School of Life and Medical Sciences

Ear Institute

A VISIONARY APPROACH TO LISTENING: DETERMINING THE ROLE OF VISION IN AUDITORY SCENE ANALYSIS

by

Huriye ATILGAN

A thesis to be presented for the degree of doctor of philosophy

Declaration

I, Huriye Atilgan, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

The study was conducted with the approval by the Committee on Animal Care and Ethical Review of UCL and licensed by the UK Home Office. The experiment was carried out in the UCL Ear Institute and Royal Veterinary College in London in the UK.

The work was done under the guidance of Dr. Jennifer Bizley at the UCL Ear Institute.

Signature:

Date:

Acknowledgment

Special thanks must go to my supervisor, Dr Jennifer Bizley, for her commitment in guiding me throughout my research project. Her continuous enthusiasm, encouragement and motivation have been contagious, and she has been a great teacher to me. Her comments, suggestions and generosity in time have been invaluable. I have been very fortunate to have a supervisor as caring and dedicated as her.

I have been fortunate to have a wonderful group throughout my time in the Bizley lab. I would like to thank Dr. Stephen Town, Dr. Katherine Wood and Dr. Gareth Jones who were instrumental in the collection of data and the completion of this research. I would like to thank them all for their constant support at every stage of this project. I would like to take this opportunity to thank the Action on Hearing Loss for the funding of my work.

I would also like to Assistant Prof Ross K Maddox from the University of Rochester and Prof Adrian KC Lee from the University of Washington for their help in the design of the project and act as an external collaborator.

I would like to express my gratitude to the staff members of RVC for their help during the operations and in the postoperative care of the animals. My unlimited thanks to a number of ferrets sacrificed for finding the solutions to the aimed questions, in this research work. Special thanks to participants for their support and willingness to volunteer in this study and making it a success. I fully acknowledge that without their participation, this research study would not have been possible.

Finally, I would like to thank my partner in life, Mehmet. The task of writing this thesis would have been so much more difficult without his love and support.

This thesis is dedicated to my dear parents, Munevver & Salih Sami Atilgan

for love and faith

Abstract

To recognize and understand the auditory environment, the listener must first separate sounds that arise from different sources and capture each event. This process is known as auditory scene analysis. The aim of this thesis is to investigate whether and how visual information can influence auditory scene analysis.

The thesis consists of four chapters. Firstly, I reviewed the literature to give a clear framework about the impact of visual information on the analysis of complex acoustic environments. In chapter II, I examined psychophysically whether temporal coherence between auditory and visual stimuli was sufficient to promote auditory stream segregation in a mixture. I have found that listeners were better able to report brief deviants in an amplitude modulated target stream when a visual stimulus changed in size in a temporally coherent manner than when the visual stream was coherent with the non-target auditory stream. This work demonstrates that temporal coherence between auditory and visual features can influence the way people analyse an auditory scene.

In chapter III, the integration of auditory and visual features in auditory cortex was examined by recording neuronal responses in awake and anaesthetised ferret auditory cortex in response to the modified stimuli used in Chapter II. I demonstrated that temporal coherence between auditory and visual stimuli enhances the neural representation of a sound and influences which sound a neuron represents in a sound mixture. Visual stimuli elicited reliable changes in the phase of the local field potential which provides mechanistic insight into this finding. Together these findings provide evidence that early cross modal integration underlies the behavioural effects in chapter II.

Finally, in chapter IV, I investigated whether training can influence the ability of listeners to utilize visual cues for auditory stream analysis and showed that this ability improved by training listeners to detect auditory-visual temporal coherence.

Contents

List of Figures

List of Tables

List of Abbreviations

A1	Primary auditory cortex
AAF	anterior auditory field
AC	auditory cortex
ADF	anterior dorsal field
AM	amplitude modulation
ANOVA	analysis of variance
AV	auditory visual
dB	decibel
ECG	electrocardiogram
EEG	electroencephalography
ERP	evoked response potential
F0	fundamental frequency
F1	first formant
fMRI	functional magnetic resonance imaging
Hz	Hertz
MEG	magnetoencephalography
MMN	mismatch negativity
PPF	posterior pseudosylvian field
PSF	posterior suprasylvian field
SNR	signal-to-noise ratio
SSY	suprasylvian cortex
VC	visual cortex
V1	primary visual cortex
2IFC	two-interval forced choice task

Chapter I: Literature Review

Our senses provide us continuous and different information about objects within our environment. In the auditory modality, there may be many sound sources active within our environment at any time and this acoustic information mixes together even before reaching our ears. It is challenging to identify all different sound sources and determine their relationship with previous events. So how does the auditory system construct, modify and maintain dynamic representations of continuous auditory stimuli within the environment, and how does it integrate with other sensory modalities? This thesis will investigate the role of auditory visual (AV) integration in the process by which the auditory system decomposes incoming complex signals into separate perceptual representations that can be identified as different objects.

This chapter is divided into four main parts in order to give a clear framework of the literature about the impact of visual information on the analysis of complex acoustic environments. The first part presents an overview of some of the perceptual principles underlying auditory scene analysis. The review is not intended to be exhaustive. Rather, it describes only those aspects of auditory scene analysis theory which are relevant to the following chapters of this thesis. The reader is directed to the book by Bregman (1990) for a comprehensive account for auditory scene analysis. The second part will introduce the current state of research on AV cross-modal correspondence and its impacts on auditory scene analysis. The third part reviews the neural correlates of AV integration, focusing on auditory cortex which is the region of interest in this thesis. All these three parts are required to give a baseline for a description of the scope of the thesis and the general hypothesis which is given in part IV.

Part I: Auditory Scene Analysis

The ability to segregate sounds produced by distinct sources is critical for noticing dangerous events and distinguishing speech in a noisy environment. In order to recognize and understand the auditory environment, the listener must first separate sounds that arise from different sources and capture each event. This process is referred to as auditory scene analysis (Bregman, 1990). An auditory stream is “the percept of a group of successive and/or simultaneous sound elements as a coherent whole, appearing to emanate from a single source" (pg. 919, Moore and Gockel, 2012). Auditory scene analysis relies on the simultaneous and sequential organization of auditory information, perceptually linking auditory events coming from the same source over time (i.e., integration), together with segregating them from sounds coming from distinct sources. Listening to sounds that unfold over time, the sounds may be perceived as a single stream (called fusion), or they may be perceived as more than one stream (called stream segregation).

Auditory stream segregation can be based on the degree of peripheral separation between sounds, in which sounds excite distinct cochlear channels, as determined by their frequencies or lateralization (van Noorden, 1975, Hartmann and Johnson, 1991). Stream segregation can also be based on the perceptual differences even when sounds excite same cochlear channel including temporal envelope (Bregman and Dannenbring, 1973), periodicity information (Vliegen and Oxenham, 1999), amplitude modulation rate (Grimault et al., 2002), phase (Roberts et al., 2002) and timbre (Iverson, 1995). Thus, any sufficient noticeable perceptual differences may lead to auditory stream segregation (see reviews for more details Moore and Gockel, 2002, Moore and Gockel, 2012).

Bregman (1990) assumes that streaming involves both primitive (bottom-up sensory process) and schema-based processes (top-down sensory process). The primitive process is a perceptual mechanism, based on perceptual attributes like frequency, intensity, or spectral content, which allow initial parallel processing of acoustic signals coming from different sources. This bottom-up mechanism decomposes the incoming sound into the distinct auditory stream.

The secondary, schema-based mechanism is organized through each listener's experiences and learned abilities during a lifetime of listening. It requires attention and is a knowledge-driven, top-down selection mechanism. It helps to match the incoming stream with memory-stored knowledge. There are different factors in forming streams such as attention and prior knowledge of the stimuli.

The rest of this part will analyse in more detail primitive processing including relevant principles of auditory perceptual organization and schema-based processing of auditory scene analyses.

Primitive Processing: Principles of perceptual organization

In his book, Bregman identifies a number of principles that the auditory system appears to use to group acoustic components together. For example, sounds sharing similar acoustic cues (i.e., frequency) are perceived as coming from the same auditory source whereas sounds with different cues are analysed as coming from distinct sources and are thus segregated by the auditory system. This part presents a brief overview of the principles underlying auditory scene analysis.

The Gestalt psychologists formulated a theory describing many of the principles of perceptual organization (eg. Koffka, 1935), they proposed a number of rules in which the brain forms mental patterns from elements of its sensory input. Although the Gestalt principles of perceptual organization were generally described first in relation to vision (see review by Wagemans et al., 2012), they are equally applicable to audition mainly in a temporal rather than spatial form. The principle of auditory perceptual organization can be categorised in the light of the Gestalt principles which are closure, good continuation, common fate and exclusive allocation.

Closure

The Gestalt principle of closure refers to a tendency to complete (close) perceptual forms. Elements tend to be grouped together if they are parts of a closed figure. Figure 1.1a shows an example how we perceive a white triangle even when it is incomplete. If part of a tone is deleted and replaced with a brief burst of random noise (Figure 1.1b, transient noise), the tone is heard to continue through the noise, even though it is not physically present (Miller and Licklider, 1950, Bregman and Dannenbring, 1973), known as the auditory continuity effect (or auditory induction or phonemic restoration). If the noise burst is absent, continuity

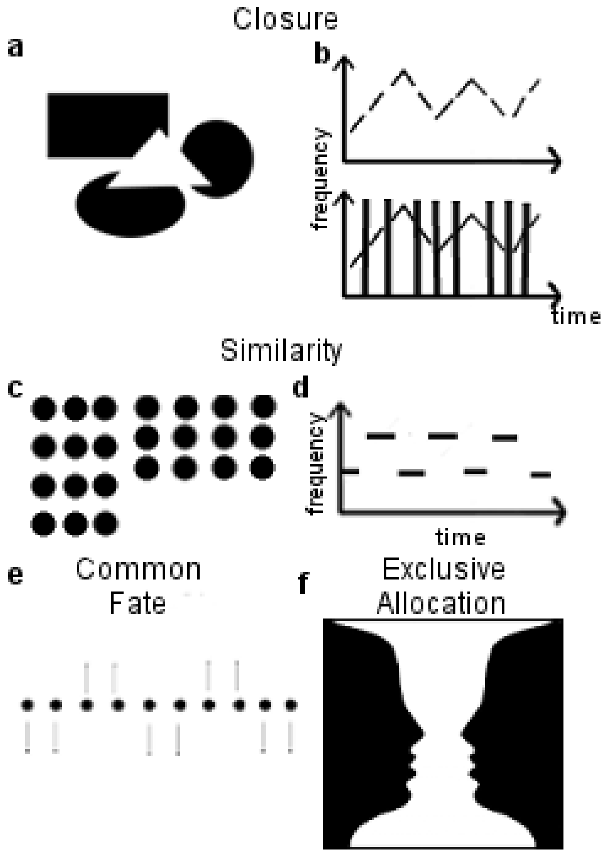

Figure 1.1 The schema for principle of perceptual organization
The principle of closure for visual perception, a, and auditory perception, b. The principle of similarity for visual perception, c, and auditory perception, d. e illustrates an example of the principal of common fate in which arrows represent the movement direction. f is the Rubin's face-vase illusion which is well-known visual example for the principle of exclusive allocation.

is abolished and a gap is heard in the tone. A similar continuity effect can be demonstrated when speech is alternated with noise bursts. In this case, the missing speech sounds are perceptually restored. Fainter sounds in speech are heard clearly when replaced by noise or louder sounds having appropriate spectral compositions (Warren et al., 1972). This is an essential perceptual "fill in gaps" mechanism since the speech of a talker is often interrupted by other sounds in the acoustic environment.

Auditory continuity phenomena are not restricted to humans, as they have been noted in cats (Sugita, 1997), birds (Seeba and Klump, 2009) and non-human primates (Petkov et al., 2003). This supports the view that these phenomena represent a widespread and fundamental perceptual-organization ability, likely to be of crucial importance for survival in diverse ecological environments where multiple sound sources are often present and need to be parsed. Samuel (1996) compared the auditory continuity effect with real words and pseudo words and found a larger amount of restoration in the real words. He also found a better restoration for words that were presented to the subjects a number of times before the test. Both results indicate that previous experience with the stimuli improves the restoration. Hence, the principle of closure is one of the rules in which the brain forms auditory objects which is innate but can be enhanced by the previous experience of the acoustic environment.

Similarity and Proximity (and Good Continuation)

The principle of similarity states that elements will be grouped if they are similar. In Figure 1.1c black circles perceived as three vertical objects and four horizontal objects because of the similarity of their arrangement. Similarly, in audition, sounds with a similar pitch, intensity, timbre or spatial location will tend to form a perceptual group. For example, van Noorden (1975) used sequences of alternation tones with frequencies A and B presented in ABA**-** or ABAB, in which A and B represent short tones with different frequency spectra (Figure 1.1d) and **-** represents a silent internal to examine the mechanism for stream segregation. When the tone interval between A and B is large, we hear two sub-sequences, the sequence of A-A-A and the sequence of B-B-B formed different perceptual streams. However, when the tone interval is small between A and B, two streams were perceived as a single stream, the whole sequence ABAB. Hence, a similarity in tone frequencies promotes perceptual auditory grouping, while dissimilarity separates them apart.

Another Gestalt principle is proximity, in parallel with the principle of similarity, which states that the closer the elements of a set are to one another, the greater is the tendency to group them perceptually. The black circles form two perceptual groups if the members of one group are closer to one another than they are to the members of the other group. In audition, acoustic components can be grouped according to their proximity in time (e.g. their onset Nakajima et al., 2000) or their proximity in frequency. For example, Bregman and Campbell (1971) presented listeners with a looping sequence of alternating high-frequency (ABC) and low-frequency sounds (123) – i.e., AB12C3- and asked participants to report the order of the tones as ABC123 (or 123ABC). When the sounds were presented slowly, subjects heard the tones in their correct sequence. However, at a faster rate of presentation, subjects failed to report them in the correct order because the high-frequency and low-frequency sounds tended to segregate into different perceptual streams. They suggested that the close proximity of the sounds in time promoted their perceptual fusion.

However, Winkler et al. (2012) speculate that it is not the raw difference, but rather the *rate of change* in sounds that alters auditory perception. If the rate of change is slower in the sounds, they will be judged more similar. This leads one to consider that in the auditory modality, the law of similarity and proximity is not separate from what the Gestalt psychologists termed as 'good continuation'. The principle of good continuation states that it is the smoothness of a change which promotes the perceptual integration of changing elements. Abrupt discontinuities are perceived as the start of something new. Bregman and Dannenbring (1973) have shown that the tendency of a sequence of high and low frequency sounds to segregate into two streams can be reduced by connecting successive sounds with frequency transitions. The principle of similarity, proximity and good continuation are based on the similar perceptual organization in which the brain forms auditory objects based on mainly on temporal rather than spatial information (change in time and frequency).

The figure-ground effect is a related example of the principle of good continuation. It is that a particular object standing out perceptually from the remainder of the scene (due to a change in the ground which promotes the perceptual grouping of the figure). When a change occurs in the acoustic environment, the attention of a listener is drawn to that change, so that it becomes the "figure" against the other sounds in the acoustic "background". For example, at a crowded cocktail party, it is possible to attend to a particular conversation while other voices form a kind of background. Similarly, when listening to a piece of polyphonic music, we attend principally to one melody at a time. Although it is suggested that the first figure ground assumption is imposed by previous experience (see review by Peterson and Skow-Grant, 2003), recent studies showed an innate, bottom-up, stimulus driven figure ground mechanism (Teki et al., 2011). Hence, this is a critical perceptual mechanism to get our attention to the new and potentially important events in the acoustic environment.

Common Fate

The common fate principle states the tendency for components of an image or a sound field to be perceived as one if they move together. In a collection of randomly moving dots, if some of the elements would begin to displace they would be perceived as part of the same object (Figure 1.1e). The same principle can be observed in the audition. In auditory displays, people tend to group sounds together if they change in pitch in a similar way. Sounds that begin and finish at the same time (or change in amplitude together and/or change in frequency together) are also likely to be perceived as related (Moore, 2012). Similarly, grouping by harmonicity can be phrased in terms of the principle of common fate. When a person speaks, the vibrations of their vocal chords generates energy at the fundamental frequency of vibration and also at integer multiples (harmonics) of this frequency. Hence, the components of a single voice can be grouped by acoustic cues that have a common spacing in frequency (i.e. harmonics) of the same fundamental.

Temporal coherence with regard to correlations over longer time windows is a version of the principle of common fate (Shamma et al., 2011). There is a tendency to be perceived as grouped together if time varying signals are coherent. People perceive the sequence of two tones in a streaming signal presented synchronously as one stream regardless of the disparity between the frequency ranges of the two tones. In addition to separation in feature space, temporal coherence between different elements in the scene is essential for segregation such that temporally incoherent patterns tend to result in a segregated percept while temporal coherence promotes integration (Elhilali et al., 2009a).

Time is an essential variable of auditory sensory inputs and time varying signals are fundamental components of acoustic information (Rosen, 1992). Some of the most important auditory grouping cues are joint temporal cues, including common onsets, offsets, and modulation profiles (Bregman, 1990, Bregman et al., 1994). Temporal structure is a crucial factor in the segmentation of complex auditory scenes through temporal coherence between elements of the auditory input (Teki et al., 2013, Elhilali et al., 2009a, Fishman and Steinschneider, 2010, Shamma et al., 2013, Micheyl et al., 2013). The principle of common fate is a critical perceptual mechanism to get potentially important temporal cues in the acoustic environment.

The Principle of Exclusive Allocation (Belongingness)

Exclusive allocation (or belongingness) refers to the principle that each element of the sensory input is only assigned to one perceptual object. The well-known visual example for this principle is Rubin's face-vase illusion (Figure1.1f). Bregman describes this process as "voting" by the grouping processes supporting one or another alternative (1990). Winkler et al. (2006) asked participants to maintain the perception of one of the two tone patterns throughout the stimulus sequences. Occasional changes violated either the selected or the alternative tone pattern, but not both at the same time. They have found that only violations of the selected pattern were represented in the auditory system in EEG recordings, suggesting that individual sounds are processed as part of only one auditory pattern at a time. However, there are some notable violations of the principle of exclusive allocation in the audition. For example, in the duplex perception, the simultaneous perception of the acoustic signal which is used for both a speech and a non-speech syllable, the same sound component can contribute to the perception of a complex sound (Whalen and Liberman, 1996) or to no object (Shinn-Cunningham et al., 2007).

To summarize, many auditory grouping principles can be described in the light of the Gestalt principles of perceptual organisation which are closure, good continuation, common fate and exclusive allocation. They are all shown to be bottom-up, innate stimulus driven principles that help the auditory system to detect and discriminate between different acoustic features. This primitive process with the basic principle of perceptual organization gives the auditory system a baseline to form the representation of auditory streams.

Schema-based processing: Auditory Stream Segregation

After the initial analysis of acoustic features, a further stage of scene analysis requires to identifying the different sound sources and grouping the sound features coming from the same source. It helps to match the incoming stream with memory-stored knowledge. It requires attention and is a knowledge-driven, top-down selection mechanism as Bregman (1990) called schema-based processes (top-down sensory process). Beyond bottom-up cues described above, there are top-down influences on the perceptual organization, including attention and prior knowledge (and contextual effect).

Attention

Tuning into a particular speaker in a noisy environment involves at least two processes: identifying and separating different sources in the environment (stream segregation) and directing attention to one behaviourally relevant task relevant one (stream selection). As studies discussed in the previous part, stream segregation by some of the aspects of acoustic features happens in an obligatory and automatic manner. Electrophysiological studies in humans provide evidence to support this view as the formation of auditory streams is preattentive (Sussman et al., 1999, Sussman et al., 2007, Macken et al., 2003, see reviews for more details,Snyder and Alain, 2007).

However, the perception of sounds as either integrated or segregated is by no means fully determined by the acoustic inputs. This is shown by the fact that perception of one and the same sequence is, to a certain extent, influenced by attentional processes. van Noorden (1975) used ABA alternating tones paradigm and found that participants could hear both streams segregated and fused depending on their active engagement. Participants' attentional focus determined the stream segregation.

Directing attention to the features of an attended stream may serve to enhance their cortical representation and thus facilitate processing at the expense of other competing input (Hillyard et al., 1973, Tiitinen et al., 1993, Alain et al., 2001, 2005, Alain, 2007). Alain et al. (2005) recorded ERP responses when participants were doing the double-vowel task during active and passive condition. In the double vowel task, a mixture of two phonetically different synthetic vowels are presented and participants were asked to identify these vowels (active condition) or passively listen to these vowels (passive condition). When two vowels are played together, one of the vowels generally sounds as a target (dominant) and the other as a distractor (non-dominant) sound. The difficulty of the task resides in identifying the non-dominant vowel, which depends on successfully breaking down the incoming vowel mixture into its elements for a comparison with representations of the each vowel (see review for more details, Alain, 2007). Double vowels are only perceived as two distinct vowels when they differ in F0, and participants' performance improved with increased difference in F0.

Alain et al. found a correlation between the improvement in performance as a function of the difference in F0 with ERP modulations, which they used as an evidence for the detection and identification of simultaneous vowels. They showed the first positive ERP modulation associated with concurrent vowel perception at about 145ms maximal over midline in active and passive condition suggesting that the first stage of auditory scene analysis extract the spectral pattern of the vowels. They showed the second ERP modulation (negative wave) at

250ms over the right and central regions of the scalp in active condition only as a second stage of the scene analysis to identify concurrent signals by grouping relevant inputs. This second modulation might be thought to be the result of mistuned harmonic detection as mistuning in complex sounds causes the perception of distinct auditory objects (Carlyon et al., 1992) however, mistuned harmonics do not generate ERP modulations (Alain et al., 2001) so the second modulation is suggested to be related to identification of vowels.

These findings support Bregman's model of two mechanisms in auditory scene analysis in which the first stage involves the analysis of acoustic signals and the second stage for top-down stream selection mechanism. There is growing evidence that attention is required to form the elements of auditory inputs to streaming together (Teki et al., 2013, Elhilali et al., 2009a, Shamma et al., 2013). However, how attention effects auditory scene analysis, it is still not clear when, where, and how attention influence the auditory stream segregation.

Prior Knowledge

Determination of whether a sung note is from a throat singer (one source) or from two simultaneous choir singers (two sources) is essentially arbitrary and depends on the listener's expectations and contextual cues (memory) and not on sensory evidence alone. The global auditory context affects how the individual sounds in the sequence are grouped in perception. The representation of acoustic signal in memory forms the basis for evaluating incoming sound information (Sussman and Steinschneider, 2006, Sussman and Gumenyuk, 2005). For example, listeners' ability to segregate concurrent sounds based on harmonicity is modulated by long term experience (in musicians, Zendel and Alain, 2009)

The use of prior knowledge in auditory perception is particularly evident in speech communication. A sentence's final word embedded in multi-talker babble is more easily detected when it is contextually predictable (Pichora-Fuller et al., 1995). Several linguistic factors also utilise prior experience influences over speech segregation and intelligibility, including semantic and syntactic (Miller et al., 1951, Borsky et al., 1998) and lexical (Cooper et al., 1978) influences. Discussing the nature of these influences is beyond the scope of this review; however, they emphasize the importance of prior knowledge in speech segregation (see reviews for more details, Zion Golumbic et al., 2012).

Winkler et al. (2003) investigated the perceptual effects of contextual manipulations on auditory grouping by using target-detection and order judgment tasks when attention was directed away from the sounds. They found good correspondence between the effects of auditory context in active and passive situations, speculating that a substantial part of contextual effect in audition can be seen even in the early grouping processing. Similarly, the effects of a prior adapting stimulus and effects of prior perception during an ABA stream segregation task were shown in ERP studies (Snyder et al., 2009).

Discussion and Conclusion

Although the above description might seem to suggest that objects are constructed through a hierarchy of processing, first grouped based on the acoustic structure and then organized across longer temporal scales, the truth is more complex. The state of the listener, from prior knowledge about a scene's content to the basic level of acoustic feature analysis, influence the perceived content of an object (Shinn-Cunningham, 2008). Higher-order features and top-down attention can alter how streams form. Rather than a hierarchical processing structure, objects are formed through interactions between processes that mutually influence one another, rather than through a sequence of processing stages (Figure 1.2). The ultimate perceptual organization of the scene depends on all evidence (Elhilali et al., 2009b, Shinn-Cunningham, 2008). Object continuity, as an example, might enhance auditory attention to the new object (Best et al., 2008).

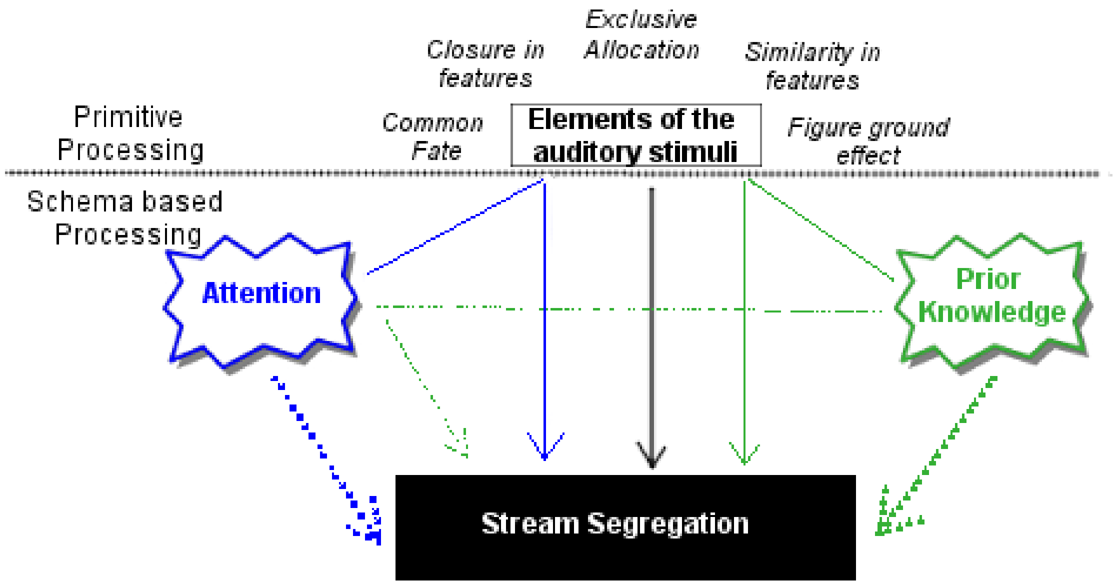

Figure 1.2 The schematic model of auditory scene analysis (see text for details)

One of the biggest limitation of the streaming literature is that studies investigating complex scene processing are focused on the influence of frequency, level or spatial location on tone-based stream perception (Bregman and Campbell, 1971, Carlyon et al., 2001, Ihlefeld and Shinn-Cunningham, 2008, Shinn-Cunningham et al., 2007, Sussman et al., 1999, Sussman et al., 2007, Elhilali et al., 2009a, Shamma et al., 2011) and how other acoustic features contribute to auditory streaming has little attention (Alain et al., 2001).

To conclude, these principles are among the fundamental properties of the perceptual system, providing the basis of our ability to make sense of the sensory signals under the influence of our focus and knowledge of our environment. Primitive grouping effects generally conform to the Gestalt principles of closure, good continuation, common fate and exclusive allocation. However, the analysis of acoustic signals relies upon temporal information and therefore perceptual grouping in the auditory domain is also crucially dependent on temporal analysis. Furthermore, attentional focus and prior knowledge of stimuli can bias the auditory perceptual grouping, suggesting that there is no sequence of processing stages but interactions across all factors consider all possible alternative groupings (Bregman, 1990, Winkler et al., 2003).

Part II: AV Integration in Auditory Scene Analysis

The aim of this thesis is to investigate the impact of visual information on the bottom-up scene analysis mechanisms as described above. In order to comprehend the visual impact on bottom-up processing, in this part, I will briefly review the fundamental findings of studies on the different feature based AV cross modal correspondence and then focus on AV integration on auditory scene analysis. How visual cues help us listen to degraded sounds, predict the auditory signal timing and segregate sounds from a mixture will be briefly mentioned and critically discussed for primitive and schema-based processing.

AV Cross Modal Correspondence

The term ‘’cross modal correspondences’’ has been used over the years by researchers in order to refer to our brain’s tendency to systematically associate certain features of stimuli, either physically present or merely imagined, across the senses. The first example of audio visual correspondences comes from is coming from bouba/kiki effect (Köhler, 1929, Ramachandran and Hubbard, 2001). People showed a preference for pairing the speech sound ‘Bouba’ with curvy shapes, and a preference for ‘Kiki’ for sharp shapes, suggesting a non-arbitrary mapping between speech sounds and the visual shape of objects. More recent studies showed that such shape-symbolism effects can arise independently of any associations between shape and sound that may be present in orthography or of any cultural influences (Bremner et al., 2013).

AV cross modal correspondences have been documented between many different pairs of AV stimulus dimensions. The summary of studies for different feature based cross modal mapping is listed in table 1.1. The table includes loudness, pitch and timbre for the auditory domain with different visual cues. The varied findings in these studies showed that same auditory feature corresponds in a different way to visual features. For example, higher pitch sounds impact visual perception to make a visual stimulus appear brighter in luminance (Marks, 1987) but smaller in size (Walker and Smith, 1985). Also, AV cross modal correspondence is task dependent. The pitch of sound did not correspond with visual elevation in a speeded detection task (Klein et al., 1987) but corresponded with higher visual elevation in a speeded classification task (Evans and Treisman, 2010).

Such AV cross modal interactions result in a benefit in behavioural performance. Gallace and Spence (2006) presented two masked grey disks at fixation, one after the other and asked participants to respond either as to whether the second variable-sized disk was larger or smaller than the first standard-sized disk (Experiment 1) or to whether the two disks were the same size or not (Experiment 2). On the majority of trials, a sound was presented in synchrony with the second disk (otherwise, no sound was presented).

The relative frequency of the sound (300 or 4500 Hz) was either congruent or incongruent with the size of the second disk (relative to the first). They showed that in both experiments, participants responded significantly more rapidly and more accurately on the congruent AV cross modal trials in which a high-frequency sound was presented with a small disk, than on the incongruent trials in which a low-frequency sound was coupled with a small disk. Similarly, people detected brief low-intensity sounds only when they were paired with a simultaneous light in a one-interval signal detection task (Lovelace et al., 2003). Evidently, behavioural performance benefits from feature based AV cross modal correspondences.

Some AV cross modal correspondences have been suggested to be more ‘natural’ and shown at very early stages of human development. For example, the correspondence between pitch-elevation and pitch-brightness are natural AV correspondence. Hence, their occurrences are more strongly internalized than others that may only develop later. Early developed AV cross modal correspondence are pre-attentive (Spence and Deroy, 2012).

Table 1.1 Summary of studies on audio-visual cross model correspondences

AV Cross modal Correspondence	*Studies*	*Stimuli*
Loudness and Visual Cues	(Marks, 1974, Marks, 1987)	Louder auditory tones with higher light intensities
	(Giannakis and Smith, 2001)	Louder auditory tones with higher colour saturations
Pitch and Visual Cues	(Melara, 1989, Parise and Spence, 2012)	Lower auditory pitch with curvy shapes and higher auditory pitch with sharp angular shapes
	(Marks, 1987)	Higher auditory pitches with brighter visual stimuli
	(Walker and Smith, 1985, Maurer and Mondloch, 2004, Parise and Spence, 2008, 2009, Parise and Spence, 2012, Evans and Treisman, 2010)	Higher auditory pitches with smaller sizes
	(Pedley and Harper, 1959, Bernstein and Edelstein, 1971, Melara and O'Brien, 1987, Klein et al., 1987, Miller, 1991, Ben-Artzi and Marks, 1995, Patching and Quinlan, 2002, Gallace and Spence, 2006, Eitan and Timmers, 2010, Evans and Treisman, 2010, Chiou and Rich, 2012, Mossbridge et al., 2011)	Higher auditory pitch with higher visual elevation
	(Mossbridge et al., 2011, Fernández-Prieto et al., 2012)	Ascending pitch with higher positions
	(Evans and Treisman, 2010)	Higher auditory pitch with high spatial frequency
Timbre and Visual Cues	(Köhler, 1929, Ramachandran and Hubbard, 2001, Maurer et al., 2006, Parise and Spence, 2012)	Curvy shape for nonsense word "Baluba" or "Bouba" and a sharp jagged shape for nonsense word "Takete" or "Kiki" (sine waves with the curvy shape and square waves with the jagged one)
	(Fernay et al., 2012)	Speaker's gender determined the size of the shape (larger shapes for male voice)
	(Giannakis, 2006)	More visual texture contrast with higher sound brightness, more visual texture periodicity with more auditory dissonance
	(Adeli et al., 2014, Schloss et al., 2012)	Soft timbres with blue, green or light grey rounded shapes, harsh timbres with red, yellow or dark grey sharp angular shapes

One of the popular well-known example of AV integration that occurs automatically regardless of the spatial location and attentional manipulations is the audio-visual ventriloquism effect. The perception of the location of a sound source can be shifted by the presence of a temporally coincident but spatially disparate visual cue (Howard and Templeton, 1966, Bertelson et al., 2000). This kind of AV integration operates in an automatic manner without observer's conscious control. Even with the exact knowledge of location of auditory and visual stimuli, and over distances too large to produce absolute spatial capture, there is still a clear bias in auditory localization towards the visual stimulus.

According to Ernst and Bulthoff (2004), under Bayesian assumptions a given modality is not always the best choice for resolving all types of problem (e.g., vision has high spatial acuity so it is good for spatial localization but has poor temporal resolution); rather each modality is most likely to provide the best sensory estimate for resolving certain types of problems, but another modality can provide a better estimate (be more appropriate) under certain circumstances. For instance, spatial localization might be more accurately performed based on auditory than on visual cues in darkness. Inputs from different modalities are weighted according to their reliability (Alais and Burr, 2004, Talsma et al., 2010) and are combined with various perceptual factors (e.g., temporal and spatial coincidence). Since the visual spatial resolution is better than the spatial resolution of the auditory system, the underlying mechanism of audio-visual ventriloquism effect might be more 'natural' and internalized, which results in the automatic preattentive process.

On the other hand, the other popular example of AV integration, McGurk effect, in which pairing conflicting visual cues and auditory speech-related cues results in reports of a novel auditory percept reflecting the synthesis of the two sensory channels (McGurk and MacDonald, 1976), is modulated by attentional and cognitive load (see review, Navarra et al., 2010a). Since McGurk effect is a speech related phenomena, the underlying mechanism of linguistic/semantic coding might be different than the mechanism for the audio-visual ventriloquism effect.

As it is seen in above examples, underlying mechanisms of AV integration on auditory processing are task-dependent and might have different mechanisms. This thesis will focus on the role of AV integration in auditory scene analysis rather than in perception more generally.

AV integration in auditory scene analysis

I will briefly outline studies providing evidence that visual cues help auditory scene analysis by enhancing our perception of degraded sound, by providing extra temporal information about the incoming auditory signal. Limited studies on how visual information helps us to segregate streams in a sound mixture studies will be critically discussed. The effect of attention and prior knowledge of stimuli on AV integration is also included to give a better understating of the influence of AV integration in auditory scene analysis.

Visual cues as complementary source

The literature of AV integration on auditory scene analysis is heavily biased by visual speech. Recognizing speech in a noisy environment is easier when the speaker's face is visible (Dodd, 1980, Sams et al., 1991). One explanation given for improvement in speech perception is that lip-reading provides relatively undistorted cues for place of articulation when transmitted acoustic information are severely degraded by noise, reverberation or combinations of these elements (Summerfield, 1987, Sumby and Pollack, 1954). The influence of visual information in speech includes lip movements as well as head, jaws and eyebrows movements and, facial, hand and body gestures (Biau and Soto-Faraco, 2013). Close temporal correspondence between the area of the mouth opening and the acoustic envelope (Chandrasekaran et al., 2009, Grant and Seitz, 2000) supported the idea that visual information not only provides linguistic information but also provides extra temporal information. Bernstein et al. (2004) found that speech detection in a noisy environment could be enhanced by synthetic and non-synthetic visual cues that were synchronized with sound intervals. They presented an auditory /ba/ sound either without any visual cues or with the original video recording, or dynamic figures. Participants were found to detect /ba/ sound the easiest when it was paired with original video recording at the hardest signal to noise ratio. Other synthetic visual stimuli increased auditory detection relative to the auditory only condition with no significant differences between the different visual stimuli.

