be in-phase with flicker (Figure 2.1, top row) or the one that would be 180° phase-shifted from flicker (Figure 2.1, second row), was presented during the no-flicker interval, whereas an unmodulated control sound was presented during the flicker interval (Figure 2.1, third row). Finally, on an *unmodulated* trial,

an unmodulated control sound was presented during both the flicker and no-flicker intervals (Figure 2.1, bottom row); this served as the baseline control condition. The four sound conditions were randomly intermixed across trials. Thus, the presence or absence of an AM sound was independent of whether a given interval contained flicker or no-flicker. Participants were informed of this fact and were instructed to ignore the sounds.

Each visual stimulus began with an LED turning on and ended with the LED turning off. During a no-flicker interval, the LED remained on for the entire duration of 1.5 s. During a flicker interval, the LED luminance was initially raised to the maximum value (as during a no-flicker interval) and then was alternated between the maximum and a lower value at a specified rate (3 Hz or 12 Hz; also 6 Hz for the five psychophysically trained observers). The maximum LED luminance, L_{max}, was 57.6 cd/m^2, while the lower luminance, L_{low}, was varied to determine flicker detection thresholds. The steady LED luminance, L_{steady}, for the no-flicker interval was scaled with the lower flicker luminance so that on each trial the flicker and no-flicker intervals appeared approximately equal in overall brightness, using the following function determined in a pilot study: $L_{steady} = L_0 + (L_{max} - L_0)/L_{max} * L_{low}$, where L_0 is the no-flicker luminance that appeared to be equal in overall brightness to the highest-amplitude flicker (between L_{max} and LED off) at rapid flicker rates (12, 24, and 36 Hz); note that it is difficult to perceptually compare the time-averaged brightness between flickered and steady stimuli when flicker rates are low.

Using an adaptive staircase procedure, QUEST, updated by Prins and Kingdom (2009), we varied L_{low} (while keeping the upper flicker luminance at L_{max}) to determine the depth of contrast modulation, $(L_{max} - L_{low})/L_{max}$ that yielded 75% accuracy in flicker detection. An experimental session consisted of 20 intermixed staircases, including 2 flicker rates (3 Hz and 12 Hz; 30 intermixed staircases including 3 Hz, 6 Hz, and 12 Hz for the five psychophysically trained observers), 2 retinal eccentricities (fovea and 56° periphery randomly on the left or right side), and 5 sound conditions (AM-in-phase trials, AM-180°-phase-shifted trials, AM-catch trials with the sound starting with the on phase and off phase, and unmodulated-sound trials). Each staircase

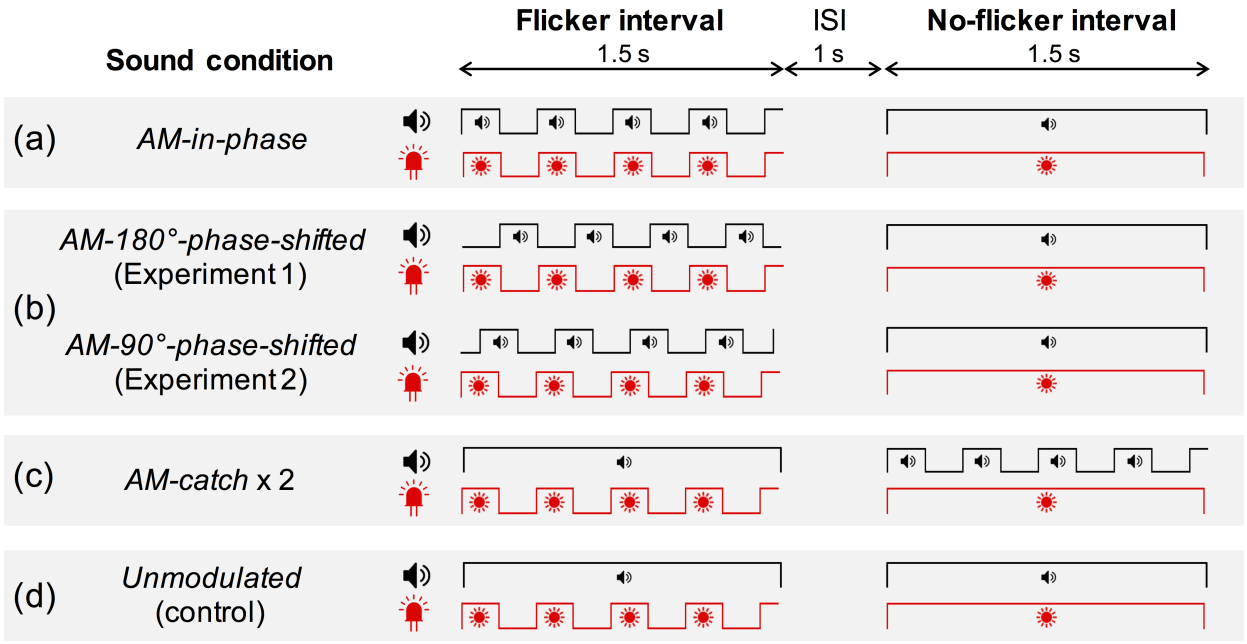

Fig. 2.1 The four sound conditions used in Experiments 1 and 2. (a) In the AM-in-phase condition, an amplitude-modulated sound in-phase with visual flicker was presented during the flicker intervals and an unmodulated sound was presented during the no-flicker intervals. (b) In the AM phase-shifted condition, the amplitude-modulated sound presented during the flicker intervals was either 180° (Experiment 1) or 90° (Experiment 2) phase-shifted relative to visual flicker. (c) In the AM-catch condition, an unmodulated sound was presented during the flicker intervals and an amplitude-modulated sound was presented during the no-flicker intervals. This condition was presented twice as frequently as the others so that visual flicker was presented with an amplitude-modulated sound with equal probability. (d) In the unmodulated-sound (control) condition, an unmodulated sound was presented during both the flicker and no-flicker intervals. All these sound conditions were presented at 3 Hz and 12 Hz in the fovea and periphery, randomly intermixed across trials.

included 25 trials for a total of 500 trials (750 trials for the five psychophysically trained observers). As noted above, four of the seven undergraduate participants underwent two sessions.

2.3.2 Analysis

In order to evaluate how sounds influenced the visual processing of flicker, it is desirable to use a measure that linearly reflected the underlying visual sensory activation. In particular, it is well known that neural contrast response functions tend to be logarithmically (rather than linearly) related to image contrast at both single-cell and neural-population levels (e.g., Albrecht and Hamilton, 1982; Campbell and Kulikowski, 1972). It is thus likely that any sound-induced changes in flicker detection thresholds (measured as changes in just-visible contrast-modulation depth) may not linearly reflect the sound-induced changes in the underlying visual sensory activation. Signal-to-noise ratio provides an appropriate measure because it is reasonable to assume that the impact of a change in sensory activation upon downstream neural processing depends on the change relative to the magnitude of the relevant noise. We thus normalized our flicker detection thresholds to the scale of the estimated noise magnitude.

In our experiments, flicker detection thresholds substantially varied depending on visual eccentricity, flicker rate, and sound conditions. Thus, the relationship between noise magnitude and threshold level can be estimated by examining the standard deviation as a function of the mean threshold. If flicker detection thresholds linearly reflected the signal-to-noise ratio of the underlying visual sensory response to flicker, the standard deviation should be constant across different mean threshold values. On the contrary, as shown in the left panels in Figure 2.2, the standard deviation systematically increased as a function of the mean threshold in all three experiments. This indicates that visual sensory response to flicker (relative to the relevant noise) was compressively related to our measure of flicker-detection threshold. We thus normalized each threshold value by scaling it to the standard deviation as follows. We first quantified the relationship between the standard deviation

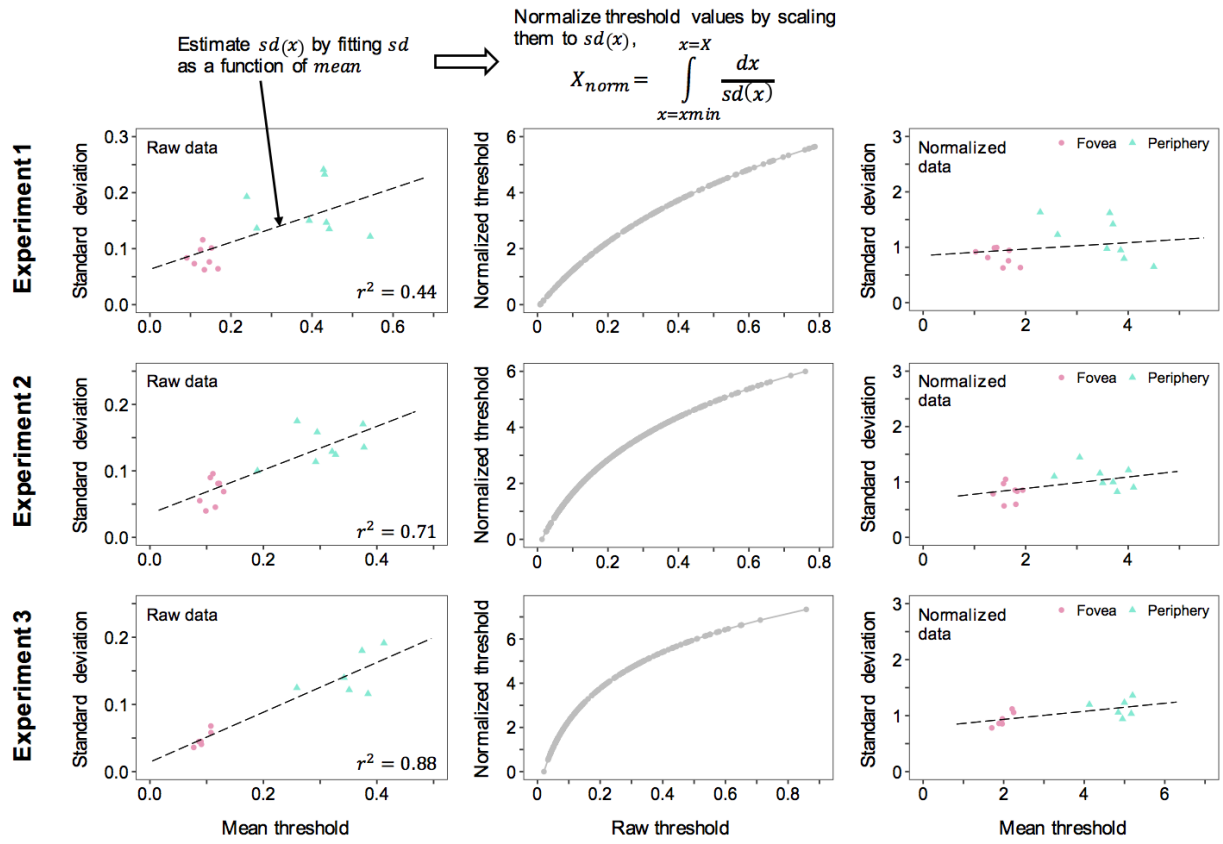

Fig. 2.2 Normalizing flicker-detection thresholds to reflect the underlying signal-to-noise ratios. As seen in the left panels, larger mean thresholds are associated with larger standard deviations (i.e., larger noise magnitudes) in the current data, indicating that the underlying signal-to-noise ration, a desired measure of visual sensory activation, scales compressively with thresholds. We thus normalized the threshold values by scaling them to the integral of the reciprocal of the standard deviation as a function of the mean threshold, $sd(x)$, captured by the linear fits show in in the left panels, so that a unit change in the normalized threshold value represented a change in the underlying visual sensory response to flicker by the relevant noise magnitude. The normalization curves are shown in the middle panels, demonstrating log-like compressive scaling. The effectiveness of the normalization can be seen in the right panels, where all mean threshold values of the normalized data are associated with the standard deviation of approximately unity. All analyses were performed on the normalized threshold values.

and mean threshold for each experiment via linear fit (as no particular non-linear relationships are evident as seen in the left panels in Figure 2.2). Each fitted line provided $sd(x)$, the standard deviation as a function of threshold (x). We then normalized each threshold value, X_i (for each condition from each participant), by scaling it to the integral of the reciprocal of x, that is,

$$X_{i,norm} = \int_{x=xmin}^{x=X_i} \frac{dx}{sd(x)} \tag{2.1}$$

where *xmin* is the minimum threshold obtained in the specific experiment. This scaling normalizes flicker detection thresholds to noise levels such that a unit change in normalized threshold represents a change in sensory response to flicker by the magnitude of the relevant noise. As seen in the middle panels in Figure 2.2, the normalization curves (i.e., normalized thresholds as a function of raw thresholds) show log-like compressive scaling. If the scaling were performed with the complete knowledge of $sd(x)$, the standard deviation of the normalized thresholds would be unity regardless of the threshold level. As shown in the right panels in Figure 2.2, the actual normalization has yielded reasonably good results with most standard deviation values clustered around unity, successfully scaling the threshold values to the relevant noise. All analyses below were performed on the normalized threshold values.

2.3.3 Results

The normalized thresholds for the detection of foveal and peripheral flicker are shown in (Figures 2.3a and 2.3b), respectively. As expected, flicker detection thresholds were overall lower in the fovea than in the periphery ($F_{(1,11)} = 200.94, p < .0001$). Importantly, the significant interaction between sound condition and retinal eccentricity ($F_{(3,33)} = 6.95, p_{GG} < .007, p_{HF} < .004$) indicates that auditory effects on visual flicker detection were different between fovea and periphery.

In the fovea (Figure 2.3a), although there was a significant main effect of flicker rate ($F_{(1,11)} = 7.05, p < .03$), suggesting overall greater sensitivity to 12 Hz than 3Hz flicker, there was no

significant main effect of sound condition ($F_{(3,33)} = .045, p > .98$). Although there was a significant interaction between sound condition and flicker rate ($F_{(3,33)} = 3.67, p < .03$), none of the AM sound conditions (AM-in-phase, AM-180°-phase-shifted, or AM-catch) significantly deviated from the unmodulated-sound (control) condition for the detection of either 3 Hz or 12 Hz flicker (t's < 1.44).

In the periphery, there was also a significant main effect of flicker rate ($F_{(1,11)} = 8.62, p < .02$), suggesting overall greater sensitivity to 12 Hz than 3 Hz flicker. Unlike in the fovea, the main effect of sound condition ($F_{(3,33)} = 9.15, p < .0002$) and its interaction with flicker rate ($F_{(3,33)} = 7.84, p < .0005$) were both significant, indicating that sounds substantially influenced flicker detection in a rate dependent manner in the periphery.

