ABSTRACT

Title of dissertation: NEUROMORPHIC MODEL FOR SOUND SOURCE SEGREGATION

Lakshmi Krishnan, Doctor of Philosophy, 2015

Dissertation directed by: Professor Shihab Shamma
Department of Electrical and Computer Engineering

While humans can easily segregate and track a speaker's voice in a loud noisy environment, most modern speech recognition systems still perform poorly in loud background noise. The computational principles behind auditory source segregation in humans is not yet fully understood. In this dissertation, we develop a computational model for source segregation inspired by auditory processing in the brain. To support the key principles behind the computational model, we conduct a series of electro-encephalography experiments using both simple tone-based stimuli and more natural speech stimulus.

Most source segregation algorithms utilize some form of prior information about the target speaker or use more than one simultaneous recording of the noisy speech mixtures. Other methods develop models on the noise characteristics. Source segregation of simultaneous speech mixtures with a single microphone recording and no knowledge of the target speaker is still a challenge.

Using the principle of temporal coherence, we develop a novel computational model that exploits the difference in the temporal evolution of features that belong to different sources to perform unsupervised monaural source segregation. While using no prior information about the target speaker, this method can gracefully incorporate knowledge about the target speaker to further enhance the segregation.Through a series of EEG experiments we collect neurological evidence to support the principle behind the model.

Aside from its unusual structure and computational innovations, the proposed model provides testable hypotheses of the physiological mechanisms of the remarkable perceptual ability of humans to segregate acoustic sources, and of its psychophysical manifestations in navigating complex sensory environments. Results from EEG experiments provide further insights into the assumptions behind the model and provide motivation for future single unit studies that can provide more direct evidence for the principle of temporal coherence.

NEUROMORPHIC MODEL FOR
SOUND SOURCE SEGREGATION

by

Lakshmi Krishnan

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Professor Shihab Shamma, Chair/Advisor
Professor Timothy Horiuchi
Professor Carol Espy Wilson
Professor Piya Pal
Professor Ramani Duraiswami, Dean's representative

Acknowledgments

I would like to thank my advisor, Professor Shihab Shamma for providing me such a rewarding graduate school experience. He has been a great mentor, always making himself available for questions and discussions and being extremely supportive. His enthusiasm and love for science has been infectious.

I would also like to thank members of my dissertation committee, Professor Carol-Espy Wilson, Professor Timothy Horiuchi, Professor Piya Pal and Professor Ramani Duraiswami. for sparing their time to be a part of my thesis defense.

I would like to express my gratitude to faculty members of the Department of Electrical and Computer Engineering, particularly Professor Jonathan Simon for his feedback on designing EEG experiments, Professor Krishnaprasad and Professor Steven Markus for providing valuable guidance during my first two years at University of Maryland. I would also like to thank Professor Cynthia Moss for providing me an opportunity to work in the BATLAB during my initial years at the University of Maryland.

I would also like to acknowledge help from some of the staff members. Tony Chan's technical help in the Brain, Behavior Science Lab is highly appreciated, as is the purchasing help from Regina King. I would like to thank the undergraduate students who assisted me with EEG data collection- Annie Thorton, Steven Nguyen, Yoga Dasari and Rodolfo Calderon.

I would like to thank my friends at University of Maryland for all the fun times - Spurthi, Anup, Sofia, Senthil, Raghav, Ganesh, Sameer, Sriram, Supriya, Rahul and Ashita. I would like to thank friends from the Neural Systems Lab - my office mate Dana Strait for all the discussions and free coffee, and Diego Elgueda, Majid Mirbhageri, Sahar Akram and Kai Lu for all their support.

I would like to thank my family for their support, my mother for her selfness and courage, my brother Ganesh for his constant encouragement, my uncles Ramesh, Rajamani and Balamani and my aunt Nirmala for urging me to come to the United States to pursue graduate school. Without their support and financial help, graduate school would have been a distant dream for me.

I would like to thank my husband, Arvind for being there for me all the time, through the ups and downs of this challenging journey and for his reassurances during times of self-doubt. His calmness and positive attitude helped me accomplish this. I would also like to thank my little boy, Ishaan for making everything even more worthwhile.

Table of Contents

List of Figures

List of Abbreviations

α	alpha
β	beta
AM	Amplitude Modulation
EEG	Electro-encephalography
aSSR	Auditory steady state response
DSS	Denoising source separation
HEOG	Horizontal electrooculogram
VEOG	Vertical electrooculogram

Chapter 1: Introduction

Humans and many animals have the remarkable ability to segregate a target talker from a noisy mixture of sound sources. Although the brain can segregate sound sources effortlessly, the performance of modern speech recognition systems degrades considerably in noisy environments. Understanding the fundamental neural mechanisms behind this remarkable perceptual phenomenon can enable better machine audition.

Auditory objects are the computational result of the auditory system's ability to extract, segregate and group spectro-temporal regularities in the acoustic environment [1, 2]. Resolving the component objects within the auditory scene depends on their temporal configuration [3]. Auditory objects have many general characteristics such as, they unfold over time i.e a single auditory object is comprised of a series of acoustic events, they possess spectro-temporal features that are separable from other auditory objects [4], possess invariance to changes in context of the spectro-temporal features [5] and the representations lend themselves to prediction of missing parts of the object [6–8].

Auditory objects are formed by means of simultaneous and sequential grouping principles that organize acoustic features into stable spectro-temporal entities [2].

Although attention is not necessary for auditory object formation, attention can influence the formation of auditory objects [9, 10].

Computational models of auditory scene analysis (CASA) have been proposed in the past to disentangle source mixtures and hence capture the functionality of this perceptual process. CASA methods for signal separation try to mimic the processing of human auditory system. Using low level features and bottom up segmentation and grouping techniques the spectrogram of a single source is extracted from the spectrogram of a mixture. Pitch, amplitude modulation and temporal continuity are used as cues to guide the segregation [11]. These algorithms can generalize to many types of signals while model based approaches [12, 13] may only work well with speakers or noise scenarios on which they have been trained. In related methods, segregation is coupled with a task such as recognition thus allowing a second stage of processing on top of the global segmentation that optimizes for the specific application scenario. While these techniques take inspiration from auditory processing in the brain, they don't seek to investigate the mechanisms behind source segregation in the brain. A different class of approaches emphasizes the biological mechanisms underlying this process and assesses both their plausibility and ability to replicate faithfully human psychoacoustics. Examples of such approaches range from models of the auditory periphery that explain how simple tone sequences may stream [14], to models that handle more elaborate sound sequences and bistable perceptual phenomena [15]. In this work, a novel computational architecture for unsupervised monaural source segregation is proposed based on the principle of extracting and grouping temporally coherent spectro-temporal features of sound sources. The key hypothesis behind the design of this model is tested through EEG experiments on human subjects. Electroencephalography(EEG) serves as a method to study perception and neural activity simultaneously. Recent MEG and EcoG studies [16–19] have shown that the envelope of the attended speech in a multi-talker scenario has higher correlation with the measured neural signal than the unattended speech. The computational principles that lead to enhancement of the neural representation of the attended speech is not yet well known. In the present work, we probe the enabling computational principles that can lead to enhanced representations of the target acoustic stimulus through systematic variation of the acoustic stimulus. In the first set of EEG experiments we use a two-tone stimulus, a classic stimulus paradigm studied in auditory streaming experiments [20,21]. While keeping the complexity of the task low, these initial studies aim at deconstructing the role of synchrony in the perception of auditory stimuli into streams or auditory objects. We then use more natural stimuli such as speech and build decoders in sensor space for estimating the attentional focus of a subject.