This visually-induced improvement in speech perception supports the idea that speech perception is dominated by auditory perception, visual cues are only beneficial when the acoustic signal is physically degraded by external noise and has a benefit on hearing-impaired listeners (Grant et al., 1998) or even when listening to a second language (Navarra et al., 2010b). However, the McGurk effect showed that the role of vision in speech is not restricted to a complementary source of information that helps to improve perception when the auditory signal is weak, but it can also change auditory perception dramatically.

Visual cues to predict auditory signal timing

Another explanation for the ability of a visual stimulus to improve speech recognition is that visual cues provide important temporal information about when to listen. Visual cues of mouth articulation precede auditory signals by ∼100–150ms (Chandrasekaran et al., 2009)and remain perceptually linked at offsets of up to 200ms if the auditory signal is temporally coherent with the visual signal (Van Wassenhove et al., 2007). This time window is enough for visual cues to have an impact on the processing of upcoming auditory signal. However, the temporal relationship between auditory and visual cues is found to be more complex than the notion that vision may lead the audition. Schwartz and Savariaux showed syllables sequence result in varying audio lead (~40ms) and visual lead (~200ms) to detect intersensory synchrony (2014). Although characterization of temporal integration window is limited, it is clear that the temporally coherent visual input is critical for auditory visual integration so that auditory processing.

Visual cues to stream segregation

Devergie et al. (2011) used an ABA- paradigm and asked participants to detect a change in the presentation rate of French vowel sequences alternating in pitch while watching short articulation videos of these vowels. Integrating all the vowels of the sequence into a single stream would result in better change detection (Gaudrain et al., 2007). Participants' detection performance was enhanced with coherent vowel videos. This finding indicates a visually-induced improvement in speech due to visual enhancement of bottom-up auditory streaming segregation without giving a clear understanding of the visual effect on stream segregation. It might be either linguistic information from articulation videos facilitating target to be separated from other acoustic signal and/or temporal information from these videos assist people to distinguish auditory streams. It also is possible that an interaction between linguistic information and temporal information is enhancing stream segregation.

Rahne and his colleagues (2007, 2008) examined the visual influence on stream segregation by generating sequences of pure tones (no linguistic information) to induce different perceptual organizations. They used alternating low and high frequency tones. While the high-frequency tones were presented in random order, the low-frequency tones together formed a sequence composed of a repeated pattern of three tones rising in pitch. This pattern was sometimes replaced by a deviant pattern of three tones decreasing in pitch. In addition, every third tone in the overall sequence was more intense (+15 dB).

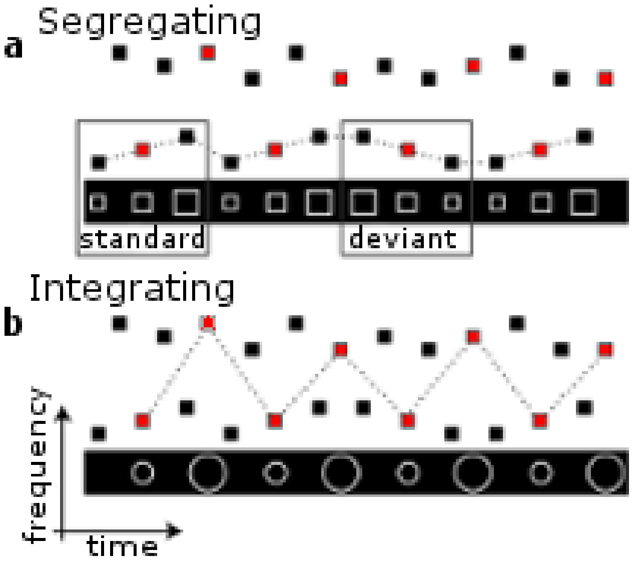

Figure 1.3 Schematic presentation of the stimuli used in Rahne et al. (2008)

a shows segregated streams (two streams; high and low frequency tones), composed of a repeated pattern on three tones rising in pitch (standard) and deviants are a pattern of three tones decreasing in pitch with a movie of square changing in sizes. **b** shows integrated stream with same tones as in a with a movie of open circles changing in size.

Detecting the deviant in a stream was assumed to be easier when the streams were perceived as segregated two streams. Large frequency differences promoted streaming into two streams, whereas the intensity changes promoted the integration of information across frequency space such that the louder tones formed one sequence and softer another. The perceptual organization was assumed to be based either on the frequency difference (lower tones into one stream and the higher tones into another stream) or the intensity difference (louder tones as one stream and the softer tones as another). Auditory sounds were presented with two temporally coherent visual conditions: open circles or squares changing in sizes. The synchrony between the change in size of circles and the change in intensity of tones was suggested to promote integration between two streams. (Figure 1.3a). While the synchrony between the change in size of the square and the change in frequency of tones was suggested to promote the segregation of tones to two streams (Figure 1.3b). Authors recorded ERP responses while subjects passively listened/watched these sequences of auditory and visual stimuli. By using the response to the deviant change in the low frequency stream as an index for how well segregated the tones sequences were, they found larger deviant responses only when the visual stimuli coincided with the segregated perception. Thus, the visual cue was found to improve the ability to perceptually organize two streams of ambiguously organised tones.

Rahne and Böckmann-Barthel (2009) extended these findings to investigate whether a visual cue enhances the segregation or integration of sound elements into streams in an inherently stable (unambiguous) condition using the same AV stimuli. Participants were asked to attend to the visual stimuli and to report which sound organization is more dominant (frequency-based segregation or intensity-based grouping) by pressing one of two buttons on a keypad. They specifically instructed participants to change buttons if the organization changed. Therefore, they modified the tone sequence of the previous ERP experiment by increasing the frequency differences of the ABA tone sequence above the temporal coherence boundary to obtain a stable segregated perceptual organization (wide frequency distance). In another condition, they lowered the frequency differences below the fission boundary to obtain a stable integrated perceptual organization (narrow frequency distance). Visual cues had no influence on the perceptual organization when auditory streams were separated by a wide frequency distance whereas when auditory streams were separated by a narrow frequency distance, larger deviant responses were found for visual cues aimed to segregate streams. They concluded that visual cues might alter perceptual organization by enhancing segregation but not in integrating elements into an inherently stable auditory organization.

However, in their AV paradigm, visual stimuli had information about the perceptual organization of the auditory scene, since each auditory stream matched with one type of visual stimulus. The visual stream provided a cue for which stream to follow, it is not clear whether differences in the perceptual organization are due to AV cross modal integration or biasing listeners' decision making process. Stimulus-induced changes in neural response patterns are a common confound in studies attempting to bridge the gap between neural activity and perception (Logothetis and Schall, 1989). To my knowledge, previous literature does not show clear evidence that visual stimuli can influence the bottom-up process of auditory scene analysis.

The effect of attention on AV interaction

In order to focus on relevant information and ignore what is irrelevant, the human mind is equipped with a selection mechanism accomplished by the cognitive function of attention. Since AV integration can occur across various stages of auditory processing, recent findings point to a complex interplay between attention and AV integration (Talsma et al., 2010).

The interaction between AV integration and attention has previously been explained both in terms of bottom-up and top-down mechanisms. Several ERP studies revealed that the integration of auditory and visual stimulus properties into a multisensory object may take place relatively early on in the processing stream (Giard and Peronnet, 1999, Molholm et al., 2002). This finding suggests that AV integration is a process that occurs largely without conscious effort. In addition, many behavioural studies have provided evidence for the hypothesis that integrating visual and auditory stimuli serves the purpose of enhancing perceptual clarity (Calvert et al., 1997, Stein et al., 1996). Similarly, according to the account of pre-attentive bottom-up interaction, the integration between auditory and visual stimuli occurs spontaneously at the very early stage of processing, then captures attention. The audio-visual ventriloquism effect occurs regardless of attentional manipulations (Bertelson et al., 2000, Vroomen et al., 2001) and further enhances spatial attention to speech sounds (Driver, 1996), suggesting that the preattentive automatic AV integration captures attention.

Alternatively, attention can limit or boost AV integration at relatively early processing. Attending to a unisensory feature of a cross-modal stimulus can direct attention to features in the other modalities (Molholm et al., 2007) and attentional focus affects AV integration by reshaping the unisensory weights across auditory and visual stimuli (Oruc et al., 2008). Similarly, the McGurk effect is found to be reduced when people are asked to attend a secondary task, suggesting attentional load lessens AV integration (Alsius et al., 2007).

The effect of prior knowledge on AV interaction

Since the representation of familiar sounds forms the basis for evaluating incoming sound information (Sussman and Steinschneider, 2006), the prior knowledge of the sounds might affect AV interaction. For example, when replacing articulatory lip-movements by non-speech visual stimuli in a McGurk paradigm, Sams et al. (1991) found no evidence of the McGurk effect in an ERP study. Similarly, participants were more likely to integrate AV signals into a McGurk effect after exposure to audio-visually congruent than incongruent speech signals (Nahorna et al., 2012).

Although listeners' ability to segregate concurrent sounds based on harmonicity is modulated by long-term musical experience as well as musical training (in musicians, Zendel and Alain, 2009), musical expertise does not enhance the use of visual input on auditory processing. Marozeau et al. (2010) investigated the effect of visual cues on stream segregation across musicians and non-musicians. They found that when no visual cue was present, musicians generally rated the melody segregation as less difficult than non-musicians. However, when the visual cue was present, difficulty ratings for musicians and non-musicians were very similar. It was found that long term musical training did not increase the advantage gained from the visual stimuli, but the effect of prior knowledge or exposure of AV stimuli effect on stream segregation is still not clear.

Discussion and Conclusion

The studies discussed above show a role for arbitrary visual cues in auditory scene analysis, but it remains unclear how this influence operates, and it is difficult to extend these conclusions to auditory processing in more ecological situations. The literature on AV perceptual organization on auditory scene analysis has either used pure tones or focused on lip reading cues for speech comprehension and the role of such ‘content/semantics-related’ associations has been the target of many studies and it is well acknowledged in the AV integration literature (Doehrmann and Naumer, 2008). Nonetheless, unlike other low-level stimulus characteristics (e.g., timing, position, etc.) these findings on the role of visual cues in scene analysis rely on pre-existing ‘prior knowledge’ and it is difficult to exactly know whether/how the participants make use of this knowledge to strategically address and solve any specific task. More studies are necessary to expand our understanding of AV integration on the primitive process underlying auditory scene analysis and specifically in the case where the visual stimulus does not explicitly convey information about how the auditory scene should be segmented (Figure 1.4).

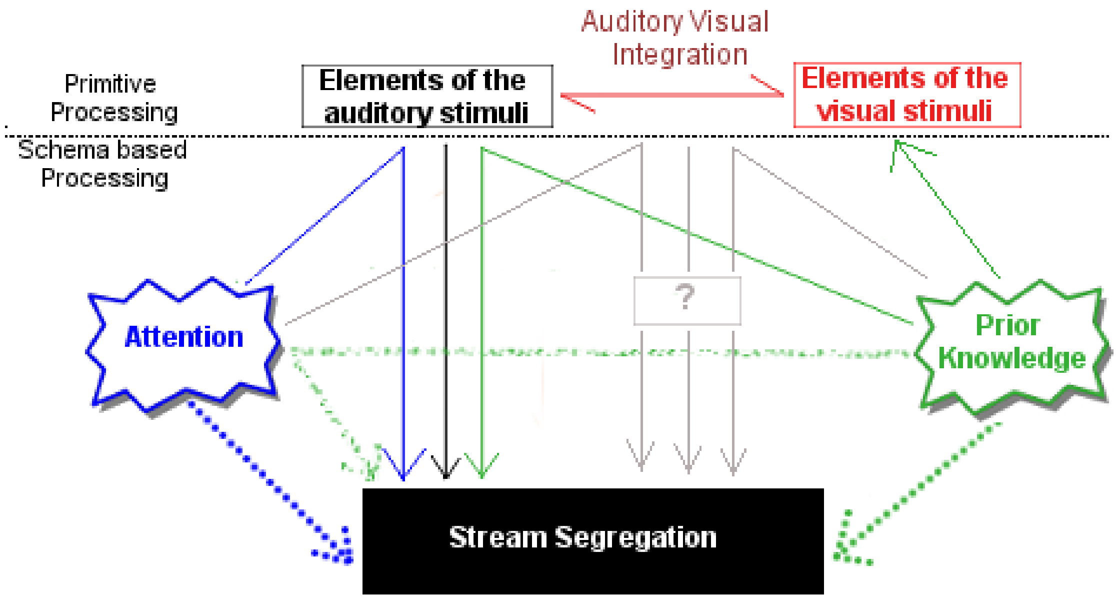
Auditory Visual Integration
Primitive Processing
Elements of the auditory stimuli
Elements of the visual stimuli
Schema based Processing
Attention
?
Prior Knowledge
Stream Segregation

Figure 2.4 Schematic representation of the AV integration on stream segregation

As it is demonstrated in Part I, the temporal coherence of sound features is one of essential step in auditory perceptual scene analysis. Similarly, for discrete stimuli, temporal coincidence is a key determinant of the likelihood of integrating information across different modalities. Temporal coherence may facilitate binding of auditory and visual stimuli into single cross modal objects which promotes auditory scene analysis.

To conclude, AV integration aids our analysis of the acoustic environment, more studies with non-speech complex stimuli are required for better understanding of the effect of visual information on auditory stream segregation which this thesis will examine as its first scope. The next section will review neural mechanisms of AV integration as the second scope of this thesis is to determine whether the integration of visual information in auditory cortex contributes auditory scene analysis.

Part III: Neural Mechanism of AV Interaction in auditory scene analysis

In many everyday situations, our senses are bombarded by many different sensory signals. To gain the most veridical, and least variable, judgment of environmental stimuli, we need to combine individual unisensory perceptual estimates that refer to the same object, while keeping separate those estimates belonging to different objects. Perception of the environment requires integration of sensory information across the senses, but how our brains combine such information is still poorly understood.

The earliest stages of cortical sensory processing were long thought to be unimodal, with multisensory processing restricted to dedicated convergence areas such as the superior temporal sulcus (STS) and prefrontal cortex (PFC) (Mesulam, 1998). However, the past decade has seen new anatomical and functional evidence for multisensory interactions even at the level of the primary sensory areas. For example, there is growing evidence for multi-sensory processing occurring in primary auditory cortex (A1) as a consequence of either visual stimulation (Calvert et al., 1997, Budinger et al., 2006, Bizley et al., 2007, Kayser et al., 2008) or somatosensory inputs (Brosch et al., 2005, Foxe et al., 2002). There are even direct connections between A1 and primary visual cortex (V1) (Falchier et al., 2002, Rockland and Ojima, 2003, Bizley et al., 2007), although the main role of such connections remains unclear.

The second aim of the thesis is to investigate the neural correlates of the impact of visual information on the analysis of acoustic scenes. The region of the interest in this study is auditory cortex since it is a potential neural substrate for bottom-up processing of auditory scene analysis. This part will review the recent findings on the representation of auditory stream segregation and AV integration in auditory cortex.

Auditory stream segregation in AC

The neural mechanisms underlying simultaneous sound segregation are poorly understood. Several lines of evidence support a role for auditory cortex in concurrent sound segregation. Single unit recording studies from primates showed that essential features of auditory organization phenomenon in humans can be seen in neural responses recorded in AC while awake primates listen to ABA tones (Micheyl et al., 2005) and that spectral and temporal information is sufficient for extracting the F0s of two simultaneously presented harmonic complex tones at the level of AC (Fishman et al., 2014). Hence, the role of AC might be automatically analysing and grouping the acoustic elements in the sound mixture (Nelken, 2004, Nelken and Bar-Yosef, 2008).

Furthermore, there are recent findings from electrophysiological recordings from AC of epileptic patients on the role of AC in selective attention to particular streams in a sound mixture. Patients were presented two distinct streams (AM tones separated by two octaves in different modulation rates) and asked either to detect noise bursts (to distract their attention away from the streams) or report the spatial change while attending to a given stream which had an extra section changing in spatial direction (Bidet-Caulet et al., 2007). Selective attention enhanced steady-state responses in the Heschl's gyrus (around A1) whereas enhanced evoked transient responses and induced gamma oscillatory activity were found in the lateral superior temporal gyrus (secondary auditory areas). Such different neural response properties suggested an attentional effect on different neural mechanisms which are related to acoustic features (bottom-up processing/ grouping of acoustic features), occurring at different latencies and regions in AC. The analysis of the sound mixture might rely on the interaction of several neurophysiological mechanisms.

The role of selective attention is often thought of as operating as a gain control mechanism, enhancing the internal representation of the attended stream and suppressing the representation of the ignored streams (Lee et al., 2014). Even at its initial stage in AC, in A1, auditory stream segregation appears to be modulated by attention to specific features of auditory stimuli, such as frequency and time (Atiani et al., 2014, Atiani et al., 2009, Fritz et al., 2005, Fritz et al., 2003). The cortical representation of sound mixture in AC does not merely reflect the external acoustic environment but instead gives rise to the perceptual aspects relevant for the listener's attentional focus (Ding and Simon, 2012, Mesgarani and Chang, 2012, Golumbic et al., 2013). However, the ability to selectively attend to one speech stream in the midst of other competing streams critically depends on how well they are perceptually segregated from one another (Shinn-Cunningham and Best, 2008).

All discussed studies provided evidence that AC has a role in the bottom-up processing of auditory stream segregation either by analysing acoustic features and/or grouping of acoustic features. Although the role of AC in auditory scene analysis is not clear, it is a potential neural substrate to investigate the impact of visual information on the bottom-up scene analysis mechanisms.

AV interaction in AC

The earliest evidence for visual input in AC came from Calvert et al.'s fMRI study (1997). They showed that silent lip reading which does not have any auditory input but auditory related visual input is sufficient to activate auditory cortex. Visual stimuli can both drive and modulate neural activity in primary and non-primary auditory cortex (Bizley et al., 2007; Chandrasekaran et al., 2013; Ghazanfar et al., 2005; Kayser et al., 2008; Perrodin et al., 2015). Visual input in AC might originate rather from the lateral involvement of other cortices and/or feedback involvement of higher areas (Bizley et al., 2007, Budinger et al., 2006) and might modulate neural responses by oscillatory activities(Kayser et al., 2010, Kayser et al., 2008).

Anatomical sources of visual input

Neurophysiological studies revealed a direct link between auditory cortices and visual cortices and higher cortical areas including prefrontal cortex (PFC) and posterior parietal cortex (PPC). Anatomical tracing has revealed that AC receives direct inputs from several visual cortical areas (Bizley et al., 2007, Budinger et al., 2006). Direct projections from the auditory cortical core (Mongolian gerbil: Budinger et al., 2000, macaque: Falchier et al., 2002) and belt areas (macaque: Rockland and Ojima, 2003) to visual cortical areas V1 and V2 were established. In order to examine both directions, Budinger et al. (2006) injected the sensitive bidirectional neuronal tracer fluorescein-labeled dextran (FD) into AI of Mongolian gerbils and investigated anatomical substrate of multisensory inputs into AI. They found that of the inputs originating from outside of the auditory pathway, 40 % of retrogradely labeled cell bodies were in cortical areas and 60% in subcortical areas. 82% of retrogradely labeled cell bodies in the cortical areas were in multisensory areas whereas 8% of in primary visual and 10% in primary somatosensory cortex, suggesting a direct connection between AI and non-auditory sensory and multisensory areas.

In a follow up study, Budinger et al. (2009) used tetramethylrhodamine-labelled dextran (TMRD), which was simultaneously injected into different frequency regions of the gerbil's AI. They examined the distribution of terminal field in V2 and compared this to the distribution of the retrogradely labelled cells in the experiment above. They showed that anterogradely labelled axons (i.e. those areas that are directly innervated by A1) were also found in the same cortical areas as the retrogradely labelled cell bodies and there were virtually no retrogradely labelled cell bodies outside the area covered by anterogradely labelled axons. Putting all the findings together, the spatial pattern of retrograde labeling supported a rather high degree of reciprocity in the connections between AI and V2 and suggesting that there is also feed forward like input of AI into V2 (Budinger et al., 2008).

The results of Bizley et al. (2007) in the ferret are consistent with Budinger et al. data (2006; 2007). Bizley et al. combined electrophysiological recordings and neuroanatomical tract tracing in ferret auditory cortex. Unisensory auditory, visual and AV neurons are found to be widely distributed in ferret AC. ~15% of units in primary auditory cortices were found to respond only to visual stimuli and this proportion was larger in secondary areas. Neural tracer injections in the EG revealed direct inputs from VC into AC, indicating a potential source of origin for these visual responses. Retrogradely labeled cells were found in visual areas 17, 18, 19, and 20, as well as the suprasylvian cortex (SSY). V1 projects sparsely to A1, whereas higher visual areas innervate auditory areas in a field-specific manner. Similarly, it is shown that posterior tonotopic non-primary fields are innervated by area 20 and anterior non-tonotopic, non-primary areas receive innervation from SSY. Consequently, there are direct connections between not only primary areas of visual and auditory cortices, but also non-primary areas of visual and auditory cortices in a field-specific manner. These anatomical connections might be the underlying reason of the modulated neural activity in AC by visual input as discussed below.

Neuronal spiking response in AC modulated by visual input

Natural sounds when accompanied with a corresponding video, visual stimuli are found to increase the reliability of the auditory cortical responses in AC. Kayser et al. (2010) showed that when monkeys are presented with naturalistic sound stimuli, accompanying visual stimulation, the mean firing rate of neurons in A1 was reduced. Moreover, the inter-trial variability of spike trains is greatly reduced, thus enhancing mutual information between stimuli and spiking patterns. This effect was significantly stronger when the auditory and the visual input were temporally coherent. Therefore coherent AV stimulation improves the reliability with which sounds are represented in AC.

Visual stimuli can speed up auditory processing by reducing timing uncertainty. Chandrasekaran et al. (2013) trained monkeys to detect visual only, auditory only and AV presentations of monkey vocalizations and showed shorter onset response latency of auditory cortical spiking activity when vocalizations accompanied with synthetic monkey agents articulating mouth movements than when only vocalizations were presented.

An investigation solely focused on spiking activity will have likely missed many of the multisensory effects. Even in well-established multisensory cortical areas, such as the superior temporal sulcus, only 23% of visually responsive single neurons are significantly influenced by auditory stimuli (Barraclough et al., 2005). Thus, establishing the relationship between spiking activity and LFPs will be particularly essential in revealing the cortical mechanisms of multisensory integration.

Sub-threshold effects of visual stimuli on auditory processing: resetting the phase of neural oscillations

Lakatos et al. (2005) examined the oscillatory structure of the EEG recorded in Auditory Cortex. They looked at the laminar profiles of synaptic activity and multiunit activity, both spontaneous and stimulus-driven, in primary AC of awake macaque monkeys. They showed the EEG hierarchically organization in which delta (1–4 Hz) phase modulates theta (4–10 Hz) amplitude, and theta phase modulates gamma (30–50 Hz) amplitude. They suggested that baseline excitability of neurons is controlled by these oscillatory hierarchies.

Lakatos and his colleagues (2007) examined the multisensory interaction in A1 by using multiunit and field potential recordings from macaque monkey and found that auditory clicks activated A1 in granular layer followed by the supra- and infragranular layers (a typical feedforward pattern) while somatosensory stimulation activated the supragranular layers faster than the other cortical layers. Also, while clicks elicited spiking activity in all layers, somatosensory input did not in any layer. They concluded that somatosensory stimulation alone does not drive auditory neurons over their action potential threshold but perhaps modulate auditory inputs by oscillatory activities.

They varied the timing of auditory clicks relative to somatosensory stimulation. They found that simultaneous representation of sensory inputs led to a nonlinear enhancement of activation in the supragranular layers and less activation in the infragranular layers, with no change in the granular layer. Co-presentation of the somatosensory and auditory stimuli resulted in a super-additive multisensory interaction at moderate auditory stimulus intensities. This interaction was largest when stimuli were presented simultaneously. They then examined the oscillatory phase distribution pre- and post-somatosensory stimulation and found that in the gamma, theta, and delta oscillations, phase distributions were essentially random before but were highly concentrated after somatosensory stimulation. They concluded that somatosensory inputs appear to reset the phase of ongoing neuronal oscillations, so that accompanying auditory inputs arrive during an ideal, high-excitability phase, and produce amplified neuronal responses.

Schroeder et al. (2008) proposed a similar phase-resetting-based mechanism in which the visual amplification of speech perception is operating through efficient modulation or ‘‘shaping’’ of ongoing neuronal oscillations. How phase and frequency flexibility of the delta oscillation are shaped by auditory stimuli has shown in Lakatos et al. study, and the findings of how onset of a speech resets the phase of the ongoing cortical oscillations in the AC (Ahissar et al., 2001), how the frequency of cortical oscillation adapts to the rate of stimulation (Lakatos et al., 2005) and how the phase of theta oscillation tracks speech and predicts speech intelligibility (Luo and Poeppel, 2007) are consistent with Schroeder et al.'s, proposal.

Accordingly, evidence for neural entrainment to the slow amplitude fluctuations by visual input in tone sequences (Lakatos et al., 2005, 2007), speech (Luo and Poeppel, 2007), natural sounds (Kayser et al., 2009, Ng et al., 2012) and frequency modulated auditory stimuli (Henry and Obleser, 2012) has been demonstrated.

Recently, Luo, Liu and Poeppel (2010) looked at the phase resetting oscillation mechanism within the different frequency bands. Participants were asked to passively watch an audio-visual movie during a MEG experiment. They reported that synchronized coordination of information across visual and auditory streams is carried by delta theta phase modulation across early sensory areas. In single trials, the phase of the 2–7 Hz delta and theta band responses carries strong and usable information which constructing the temporal structure of the stimulus in both sensory modalities. They suggested that delta-theta phase modulation across early sensory areas plays an important ‘‘active’’ role in continuously tracking naturalistic audio-visual streams as well as carrying dynamic multi-sensory information.

Discussion and Conclusion

To sum up, there is enhanced activation in the AC during AV stimulation. It might result in direct lateral input from other sensory cortices and/or feedback involvement of higher areas by resetting the phase of neural oscillations.

The studies discussed above, clearly illustrate a role for AC in auditory scene analysis and the presence of visual innervation. However, how visual cues influence auditory processing in AC and the role of AC in AV integration on auditory scene analysis is not clear. More studies are required to understand how visual input effect auditory scene analysis and to expand our understanding of the role of AC in AV integration on the analysis of auditory scenes.

Part IV: Scope and Hypotheses

Our understanding of AV integration on bottom-up processing of stream segregation is limited (Figure 1.4). The aim of this thesis is to investigate the effect of visual information on auditory stream segregation to fill this gap in the literature with complex auditory stimuli with no linguistic information and task irrelevant visual stimuli.

First, I examined psychophysically whether temporal coherence between auditory and visual stimuli was sufficient to promote auditory stream segregation (Chapter II). I speculated that modulating a visual stimulus coherently with one auditory stream in a mixture would cause the temporally coherent auditory and visual stimuli to bind together. I hypothesised that there would be a consequent improvement in performance in an auditory selective attention task either when the target auditory stream was bound with the visual stimulus or when an auditory stream regardless of the attentional focus was bound with the visual stimulus.

In this study (performed in collaboration with Dr Ross Maddox and Prof KC Lee (University of Washington, Seattle, published as Maddox et al. (2015)), I have shown that listeners were better able to report brief deviants in the target stream when a visual stream was temporally coherent with the target stream and performance was impaired with the visual stream was temporally coherent with the non-target auditory stream. I speculated that when there is cross-modal temporal coherence between a feature of an auditory and visual stream, those features are bound and this results in a cross-modal object.

In chapter III, I investigated the neural correlate of temporal coherence driven binding of auditory and visual features which results in a cross modal objects. To do so, I have investigated the neuronal representation of the early integration of auditory and visual stimuli by recording in awake and anaesthetised ferret AC. At the bottom up level, visual inputs might promote the auditory scene by enhancing the temporally coherent sound and provide information for higher cortical areas to perceptually bind auditory and visual information. If so, I hypothesise that temporally coherent AV stimuli are better represented in AC whereas, visual inputs might promote the auditory scene by facilitating the formation of cross-modal objects. If so, I hypothesise that temporally coherent AV stimuli are better represented in AC with enhanced coding for any feature associated with the cross-modal object including those that are orthogonal to the features that bind them.

I have shown that visual stimuli elicit reliable changes in the phase of the local field potential and an enhanced spike-based representation of acoustic mixture within auditory cortex. I have provided mechanistic insight into how auditory and visual information are bound together to form coherent perceptual objects.

Finally, in Chapter IV, I have examined the great individual differences in subjects' ability to benefit from temporal coherence between the auditory and visual streams. In Chapter II, I observed that participants were highly variable in the extent to which visual stimuli influenced perception. In order to investigate whether these differences in participants' performance is due to their ability to use AV coherence, I have tested participants performance on AV temporal coherence after short AV training. This chapter aimed to determine whether (i) exposure to temporally coherent AV stimuli or (ii) actively discriminating AV temporal coherence influenced the ability of listeners to exploit visual information for auditory scene analysis. I hypothesised an improved ability to use visual cues for scene analysis. Specifically that either participants' performance will be improved by target coherent visual stimuli and will be degraded by distractor coherent visual stimuli, or a greater ability to exploit temporal coherence such that both target and distractor coherent visual stimuli will improve (as in both cases the visual stimulus helps listeners to segregate the streams more effectively) relative to the independent condition.

Chapter II: The role of temporally coherent visual information in the formation of auditory streams

Introduction

In everyday listening, we are bombarded with many sounds coming from different sources, which degrade our ability to focus on only one specific sound. However, we can often see the source of an incoming sound, which might potentially enhance our listening ability. For example, being able to see a talker's face drastically improves our speech intelligibility (Dodd, 1980, Sams et al., 1991, Bernstein and Grant, 2009). Although facing a great amount of sensory information in multiple modalities might be overwhelming and confusing, if the multisensory information is properly integrated, we can utilize redundancies within and across modalities, and hence gain a better understanding of these complicated environments.

In the psychoacoustics literature, attention is often thought of as operating as a gain control mechanism, enhancing the internal representation of the attended stream and suppressing the representation of the ignored streams (Hillyard et al., 1998, Lee et al., 2014). Consistent with this view, several imaging studies have demonstrated that signal activity in early sensory areas such as in the auditory cortices is modulated by attention (fMRI: Grady et al., 1997, Petkov et al., 2004, Woods et al., 2009, ERPs: Hillyard et al., 1998) and spectro-temporal features of speech are better represented in the cortical responses when the speech is attended to in a mixture compared to when it is ignored (Ding and Simon, 2012, Mesgarani and Chang, 2012). However, the interaction between attention and AV integration is not clear. AV integration might lead to saliency-based selective attention which results in better internal representation, or they both might be parallel mechanisms that might help our auditory stream formation.

In this study, I used relatively long duration (14 seconds) artificial vowel sounds with amplitude modulation that was generated with a noisy envelope low-pass filtered at 7 Hz. This frequency range is within the ethologically relevant modulation frequency range

(Chandrasekaran et al., 2009), the resulting stimuli were naturalistic time-varying signals whose temporal envelope shares properties of speech but lacked any linguistic content.

Participants were simultaneously presented two such streams, each of which had a different pitch and timbre, and contained short ‘deviants’ where the timbre of the vowel briefly changed. They were asked to attend to one of the two streams and to report the deviants in the target stream while ignoring those in the distractor stream. I tested listeners with this auditory selective attention task that required they detect timbre deviants in complex sounds (artificial vowels), whereas our collaborators tested the ability of listeners to report frequency deviants in pure tones.

In order to perform in this auditory selective attention task, participants had to be able to segregate the auditory streams and then selectively attend to the target stream to detect deviants. Attending to some sound sources while ignoring others requires that sound elements are appropriately grouped in order that *selective* attention can operate to focus on one object or stream from multiple competing sound sources. Visual cues may provide an additional source of information that facilitates grouping and enables listeners to successfully attend to target sounds.

I asked whether a radius-modulated visual stimulus (which itself contained no auditory task-relevant information) could enhance the ability of listeners to detect brief deviants occurring in a target stream while ignoring those in a distractor stream. The visual stimulus was designed such that radius changed in a way that was temporally coherent with one of the sound streams, or independent of both.

This allowed an investigation of the role that temporal coherence might play in audio-visual integration in auditory scene analysis by creating cross-modal objects. Importantly, the dimension that linked auditory and visual streams (i.e. auditory intensity and visual radius size) was independent of the feature that participants were required to detect (auditory timbre) so the visual stimulus offered no additional information to assist listeners in detecting deviants.

I have speculated either that there will be no influence of temporal coherence in scene analysis so that no differences in listener's performance or that temporal coherence driven binding will create a cross-modal object which might be result in;

(i) Enhancing the listener's ability to detect brief timbre deviants when the visual stimulus is coherent with the target stream and disrupting their ability when visual stimuli is coherent the distractor stream, or

(ii) Enhancing the listener's ability to detect brief timbre deviants when the visual stimulus is coherent with either the target stream or the distractor stream as either case participants will be better able to separate the streams.

Methods and Materials

Participants

Twenty five healthy subjects (age range 18–34 years; mean age 27 years; 11 males) participated in the study. They were paid for their participation in the study and gave written informed consent to the study approved by the Ethics Committee of the University College London (ref: 5139). Five participants could not complete the minimum requirement of 70% correct in vowel detection threshold task and therefore did not go on to perform the main AV selective attention task. The data from four of the participants were excluded from analysis due to very low performance (d'<0.8) in the AV selective attention task.

Stimuli

Auditory Stimuli were artificial vowel sounds that were created in Matlab (MathWorks, USA), based on an algorithm adapted from Malcolm Slaney's Auditory Toolbox (http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/). Artificial vowel sounds were generated by band-pass filtering click trains (the repetition rate of which determines the fundamental frequency (F0), or perceived pitch). The centre frequency of the band pass filters determines the location of 'formants' or peaks in the energy spectrum, which determine the identity of the vowel sound. Spoken vowels form clusters according to phonetic identity within a space defined by the location of the first (F1) and second formant (F2). Two artificial vowels presented at a fixed fundamental frequency (F0) were used for reference vowel streams; [u] (F1-F4 460, 1105, 2857, 4205 Hz) presented at a F0 of 175 Hz, and [a] (F1–F4 at 936, 1551, 2975, 4263 Hz) at 195 Hz F0 (Figure 2.1b)

The two vowel streams were independently amplitude modulated with low pass <7 Hz noise envelopes (Chandrasekaran et al., 2009). A null envelope was created firstly by setting all amplitudes of frequency bins 0Hz <bins<7 Hz to unity and others to zero. At an audio sampling rate of 24.414 Hz, all non-zero bins were given a random phase from a uniform distribution between 0 and 2π, the corresponding frequency bins across Nyquist frequency were set to the complex conjugates to maintain Hermitian symmetry, and the inverse Fourier transform

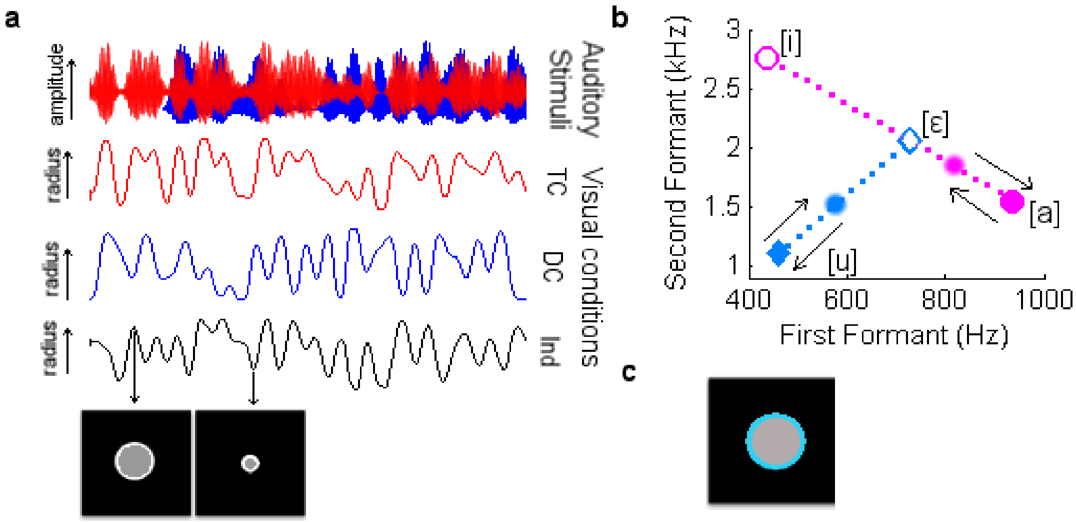

Figure 2.1 Stimuli

Panel a shows a 3 second segment of the auditory stimuli and the temporal envelopes with which the visual stimulus could be modulated. The target stream is shown in red, and starts 1 second earlier than the distractor stream (blue). There were three visual condition. Target coherent condition (TC); in which the envelope that modulated the radius of the visual stimulus changed coherently with the target stream amplitude modulation envelope, distractor coherent condition (DC); the visual envelope changed coherently with the distractor stream envelope and independent condition (Ind); the visual envelope was independent of the envelope of both auditory streams.
Panel b shows the first and second formant of the reference vowels (filled markers), deviant vowels (open markers; Blue diamonds: A1; Pink circles: A2) Dots trajectory represent morph steps and small filled circles illustrate the morph step used as deviant. Panel c shows visual deviants that were generated by changing the color of the circumference of the disk to cyan.

was computed yielding a time domain envelope. A second and third envelope were created using the same method and orthogonalised using a Gram-Schmidt procedure, which is a numerical analysis method for orthonormalising a set of vectors in an inner product space, most commonly the Euclidean space R^n equipped with the standard inner product. Each envelope was then normalized so that it spanned the interval [0, 1] and then sine-transformed [$y = \sin^2(\pi x/2)$] so that the extremes were slightly accentuated.