It is apparent from Figure 2.3b that the sound conditions little influenced the detection of 3 Hz flicker; none of the AM sound conditions significantly changed 3 Hz flicker detection thresholds relative to the unmodulated-sound control condition (t's < 0.71). In contrast, for the detection of 12 Hz flicker, the AM sound conditions had substantial impact. Both the in-phase AM sounds ($t_{(11)} = 4.69, p < .0007$) and 180°-phase-shifted AM sounds ($t_{(11)} = 3.49, p < .006$) significantly lowered 12 Hz flicker detection thresholds relative to the unmodulated (control) sounds, with no significant difference ($t_{(11)} = .83, p > .42$) between the in-phase and 180°-phase-shifted conditions. Thus, AM sounds presented during the flicker intervals selectively facilitated the detection of the faster (12 Hz) flicker, whether the crossmodal phase was aligned or 180° shifted. This indifference to a 180° phase-shift is consistent with the interpretation that the crossmodal effect is mediated by the frequency-doubling populations of auditory and visual neurons (see the Introduction).

Finally, presenting AM sounds during the no-flicker intervals (AM-catch condition) significantly elevated thresholds for the detection of 12 Hz flicker relative to the unmodulated-sound (control) condition ($t_{(11)} = 4.15, p < .002$). This result is consistent with the interpretation that 12Hz (but not 3 Hz) AM sounds induced illusory flicker during the no-flicker intervals that competed with the actual flicker presented during the flicker intervals.

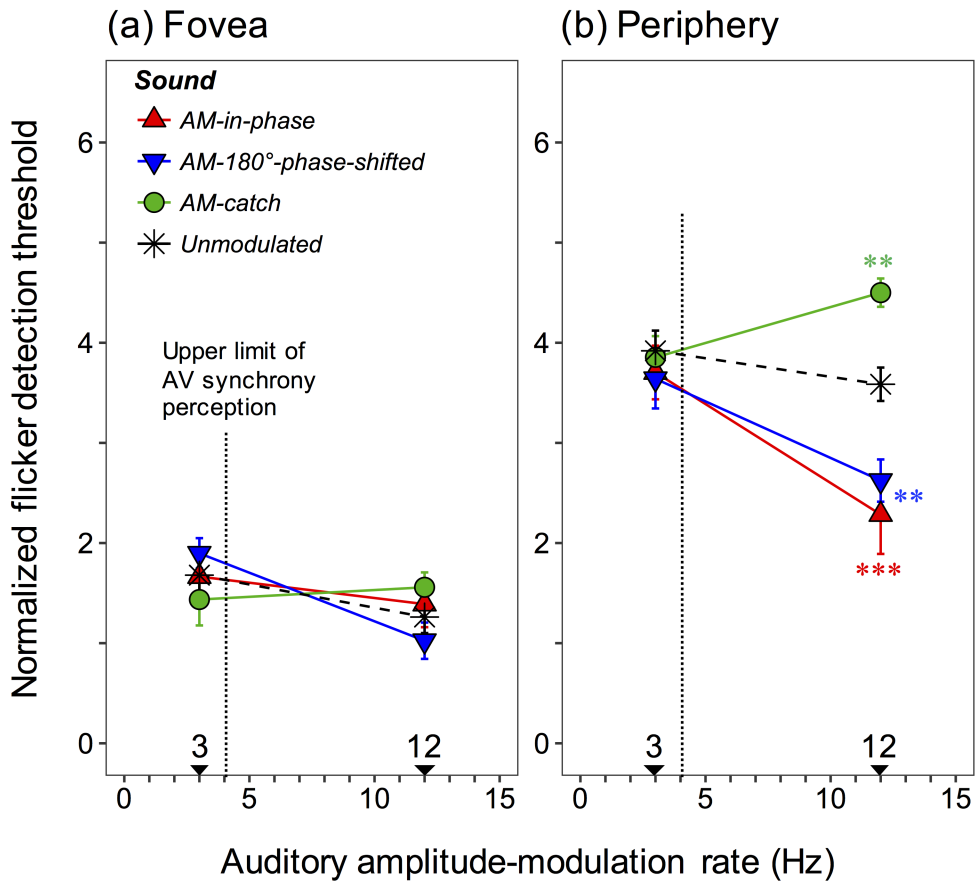

Fig. 2.3 Results of Experiment 1. Flicker detection thresholds (contrast modulation yielding 75% accuracy) normalized to reflect the signal-to-noise ratio (see Figure 2.2), are plotted for the detection of 3 *Hz* and 12 *Hz* flicker in the fovea (a) and periphery (b), under four sound conditions: AM-in-phase (in-phase AM sound presented during flicker intervals), AM 180°-phase-shifted (180°-phase-shifted AM sound presented during flicker intervals), AM-catch (AM sound presented during no-flicker intervals), and Unmodulated (steady sound presented during both flicker and no-flicker intervals). The upper limit of AV synchrony perception is based on Fujisaki & Nishida, 2007. The error bars represent ± 1 SEM adjusted for within-participant comparisons (Morey, 2008)

There is a concern that all these effects might reflect participants having had a general bias to choose the intervals during which AM sounds were presented as the flickered intervals over the intervals during which unmodulated sounds were presented. This response bias interpretation seems untenable for the following reasons. First, participants had no reason to choose the AM-sound intervals as flicker intervals because unmodulated sounds were presented in both intervals on the control trials and AM sounds were presented during the flicker and no-flicker intervals with equal probability across the remaining trials. Participants were also instructed to ignore the sounds to focus on visual flicker detection. Second, the AM-catch condition had no effect in the fovea whether 3 Hz or 12 Hz; nor did it influence the detection of 3 Hz flicker in the periphery; it selectively elevated thresholds for the detection of 12 Hz flicker in the periphery. We could think of no reason why participants would exercise response bias selectively in this specific case especially as the foveal and peripheral locations as well as 3 Hz and 12 Hz modulations were randomly intermixed across trials. It might be reasonable to speculate that participants might rely on response bias in difficult conditions; however, flicker detection thresholds were actually lower for 12 Hz flicker than for 3 Hz flicker in the unmodulated-sound control condition in both fovea and periphery (Fig 2.3). Nevertheless, we addressed the response bias concern in the following experiments.

Overall, the results suggest that AM sounds selectively influence the detection of fast (12 Hz) peripheral flicker well above the limit of perceptual judgments of auditory-visual synchronization, but do not influence the detection of foveal flicker (fast or slow) or slow (3 Hz) peripheral flicker within the limit of perceptual judgments of auditory-visual synchronization. The threshold-lowering effect of the AM-in-phase condition is consistent with the interpretation that auditory amplitude-modulation signals augment visual flicker signals, whereas the threshold-elevating effect of the AM-catch condition is consistent with the interpretation that auditory amplitude-modulation signals induce illusory visual flicker, with both crossmodal effects selectively influencing the processing of fast (e.g., 12 Hz) flicker in the periphery. The insensitivity of the threshold-lowering effect to a 180° auditory-visual phase-shift suggests that the underlying crossmodal interactions are primarily

mediated by the frequency-doubling populations of auditory and visual neurons. Nevertheless, the results do not completely rule out the alternative interpretation that participants somehow selectively committed response bias only when fast flicker was presented in the periphery.

2.4 Experiment 2

One goal of this experiment was to replicate the finding from Experiment 1 that AM sounds influenced flicker detection selectively for a fast rate of 12 Hz (not 3 Hz) in the periphery (not fovea). A second goal was to investigate the potential involvement of the frequency-doubling populations of auditory and visual neurons suggested by the results that the threshold-lowering effect of AM sounds for the detection of 12 Hz flicker in the periphery was relatively insensitive to a 180° crossmodal phase-shift. A 90° phase-shift between an AM sound and visual flicker should abolish the crossmodal effect because responses of the auditory and visual frequency-doubling neurons would be out of phase. To test this prediction, we replaced the AM-180°-phase-shifted condition with the AM-90°-phase-shifted condition. Note that a demonstration of 90° phase sensitivity would also provide direct evidence against the response bias account discussed above.

2.4.1 Method

Participants

Twenty-seven Northwestern University undergraduate students gave informed consent to participate for partial course credit. They had normal or corrected- to-normal vision and normal hearing, and were tested individually in a dimly lit room.

Stimuli and Procedures

The stimuli and procedures were the same as in Experiment 1, except that all participants were tested in only one session and the AM-180°-phase-shifted condition was replaced with the AM-90°-phase-shifted condition (Figure 2.3b).

2.4.2 Results

As in Experiment 1, flicker detection thresholds were overall lower in the fovea than in the periphery ($F_{(1,26)} = 147.29, p < .0001$), and the AM sounds selectively influenced flicker detection in the periphery, supported by the significant interaction between sound condition and retinal eccentricity ($F_{(3,78)} = 7.46, p < .0002$).

In the fovea (Figure 2.4a), there was a main effect of flicker rate ($F_{(1,26)} = 15.07, p < .0007$), suggesting overall greater sensitivity to 12 Hz than 3 Hz flicker, but there were no significant effects of sound condition either as a main effect ($F_{(3,78)} = 1.57, p > .20$) or as its interaction with flicker rate ($F_{(3,78)} = 1.61, p > .19$).

In the periphery (Figure 2.4b), there was also a main effect of flicker rate ($F_{(1,26)} = 7.40, p < .02$), suggesting overall greater sensitivity to 12 Hz than 3 Hz flicker. Importantly, the main effect of sound condition ($F_{(3,78)} = 13.01, p_{GG} < .0001, p_{HF} < .0001$) and its interaction with flicker rate ($F_{(3,78)} = 3.47, p < .03$) were both significant, indicating that sounds substantially influenced flicker detection in a rate dependent manner in the periphery. As in Experiment 1, the AM sounds selectively influenced the detection of 12 Hz flicker. None of the AM-sound conditions were significantly different from the unmodulated-sound control condition for the detection of 3Hz flicker (t's < 1.55). For the detection of 12Hz flicker, the AM-in-phase condition significantly lowered ($t_{(26)} = 3.54, p < .002$) and the AM-catch condition significantly elevated ($t_{(26)} = 3.11, p < .005$) thresholds as in Experiment 1. The critical AM-90°-phase-shifted condition did not significantly lower 12 Hz flicker-detection threshold relative to the unmodulated control ($t_{(26)} = 1.44, p > .16$)

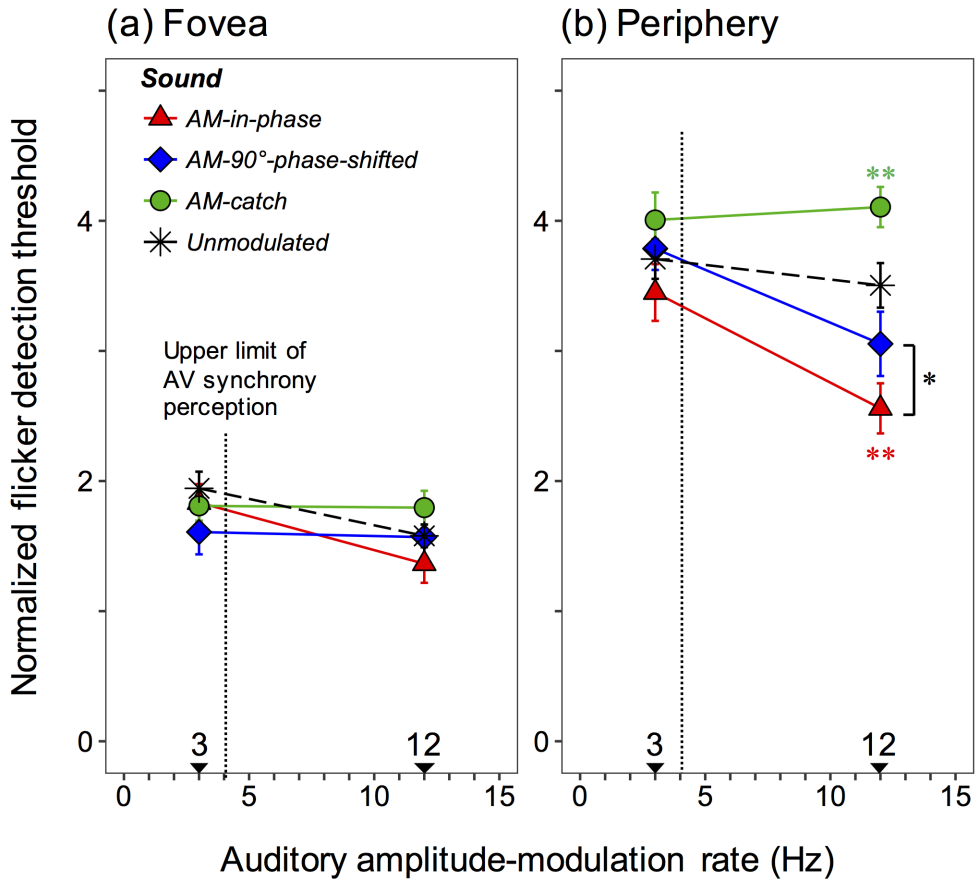

Fig. 2.4 Results of Experiment 2. Flicker detection thresholds (contrast modulation yielding 75% accuracy) normalized to reflect the signal-to-noise ratio, are plotted for the detection of 3 Hz and 12 Hz flicker in the fovea (a) and periphery (b), under four sound conditions: AM-in-phase (in-phase AM sound presented during flicker intervals), AM 90°-phase-shifted (90°-phase-shifted AM sound presented during flicker intervals), AM-catch (AM sound presented during no-flicker intervals), and Unmodulated (steady sound presented during both flicker and no-flicker intervals). The upper limit of AV synchrony perception is based on Fujisaki & Nishida, 2007. The error bars represent ± 1 SEM adjusted for within-participant comparisons (Morey, 2008)

while the AM-in-phase condition significantly lowered 12 Hz flicker-detection threshold relative to the AM-90°-phase-shifted condition ($t_{(26)} = 2.22, p < .04$), indicative of 90° phase sensitivity of the crossmodal effect. Overall we replicated Experiment 1 in that AM sounds selectively influenced the detection of fast (12 Hz, not 3 Hz) flicker in the periphery, with AM sounds presented in-phase with visual flicker during the flicker intervals lowering detection thresholds (relative to the unmodulated control sound) while the same AM sounds presented during the no-flicker intervals elevating detection thresholds. Crucially, 90° phase-shifting of AM sounds eliminated the facilitative effect, consistent with the mediation by the frequency-doubling populations of auditory and visual neurons. The 90° phase sensitivity has also provided direct evidence against the possibility that the facilitative effect of an in-phase AM sound might be explained by response bias.

2.5 Experiment 3

In Experiments 1 and 2, we presented AM sounds either during the flicker intervals or during the no-flicker intervals. We used this design because we hypothesized that an AM sound might augment visual flicker signals when it is synchronized with flicker, but it might also induce illusory visual flicker when a steady visual stimulus is viewed. The results from Experiments 1 and 2 are consistent with the possibility that for the processing of fast (12 Hz) flicker in the periphery, AM sounds augment visual flicker signals as well as induce illusory visual flicker. The 90° phase sensitivity of the threshold-lowering effect demonstrated in Experiment 2 ruled out a response-bias interpretation at least for the crossmodal augmentation effect. The goal of this experiment was to replicate the crucial 90° phase sensitivity, to further rule out a response-bias interpretation, and to evaluate the relative strength of the crossmodal flicker augmentation effect and the illusory flicker induction effect.