In the Chapter 2, the computational framework for monaural unsupervised source segregation is presented. In Chapter 3, the series of EEG experiments with tone stimulus is described. In Chapter 4, a technique to decode attention of human subjects while they listen to a speech mixture is presented. Chapter 5 concludes the thesis and presents future directions.

Chapter 2: Computational model for segmenting cluttered auditory scenes.

2.1 Introduction

Humans and many animals can effortlessly navigate complex sensory environments, segregating and attending to one desired target source while suppressing distracting and interfering others. In this chapter we present an algorithmic model that can accomplish this task with no prior information or training on complex signals such as speech mixtures, and speech in noise and music. The model accounts for this ability relying solely on the temporal coherence principle, the notion that perceived sources emit coherently modulated features that evoke coincident cortical response patterns. It further demonstrates how basic cortical mechanisms common to all sensory systems can implement the necessary representations and adaptive computations.

Humans and animals can attend to a sound source and segregate it rapidly from a background of many other sources, with no learning or prior exposure to the specific sounds. For humans, this is the essence of the well-known *cocktail party problem* in which a person can effortlessly conduct a conversation with a new acquaintance in a crowded and noisy environment [2, 22]. For frogs, songbirds, and penguins, this ability is vital for locating a mate or an offspring in the midst of a loud chorus [23, 24]. This capacity is matched by comparable object segregation feats in vision and other senses [25, 26], and hence understanding it will shed light on the neural mechanisms that are fundamental and ubiquitous across all sensory systems.

Computational models of auditory scene analysis have been proposed in the past to disentangle source mixtures and hence capture the functionality of this perceptual process. The models differ substantially in flavor and complexity depending on their overall objectives. For instance, some rely on prior information to segregate a specific target source or voice, and are usually able to reconstruct it with excellent quality [27]. Another class of algorithms relies on the availability of multiple microphones and the statistical independence among the sources to separate them, using for example ICA approaches or beam-forming principles [28]. Others are constrained by a single microphone and have instead opted to compute the spectrogram of the mixture, and then to decompose it into separate sources relying on heuristics, training, mild constraints on matrix factorizations [29–31], spectrotemporal masks [32], and gestalt rules [2, 33, 34]. A different class of approaches emphasizes the biological mechanisms underlying this process, and assesses both their plausibility and ability to replicate faithfully the psychoacoustics of stream segregation (with all their strengths and weaknesses). Examples of the latter approaches include models of the auditory periphery that explain how simple tone sequences may stream [14, 35, 36], how pitch modulations can be extracted and used to segregate sources of different pitch [37–39], and models that handle more elaborate sound sequences and bistable perceptual phenomena [30, 40–42]. Finally, of particular relevance here are algorithms that rely on the notion that features extracted from a given sound source can be bound together by correlations of intrinsic coupled oscillators in neural networks that form their connectivity online [42, 43]. It is fair to say, however, that the diversity of approaches and the continued strong interest in this problem suggest that no algorithm has yet achieved sufficient success to render the "cocktail party problem" solved from a theoretical, physiological, or applications point of view.

While our approach echoes some of the implicit or explicit ideas in the above-mentioned algorithms, it differs fundamentally in its overall framework and implementation. It is based on the notion that perceived sources (sound streams or objects) emit features , that are modulated in strength in a largely temporally coherent manner and that they evoke highly correlated response patterns in the brain. By clustering (or grouping) these responses one can reconstruct their underlying source, and also segregate it from other simultaneously interfering signals that are uncorrelated with it.

This simple principle of *temporal coherence* has already been shown to account experimentally for the perception of sources (or streams) in complex backgrounds [44–48]. However, this is the first detailed computational implementation of this idea that demonstrates how it works, and why it is so effective as a strategy to segregate spectrotemporally complex stimuli such as speech and music. Furthermore, it should be emphasized that despite apparent similarities, the idea of temporal coherence differs fundamentally from previous efforts that invoked correlations and synchronization in the following ways [49–53]: (1) coincidence here refers to that among modulated feature channels due to slow stimulus power (envelope) fluctuations, and not to any *intrinsic* brain oscillations; (2) coincidences are strictly done at cortical time-scales of a few hertz, and not at the fast pitch or acoustic frequency rates often considered; (3) coincidences are measured among modulated cortical features and perceptual attributes that usually occupy well-separated channels, unlike the crowded frequency channels of the auditory spectrogram; (4) coincidence must be measured over multiple time-scales and not just over a single time-window that is bound to be too long or too short for a subset of modulations; and finally (5) the details we describe later for how the coincidence matrices are exploited to segregate the sources are new and are critical for the success of this effort. For all these reasons, the simple principle of temporal coherence is not easily implementable. Our goal here is to show how to do so using plausible cortical mechanisms able to segregate realistic mixtures of complex signals.

As we shall demonstrate, the proposed framework mimics human and animal strategies to segregate sources with no prior information or knowledge of their properties. The model can also gracefully utilize available cognitive influences such as attention to, or memory of specific attributes of a source (e.g., its pitch or timbre) to segregate it from its background. We begin with a sketch of the model stages, with emphasis on the unique aspects critical for its function. We then explore how separation of feature channel responses and their temporal continuity contribute to source segregation, and the potential helpful role of perceptual attributes like pitch and location in this process. Finally, we extend the results to the segregation of complex natural signals such as speech mixtures, and speech in noise or music.