Within each trial, two amplitude modulated vowel “streams” were presented simultaneously, in which one stream started one second earlier (target stream) than the other stream (distractor stream, Figure 2.1a top schema). Each trial consisted of a 14 second target stream and 13 second distractor stream. 200ms long timbre deviants were embedded into these two amplitude modulated vowel streams.

Timbre deviants were created by smoothly morphing from a reference vowel along a trajectory in F1 and F2 space from the reference vowel to the deviant vowel (Figure 2.1b). Briefly, the trajectory F1 and F2 space from reference to deviant vowel were divided into forty steps, resulting in forty timbre morphs of increasing difficulty. Morphs were made by generating a series of closely spaced vowel sounds which were stitched together with linear onset and offset ramps such that the resulting sound maintained a constant pitch and the timbre smoothly varied to (and from) the deviant identity. The larger the step size, the easier to detect the change between the reference and deviant vowel. For example, figure 2.1b shows the F1 and F2 space of timbre deviants. For the [u] stream the deviant vowel was [e] (F1-F4: 730, 2058, 2857, 4205 Hz, F0: 175Hz, in blue), and for the [a] stream the deviant was [i] (F1-F4: 437, 2761, 2975, 4263 Hz, F0: 195Hz, in purple). Forty timbre morphs between reference vowel (filled circle) and deviant vowel (open circle) were shown with dots. The threshold timbre step size along the reference-deviant timbre morph (filled small circle) for each vowel was measured for each participant with a vowel detection threshold task (as described in more detail below). Fixed step sizes were then used in the main experiments for all trials. Typical thresholds (70%) for subjects were 12.25% along the trajectory from reference to target (Fig 2.2, discussed in more detail in the results). In our collaborators' study, the higher frequency stream was more salient and so was attenuated 3 dB so that both streams were of equivalent perceived loudness. Frequency deviants were 100 ms sinusoidal carrier frequency deflections with a frequency modulation of 1.5 semitones. Streams were calibrated to be 65 dB SPL (RMS normalized) using a Brüel & Kjær artificial ear and presented against a low level of background noise (54 dB SPL).

Visual stimuli consisted of a radius-modulated light grey disc on a black background. The disk was presented centrally (at fixation) and subjects were asked to report occasional visual deviants where the circumference of the ring briefly flashed cyan (Figure 2.1c). Its maximum size subtended 2.5 degrees of visual space and was presented on a LED (LG 24EN43) monitor with a refresh rate of 60 Hz. Subjects sat a distance of 60 cm, with their heads immobilized on a chin rest. On each trial, the disc radius changed in one of three temporal relationships:

coherent with changes in the amplitude of the target stream, coherent with amplitude changes in the distractor stream or independent of both auditory streams (Figure 2.1a).

There were 96 trials with 480 deviants (192 target deviants, 192 masker deviants, 96 visual deviants). Each trial lasted 14 seconds (target stream, 13 second distractor stream). On average there were two deviants per auditory stream and one deviant per visual. Deviants were placed pseudo randomly in the envelope. Across streams (auditory and visual) deviants could not occur within 1 second of one another. Deviants always occurred at times when all three envelopes had a value of >0.7 normalized amplitude to ensure that the signal to noise ratio of the two streams was matched. Visual envelopes were created by subsampling the auditory envelope at the monitor frame-rate of 60 Hz, starting with the first auditory stream so that auditory amplitude corresponded with the disc radius at the beginning of each frame. All streams ended simultaneously.

In the experiment there were three AV coherence conditions, (1) target coherent condition (TC); in which the envelope that modulated the radius of the visual stimulus changed coherently with the target stream amplitude modulation envelope, (2) distractor coherent condition (DC); the visual envelope changed coherently with the distractor stream envelope (3) independent condition (Ind); the visual envelope was independent of the envelope of both auditory streams. (Figure 2.1a).

Procedure

MatLab version 7.7.0.471 2011b (The MathWorks) equipped with Psych Toolbox 3.0 was used for stimulus and protocol control as well as to acquire all behavioral data. Temporally precise presentation of auditory and visual stimuli was achieved using Real-Time processors version 2.1 (Tucker-Davis Technologies, Alachua, FL) for use in PsychToolbox to ensure low-latency. Sound stimuli were presented via headphones (HD 555, Sennheiser, Wedemark, Germany). Participants were asked to run a short vowel detection threshold pre-test (~30mins) in order to establish their thresholds for detecting timbre differences so that difficulty could be matched across participants.

In the vowel detection threshold test, eight timbre morphs of increasing difficulty were generated for each reference vowel, varying from 5% to 22.5% along the reference-deviant vowel trajectory. This range was chosen following pilot studies in which timbre morphs were presented across a wider range (2.5% to 50%). As in the main experiment, subjects were required to press a button if they heard a timbre deviant. From the resulting data hit rates were calculated and thresholds (70% hit rate) were estimated for both vowels independently from a fitted psychometric function (e.g. Figure 2.2). The threshold (70% correct) timbre step size for each vowel was then used in the main experiments. This ensured that the deviants were equivalently difficult to detect for both vowels, across all participants.

In the auditory selective attention task, participants were instructed to listen for deviants in the target stream and ignore deviants in the distractor stream. The target stream was cued by starting one second earlier than the distractor stream. The visual stream started at the same time as the target stream, and either changed coherently with target stream or distractor stream or was independent of both. Participants were asked to press a button whenever they detected a deviant in the attended stream. In order to ensure that participants were watching the visual stimulus, they were instructed to also press a button if they saw a brief change in colour in the ring surrounding the visual stimulus.

Data Analysis

I used signal detection analysis (Swets, 1964) to calculate participants' performance in the auditory selective attention task. Within this framework, responses are assumed to fall into one of four categories. 1: (1) hits: a button press within 1 second of a deviant in the target stream (2) false alarms: a response within 1 second of a deviant in the distractor stream (3) misses: failure to detect a deviant in the target stream, and (4) correct rejections: not reporting deviants in the distractor stream. Signal detection analysis employs the proportions of such responses in the calculation of separate indexes for accuracy and response bias. Accuracy can be indexed by the number of hits relative to the number of false alarms, which is known as d prime (d'). An accurate individual, then, is not the one scoring the most hits, but the one scoring the most hits relative to false alarms. False alarms to non-deviants were excluded from analysis as they were less than 0.3% of all responses (mean = 1.4 response, std= 1.2 response) for each participant.

While d' and bias are each dependent on hit rate and false alarm rate, there is no dependence between d' and bias, or between false alarm rate and hit rate, and none of these should be dependent on visual hit rate. While the MANOVA might provide a bit more statistical power, the separate ANOVAs run here for hit rates, false alarm, d’, bias and hit rates are easier to interpret while being more specific. We specifically explore the hit rates and false alarms so that we can interpret changes in d’ to determine superior d’ values result from higher hit rates, lower false alarms, or a combination of these factors.

Statistical analyses

I combined our dataset with that of our collaborators (Ross Maddox and KC Lee, University of Washington, Seattle) who performed an identical experiment using pure tones (instead of vowels) and requiring subjects to detect frequency deviants. Our findings on combined datasets are published see Maddox et al (2015).

We used a mixed analysis of variance (ANOVA) to test for differences in the d’, hit rates, false alarm, and visual hit rates across AV coherence condition and deviants stimulus type (timbre deviant and pitch deviant). We tested for differences in the d’, hit rates, false alarm, and visual hit rates using a one-way ANOVA. Statistical tests were performed using the SPSS statistical software (version 20.0, IBM Corp., Armonk, NY, USA)

Results

In order to examine the performance differences across three AV coherence conditions, the hit rates, false alarms, d’ and reaction times of 16 participants were calculated for auditory detection performance. Hit rates were also calculated for visual deviant detection. Participants who achieved a d’ less than 0.8 on average and/or visual hit rate less than 70% were excluded from data analysis.

F0 Determining individual timbre deviant detection thresholds

In order to control participants’ ability to detect the timbre change in an ongoing auditory stimulus, participants were asked to run a short vowel detection threshold for 2 different timbre deviants across 2 F0s and 8 different morph sizes. I have calculated fitted psychometric functions for each individual performance for both artificial vowels across all morph steps (e.g Figure 2.2a) and found no significant differences in morph size between timbre deviants (pairwise t-test, t=6.765, p=0.876). The mean (±SEM) 70% correct detection thresholds morph step for [e] in [u] stream were 42 ±7 Hz for F1, 143 ±24 Hz for F2 and [i] in [a] stream expressed in Hz (Figure 2.2b) were 70 ±12 Hz for F1, 168 ±32 Hz for F2.

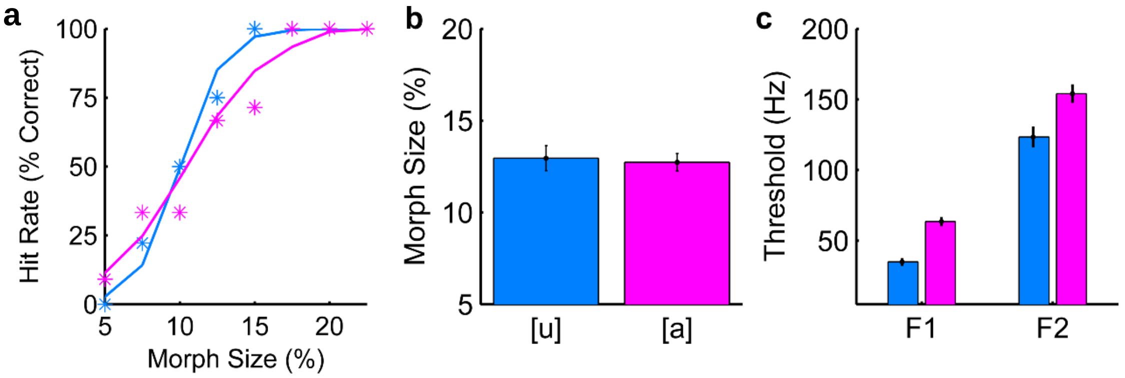

Figure 2.2 Estimating timbre change detection threshold

Panel a shows an example of one participants' threshold data. Each line shows the fitted line for each stream (magenta-[a], blue-[u]). Panel b shows the mean ±SEM morph size (% difference from reference and target vowel) of all participants. A given % change in morph size for a vowel results in formant-dependent differences when measured in Hz. c The mean ±SEM morph size for all participants expressed in Hz for both first and second formants.

Auditory-visual temporal coherence enhances performance in an auditory selective attention task

In order to examine the influence of AV temporal coherence, I have compared discriminability (d') values for three AV coherent conditions across participants. People were better able to detect deviants when visual stimuli were coherent to the target stream compared to an independent visual stimulus, or one that was coherent with the distractor stream. Improved performance was observed regardless of deviant stimulus type for timbre and frequency deviant detection. The results for each of the three AV coherence conditions are shown across all subjects for the timbre and frequency detection tasks in Figure 2.3.

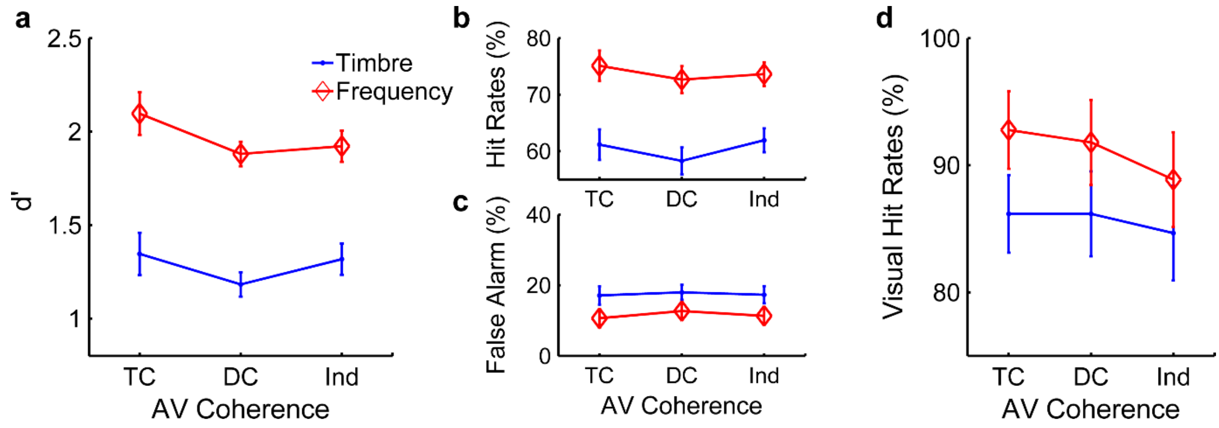

Figure 2.3 The effect of visual information on selective attention task (for timbre and frequency deviants)

a, b, c, shows the mean and ±SEM of d’, hit rates and false alarm for three AV coherence conditions in the auditory deviant detection respectively. Panel d shows the mean hit rates for visual stimuli detection in all conditions with ±SEM.

I ran an ANOVA on the combined dataset (timbre and frequency deviant data) for d’ with factors of auditory deviants stimulus type (timbre deviant and frequency deviant) and audio-visual coherence (target coherent, distractor coherent, independent), and found a significant between-groups effect of deviants stimulus type (F(1,30) = 9.36, p = 0.005) and a significant within-subjects effect of coherence (F(2,60) = 4.28, p = 0.018). There was no interaction between these two factors (p = 0.60), indicating the generality of the effect of cross modal coherence on the auditory selective attention task to different features and experimental setups, as well as different task difficulties (as the performance was significantly better with frequency deviants versus timbre deviants). The mean d’ and hit rates and across-subject mean d’ and hit rates in each condition relative to each subjects overall mean are shown in Figure 2.4. Post hoc comparison revealed that participants were better able to detect the deviants when the visual stimulus was coherent with the target auditory stream than with the distractor stream (TC>DC p = 0.0049, Bonferroni corrected α = 0.017).

Similar results were observed with an ANOVA for hit rates. There was a significant effect of event type (F (1, 30) = 10.1, p = 0.0034) and also of coherence (F (2, 60) = 1.286, p = 0.0497). Post hoc comparison revealed a significant difference between hit rate in the target coherent and distractor coherent conditions (p = 0.011). These results showed that a visual stimulus which is temporally coherent with a target steam increases hit rates when compared to an independent visual stimulus, or one that is coherent with the distractor stream (Figure 2.3b-c). Participants were equally good at detecting the flashing ring in all three conditions (overall hit rate was =89.5% for TC, 89.74% for DC, 84.17% for Ind, Figure 2.3d).

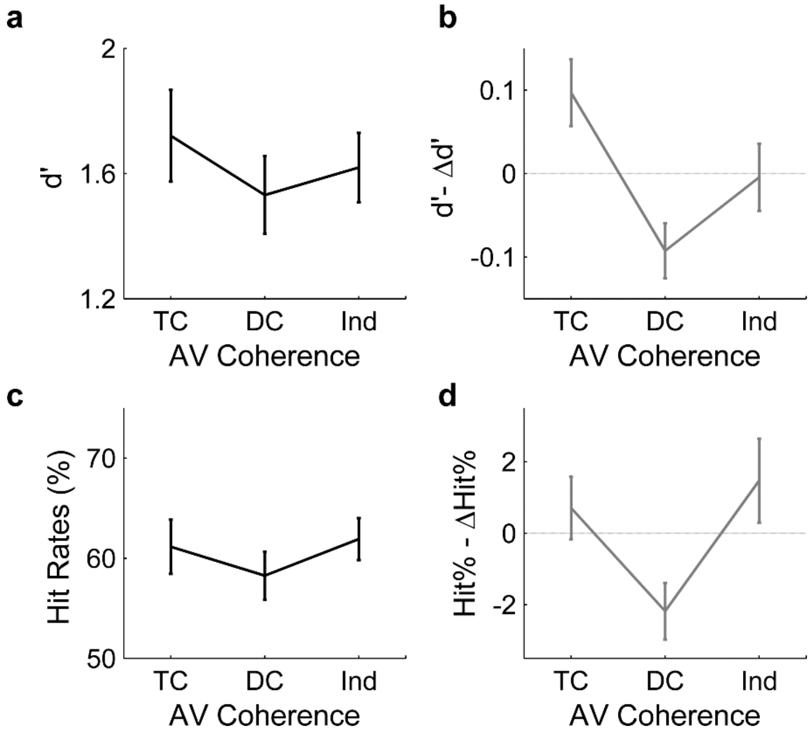

Figure 2.4 Behavioural measure shown in d’ and hit rate.

Left: mean ±SEM for each AV coherence condition across all participants including both experiments (timbre deviant detection as well as frequency deviant detection). Right: normalized mean ±SEM for all participants indicating within-subjects effects.

AV temporal coherence has a stronger effect in more difficult listening conditions

Auditory streams had fixed F0s, ([a] - F0 = 195Hz and [u] - F0=175Hz). Participants showed better performance (larger d’ values) in detecting deviants in [a] target stream trials compared to [u] target stream trials (Pairwise t-test, t = 0.265, p = 0.792), suggesting that participants found [u] target stream trials easier than [a] target stream trials. 2 way repeated ANOVA with factors target stream type ([u] and [a]) and audio-visual coherence (target coherent, distractor coherent, independent) revealed a significant effect of target stream type (F (1, 60) =28.96, p< .001) but no effect of audio-visual coherence (F (1, 60) =1.96, p=0.079).

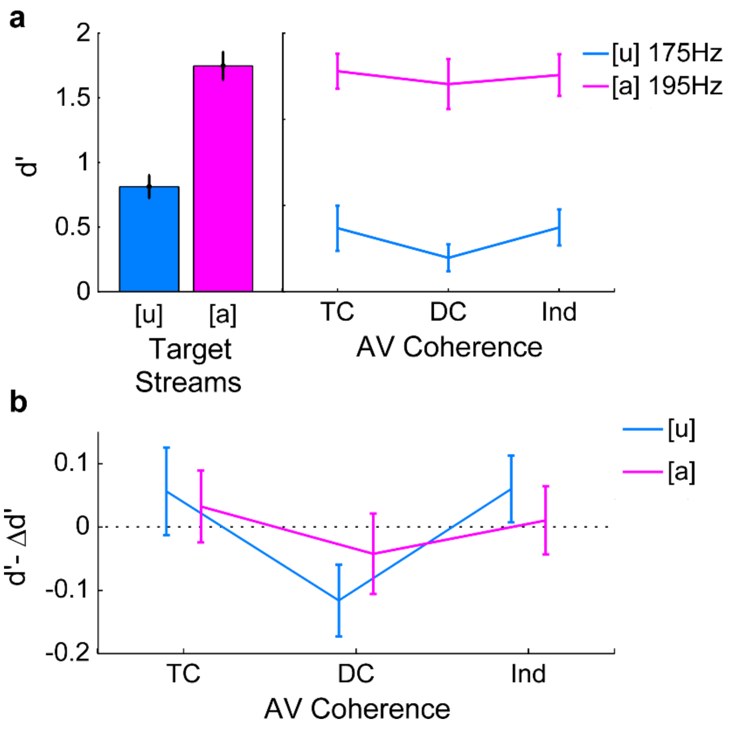

Figure 2.5 Larger effect in harder condition

Panel a shows the mean d` ±SEM for auditory deviant detection for [u] target or [a] target stream and mean d` ±SEM for different AV coherence conditions. Panel b shows d` differences.

There was a significant interaction between target stream type and audio-visual coherence (F (1, 60) =8.96, p< .001). In order to examine the audio-visual coherence effect on the auditory selective attention task, the across-subject means in each condition relative to each subject's overall mean was calculated in Figure 2.4b.

Post-hoc comparisons revealed that when the [u] stream was the target participants were better able to detect deviants when the visual stimulus was coherent with the target auditory stream than distractor stream (TC>DC p = 0.0049, Bonferroni corrected α = 0.017) in but this effect was not present when the [a] stream was the target (TC and DC p = 0.0049, Bonferroni corrected α = 0.017). Since participants' performance were poorer on [u] stream trials this suggests that the impact of AV coherence on stream segregation is affected by task difficulty.

Sound Alone Condition

In order to compare the AV coherence effect of AV selective attention task with a sound alone selective attention task, I ran an additional experiment. I have collected data from 7 participants, in which there were four visual coherence conditions including target coherent, distractor coherent, independent and sound alone. First, I have run an ANOVA to replicate the previous AV coherence effect on deviant detection. However, I failed to observe a significant effect of AV coherence (Figure 2.6a, F (3, 27) = 1.04, p = 0.604) due to the small sample size. A power analysis indicated that I would require 10 subjects in order to perform a within group comparison, providing a 80% chance of detecting a difference in d prime value of 0.163 (as observed in the main experiment) at the 5% level of statistical significance.

Pairwise comparison revealed that, participants were better able to detect deviants when visual stimuli were coherent to the target stream compared to one that was coherent with the distractor stream, as shown in the previous experiment (Pairwise t-test, t_{12}=1.34, p=0.003). However, there were no difference in performance when visual stimuli was coherent to target stream compared to independent visual stimulus (Pairwise t-test, t_{12}=0.205, p=0.841) or without visual stimulus (sound alone condition; Pairwise t-test, t_{12}=0.327, p=0.748). Similarly, there were no differences in hit rates for visual detection across AV coherence conditions (F (2, 24) = 0.96, p = 0.9607).

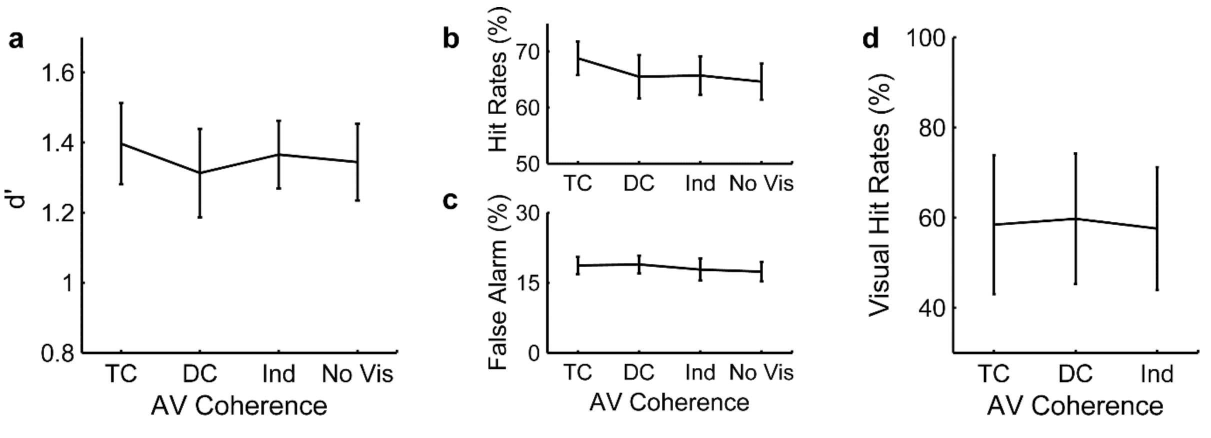

Figure 2.6: The effect of visual information on selective attention task – With sound alone condition

a, b, c, shows the mean and ±SEM of d', hit rates and false alarm for four AV coherence conditions in the auditory deviant detection respectively. Panel d shows the mean hit rates for visual stimuli detection in all conditions with ±SEM.

While robust statistical comparisons were not possible with this dataset, the pairwise comparisons above indicate that the trend within this subpopulation is similar to that in the main experiment with TC significantly better than DC. Within these subjects the A only condition appeared intermediary to the TC and DC. In order to explore this further I plotted scatter plots comparing the A only condition to both TC and DC conditions (Figure 2.7). For 5/7 participants the TC performance was slightly higher than their A only condition, and 5/7 participants performed slightly worse in the DC condition when compare to sound alone. Together, the trends in these data suggest that the sound alone condition is likely to be intermediary - that is, a visual stimulus that is temporally coherent with the target stream boosts performance relative to no visual stimulus, and visual stimulus that is temporally coherent with a distractor stream impairs performance. However, due to the limited sample size more data will need to be collected (at least an additional 3 participants) to confirm these trends statistically.

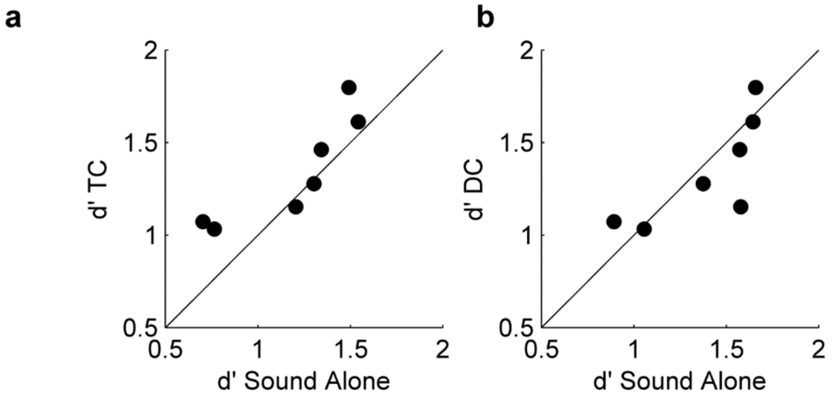

Figure 2.7 Correlation between AV condition and sound alone condition.

Panel a shows the scatter plot of d' values for target coherent (TC) condition and sound alone condition. Panel b shows distractor coherent (DC) versus sound alone condition. Each circle shows one participants performance (n=7).

Discussion

In this study, participants were more sensitive to deviants when the visual stimulus was temporally coherent with the target auditory stream than when it was coherent with the distractor. Their performance was improved when the visual stimulus had a modulation envelope that was temporally coherent with the target stream and impaired when the visual stimulus was temporally coherent with the distractor stream. This effect was small, but robust across participants. When combined with the paired study performed by our collaborators, listeners were significantly better at detecting auditory deviants when a visual stimulus was coherently modulated with the target stream compared to when modulated with the distractor stream. This result suggests that a temporally-modulated visual stimulus, which itself contains no auditory task relevant information, can influence the ability of listeners to segregate two concurrently presented competing auditory stimuli.

My stimuli were complex sounds (vowels), while our collaborators used pure tones, in an otherwise identical experiment. Both datasets had similar trends, although their d' values were slightly higher than mine on average (they did not individualize the difficult of their deviant sounds with a threshold task as I did). Nevertheless, across all 32 participants, there was significantly improved performance when the visual stimulus was coherent with the target stream (Figure 2.3) compared to when the visual stimulus was coherent with the distractor. However, individual subjects were variable in the magnitude of the observed improvement. I, therefore, conducted an experiment to explore whether training can lead to an enhanced ability of subjects to benefit from the visual stimulus. This is described in Chapter IV.

The more difficult the task is, the larger the visual effect on auditory processing

Participants showed better detection ability in [a] stream than [u] stream (Figure 2.4). There might be two reasons; *(i)* [a] and [i] vowels have more distinct perceptual boundaries. One possible reason, therefore, is that the higher F0 of [a] stream might have an interaction with formants of the vowel, leading easier detection. Another possibility is that while I calibrated the RMS energy of the two vowels, the particular spectral features of the [a] vowel mean that it had a higher energy level around the formant regions leading to a higher signal: noise ratio.

Consistent with the literature on visual cues as complementary sources for auditory processing, the trend of the visual stimuli effect on stream segregation appeared to be strongest for the [u] vowel stream, in which participants found it harder to detect the deviants. This is consistent with the idea that we benefit from visual cues in challenging conditions (see Chapter I-Part I; Grant and Seitz, 2000, Sumby and Pollack, 1954).

Visual cues with no auditory task-relevant information enhance auditory stream segregation

Since the visual stimulus does not provide any information about when, or in which stream, the timbre deviants occurred, the benefit in a performance that I observed likely stems from an improved ability for listeners to segregate the streams and/or successfully apply selective attention to the target stream. Forming a stream also requires binding of the parallel perceptual attributes of its tokens, to the exclusion of those belonging to competing streams. Temporal coherence across different sensory modalities might support automatic cross modal binding (Shamma et al., 2011) and asserts that any sequences of attributes that are temporally correlated will bind and form a stream segregated from uncorrelated tokens of perceptually different attributes. A temporally coherent visual stimulus unifies with the target stream and becomes a part of the same ‘object’, so that does not disturb the performance to detect the auditory (and/or visual) deviant in an on-going competing stream. However, when it unifies with distractor stream (temporally coherent with distractor stream), attention must be divided across two objects: the target auditory stream and the visual stream, likely leading to disrupted performance.

Coherence driven object formation

The conceptual model of processing were proposed (Figure 2.6, adapted from Maddox et al. 2015) that when there is cross-modal temporal coherence between a feature of an auditory and visual stream, those features are bound and this results in a cross-modal object (just as two auditory or visual streams with the same envelope very likely would have bound together as well). The definition of ‘an object’ is controversial, but one of the criteria that defines an object is that the stimuli with all features if perceived as one object, would be influenced as a whole (Blaser et al., 2000, Shinn-Cunningham, 2008). My result illustrated that temporal coherence between amplitude modulation and visual size enhanced the timbre of the auditory stimuli, although this feature of auditory stimuli had no relation to the visual stimuli. This suggests a coherence driven binding of auditory and visual features forms a cross-modal object.

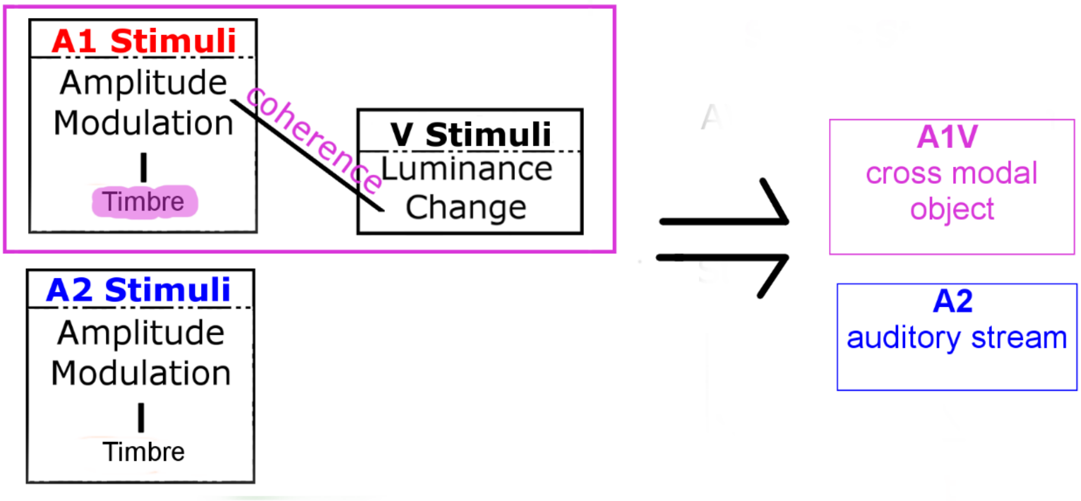

Figure 2.7 The conceptual model of cross-modal object formation (adapted from Maddox et al. 2015)

To sum up, the study provides psychophysical evidence on the influence of temporally coherent visual stimuli on auditory stream segregation. Participants' performance was best when the visual stimulus was coherent with the target stream, and their performance was worse when the visual stimulus was coherent with the non-target stream. These findings shows the role that temporal coherence plays in AV integration in auditory scene analysis by creating a cross-modal object.

These findings are discussed in the light of the basic principle of AV binding without linguistic context. The enhancement in performance seen with temporally coherent visual stimuli is relatively modest and rather variable between subjects. Nevertheless, since a 1 dB increase in SNR results in a 10-20% increases in intelligibility (Brand and Kollmeier, 2002) the small effect I observe here may provide significant performance benefits in adverse listening conditions.

However, the underlying mechanism of how temporally coherent visual input promotes stream segregation is still not clear: specifically, whether temporal coherence between auditory and visual stimuli lead to saliency-based selective attention (Ding and Simon, 2012, Mesgarani and Chang, 2012), resulting in a better internal representation of the auditory stream or coherence driven binding between auditory and visual stimuli result in cross-modal object. In order to examine the neural correlates of coherence driven binding between auditory and visual stimuli, I conducted a neurophysiology experiment, as discussed in chapter III.

Chapter III: Early integration of visual information in auditory cortex promotes auditory stream segregation

Introduction

When listening to a sound of interest, we frequently look at the source. However, how auditory and visual information are integrated into a coherent perceptual object is unknown. As I have found in Chapter II, the temporal properties of a visual stimulus can bias the perceptual organisation of a sound scene and enhance or impair listening performance depending on whether the visual stimulus is temporally coherent with a target or distractor sound stream. Together, these behavioural results suggest that temporal coherence between auditory and visual stimuli promotes binding of cross modal features to enable the formation of an AV object (Bizley et al., 2016).

As it is discussed in Chapter I, visual stimuli can modulate neural activity in the auditory pathway. In this chapter, I investigate the role of visual activity in stream segregation in auditory cortex (AC). One hypothesis is that the early integration of cross-sensory information provides a bottom-up substrate for the binding of multisensory stimulus features into a single perceptual object (Bizley et al., 2016). Bizley et al. have recently argued that in order to distinguish binding (defined as the process that underpins perceptual object formation) from other types of multisensory integration, behavioural and neural experiments must demonstrate a benefit in a stimulus feature orthogonal to the features that link cross-modal stimuli. As I have demonstrated in Chapter II, attending to the feature of a visual stream (color) enhanced the perception of a feature of the auditory stream (auditory timbre) in which those two features on their own were orthogonal. Such enhancements were suggested to occur through a chain of bound features within cross-modal objects.

In order to examine the neural correlates of coherence-driven binding between auditory and visual stimuli, I have recorded from AC of anesthetized ferrets which allowed me to rule out any top-down effects of attention and perform additional control analysis, and recorded from AC of awake passively listening ferrets which allowed me to measure neural activity free from confounds of pharmacological manipulation. I have used a modified version of naturalistic time-varying auditory and visual stimuli used in chapter II. The auditory streams were two artificial vowels with a distinct pitch and timbre (denoted A1: [u], F0 = 175 Hz and A2: [a], F0 = 195Hz). Each vowel was independently amplitude modulated with a noisy low-pass (<7 Hz) envelope. A full-field, luminance-modulated visual stimulus accompanied the auditory stimuli with temporal dynamics that matched one of the two auditory streams.

Firstly, I investigated how temporal coherence between auditory and visual stimuli influences the representation of a single auditory stream. To do this, I tested stimulus conditions in which a single auditory-visual stimulus pair was presented, where the auditory and visual streams could be temporally coherent (A1V1, A2V2) or independently modulated (A1V2, A2V1). I hypothesise that temporally coherent AV stimuli are better represented in AC by enhanced spiking activity in auditory cortical neurons based on the direct link from visual cortex (Bizley et al., 2007, Budinger et al., 2006, Kayser et al., 2008) and enhanced across trial reliability as a result of phase resetting by coherent visual stimuli (Lakatos et al., 2008).

Then, I investigated how temporal coherence between AV stimuli influences the representation of an auditory scene composed of two competing auditory streams (dual stream condition) in which both auditory streams were presented and the visual stimulus was temporally coherent with one of the auditory streams (A12V1 or A12V2, Fig. 3.1). I hypothesised that the auditory stream in the sound mixture that was temporally coherent to visual stimuli is better represented in AC.