We presented an identical AM sound during both the flicker and no-flicker intervals on a given trial, so that response bias could not play a role. In the AM-in-phase condition, an amplitude-

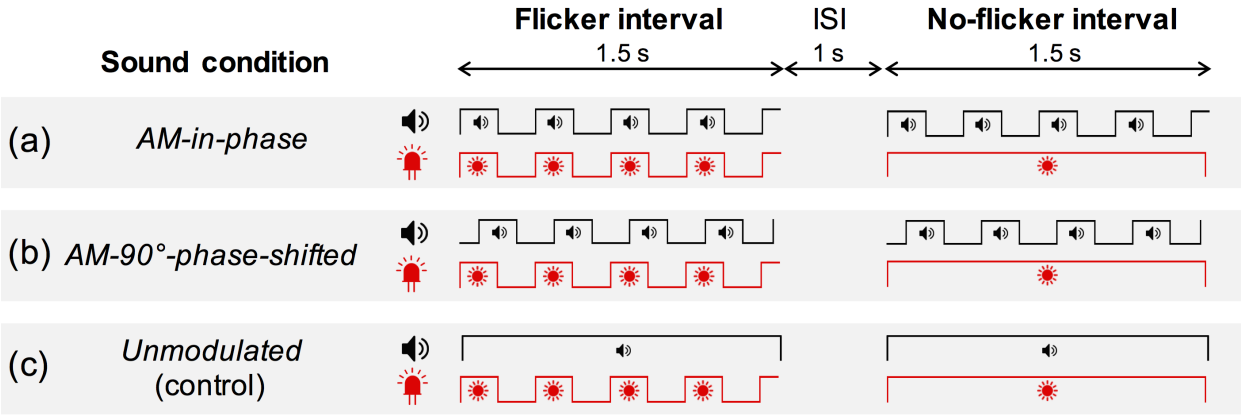

Fig. 2.5 Sound Conditions used in Experiment 3. In this experiment, any amplitude-modulated sound was presented during both the flicker and no-flicker intervals (a) In the AM-in-phase condition, an amplitude-modulated sound in-phase with visual flicker was presented during the flicker intervals and the same amplitude-modulated sound was also presented during the no-flicker intervals. (b) In the AM-90°-phase-shifted condition, an amplitude-modulated sound 90° phase-shifted relative to visual flicker was presented during the flicker intervals and the same amplitude-modulated sound was also presented during the no-flicker intervals. (c) In the unmodulated-sound (control) condition, an unmodulated sound was presented during both the flicker and no-flicker intervals. All these sound conditions were presented at 3 *Hz* and 12 *Hz* in the fovea and periphery, randomly intermixed across trials.

modulated sound in-phase with visual flicker was presented during the flicker intervals (as in Experiments 1 and 2), and the same amplitude-modulated sound was also presented during the no-flicker intervals (Figure 2.5a). In the AM-90°-phase-sifted condition, an amplitude-modulated sound 90° phase-shifted relative to visual flicker was presented during the flicker intervals, and the same amplitude-modulated sound was presented during the no-flicker intervals (Figure 2.5b). The unmodulated-sound control condition was the same as in Experiments 1 and 2, where an identical unmodulated sound was presented during both the flicker and no-flicker intervals (Figure 2.5c).

In the AM-in-phase condition, an AM sound presented during the flicker intervals would augment the synchronized flicker signals whereas the same AM sound presented during the no-flicker intervals might induce illusory flicker that would compete with the flicker signals presented during the flicker intervals. Thus, only if the flicker augmentation effect during the flicker intervals was reliably larger than any flicker induction effect during the no-flicker intervals, the AM-in-phase

condition would lower flicker detection thresholds relative to the unmodulated-sound control condition. In the AM-90°-phase-shifted condition, a phase-misaligned AM sound presented during the flicker intervals would not substantially augment the flicker signals whereas the same AM sound presented during the no-flicker intervals might induce illusory flicker. If the induced illusory flicker during the no-flicker trials was reliably larger than any small flicker augmentation effect produced by a 90°-phase-shifted AM sound during the flicker intervals, the AM-90°-phase-shifted condition would elevate flicker detection thresholds relative to the unmodulated-sound control condition. Thus, a reliable threshold reduction in the AM-in-phase condition would provide evidence for the augmentation of flicker signal by a synchronized AM sound over and above any effect of illusory flicker induction by an AM sound, whereas a reliable threshold elevation in the AM-90°-phase-shifted condition would provide evidence for the induction of illusory flicker by an AM sound over and above any small augmentation of flicker signals by a phase-misaligned AM sound.

2.5.1 Method

Participants

Seventeen Northwestern University undergraduate students gave informed consent to participate for partial course credit. They had normal or corrected- to-normal vision and normal hearing, and were tested individually in a dimly lit room.

Stimuli and Procedures

The stimuli and procedures were the same as in Experiment 2, except that an identical amplitude-modulated sound was presented during both the flicker and no-flicker intervals in the AM-in-phase and AM-90°-phase-shifted conditions (Figure 5), and the AM-catch condition was removed.

2.5.2 Results

Consistent with the results from Experiments 1 and 2, flicker detection thresholds were overall lower in the fovea than in the periphery ($F_{(1,16)} = 254.00, p < .0001$). Partially due to the fact that the AM-in-phase condition was virtually identical to the unmodulated-sound control condition in the fovea (Figure 2.6a) and the AM-90°-phase-shifted condition was virtually identical to the unmodulated-sound control condition in both the periphery (Figure 2.6b), the interaction between sound condition and retinal eccentricity was not statistically significant ($F_{(2,32)} = 5.48, p < .07$). Nevertheless, the results have provided clear evidence regarding the hypotheses we considered.

In the fovea, there were no significant main effects of sound condition or retinal eccentricity, or any interaction between them (F's < 2.87) (Figure 2.6a). Interestingly, the threshold for detecting 3 Hz flicker was significantly lower in the AM-90°-phase-shifted condition relative to both the unmodulated-sound control condition ($t_{(16)} = 2.80, p < .02$) and the AM-in-phase condition ($t_{(16)} = 3.15, p < .007$). A similar trend is seen in Figure 2.4a (left side) from Experiment 2. We do not have an reasonable explanation as to why, specifically for the detection of 3 Hz flicker in the fovea, a 90°-phase-shifted AM sound would be beneficial while a synchronized AM sound would have no effect. For the detection of 12 Hz flicker in the fovea, AM sounds produced no significant effects whether they were in-phase or 90°-phase-shifted relative to flicker (t's < 0.39). Overall, the three experiments have consistently demonstrated that AM sounds make minor to little impact on flicker detection in the fovea.

In the periphery, AM sounds selectively influenced the detection of 12 Hz flicker as indicated by the significant interaction between sound condition and flicker rate ($F_{(2,32)} = 5.03, p < .02$). Neither types of AM sounds influenced the detection of 3 Hz flicker relative to the unmodulated control sound (t's < 0.89; Figure 2.6b, left side). In contrast, the in-phase AM sound significantly lowered the detection threshold for 12 Hz flicker relative to both the unmodulated control sound ($t_{(16)} = 3.98, p < .002$) and the 90°-phase-shifted AM sound ($t_{(16)} = 2.98, p < .009$) (Figure 2.6b,

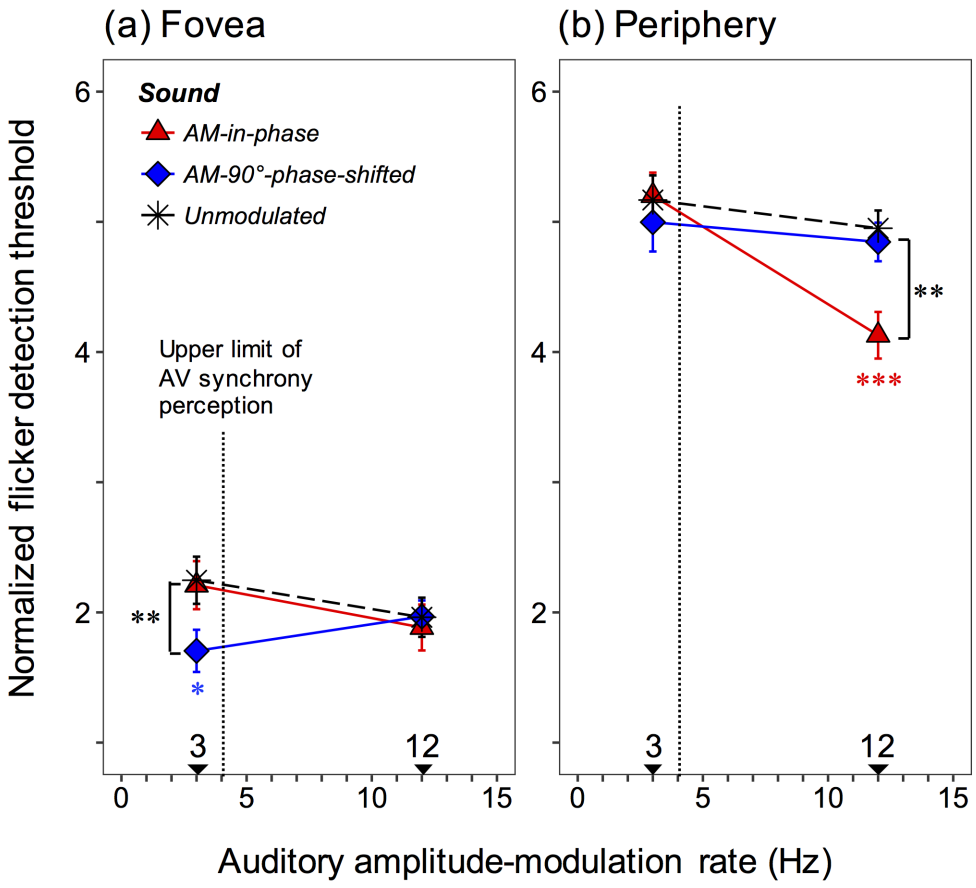

Fig. 2.6 Results of Experiment 3. Flicker detection thresholds (contrast modulation yielding 75% accuracy) normalized to reflect the signal-to-noise ratio, are plotted for the detection of 3 Hz and 12 Hz flicker in the fovea (a) and periphery (b), under three sound conditions: AM-in-phase (in-phase AM sound presented during flicker intervals and the same AM sound presented during no-flicker intervals), AM 90°-phase-shifted (90°-phase-shifted AM sound presented during flicker intervals and the same AM sound presented during no-flicker intervals), and Unmodulated (steady sound presented during both flicker and no-flicker intervals). The upper limit of AV synchrony perception is based on Fujisaki & Nishida, 2007. The error bars represent ± 1 SEM adjusted for within-participant comparisons (Morey, 2008)

right side), while the 90°-phase-shifted AM sound produced little effect relative to the unmodulated control sound ($t_{(16)} = 0.57, p > .57$). We have thus replicated the results from Experiment 2 that AM sounds selectively facilitate the detection of 12 Hz (not 3 Hz) flicker in the periphery with 90° phase sensitivity.

Because AM sounds were presented during both the flicker and no-flicker intervals in this experiment, the result has conclusively demonstrated that an in-phase AM sound augments 12 Hz visual flicker signals in the periphery over and above any illusory flicker that might have been induced during the no-flicker intervals. However, the result did not provide conclusive evidence regarding the induction of illusory flicker. Because the threshold for the detection of 12 Hz flicker was not elevated in the 90°-phase-shifted condition relative to the unmodulated-sound control condition, it is unclear whether the detrimental effect of illusory flicker induction during the no-flicker intervals was offset by a small augmentation of flicker signals by the 90°-phase-shifted AM sound during the flicker intervals, or there was no reliable induction of illusory flicker.

2.6 Discussion

While previous auditory-visual research primarily focused on single crossmodal events in isolation, we investigated how dynamic auditory and visual stimuli interacted. We asked several specific questions. Do amplitude-modulated (AM) sounds improve visual flicker sensitivity, generate illusory visual flicker, or both? Do any of these dynamic crossmodal interactions differ between the fovea and periphery? How do dynamic auditory-visual interactions depend on modulation rate, crossmodal temporal phase, and the ability to perceptually judge auditory-visual temporal alignment?

All three experiments consistently demonstrated that AM sounds selectively influenced flicker detection in the periphery. This result is consistent with the inverse-effectiveness principle of crossmodal interactions (e.g., Stein and Stanford, 2008) because photopic vision is generally

less precise in the periphery relative to the fovea (see Strasburger et al. (2011), for a review). The peripheral specificity of these results may also suggest that dynamic crossmodal interactions are mediated by neural connections from auditory cortex that primarily target the peripheral representations within visual cortex, as has been demonstrated in non-human mammals (e.g., Cappe and Barone, 2005; Clavagnier et al., 2004; Falchier et al., 2002; Hall and Lomber, 2008; Rockland and Ojima, 2003).

AM sounds improved peripheral detection of 12Hz but not 3Hz flicker across all three experiments. Thus, AM sounds improved flicker detection selectively for rates well above the temporal limit of consciously tracking auditory-visual synchrony (4 Hz; Fujisaki and Nishida (2005)). This suggests that there are at least two separate mechanisms through which AM sounds influence visual flicker perception, slow crossmodal-binding mechanisms that support conscious tracking of auditory-visual synchrony, and fast mechanisms that facilitate the detection of subtle flicker. When multiple items are flickering at similar rates but in different phases from one another, none of them would stand out. However, when an AM sound (e.g., auditory pulses) is presented in synchrony with one of the flickering items, that item would stand out and attract attention (e.g., Van der Burg et al. (2008)). This synchrony-based crossmodal attention capture is limited to low flicker rates (1-2 Hz), and is absent when flicker rates are higher than the temporal limit (4 Hz) of consciously tracking auditory-visual synchrony (Fujisaki et al., 2006; Fujisaki and Nishida, 2005; Keetels and Vroomen, 2012). These results suggest that the relatively slow (< 4 Hz) crossmodal-binding mechanisms mediate conscious tracking of auditory-visual synchrony and synchrony-based crossmodal attention capture. In contrast, we have demonstrated that AM sounds selectively increase peripheral sensitivity to 12 Hz (but not 3 Hz) flicker. At 12 Hz, the 90° phase sensitivity of this effect demonstrated in Experiments 2 and 3 (Figures 2.4b and 2.6b) corresponds to crossmodal temporal precision of 20 ms, suggesting that AM sounds facilitate the detection of dynamic visual signals through fast auditory input to visual cortex, potentially through direct neural connections from auditory cortex to the peripheral representations of visual cortex (see above). The fact that

this fast dynamic interaction does not enable conscious tracking of auditory-visual synchrony is consistent with the fact that neural processing within primary visual cortex does not necessarily generate awareness (e.g., Blake and Fox, 1974; Blake et al., 2006; Leopold and Logothetis, 1999; Sweeny et al., 2011).