2.2 Results

The temporal coherence algorithm consists of an auditory model that transforms the acoustic stimulus to its cortical representation (Figure 1A). A subsequent stage computes a coincidence matrix (C-matrices in Figure 1B) that summarizes the pair-wise coincidences (or correlations at zero-lag) between all pairs of responses making up the cortical representation. A final auto-encoder network is then used to decompose the coincidence matrix into its different streams. The use of the cortical representation here is extremely important as it provides a multiresolution view of the signal's spectral and temporal features, and these in turn endow the process with its robust character. Details of these auditory transformations are described elsewhere [54], and summarized in **Methods** below for completeness.

Extracting streams from the coincidence matrices

The critical information for identifying the perceived sources is contained in the instantaneous coincidence among the feature channel pairs as depicted in the C-matrices (Figure 1B). At each modulation rate ω_i, the coincidence matrix at time t is computed by taking the outer product of all cortical frequency-scale (f, Ω) outputs $(X(t, f; \Omega, \omega_i))$. Such a computation effectively estimates simultaneously the "average coincidence" over the time window implicit in each ω_i rate, i.e., at different temporal resolutions, thus retaining both short- and long-term coincidence

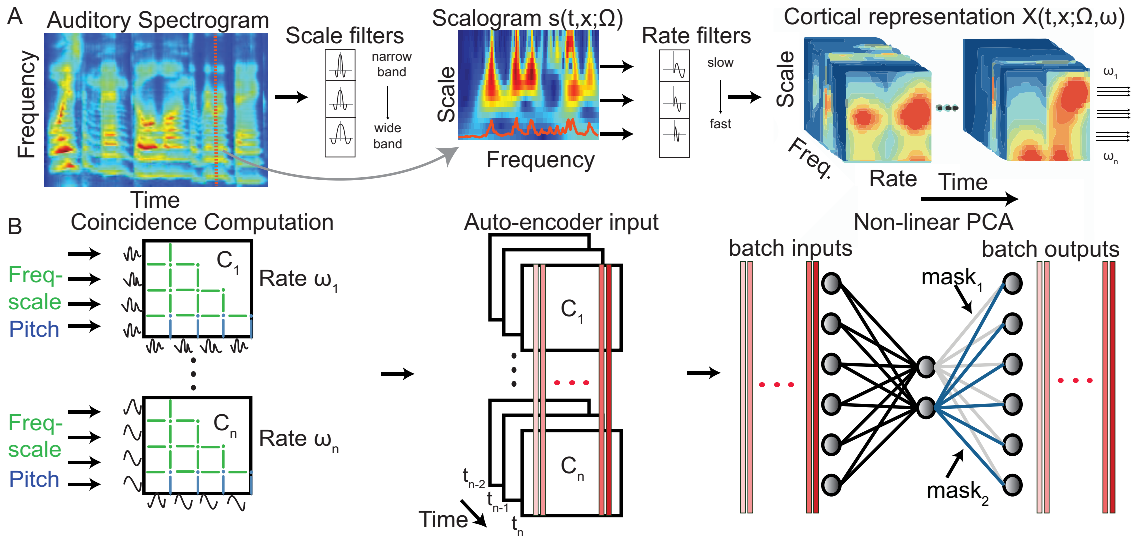

Figure 2.1: **The temporal coherence model consists of two stages**(A) Transformation of sound into a cortical representation [54]: It begins with a computation of the auditory spectrogram (left panel), followed by an analysis of its spectral and temporal modulations in two steps (middle and right panels, respectively): a multi-scale (or a multi-bandwidth) wavelet analysis along the spectral dimension to create the frequency-scale responses, $s(t,x;\Omega)$, followed by a wavelet analysis of the *modulus* of these outputs to create the final cortical outputs $X(t,x;\Omega,\omega)$ (right panel). (B) *Coincidence and clustering*: The cortical outputs at each time-step are used to compute a family of coincidence matrices (left panel). Each matrix (C_i) is the outer product of the cortical outputs $X(t,x;\Omega,\omega_i)$ (i.e., separately for each modulation rate ω_i). The C-matrices are then stacked (middle panel) and simultaneously decomposed by a nonlinear auto-encoder network (right panel) into two principal components corresponding to the foreground and background masks which are used to segregate the cortical response.

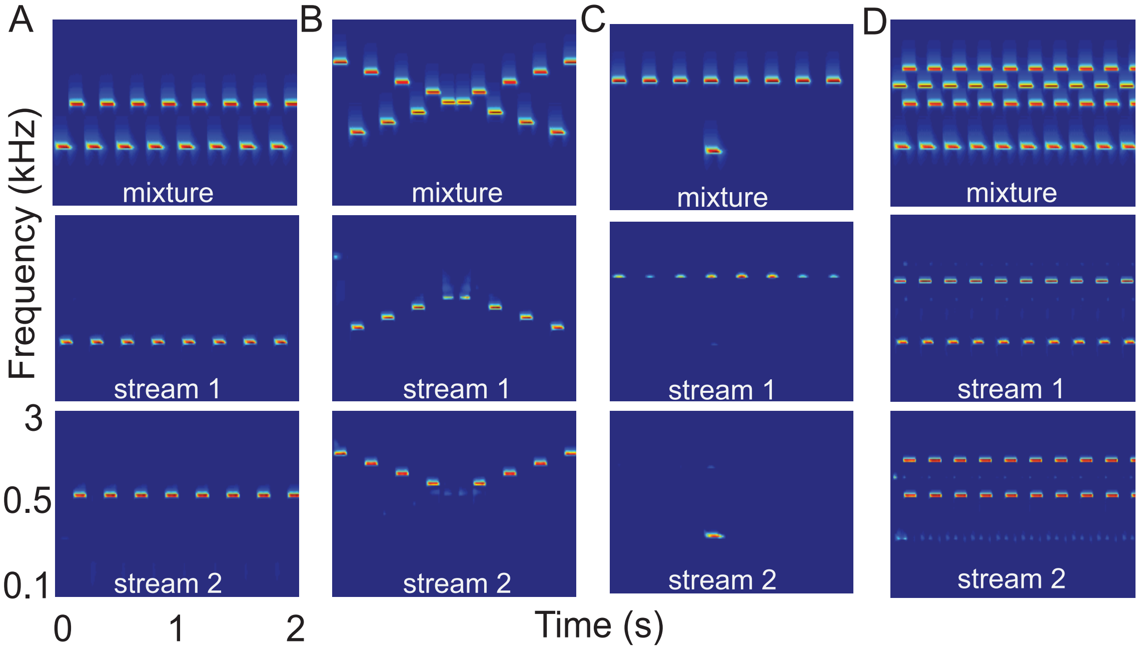

Figure 2.2: **Stream segregation of tone sequences and complexes**. Top row of panels represent the "mixture" audio whose two segregated streams are depicted in the middle and bottom rows. (A) The classic case of the well-separated alternating tones (top panel) becoming rapidly segregated into two streams (middle and bottom panels). (B) Continuity of the streams causes the crossing alternating tone sequences (top) to bounce maintaining an upper and a lower stream (middle and bottom panels). (C) Continuity also helps a stream maintain its integrity despite a transient synchronization with another tone. (D) When a sequence of tone complexes becomes desynchronized by more than 40 ms (top panel), they segregate into different streams despite a significant overlap (middle and bottom panels).