Finally, I investigated how temporally coherent AV features influence the representation of the orthogonal features of the auditory scene in AC. To do this, both auditory streams with timbre deviants were presented and the visual stimulus was temporally coherent with one of the auditory streams. I hypothesised that timbre deviants in the auditory stream that was temporally coherent to visual stimuli were better represented in AC with enhanced coding for any feature associated with the cross-modal object including those that are orthogonal to the features that bind them.

An important prediction, therefore, is that a neural response to demonstrate binding, not only should the neural encoding of stimulus features that bind auditory and visual streams to be enhanced, but that there should be an enhancement in the representation of stimulus features orthogonal to those that determine binding. Here I tested the hypothesis that the temporal coherence between auditory and visual stimuli can determine the neuronal representation of an auditory scene.

Methods and Materials

Animal preparation

The experiments were approved by the Committee on Animal Care and Ethical Review of University College London and The Royal Veterinary College, and licensed by the UK Home Office, in accordance with the Animals (Scientific Procedures) Act 1986. Neural responses were recorded in 11 adult female ferrets who were chronically implanted with recording electrodes and passively listening/watching stimuli. These animals were trained in various listening tasks for other studies. An additional 5 adult female ferrets (Mustela putorius furo) were used to record responses under anaesthesia. Regular otoscopic examinations were carried out to ensure that both ears of the animals were clean and healthy.

Electrophysiological recordings under anaesthesia

Animal preparation: Ferrets were anesthetized with a single dose of a mixture of medetomidine (Domitor; 0.022mg/kg/h; Pfizer, Sandwich, UK) and ketamine (Ketaset; 5mg/kg/h; Fort Dodge Animal Health, Southampton, UK). The animal was intubated and the left radial vein was cannulated in order to provide a continuous infusion (5 ml/h) of a mixture of medetomidine and ketamine in lactated Ringer's solution augmented with 5% glucose, atropine sulfate (0.06 mg/kg/h; C-Vet Veterinary Products) and dexamethasone to reduce cerebral oedema (0.5 mg/kg/h, Dexadreson; Intervet, UK) in Hartmann's solution. The ferret was intubated, placed on a ventilator (Harvard Model 683 small animal ventilator; Harvard Apparatus) and supplemented with oxygen. Body temperature at around 38°C, end-tidal CO2, and the electrocardiogram were monitored throughout the experiment. The eyes were protected with a zero-refractive power contact lens. Experiments typically lasted between 36 and 56 h.

The ferret was placed in a stereotaxic frame in order to implant a bar to the skull to enable the subsequent removal of the stereotaxic frame. The left temporal muscle was largely removed, and the suprasylvian and pseudosylvian sulci were exposed by a craniotomy, exposing auditory cortex (Kelly et al., 1986). A metal bar was cemented and screwed into the right side of the skull, holding the head without further need of a stereotaxic frame. The dura was removed over auditory cortex and the brain protected with 3% agar solution. The animal was then transferred to a small table in a soundproof and darkened chamber (Industrial Acoustics, Winchester, UK).

Neural activity was recorded with multisite silicon electrodes (Neuronexus Technologies, Ann Arbor, MI) in a 1x 16, 2x 16 or 4x 8 (shank x number of sites; 100 µm site spacing) configuration. Electrodes were positioned so that they entered the cortex approximately orthogonal to the surface of the suprasylvian gyrus. Neural recordings were obtained using TDT System III hardware (RZ2 data acquisition system) with custom written software in Open Project (Tucker-Davis Technologies, Alachua, FL) and MATLAB (Mathworks, Natick, USA).

Stimulus Presentation: Sounds were generated using TDT system III hardware with custom written software in MATLAB, and presented through customized Panasonic RPHV297 headphone drivers (Bracknell, UK). Closed-field calibrations were performed using a one-eighth inch condenser microphone (Bruel and Kjær), placed at the end of a model ferret ear canal, to create an inverse filter that ensured the driver produced a flat (<5 dB) output. Visual stimuli were presented with a white Light Emitting Diode (LED) which was placed roughly 10 cm from the contralateral eye so that it illuminated virtually the whole contralateral visual field.

Electrophysiological recordings for awake passively listening ferrets

Animal Preparation: Animals were bilaterally implanted with WARP-16 drives (Neuralynx, Montana, USA) loaded with tungsten electrodes (FHC, Bowdoin, USA) under general anaesthesia (medetomidine and ketamine induction, as above, isoflurane maintenance 1-3%). Craniotomies were made over left and right auditory cortex, a small number of screws were inserted into the skull for anchoring and grounding the arrays, and the WARP-16 drive was anchored with dental acrylic. WARP drives were protected with a capped well. Animals were allowed to recover for a week before the electrodes were advanced into auditory cortex. Pre-operative, peri-operative and post-operative analgesia were provided to animals under veterinary advice. Full surgical methods for recording from awake animals are available in Bizley et al, 2013.

Ferrets were water restricted prior to testing; on each day of testing, subjects received a minimum of 60ml/kg of water either during testing or supplemented as a wet mash made from water and ground high-protein pellets. Subjects were tested once in a month in morning or afternoon session. Test sessions lasted between 10 and 50 minutes and ended when the animal lost interest in performing the task. The weight and water consumption of all animals was measured throughout the experiment.

Ferrets were passively exposed to stimuli in a customized pet cage (69 x 42 x 52 cm, length x width x height) within a sound-attenuating chamber lined with sound-attenuating foam. The floor of the cage was made from plastic, with an additional plastic skirting into which three spouts (center, left and right) were inserted. Each spout contained an infra-red sensor (OB710, TT electronics, UK) that detected nose-pokes and an open-ended tube through which water could be delivered.

Sound stimuli were presented through two loud speakers (Visaton FRS 8) positioned on the left and right sides of the head at equal distance and approximate head height. These speakers produce a flat response (±2 dB) from 200Hz to 20 kHz, with an uncorrected 20 dB drop-off from 200 to 20 Hz when measured in an anechoic environment using a microphone positioned at a height and distance equivalent to that of the ferrets in the testing chamber. An LED was also mounted above the center spout. Animals were freely rewarded with water from the centre spout, and recording was terminated when animals ceased facing forwards and drinking from the spout.

Stimulus Presentation: Data acquisition and stimulus generation were all automated using custom software running on computers, which communicated with TDT real-time signal processors (RZ2 and RZ6, Tucker-Davis Technologies, Alachua, FL). Sound levels were calibrated using a Brüel and Kjær (Norcross, GA) sound level meter and free-field ½-inch microphone (4191). Visual stimuli were delivered by illuminating the spout with a white LED which provided full field illumination. The animals were not required to do anything other than maintain their heads in position at the spout in a darkened chamber.

Stimuli

Auditory stimuli were artificial vowel sounds that were created in Matlab (MathWorks, USA). In awake recordings, the duration was reduced to 3 seconds in order to collect sufficient repetitions of all stimuli and to ensure animals were facing forwards for the whole trial duration. In the anesthetised recording stimulus streams were 14 seconds long, (Maddox et al., 2015) but in order to directly compare anesthetised and awake data in both cases only the first 3 seconds were analysed (will be discussed in detail, see result section). Auditory streams were presented at 65 dB SPL (Figure 3.1a).

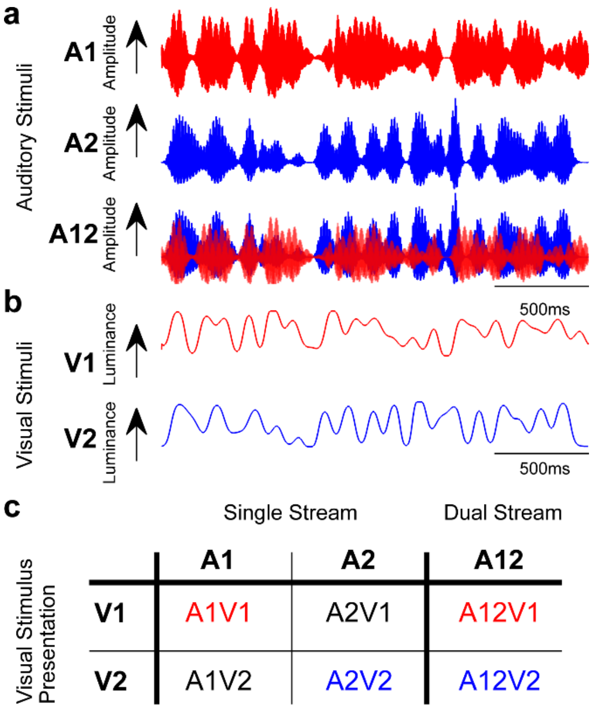

Figure 3.1 Experimental design

a Auditory stimuli; two artificial vowels (A1 and A2), with distinct pitch and timbres, were independently amplitude modulated at <7Hz. Auditory stimuli were presented either separately (top, middle) or in competition (bottom panel) in the presence of luminance modulated visual stimuli **b** the temporal dynamics of which matched one of the two auditory streams. **c** illustrates the stimulus combinations that were tested.

Stimulus A1 was the vowel [u] (formant frequencies F1-4: 460, 1105, 2857, 4205 Hz, F0= 195Hz), A2 was [a] (F1-4: 936, 1551, 2975, 4263 Hz, F0= 175Hz). Streams were amplitude modulated with a noisy <7 Hz envelope. Unless specifically noted, the identity of the auditory stream remained fixed throughout the trial. However, I also recorded responses to auditory streams that included brief timbre deviants (which were the targets in the auditory selective attention task for human listeners). Details of the vowel streams and timbre deviants can be found in Chapter II, Stimuli. As in the previous study, timbre deviants were 200ms epochs in which the identity of the vowel smoothly changed by smoothly changing the first and second formant frequencies to and from those identifying another vowel. Stream A1 was morphed to/from [ɛ] (730, 2058, 2857, 4205 Hz) and A2 to/from [i] (437, 2761, 2975, 4263 Hz).

Visual stimuli were generated using a luminance-modulated LED whose luminance was modulated with dynamics that matched the amplitude modulation applied to A1 or A2. In single stream conditions a single auditory and single visual stream were presented, in dual stream conditions, both auditory streams were presented simultaneously from both speakers (i.e. there are no spatial cues to facilitate segregation) accompanied by a single visual stimulus (Figure 3.1b).

There were 4 single stream conditions (A1V1, A1V2, A2V1, and A2V2) and 4 dual stream conditions (A12V1, A12V2, A12V1$_{dev}$ A12V2$_{dev}$) which were presented pseudo randomly. In the anesthetised recordings, each was presented 20 times.

Data acquisition

In the awake dataset, where recording duration was limited by how long the ferret would remain at the central location, an average of 20 repetitions were collected (minimum: 14). Main auditory visual stimuli set recorded from 522 driven units when animals were awake. Two additional extended stimuli set were recorded. One with no visual condition of single and dual stream from 58 units and other extended stimuli set consisted 3 different amplitude modulation of stimuli (>7Hz, >12Hz and >17Hz) recorded from 92 units.

In the anaesthetised recordings, we recorded the main auditory visual stimulus set from 1198 driven units. Additionally, pure tone stimuli (150 Hz to 19 kHz in 1/3-octave steps, from 10 to 80 dB SPL in 10 dB, 100 ms in duration, 5 ms cosine ramped) were also presented. These allowed us to both to characterize individual units and to determine tonotopic gradients, so as to confirm the cortical field in which any given recording was made. Finally, broadband noise bursts and diffuse light flashes (100 ms duration, 70 dB SPL) were presented and used to classify a stimulus as auditory, visual or auditory visual.

Current source-density (CSD) analysis was applied to field potential data recorded across cortical layers using the inverse CSD method [29]. CSD analysis identifies current sinks and sources in the extracellular space and was used to estimate the layer of the recorded units and to determine changes in activity in different layers of cortex.

Data Analysis

On each electrode wide-band, voltage traces were recorded using TDT System III hardware (RX8 and RZ2) and OpenEx software (Tucker-Davis Technologies, Alachua, FL) with a sample rate of 25 kHz. For extraction of action potentials, data were bandpass filtered between 300 and 5000 Hz and motion artefacts were removed using a decorrelation procedure applied to all voltage traces recorded from the same microelectrode array in a given session. Spiking activity from individual neurons was extracted offline and sorted with a spike-sorting algorithm (WaveClus) (Quiroga et al., 2004). Spikes were sorted based on their waveform and inter-spike intervals (<1ms). Spikes that did not pass these criteria were considered to be MU activity. For local field potentials, signals were filtered between 2 Hz and 100Hz and digitized at 1kHz.

I performed two analyses to evaluate the effect of visual input on spiking and sub-threshold neural activity. I used a cross-validated Euclidean distance based pattern classifier (Schnupp et al., 2006) to determine whether the neuronal responses to different stimuli could be discriminated. Spiking responses were binned (20 ms resolution). The average across-repetition response to each stimulus (minus the to-be-classified response) was used as a template and the response to a single stimulus presentation was classified by calculating the Euclidean distance between itself and the template sweeps and assigning it to the closest template. To determine whether the classifier performed significantly better than that expected by chance a 1000 iteration permutation test was performed where trials were drawn (with replacement) from the observed data and randomly assigned to a stimulus. A neural response was considered to be significant if the observed value exceeded the 95th of the modelled distribution.

Such approach allowed me to classify units according to whether their responses could discriminate two auditory stimuli based on the amplitude modulation of sound (A1 vs A2) regardless of visual presentation, termed “auditory classified unit” (Figure 3.2a, b) and/or whether they could discriminate two visual presentations based on temporal envelope of visual stimuli (V1 vs V2) regardless of auditory presentation, termed “visual classified unit” (Fig. 3.2c, d).

This approach was extended to classify dual stream responses by using the average response to each of the temporally coherent AV stimuli as templates. The performance was (arbitrarily) expressed as the proportion of responses classified as being from the A1, and compared for the two different visual conditions (Figure 3.5).

Phase/power dissimilarity analysis Local field potential recordings were considered for all sites at which there was a significant driven spiking response, irrespective of whether that response could discriminate auditory or visual stream identity. For the single stream trials, I computed a single stream Phase Dissimilarity Index (PDI), which characterizes the consistency and uniqueness of the temporal phase/power pattern of neural responses to continuous auditory stimuli (Luo and Poeppel, 2007). This analysis compares the phase (or power)-consistency across repetitions of the same stimulus with a baseline of phase-consistency across trials in which different stimuli were presented.

In the first stage of PDI analysis, I obtained a time-frequency representation of each response using wavelet decomposition with complex 7-cycle Morlet wavelets in 0.5 steps between 2.5–45 Hz, resulting in 86 frequency points. Next, I calculated the inter-trial phase-coherence value (ITPC; Equ.1) at each time-frequency point, across all trials in which the same stimulus was presented. For each frequency band, the ITPC time-course was averaged over the duration of the analysis window and across all repetitions to obtain the average within-stimulus ITPC.

$$ITPC_{t,f} = \left|\frac{\sum_{k=1}^{N} e^{i\theta_{k,t,f}}}{N}\right| \qquad \text{Equ.1}$$

In which N is equal to the number of trials, and θ is the phase of trial *k* at a given frequency (*f*) and time (*t*). The across-stimuli ITPC was estimated using the same approach but using shuffled data, such that the ITPC was computed across randomly selected trials in which different visual stimuli were presented. The single stream phase dissimilarity index (PDI) is computed as the difference between the ITPC value calculated for *within visual* trials and the ITPC values calculated across visual trials (Equ.2). Large phase-dissimilarity values indicate that the responses to individual stimuli have a highly consistent time course as evidenced in the response to single trials. PDI values were calculated for each stimulus and then averaged across stimuli to calculate values for temporally coherent and temporally independent auditory visual stimuli. Single stream PDI was greater if the *within_vis* value was significantly larger than *across_vis* value (pairwise ttest, p<0.05 Bonferroni correction).

$$Single\ Stream\ PDI = \frac{\sum_{j=1}^{N} ITPCij{,}within_{vis}}{N} - \frac{\sum_{j=1}^{N} ITPCij{,}across_{vis}}{N} \qquad \text{Equ.2}$$

Large positive PDI values indicate that responses to individual stimuli have a highly consistent response on single trials. Single stream PDI values were calculated for each stimulus type and then averaged across stimuli to calculate values for temporally coherent and temporally independent auditory visual stimuli. Single stream PDI was positive if within stimulus ITPC was larger than across-stimulus ITPC (pairwise t-test, p<0.05 Bonferroni correction for 86 frequencies points) and was considered significant if a minimum of 2 adjacent bins exceeded the corrected threshold.

Dual stream phase dissimilarity index (dual stream PDI) values were calculated by extending this approach for dual stream stimuli with the goal of determining how the temporal envelope of the visual stimulus influences the neural response to a sound mixture. To this end, we calculated the *within-dual ITPC* from the A12V1 trials and A12V2 trials separately and *across-dual ITPC* by randomly selecting trials from both stimuli (Equ.3). The within-dual and across-dual ITPCs were then averaged over time and subtracted to yield the dual stream PDI (Equ.3).

$$Dual\ Stream\ PDI = \frac{\sum_{j=1}^{N} ITPCij{,}within_{vis}}{N} - \frac{\sum_{j=1}^{N} ITPCij{,}across_{vis}}{N} \qquad \text{Equ.3}$$

Positive dual stream PDI values indicate that the time course of the neural responses was influenced by visual input, despite the identical acoustic input. We determined whether the dual stream PDI was greater if the *within_dual ITPC* was significantly larger than *across_dual ITPC* (pairwise t-test, p<0.05 Bonferroni correction, as above).

Timbre deviant analysis***:*** In order to determine how a visual stimulus influenced the ability to decode timbre deviants embedded within the auditory streams we used the cross-validated pattern classifier described above for analysing single stream stimuli to discriminate deviant from no-deviant trials. Responses were considered over the 200 ms time window that the deviant occurred (or the equivalent point in the no-deviant stimulus) binned with a 10 ms resolution. Significance was assessed by a 1000 iteration permutation test in which trials were randomly drawn with replacement from deviant and no-deviant responses. The discrimination score was calculated as the proportion of correctly classified trials.

Results

I recorded neuronal responses in the auditory cortex of awake (n = 9, 221 driven single units 311 driven multi-units) and medetomidine-ketamine anesthetised ferrets (n = 5, 1198 driven units, 426 driven single units, 772 driven multi units) in response to naturalistic time-varying auditory and visual stimuli adapted from Chapter II. The auditory streams were two amplitude modulated artificial vowels with a distinct pitch and timbre (Figure 3.1). A full-field, luminance-modulated visual stimulus accompanied the auditory stimuli with temporal dynamics that matched one of the two auditory streams. I first tested the hypothesis that temporal coherence between auditory intensity and visual luminance would enhance the neural encoding of the auditory stream in AC. To do this, I analysed neuronal responses to single stream conditions which a single auditory-visual stimulus pair was presented, where the auditory and visual streams could be temporally coherent (A1V1, A2V2) or independently modulated (A1V2, A2V1, Figure 3.1).

Spike patterns in auditory cortex differentiate dynamic auditory-visual stimuli

I first classified neurons in auditory cortex according to whether they were dominantly modulated by auditory or visual stimulus dynamics using the responses to single stream stimuli. To determine whether auditory dynamics reliably modulated spiking I used a spike-pattern classifier (Schnupp et al., 2006) to decode the auditory stream identity, collapsed across visual stimuli (e.g. responses to A1V1 and A1V2 stimuli were combined and compared to those elicited by A2V1 and A2V2 stimuli). An identical approach was taken in order to determine whether a response was reliably modulated by visual stimulus identity (i.e. A1V1 and A2V1 stimuli were combined and compared to those elicited by A1V2 and A2V2). Neuronal responses which could be decoded at a level better than chance (estimated with a bootstrap resampling) were classified as auditory (Figure 3.2a-b) or visual (Figure 3.2c-d) respectively. 39% (210/532) of units recorded in the awake dataset were classified as auditory, 11% (59/532) as visual, and only 0.3% (2/532) were both auditory and visual (AV). Overall a smaller proportion of units were classified in the anesthetised dataset: 20% (242/1198) were classified as auditory, 7% (82/1198) classified as visual, and 0.7% (9/1198)

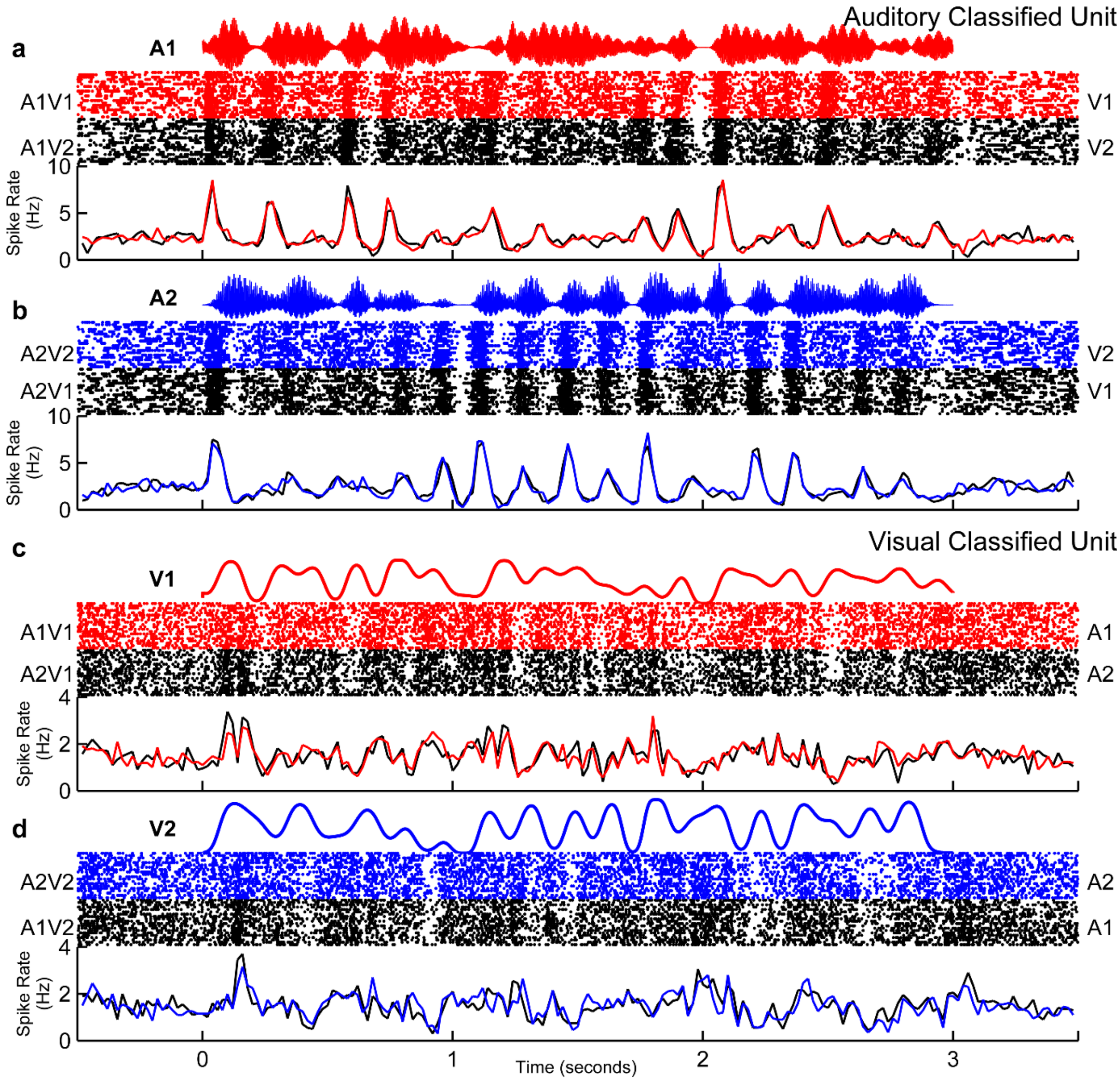

Figure 3.2 Auditory-visual temporal coherence enhances neural coding in auditory cortex

A pattern classifier was used to determine whether neuronal responses could be decoded according to the auditory or visual temporal dynamics. The responses to all four AV combinations are shown for two example units, with responses grouped according to the identity of the auditory stream (**a**, **b**, auditory unit) or visual stream (**c**, **d**, visual unit). In each case the stimulus amplitude/luminance waveform is shown in the top panel with the resulting rasters and peristimulus histogram (PSTH) below.

as AV. During recordings made under anaesthesia, I was able to record additional control stimuli including responses to noise bursts and light flashes that have previously been used to map AV responses in auditory cortex (Bizley et al., 2007). Using such stimuli allow me to assess whether units respond to the onset of a discrete stimulus event, whereas the pattern classifier approach assesses whether you can distinguish the temporal envelopes of two time-varying stimuli. When estimated with noise bursts and light flashes the proportions of auditory, visual and auditory visual units in line with previous studies: 66% of units were driven by noise bursts, 15% by light flashes and 14% responsive to auditory and visual features of stimuli (n = 504, p<0.001).

Temporal coherence between auditory and visual streams enhances neural coding

I hypothesised more reliable neural responses would be elicited by temporally coherent auditory-visual stimuli than by temporally independent stimulus combinations, irrespective of whether a unit's responses were classified as discriminating auditory or visual dynamics. I therefore compared classification accuracy for temporally coherent (A1V1 vs. A2V2) and temporally independent (A1V2 vs. A2V1) stimulus pairs. As predicted, the identity of temporally coherent AV stimuli were better decoded than temporally independent ones (Figure 3.3a, pairwise t-test, awake recordings, pairwise t-test, Auditory Classified n = 210, t_{418} = 34.277, p<0.001; Visual Classified n = 59, t_{116} = 13.327, p<0.001; All Classified n = 271, t_{540} = 35.196, p<0.001; Figure 3.3b, anesthetised recordings, Auditory Classified n = 242, t_{482} =27.631, p<0.001; Visual Classified n = 82, t_{162} = 22.907, p<0.001; All Classified n = 324, t_{660} =33.149, p<0.001).

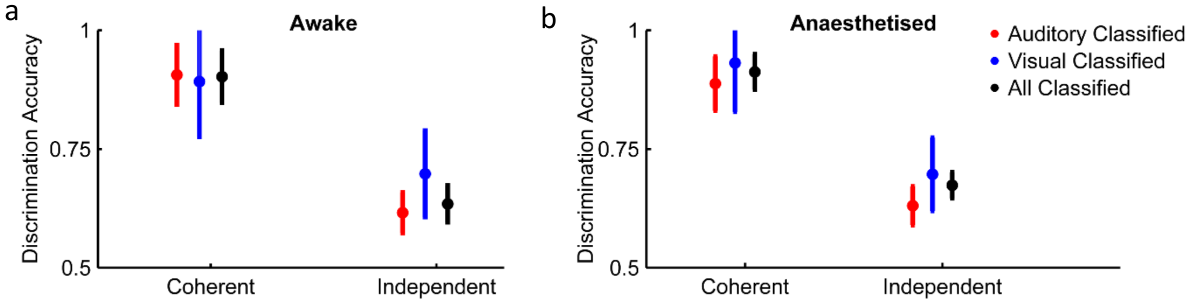

Figure 3.3 Decoder performance in population level

a,b: Decoder performance (mean ± SEM) for discriminating stimulus identity (coherent: A1V1 vs. A2V2; independent: A1V2 vs. A2V1). Unit activity more accurately represented temporally coherent AV stimuli than independently modulated stimuli in both awake (a) and anesthetised (b) datasets. Population discrimination accuracy (mean ± SEM) using the spike-pattern classifier for discriminating coherent and independent auditory visual stimuli in awake (a) and anaesthetised (b) datasets. Pairwise comparisons for coherent versus independent all classified p<0.001 (see results).

What might underlie the enhanced discriminability observed for temporally coherent stimuli? I hypothesised that sub-threshold visual inputs in auditory cortex could modulate spiking activity and lead to more robust spiking responses and subsequently better decoding.

Dynamic visual stimuli elicit reliable changes in the phase of the LFP

Reliable changes in the evoked LFP were evident both in the voltage trace and in time-frequency plots for auditory classified (Figure 3.4a, b) and visual classified (Figure 3.5a,b) units.

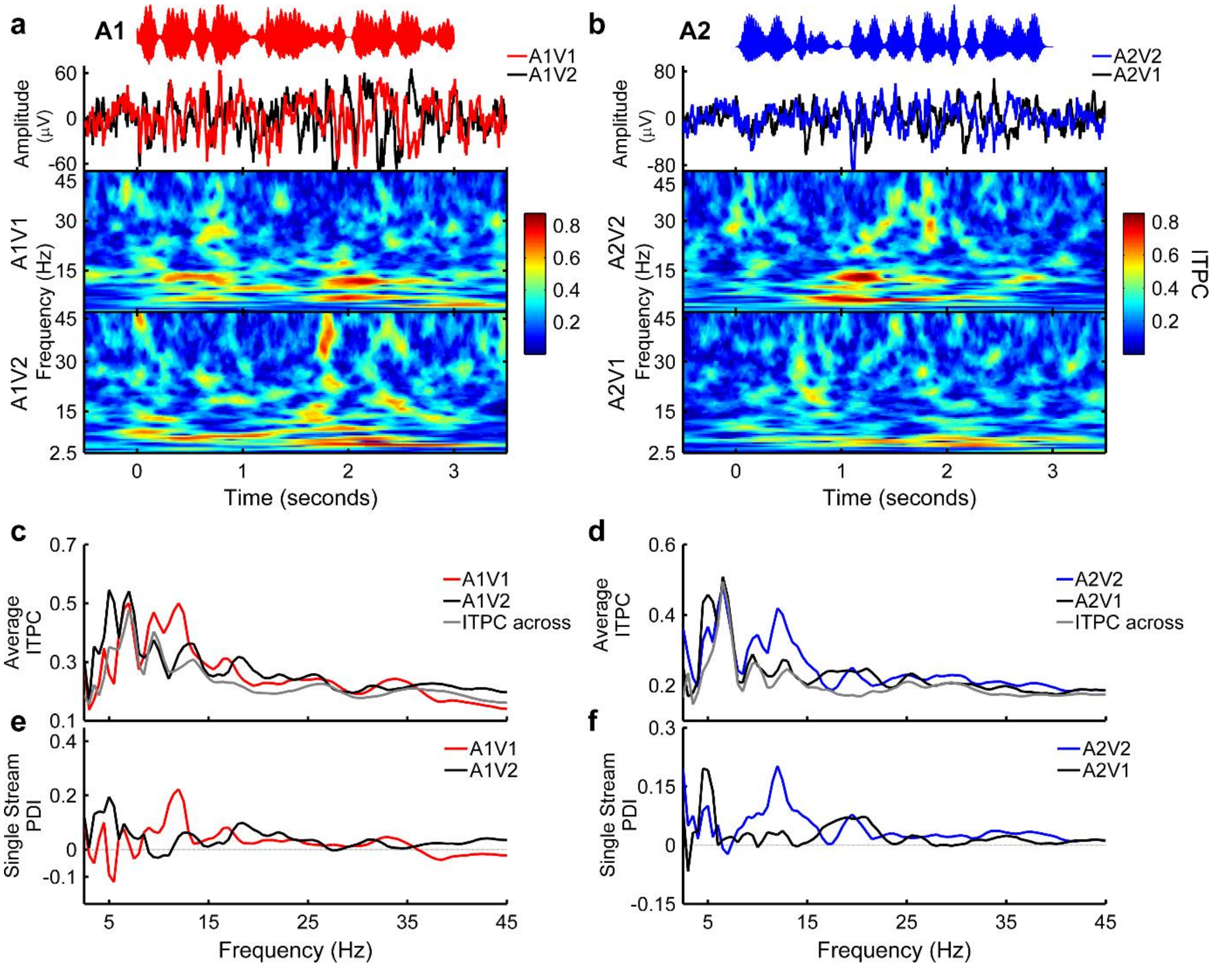

Figure 3.4 Temporal coherence between auditory and visual stimuli elicit reliable changes in the phase of the local field potential

a, b Local field potential (LFP) recordings in response to single stream stimuli. Responses to A1 are shown in (a) and A2 in (b). The amplitude waveforms of the stimuli are shown in the top panel, with the evoked LFP (for both visual stimuli) and resulting intertrial phase coherence values beneath for a typical recording site at which multiunit spiking activity accurately encoded auditory stream identity ('auditory classified' in figure 2). **c, d** Inter trial phase coherency (ITPC) values were calculated for both coherent and independent AV stimuli and compared to a null distribution (ITPC across). Single stream (SS) phase dissimilarity values were calculated by comparing ITPC values to the null distribution (**e, f**).

I calculated phase and power dissimilarity functions for stimuli with identical auditory signals but differing visual ones (Luo and Poeppel, 2007); briefly, if the phase (or power) within a particular frequency of a neural signal can discriminate between two stimuli, then the phase (or power) values across repetitions of the same stimulus will be more similar than across repetitions of randomly chosen stimuli.

I, therefore, calculated the ITPC for LFP responses to a single stimulus combination (e.g. each of A1V1, A1V2, A2V2, A2V1 yielding 'within-group' values) and compared the resulting values for trials that had the same auditory signal but randomly selected visual stimuli (i.e. for A1V1 and A1V2 within-group values were compared to across-group values calculated by randomly drawing trials that were either A1V1 or A1V2, Figure 4c, d). The cross-trial coherence was

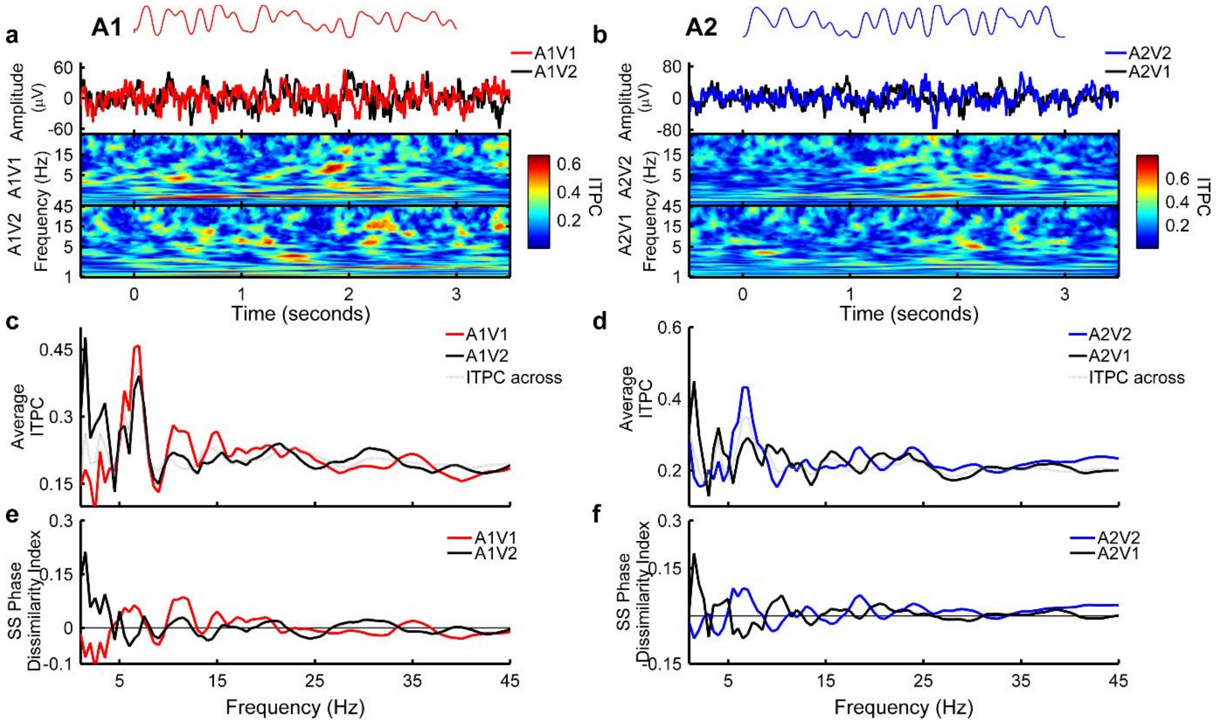

Figure 3.5 Temporally coherent visual stimuli increase local field potential reliability in visual classified unit.

a,b Local field potential (LFP) recordings in response to single stream stimuli. Responses to A1 are shown in (a) and A2 in (b). The envelopes of luminance change are shown in the top panel, with the evoked LFP (for both visual stimuli) and inter-trial phase coherence plots beneath for an example recording site at which multiunit spiking activity accurately encoded envelope of luminance change in visual stimuli ('visual classified' in figure 2). **c, d** Inter trial phase coherency (ITPC) values were calculated for both coherent and independent AV stimuli and compared to a null distribution (ITPC across, see methods). Single stream (SS) phase dissimilarity values were calculated by comparing ITPC values to the null distribution (**e, f**).

calculated as a function of frequency and compared for within and across group signals to yield a single stream phase dissimilarity index (single stream PDI, Figure 3.4e-f for auditory classified unit and Figure 3.5e-f for visual classified unit). Positive single stream PDI values support the hypothesis that the temporal dynamics of the visual stimulus elicits reliable changes in the phase of the on-going field potential. Importantly, because the across-group trials have the same auditory stimulus as the within group trials, any significant phase dissimilarity values can only result only from effects elicited by the visual stimulus.