The characteristics of auditory-visual phase dependency provided an additional insight into underlying neural interactions. Both auditory and visual cortices include subpopulations of neurons that primarily respond to either stimulus onsets or offsets—the frequency-following neurons—and those that respond to both stimulus onsets and offsets—the frequency-doubling neurons (e.g., Benucci et al., 2007; Hubel and Wiesel, 1968; Kim et al., 2011; Qin et al., 2007; Recanzone, 2003). The fact that the in-phase AM sounds and 180°-phase-shifted AM sounds both equivalently facilitated flicker detection (Experiment 1), but the 90°-phase-shifted AM sounds did not (Experiments 2 and 3), suggests that the dynamic crossmodal interactions are mediated by the input from auditory cortical frequency-doubling neurons to visual cortical frequency-doubling neurons (or involving frequency-doubling neurons in at least one modality) because frequency-doubling responses would have been in-phase whether AM sounds were in- with or 180° phase-shifted from visual flicker but out-of-phase when AM sounds were 90° phase-shifted from visual flicker; frequency-following responses would have been out-of-phase when AM sounds were 180° phase-shifted from visual flicker.

Turning to the question of whether AM sounds induced illusory flicker on static visual stimuli, 12 Hz AM sounds presented during the no-flicker intervals in the AM-catch condition (Experiments 1 and 2) elevated flicker detection thresholds, but 3 Hz AM sounds did not. A plausible interpretation is that AM sounds crossmodally induced illusory flicker when the modulation rate was sufficiently fast (e.g., 12 Hz). Nevertheless, we cannot conclusively rule out an alternative interpretation that the threshold elevation effects in the AM-catch condition obtained in Experiments 1 and 2 were due to a response bias (to choose AM-sound present intervals as flicker intervals), though it seems unlikely that response bias somehow occurred only when 12 Hz (but not 3 Hz) flicker was

presented in the periphery (not fovea) despite the fact that all trial types were randomly intermixed, AM sounds were presented during the flicker and no-flicker intervals with equal probability, and participants were instructed to ignore sounds. Unfortunately, although Experiment 3 that excluded response bias by design provided strong evidence for 90° phase-sensitive crossmodal augmentation of visual flicker signals, it did not provide positive evidence for the induction of illusory flicker.

Finally, the current results may have relevance to a phenomenon known as the sound-induced flash illusion (SIFI) in which a single visual flash appears to be two flashes when it is accompanied by a rapid succession of two auditory pulses (Shams et al., 2000, 2001). Consistent with the current result, the SIFI was shown to be minimal when the stimulus onset asynchrony between the sequential auditory clicks was longer than 100 ms, that is, when the rate of the auditory clicks was slower than 10Hz (Apthorp et al., 2013). Does the SIFI occur as a result of auditory-induced visual flicker? This possibility was suggested by Wilson (1987). Alternatively, the SIFI may be driven by auditory influences on the perceived number of visual flashes owing to the greater temporal resolution of auditory processing (Apthorp et al., 2013). Consistent with the latter possibility, SIFIs have been reliability demonstrated with high-contrast stimuli (e.g., Apthorp et al., 2013; Kumpik et al., 2014), whereas our results suggest that a 12 Hz AM sound would have induced (if any) rather weak visual flicker equivalent to only 13% contrast modulation (the threshold elevation in the AM-catch condition averaged across Experiments 1 and 2). Further, although a contribution of sound-induced visual flicker to the SIFI would predict stronger SIFIs for lower-contrast visual stimuli, SIFIs have been shown to be equivalent (or even stronger) for higher-contrast visual stimuli (Kaposvári et al., 2014). In addition, although we obtained no evidence of flicker induction in the fovea, reliable SIFIs have been demonstrated in the fovea (Kaposvári et al., 2014), though SIFIs tend to be stronger in the periphery (Kumpik et al., 2014). Thus, the mechanisms through which fast (e.g., 12 Hz) AM sounds facilitate flicker detection or might induce illusory flicker seem to be different than the mechanisms through which dynamic auditory stimuli influence the perceived

number of flashes or rate of visual flicker (e.g., Gebhard and Mowbray, 1959; Recanzone, 2003; Shams et al., 2000, 2002; Shipley, 1964; Wada et al., 2003).

In summary, we have demonstrated that AM sounds facilitate visual flicker detection selectively in the periphery (but not in the fovea) for rates well above (but not below) the temporal limit of synchrony-based auditory-visual binding and conscious tracking of auditory-visual synchrony, with sensitivity to a 90° phase-shift (but no sensitivity to a 180° phase-shift). This pattern of results combined with prior results suggest that AM sounds influence the perception of visual dynamics through three distinct mechanisms: (1) slow crossmodal-binding mechanisms that mediate conscious tracking of crossmodal synchrony and synchrony-based crossmodal attention capture (e.g., Fujisaki et al., 2006; Fujisaki and Nishida, 2005; Keetels and Vroomen, 2012; Van der Burg et al., 2008), (2) crossmodal temporal-integration mechanisms that influence the perceived number and rate of visual flashes by strongly weighting the more reliable auditory temporal processing (e.g., Apthorp et al., 2013; Gebhard and Mowbray, 1959; Kaposvári et al., 2014; Kumpik et al., 2014; Recanzone, 2003; Shams et al., 2000, 2002; Shipley, 1964; Wada et al., 2003), and (3) fast sensory mechanisms, likely mediated by direct neural connections from auditory cortex targeting the peripherally tuned frequency-doubling neurons in visual cortex, boosting visual cortical responses to subtle flicker in a phase-specific manner (the current results).

Chapter 3

Silent lip reading generates speech signals in auditory cortex

3.1 Introduction

Speech perception is inherently multisensory with sounds and facial movements both providing relevant information. One way in which visual speech movements facilitate auditory speech processing (Sumby, 1954) is by providing a temporal cue; for example, rapidly processed visual motion signals may reset the phase of the ongoing oscillatory auditory cortical activity (Kayser et al., 2008) to be optimal for processing upcoming speech signals (Schroeder et al., 2008). An intriguing possibility is that visual processing of speech movements may also convey linguistic information to auditory cortex. There is some fMRI evidence suggesting that speech- and object-relevant visual information is conveyed to auditory cortex (Calvert et al., 1997; Hall et al., 2005; Meyer et al., 2010; Pekkola et al., 2005). Whereas these prior studies focused on spatial patterns of fMRI BOLD activity, we focused on the dynamics of electrophysiological activity. Because visual speech movements and vocalization are intrinsically dynamic and temporally correlated, we hypothesized that viewing silent speech movements may generate phoneme-specific dynamic neural

activity in auditory cortex. To monitor auditory cortical activity with millisecond resolution, we recorded electrocorticographic activity from depth electrodes implanted within primary/secondary auditory cortex in epilepsy patients. Patients were presented with one of four representative auditory phonemes (ba, da, ta, or tha) or silent videos showing mouth movements articulating those phonemes. Using an ensemble of deep convolutional neural networks to decode temporal patterns of neural activity, we demonstrate that viewing speech movements generates dynamic activity in auditory cortex similar to that generated while listening to the articulated phonemes. These results highlight a remarkably content-specific exchange of dynamic sensory information at early stages of the cortical processing hierarchy.

3.2 Experiment 4

3.2.1 Method

Patients

Two patients (One female and one male) with epilepsy (36 and 32 years of age, respectively) participated in this study during invasive work-up for medically intractable seizures. They participated during stable periods between seizures, either before or after seizures had been recorded and characterized using ECoG monitoring from chronically implanted depth electrodes (5 mm center-to-center spacing, 2 mm diameter). Electrodes were placed according to the clinical needs of the patients. Written consent was obtained from each patient according to the direction of the institutional review board at the University of Chicago.

MRI and CT Acquisition and Processing

A preoperative T1-weighted MRI and a postoperative CT scan were acquired for each patient to aid in localization of electrodes. Cortical reconstruction and volumetric segmentation of each patient's

MRI was performed with the Freesurfer image analysis suite, which is documented and available for download online (http://surfer.nmr.mgh.harvard.edu/; Dale et al., 1999; Fischl et al., 1999). Postoperative CT scans were registered to the T1-weighted MRI through SPM and electrodes were localized along the Freesurfer cortical surface using customized open-source software developed in our laboratories (available for download online https://github.com/towle-lab/electrode-registration-app/).

Behavioral Task

Patients were seated in a hospital bed or nearby chair. Stimuli were delivered using a laptop computer (using PsychToolbox; Brainard, 1997; Pelli, 1997) and a pair of free-field speakers. On each trial, patients were presented with one of four phonemes [*ba, da, ta, tha*] (an auditory trial), or a silent video showing the mouth portion of a face articulating one of the phonemes (a visual trial). Each stimulus lasted 1000 ms. They were also presented with audiovisual stimuli (a phoneme and video played concurrently) on some of the trials, but those trials were not relevant to the aim of the current study. Videos used in phoneme and viseme trials were recorded by a trained actor and included 10 instances of each phoneme for a total of 40 movies. Each movie (the audio portion for an auditory trial and the video portion for a visual trials) was presented twice in each of the phoneme and viseme conditions, totaling 80 trials per condition (10 video or audio clips x 4 phonemes x 2 repetitions); all conditions were randomly intermixed across trials. Each trial was advanced only after patients reported aloud which phoneme was spoken or articulated (via lip movements).

ECoG Analysis

Electrocorticographic (ECoG) signals were recorded at a sampling rate of 1024 Hz. Data from electrodes near regions previously or subsequently surgically resected were removed from analyses, as were excessively noisy electrodes (with overall amplitude variability exceeding three standard

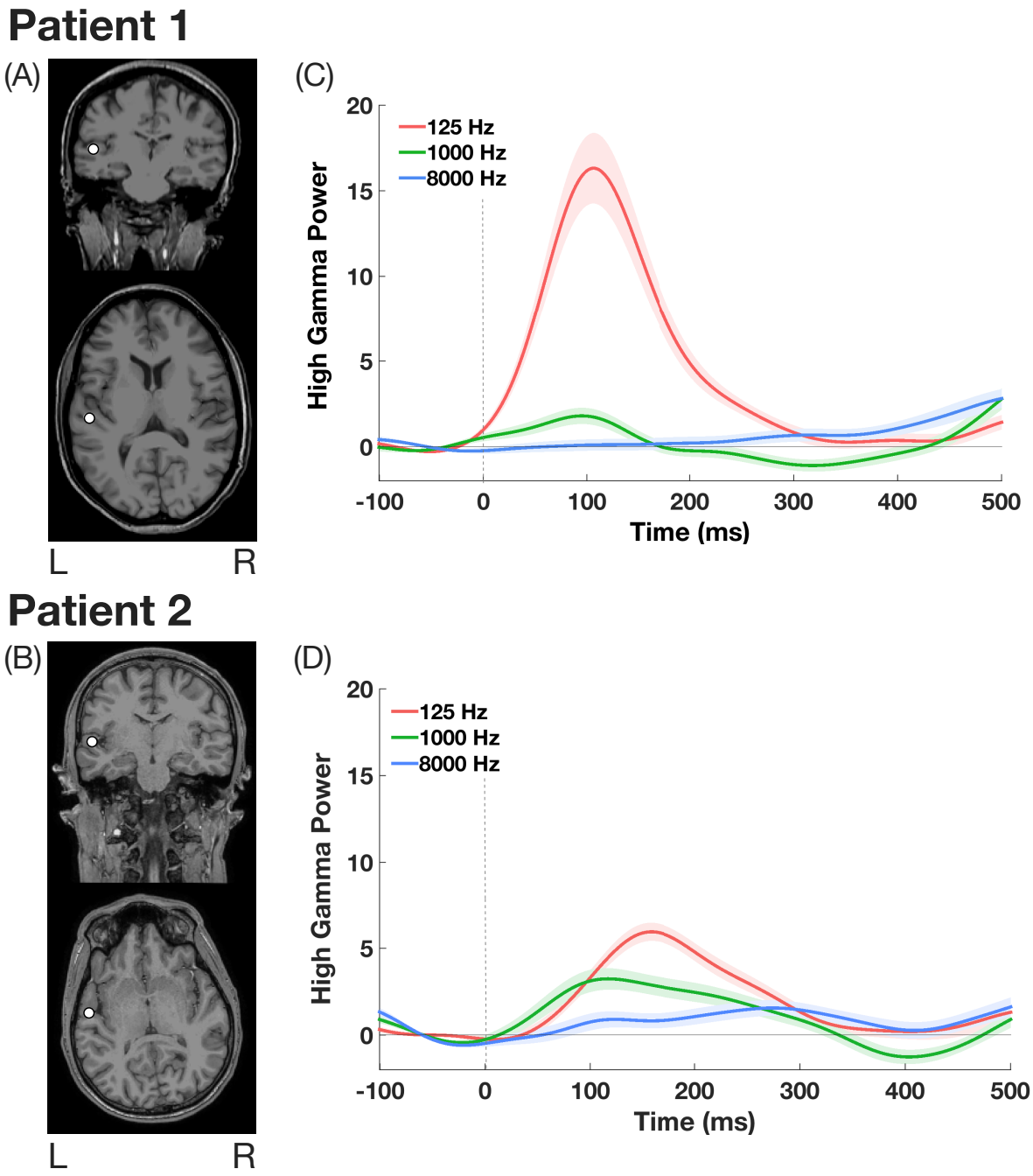

Fig. 3.1 Electrodes were selected based on both anatomical and functional criteria. (A, B) Anatomical location in left superior temporal sulcus and (C, D) the dependence of response amplitudes upon auditory frequency confirm the localization of the selected electrodes in early auditory cortex for both patients. High gamma power was normalized to baseline period. Shaded region represents 95% confidence interval.