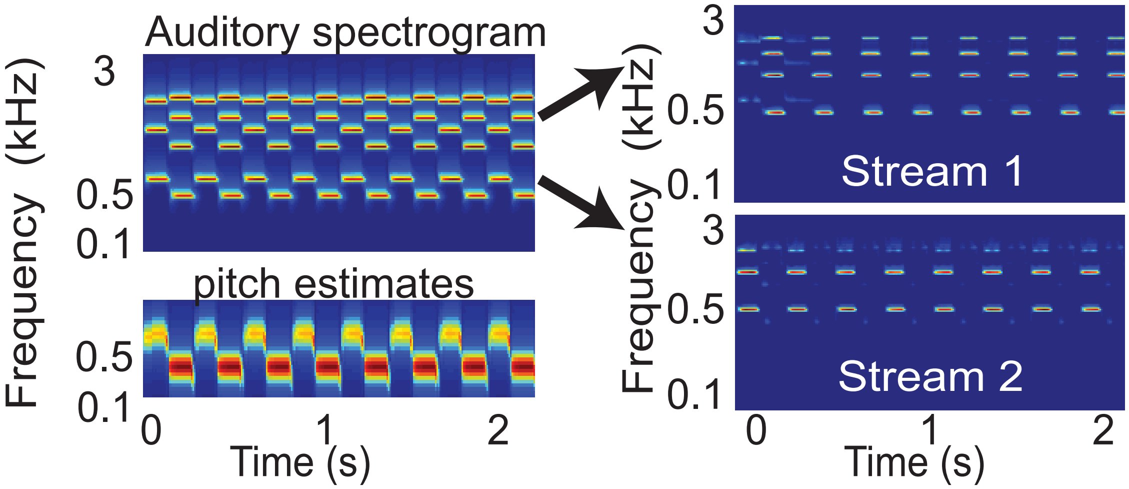

Figure 2.3: **Segregation of harmonic complexes by the temporal coherence model**. (A) A sequence of alternating harmonic complexes (pitches = 500 and 630 Hz). (B) The complexes are segregated using all spectral and pitch channels. Closely spaced harmonics (1890, 2000 Hz) mutually interact and hence their channels are only partially correlated with the remaining harmonics, becoming weak or may even vanish in the segregated streams.

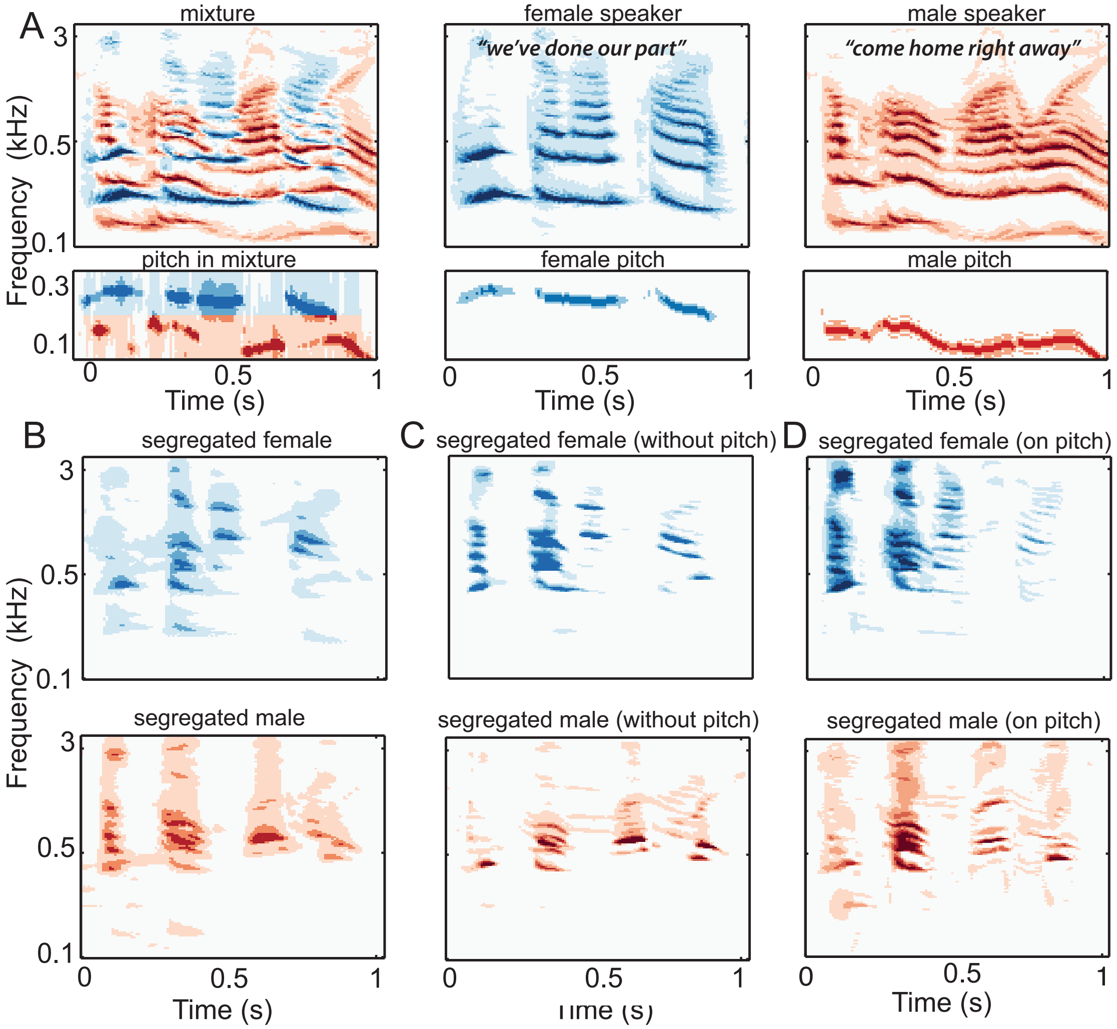

Figure 2.4: **Segregation of speech mixtures**. (A) Mixture of two sample utterances (left panel) spoken by a female (middle panel) and male (right panel); pitch tracks of the utterances are shown below each panel. (B) The segregated speech using all C-matrix columns. (C) The segregated speech using only coincidences among the frequency-scale channels (*no pitch* information). (D) The segregated speech using the channels surrounding the pitch channels of the female speaker as the anchor.