To determine at what frequencies the across-trial phase reliability was significantly non-zero, I performed a pairwise comparison of the $\text{ITPC}_{\text{within}}$ (averaged across coherent Figure 3.6a, d and independent stimuli Figure 3.6b, e) and $\text{ITPC}_{\text{across}}$ conditions (Bonferroni corrected for 45 frequencies). In the awake dataset, this yielded a restricted range of frequencies between 10.5Hz and 21 Hz (Figure 3.6c), whereas in the anesthetised dataset all frequencies tested

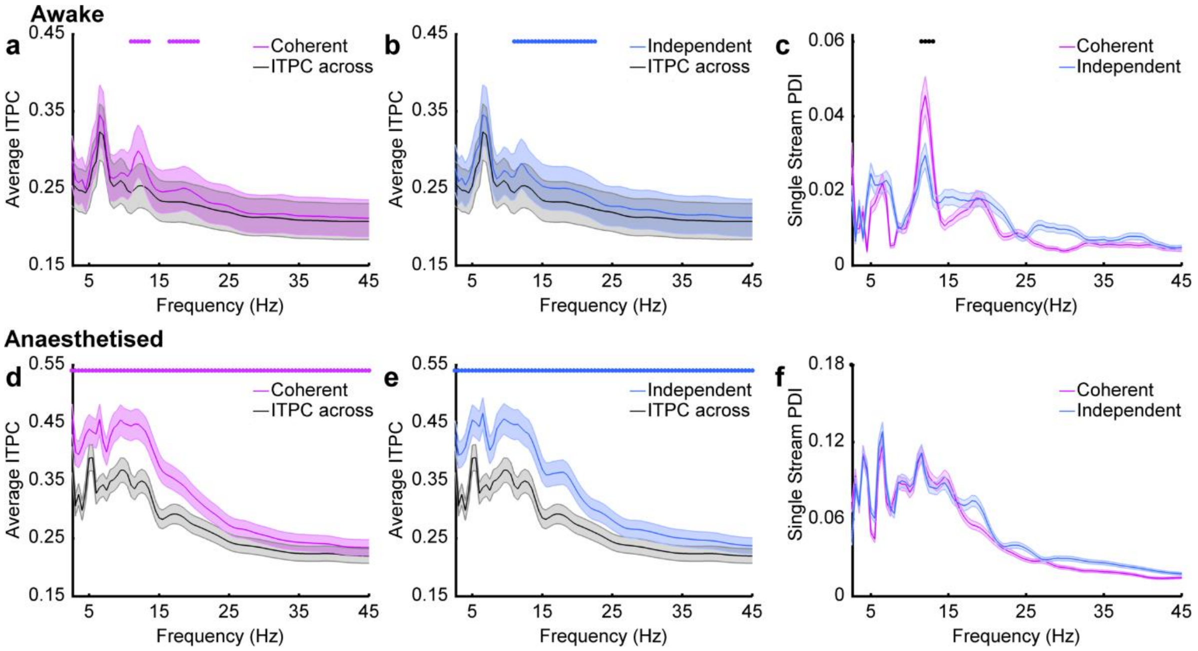

Figure 3.6 Visual stimuli elicit reliable changes in LFP phase in awake and anesthetised animals.

Mean inter-trial phase coherence (ITPC) values across frequency for coherent (a, d) and independent (b, e) conditions. Dots indicate frequencies at which the ITPC values were significantly greater than chance (permutation test, p = 0.0012, Bonferroni corrected for 43 frequencies). c f: Mean (±SEM) single stream phase dissimilarity index (PDI) values for coherent and independent stimuli in awake (c) and anaesthetised (f) animals. Black dots indicate frequencies at which the coherent stream PDI is significantly greater than in the independent conditions (p<0.001).

had SS PDI values that were significantly greater than zero (Figure 3.6f). I then asked whether there were any frequencies at which temporally coherent stimuli yielded greater SS PDI values by performing a pairwise comparison of values yielded from temporally coherent and independent stimuli, for all frequency points. Only in the 11-14 Hz and 19-20 Hz band, in the awake dataset were SS PDI values significantly higher when stimuli were temporally coherent than in the independent case. In the anaesthetised dataset there were no frequencies at which the phase dissimilarity index for the coherent stimuli was greater than for the independent stimuli.

To assess whether the observed changes in the phase of the field potential are accompanied by corresponding changes in power, I repeated these analyses to yield the "SS power dissimilarity index," characterizing the difference in the across-trial power coherence between "within-condition" and "across-condition" signals. Visual stimuli did not elicit reliable changes in LFP power across trials and values obtained were not significant for any frequency tested in either dataset (Figure 3.7). Therefore sub-threshold visual inputs modulate the phase of the on-going oscillations in auditory cortex providing a mechanism for enhancing the neural coding of temporally coherent auditory and visual signals.

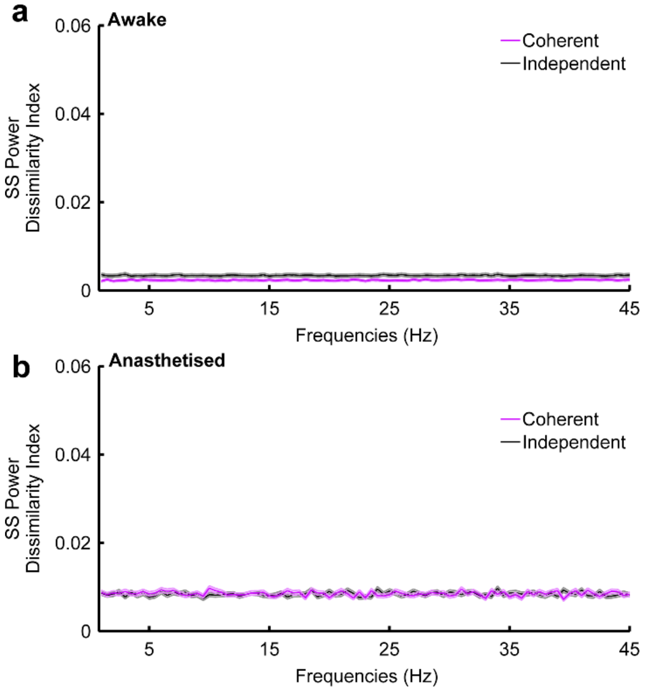

Figure 3.7 Dynamic visual stimuli do not elicit reliable changes in the power of the LFP

a, b: Mean (±SD) single stream phase dissimilarity index values for coherent (A1V1 and A2V2) and independent (A1V2 and A2V1) stimuli at all classified units recording for awake (a) and anaesthetised (b) datasets.

Visual information enhances the representation of the temporally coherent auditory stream in a sound mixture

Sound sources in the world must be reconstructed from their representation at the cochlear requiring that the auditory brain segregate the overlapping representation elicited by competing sound sources. Following the results above, which suggest that visual stimuli can enhance the representation of a temporally coherent sound, I asked whether the temporal dynamics of a visual stimulus could enhance the representation of one sound in a mixture. I therefore recorded responses to sounds A1 and A2 presented simultaneously (A12, 'dual stream') with a visual stimulus that was temporally coherent with one or other auditory stream. I extended the pattern classification approach used earlier by using the temporally coherent single stream conditions (A1V1 and A2V2) as templates for decoding dual stream

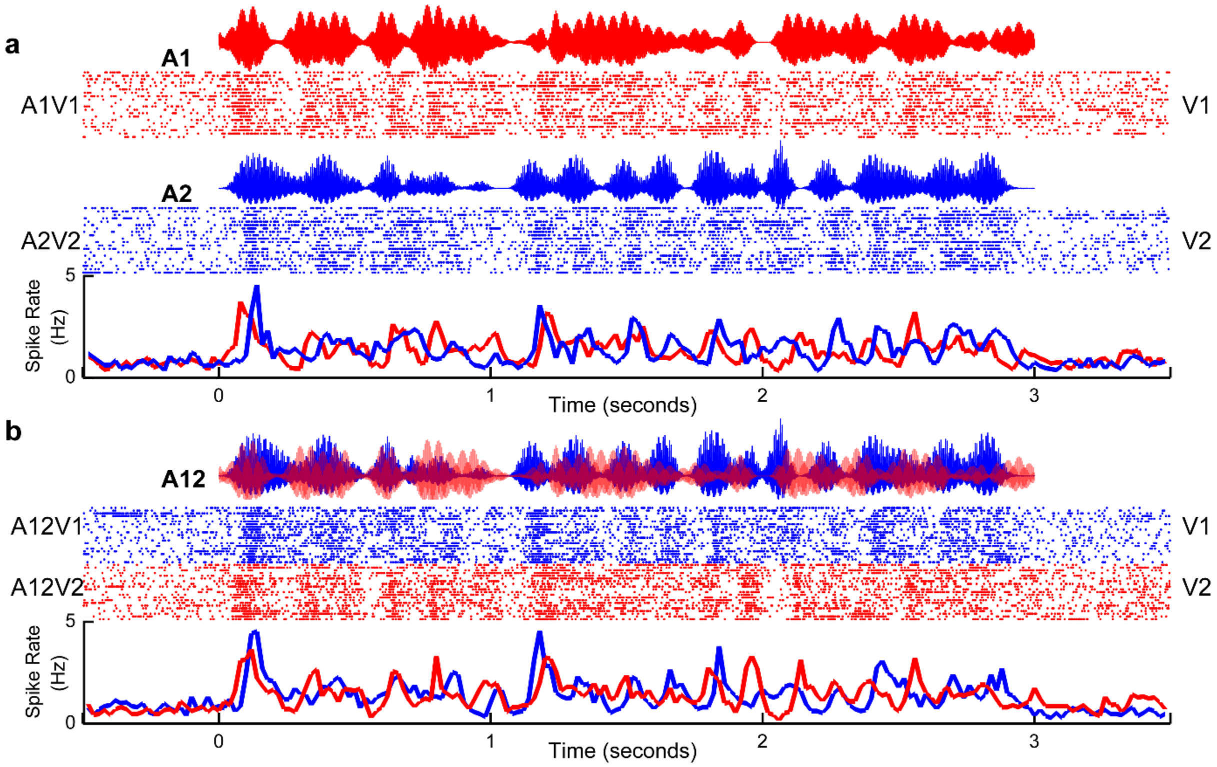

Figure 3.8 Visual stimuli can determine which sound stream auditory cortical neurons follow in a mixture

The spiking responses of an example unit are shown to a, single stream coherent stimuli A1V1 (top) and A2V2 (bottom). b, responses to stimuli comprised of two simultaneously presented auditory streams (A12) and one visual stream (either V1, top, or V2, bottom). In each panel rasters and PSTHs are shown. This example unit was a visual classified unit recorded from an awake animal. 78% of responses classified as A1 when visual stimuli was A1V1 and 76% of responses classified as A2 when visual stimuli was A2V2.

responses with decoding performance expressed as the proportion of responses classified as being from the A1 stream.

Figure3.8 illustrates this approach: the single stream templates used for decoding are shown in 3.8a and the spiking responses elicited to the dual stream stimuli are shown in 3.8b. The decoder classified responses to the dual stream stimuli as predominantly A1 when the visual stimulus was V1 (78% of trials classified as A1), and A2 when the visual stimulus was V2 (76% of responses classified as A1). The responses of this unit classified AV stimuli according to their visual dynamics, but many auditory units showed similar response properties (Figure 3.9).

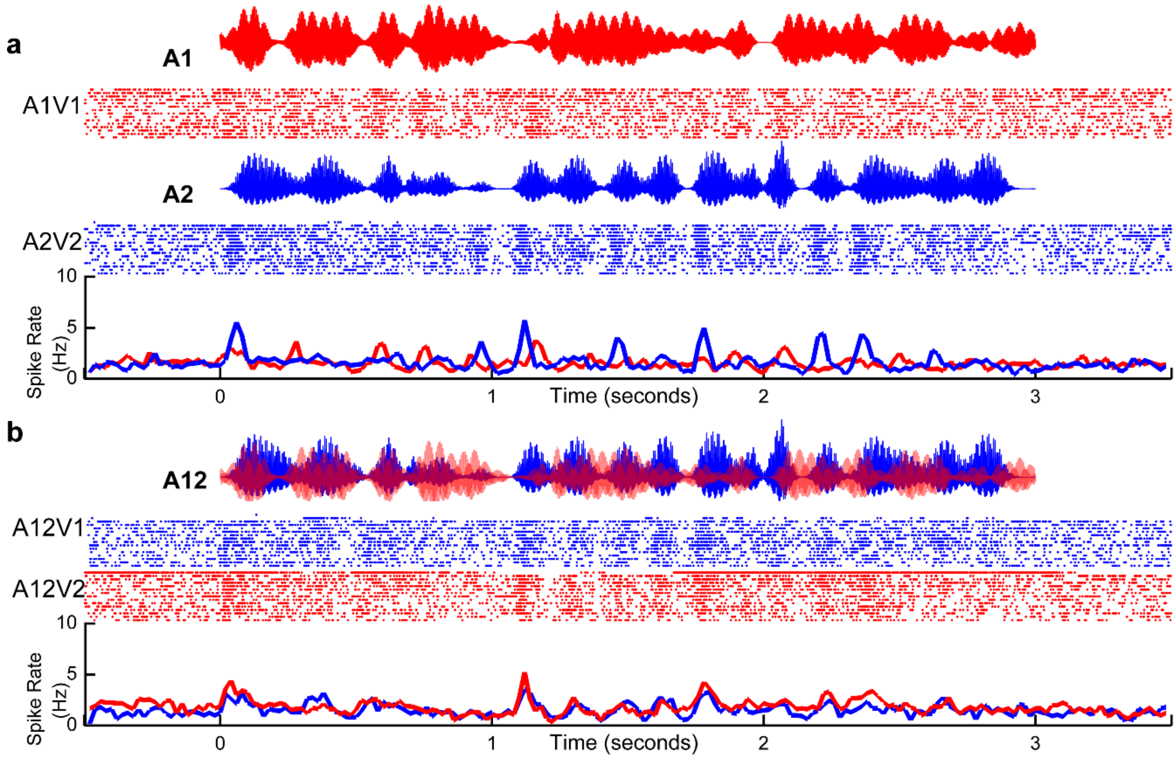

Figure 3.9 Visual stimuli can determine which sound stream auditory cortical neurons follow in a mixture

The spiking responses of an example unit are shown to a, single stream coherent stimuli A1V1 (top) and A2V2 (bottom). b, responses to stimuli comprised of two simultaneously presented auditory streams (A12) and one visual stream (either V1, top, or V2, bottom). In each panel rasters and PSTHs are shown. This example unit was an auditory classified unit recorded from an awake animal. When the visual stimulus was V1 67% of trials were classified as A1V1, when the visual stimulus was V2 only 71% of trials were classified as A2V2.

These examples were born out at the level of the population: the temporal property of the visual stimulus was able to bias which of the two competing auditory streams was represented by auditory cortical neurons. This finding was robust in both awake (Figure 3.10a, b pairwise t-test, all classified: t_{540} = 6.0737, p<0.00) and anesthetised datasets (Figure 3.10c, d t_{660} = 9.514, p<0.001), suggesting that this effect was not mediated by attention.

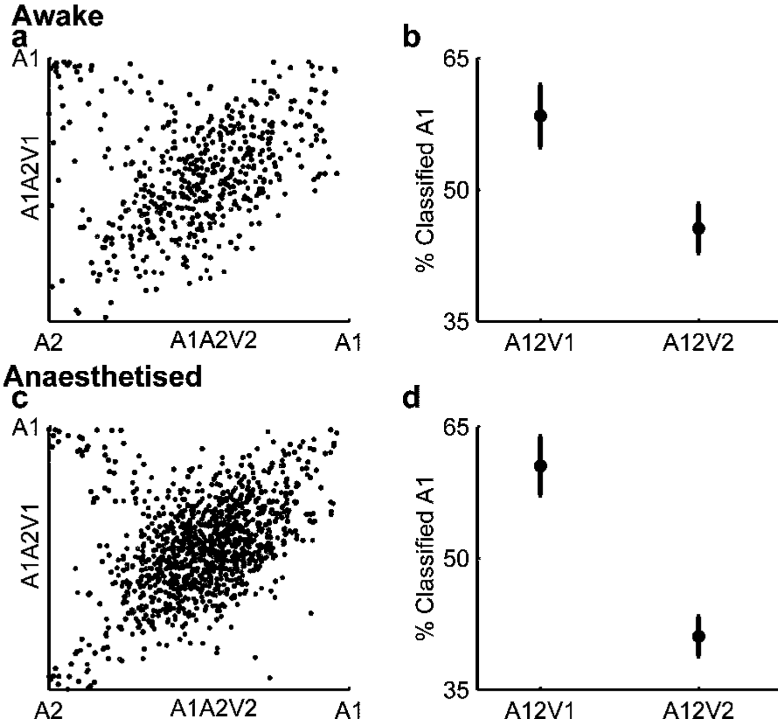

Figure 3.10 When two sounds are presented in competition auditory neurons respond preferentially to the auditory stream with which the visual stimulus was coherent

Neural responses to dual stream stimuli were classified as either A1 or A2 (using templates from single stream conditions, see Figure 3.8). **a,c** show the classification performance of all neurons (expressed as the proportion of trials classified as ‘A1) in dual stream condition (each dot = one neuron) for each visual condition. **b**,**d***:* Mean (± SEM) values for units recorded in auditory cortex. **a**,**b** awake dataset, **c**,**d**, anesthetized dataset. The pairwise comparison revealed significant effect across visual conditions (p<0.001).

No preference was found in no visual condition

In order to exclude the speculation of decoding the identity of the visual stimulus from a subpopulation of visual units as I have used auditory-visual templates (A1V2 and A2V2), in an additional experiment (n = 58 units, in passively listening awake ferrets) I recorded responses to auditory-only sounds (A1 and A2) presented simultaneously (A12), in a 'dual stream' condition with a visual stimulus that was temporally coherent with one or other auditory stream (A12V1 or A12V2). I extended the pattern classification approach used above by taking the responses to two sounds and the temporally coherent single stream conditions as templates for decoding dual stream responses and analysed responses to mixed auditory streams with no visual stimulus (A12) using responses either to coherent single stream stimuli (A1V1, A2V2).

A two-way repeated measures ANOVA on the decoder responses with factors of visual stream (V1, V2, no visual), and template type (AV or A) demonstrated a significant effect of visual stream identity on dual stream decoding (Fig. 3.11, F(2, 528) = 19.320, p <0.001), but there was no effect of template type (F(1, 528) = 0.073, p = 0.787) or interaction between factors (F(2, 528) = 0.599, p = 0.550). Post-hoc comparison across units revealed that without visual stimulation there was no tendency to respond preferentially to either stream but that visual stream identity significantly influenced classification of dual stream responses.

These data demonstrated that in the absence of visual stimulation there was no tendency across units to respond preferentially to either stream and that when dual stream auditory responses (in the presence of visual stimulation) were decoded with an auditory-only template a bias in favour of representing the temporally coherent stream persisted (Figure 3.11).

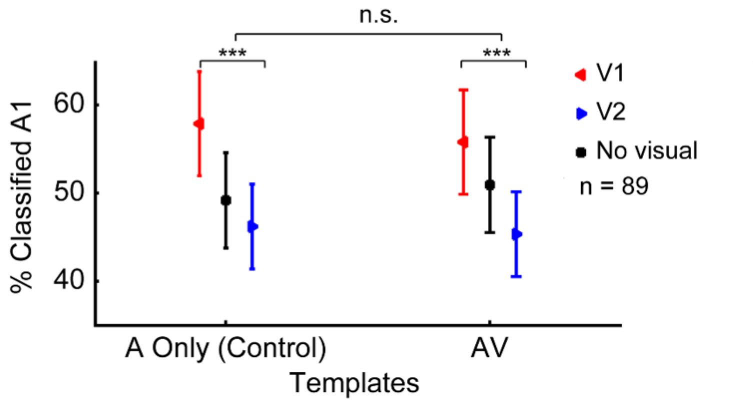

Figure 3.11 No preference was found in no visual condition.

Mean (± SEM) values for these units. Pairwise comparison revealed significant effect across visual conditions in both datasets (p<0.001).

Preference to the auditory stream with which stimuli was coherent is not field or layer specific

In the anesthetised animal, we recorded responses to pure tone stimuli to determine tonotopic gradients, so as to confirm the cortical field in which any given recording was made and LFPs were subjected to current source density (CSD) analysis to identify the cortical layers. Preference to the auditory stream with visual stimuli coherent was found in supragranular layers (SG, t_{700} = 5.686, p<0.001), granular layer (G, t_{878} = 3.481 p<0.001) and intragranular layer (IG, t_{690} = 4.418, p<0.001), suggesting that such a visual effect is not layer specific. A two-way ANOVA across visual condition and cortical layers showed only a significant effect of visual condition (F (1, 2273) = 64.288, p<0.001) but no layers effect. (F (1, 2273) = 0.91, p=0.404; visual condition x fields: F (1, 2273) = 2.679, p = 0.068).

Similarly, (A1, t_{798} = 5.435, p<0.001; AAF, t6$_{36}$ = 4.302, p<0.001; PPF, t_{510} = 3.609, p<0.001; PSF, n=159, t = 0.932, p =0 .352). A two-way ANOVA across visual condition and cortical fields showed only a significant effect of visual condition (Visual condition: F (1, 2267) = 40.301,

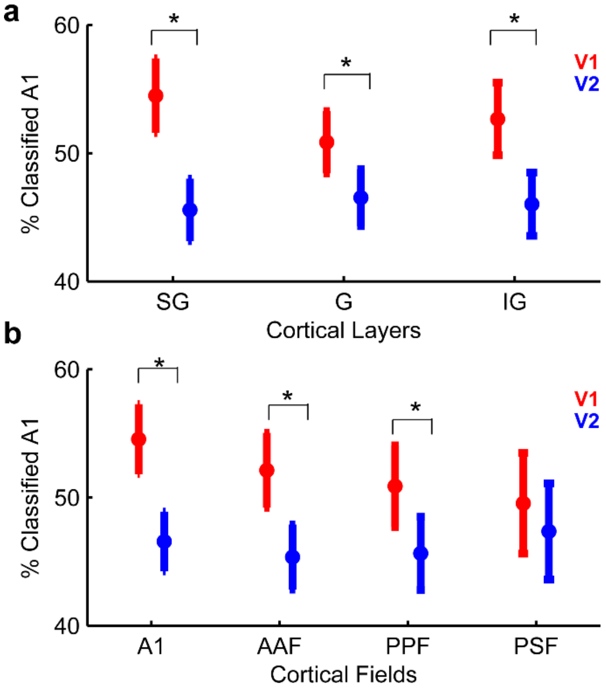

Figure 3.12 Preference to the auditory stream with which the visual stimulus was coherent is not layer specific nor field specific.

Mean (± SEM) values of the proportion of responses classified as A1 when visual stimuli were V1 or V2 are shown across different cortical layers **a**, and cortical fields **b**.

p<0.001; fields: F (1, 2267) = 1.937, p= 0.121; visual condition x fields: F (1, 2267) = 1.804, p = 0.144).

Enhancement representation of the auditory steam which was temporally coherent with the visual stimuli in auditory only responsive units

In the anesthetised animal, we recorded responses to 100 ms noise bursts and/or LED flashes in order to characterise units as either auditory, visual or auditory-visual based on a 2-way ANOVA on spike counts calculated over a 200ms window with auditory and visual stimulus as factors. Units in which both auditory and visual stimuli significantly modulated spiking or in which there was a significant auditory x visual interaction term were classified as auditory visual (Figure 3.13). This analysis yielded auditory only (red n = 177), visual only (blue, n = 130) and auditory visual driven units (grey, n = 150). I then used these classifications to determine whether the influence of visual stimuli on the representation of an auditory scene was seen across all neuronal subtypes. The enhancement of the representation of the auditory stream which was temporally coherent with the visual stimulus was observed in visual, auditory-visual and auditory neurons (Pairwise t-test, p<0.001 with correction).

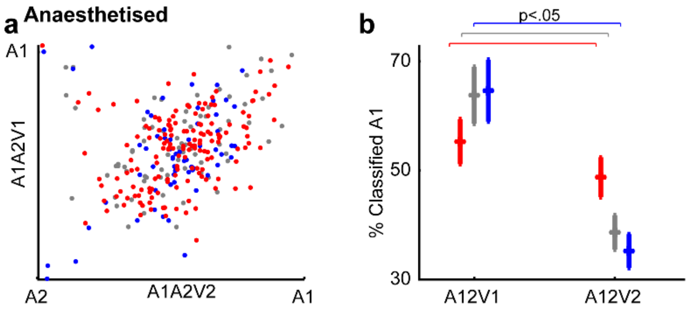

Figure 3.13 Selective representation can be seen in auditory only, visual only and auditory visual units under anaesthesia.

a, The proportion of responses classified as A1 when visual stimuli as V1 or V2 is plotted with each unit color coded according to whether it was classified as A, V or AV. **b**, mean (± SEM) values across the population. A pairwise comparison revealed a significant effect in all subgroups (p<0.05).

Selective representation of AV temporal coherence on stream segregation

Since temporally coherent AV stimuli elicited more reliable sub-threshold activity, as assessed by the across-trial phase coherence, I reasoned that this may provide the mechanism by which the visual stimulus could shape the neural representation of the auditory scene. To determine whether this was the case, I again measured the across trial phase coherence calculating for neural responses to identical auditory stimuli but differing visual conditions (i.e. A12V1 or A12V2, 'within group') and trials that had the same auditory signal but randomly selected visual stimuli (A12 randomly drawn trials that were either V1 or V2, 'across group'). Since the auditory streams are identical in all cases, significant phase-selectivity values indicate that the

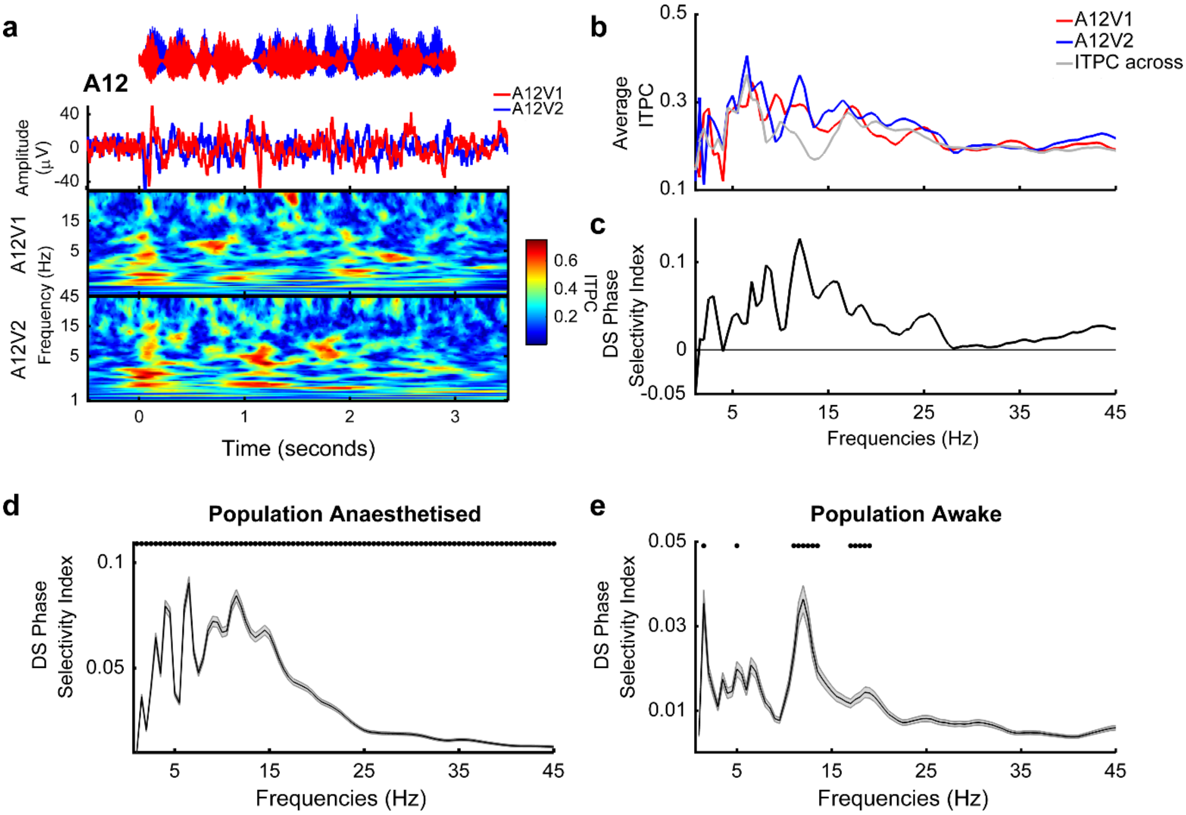

Figure 3.14 Visual stimuli elicits reliable phase patterns in auditory cortex to shape auditory scene analysis

a, stimulus waveforms are shown in the top panel, with the evoked LFP (for both visual stimuli) and resulting inter-trial phase coherency plots beneath for a typical recording site in response to dual stream stimuli. **b**, ITPC values were calculated across frequency for A12V1 (red) and A12V2 stimuli (blue) and for a null distribution (green, shuffled trials) and used to determine dual stream (DS) phase selectivity index values (**c**)**. d, e** Mean (± SEM) DS phase selectivity index values for awake and anaesthetised datasets. Symbols indicate where the DS phase selectivity index was significant (pairwise ttest, p<0.05 with correction).

time course of the neural response is significantly influenced by the dynamics of the visual stimulus. In the awake dataset, significant phase selectivity was once again found at 12-14Hz (Figure 3.14e, $p< 0.001$, and paired t-test between ITPC$_{\text{within-visual}}$ and ITPC$_{\text{across-visual}}$ values, with Bonferroni correction for 49 frequencies). Increased phase reliability at alpha frequencies was observed independently of the amplitude modulation rate of the auditory stimulus (Figure 3.15a-c). In the anesthetised dataset, I found significant phase selectivity across all frequencies tested (Fig. 3.14d, $p< 0.001$).

Phase selectivity was observed in visual, auditory-visual and auditory neurons (Figure 3.15) which characterized based on whether units respond to the onset of a discrete stimulus event, as described above.

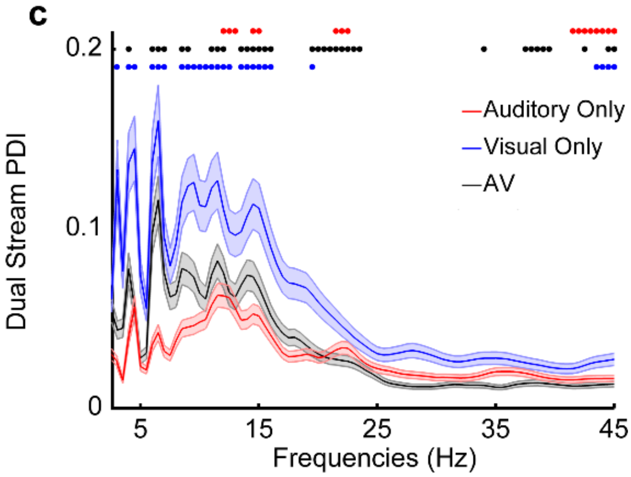

Figure 3.15 Selective representation can be seen in auditory only, visual only and auditory visual units under anaesthesia.

Mean (± SEM) DS phase selectivity index values for all subgroups. Symbols indicate where the DS phase selectivity index was significant (pairwise ttest, p<0.05 with correction).

Significant alpha range are independent of amplitude modulation rate

In an additional control experiment (n = 58 units, in passively listening awake ferrets), I have recorded same set of stimuli for dual stream condition (A1V1, A2V2m A12V1m A12V2) with 3 different amplitude modulation rates (7Hz, 12Hz and 17Hz) in order to examine the stimuli driven change in frequency bands (Figure 3.16). Pairwise comparison (p<0.05 with Bonferroni correction) revealed that in all three cases significant phase coherence was seen between 10Hz-11.5Hz, 19Hz-20Hz and 24-26 Hz.

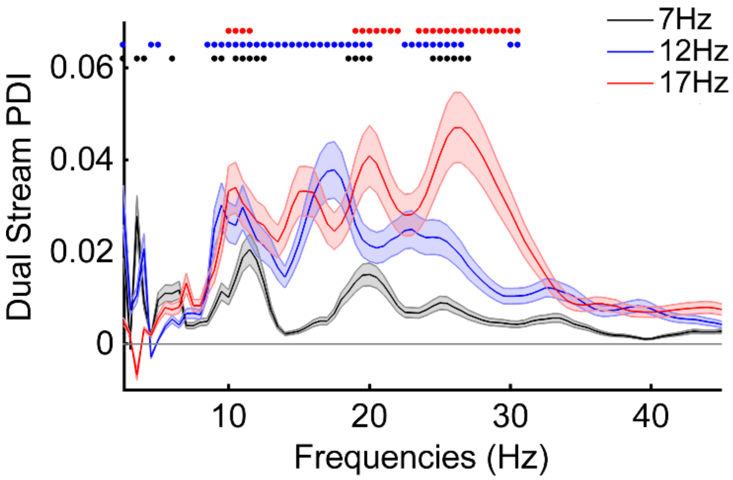

Figure 3.16 Dual stream PDI values in the alpha range are independent of amplitude modulation rate.

Mean (± SEM) DS phase selectivity index values for three different amplitude modulation rates (7Hz, 12Hz and 17Hz. Symbols indicate where the DS phase selectivity index was significant (pairwise ttest, p<0.05 with correction).

Neural responses to auditory timbre deviants are enhanced when changes in visual luminance and auditory intensity are temporally coherent

A hall-mark of an object-based rather than feature-based representation is that all stimulus features are bound into a unitary perceptual construct, including those features which do not directly mediate binding (Desimone and Duncan, 1995). I predicted that binding across modalities would be promoted via synchronous changes in auditory intensity and visual luminance and observed that the temporal dynamics of the visual stimulus enhanced the representation of temporally coherent auditory streams. To determine whether temporal synchrony of visual and auditory stimulus components also enhanced the representation of orthogonal stimulus features and thus fulfil a key prediction of binding, I introduced brief timbre perturbations into our dual stream stimuli (n = 4 deviants, two in A1 and two in A2). Such deviants could be detected by human listeners and were better detected when the auditory stream in which they were embedded was temporally coherent with an accompanying visual stimulus (Maddox et al., 2015). I hypothesised that despite containing no information about the occurrence of deviants, a temporally coherent visual stimulus would enhance the representation of changes in timbre in the responses of auditory cortical neurons.

To isolate neural responses to the timbre change from those elicited by the on-going amplitude modulation, I extracted the 200ms epochs of the neuronal response during which the timbre deviant occurred and compared these to epochs from responses to otherwise identical stimuli without deviants. I observed that the spiking activity of many units differed between deviant and no-deviant trials (e.g. Figure 3.17a) and so I used a pattern-classifier approach to estimate the presence/absence of a timbre deviant in a given response window. I first considered the influence of temporal coherence between auditory and visual stimuli on the representation of timbre deviants in the single stream condition (A1V1, A1V2 etc.).