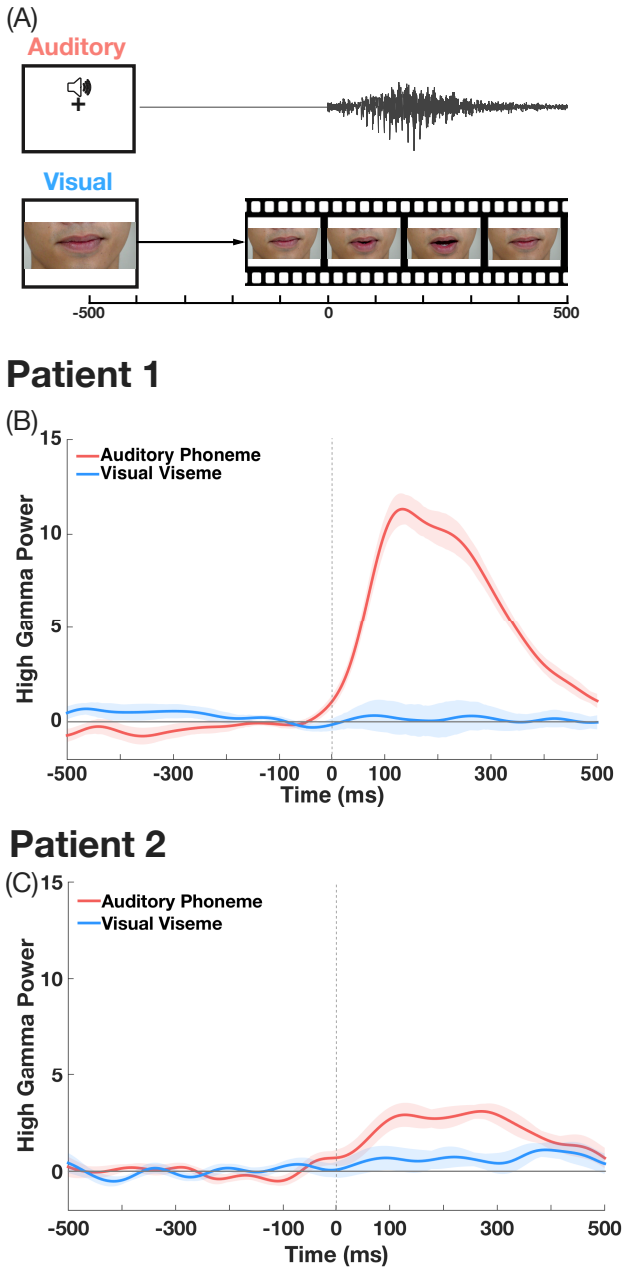

Fig. 3.2 (A) Illustrations of an auditory trial and a visual trial. Participants viewed a fixation screen during an auditory trial. A visual trial began with a static face, followed by a movie of lip movements. (B, C) Average high gamma activity (reflective of synaptic activity and spike rates) from the auditory cortical electrodes in response to phonemes (red curves) and visemes (blue curves) for patient 1 (B) and patient 2 (C). High gamma power was normalized to 500 ms baseline period. Shaded region represents 95% confidence interval.

deviations). ECoG signals were common average referenced, and 500ms epochs were extracted beginning at each phoneme or viseme onset (marked on ECoG recordings with a voltage isolated TTL trigger). Raw ECoG data was used for classification. Electrodes within primary auditory cortex were selected for analyses based on their anatomical location as well as the strength of auditory evoked potentials (see figure 3.1).

Classification of electrocorticographic activity was conducted using an ensemble of deep convolutional neural networks (CNNs). CNNs are non-linear networks comprising multiple layers designed to extract high-dimensional features from multivariate time series (LeCun et al., 2015). We utilized an ensemble method as they have been shown to often outperform single models (Dietterich, 2000). Eight CNNs were included in the ensemble, each constructed with a different architecture based on the VGG network (Simonyan and Zisserman, 2014) with 2-3 convolutional layers before max pooling. All convolutional layers performed 1-dimensional convolutions with parametric rectified linear unit activation (He et al., 2015b). Each model had a different number of filters, filter length, and depth to promote diversity in the ensemble. We utilized dropout between layers to reduce overfitting (Srivastava et al., 2014). All models were built and trained using Keras (Chollet, 2015), running on the Theano framework.

Classification accuracy was determined by performing 30 Monte Carlo iterations of randomly partitioning the phoneme or viseme epochs of ECoG data into training (80%, of which 20% were used for validation), and test sets (20%). Within each randomly partitioned iteration, each of the eight CNNs was trained using the Adam optimization algorithm (Kingma and Ba, 2015) with a learning rate of 0.001 for a maximum of 50 epochs or until loss, as calculated via cross entropy (Golik et al., 2013) on the separate validation set, reached a minimum. To increase the amount of available data for training, new samples were generated for each phoneme or viseme within the training set using a convolutional variational autoencoder 3.4. The autoencoder was trained to learn the latent parameters of the underlying probability distribution. We then selected random samples from this latent normal distribution and used a generative model to map this distribution to the

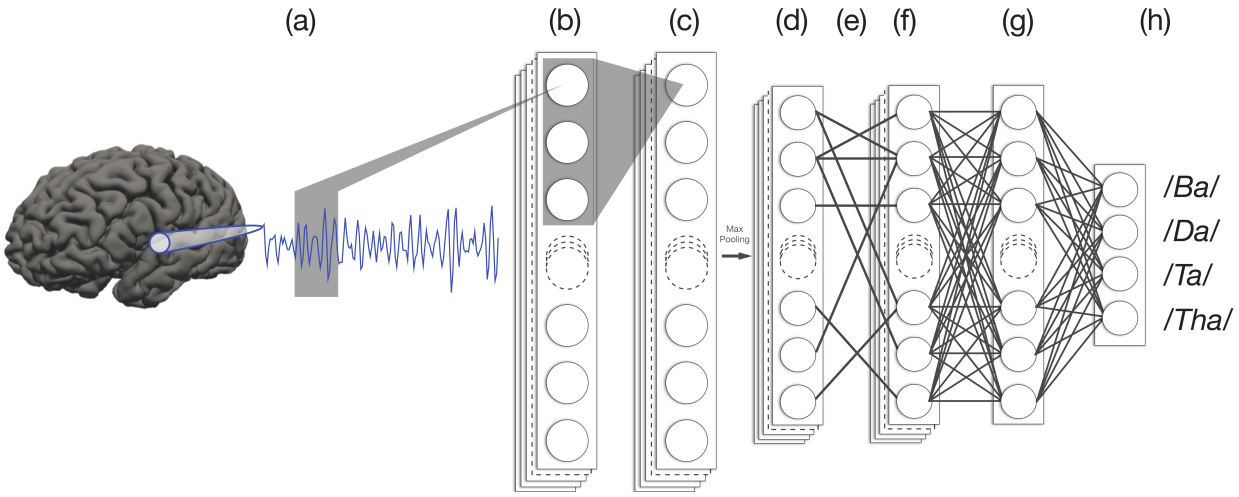

Fig. 3.3 High-dimensional non-linear features are derived from the input time-series without requiring specification of pre-defined features. (A) Input to the network was a half-second raw ECoG signal from auditory cortex in response to a phoneme or viseme trial. (B) The output of a hidden unit (represented above by a circle) in the initial convolutional layer was computed as the dot product between a learnable filter and a time-segment of the input signal. The learnable filter was slid across the input signal to produce an activation map (represented above by a single rectangular layer). A set of activation maps is derived (displayed as multiple rectangular layers stacked along the depth dimension), each of which represents a distinct time invariant temporal feature. (C) The next convolutional layer receives input from the previous convolutional layer. (D) In max pooling layers, data were subsampled, allowing deeper layers of the network access to information spanning longer time periods. (E) Dropout regularization was utilized between max pooling and deeper convolutional layers to minimize overfitting during training. (G) The output from the last convolutional layer was flattened and input to a densely connected layer. (H) At the final classification layer, a four unit densely connected layer with softmax activation corresponding to the phoneme or viseme (ba, da, ta, or tha) output the predicted label of the input signal.

original input space. We generated 2000 new phoneme or viseme epochs for classification training because classification accuracy on the test set (always entirely consisting of the actual data epochs) plateaued near this value. Each trained CNN then predicted the identities of the phonemes or visemes on the separate and unseen test set. Ensemble predictions were determined via hard-voting, a technique by which the predicted identity of the phoneme or viseme is determined by selecting the majority vote across all eight model predictions. The ensemble's accuracy for each randomly partitioned iteration was calculated as the proportion of the phonemes or visemes in the test set that were correctly identified. The 30 randomly partitioned iterations generated 30 accuracy values for each condition with the variability reflecting the reliability of the ensemble model. To evaluate the statistical significance of decoding, we performed an additional 30 randomly partitioned iterations on label-shuffled data (i.e., the phoneme or viseme labels were randomly shuffled across epochs) for each condition to perform t-tests (2-tailed at $alpha = 0.05$) with iteration as the random effect. We chose 30 iterations because an error distribution of the mean has been shown to approach normality for sample sizes of 30 or larger.

3.2.2 Results and Discussion

Several studies have demonstrated that viewing speech movements modulates auditory cortical response during auditory speech perception (Besle et al., 2009, 2008, 2004; Hall et al., 2005; Nath and Beauchamp, 2012). It has also been demonstrated that viewing silent speech movements alone can crossmodally activate auditory cortex (Calvert et al., 1997; Hall et al., 2005; Pekkola et al., 2005). For example, fMRI BOLD activation in auditory cortex was significantly greater in response to vocalization-related facial movements than to nonlinguistic facial movements (Calvert et al., 1997) or to moving circles (Pekkola et al., 2005), with the amount of auditory cortical activation from speech movements correlated with speech reading performance (Hall et al., 2005). These fMRI results suggest that visual processing of speech movements conveys

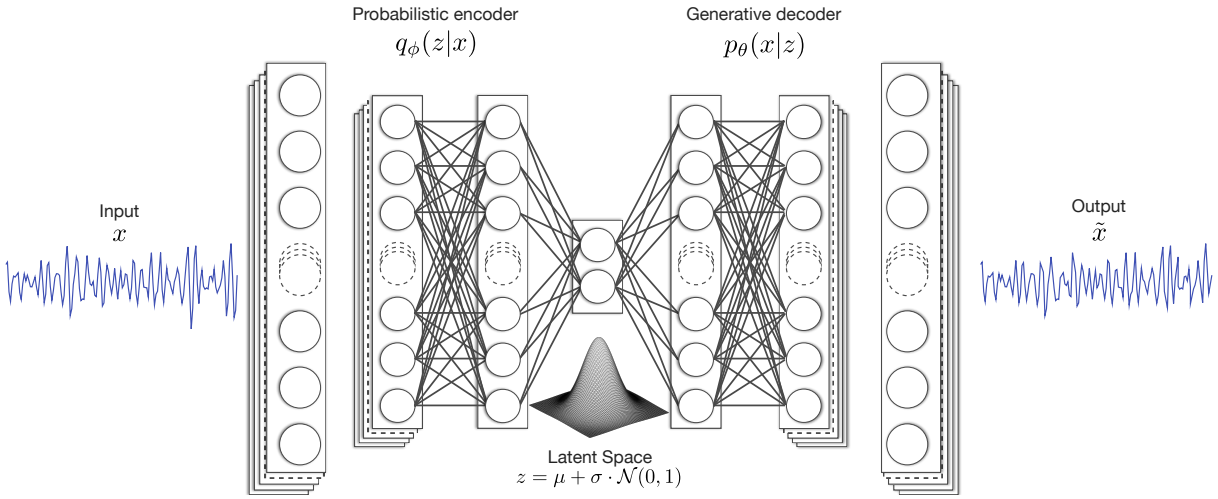

Fig. 3.4 A convolutional variational autoencoder was used to generate novel synthetic trials to increase the number of training samples. The autoencoder is trained to learn the joint distribution of over the input data and a set of latent variables representing the normal probability distribution. By sampling from this distribution following training, the generative decoder can be used to create synthetic trials.

speech-related information to auditory cortex. However, these results did not elucidate the content of visual speech-movement information conveyed to auditory cortex. Because speech signals are intrinsically dynamic, phonemic information coded in early auditory processing is likely to be contained in the dynamics of neural activity. Prior fMRI studies (Calvert et al., 1997; Hall et al., 2005; Pekkola et al., 2005) did not have sufficient temporal resolution to reveal such temporal coding. To elucidate the dynamic content of auditory cortical activity crossmodally evoked by visual speech movements, we recorded electrocorticographic (ECoG) activity from macroscopic depth electrodes implanted within left primary/secondary auditory cortices in two epilepsy patients.

Auditory cortical electrodes were identified based on their anatomical locations as well as the strength of auditory evoked potentials. The selected electrodes exhibited clear frequency selectivity (Fig. 3.1), confirming their locations in early auditory cortex (Moerel et al., 2014).

The patients heard four representative auditory phonemes (*ba, da, ta, and tha*) or saw silent videos of speech movements articulating these phonemes, visemes. Average stimulus-evoked high-

gamma power shows that phonemes strongly activated auditory cortex while visemes generated little overall activation (Fig. 3.2). However, hidden in these average evoked responses is information about individual phonemes and visemes that may be represented in the complex temporal patterns of evoked activity. Specifically, although visemes do not strongly evoke auditory neural activity, they may still reliably modulate the temporal pattern of auditory neural activity in a viseme-specific manner.

To determine the degree to which the dynamics of auditory cortical response to phonemes and visemes contain phoneme-specific information, we attempted to classify heard phonemes or seen visemes based on the temporal patterns of auditory neural activity using an ensemble of deep convolutional neural networks (CNNs; see Methods). CNNs have recently been shown to achieve remarkable success on many challenging tasks involving classification of complex, multidimensional datasets, including image (e.g., face) recognition and natural language processing (LeCun et al., 2015). An advantage of CNNs over other pattern-classification algorithms is that CNNs allow the decoding of information directly from raw data without having to pre-define specific features to be fed to a pattern classifier. As such, CNNs can be used as an effective tool for determining whether specific information is contained in complex signals. CNNs are comprised of multiple convolutional layers with each layer deriving multiple time-invariant dynamic features (or location-invariant spatial features in other applications) from the input signal (Fig. 3.3).

We trained an ensemble of CNNs to classify the four phonemes or visemes on a trial-by-trial basis for a subset (80%) of the ECoG data. Additional training samples were generated using a convolutional variational autoencoder trained on multiple trials of the same phoneme (see methods for additional information). The ensemble's classification accuracy was evaluated by testing its performance on the remaining ECoG data (20%; see methods). Because each independent trial could be labeled as one of four possible phonemes or visemes (ba, da, ta, and tha), the a priori chance-level for classification would be 25%. However, because actual chance-level classification accuracy depends on sample size (Combrisson and Jerbi, 2015), we used a label-shuffling procedure

to generate null distributions to determine statistically significant levels of classification. Successful classification would indicate that temporal patterns of auditory cortical activity contain reliable information about the identity of the four phonemes or visemes.