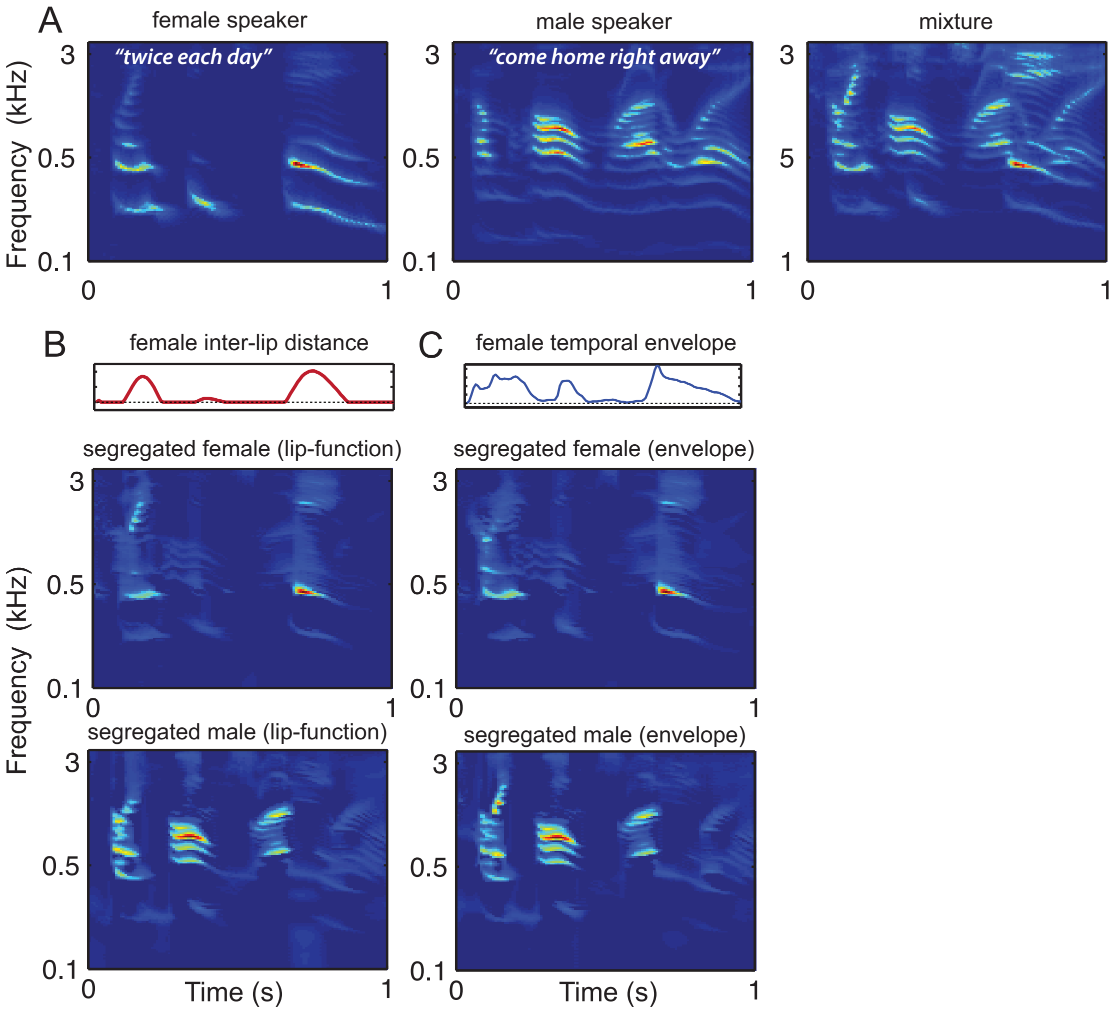

Figure 2.5: **Segregation of speech utterances based on auxiliary functions.** (A) Mixture of two sample utterances (right panel) spoken by a female (left panel) and male (middle panel) speakers; (B) The inter-lip distance of the female saying *"twice each day"*used as the anchor to segregate the mixture into its target female (middle panel) and the remaining male speech (bottom panel); (C) The envelope of the female speech *"twice each day"* used as anchor to segregate the mixture into its target female speaker (middle panel) and the remaining male speech (bottom speech).

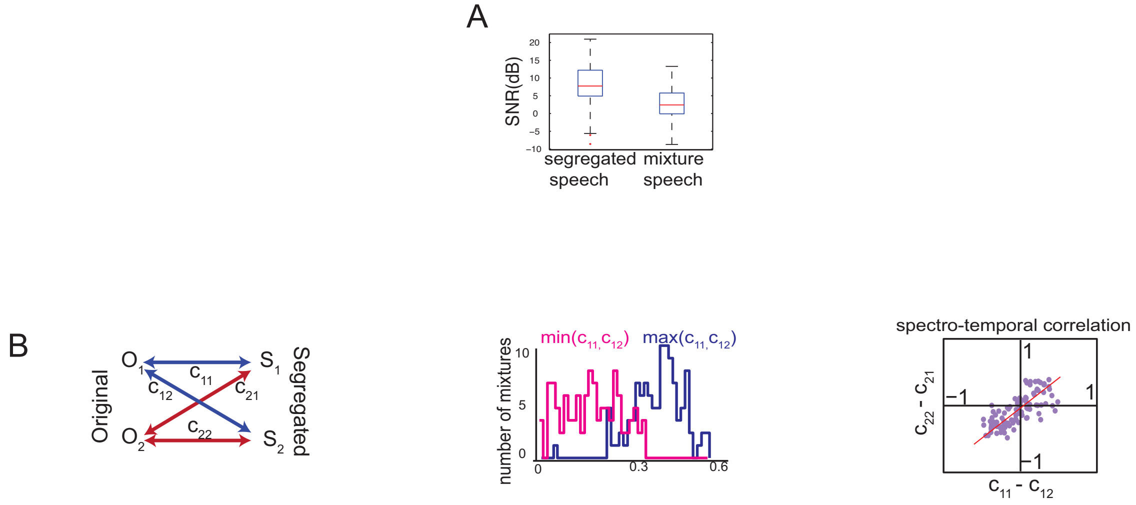

Figure 2.6: **Signal to Noise Ratio. (A)** Box plot of the SNR of the segregated speech and the mixture over 100 mixtures from the TIMIT corpus. **(B)** (Top) Notation used for coincidence measures computed between the original and segregated sentences plotted in panels below. (Middle) Distribution of coincidence in the cortical domain between each segregated speech and its corresponding original version (violet) and original interferer (magenta). 100 pairs of sentences from the TIMIT corpus were mixed together with equal power. (Bottom) Scatter plot of difference between correlation of original sentences with each segregated sentence demonstrates that the two segregated sentences correlate well with different original sentences.