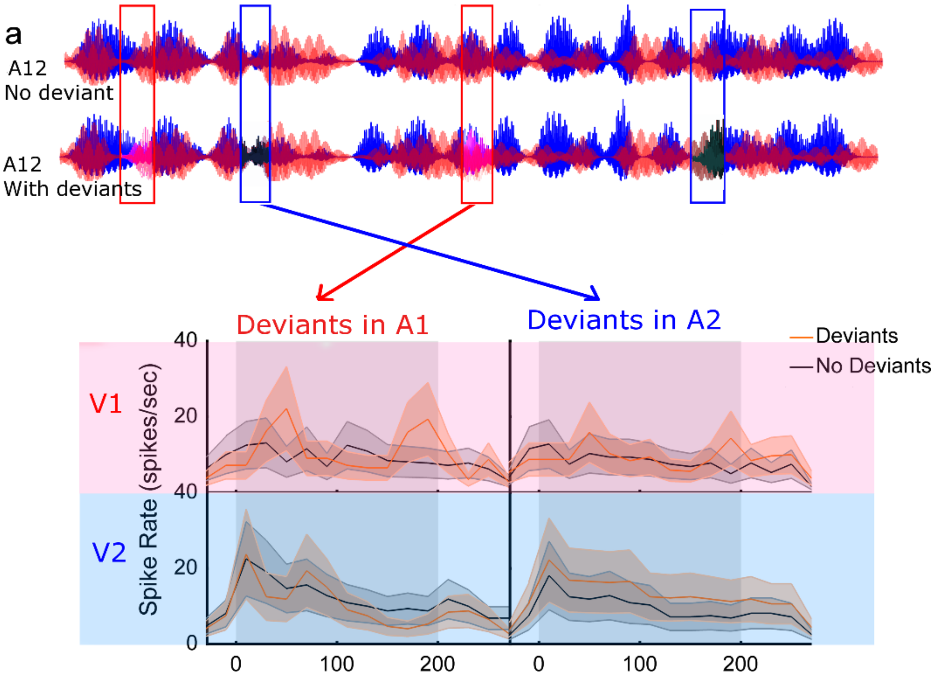

Figure 3.17 Example unit (from the awake dataset) showing the influence of visual temporal coherence on spiking responses to dual stream stimuli with or without deviants embedded.

Shaded rectangles indicate the 200 ms window over which the timbre deviant occurred and over which analysis was conducted.

I found that a greater proportion of units detected at least one deviant when the auditory stream in which deviants occurred was temporally coherent with the visual stimulus relative to the temporally independent condition. This was true both for awake (Figure 3.18a; Pearson chi-square statistic, χ2 = 322.617, p < 0.001) and anesthetised animals (Figure 3.18d; χ2 = 288.731, p < 0.001). For units that discriminated at least one deviant, discrimination scores were significantly higher when accompanied by a temporally coherent visual stimulus (Figure 3.18b, awake dataset, pairwise t-test t300 = 3.599 p<0.001; Figure 3.18e, anesthetised data t262 = 4.444 p<0.001).

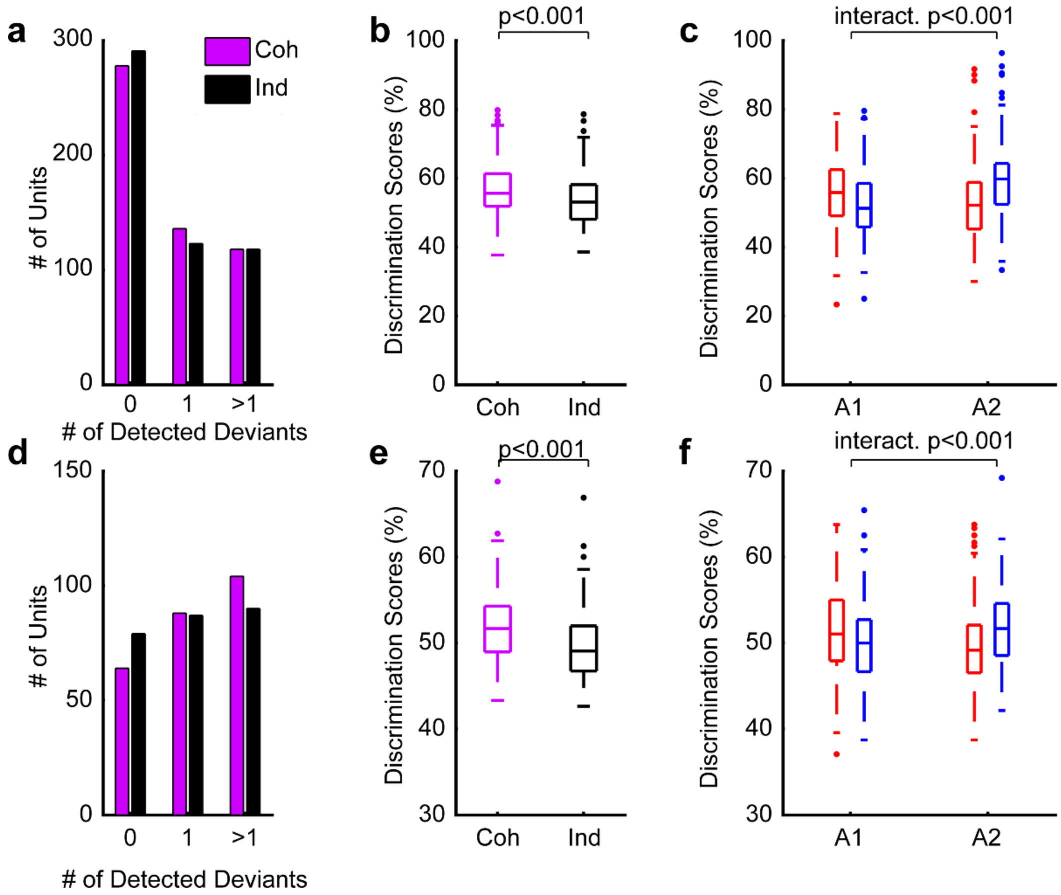

Figure 3.18 Temporally coherent changes in visual luminance and auditory intensity enhance the coding of another auditory feature.

a, Histogram showing the number of units in which spike responses could discriminate trials in which deviants occurred. Two deviants were included in each auditory stream giving a possible maximum of four in awake dataset. **b**, Box plots showing the timbre deviant discrimination scores in the single stream condition across different visual conditions (Coh: coherent, ind: independent). The boxes show the upper and lower quartile values, and the horizontal lines at their “waist” indicate the median. Awake dataset. **c**, Discrimination scores for timbre deviant detection in dual stream stimuli. Discrimination scores are plotted according to the auditory stream in which the deviant occurred and the visual stream that accompanied the sound mixture. **d-f** show the same as b-d but for the anesthetised dataset

Across the population of units, I performed a two-way repeated measures ANOVA on discrimination performance with visual condition (V1/V2) and the auditory stream in which the deviants occurred (A1/A2) as factors. I predicted that enhancement of the representation of timbre deviants in the temporally coherent auditory stream would be revealed as a significant interaction term. Significant interactions were seen in both the awake (Figure 3.18c, F (1, 600) = 29.138, p<0.001) and anesthetised datasets (Figure 3.18f, F (1, 524) = 16.652, p<0.001). We also observed significant main effects of auditory and visual conditions in awake (main effect of auditory stream, F (1, 600) = 4.565, p = 0.033; main effect of visual condition, F (1, 600) = 2.650, p = 0.010) but not anesthetised animals (main effect of auditory stream, F (1,524) = 0.004, p = 0.948; main effect of visual condition, F (1, 524) =1.355, p = 0.245). Thus, these findings suggested that a temporally coherent visual stimulus can enhance the representation of features (here auditory timbre) orthogonal to those that promote binding between auditory and visual streams. This finding is consistent with our model of cross-modal binding and so these data fulfil our definition of binding.

Discussion

Here I provide mechanistic insight into how auditory and visual information are bound together in the auditory cortex. The data presented here demonstrated that visual stimuli elicit reliable changes in the phase of the local field potential and an enhanced spike-based representation of auditory information within auditory cortex. When two sounds are presented in competition within an auditory scene, the representation of the stream that is temporally coherent with the visual stimulus is enhanced. Such enhanced representation was shown across different cortical layers including superficial and deep layers, different cortical fields and different unit types, suggesting that it is a general phenomenon in the auditory cortex.

Importantly, this enhancement is not restricted to the encoding of the amplitude changes that bind auditory and visual information in the anaesthetised animals, the encoding of auditory timbre, a stimulus dimension orthogonal to the dimensions that link auditory and visual stimuli in the anaesthetised. These data provide a physiological underpinning for the advantage conferred upon listeners in an auditory selective-attention task when a visual stimulus is temporally coherent with the target auditory stream. However, I could not observe such effect in awake animals, it might be either because of no controlled listening condition in awake recordings or the top-down connections has inhibited such an effect in AC.

In the awake animal, the impact of visual stimulation on LFP phase reliability was smaller than in the anesthetised animal and was restricted to a narrower range of frequencies, consistent with a dependence of oscillatory activity on behavioural state (Tukker et al., 2007, Voloh and Womelsdorf, 2016, Wang, 2010). Since the neural correlates of multisensory binding are evident in the anesthetised animal, the specific increase in alpha phase reliability that occurred in awake animals in response to temporally coherent auditory-visual stimulus pairs (Fig. 4c & 7e) may indicate an attention-related signal triggered by temporal coherence between auditory and visual signals. Phase resetting or synchronisation of alpha phase has been associated both with enhanced functional connectivity (Voloh and Womelsdorf, 2016) and as a top-down predictive signal for upcoming visual information (Samaha et al., 2015). Disambiguating these possibilities would require simultaneous recordings in auditory and visual cortex and/or recording during the performance of a task designed to explicitly manipulate attention.

Although AV effect was observed across layers, fields, and anesthetic states, it is important to point out that many other aspects of neural coding for sound are not expressed similarly across these dimensions. For example, temporal encoding (Christianson et al., 2011), spectrotemporal receptive field complexity (Atencio et al., 2009), and connectivity patterns (Llano and Sherman, 2009; Oviedo et al., 2010) vary considerably between cortical laminae. Similarly, anesthesia inevitably has an influence on spiking activity, therefore, many aspects of cortical coding of sound stimuli (Wang et al., 2005; Petkov et al. 2007), and comparison of the LFP and unit firing can reveal many diverse aspects of stimulus coding (Lakatos et al., 2007; O'Connell et al., 2011). Although neurons in auditory cortex gave stronger responses to the sound onset and offset in awake animals, the capacity of neurons were similar in anaesthetised animals to follow rapid fluctuations in the stimulus waveform (Mickey and Middlebrooks, 2003). The smaller LFP phase reliability in awake animals might also due to the methodological differences in recordings made under awake and anaesthetised conditions.

Our data provide compelling evidence that one role for the early integration of visual information into auditory cortex is to resolve competition between multiple sound sources within an auditory scene. While previous studies have demonstrated a role for visual information in conveying lip movement information to auditory cortex (Golumbic et al., 2013, Chandrasekaran et al., 2013, Ghazanfar et al., 2005, Crosse et al., 2015), here I suggest a more general phenomena whereby visual temporal cues facilitate auditory scene analysis through the formation of cross-sensory objects. The origin of the visual inputs is an open question but both visual cortical and sub-cortical structures innervate tonotopic auditory cortex (Budinger et al., 2006, Bizley et al., 2007) and visual responses in auditory cortex can be disrupted through the inactivation of visual cortex in the ferret (Town, Wood, Atilgan and Bizley, unpublished results).

Temporal coherence between sound elements has been proposed as a fundamental organisational principle for auditory cortex (O'Sullivan et al., 2015, Elhilali et al., 2009a, Teki et al., 2016, Sohoglu and Chait, 2016) and here I extend this principle to the formation of cross modal constructs. The demonstration that temporal coherence between auditory and visual streams enhances the encoding of binding-orthogonal stimulus features is robust evidence that the effects I observe here are the neural correlates of multisensory binding (Bizley et al., 2016). That these effects are observed in the anesthetised auditory cortex is supportive of a bottom-up mechanism that promotes the formation of cross modal objects and provides a substrate on which selective attention can subsequently operate.

Chapter IV: Perceptual learning influences the ability of listeners to utilize visual cues to separate competing auditory streams

Introduction

Recent research in the study of AV integration has focused on the modulating effects of different forms of experience, including training. There has been extensive investigation on how experience during development affects multisensory processing (for an extensive review see (Spence and Deroy, 2012). As discussed in chapter I, some AV cross-modal correspondences seem to be innate or established very early in life, while others are generated via learning through experience. Results consistently show that AV cross-modal interactions that rely on temporal or spatial relations of the stimuli are already present early in life, while ones that rely on semantic relations are gradually developed later in life on the basis of exposure to relevant experience (Navarra et al., 2010b). However, it is still not clear how innate AV cross modal interactions are influenced by short term training.

In chapter II, we demonstrated that when the size of a visual stimulus was coherently modulated with a target auditory stream, human listeners were better able to report brief deviants in the target stream, than when the visual stream was coherent with the non-target auditory stream (Maddox et al., 2015, Chapter II). Auditory visual integration in early auditory cortex provides a potential mechanism for this advantage (Chapter III). However, there were great individual differences in subjects' ability to benefit from temporal coherence between the auditory and visual streams. Some people apparently benefit more from visual cues than others. In this chapter, I aimed to firstly determine whether these differences in participants' performance is due to their ability to detect AV coherence. The second aim of this study is to determine whether (i) exposure to temporally coherent AV stimuli or (ii) actively discriminating AV temporal coherence influenced the ability of listeners to exploit visual information for auditory scene analysis.

To address this issue, we recruited subjects randomly and assigned them to one of three groups. In the active training group, participants were trained on a two-interval forced choice task (2IFC) task in which they had to report in which interval a temporally coherent auditory visual stimulus pair was presented. In the exposure group, participants were trained on a two-interval 2IFC modulation rate discrimination task in which they were always exposed to temporally coherent AV stimuli while they were actively engaged with modulation rate discrimination. The third group of subjects did not perform any training. In each case, participants performed the auditory selective attention task described in Chapter II, and an additional test, AV coherency threshold test, to determine their sensitivity to temporal coherence between auditory and visual stimuli. In order to compare the effects of training these tests were completed twice, before and after training.

This study allowed me to examine whether the ability to discriminate temporal coherence between auditory and visual stimuli improves with training. In Chapter II, I initially hypothesised that temporal coherence driven binding will create a cross-modal object which might result in either *(i)* enhancing listener's ability to detect brief timbre deviants when the visual stimulus is coherent with the target stream and disrupting their ability when visual stimuli is coherent the distractor stream, or *(ii)* enhancing listener's ability to detect brief timbre deviants when the visual stimulus is coherent with either the target stream or the distractor stream as either case participants will be better able to separate the streams. I showed that first hypothesis (i) was the case. Here, I predicted either that;

(i) this effect would be enhanced such that the difference between target coherent and distractor coherent became greater or,

(ii) this effect will change in a pattern consistent with the second hypothesis - i.e. greater ability to exploit temporal coherence such that both coherent visual stimuli regardless of target or distractor auditory stream would enhance listeners performance by to segregate the streams more effectively relative to the independent condition.

Visual experience may help guide adaptation to changed auditory input as may be the case when the hearing is restored with a hearing aid or cochlear implant. AV training may help listeners adapt more effectively to hearing prosthetics (Isaiah et al., 2014) and so it is potentially of clinical relevance to establish whether listeners can improve their ability to exploit visual information.

Methods and Materials

Participants

42 adults (age range 18–34 years; mean age 28 years; 11 males) with normal hearing and normal or corrected-to-normal vision, participated in the study. Participants were randomly allocated to three groups and 12 participants in each group (N = 36, 6 participants were excluded after pre-test, see below). They were paid for their participation in the study and gave written informed consent to the study approved by the Ethics Committee of the University College London (ref: 5139). The data from four of the participants were excluded from analysis due to very low performance (across condition d'<0.8) and another two participants due to the low detection hit rates in visual detection task (<70%).

Stimuli

For the **auditory selective attention task,** the same set of auditory (and visual) stimuli were used as described in Chapter II. Briefly, two 14 second artificial vowel sounds with different F0s were used (175 and 195Hz, counter-balanced) with 200ms timbre deviants ([e] deviant in [u] stimuli and [i] deviant in [a] stimuli]. The two vowel streams were independently amplitude modulated with low pass <7 Hz envelopes. Unlike in Chapter II, I did not measure each subjects' ability to detect timbre deviants in order to set a difficulty threshold but instead used the average morph size for both timbre (12.5%) from the experiment in Chapter II and set a fixed difficulty level across all participants. For [i] deviants in [a] stimuli these corresponded to a maximum shift of 42 Hz in F1 frequency and 143 Hz for F2, and for [e] deviants in [u] stream there was a maximum shift of 75 Hz for F1, 196 Hz for F2 (adapted from Chapter II threshold result).

Visual stimuli consisted of a radius-modulated light grey disc on a black background. The disk was presented centrally (at fixation) and subjects were asked to report occasional visual deviants where the circumference of the ring briefly flashed cyan (Figure 2.1c). There were 96 trials with 480 deviants (192 target deviants, 192 masker deviants, 96 visual deviants). The mean number of the deviants per auditory stream was two and the mean number of the deviant per visual stream was one.

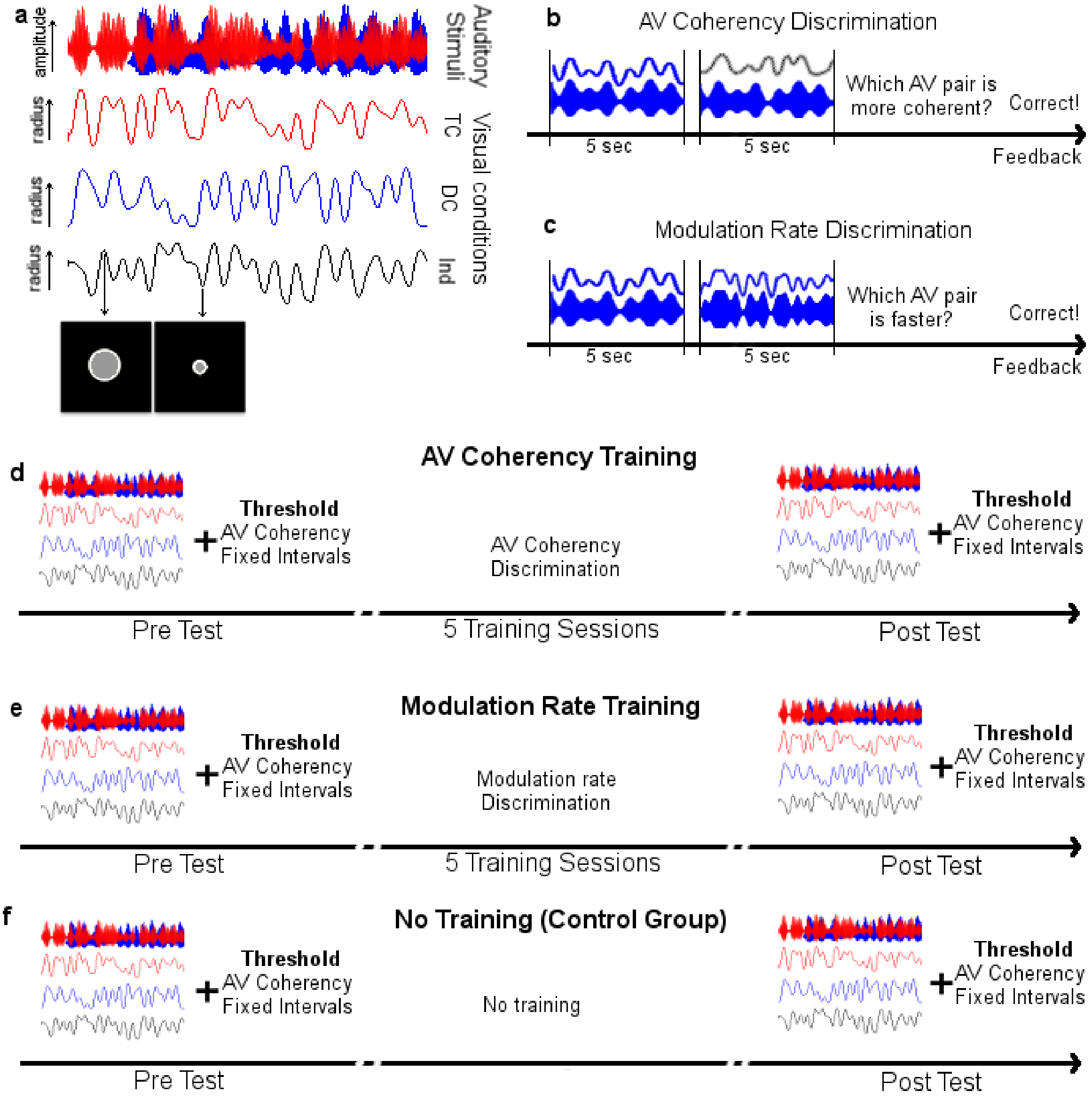

Figure 4. 1 The stimuli and schemas of training paradigms

Panel a shows the stimuli used in the auditory selective attention task, showing the initial 3 second segment of the auditory stimuli (full duration 14 seconds) and the temporal envelopes with which the visual stimulus could be modulated. The target stream is shown in red, and starts 1 second earlier than the distractor stream (blue). Panel b shows AV coherency discrimination paradigm, presenting one AV pair that auditory and visual stimuli was generated independently and one coherent AV pair that auditory and visual stimuli was generated with a constant coherency profile. Panel c shows the modulation rate discrimination paradigm, presenting two different AV pair in which AM vowel sounds accompanied with temporally coherent visual stimuli. Panel d, e and f show the procedure followed for AV coherency training, modulation rate training and control groups respectively. See text for details.

Stimuli for AV temporal coherency discrimination task: Two five second artificial vowel sounds were consecutively presented with 0.5 sec inter-stimulus interval with radius-modulated light grey disc on a black background. Both sounds were either [u] or [a], randomly selected for each trial. As in the main experiment, the envelope of the two auditory stimuli was amplitude modulated with low pass <7 Hz envelopes. The radius modulation in one of the visual stimuli was always orthogonally independent with the envelope of one of the auditory stimuli, while the other maintained some degree of temporal coherence. This could be fully temporally coherent (value of 100% coherence), or, by multiplying the temporally coherent envelope with another, independent, envelope, envelopes with varying coherence were generated.

For the AV coherency threshold test which all participants performed, stimuli were generated from 100% coherent in 10% steps to 10 % coherent and subjects performed 20 trials at each coherence level. In the sessions of AV coherency training, a three-down one-up rule was used to determine the coherence level of the stimulus in the next trial. In the first training session, the first stimuli generated at 100% coherent (easy to discriminate from the orthogonally independent envelope) and differed 10% steps for six reversals followed by 5% steps for following six reversals and 2.5% steps in the rest of the reversals. The procedure was terminated at 18 reversals unless a maximum of 150 trials was reached first. For the 2th-5th training session, the first stimuli generated with the average coherence level of the last ten reversal in the previous session. Each training session lasted not more than 40mins.

Stimuli for modulation rate discrimination task: Two five second artificial vowel sounds were presented with a 0.5 sec silent interval paired with radius-modulated light grey disc on a black background. Both sounds were either [u] or [a], randomly selected for each trial and was always coherent with the visual stimuli. One envelope was always generated with a 7 Hz cut off rate, whereas the other was generated with a higher rate (maximum AM cut off rate = 11Hz). In the sessions of modulation rate training, a three-down one-up rule was used to determine the modulation rate of the stimulus in the next trial. In the first session, the first stimulus was generated at the maximum modulation rate, and differed in modulation rate by 1Hz for the first six reversals and 0.5 Hz for the next six reversals and 0.25Hz for the rest of the trials. The procedure was terminated at 18 reversals unless a maximum of 150 trials was reached first. In each consequent session, the first stimulus was generated with the average coherence level of the last ten reversals in the previous session.

Procedure

The experiment consisted of three phases; pre-test, training sessions and post-test. In the pre-test the auditory selective attention task was performed, followed by the AV temporal coherency threshold test. It took 60-90 mins. In the five training sessions, each taking around 40 mins over a maximum of two weeks, participants carried out either AV coherency training or modulation rate training. Finally, the post-test was identical to the pre-test. Participants in the control group did no training sessions but performed the pre-test and post-test within 2 weeks (mean ± SD = 5days ± 3). Figure 4.1d-e and f illustrate the procedure for AV coherency training, modulation rate training and control respectively.

Pre-Post testing auditory selective attention task

Details of the auditory selective attention task can be found in the Stimuli section- Chapter II. Briefly, participants were instructed to selectively attend to the target stream, while ignoring the distractor stream. The target stream was cued by starting 1 second earlier than the distractor stream. Participants were asked to press a button whenever they detected a timbre deviant in the attended stream. Simultaneously, a single central visual stimulus was presented. Visual deviants, where the edge of the grey disk briefly changed cyan, were presented. Subjects were instructed to press a button if a visual deviant occurred.

There were three audio-visual coherence conditions; (1) Target congruent condition (TC); in which the envelope of the visual stimuli was matched to the target stream amplitude modulation envelope, (2) Distractor congruent condition (DC); the visual envelope matched the distractor stream envelope (3) Independent condition (Ind); the visual envelope was independent of the envelope of both auditory streams. (Figure 4.1a). Deviants were placed pseudo-randomly in the envelope. Across streams (auditory and visual) deviants could not occur within 1 second of one another. Deviants always occurred at times when all three envelopes had a value of >0.7 normalized amplitude to ensure that the signal-to-noise ratio of the two streams was matched.

AV coherency discrimination task (threshold and training sessions)

AV coherency discrimination thresholds were determined psychoacoustically in a two-interval forced choice task (2IFC) presenting one AV pair in which auditory and visual stimuli were generated independently and one AV pair in which auditory and visual stimuli had some degree of temporal coherency. Subjects were tested with 10 coherency levels for 20 repetitions each, in random order. Participants pressed "1" or "2" on the press box to indicate the interval of the coherent AV pair (Figure 4.1b). The same task was used for AV coherency training in which the coherency level of stimuli was generated based on three-up one-down rule. Feedback was provided on each trial in both threshold and training sessions.

Modulation rate discrimination task

The modulation rate discrimination task was a 2IFC task that adapts based on a three-down one-up rule. Participants were played two different AV pairs in which AM vowel sounds were accompanied by temporally coherent visual stimuli and asked to decide which one sounded faster. Participants pressed "1" or "2" on the press box to indicate the interval of the faster AV pair (Figure 4.1c). Feedback was provided on each trial in both threshold and training sessions.

Data analysis and Statistical analysis

As in chapter II, we measured performance by calculating d'.

We used a repeated-measures analysis of variance (ANOVA) to test for differences in the d', hit rates, false alarm, and visual hit rates across AV coherence condition and training (pre-test and post-test) and a mixed ANOVA for AV coherence condition, training and experimental group. Post-hoc comparisons were used to test the differences in the d', hit rates, false alarm, and visual hit rates.

Results

In order to examine whether the ability to discriminate temporal coherence between auditory and visual stimuli improves with training, I calculated coherence thresholds and before and after training in all three experimental groups (AV coherency training, modulation rate training and control group), this allow me to quantify the change in the ability to detect temporal coherence between auditory and visual stimuli. Then, hit rates, false alarms and d' in auditory selective attention task were calculated to assess the training effect on different AV coherence conditions.

AV temporal coherence trained group

In order to examine how actively discriminating AV temporal coherence affects the ability to detect AV coherency and whether this influences the ability to use visual stimuli to promote auditory scene analysis, twelve participants performed five AV coherency discrimination training sessions. My first aim was to establish whether this training had improved their ability to detect temporal coherence between auditory and visual streams. For each training session, I have calculated the mean coherency values and found that after five training sessions, 10/12 participants showed lower threshold values in the last session (S5) than the first session (S1). Pairwise comparison revealed that AV temporal coherence training significantly enhanced participants' AV coherence detection ability (Figure 4.2a, pairwise t-test on S1 and S5 coherency values, t_{22} = 2.961, p=0.007).

Correspondingly, the performance of these subjects in the pre and post AV coherency threshold test also showed a drop in their threshold (Figure 4.2b, Pairwise t-test, t_{22} =3.081, p=0.005). I found significant correlation between the changes in coherency values in training and threshold (Figure 4.2c, r = 0.632, p=0.027). In conclusion, training on a task that required that subjects detect AV cross-modal temporal coherence enhanced their ability to detect the temporal coherence between auditory and visual stimuli.

In order to examine the training effect on the AV selective attention task, I have compared resulting discriminability (d') values for participants' pre-test and post-test. Subjects were more sensitive to deviants after training. A two-way repeated ANOVA was conducted for d',

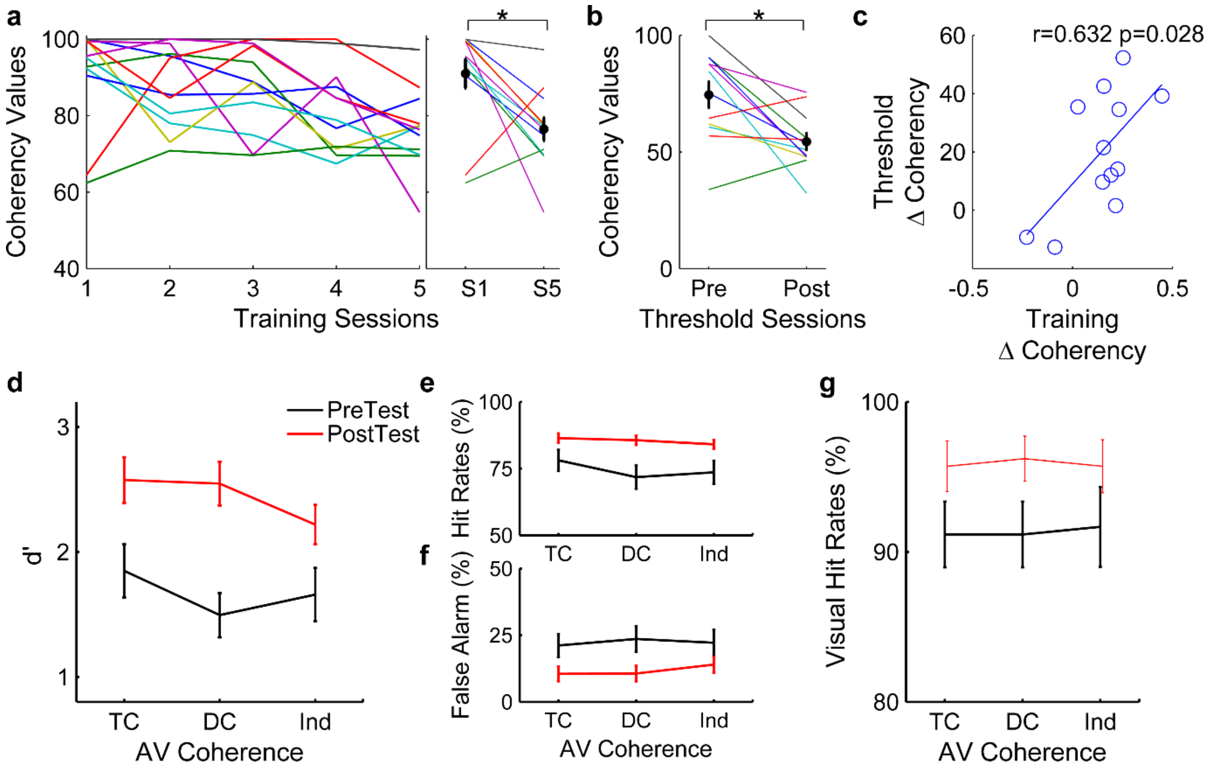

Figure 4. 2 The results of training sessions and testing in AV coherency training.

Panel a shows mean coherency values for 12 participants for five training sessions. The right panel shows the mean coherency values for the first training session (S1) and last session (S5) for all participants in color and the average of all participants (± SEM) in black. b shows all coherency values for pre and post threshold in color with the average of all participants in black (±SEM) c shows the scatter plot for the differences in coherency values in threshold (pre and post) and training (S1 and S5). d, e, f, shows the mean and SEM of d', hit rates and false alarm for three AV coherence conditions in the auditory deviant detection respectively. Panel d shows the mean hit rates for visual stimuli detection in all conditions with SEM.

bias, hit rate, false alarm rate, and visual hit rate with factors of training (pre-test and post-test) and audio-visual coherence condition (target coherent, distractor coherent, independent). For d', there was a significant effect of training (F (1, 71) = 9.39, p = 0.006), AV coherence (F (2, 71) = 9.13, p<0.001) and interaction between training and AV coherence (F (2, 71) = 7.26, p = 0.002). Post-hoc comparison (p<0.05) across AV coherence condition in the pre-test revealed that subjects performed better when the visual stimulus was coherent with the target auditory stream vs the distractor auditory stream (TC > DC; p = 0.0031, Bonferroni-corrected α = 0.017). Similar results were obtained for hit rates (see Table 4.1 for statistical values in detail). This results are consistent with previous findings that a visual stimulus which is temporally coherent with a target steam increases hit rates and decreases false alarms when compared to one that is coherent with the distractor stream (Figure 4.2d-f).

	AV Coherence		*Training*		*Interaction between AV Coherence and training*	
	F	p	F	p	F	p
d’	9.125	<.001	9.393	0.006	7.258	0.002
Hit Rates	6.660	0.002	6.118	0.021	3.210	0.004
False Alarm	2.481	0.095	3.838	0.062	2.755	0.074
Visual hit rates	0.027	0.972	3.300	0.083	0.083	0.920

Table 4. 1 The results of two-way repeated measures ANOVA for each variables (p< 0.05 in bold)

Post-hoc comparison (p<0.05) across AV coherence in post-test revealed that subjects performed better when the visual stimulus was coherent with either the target auditory or the distractor auditory stream than the independent stream (TC >Ind, p = 0.0046; DC> Ind, p = 0.0055, Bonferroni-corrected α = 0.017). Consistent with my second hypothesis, after AV coherency training, a visual stimulus which is temporally coherent with either a target stream or distractor stream increases hit rates and decreases false alarms when compared to an independent visual stimulus. Therefore participants’ performance is enhanced when the visual stimuli was coherent either a target or a distractor stream, suggesting that when a visual stimulus is temporally coherent with one sound in a mixture listeners are better able to separate the scene into two streams (which subsequently improves their performance in the auditory selective attention task).

Participants were equally good at detecting the flashing ring in all three AV coherence conditions before and after the training. (Figure 2.3d, two way repeated ANOVA: AV coherence: F (2, 71) = 0.027, p = 0.972, training effect: F (2, 71) = 3.300, p = 0.083).

AV temporal coherency exposed group

In order to examine whether exposure to temporally coherent AV stimuli will result in similar enhancements in AV coherence ability and the ability to use these skills in auditory scene analysis, twelve participants performed five modulation rate discrimination training sessions. Pairwise comparison between the first and last session revealed that training significantly increased participants' ability to discriminate modulation rate (Figure 4.3a, t_{22} = 4.529, p<0.001). However, while some subjects' AV coherency thresholds dropped between pre and post-test (Fig 4.3b) across the group there was no significant difference in AV coherency threshold values (Figure 4.3b, pairwise t-test, t $_{22}$ = 1.69, p = 0.104). We found statistically no correlation between the changes in modulation rate in training and coherency values in threshold (Figure 4.3c, r = 0.392, p = 0.207), suggesting that modulation rate training showed no effect on participants' ability to detect the temporal coherence between auditory and visual stimuli.

In modulation rate training, although participants were not trained on AV coherency, they were exposed to temporally coherent AV stimuli and to the target vowel sounds with the same pitch and timbre as those used in the auditory selective attention task. In order to examine the exposure effect on the auditory selective attention task and whether this influences the ability to use visual stimuli to promote auditory scene analysis, I have compared discriminability values for three AV coherence conditions in the auditory selective attention task for the pre-test and post-test. Subjects were more sensitive to deviants after training and when the visual stimulus was coherent with the target auditory stream than when it was coherent with the distractor. For d', two way repeated ANOVA with factors of training (pre-test and post-test) and AV coherence condition (target coherent, distractor coherent, independent) revealed a significant effect of training (F(1,71)=5.31, p = 0.009) and visual condition (F(2,71) = 3.69, p = 0.044), but no interaction (F(2,71)=0.17, p=0.844).