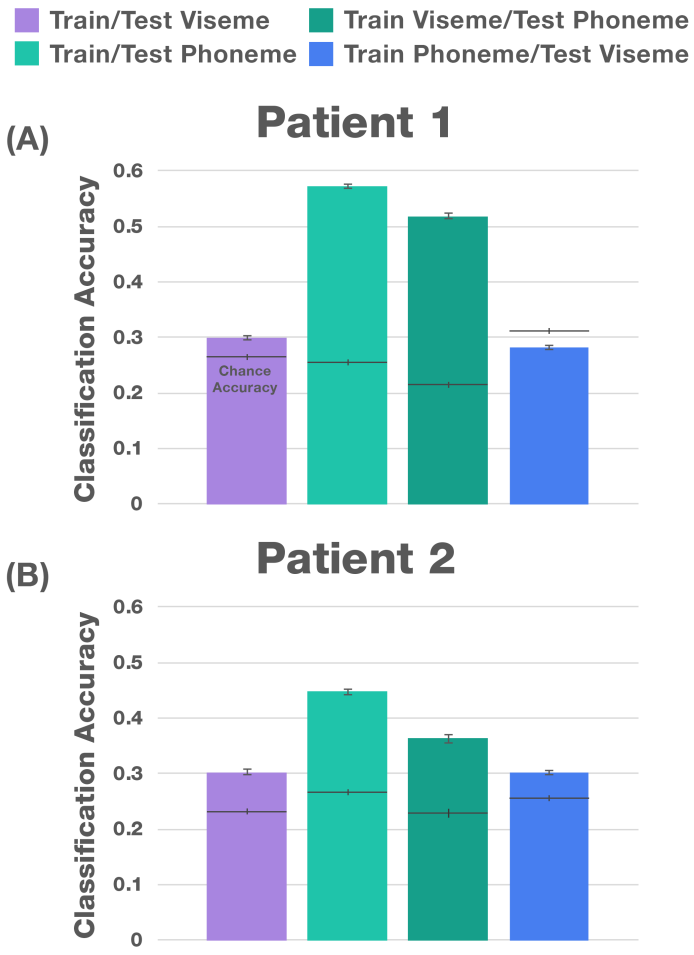

Fig. 3.5 Accuracy for classifying phonemes and visemes based on 0-500ms post-stimulus activity from the auditory cortical electrodes for patient 1 (A) and patient 2 (B). Dashed lines indicate classification accuracies for the corresponding label-shuffled data. The error bars represent ±1 standard error of the mean with Monte Carlo iteration as the random effect (see methods for details).

We successfully classified phoneme evoked activity well above the null distribution (Fig. 4), yielding classification accuracy of 57.3% ($t_{(58)} = 55.60, p < 0.001$, against the label-shuffled accuracy of 25.5%) for patient 1 and 44.8% ($t_{(58)} = 25.97, p < 0.001$, against the label-shuffled

accuracy of 26.8%) for patient 2. For visemes, the CNN ensemble yielded classification accuracy of 29.8% patient 1 ($t_{(58)} = 6.69, p < 0.001$, against the label-shuffled accuracy of 26.5%) and accuracy of 30.3% ($t_{(58)} = 10.31, p < 0.001$, against the label-shuffled accuracy of 23.3%) for patient 2. Classification accuracy was significantly greater for phonemes than for visemes (patient 1, $t_{(58)} = 49.46, p < 0.001$; patient 2, $t_{(58)} = 21.81, p < 0.001$), indicating that dynamic auditory cortical activity more reliably conveys auditory than visual information.

Interestingly, our results demonstrate a consistent pattern of crossmodal asymmetry. Despite significant, but near chance decoding of viseme evoked auditory cortical activity in one patient, our initial evidence suggests visemes may generate noisy signals in auditory cortex that diffusely contains phoneme-specific information. To better understand this possibility, we asked whether the dynamic auditory cortical activity evoked by visual speech movements is similar to that evoked by the corresponding phonemes. If yes, a CNN ensemble trained to identify viseme-specific dynamic auditory cortical activity should also be able to identify the corresponding phoneme-specific activity. This was indeed the case, with mean classification accuracy of 51.8% ($t_{(58)} = 48.45, p < 0.001$, against the label-shuffled accuracy of 21.4%) for patient 1 and 36.3% ($t_{(58)} = 17.26, p < 0.001$, against the label-shuffled accuracy of 23.0%) for patient 2. Importantly, the CNN ensemble trained to classify visemes classified phonemes with greater accuracy than visemes themselves (patient 1, $t_{(58)} = 34.87, p < 0.001$; patient 2, $t_{(58)} = 7.55, p < 0.001$). In contrast, as demonstrated above, the viseme-trained CNN ensemble was barely able to classify new examples of viseme-evoked activity, and a phoneme-trained CNN ensemble was unable to classify viseme-evoked activity. If visemes generate a noisy representation of phoneme-evoked activity in auditory cortex, a classifier trained with visemes would learn the overlapping consistent activation pattern corresponding to phoneme-specific dynamic activity, but the classification accuracy for visemes would be relatively low due to the low consistency and SNR provided by viseme-evoked activity. However, a classifier trained on viseme-evoked activity should perform well on decoding phoneme-evoked activity due to the consistent overlapping feature space. Thus, our result is consistent with the idea that visemes

and phonemes evoke similar phoneme-specific dynamic activity in auditory cortex, except that viseme-evoked activity reflects a noisy representation low in SNR.

A number of recent studies have suggested that multisensory convergence is widespread throughout the cortex so that crossmodal signals can modulate or even drive responses in primary sensory areas of the brain (Shams et al., 2005; Watkins et al., 2007, 2006). While much research has focused on crossmodal modulations of the strength and/or reliability of sensory signals, some studies have demonstrated crossmodal exchanges of content-related information in early cortical processing. In particular, fMRI studies have shown that visual speech movements generate BOLD activity in auditory cortex (Meyer et al., 2010; Pekkola et al., 2005) with the amount of activation correlated with speech reading performance (Hall et al., 2005), suggesting that visual processing of speech movements convey speech-related information to auditory cortex. Another fMRI study has demonstrated that silent videos of sound-implying objects can be classified based on the spatial pattern of BOLD activity in auditory cortex, suggesting that visual processing of objects conveys object-related information to auditory cortex (Meyer et al., 2010). However, no prior study has answered the question of whether visual signals can activate auditory cortex in a content-specific manner, generating object- or speech-specific neural activity similar to that evoked by auditory stimuli.

In natural speech perception, visual speech movements and auditory phonemes provide dynamic signals that are temporally correlated, implying the relevance of temporal coding. Thus, whereas prior investigations of crossmodal sensory activation focused on spatial patterns of BOLD activity using fMRI, we focused on dynamic patterns of electrophysiological activity using ECoG. Using ensembles of CNNs to decode the content of temporal patterns of neural activity, we have demonstrated that viewing visual speech movements generate dynamic auditory cortical activity that is virtually equivalent (albeit lower in signal-to-noise ratio) to that evoked by hearing the corresponding phonemes. The fact that our results are based on the initial 500 ms of stimulus-evoked electrophysiological activity suggests that visual speech movements are concurrently translated

into phoneme-specific dynamic neural activity in auditory cortex. It is as if one hears speech while lip reading. Whereas prior research suggested that visual processing of facial movements facilitates speech perception by providing temporal cues (Kayser et al., 2008; Schroeder et al., 2008), our results suggest that it can also facilitate speech perception by dynamically augmenting auditory speech signals. The current results add a new insight to the growing body of literature on crossmodal interactions in early sensory processing by demonstrating that those interactions include exchanges of content-specific information.

Chapter 4

Recurrent convolutional neural networks for electrophysiologic signal decoding

4.1 Introduction

Neuronal activity in the brain gives rise to extracellular electric and magnetic activity which can be recorded from electrodes with high temporal resolution. Electric potentials and magnetic fields can be recorded in humans and animals both non-invasively from the scalp (electroencephalography [EEG], magnetoencephalography [MEG]) and invasively via subdural electrodes on the surface of the cortex (electrocorticography [ECoG]) or electrodes implanted within the brain (local field potentials [LFP]) (Buzsáki et al., 2012). These recordings can be utilized to provide information about neural interactions and computation, and therefore have been widely utilized by both clinicians and researchers.

Analysis of electrophysiological data is encumbered by properties of the signal including the dynamic non-stationary nature of brain activity as well as the relatively poor signal-to-noise ratio of non-invasive recordings. As such, common analysis techniques such as event related potentials (ERPs) or time-frequency analysis rely upon averaging across multiple trials and many subjects.

However, by averaging ERPs over trials and subjects, reliable brain dynamics can be masked (Gaspar et al., 2011; Stewart et al., 2014).

Multivariate machine learning algorithms allow detection of consistent patterns of neural activity on a trial-by-trial basis (Stewart et al., 2014). These algorithms have been utilized for a variety of purposes including automated sleep scoring (Gao et al., 2016; Längkvist et al., 2012; Sunagawa et al., 2013), epilepsy diagnosis (Alkan et al., 2005; Mirowski et al., 2009, 2008), brain-computer interface technologies (Lotte et al., 2007), as well as for basic research (Müller et al., 2008; Valenzi et al., 2014; Wang et al., 2014). Conventional approaches to electrophysiological classification involve the use of supervised learning algorithms such as support vector machines (Crisler et al., 2008; Mirowski et al., 2008), decision trees (Brankačk et al., 2010; Gao et al., 2016; Mossbridge et al., 2013), naïve bayes (Gao et al., 2016; Rytkönen et al., 2011), and neural networks (Alkan et al., 2005; Baldwin and Penaranda, 2012; Subasi and Erçelebi, 2005). Such approaches, however, often require labor-intensive preprocessing and extraction of expertly-selected features which widely vary across studies.

Here, we describe a technique for classifying electrophysiological data end to end on raw signals without requiring feature extraction, using an ensemble of recurrent convolutional neural networks. We demonstrate accuracy exceeding conventional machine learning methods. A python-based open-source toolbox for performing these analyses is provided.

4.2 Methods

4.2.1 Convolutional Neural Networks

Convolutional neural networks (CNNs) have recently been shown to achieve remarkable success on a variety of challenging tasks involving complex, multidimensional datasets, including image recognition, facial recognition, and natural language processing (LeCun et al., 2015). They are

similar to standard multi-layer perceptron (MLP) networks in that they are comprised of multiple stacked layers. However, in a CNN, each convolutional layer is comprised of multiple kernels (or filters), each of which computes the dot product between its kernel (or filter) and the input, rather than performing the complete matrix multiplication required in a fully connected MLP. Parameters are shared within each kernel, allowing the kernels to extract location- or time-invariant spatial or temporal features from an input signal. Importantly, each filter learns to extract features during training, thereby allowing CNNs to decode information directly from raw data without having to pre-define specific features to be fed to a pattern classifier (LeCun et al., 2015).

4.2.2 Recurrent Convolutional Neural Networks

In a standard CNN, connectivity is comprised of only feed-forward inter-layer connections. As a result, layers at early stages of the network cannot access higher order contextual features of the input data. At deeper layers, following subsampling via maxpooling or strided convolution, hidden units have large enough spatial or temporal receptive field sizes to access higher order features with a larger spatial or temporal scale, yet the feed-forward architecture does not permit this information to be utilized by earlier layers.

An alternative architecture proposed by Liang and Hu (2015), the recurrent convolutional neural network (RCNN), introduces recurrent intra-layer connectivity to a standard CNN. The RCNN is comprised of multiple stacked recurrent convolutional layers (RCLs)(Fig. 4.1), each of which is a subnetwork comprised of both feed-forward and recurrent connectivity (Fig. 4.2). The RCL structure, which provides inputs from earlier layers to later layers within the subnetwork, resembles the standard RNN computation (see Fig. 4.2) and can be reformulated as a recurrent structure highly similar to a conventional RNN (Liao and Poggio, 2016). The RCNN therefore represents a system with multiple sequentially connected recurrent sub-systems. The intra-layer connectivity within each RCL allows the subnetwork to evolve over time and provides multiple paths from

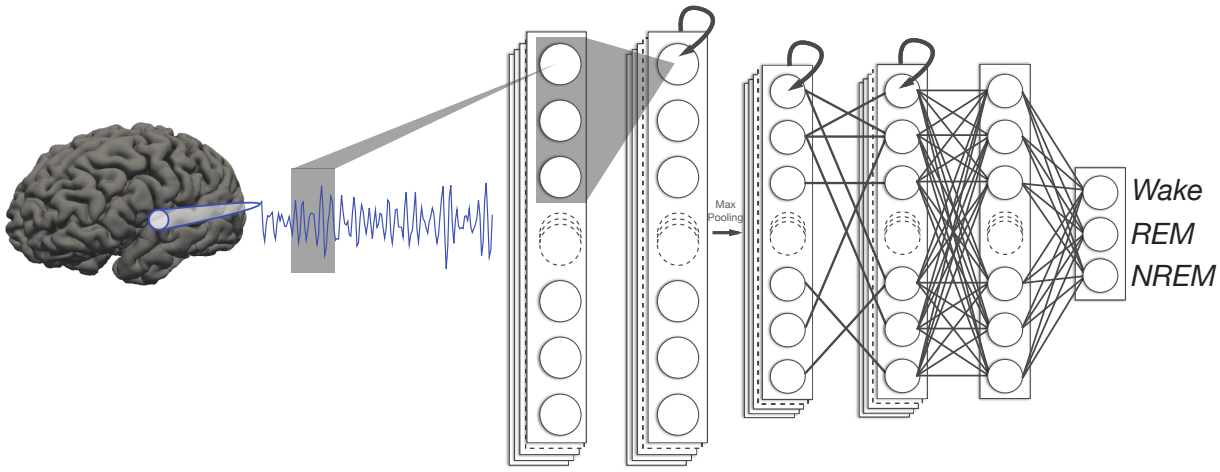

Fig. 4.1 Simplified architecture of a sample Recurrent Convolutional Neural Network (RCNN). High-dimensional non-linear features are derived from input raw amplitude electrophysiologic signals. Recurrent intra-layer connectivity allows time evolving activity over static inputs that is modulated by local contextual information

input to output, allowing deep networks with a reduced number of hyperparameters that remain less prone to vanishing gradients than standard CNNs. Several state of the art convolutional neural networks based on similar architectures (e.g., residual networks, fractal networks) have recently been shown to achieve remarkable success on image recognition tasks, taking first place in the 2015 ImageNet Large Scale Visual Recognition Challenge (He et al., 2015a; Szegedy et al., 2016).

We provide a set of Python functions to easily build and use RCNN networks for electrophysiologic signal decoding. Our implementation of the RCNN architecture follows that of Liang and Hu (2015), with an initial convolutional layer followed by stacks of RCLs before max pooling. A final dense layer with softmax activation is used for classification. Hyperparameters can be modified in the model initialization as demonstrated in several examples in the accompanying code. For all of the following examples, we utilize He weight initialization throughout the network, with weights sampled from a Gaussian distribution with rectified linear unit scaling factors (He et al., 2015b). We utilized the Adam optimization algorithm (Kingma and Ba, 2015) with a learning rate of 0.001, and light L2 regularization of 0.0025.