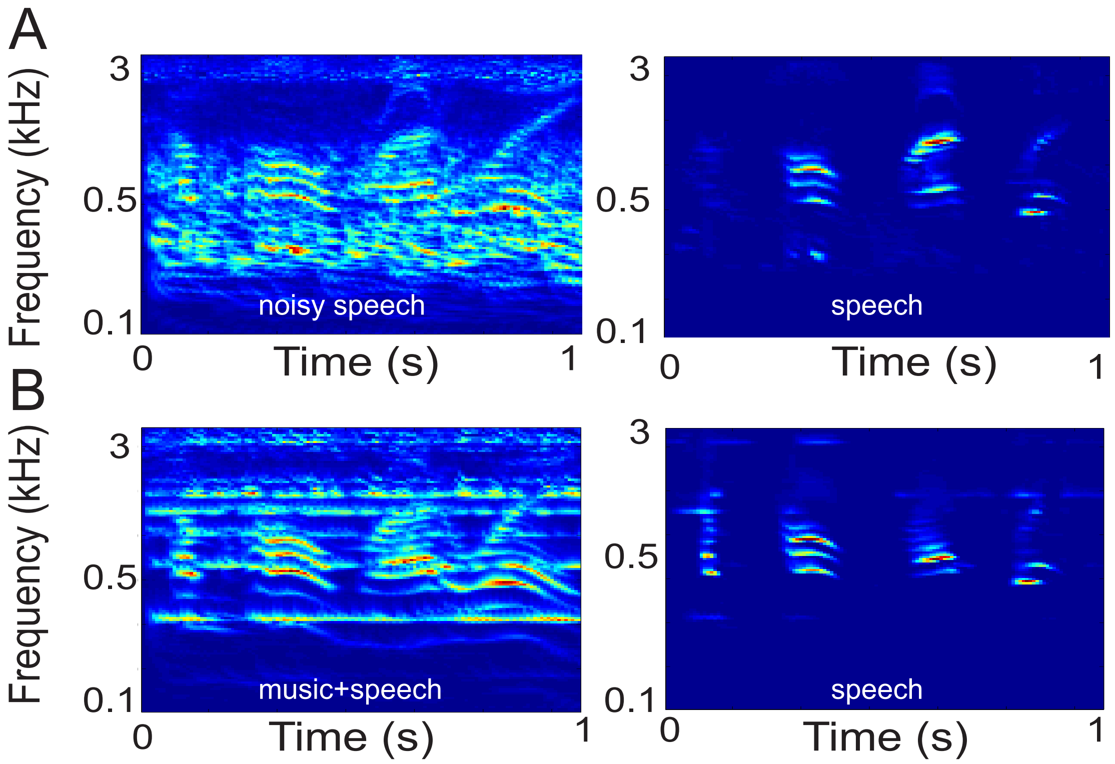

Figure 2.7: **Extraction of speech from noise and music**. (A) Speech mixed with street noise of many overlapping spectral peaks (left panel). The two signals are uncorrelated and hence can be readily segregated and the speech reconstructed (right panel). (B) Extraction of speech (right panel) from a mixture of speech and a sustained oboe melody (left panel).

measures crucial for segregation. Intuitively, the idea is that responses from pairs of channels that are strongly positively correlated should belong to the same stream, while channels that are uncorrelated or anti-correlated should belong to different streams. This decomposition need not be all-or-none, but rather responses of a given channel can be parceled to different streams in proportion to the degree of the average coincidence it exhibits with the two streams. This intuitive reasoning is captured by a factorization of the coincidence matrix into two uncorrelated streams by determining the direction of maximal incoherence between the incoming stimulus patterns. One such factorization algorithm is a nonlinear principal component analysis (nPCA) of the C-matrices [55], where the principal eigenvectors correspond to masks that select the channels that are positively correlated within a stream, and parcel out the others to a different stream. This procedure is implemented by an auto-encoder network with two rectifying linear hidden units corresponding to foreground and background streams as shown in Fig.1B (right panel). The weights computed in the output branches of each unit are associated with each of the two sources in the input mixture, and the number of hidden units can be automatically increased if more than two segregated streams are anticipated. The nPCA is preferred over a linear PCA because the former assigns the channels of the two (often anti-correlated) sources to different eigenvectors, instead of combining them on opposite directions of a single eigenvector [56].

Another key innovation in the model implementation is that the nPCA decomposition is performed not directly on the input data from the cortical model (which are modulated at ω_i rates), but rather on the columns of the C-matrices whose entries are either stationary or vary slowly regardless of the ω_i rates of the coincident channels. These common and slow dynamics enables stacking *all* C-matrices into one large matrix decomposition (Fig.1B). Specifically, the columns of the stacked matrices are applied (as a batch) to the auto-encoder network at each instant t with the aim of computing weights that can reconstruct them while minimizing the mean-square reconstruction error. Linking these matrices has two critical advantages: It ensures that the pair of eigenvectors from each matrix decomposition is consistently labeled across all matrices (e.g., source 1 is associated with eigenvector 1 in all matrices); It also couples the eigenvectors and balances their contributions to the minimization of the MSE in the auto-encoder. The weight vectors thus computed are then applied as masks on the cortical outputs $X(t, f; \Omega, \omega)$. This procedure is repeated at each time step as the coincidence matrices evolve with the changing inputs.

Role of feature separation, temporal continuity, and pitch in source segregation

The separation of feature responses on different channels and their temporal continuity are two important properties of the model that allow temporal coherence to segregate sources. Several additional perceptual attributes can play a significant role including pitch, spatial location, and timbre. Here we shall focus on pitch as an example of such attributes.

Feature separation: This refers to the notion that for two sounds to be segregated, it is necessary (but insufficient) that their features induce responses in mostly different auditory channels. Temporal coherence then serves to bind the coincident channels and segregate them as one source. For example, the tone sequences of Figure 2A,B are well separated at the start, and are alternating and hence non-coincident. The sequences therefore quickly stream apart perceptually and become two segregated streams of high and low tones [2]. When the tones approach each other and their responses interact (as in Fig.2B), the channels become more coherent and the segregation fails, as is evident by the middle tones becoming momentarily attenuated in the two segregated sequences [44].