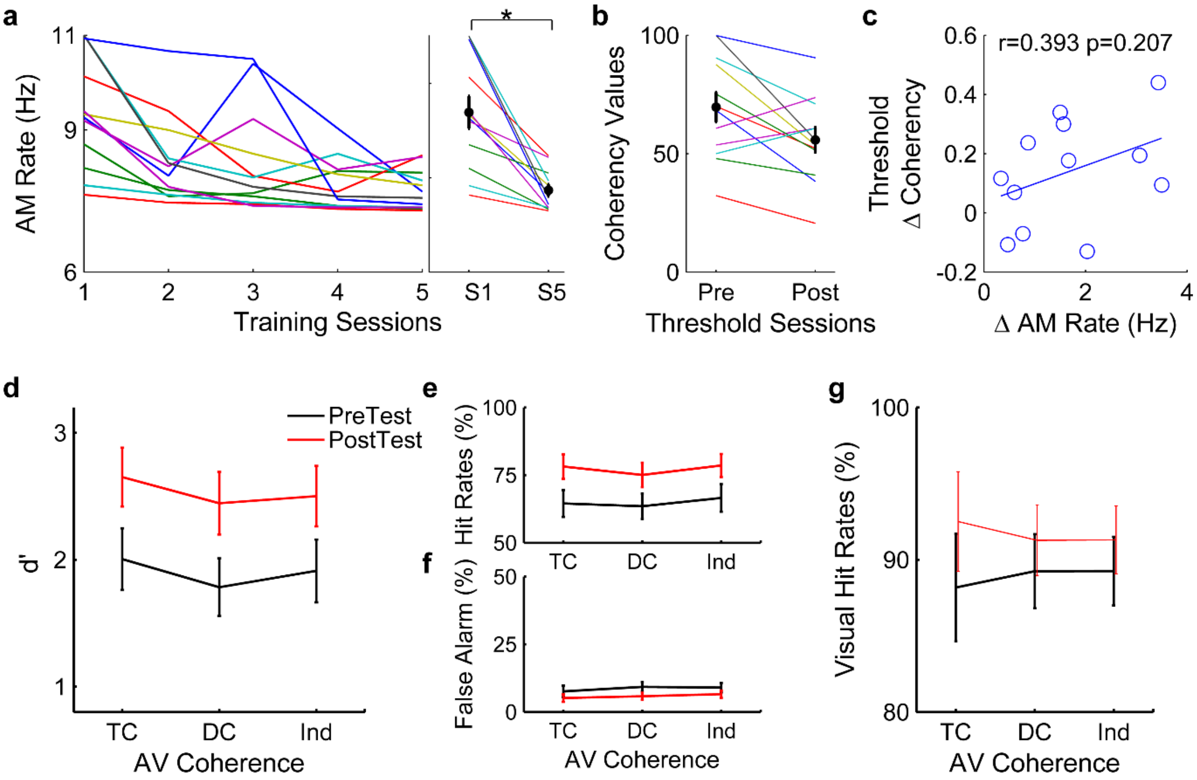

Figure 4.3 The results of training sessions and testing in modulation rate training (AV coherency exposed group)

Panel a shows mean modulation rate thresholds for 12 participants for five training sessions. The right panel shows the mean modulation rate threshold for the first training session (S1) and the last session (S5) for all participants (n=12) in color and the average of all participants (± SEM) in black. b shows all coherency values for pre and post threshold in color with the average of all participants in black (± SEM) c shows the scatter plot for the differences in coherency values in threshold (pre and post) and modulation rate in training (S1 and S5). d, e, f, shows the mean and SEM of d', hit rates and false alarm for three AV coherence conditions in the auditory deviant detection respectively. Panel d shows the mean hit rates for visual stimuli detection in all conditions with SEM.

Post-hoc comparison ($p<0.05$) across AV coherence in pre-test and post-test revealed that subjects performed better when the visual stimulus was coherent with the target auditory stream vs the distractor auditory stream (Figure 4.3d-f, Pre-test: TC > DC, p = 0.0049; Post-test: TC>DC, p= 0.0063, Bonferroni-corrected α = 0.017). Similar results were obtained for hit rates and false alarm (see Table 4.2 for statistical values in detail). Therefore, this suggests an overall improvement in performance after AM rate training in all three AV coherence conditions, but no change in the way in which subjects are able to exploit visual cues.

Consistent with the findings of AV coherency training, participants were equally good at detecting the flashing ring in all three AV coherence conditions and AV coherency training had no effect. (Figure 2.3d; AV coherence: F (2, 71) = 0.002, p = 0.998, training effect: F (2, 71) = 0.640, p = 0.432).

	AV Coherence		*Training*		*Interaction between AV Coherence and training*	
	F	p	F	p	F	p
d’	5.307	0.009	3.688	0.037	0.171	0.844
Hit Rates	3.536	0.037	3.676	0.038	0.386	0.682
False Alarm	2.377	0.105	1.629	0.215	0.366	0.695
Visual hit rates	0.002	0.998	0.640	0.432	0.456	0.637

Table 4. 2 The results of two-way repeated measures ANOVA for each variables (p<0.05 in bold)

Control group

In order to exclude the possibility that changes resulted simply from a practice effect, another twelve participants had no training (Control group) and did only a pre-test and post-test, each including the auditory selective attention task and the AV coherence threshold test. The ability of this group to detect auditory visual coherence did not change between pre-test and post-test (Figure 4.4c, pairwise t-test, t_{22} =0.234, p=0.817). A two-way repeated ANOVA with

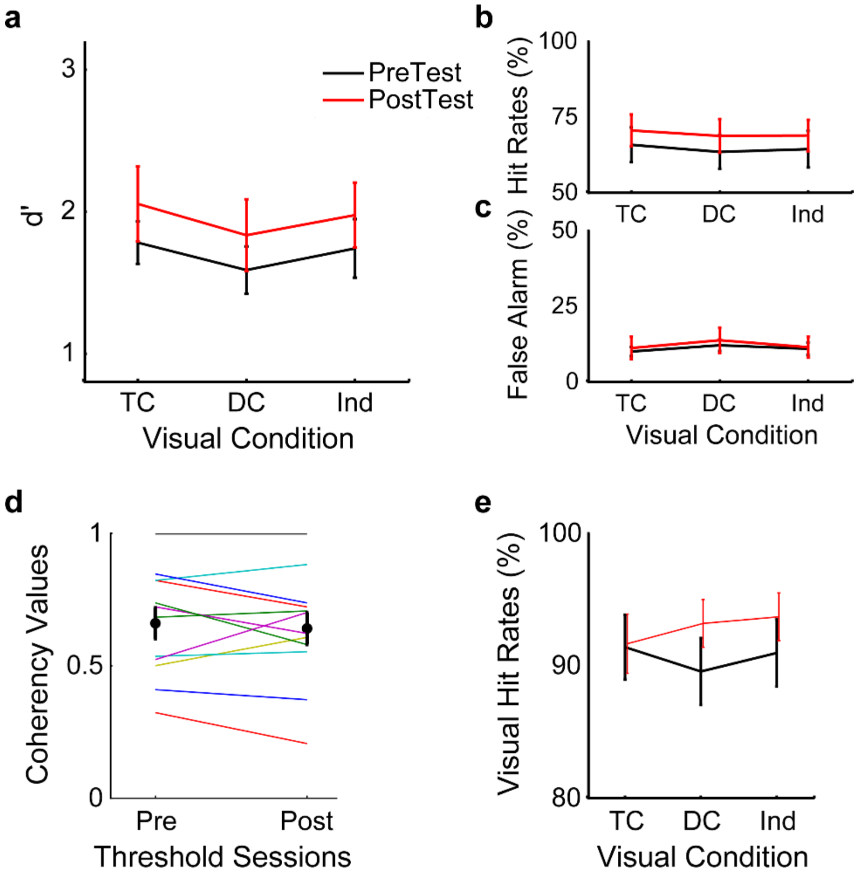

Figure 4. 4 No pre-test-posttest differences in the no-training control group

a, b, c shows the mean and SEM of d', hit rates and false alarm for three AV coherence conditions in the auditory deviant detection respectively. d shows all coherency values for pre and post threshold in color with the average of all participants in black (± SEM). Panel e shows the mean hit rates for visual stimuli detection in all conditions with SEM.

factors of training (pre-test and post-test) and AV coherence (target coherent, distractor coherent, independent) revealed a significant effect of AV coherence condition (F (2,71) = 4.600, p = 0.015), but not training (F (1,71) = 0.730, p = 0.402) or any interaction (F(1,71) = 0.039, p = 0.961). Post-hoc comparison (p<0.05) revealed that participants performed better when the visual stimulus was coherent with the target auditory stream vs the distractor auditory stream in pre-test and post-test (Figure 4.4 a-c, Pre-test: TC > DC, p = 0.0032; Post-test: TC>DC, p= 0.013, Bonferroni-corrected α = 0.017). Therefore, these findings suggests that the training effect we observed in AV coherency training and modulation rate training groups is not due to a simple practice effect.

Training enhances ability to use AV coherence

To directly compare the data between the three experimental groups, I have quantified the mean across AV coherence condition d' values. A 2 x 3 way mixed ANOVA with factors of training (pre-test and post-test) and experimental group (AV coherency, modulation rate and control) revealed significant effect of training (F(1,71)=10.66, p=0.002) but not experimental group (F(2,71) =1.75, p=0.181). There was no interaction between experimental group and training (F (2, 71) = 0.86, p = 0.427). There was an increased in mean d' after AV coherency and AM modulation training. Pairwise comparison between pre-test and post-test for three experimental group, revealed that d' in post-test is significantly larger in the AV coherency group (t_{22}=3.065, p=0.006) as well as the modulation rate group (t_{22}=1.920, p=0.034) but not in the control group (Figure 4.5a; Control: t_{22}=0.854, p=0.402).

Similarly, I have compared AV coherence threshold values for the three experimental group. A 2 by 3 way mixed ANOVA showed no significant effect of experimental group: (F (2, 71) = 0.79, p=0.459; nor of training (F (1, 71) = 2.49, p=0.119; interaction: F (2, 71) = 1.41, p = 0.250). Despite there being no significant effect, there appeared to be a drop in coherency threshold only for the AV coherency group. Pairwise comparison between pre-test and post-test for three experimental groups, revealed that coherency thresholds were significantly decreased only after AV coherency training (Figure 4.5b: t_{22} =3.081, p=0.005) but not AM rate training (t_{22}=1.69, p=0.104) or in the control group (t_{22} =0.234, p=0.817).

Participants' performance was correlated with their change in ability to detect temporal coherence between auditory and visual stimuli. Participants with larger change in their AV threshold across all experiment group showed larger improvement in their performance (r =0.353, p=0.0347). In the AV coherency training group, decrease in coherency values was correlated with the increase in mean d' across coherence condition (r= 0.589, p = 0.04438), suggesting that actively discriminating AV temporal coherence improves the ability to exploit visual stimuli to promote auditory scene analysis.

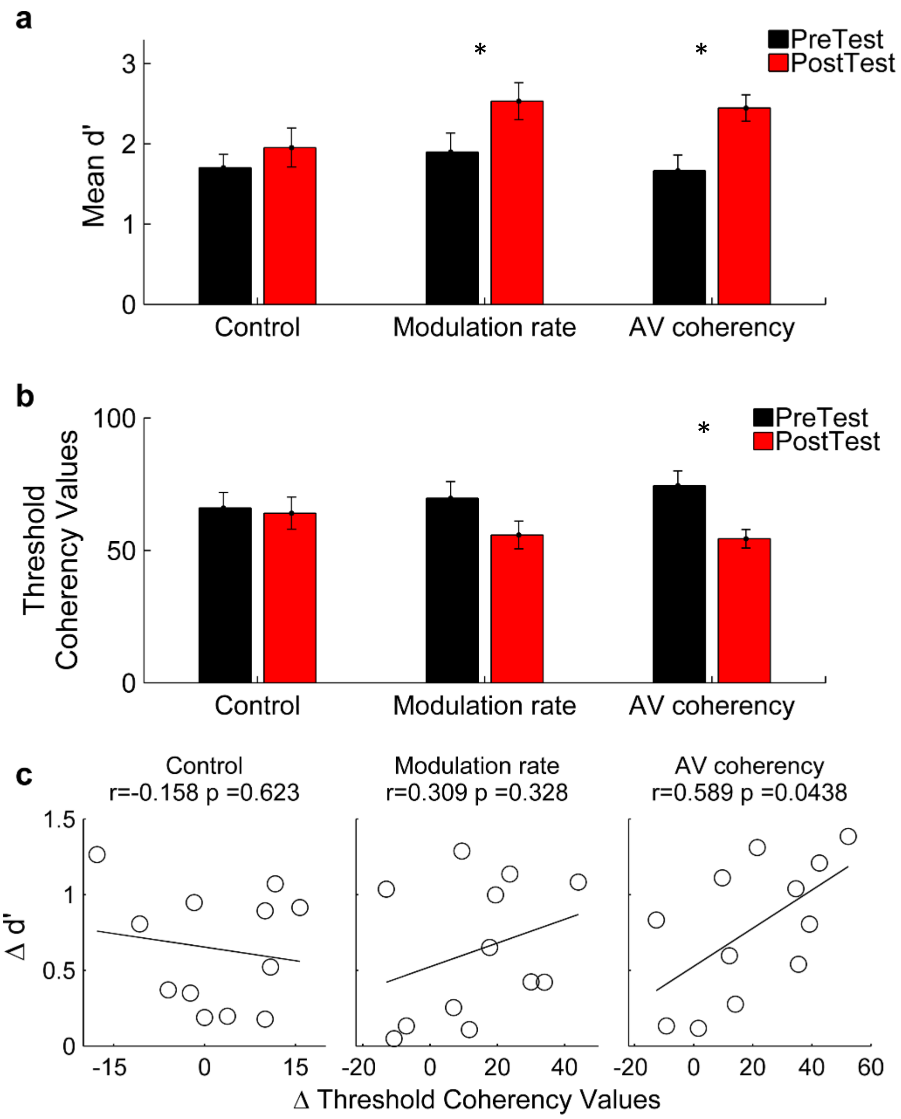

Figure 4. 5 Training enhances ability to use temporal coherence between auditory and visual stimuli.

a, b Bar plots shows the mean d’ in the auditory selective attention task and coherency values in the threshold task for pre and post-test for three experimental groups. Asterisks show significant pairwise comparisons. c shows the correlation between the change in d’ and the change in coherency values across pre-test and post-test.

To isolate the effect of training on the ability to use visual cues, Figure 4.6 show the across-subject means in each condition relative to each subject's overall mean (panel a is equivalent to Figure 4.2d, b to Figure 4.3d and c to Figure 4.4a). This illustrates that participants’ performance across visual conditions were same in control and modulation rate training, in which participants were more sensitive to deviants when the visual stimulus was coherent with the target stream. Modulation rate training did not appear to cause a difference in the ability to exploit visual cues as the difference between TC and DC was no different in pre-test and post-test data. In contrast, after training the AV coherence trained group showed a very different pattern of responses with both target and distractor coherent conditions having superior performance to the independent condition. This implies that this group were better able to exploit visual cues to segregate the two competing auditory streams.

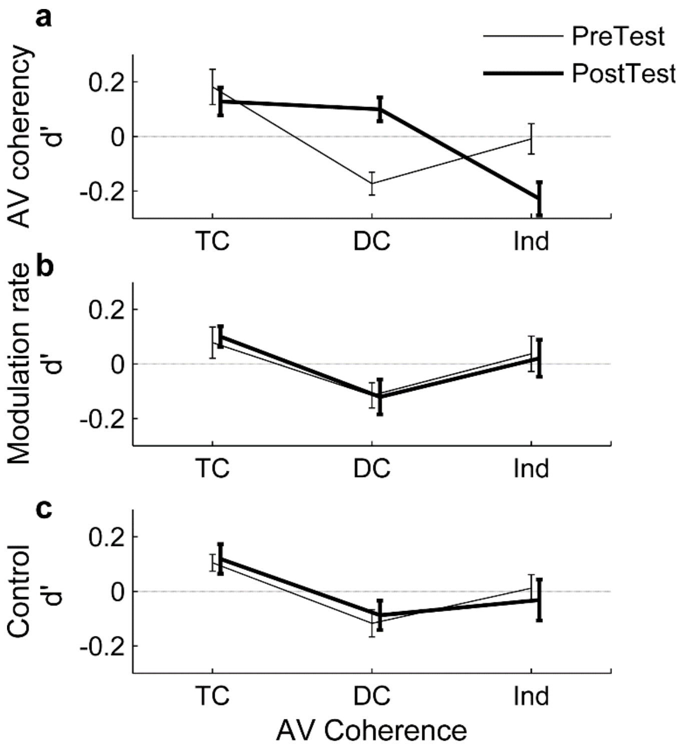

Figure 4.6 The across-subject means in each condition relative to each subject's overall mean for three experimental groups.

Discussion

In this study, subjects were either actively trained to detect AV temporal coherence or exposed to stimuli in which auditory and visual elements were always temporally coherent but subjects were required to discriminate modulation rates. I demonstrated that AV coherence thresholds of those in the AV coherence training group were reduced after five relatively short training sessions and that actively discriminating AV temporal coherency enhanced the ability of listeners to detect the temporal coherence between auditory and visual stimuli. Furthermore, in the auditory selective attention task, participants' performance on reporting deviants were better when the visual steam was coherent with either target or non-target auditory stream than when the visual stimuli was independent suggesting that training changed the way in which visual cues were exploited when performing the ASA task. Hence, listeners were better able to exploit visual cues to segregate the two competing auditory streams.

However, AV coherence thresholds were not reduced by being exposed to AV temporal coherency. Overall performance improved equally across the three visual conditions. Participants in modulation rate training could have solved the modulation rate discrimination task by auditory alone, visual alone or AV information. There was no requirement for listeners to integrate auditory and visual stimuli. This group were always exposed to the modulated vowel sounds while performing the modulation rate task and therefore the improvement in performance is likely due to experience enabling improved timbre detection rather than to any change in the way in which listeners use visual information.

These findings support the idea that participants' auditory performance with visual cues is influenced by their ability to detect AV temporal coherency. As I have speculated, the great individual differences in subjects' performance in the auditory selective attention task we observed in Chapter II might due to their ability to detect AV temporal coherency.

Consistent with my previous data, with no training, when the visual stimulus was coherently modulated with a target auditory stream, participants were better able to report brief timbre deviants in the target stream, than when the visual stream was coherent with the non-target auditory stream (Maddox et al., 2015). A short training on AV temporal coherence detection altered the pattern of responses to AV coherence. Participant's performance on reporting deviants were better when the visual steam was coherent with either target or non-target auditory stream than when the visual stimuli was independent suggesting that training enhanced the ability to exploit visual cues even without selectively attending to auditory stream.

The pattern of change observed after AV coherency training may be due to the awareness of the task requirements. When participants were asked to discriminate temporally coherent AV stimuli in AV temporal coherent discrimination task, they are being conscious about the temporal relation between auditory amplitude modulation and visual size. Such awareness could have an effect on participants' performance. However, the same task paradigm was used for the threshold test in modulation rate training group as well as control group, and the pattern change only occurred in AV coherency training. Hence, I believe that the pattern change is due to the enhancement in ability to detect temporal coherence between auditory and visual cues and not simply due to an increased awareness of task design.

Training enhances coherence driven AV binding

In cross-modal objects, inputs from different modalities are weighted according to their reliability (Ernst and Bülthoff, 2004) and cross-modal correspondence is strongest when evidence supports information from both modalities as originating from the same object. (Bizley et al., 2012). The training in this study entailed actively discriminating AV temporal coherency which enhanced the ability to detect temporal coherence between sensory modalities, suggesting that an enhanced ability to detect temporal coherence might increase the reliability of the information that both modalities originating from the same object and let these information from both modalities be perceived as one object.

There were two speculations about how temporal coherence driven binding creates a cross-modal object. It is either (i) enhancing listener's ability to detect brief timbre deviants when the visual stimulus is coherent with the target stream and disrupting their ability when visual stimuli is coherent the distractor stream or (ii) enhancing listener's ability to detect brief timbre deviants when the visual stimulus is coherent with either the target stream or the distractor stream. Here, I show that that temporal coherence between the visual stimulus and either auditory stream leads to better performance relative to a condition in which the visual stimulus is independent with both. This suggests that in either case you are able to parse the auditory scene more effectively and therefore improve in performance of the auditory selective attention task.

Evidence for bottom–up processing in cross-model object formation leads to the prediction that there should be a bi-directional effect such that participants show enhanced visual detection. Unfortunately, visual hit rates are at ceiling so it is hard to detect the difference for AV coherence condition. Since my auditory selective attention task is designed to test visual impact on auditory scene analysis as discussed in discussion section in chapter II. Further experiments designed to test for a bidirectional effect are required.

Changes in multisensory processes do not originate solely from multisensory training. Recent evidence indicates that unisensory visual training can also have an impact on multisensory processing (Stevenson et al., 2013) in that it narrows the AV temporal binding window (Powers Iii et al., 2016).

Biological implementation

Three models have been proposed as the explanation for plasticity induced by multisensory learning (Shams and Seitz, 2008, Driver and Noesselt, 2008). The first model suggests that multisensory learning results in changes in unisensory cortical structures involved in the multisensory task (Stevenson et al., 2013, Powers Iii et al., 2016). The second model suggests an alteration in the interconnection of the unisensory structures (Giraud and Poeppel, 2012) and the final view that training alters the cortical fields known to integrate multisensory information responsible for integrating the stimuli (Campanella and Belin, 2007). A training study using MEG (Paraskevopoulos et al., 2012) argued that plasticity due to short-term multisensory training alters the function of separate multisensory structures, and not merely the unisensory ones, along with their interconnection. Musically un-trained participants were trained to play tone sequences from visually presented patterns in a music notation-like system (audio, visual and somatosensory training), while another group attentively listened to the recordings of the first group's sessions (AV training). The cortical MMN responses of AV, an auditory and a visual were observed before and after the training with an enhancement of the AV MMN, while there was no significant effect on the auditory and visual MMN. They reported a region in the right superior temporal gyrus which was affected by input from all three modalities during the training procedure in such a way that the neural plasticity effect of short-term multisensory training modified its function. However, the underlying mechanism of multisensory training enhancement in multisensory processing is still not clear.

One of the consequences of the cross-modal plasticity arising from hearing loss is that AV integration by human cochlear implant (CI) users is often abnormal. Thus, CI users especially CI users with less proficient at speech recognition tend to rely more heavily on visual cues when presented with incongruent AV speech (Schorr et al., 2005, Tremblay et al., 2010). Our neurological evidence that cortical processing of auditory scene analysis is enhanced by coherent visual cues (Chapter III) and behavioural evidence that temporal coherence between auditory and visual stimuli can enhance listeners' ability to segment an auditory scene (Chapter III) , suggests a facilitative role of vision in helping auditory scene analysis in human CI users. These findings support investigation of a similar training paradigm in human CI users.

To sum up, this study allowed me to examine whether the ability to discriminate temporal coherence between auditory and visual stimuli improves with training. I showed that an improved ability to detect temporal coherence between auditory and visual stimuli result in a greater ability to exploit temporal coherence such that both coherent visual stimuli regardless of target or distractor auditory stream improved listeners performance by helping to segregate the streams more effectively relative to the independent condition.

Chapter V: Discussion and Conclusions

Discussion

This research was comprised of a series of human psychophysical and animal neurophysiological studies investigating the impact of visual information on the bottom-up auditory scene analysis. The project had a focus on the role of the temporal coherent visual cues in auditory scene analysis, or promoting the ability to segregate sounds from a mixture and demonstrated that temporal coherence between auditory and visual stimuli can enhance listeners' ability to segment an auditory scene, and influence the representation of sounds in auditory cortex.

The reasoning behind the AV stimuli

Previous studies exploring the role of vision in auditory scene analysis either used very simple sounds or sounds in which linguistic factors come in to play. Visual speech is a complex stimulus including semantically related associations, and unlike other low level auditory features, these associations rely on pre-existing knowledge and it is too complex to know whether they are specific to speech or whether some of them are more general phenomena (such as temporal coherence) and how listeners make use of this knowledge for their perceptual processing.

In order to conduct human and animal studies with similar stimuli, I have used ongoing artificial vowel sounds with amplitude modulation with a noisy envelope low-pass filtered at 7 Hz. These parameters were chosen within the ethologically relevant modulation frequency range of human hearing (Chandrasekaran et al., 2009). Denison et al. showed that coherence discrimination was better for the unpredictable sequences than for predictable ones (2013), unlike other studies train a repeating sequence of sound, the noisy dynamics of stimulus features was added to enhance the bound percept and to make things more naturalistic.

I chose to use artificial vowels because not only do they carry complex acoustic information like natural sounds for human psychophysics but they allow full control to parametrically vary perceptual features. Moreover, ferrets, the animal model of this study, are easily trained in vowel discrimination tasks (based on performance (Town et al., 2015, Walker et al., 2011b). In the second chapter, I have investigated psychophysically whether temporal coherence between auditory and visual stimuli was sufficient to promote auditory stream segregation. I have used an auditory selective attention task in which participants were instructed to report deviants in the target auditory stream while watching the visual stimulus and ignoring the distractor stream. The visual stimulus was designed such that its temporal envelope was either coherent with one of the sound streams, or independent of both. This allowed an investigation of the role that temporal coherence might play in AV integration and whether visual information can promote stream segregation.

As I hypothesised, modulating a visual stimulus coherently with one auditory stream in a mixture caused the temporally coherent AV stimuli to bind together resulting better performance in an auditory selective attention task. Notably, the temporal coherence between auditory feature (amplitude modulation of the sound) and visual feature (size of the grey disc) is not-task relevant as participants are asked to detect the timbre of the auditory stimuli and color of the visual stimuli. Any binding between auditory and visual stimuli and/or any enhancement of an auditory feature would be about an uninformative visual stimulus.

I have shown that listeners were better able to report brief deviants in the target stream when a visual stream was temporally coherent with the target stream and performance was impaired when the visual stream was temporally coherent with the non-target auditory stream. These findings, with well control non-linguistic stimuli provide novel evidence that temporally coherent visual cues promote auditory stream segregation. We have speculated that improvement in performance is due to coherence driven binding of auditory and visual stimuli which are subsequently perceived as one cross-modal object.

Cross-modal object formation

In our conceptual model of processing underlying these results, we suggested that when there is cross-modal temporal coherence between a feature of an auditory and visual stream, those features are bound and this results in a cross-modal object (just as two auditory streams with the same envelope very likely would have bound together as well). The definition of 'an object' is a controversial issue, but one of the criteria defining an object is that the stimuli with all features if perceived as one object, would be influenced as a whole (Bizley et al., 2016, Blaser et al., 2000). My result illustrated that temporal coherence between amplitude modulation and visual size enhanced the timbre of the auditory stimuli, although this feature of auditory stimuli had no relation to the visual stimuli. This is consistent with the idea that a coherence driven binding of auditory and visual features forms a cross-modal object.

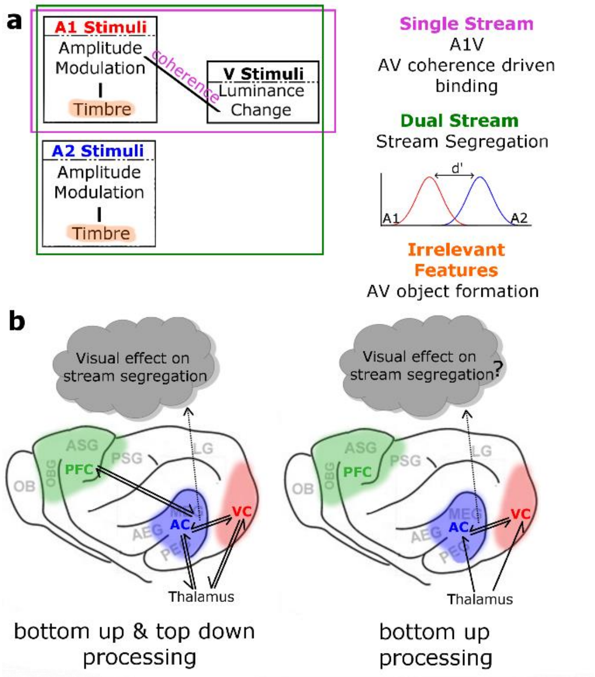

Figure 5.1 The conceptual model of visual cues impact on bottom-up level auditory scene analysis by AV temporal coherence driven cross-modal object formation.

Coherence driven binding in auditory cortex

In order to understand the biological implementation of the impact of visual input on auditory scene analysis, I have conducted neurophysiological studies on ferret auditory cortex (Chapter III). Ferrets have excellent low frequency hearing (Sumner and Palmer 2012) and are able to identify the direction of pitch changes (Walker et al., 2009) and detect changes in pitch or spectral timbre of artificial vowels (Walker et al., 2011a). They can generalize over a range of pitches and maintain performance in the presence of background noise (Bizley et al., 2013a, Town et al., 2015).

This observation suggests that ferrets are a suitable species in which to study neural process underlying the translation of sound acoustics into perceptual features (Bizley et al., 2013b). Furthermore, the neural responses of ferret neurons in AC are sufficiently rich to encode and discriminate phoneme classes, suggesting that ferrets may have a similar mechanism to humans for general acoustic representation to learn boundaries for categorical sound identification (Mesgarani et al., 2008) Furthermore, neurons throughout ferret AC are sensitive to sound timbre and when artificial vowel stimuli were varied simultaneously in different acoustic features (e.g. pitch, location) as well as timbre, neural responses were sensitive to multiple sound features (Bizley et al., 2009).

A role for AC in the bottom-up processing of auditory stream segregation was demonstrated either by analysing acoustic features and/or grouping of acoustic features (Micheyl et al., 2005, Fishman et al., 2014, Bidet-Caulet et al., 2007). Stream segregation appears to be modulated by attention to specific features of auditory stimuli (Atiani et al., 2014, Atiani et al., 2009, Fritz et al., 2005, Fritz et al., 2003) and modulated by attentional focus (Ding and Simon, 2012, Mesgarani and Chang, 2012, Golumbic et al., 2013). However, how multisensory input affects the representation of auditory streams in AC has not been demonstrated yet.

Since visual stimuli can both drive and modulate neural activity in primary and non-primary auditory cortex (Bizley et al., 2007; Chandrasekaran et al., 2013; Ghazanfar et al., 2005; Kayser et al., 2008; Perrodin et al., 2015), I have speculated that visual inputs would boost auditory scene analysis by enhancing the representation of the temporally coherent stream and/or facilitating the formation of cross-modal objects.

To do so, I have recorded a modified version of the stimuli used in human psychophysics from passively awake ferrets and anaesthetised ferrets. Recordings in anesthetized animals allowed me to rule out any top-down effects of attention while recordings in awake passively listening animals allowed me to measure neural activity free from confounds of pharmacological manipulation.

I have demonstrated that temporally coherent visual stimuli elicit reliable changes in the phase of the local field potential revealing that visual information reliably modulated the phase of oscillatory activity of neurons in AC. Furthermore, the enhanced spike-based representation of auditory stream that is temporally coherent with the visual stimulus supports my speculation that visual inputs boost auditory scene analysis by both enhancing the representation of the temporally coherent stream and facilitating the formation of cross-modal objects. These findings provide strong evidence that one role for the early integration of temporally coherent visual information into AC is to resolve competition between multiple sound sources within an auditory scene.

Stronger coherence driven binding by perceptual learning

Finally, I have investigated the effects of short-term training effect on the ability of listeners to exploit visual information for auditory scene analysis. Participants were either actively engaged with AV temporal coherency by training on an AV coherency discrimination task in which they had reported which visual stimulus was temporally coherent to auditory stimuli, or exposed to AV temporal coherency by training on modulation rate task in which they reported which temporally coherent AV stimulus had a larger modulation. I demonstrated that AV coherence thresholds of those participants in the AV coherence training group were decreased by a short training period and these participants' performance in the auditory selective attention task was enhanced, but the pattern changed such that both conditions had improved performance relative to the independent condition.

Hence, listeners were better able to exploit visual cues to segregate the two competing auditory streams after actively engaged with AV temporal coherency. These findings suggest that listeners can learn to exploit visual cues to auditory scene analysis by training on a simple sensory discrimination (temporal coherence discrimination task).

Perceptual learning (Samuel and Kraljic, 2009) in this context can be defined as a process by which listeners alter their ability to exploit visual cues to promote stream segregation based on the context those sounds occur. In the training sessions, participants were asked to detect the temporal coherence between auditory and visual stimuli. This results in stronger coherence binding of auditory and visual stimuli leading to better stream segregation in the auditory selective attention task.

Findings presented in this thesis are critical as they demonstrate a facilitative role of vision in helping auditory scene analysis. Normal hearing listeners in silent listening conditions might not need to augment listening with visual information. However, once listening becomes more challenging, an optimal strategy might be to integrate reliable visual cues in order to enhance listening. The ability to accurately and appropriately group sound elements is an essential part of comprehending speech in noise. Here, I have provided evidence that temporally coherent visual information helps us in stream segregation and this benefit can be enhanced by short training. Moreover, cochlear implant (CI) users especially CI users with less proficient in their implants tend to rely more heavily on visual cues when presented with incongruent AV speech (Schorr et al., 2005, Tremblay et al., 2010). The enhancement in performance seen with temporally coherent visual stimuli is relatively modest and rather variable between subjects. Nevertheless, since a 1 dB increase in SNR results in a 10-20% increases in intelligibility (Brand and Kollmeier, 2002), any effect I presented here may provide significant performance benefits in adverse listening conditions especially listeners with hearing impairment and CI users.

Conclusions

This research by combining human psychophysical and ferret neurophysiological studies provided compelling evidence on the impact of visual information on the bottom-up auditory scene analysis by coherence driven binding. Neural correlates of visually-enhanced auditory stream segregation are observed in the absence of a behavioural task (and even under anesthesia), suggesting that attention is not required for these effects. Finally, the demonstration that visual stimuli can modify the representation of an auditory scene in auditory cortex provides evidence that one role for visual innervation of auditory cortex is in performing auditory scene analysis.

Future Directions

Although I have provided evidence on how AV integration has an impact on bottom-up processing auditory stream segregation and such that impact can be modulated by attention and prior knowledge, the truth is, however, more complex. The AV paradigm used in this research is a strong tool to investigate cross-model object formation. Well-controlled studies across different perceptual attributes are needed to examine the cross-modal effect son auditory scene analysis. It is also important to replicate these findings with a different paradigm to show the improvement by visual information in auditory scene analysis generalises to different situations.

In chapter II, I have demonstrated a bottom-up mechanism that promotes the formation of cross-modal objects, but the underlying mechanism is still not clear. It might be an attention-related signal trigged by temporally coherent AV stimuli (Shamma et al., 2011) or enhanced functional connectivity by phase resetting as a top-down predictive signal for upcoming visual information (Samaha et al., 2015). In order to disambiguate the neural mechanism, during the anaesthetised recordings, simultaneous multi-electrode array recordings were made in the auditory and visual cortex. The data from visual cortex was collected, but time constraints mean that it has not yet been analysed. I would speculate that if temporal coherence binding leads to the formation of cross-modal objects, there will be enhanced functional connectivity between auditory cortex and visual cortex for the auditory stream which is coherent to visual stimuli. Neurophysiological studies across different cortical regions including parietal cortex and prefrontal cortex are needed to understand the underlying mechanism of AV integration in auditory scene analysis and whether these bottom-up effects interact with top-down factors, especially in the context of a behavioural task.

Our neurological evidence on that cortical processing of auditory scene analysis are enhanced by coherent visual cues (Chapter III) and behavioural evidence that temporal coherence between auditory and visual stimuli can enhance listeners' ability to segment an auditory scene (Chapter III and IV), suggests a facilitative role of vision in helping auditory scene analysis in human CI users. These findings support investigation of a similar training paradigm in human CI users.

Reference

ADELI, M., ROUAT, J. & MOLOTCHNIKOFF, S. 2014. Audiovisual correspondence between musical timbre and visual shapes. *Frontiers in Human Neuroscience,* 8.

AHISSAR, E., SOSNIK, R., BAGDASARIAN, K. & HAIDARLIU, S. 2001. Temporal frequency of whisker movement. II. Laminar organization of cortical representations. *Journal of Neurophysiology,* 86**,** 354-367.

ALAIN, C. 2007. Breaking the wave: Effects of attention and learning on concurrent sound perception. *Hearing Research,* 229**,** 225-236.

ALAIN, C., ARNOTT, S. R. & PICTON, T. W. 2001. Bottom–up and top–down influences on auditory scene analysis: Evidence from event-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance,* 27**,** 1072.

ALAIN, C., REINKE, K., HE, Y., WANG, C. & LOBAUGH, N. 2005. Hearing two things at once: neurophysiological indices of speech segregation and identification. *Journal of Cognitive Neuroscience,* 17**,** 811-818.

ALAIS, D. & BURR, D. 2004. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology,* 14**,** 257-262.

ALSIUS, A., NAVARRA, J. & SOTO-FARACO, S. 2007. Attention to touch weakens audiovisual speech integration. *Experimental Brain Research,* 183**,** 399-404.

ATIANI, S., DAVID, S. V., ELGUEDA, D., LOCASTRO, M., RADTKE-SCHULLER, S., SHAMMA, S. A. & FRITZ, J. B. 2014. Emergent Selectivity for Task-Relevant Stimuli in Higher-Order Auditory Cortex. *Neuron,* 82**,** 486-499.