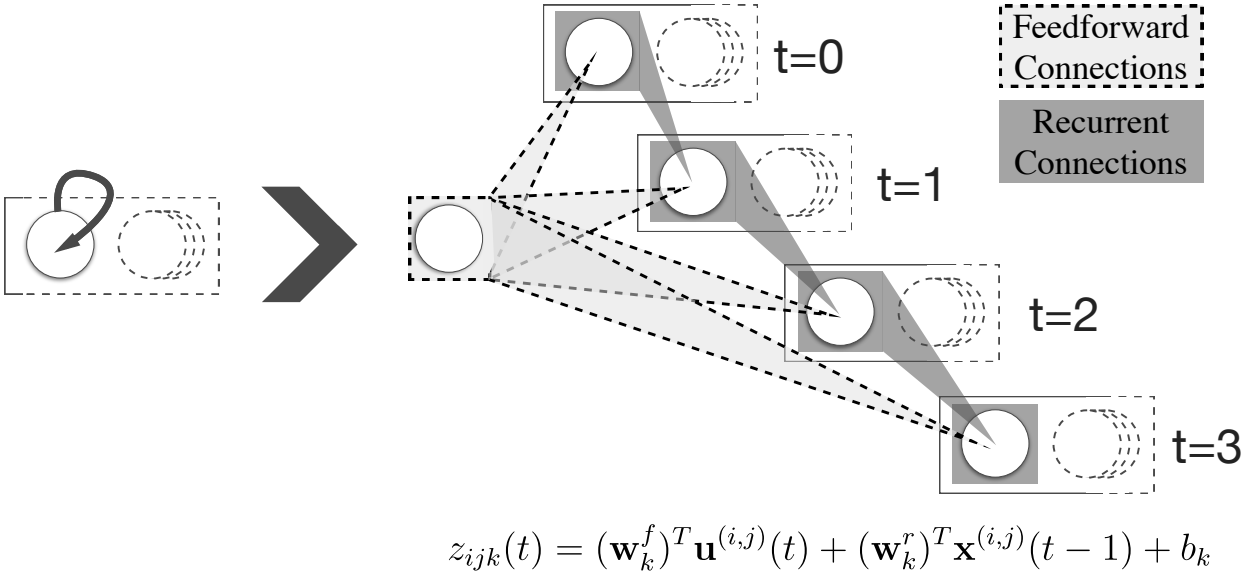

Fig. 4.2 Connectivity within a recurrent convoltional layer (RCL) is comprised of both feedforward and recurrent connectivity. Input from earlier layers is provided directly to later layers by means of skip connections (feedforward connections to later timepoints).

4.2.3 RCNN Ensembles

In an ensemble classifier, several instances of a single model or multiple models each classify the target data and the consensus between models is determined for final classification. Ensembling is commonly used with weak learners such as decision trees, and has been shown to substantially improve generalizability and robustness over single models (Breiman, 1998, 2001; Geurts et al., 2006). More recently, ensembles of strong learning models such as deep neural networks have also been shown to provide similar improvements in accuracy (Deng and Platt, 2014), and are therefore becoming increasingly popular.

Multiple methods for combining predictions are used by researchers to improve predictive models. In the accompanying code, we provide utilities for ensemble decision making including hard-voting, averaging, and stacking of meta-models. In a hard-voting ensemble, each model predicts the class of the target data, and the majority vote is selected as the final predicted class. In an averaging ensemble (or 'soft-voting' ensemble), each model predicts the probability that

a target sample belongs to a particular class. The probabilities are then averaged across models, and the maximum class probability is selected as the final predicted class. A stacked ensemble consists of two stages. First level models (in this case, multiple RCNN models) predict classes (or class probabilities) for target data. These first level predictions are then used as features for second level meta-models, from which the majority predicted class is selected as the final prediction of the stacked model. We provide examples of each type of model ensemble in the accompanying code.

4.3 Experiment 5: Automatic Sleep Scoring in Mice

4.3.1 Methods

Detailed methods for collecting EEG/EMG recordings were previously reported by Gao et al. (2016). In summary, mice were implanted with EEG and EMG electrodes for sleep recording. Recordings were divided into 10 second epochs with a sampling rate of 1000 Hz (downsampled to 200 Hz) with each 24h recording consisting of 8640 epochs.

Sleep data were scored by Gao et al. (2016). In their original experiment, each recording was scored by two human experts: the primary scorer was used to train the classifiers and compare computer-human agreement, while the secondary scorer was used to compare human-human agreement.Using PAL 8200 Sleep Score software (Pinnacle Technologies, Lawrence, KS), the scorer viewed each 10 second epoch and labeled it as either Wake, rapid eye movement sleep (REM), or non-rapid eye movement sleep (NREM), or excluded it from analysis if the signal contained a major artifact.

In the present experiment, classification of the EEG/EMG activity was conducted using an ensemble of deep recurrent convolutional neural networks (RCNNs). Three RCNN models were included in the ensemble, each with a different number of RCLs and filters. We performed sample-wise normalization of data between 0 and 1. Classification accuracy was determined by performing

5 Monte Carlo iterations of randomly partitioning EEG/EMG data into training (80%, of which 20% were used for validation), and test sets (20%). For evaluation of classifier accuracy, epochs are considered as errors if in disagreement with human classification (based on the primary scorer). Each model was trained for a maximum of 200 epochs, or until loss, as calculated via cross entropy on the separate validation set, reached a minimum.

Additional models included for comparison include random forest, naïve bayes, and support vector machine. We also included two standard non-recurrent convolutional neural network architectures. The first, a small network comprised of 3 convolutional layers each with a 3-sample wide kernel, with max pooling following each convolutional layer (*Conv-small*), and a larger network following the VGG architecture (Simonyan and Zisserman, 2014) with 5 consecutive convolutional blocks comprised of 2 stacked convolutional layers followed by max pooling (*Conv-large*). We used paired t-tests on accuracy values of 5 folds for statistical comparisons. We used one sample t-tests for comparisons to the original study presented in Gao et al. (2016).

Ensemble predictions were determined via soft-voting (in the accompanying code, we provide examples of multiple ensemble techniques). The ensemble's accuracy for each randomly partitioned iteration was calculated as the proportion of the trials in the test set that were correctly identified. The five randomly partitioned iterations generated five accuracy values for each condition with the variability reflecting the reliability of the ensemble model.

4.3.2 Results

As described by Gao et al. (2016), a second human scorer disagreed with the original scorer with an average error rate of 0.046. Therefore, we define human accuracy on this sleep classification task is 95.4%.

In the following comparisons, we utilize raw un-preprocessed amplitude data (normalized between 0 and 1). Standard classification algorithms, including several included in the original

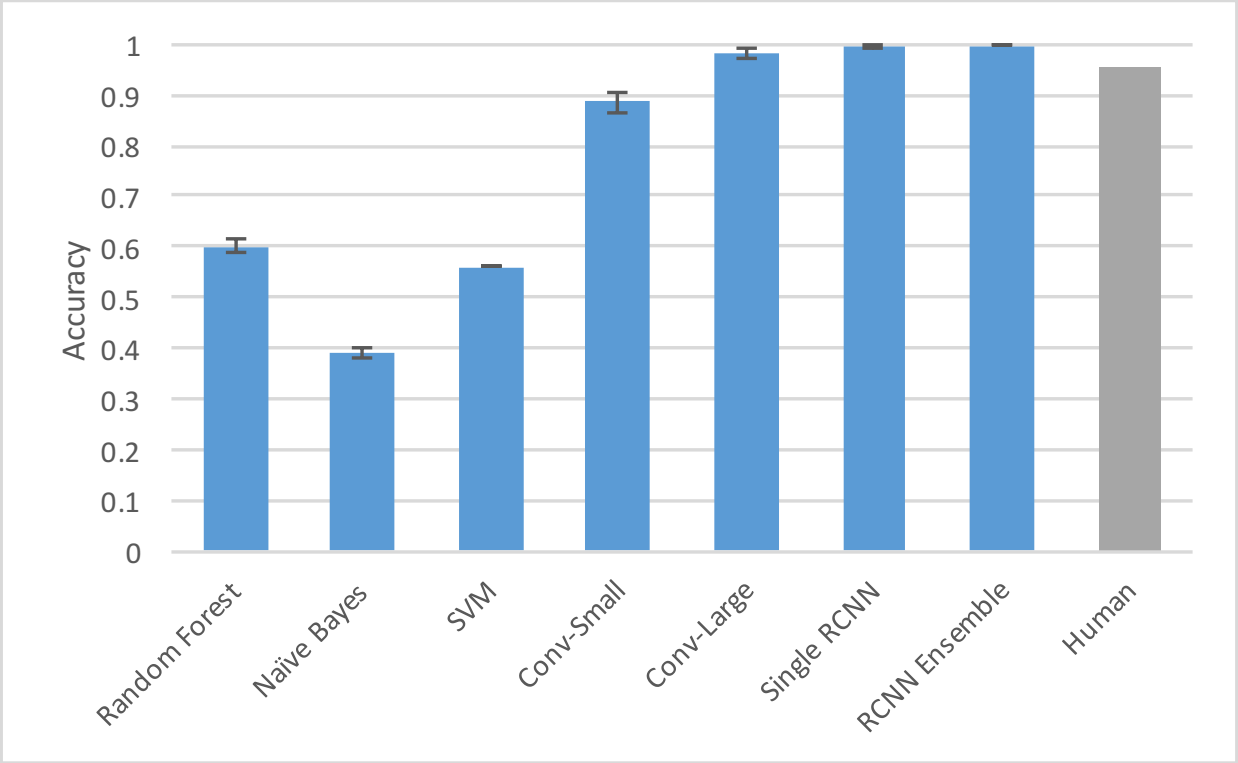

Fig. 4.3 Mean accuracy of classifiers as well as human accuracy as measured via consistency with a second scorer. Convolutional neural network models outperform standard models. Ensembled RCNN models achieve extremely high accuracy rates (99.7%), significantly outperforming all other models, as well as humans on this task.

study, performed poorly on this raw data. Among single classifiers, random forest was the most accurate, achieving accuracy of 60.2% ($SD = .01$). SVM performed slightly worse, with accuracy of 56.1% ($SD = .0002$). Naïve bayes was the least accurate classifier, achieving 39.1% accuracy ($SD = .01$).

A simple convolutional neural network (Conv-small) significantly outperformed all standard classification algorithms on raw data ($t_{(4)} = 33.81, p < .0001$ compared to the highest performing standard algorithm, random forest), achieving accuracy of 88.6% ($SD = .02$). Increasing the capacity of the convolutional neural network and implementing a more standard architecture (Conv-large) led to a significant increase in performance ($t_{(4} = 14.04, p = .0001$), achieving 98.1% accuracy ($SD = .01$), which is significantly greater than human accuracy ($t_{(4)} = 5.51, p = .005$).

A single recurrent convolutional neural network significantly outperformed Conv-large ($t_{(4)} = 3.44, p = .026$), achieving accuracy of 99.4% ($SD = .003$), which is also significantly greater than human accuracy ($t_{(4)} = 29.9, p < .0001$). An ensemble of three recurrent convolutional networks, each with a different number of RCLs and filters, achieved accuracy of 99.7% (SD = .0009), significantly outperforming the single RCNN model ($t_{(4)} = 3.78, p = .02$). Consistent with Gao et al. (2016), REM trials were the most difficult to accurately decode, with the RCNN ensemble incorrectly classifying REM trials as Wake trials on 3% of trials.

Table 4.1 Experiment 5: RCNN Ensemble Classification Metrics

Condition	Precision	Recall	f1 Score	Support
Wake	1.0	1.0	1.0	4803
NREM	1.0	1.0	1.0	3385
REM	0.99	0.97	0.98	432
Avg / Total	1.0	1.0	1.0	8620

Table 4.2 Experiment 5: RCNN Ensemble Confusion Matrix

		Predicted Label		
		Wake	NREM	REM
True Label	Wake	4796	1	6
	NREM	3	3382	0
	REM	13	0	419

4.4 Discussion

These results demonstrate a remarkable effectiveness of using recurrent convolutional neural networks for classifying electrophysiologic data end-to-end using raw amplitude signals. The

recurrent connectivity implemented in these models may perform particularly well due to the time-varying nature of such signals. We demonstrate improved accuracy in automatic sleep scoring compared to state-of-the-art techniques Gao et al. (2016), while reducing the difficulty and potential bias inherent in feature preprocessing. By combining the predictions of multiple well-performing models, we achieve near-zero error rates on automated sleep scoring, significantly outperforming human accuracy as measured via inter-rater consistency.

In the present study, we trained subject-specific models, which would require the building of an initial training set for each subject. Future research to explore the effectiveness of subject-general models may provide a more broadly useful tool for automated sleep scoring. Additionally, models are trained in an offline fashion and training is highly time consuming, limiting their potential usefulness for brain-machine interfaces.

While the applicability of these models to human sleep staging has not yet been evaluated, they have proven highly capable of classifying human electrocorticographic data (see chapter 3). Additional experiments to classify human EEG and MEG will provide a critical validation of these techniques.

In conclusion, recurrent convolutional neural networks are a highly effective method for classifying electrophysiologic signals in an end-to-end fashion. These techniques for trial-by-trial decoding of electrophysiologic data are critical for variety of purposes including automated sleep scoring (Gao et al., 2016; Längkvist et al., 2012; Sunagawa et al., 2013), epilepsy diagnosis (Alkan et al., 2005; Mirowski et al., 2009, 2008), as well as for understanding fundamental questions about neural processing (Müller et al., 2008; Valenzi et al., 2014; Wang et al., 2014).

The code used for these experiments is available at https://github.com/jacobzweig/RCNN_Toolbox

Chapter 5

Conclusion

Across two experiments I demonstrate early multisensory interactions with converging evidence pointing towards potential reliance upon direct neural pathways between auditory and visual corticies. In Chapter 2, I investigated early multisensory interactions and the potential neural populations that are targeted via auditory influences on visual flicker detection. By psychophysically investigating the effects of amplitude modulated sounds on flicker sensitivity, I demonstrated evidence for rapid (operating beyond the temporal limit of consciously tracking auditory-visual synchrony) phase-specific auditory influences on early dynamic visual processing. The spatiotemporal characteristics of the behavioral results combined with the relevant knowledge from neuroanatomy and neurophysiology suggest that the underlying interactions are mediated by direct neural connections from auditory cortex that target the peripherally-tuned frequency-doubling neurons in visual cortex.

In Chapter 3, I extended my investigation of early multisensory integration processes by exploring how visual processing of speech includes rapid and highly specific exchanges of information. While previous evidence from neuroimaging studies in humans suggested that multisensory interactions during speech might occur at the level of primary sensory cortex, the poor temporal resolution of fMRI had limited our understanding of the content of visually driven activity in auditory cortex during speech processing. Using deep convolutional neural networks applied to data from direct

neural recordings in human surgical patients, I demonstrated that silent lip reading activates primary auditory cortex in a manner similar to auditory speech. These results add new insight into the growing body of literature on crossmodal interactions in early sensory processing by demonstrating that those interactions include exchanges of highly detailed content-specific information.

Finally, in Chapter 4 I demonstrated a novel technique for probing these types of questions using end-to-end decoding of electrophysiologic signals. Using this technique provides state-of-the-art accuracy while simplifying the decoding pipeline by removing the need for extensive feature preprocessing steps. I show successful automated sleep classification in mice that exceeds human accuracy. Additionally, I provided an open-source toolbox with documentation and several examples of advanced decoding and model-ensembling techniques with simple abstraction.

The combined set of studies shed light on the early and highly-specific nature of multisensory interactions, a dramatic departure from the traditional view of the sensory processing hierarchy. The multivariate decoding techniques demonstrated in these studies will continue to advance our understanding of the dynamic relationship between neural processes and perception.

References

Alais, D. and Burr, D. (2004). Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, 14(3):257–262.

Albrecht, D. G. and Hamilton, D. B. (1982). Striate cortex of monkey and cat: contrast response function. *Journal of neurophysiology*, 48(1):217–237.