Temporal Continuity: The relatively slow dynamics of the cortical rate-filters (tuned at 2-16 Hz) confer this important property on streams. Specifically, the C-matrix entries inherit the dynamics of their rate-filters and hence change only as fast as the rate of their inputs, exhibiting an *inertia* or continuity. This explains why a tone sequence of rapidly alternating tones across two frequency channels splits into two streams each composed of slowly changing or stationary tones. By contrast, when a tone sequence changes its frequencies slowly, a stream can track the slow change and maintain the ongoing organization (as demonstrated by the slowly varying upper and lower frequency streams of the "bouncing-tone" sequence in Fig.2B). Another example is when a new distant-frequency tone suddenly appears in a sequence, the C-matrix entries cannot track it rapidly enough causing the sequence to segregate and form a new stream that perceptually pops-out of the ongoing background (Fig.2C). Finally, the bandpass character of cortical rate-filtering enhances the response to tone onsets (relative to their sustained portions), and hence repeated desynchronization of *onsets* is sufficient to segregate tone sequences despite extensive overlap as seen in Fig.2D. These same phenomena are commonly seen with mixtures of more complex signals such as speech and music where the continuity of different streams is maintained despite transient synchronization and overlap.

How pitch contributes to segregation: Harmonic complexes evoke pitch percepts at their fundamental and are commonly found in speech and music (see **Methods** for details). Fig. 3A illustrates how two such alternating complexes with different pitches (500 Hz and 630 Hz) form two streams. Aside from the spectral channels, we also plot the pitch of the complexes alternating below the spectrograms. These pitch-grams contribute to the coincidence matrices much the same way any spectral channel does, i.e., as part of the feature vector that defines the emissions of each source, Thus, despite having some closely spaced harmonics (1890, 2000 Hz), the two complexes are sufficiently different (both in pitch and spectral components) that they remain largely uncorrelated and hence are readily segregated. The C-matrices in this simulation utilize all spectral and pitch channels. However, not all these channels are necessary as comparable segregation is achieved based only on the spectral scale-frequency inputs. Since the pitch channels are correlated with their own spectral harmonics, it is sufficient to compute the nPCA decomposition only on the columns of the pitch channels in the C-matrices (see **Methods** for more details) to segregate the two complex sequences. Using coincidences between spectral scale-frequency inputs or coincidences with respect to pitch channels alone also yield similar segregation.In fact, if the pitch range of one harmonic complex is known (e.g., the pitch of the first complex is in the range 450 to 550 Hz), then its stream can be readily extracted by iterating the auto-encoder on the columns of the C-matrix that lie *only* in this pitch range. All these variations illustrate that the C-matrices can be exploited in various ways to segregate sources depending on availability of the different sound attributes, and that even partial information is often sufficient to form the streams and bind all their correlated components together. For example, if the location information is extracted and is available to the C-matrices (as with the pitch-grams), then they can be exploited in parallel with, and in a manner exactly analogous to the pitch. Temporal coherence can similarly help segregate speech using co-modulated signals of other modalities as in lip-reading.

Segregating speech from mixtures

Speech mixtures share many of the same characteristics already seen in the examples of Fig.2 and Fig.3. For instance, they contain harmonic complexes with different pitches (e.g., males versus females) that often have closely spaced or temporally overlapped components. Speech also possesses other features such as broad bursts of noise immediately followed or preceded by voiced segments (as in various consonant-vowel combinations), or even accompanied by voicing (voiced consonants and fricatives). In all these cases, the syllabic onsets of one speaker synchronize a host of channels driven by the harmonics of the voicing, and that are desynchronized (or uncorrelated) with the channels driven by the other speaker. Fig. 4A depicts the clean spectra of two speech utterances (middle and right panels) and their mixture (left panel) illustrating the harmonic spectra and the temporal fluctuations in the speech signal at 3-7 Hz that make speech resemble the earlier harmonic sequences. The pitch tracks associated with each of these panels are shown below them.

Fig. 4B illustrates the segregation of the two speech streams from the mixture using all available coincidence among the spectral (frequency-scale) and pitch channels in the C-matrices. The reconstructed spectrograms are not identical to the originals (Fig.4A), an inevitable consequence of the energetic masking among the crisscrossing components of the two speakers. Nevertheless, with two speakers there are sufficient gaps between the syllables of each speaker to provide clean, unmasked views of the other speaker's signal [57]. If more speakers are added to the mix, such gaps become sparser and the amount of energetic masking increases, and that is why it is harder to segregate one speaker in a crowd if they are not distinguished by unique features or a louder signal. An interesting aspect of speech is that the relative amplitudes of its harmonics vary widely over time reflecting the changing formants of different phonemes. Consequently, the saliency of the harmonic components changes continually, with weaker ones dropping out of the mixture as they become completely masked by the stronger components. Despite these changes, speech syllables of one speaker maintain a stable representation of a sufficient number of features from one time instant to the next, and thus can maintain the continuity of their stream. This is especially true of the pitch (which changes only slowly and relatively little during normal speech). The same is true of the spectral region of maximum energy which reflects the average formant locations of a given speaker, reflecting partially the timbre and length of their vocal tract. Humans utilize either of these cues alone or in conjunction with additional cues to segregate mixtures.

For instance, to segregate speech with overlapping pitch ranges (a mixture of male speakers), one may rely on the different spectral envelopes (timbres), or on other potentially different features such as location or loudness. Humans can also exploit more complex factors such as higher-level linguistic knowledge and memory as we discuss later.

In the example of Fig.4C, the two speakers of Fig.4A are segregated based on the coincidence of only the spectral components conveyed by the frequency-scale channels. The extracted speech streams of the two speakers resemble the original unmixed signals, and their reconstructions exhibit significantly less mutual interference than the mixture as quantified later. Finally, as we discuss in more detail below, it is possible to segregate the speech mixture based on the pattern of correlations computed with one "anchor " feature such as the pitch channels of the female, i.e., using only the columns of the C-matrix near the female pitch channels as illustrated in Fig.4D.

Exactly the same logic can be applied to any auxiliary function that is co-modulated in the same manner as the rest of the speech signal. For instance, one may "look" at the lip movements of a speaker which open and close in a manner that closely reflects the instantaneous power in the signal (or its envelope) as demonstrated in [58]. These two functions (inter-lip distance and the acoustic envelope) can then be exploited to segregate the target speech much as with the pitch channels earlier. Thus, by simply computing the correlation between the lip function (Fig.5B) or the acoustic envelope (Fig.5C) with all the remaining channels, an effective mask can be readily computed to extract the target female speech (and the background male speech too). This example thus illustrates how in general any other co-modulated features of the speech signal (e.g., location, loudness, timbre, and visual signals such as lip movements can contribute to segregation of complex mixtures.