ATIANI, S., ELHILALI, M., DAVID, S. V., FRITZ, J. B. & SHAMMA, S. A. 2009. Task Difficulty and Performance Induce Diverse Adaptive Patterns in Gain and Shape of Primary Auditory Cortical Receptive Fields. *Neuron,* 61**,** 467-480.

BARRACLOUGH, N. E., XIAO, D., BAKER, C. I., ORAM, M. W. & PERRETT, D. I. 2005. Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience,* 17**,** 377-391.

BEN-ARTZI, E. & MARKS, L. E. 1995. Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics,* 57**,** 1151-1162.

BERNSTEIN, I. H. & EDELSTEIN, B. A. 1971. Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology,* 87**,** 241.

BERNSTEIN, J. G. & GRANT, K. W. 2009. Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America,* 125**,** 3358-3372.

BERNSTEIN, L. E., AUER, E. T. & TAKAYANAGI, S. 2004. Auditory speech detection in noise enhanced by lipreading. *Speech Communication,* 44**,** 5-18.

BERTELSON, P., VROOMEN, J., DE GELDER, B. & DRIVER, J. 2000. The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics,* 62**,** 321-332.

BEST, V., OZMERAL, E. J., KOPCO, N. & SHINN-CUNNINGHAM, B. G. 2008. Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences of the United States of America,* 105**,** 13174-13178.

BIAU, E. & SOTO-FARACO, S. 2013. Beat gestures modulate auditory integration in speech perception. *Brain and Language,* 124**,** 143-152.

BIDET-CAULET, A., FISCHER, C., BESLE, J., AGUERA, P. E., GIARD, M. H. & BERTRAND, O. 2007. Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. *Journal of Neuroscience,* 27**,** 9252-9261.

BIZLEY, J. K., MADDOX, R. K. & LEE, A. K. 2016. Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. *Trends in Neurosciences*, 39(2), 74-85.

BIZLEY, J. K., NODAL, F. R., BAJO, V. M., NELKEN, I. & KING, A. J. 2007. Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cerebral Cortex,* 17**,** 2172-2189.

BIZLEY, J. K., SHINN-CUNNINGHAM, B. G. & LEE, A. K. C. 2012. Nothing Is Irrelevant in a Noisy World: Sensory Illusions Reveal Obligatory within-and across-Modality Integration. *Journal of Neuroscience,* 32**,** 13402-13410.

BIZLEY, J. K., WALKER, K. M., KING, A. J. & SCHNUPP, J. W. 2013a. Spectral timbre perception in ferrets: discrimination of artificial vowels under different listening conditions. *The Journal of the Acoustical Society of America,* 133**,** 365-376.

BIZLEY, J. K., WALKER, K. M. M., NODAL, F. R., KING, A. J. & SCHNUPP, J. W. H. 2013b. Auditory Cortex Represents Both Pitch Judgments and the Corresponding Acoustic Cues. *Current Biology,* 23**,** 620-625.

BIZLEY, J. K., WALKER, K. M. M., SILVERMAN, B. W., KING, A. J. & SCHNUPP, J. W. H. 2009. Interdependent Encoding of Pitch, Timbre, and Spatial Location in Auditory Cortex. *Journal of Neuroscience,* 29**,** 2064-2075.

BLASER, E., PYLYSHYN, Z. W. & HOLCOMBE, A. O. 2000. Tracking an object through feature space. *Nature,* 408**,** 196-199.

BORSKY, S., TULLER, B. & SHAPIRO, L. P. 1998. “How to milk a coat:” The effects of semantic and acoustic information on phoneme categorization. *The Journal of the Acoustical Society of America,* 103**,** 2670-2676.

BRAND, T. & KOLLMEIER, B. 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *The Journal of the Acoustical Society of America,* 111**,** 2801-2810.

BREGMAN, A. 1990. Auditory Scene Analysis: The perceptual organization of sound. 1990. MIT Press, Cambridge, MA.

BREGMAN, A. S., AHAD, P. A. & KIM, J. 1994. Resetting the pitch-analysis system. 2. Role of sudden onsets and offsets in the perception of individual components in a cluster of overlapping tones. *The Journal of the Acoustical Society of America,* 96**,** 2694-2703.

BREGMAN, A. S. & CAMPBELL, J. 1971. Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology,* 89**,** 244.

BREGMAN, A. S. & DANNENBRING, G. L. 1973. The effect of continuity on auditory stream segregation. *Perception & Psychophysics,* 13**,** 308-312.

BREMNER, A. J., CAPAROS, S., DAVIDOFF, J., DE FOCKERT, J., LINNELL, K. J. & SPENCE, C. 2013. “Bouba” and “Kiki” in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners. *Cognition,* 126**,** 165-172.

BROSCH, M., SELEZNEVA, E. & SCHEICH, H. 2005. Nonauditory events of a behavioral procedure activate auditory cortex of highly trained monkeys. *The Journal of Neuroscience,* 25**,** 6797-6806.

BUDINGER, E., HEIL, P., HESS, A. & SCHEICH, H. 2006. Multisensory processing via early cortical stages: connections of the primary auditory cortical field with other sensory systems. *Neuroscience,* 143**,** 1065-1083.

BUDINGER, E., HEIL, P. & SCHEICH, H. 2000. Functional organization of auditory cortex in the Mongolian gerbil (Meriones unguiculatus). III. Anatomical subdivisions and corticocortical connections. *European Journal of Neuroscience,* 12**,** 2425-2451.

BUDINGER, E., LASZCZ, A., LISON, H., SCHEICH, H. & OHL, F. W. 2008. Non-sensory cortical and subcortical connections of the primary auditory cortex in Mongolian gerbils: bottom-up and top-down processing of neuronal information via field AI. *Brain Research,* 1220**,** 2-32.

CALVERT, G. A., BULLMORE, E. T., BRAMMER, M. J., CAMPBELL, R., WILLIAMS, S. C., MCGUIRE, P. K., WOODRUFF, P. W., IVERSEN, S. D. & DAVID, A. S. 1997. Activation of auditory cortex during silent lipreading. *Science,* 276**,** 593-596.

CAMPANELLA, S. & BELIN, P. 2007. Integrating face and voice in person perception. *Trends in Cognitive Sciences,* 11**,** 535-543.

CARLYON, R. P., CUSACK, R., FOXTON, J. M. & ROBERTSON, I. H. 2001. Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance,* 27**,** 115.

CARLYON, R. P., DEMANY, L. & SEMAL, C. 1992. Detection of across-frequency differences in fundamental frequency. *The Journal of the Acoustical Society of America,* 91**,** 279-292.

CHANDRASEKARAN, C., LEMUS, L. & GHAZANFAR, A. A. 2013. Dynamic faces speed up the onset of auditory cortical spiking responses during vocal detection. *Proceedings of the National Academy of Sciences,* 110**,** E4668-E4677.

CHANDRASEKARAN, C., TRUBANOVA, A., STILLITTANO, S., CAPLIER, A. & GHAZANFAR, A. A. 2009. The natural statistics of audiovisual speech. *PLoS Computational Bioogyl,* 5**,** e1000436.

CHIOU, R. & RICH, A. N. 2012. Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception,* 41**,** 339-353.

COOPER, W. E., PACCIA, J. M. & LAPOINTE, S. G. 1978. Hierarchical coding in speech timing. *Cognitive Psychology,* 10**,** 154-177.

CROSSE, M. J., BUTLER, J. S. & LALOR, E. C. 2015. Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *The Journal of Neuroscience,* 35**,** 14195-14204.

DENISON, R. N., DRIVER, J. & RUFF, C. C. 2013. Temporal structure and complexity affect audio-visual correspondence detection. *Frontiers in Psychology,* 3.

DESIMONE, R. & DUNCAN, J. 1995. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience,* 18**,** 193-222.

DEVERGIE, A., GRIMAULT, N., GAUDRAIN, E., HEALY, E. W. & BERTHOMMIER, F. 2011. The effect of lip-reading on primary stream segregation. *Journal of the Acoustical Society of America,* 130**,** 283-291.

DING, N. & SIMON, J. Z. 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences,* 109**,** 11854-11859.

DODD, B. 1980. Interaction of auditory and visual information in speech perception. *British Journal of Psychology,* 71**,** 541-549.

DRIVER, J. 1996. Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, 381 (6577), 66.

DRIVER, J. & NOESSELT, T. 2008. Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron,* 57**,** 11-23.

EITAN, Z. & TIMMERS, R. 2010. Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition,* 114**,** 405-422.

ELHILALI, M., MA, L., MICHEYL, C., OXENHAM, A. J. & SHAMMA, S. A. 2009a. Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron,* 61**,** 317-329.

ELHILALI, M., XIANG, J. J., SHAMMA, S. A. & SIMON, J. Z. 2009b. Interaction between Attention and Bottom-Up Saliency Mediates the Representation of Foreground and Background in an Auditory Scene. *Plos Biology,* 7.

ERNST, M. O. & BÜLTHOFF, H. H. 2004. Merging the senses into a robust percept. *Trends in Cognitive Sciences,* 8**,** 162-169.

EVANS, K. K. & TREISMAN, A. 2010. Natural cross-modal mappings between visual and auditory features. *Journal of Vision,* 10**,** 6.

FALCHIER, A., CLAVAGNIER, S., BARONE, P. & KENNEDY, H. 2002. Anatomical evidence of multimodal integration in primate striate cortex. *The Journal of Neuroscience,* 22**,** 5749-5759.

FERNAY, L., REBY, D. & WARD, J. 2012. Visualized voices: a case study of audio-visual synesthesia. *Neurocase,* 18**,** 50-56.

FERNÁNDEZ-PRIETO, I., VERA-CONSTÁN, F., GARCÍA-MORERA, J. & NAVARRA, J. 2012. Spatial recoding of sound: Pitch-varying auditory cues modulate up/down visual spatial attention. *Seeing and Perceiving,* 25**,** 150-151.

FISHMAN, Y. I. & STEINSCHNEIDER, M. 2010. Neural correlates of auditory scene analysis based on inharmonicity in monkey primary auditory cortex. *The Journal of Neuroscience,* 30**,** 12480-12494.

FISHMAN, Y. I., STEINSCHNEIDER, M. & MICHEYL, C. 2014. Neural representation of concurrent harmonic sounds in monkey primary auditory cortex: implications for models of auditory scene analysis. *The Journal of Neuroscience,* 34**,** 12425-12443.

FOXE, J. J., WYLIE, G. R., MARTINEZ, A., SCHROEDER, C. E., JAVITT, D. C., GUILFOYLE, D., RITTER, W. & MURRAY, M. M. 2002. Auditory-somatosensory multisensory processing in auditory association cortex: an fMRI study. *Journal of Neurophysiology,* 88**,** 540-543.

FRITZ, J., ELHILALI, M. & SHAMMA, S. 2005. Active listening: Task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. *Hearing Research,* 206**,** 159-176.

FRITZ, J., SHAMMA, S., ELHILALI, M. & KLEIN, D. 2003. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience,* 6**,** 1216-1223.

GALLACE, A. & SPENCE, C. 2006. Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics,* 68**,** 1191-1203.

GAUDRAIN, E., GRIMAULT, N., HEALY, E. W. & BÉRA, J.-C. 2007. Effect of spectral smearing on the perceptual segregation of vowel sequences. *Hearing Research,* 231**,** 32-41.

GHAZANFAR, A. A., MAIER, J. X., HOFFMAN, K. L. & LOGOTHETIS, N. K. 2005. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience,* 25**,** 5004-5012.

GIANNAKIS, K. 2006. A comparative evaluation of auditory-visual mappings for sound visualisation. *Organised Sound,* 11**,** 297-307.

GIANNAKIS, K. & SMITH, M. 2001. Imaging soundscapes: Identifying cognitive associations between auditory and visual dimensions. *Musical Imagery***,** 161-179.

GIARD, M. H. & PERONNET, F. 1999. Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience,* 11**,** 473-490.

GIRAUD, A.-L. & POEPPEL, D. 2012. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience,* 15**,** 511-517.

GOLUMBIC, E. M. Z., DING, N., BICKEL, S., LAKATOS, P., SCHEVON, C. A., MCKHANN, G. M., GOODMAN, R. R., EMERSON, R., MEHTA, A. D. & SIMON, J. Z. 2013. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. *Neuron,* 77**,** 980-991.

GRADY, C. L., VAN METER, J. W., MAISOG, J. M., PIETRINI, P., KRASUSKI, J. & RAUSCHECKER, J. P. 1997. Attention-related modulation of activity in primary and secondary auditory cortex. *Neuroreport,* 8**,** 2511-2516.

GRANT, K. W. & SEITZ, P.-F. 2000. The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America,* 108**,** 1197-1208.

GRANT, K. W., WALDEN, B. E. & SEITZ, P. F. 1998. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America,* 103**,** 2677-2690.

GRIMAULT, N., BACON, S. P. & MICHEYL, C. 2002. Auditory stream segregation on the basis of amplitude-modulation rate. *The Journal of the Acoustical Society of America,* 111**,** 1340-1348.

HARTMANN, W. M. & JOHNSON, D. 1991. Stream segregation and peripheral channeling. *Music Perception: An Interdisciplinary Journal,* 9**,** 155-183.

HENRY, M. J. & OBLESER, J. 2012. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proceedings of the National Academy of Sciences,* 109**,** 20095-20100.

HILLYARD, S. A., HINK, R. F., SCHWENT, V. L. & PICTON, T. W. 1973. Electrical signs of selective attention in the human brain. *Science,* 182**,** 177-180.

HILLYARD, S. A., VOGEL, E. K. & LUCK, S. J. 1998. Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society of London B: Biological Sciences,* 353**,** 1257-1270.

HOWARD, I. P. & TEMPLETON, W. B. 1966. Human spatial orientation.

IHLEFELD, A. & SHINN-CUNNINGHAM, B. 2008. Disentangling the effects of spatial cues on selection and formation of auditory objects. *Journal of the Acoustical Society of America,* 124**,** 2224-2235.

ISAIAH, A., VONGPAISAL, T., KING, A. J. & HARTLEY, D. E. H. 2014. Multisensory Training Improves Auditory Spatial Processing following Bilateral Cochlear Implantation. *The Journal of Neuroscience,* 34**,** 11119-11130.

IVERSON, P. 1995. Auditory stream segregation by musical timbre: effects of static and dynamic acoustic attributes. *Journal of Experimental Psychology: Human Perception and Performance,* 21**,** 751.

KAYSER, C., LOGOTHETIS, N. K. & PANZERI, S. 2010. Visual enhancement of the information representation in auditory cortex. *Current Biology,* 20**,** 19-24.

KAYSER, C., PETKOV, C. I. & LOGOTHETIS, N. K. 2008. Visual modulation of neurons in auditory cortex. *Cerebral Cortex,* 18**,** 1560-1574.

KAYSER, C., PETKOV, C. I. & LOGOTHETIS, N. K. 2009. Multisensory interactions in primate auditory cortex: fMRI and electrophysiology. *Hearing Research,* 258**,** 80-88.

KLEIN, R., BRENNAN, M. & GILANI, A. Covert cross-modality orienting of attention in space. annual meeting of the Psychonomic Society, Seattle, WA, 1987.

KOFFKA, K. 1935. Principles of GestaltpsychologyHarcourt Brace. *New York*.

KÖHLER, W. 1929. Gestalt psychology.

LAKATOS, P., CHEN, C.-M., O'CONNELL, M. N., MILLS, A. & SCHROEDER, C. E. 2007. Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron,* 53**,** 279-292.

LAKATOS, P., KARMOS, G., MEHTA, A. D., ULBERT, I. & SCHROEDER, C. E. 2008. Entrainment of neuronal oscillations as a mechanism of attentional selection. *science,* 320**,** 110-113.

LAKATOS, P., SHAH, A. S., KNUTH, K. H., ULBERT, I., KARMOS, G. & SCHROEDER, C. E. 2005. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology,* 94**,** 1904-1911.

LEE, A. K., LARSON, E., MADDOX, R. K. & SHINN-CUNNINGHAM, B. G. 2014. Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hearing Research,* 307**,** 111-120.

LOGOTHETIS, N. K. & SCHALL, J. D. 1989. Neuronal correlates of subjective visual perception. *Science,* 245**,** 761-763.

LOVELACE, C. T., STEIN, B. E. & WALLACE, M. T. 2003. An irrelevant light enhances auditory detection in humans: a psychophysical analysis of multisensory integration in stimulus detection. *Cognitive Brain Research,* 17**,** 447-453.

LUO, H., LIU, Z. & POEPPEL, D. 2010. Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS Biol,* 8**,** e1000445.

LUO, H. & POEPPEL, D. 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron,* 54**,** 1001-1010.

MACKEN, W. J., TREMBLAY, S., HOUGHTON, R. J., NICHOLLS, A. P. & JONES, D. M. 2003. Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology: Human Perception and Performance,* 29**,** 43.

MADDOX, R. K., ATILGAN, H., BIZLEY, J. K. & LEE, A. K. 2015. Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. e*life,* 4**,** e04995.

MARKS, L. E. 1974. On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American journal of psychology***,** 173-188.

MARKS, L. E. 1987. On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance,* 13**,** 384.

MAROZEAU, J., INNES-BROWN, H., GRAYDEN, D. B., BURKITT, A. N. & BLAMEY, P. J. 2010. The Effect of Visual Cues on Auditory Stream Segregation in Musicians and Non-Musicians. *Plos One,* 5.

MAURER, D. & MONDLOCH, C. 2004. Neonatal synesthesia: A reevaluation. *Synesthesia: Perspectives from Cognitive Neuroscience***,** 193-213.

MAURER, D., PATHMAN, T. & MONDLOCH, C. J. 2006. The shape of boubas: sound–shape correspondences in toddlers and adults. *Developmental Science,* 9**,** 316-322.

MCGURK, H. & MACDONALD, J. 1976. Hearing lips and seeing voices. *Nature,* 264**,** 746-748.

MELARA, R. D. 1989. Dimensional interaction between color and pitch. *Journal of Experimental Psychology: Human Perception and Performance,* 15**,** 69.

MELARA, R. D. & O'BRIEN, T. P. 1987. Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General,* 116**,** 323.

MESGARANI, N. & CHANG, E. F. 2012. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature,* 485**,** 233-236.

MESGARANI, N., DAVID, S. V., FRITZ, J. B. & SHAMMA, S. A. 2008. Phoneme representation and classification in primary auditory cortex. *Journal of the Acoustical Society of America,* 123**,** 899-909.

MESULAM, M.-M. 1998. From sensation to cognition. *Brain,* 121**,** 1013-1052.

MICHEYL, C., KREFT, H., SHAMMA, S. & OXENHAM, A. J. 2013. Temporal coherence versus harmonicity in auditory stream formation. *Journal of the Acoustical Society of America,* 133**,** El188-El194.

MICHEYL, C., TIAN, B., CARLYON, R. P. & RAUSCHECKER, J. P. 2005. Perceptual Organization of Tone Sequences in the Auditory Cortex of Awake Macaques. *Neuron,* 48**,** 139-148.

MILLER, G. A., HEISE, G. A. & LICHTEN, W. 1951. The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology,* 41**,** 329.

MILLER, G. A. & LICKLIDER, J. C. 1950. The intelligibility of interrupted speech. *The Journal of the Acoustical Society of America,* 22**,** 167-173.

MILLER, J. 1991. Channel interaction and the redundant-targets effect in bimodal divided attention. *Journal of Experimental Psychology: Human Perception and Performance,* 17**,** 160.

MOLHOLM, S., MARTINEZ, A., SHPANER, M. & FOXE, J. J. 2007. Object-based attention is multisensory: co-activation of an object's representations in ignored sensory modalities. *European Journal of Neuroscience,* 26**,** 499-509.

MOLHOLM, S., RITTER, W., MURRAY, M. M., JAVITT, D. C., SCHROEDER, C. E. & FOXE, J. J. 2002. Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive Brain Research,* 14**,** 115-128.

MOORE, B. C. 2012. *An introduction to the psychology of hearing*, Brill.

MOORE, B. C. & GOCKEL, H. 2002. Factors influencing sequential stream segregation. *Acta Acustica United with Acustica,* 88**,** 320-333.

MOORE, B. C. & GOCKEL, H. E. 2012. Properties of auditory stream formation. *Philosophical Transactions of the Royal Society BB,* 367**,** 919-931.

MOSSBRIDGE, J. A., GRABOWECKY, M. & SUZUKI, S. 2011. Changes in auditory frequency guide visual–spatial attention. *Cognition,* 121**,** 133-139.

NAHORNA, O., BERTHOMMIER, F. & SCHWARTZ, J.-L. 2012. Binding and unbinding the auditory and visual streams in the McGurk effect. *The Journal of the Acoustical Society of America,* 132**,** 1061-1077.

NAKAJIMA, Y., SASAKI, T., KANAFUKA, K., MIYAMOTO, A., REMIJN, G. & TEN HOOPEN, G. 2000. Illusory recouplings of onsets and terminations of glide tone components. *Perception & Psychophysics,* 62**,** 1413-1425.

NAVARRA, J., ALSIUS, A., SOTO-FARACO, S. & SPENCE, C. 2010a. Assessing the role of attention in the audiovisual integration of speech. *Information Fusion,* 11**,** 4-11.

NAVARRA, J., ALSIUS, A., VELASCO, I., SOTO-FARACO, S. & SPENCE, C. 2010b. Perception of audiovisual speech synchrony for native and non-native language. *Brain Research,* 1323**,** 84-93.

NELKEN, I. 2004. Processing of complex stimuli and natural scenes in the auditory cortex. *Current Opinion in Neurobiology,* 14**,** 474-480.

NELKEN, I. & BAR-YOSEF, O. 2008. Neurons and objects: the case of auditory cortex. *Frontiers in Neuroscience,* 2**,** 107.

NG, B. S. W., SCHROEDER, T. & KAYSER, C. 2012. A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. *The Journal of Neuroscience,* 32**,** 12268-12276.

O'SULLIVAN, J. A., SHAMMA, S. A. & LALOR, E. C. 2015. Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. *The Journal of Neuroscience,* 35**,** 7256-7263.

ORUC, I., SINNETT, S., BISCHOF, W. F., SOTO-FARACO, S., LOCK, K. & KINGSTONE, A. 2008. The effect of attention on the illusory capture of motion in bimodal stimuli. *Brain Research,* 1242**,** 200-208.

PARASKEVOPOULOS, E., KUCHENBUCH, A., HERHOLZ, S. C. & PANTEV, C. 2012. Evidence for training-induced plasticity in multisensory brain structures: an MEG study. *PloS One,* 7**,** e36534.

PARISE, C. & SPENCE, C. 2008. Synesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience Letters,* 442**,** 257-261.

PARISE, C. V. & SPENCE, C. 2009. 'When Birds of a Feather Flock Together': Synesthetic Correspondences Modulate Audiovisual Integration in Non-Synesthetes. *Plos One,* 4.

PARISE, C. V. & SPENCE, C. 2012. Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. *Experimental Brain Research,* 220**,** 319-333.

PATCHING, G. R. & QUINLAN, P. T. 2002. Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology: Human Perception and Performance,* 28**,** 755.

PEDLEY, P. E. & HARPER, R. S. 1959. Pitch and the vertical localization of sound. *The American Journal of Psychology***,** 447-449.

PETERSON, M. A. & SKOW-GRANT, E. 2003. Memory and learning in figure–ground perception. *Psychology of Learning and Motivation,* 42**,** 1-35.

PETKOV, C. I., KANG, X. J., ALHO, K., BERTRAND, O., YUND, E. W. & WOODS, D. L. 2004. Attentional modulation of human auditory cortex. *Nature Neuroscience,* 7**,** 658-663.

PETKOV, C. I., O'CONNOR, K. N. & SUTTER, M. L. 2003. Illusory sound perception in macaque monkeys. *The Journal of Neuroscience,* 23**,** 9155-9161.

PICHORA-FULLER, M. K., SCHNEIDER, B. A. & DANEMAN, M. 1995. How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America,* 97**,** 593-608.

POWERS III, A. R., HILLOCK-DUNN, A. & WALLACE, M. T. 2016. Generalization of multisensory perceptual learning. *Scientific Reports,* 6**,** 23374.

QUIROGA, R. Q., NADASDY, Z. & BEN-SHAUL, Y. 2004. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Computation,* 16**,** 1661-1687.

RAHNE, T., BÖCKMANN, M., VON SPECHT, H. & SUSSMAN, E. S. 2007. Visual cues can modulate integration and segregation of objects in auditory scene analysis. *Brain Research,* 1144**,** 127-135.

RAHNE, T. & BÖCKMANN-BARTHEL, M. 2009. Visual cues release the temporal coherence of auditory objects in auditory scene analysis. *Brain Research,* 1300**,** 125-134.

RAHNE, T., DEIKE, S., SELEZNEVA, E., BROSCH, M., KONIG, R., SCHEICH, H., BOCKMANN, M. & BRECHMANN, A. 2008. A multilevel and cross-modal approach towards neuronal mechanisms of auditory streaming. *Brain Research,* 1220**,** 118-131.

RAMACHANDRAN, V. S. & HUBBARD, E. M. 2001. Synaesthesia--a window into perception, thought and language. *Journal of Consciousness Studies,* 8**,** 3-34.

ROBERTS, B., GLASBERG, B. R. & MOORE, B. C. 2002. Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. *The Journal of the Acoustical Society of America,* 112**,** 2074-2085.

ROCKLAND, K. S. & OJIMA, H. 2003. Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology,* 50**,** 19-26.

ROSEN, S. 1992. Temporal Information in Speech: Acoustic, Auditory and Linguistic Aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* 336**,** 367-373.

SAMAHA, J., BAUER, P., CIMAROLI, S. & POSTLE, B. R. 2015. Top-down control of the phase of alpha-band oscillations as a mechanism for temporal prediction. *Proceedings of the National Academy of Sciences,* 112**,** 8439-8444.

SAMS, M., AULANKO, R., HÄMÄLÄINEN, M., HARI, R., LOUNASMAA, O. V., LU, S.-T. & SIMOLA, J. 1991. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters,* 127**,** 141-145.

SAMUEL, A. G. 1996. Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General,* 125**,** 28.

SAMUEL, A. G. & KRALJIC, T. 2009. Perceptual learning for speech. *Attention, Perception, & Psychophysics,* 71**,** 1207-1218.

SCHLOSS, K. B., STRAUSS, E. D. & PALMER, S. E. 2012. Object color preferences. *Journal of Vision,* 12**,** 66-66.

SCHNUPP, J. W., HALL, T. M., KOKELAAR, R. F. & AHMED, B. 2006. Plasticity of temporal pattern codes for vocalization stimuli in primary auditory cortex. *The Journal of Neuroscience,* 26**,** 4785-4795.

SCHORR, E. A., FOX, N. A., VAN WASSENHOVE, V. & KNUDSEN, E. I. 2005. Auditory-visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 18748-18750.

SCHROEDER, C. E., LAKATOS, P., KAJIKAWA, Y., PARTAN, S. & PUCE, A. 2008. Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences,* 12**,** 106-113.

SCHWARTZ, J.-L. & SAVARIAUX, C. 2014. No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology,* 10**,** e1003743.

SEEBA, F. & KLUMP, G. M. 2009. Stimulus familiarity affects perceptual restoration in the European starling (Sturnus vulgaris). *PLoS One,* 4**,** e5974.

SHAMMA, S., ELHILALI, M., MA, L., MICHEYL, C., OXENHAM, A. J., PRESSNITZER, D., YIN, P. B. & XU, Y. B. 2013. Temporal Coherence and the Streaming of Complex Sounds. *Basic Aspects of Hearing: Physiology and Perception,* 787**,** 535-543.

SHAMMA, S. A., ELHILALI, M. & MICHEYL, C. 2011. Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences,* 34**,** 114-123.

SHAMS, L. & SEITZ, A. R. 2008. Benefits of multisensory learning. *Trends in Cognitive Sciences,* 12**,** 411-417.

SHINN-CUNNINGHAM, B. G. 2008. Object-based auditory and visual attention. *Trends in Cognitive Sciences,* 12**,** 182-186.

SHINN-CUNNINGHAM, B. G. & BEST, V. 2008. Selective attention in normal and impaired hearing. *Trends in Amplification*, 12(4), 283-299.

SHINN-CUNNINGHAM, B. G., LEE, A. K. & OXENHAM, A. J. 2007. A sound element gets lost in perceptual competition. *Proceedings of the National Academy of Sciences,* 104**,** 12223-12227.

SNYDER, J. S. & ALAIN, C. 2007. Toward a neurophysiological theory of auditory stream segregation. *Psychological Bulletin,* 133**,** 780.

SNYDER, J. S., CARTER, O. L., HANNON, E. E. & ALAIN, C. 2009. Adaptation Reveals Multiple Levels of Representation in Auditory Stream Segregation. *Journal of Experimental Psychology-Human Perception and Performance,* 35**,** 1232-1244.

SOHOGLU, E. & CHAIT, M. 2016. Detecting and representing predictable structure during auditory scene analysis. *eLife,* 5**,** e19113.

SPENCE, C. & DEROY, O. 2012. Crossmodal correspondences: Innate or learned? *i-Perception,* 3**,** 316.

STEIN, B. E., LONDON, N., WILKINSON, L. K. & PRICE, D. D. 1996. Enhancement of perceived visual intensity by auditory stimuli: a psychophysical analysis. *Journal of Cognitive Neuroscience,* 8**,** 497-506.

STEVENSON, R. A., WILSON, M. M., POWERS, A. R. & WALLACE, M. T. 2013. The effects of visual training on multisensory temporal processing. *Experimental brain research. Experimentelle Hirnforschung. Experimentation Cerebrale,* 225**,** 479-489.

SUGITA, Y. 1997. Neuronal correlates of auditory induction in the cat cortex. *Neuroreport,* 8**,** 1155-1159.

SUMBY, W. H. & POLLACK, I. 1954. Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America,* 26**,** 212-215.

SUMMERFIELD, Q, 1987. Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (Eds.), Hearing by Eye: The Psychology of Lip-reading, Lawrence Erlbaum, London, 1987, pp. 3-51.

SUSSMAN, E., RITTER, W. & VAUGHAN, H. G. 1999. An investigation of the auditory streaming effect using event-related brain potentials. *Psychophysiology,* 36**,** 22-34.

SUSSMAN, E. & STEINSCHNEIDER, M. 2006. Neurophysiological evidence for context-dependent encoding of sensory input in human auditory cortex. *Brain Research,* 1075**,** 165-174.

SUSSMAN, E. S. & GUMENYUK, V. 2005. Organization of sequential sounds in auditory memory. *Neuroreport,* 16**,** 1519-1523.

SUSSMAN, E. S., HORVÁTH, J., WINKLER, I. & ORR, M. 2007. The role of attention in the formation of auditory streams. *Perception & Psychophysics,* 69**,** 136-152.

SWETS, J. A. 1964. Signal Detection and Recognition in Human Observers: Contemporary Readings.

TALSMA, D., SENKOWSKI, D., SOTO-FARACO, S. & WOLDORFF, M. G. 2010. The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences,* 14**,** 400-410.

TEKI, S., BARASCUD, N., PICARD, S., PAYNE, C., GRIFFITHS, T. D. & CHAIT, M. 2016. Neural Correlates of Auditory Figure-Ground Segregation Based on Temporal Coherence. *Cerebral Cortex,* 26**,** 3669-3680.

TEKI, S., CHAIT, M., KUMAR, S., SHAMMA, S. & GRIFFITHS, T. D. 2013. Segregation of complex acoustic scenes based on temporal coherence. e*life,* 2.

TEKI, S., CHAIT, M., KUMAR, S., VON KRIEGSTEIN, K. & GRIFFITHS, T. D. 2011. Brain bases for auditory stimulus-driven figure–ground segregation. *The Journal of Neuroscience,* 31**,** 164-171.

TIITINEN, H., SINKKONEN, J., REINIKAINEN, K., ALHO, K., LAVIKAINEN, J. & NÄÄTÄNEN, R. 1993. Selective attention enhances the auditory 40-Hz transient response in humans.59-60.

TOWN, S. M., ATILGAN, H., WOOD, K. C. & BIZLEY, J. K. 2015. The role of spectral cues in timbre discrimination by ferrets and humans. *The Journal of the Acoustical Society of America,* 137**,** 2870-2883.

TREMBLAY, C., CHAMPOUX, F., LEPORE, F. & THÉORET, H. 2010. Audiovisual fusion and cochlear implant proficiency. *Restorative Neurology and Neuroscience,* 28**,** 283-291.

TUKKER, J. J., FUENTEALBA, P., HARTWICH, K., SOMOGYI, P. & KLAUSBERGER, T. 2007. Cell type-specific tuning of hippocampal interneuron firing during gamma oscillations in vivo. *The Journal of Neuroscience,* 27**,** 8184-8189.

VAN NOORDEN, L. P. A. S. 1975. Temporal coherence in the perception of tone sequences.

VAN WASSENHOVE, V., GRANT, K. W. & POEPPEL, D. 2007. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia,* 45**,** 598-607.

VLIEGEN, J. & OXENHAM, A. J. 1999. Sequential stream segregation in the absence of spectral cues. *The Journal of the Acoustical Society of America,* 105**,** 339-346.

VOLOH, B. & WOMELSDORF, T. 2016. A Role of Phase-Resetting in Coordinating Large Scale Neural Networks During Attention and Goal-Directed Behavior. *Frontiers in Systems Neuroscience,* 10.

VROOMEN, J., BERTELSON, P. & DE GELDER, B. 2001. Directing spatial attention towards the illusory location of a ventriloquized sound. *Acta Psychologica,* 108**,** 21-33.

WAGEMANS, J., ELDER, J. H., KUBOVY, M., PALMER, S. E., PETERSON, M. A., SINGH, M. & VON DER HEYDT, R. 2012. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin,* 138**,** 1172.

WALKER, K. M. M., BIZLEY, J. K., KING, A. J. & SCHNUPP, J. W. H. 2011a. Cortical encoding of pitch: Recent results and open questions. *Hearing Research,* 271**,** 74-87.

WALKER, K. M. M., BIZLEY, J. K., KING, A. J. & SCHNUPP, J. W. H. 2011b. Multiplexed and Robust Representations of Sound Features in Auditory Cortex. *Journal of Neuroscience,* 31**,** 14565-14576.

WALKER, K. M. M., SCHNUPP, J. W. H., HART-SCHNUPP, S. M. B., KING, A. J. & BIZLEY, J. K. 2009. Pitch discrimination by ferrets for simple and complex sounds. *Journal of the Acoustical Society of America,* 126**,** 1321-1335.

WALKER, P. & SMITH, S. 1985. Stroop interference based on the multimodal correlates of haptic size and auditory pitch. *Perception,* 14**,** 729-736.

WANG, X.-J. 2010. Neurophysiological and computational principles of cortical rhythms in cognition. *Physiological Reviews,* 90**,** 1195-1268.

WARREN, R. M., OBUSEK, C. J. & ACKROFF, J. M. 1972. Auditory induction: Perceptual synthesis of absent sounds. *Science,* 176**,** 1149-1151.

WHALEN, D. & LIBERMAN, A. M. 1996. Limits on phonetic integration in duplex perception. *Perception & Psychophysics,* 58**,** 857-870.

WINKLER, I., DENHAM, S., MILL, R., BŐHM, T. M. & BENDIXEN, A. 2012. Multistability in auditory stream segregation: a predictive coding view. *Philosophical Transactions of the Royal Society of London B: Biological Sciences,* 367**,** 1001-1012.

WINKLER, I., SUSSMAN, E., TERVANIEMI, M., HORVÁTH, J., RITTER, W. & NÄÄTÄNEN, R. 2003. Preattentive auditory context effects. *Cognitive, Affective, & Behavioral Neuroscience,* 3**,** 57-77.

WINKLER, I., VAN ZUIJEN, T. L., SUSSMAN, E., HORVÁTH, J. & NÄÄTÄNEN, R. 2006. Object representation in the human auditory system. *European Journal of Neuroscience,* 24**,** 625-634.

WOODS, D. L., STECKER, G. C., RINNE, T., HERRON, T. J., CATE, A. D., YUND, E. W., LIAO, I. & KANG, X. 2009. Functional maps of human auditory cortex: effects of acoustic features and attention. *PLoS One,* 4**,** e5183.

ZENDEL, B. R. & ALAIN, C. 2009. Concurrent sound segregation is enhanced in musicians. *Journal of Cognitive Neuroscience,* 21**,** 1488-1498.

ZION GOLUMBIC, E. M., POEPPEL, D. & SCHROEDER, C. E. 2012. Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. *Brain and Language,* 122**,** 151-161.