Alkan, A., Koklukaya, E., and Subasi, A. (2005). Automatic seizure detection in EEG using logistic regression and artificial neural network. *Journal of Neuroscience Methods*, 148(2):167–176.

Apthorp, D., Alais, D., and Boenke, L. T. (2013). Flash illusions induced by visual, auditory, and audiovisual stimuli. *Journal of Vision*, 13(5):1–15.

Baldwin, C. L. and Penaranda, B. N. (2012). Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *NeuroImage*, 59(1):48–56.

Benucci, A., Frazor, R. A., and Carandini, M. (2007). Standing Waves and Traveling Waves Distinguish Two Circuits in Visual Cortex. *Neuron*, 55(1):103–117.

Besle, J., Bertrand, O., and Giard, M. H. (2009). Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex. *Hearing Research*, 258(1-2):143–151.

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., and Giard, M.-H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(52):14301–14310.

Besle, J., Fort, A., Delpuech, C., and Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *The European journal of neuroscience*, 20(8):2225–34.

Blake, R. and Fox, R. (1974). Adaptation to invisible gratings and the site of binocular rivalry suppression. *Nature*, 249:488–490.

Blake, R., Tadin, D., Sobel, K. V., Raissian, T. a., and Chong, S. C. (2006). Strength of early visual adaptation depends on visual awareness. *Proceedings of the National Academy of Sciences of the United States of America*, 103(12):4783–8.

Bolognini, N., Senna, I., Maravita, A., Pascual-Leone, A., and Merabet, L. B. (2010). Auditory enhancement of visual phosphene perception: the effect of temporal and spatial factors and of stimulus intensity. *Neuroscience letters*, 477:109–114.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial vision*, 10(4):433–6.

Brankačk, J., Kukushka, V. I., Vyssotski, A. L., and Draguhn, A. (2010). EEG gamma frequency and sleep–wake scoring in mice: Comparing two types of supervised classifiers. *Brain Research*, 1322:59–71.

Breiman, L. (1998). Arcing classifiers. *Annals of Statistics*, 26(3):801–849.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Burr, D., Banks, M. S., and Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research*, 198(1):49–57.

Buzsáki, G., Anastassiou, C. a., and Koch, C. (2012). The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, 13(6):407–420.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D., and David, a. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, 276(5312):593–596.

Campbell, F. W. and Kulikowski, J. J. (1972). The visual evoked potential as a function of contrast of a grating pattern. *The Journal of physiology*, 222(2):345–356.

Cappe, C. and Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *European Journal of Neuroscience*, 22(11):2886–2902.

Chollet, F. (2015). Keras. https://github.com/fchollet/keras.

Clavagnier, S., Falchier, A., and Kennedy, H. (2004). Long-distance feedback projections to area V1: implications for multisensory integration, spatial awareness, and visual consciousness. *Cognitive, affective & behavioral neuroscience*, 4(2):117–126.

Combrisson, E. and Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods*, 250:126–136.

Crisler, S., Morrissey, M. J., Anch, A. M., and Barnett, D. W. (2008). Sleep-stage scoring in the rat using a support vector machine. *Journal of Neuroscience Methods*, 168(2):524–534.

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194.

Deng, L. and Platt, J. (2014). Ensemble Deep Learning for Speech Recognition. *Research.Microsoft.Com*, (September):1915–1919.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple classifier systems*, pages 1–15.

Ernst, M. O. and Bülthoff, H. H. (2004). Merging the senses into a robust percept.

Falchier, A., Clavagnier, S., Barone, P., and Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 22:5749–5759.

Fischl, B., Sereno, M. I., and Dale, A. M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207.

Fujisaki, W., Koene, A., Arnold, D., Johnston, A., and Nishida, S. (2006). Visual search for a target changing in synchrony with an auditory signal. *Proceedings. Biological Sciences / The Royal Society*, 273(12):865–874.

Fujisaki, W. and Nishida, S. (2005). Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. In *Experimental Brain Research*, volume 166, pages 455–464.

Gao, V., Turek, F., and Vitaterna, M. (2016). Multiple classifier systems for automatic sleep scoring in mice. *Journal of neuroscience methods*, 264:33–39.

Gaspar, C. M., Rousselet, G. A., and Pernet, C. R. (2011). Reliability of ERP and single-trial analyses. *NeuroImage*, 58(2):620–629.

Gebhard, J. W. and Mowbray, G. H. (1959). On discriminating the rate of visual flicker and auditory flutter. *The American journal of psychology*, 72:521–529.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Ghazanfar, A. a. and Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in cognitive sciences*, 10(6):278–285.

Golik, P., Doetsch, P., and Ney, H. (2013). Cross-entropy vs. Squared error training: A theoretical and experimental comparison. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1756–1760.

Hall, A. J. and Lomber, S. G. (2008). Auditory cortex projections target the peripheral field representation of primary visual cortex. *Experimental brain research. Experimentelle Hirnforschung. Experimentation cerebrale*, 190:413–430.

Hall, D. A., Fussell, C., and Summerfield, A. Q. (2005). Reading fluent speech from talking faces: typical brain networks and individual differences. Technical Report 6.

He, K., Zhang, X., Ren, S., and Sun, J. (2015a). Deep Residual Learning for Image Recognition. *Arxiv.Org*, 7(3):171–180.

He, K., Zhang, X., Ren, S., and Sun, J. (2015b). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv preprint*, pages 1–11.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–43.

Kaposvári, P., Bognár, A., Csibri, P., Utassy, G., and Sáry, G. (2014). Fusion and fission in the visual pathways. *Physiological research*, 63(5):625–35.

Kayser, C. and Logothetis, N. K. (2007). Do early sensory cortices integrate cross-modal information? *Brain structure & function*, 212:121–132.

Kayser, C., Petkov, C. I., and Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex*, 18(7):1560–1574.

Keetels, M. and Vroomen, J. (2012). Perception of Synchrony between the Senses.

Kim, Y.-J., Grabowecky, M., Paller, K. a., and Suzuki, S. (2011). Differential roles of frequency-following and frequency-doubling visual responses revealed by evoked neural harmonics. *Journal of cognitive neuroscience*, 23(8):1875–86.

Kingma, D. P. and Ba, J. L. (2015). Adam: a Method for Stochastic Optimization. *International Conference on Learning Representations*, pages 1–13.

Kumpik, D. P., Roberts, H. E., King, A. J., and Bizley, J. K. (2014). Visual sensitivity is a stronger determinant of illusory processes than auditory cue parameters in the sound-induced flash illusion. *Journal of vision*, 14(7):1–14.

Längkvist, M., Karlsson, L., and Loutfi, A. (2012). Sleep Stage Classification Using Unsupervised Feature Learning.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Leopold, D. A. and Logothetis, N. K. (1999). Multistable phenomena: Changing views in perception.

Liang, M. and Hu, X. (2015). Recurrent Convolutional Neural Network for Object Recognition. *Cvpr*, (Figure 1).

Liao, Q. and Poggio, T. (2016). Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex. *arXiv preprint*, (047):1–16.

Lotte, F., Congedo, M., Lécuyer, a., Lamarche, F., and Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of neural engineering*, 4(2):R1–R13.

Martuzzi, R., Murray, M. M., Michel, C. M., Thiran, J.-P., Maeder, P. P., Clarke, S., and Meuli, R. A. (2007). Multisensory interactions within human primary cortices revealed by BOLD dynamics. *Cerebral cortex (New York, N.Y. : 1991)*, 17:1672–1679.

Meyer, K., Kaplan, J. T., Essex, R., Webber, C., Damasio, H., and Damasio, A. (2010). Predicting visual stimuli on the basis of activity in auditory cortices. *Nature Neuroscience*, 13(6):667–668.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, 14(2):247–279.

Mirowski, P., Madhavan, D., LeCun, Y., and Kuzniecky, R. (2009). Classification of patterns of EEG synchronization for seizure prediction. *Clinical Neurophysiology*, 120(11):1927–1940.

Mirowski, P. W., LeCun, Y., Madhavan, D., and Kuzniecky, R. (2008). Comparing SVM and convolutional networks for epileptic seizure prediction from intracranial EEG. In *Proceedings of the 2008 IEEE Workshop on Machine Learning for Signal Processing, MLSP 2008*, pages 244–249.

Mishra, J., Martínez, A., and Hillyard, S. A. (2010). Effect of attention on early cortical processes associated with the sound-induced extra flash illusion. *Journal of cognitive neuroscience*, 22:1714–1729.

Mishra, J., Martinez, A., Sejnowski, T. J., and Hillyard, S. A. (2007). Early cross-modal interactions in auditory and visual cortex underlie a sound-induced visual illusion. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27:4120–4131.

Moerel, M., De Martino, F., and Formisano, E. (2014). An anatomical and functional topography of human auditory cortical areas. *Frontiers in Neuroscience*, 8(8 JUL):1–14.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., and Foxe, J. J. (2002). Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive Brain Research*, 14(1):115–128.

Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, 4(2):61–64.

Mossbridge, J. a., Grabowecky, M., Paller, K. a., and Suzuki, S. (2013). Neural activity tied to reading predicts individual differences in extended-text comprehension. *Frontiers in human neuroscience*, 7(November):655.

Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., and Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: From brain–computer interfacing to mental state monitoring. *Journal of Neuroscience Methods*, 167(1):82–90.

Nath, A. R. and Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage*, 59(1):781–787.

Noesselt, T., Fendrich, R., Bonath, B., Tyll, S., and Heinze, H. J. (2005). Closer in time when farther in space - Spatial factors in audiovisual temporal integration. *Cognitive Brain Research*, 25(2):443–458.

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H.-J., and Driver, J. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27:11431–11441.

O'Connor, K. N., Johnson, J. S., Niwa, M., Noriega, N. C., Marshall, E. A., and Sutter, M. L. (2011). Amplitude modulation detection as a function of modulation frequency and stimulus duration: Comparisons between macaques and humans. *Hearing Research*, 277(1-2):37–43.

Ortega, L., Guzman-Martinez, E., Grabowecky, M., and Suzuki, S. (2014). Audition dominates vision in duration perception irrespective of salience, attention, and temporal discriminability. *Attention, perception & psychophysics*, 76(5):1485–502.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., and Sams, M. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T. *Neuroreport*, 16(2):125–128.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial vision*, 10(4):437–42.

Prins, N. and Kingdom, F. A. A. (2009). Palamedes: Matlab routines for analyzing psychophysical data.

Qin, L., Chimoto, S., Sakai, M., Wang, J., and Sato, Y. (2007). Comparison Between Offset and Onset Responses of Primary Auditory Cortex. *Journal of Neurophysiology*, 97(5):3421–3431.

Recanzone, G. H. (2000). Spatial processing in the auditory cortex of the macaque monkey. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):11829–11835.

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of neurophysiology*, 89(2):1078–1093.

Regan, D. and Spekreijse, H. (1977). Auditory visual interactions and the correspondence between perceived auditory space and perceived visual space. *Perception*, 6(2):133–138.

Rockland, K. S. and Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, 50:19–26.

Rytkönen, K. M., Zitting, J., and Porkka-Heiskanen, T. (2011). Automated sleep scoring in rats and mice using the naive Bayes classifier. *Journal of Neuroscience Methods*, 202(1):60–64.

Schroeder, C. E. and Foxe, J. (2005). Multisensory contributions to low-level, 'unisensory' processing. *Current opinion in neurobiology*, 15(4):454–8.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, 12(3):106–113.

Shams, L., Iwaki, S., Chawla, A., and Bhattacharya, J. (2005). Early modulation of visual cortex by sound: an MEG study. *Neuroscience letters*, 378(2):76–81.

Shams, L., Kamitani, Y., and Shimojo, S. (2000). What you see is what you hear. *Nature*, 408(6814):788.

Shams, L., Kamitani, Y., and Shimojo, S. (2002). Visual illusion induced by sound. In *Cognitive Brain Research*, volume 14, pages 147–152.

Shams, L., Kamitani, Y., Thompson, S., and Shimojo, S. (2001). Sound alters visual evoked potentials in humans. *Neuroreport*, 12(17):3849–3852.

Shipley, T. (1964). Auditory flutter driving of visual flicker. *Science (New York, N.Y.)*, 145(3638):1328–1330.

Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv*, pages 1–14.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15:1929–1958.

Stein, B. E., London, N., Wilkinson, L. K., and Price, D. D. (1996). Enhancement of Perceived Visual Intensity by Auditory Stimuli: A Psychophysical Analysis.

Stein, B. E. and Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature reviews. Neuroscience*, 9(4):255–66.

Stewart, A. X., Nuthmann, A., and Sanguinetti, G. (2014). Single-trial classification of EEG in a visual object task using ICA and machine learning. *Journal of Neuroscience Methods*, 228:1–14.

Strasburger, H., Rentschler, I., and Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of vision*, 11(5):13.

Subasi, A. and Erçelebi, E. (2005). Classification of EEG signals using neural network and logistic regression. *Computer Methods and Programs in Biomedicine*, 78(2):87–99.

Sumby, W. H. (1954). Visual Contribution to Speech Intelligibility in Noise.

Sunagawa, G. A., S, H., Shimba, S., Urade, Y., and Ueda, H. R. (2013). FASTER: An unsupervised fully automated sleep staging method for mice. *Genes to Cells*, 18(6):502–518.

Sweeny, T. D., Grabowecky, M., and Suzuki, S. (2011). Awareness becomes necessary between adaptive pattern coding of open and closed curvatures. *Psychological science*, 22(7):943–50.

Szegedy, C., Ioffe, S., and Vanhoucke, V. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Arxiv*, page 12.

Valenzi, S., Islam, T., Jurica, P., and Cichocki, A. (2014). Individual Classification of Emotions Using EEG. *Journal of Biomedical Science . . .* , (June):604–620.

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of experimental psychology. Human perception and performance*, 34(5):1053–65.

Van Hateren, J. H. (1993). Spatiotemporal contrast sensitivity of early vision. *Vision Research*, 33(2):257–267.

Wada, Y., Kitagawa, N., and Noguchi, K. (2003). Audio-visual integration in temporal perception. In *International Journal of Psychophysiology*, volume 50, pages 117–124.

Wang, X.-W., Nie, D., and Lu, B.-L. (2014). Emotional state classification from EEG data using machine learning approach. *Neurocomputing*, 129(August):94–106.

Watkins, S., Shams, L., Josephs, O., and Rees, G. (2007). Activity in human V1 follows multisensory perception. *NeuroImage*, 37(2):572–8.

Watkins, S., Shams, L., Tanaka, S., Haynes, J.-D., and Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. *NeuroImage*, 31(3):1247–56.

Wilson, J. T. (1987). Interaction of simultaneous visual events. *Perception*, 16(3):375–383.

Zangenehpour, S. and Zatorre, R. J. (2010). Crossmodal recruitment of primary visual cortex following brief exposure to bimodal audiovisual stimuli. *Neuropsychologia*, 48:591–600.