The performance of the model is quantified with a database of 100 mixtures formed from pairs of male-female speech randomly sampled from the TIMIT database (Fig.6) where the spectra of the clean speech are compared to those of the corresponding segregated versions. The signal-to-noise ratio is computed as

$$SNR\_segregated\_speech = \max(10 * \log(\frac{|S_1 * O_1|^2}{|S_1 * O_2|^2}), 10 * \log(\frac{|S_2 * O_1|^2}{|S_2 * O_2|^2})) \quad (2.1)$$

$$SNR\_mixture = 10 * \log(\frac{|M * O_1|^2}{|M * O_2|^2}) \quad (2.2)$$

where S_1, S_2 are the cortical representations of the segregated sentences and O_1, O_2 are the cortical representations of the original sentences and M is the cortical representation of the mixture. Average SNR improvement was 6dB for mixture waveforms mixed at 0dB.

Another way to demonstrate the effectiveness of the segregation is to compare the match between the segregated samples and their corresponding originals. This is evidenced by the minimal overlap in Fig.6B (middle panel) across the distributions of the coincidences computed between each segregated sentence and its original version versus the interfering speech. To compare directly these coincidences for each pair of mixed sentences, the difference between coincidences in each mixture are scatter-plotted in the bottom panel. Effective pairwise segregation (e.g., not extracting only one of the mixed sentences) places the scatter points along the diagonal.

Segregating speech from music and noise In principle, segregating mixtures does not depend on them being speech or music, but rather that the signals have different spectrotemporal patterns and exhibit a continuity of features. Fig. 7A illustrates the extraction of a speech signal from a highly overlapping temporally modulated street noise background. The same speech sample is extracted from a mixture with music in Fig. 7B. As explained earlier, this segregation (psychoacoustically and in the model) becomes more challenging in the absence of "clean looks", as when the background is an unmodulated white noise or babble that energetically masks the target speech.

Attention and memory in streaming

So far, attention and memory have played no direct role in the segregation, but adding them is relatively straightforward. From a computational point of view, attention can be interpreted as a focus directed to one or a few features or feature subspaces of the cortical model which enhances their amplitudes relative to other unattended features. For instance, in segregating speech mixtures, one might choose to attend specifically to the high female pitch in a group of male speakers (Fig.4D), or to attend to the location cues or the lip movements (Fig.5C) and rely on them to segregate the speakers. In these cases, only the appropriate subset of columns of the C-matrices are needed to compute the nPCA decomposition (Fig.1B). This is in fact also the interpretation of the simulations discussed in Fig. 3 for harmonic complexes. In all these cases, the segregation exploited only the C-matrix columns marking coincidences of the attended *anchor* channels (pitch, lip, loudness) with the remaining channels.

Memory can also be strongly implicated in stream segregation in that it constitutes *priors* about the sources which can be effectively utilized to process the C-matrices and perform the segregation. For example, in extracting the melody of the violins in a large orchestra, it is necessary to know first what the timbre of a violin is before one can turn the attentional focus to its unique spectral shape features and pitch range. One conceptually simple way (among many) of exploiting such information is to use as 'template' the average auto-encoder weights (masks) computed from iterating on clean patterns of a particular voice or instrument, and use the resulting weights to perform an initial segregation of the desired source by applying the mixture to the stored mask directly.

2.3 Discussion

A biologically plausible model of auditory cortical processing can be used to implement the perceptual organization of auditory scenes into distinct auditory objects (streams). Two key ingredients are essential: (1) a multidimensional cortical representation of sound that explicitly encodes various acoustic features along which streaming can be induced; (2) clustering of the temporally coherent features into different streams. Temporal coherence is quantified by the coincidence between all pairs of cortical channels, slowly integrated at cortical time-scales as described in

Fig. 1. An auto-encoder network mimicking Hebbian synaptic rules implements the clustering through nonlinear PCA to segregate the sound mixture into a foreground and a background.

The temporal coherence model segregates novel sounds based exclusively on the ongoing temporal coherence of their perceptual attributes. Previous efforts at exploiting explicitly or implicitly the correlations among stimulus features differed fundamentally in the details of their implementation. For example, some algorithms attempted to decompose directly the channels of the spectrogram representations [59] rather than the more distributed multi-scale cortical representations. They either used the fast phase-locked responses available in the early auditory system [60], or relied exclusively on the pitch-rate responses induced by interactions among the unresolved harmonics of a voiced sound [61]. Both these temporal cues, however, are much faster than cortical dynamics (> 100 Hz) and are highly volatile to the phase-shifts induced in different spectral regions by mildly reverberant environments. The cortical model instead naturally exploits multi-scale dynamics and spectral analyses to define the structure of all these computations as well as their parameters. For instance, the product of the wavelet coefficients (entries of the C-matrices) naturally compute the running-coincidence between the channel pairs, integrated over a time-interval determined by the time-constants of the cortical rate-filters (Fig.1 and **Methods**). This insures that all coincidences are integrated over time intervals that are commensurate with the dynamics of the underlying signals and that a balanced range of these windows are included to process slowly varying (2 Hz) up to rapidly changing (16 Hz) features.

The biological plausibility of this model rests on physiological and anatomical support for the two postulates of the model: a cortical multidimensional representation of sound and coherence-dependent computations. The cortical representation is the end-result of a sequence of transformations in the early and central auditory system with experimental support discussed in detail in [54]. The version used here incorporates only a frequency (tonotopic) axis, spectrotemporal analysis (scales and rates), and pitch analysis [62]. However, other features that are pre-cortically extracted can be readily added as inputs to the model such as spatial location (from interaural differences and elevation cues) and pitch of unresolved harmonics [63].

The second postulate concerns the crucial role of temporal coherence in streaming. It is a relatively recent hypothesis and hence direct tests remain scant. Nevertheless, targeted psychoacoustic studies have already provided perceptual support of the idea that coherence of stimulus-features is *necessary* for perception of streams [46–48, 64]. Parallel physiological experiments have also demonstrated that coherence is a critical ingredient in streaming and have provided indirect evidence of its mechanisms through rapidly adapting cooperative and competitive interactions between coherent and incoherent responses [45, 65]. Nevertheless, much more remains uncertain. For instance, where are these computations performed? How exactly are the (auto-encoder) clustering analyses implemented? And what exactly is the role of attentive listening (versus pre-attentive processing) in facilitating the various computations? All these uncertainties, however, invoke coincidence-based computations and adaptive mechanisms that have been widely studied or postulated such as coincidence detection and Hebbian associations [66, 67].

Dimensionality-reduction of the coincidence matrix (through nonlinear PCA) allows us effectively to cluster all correlated channels apart from others, thus grouping and designating them as belonging to distinct sources. This view bears a close relationship to the predictive clustering-based algorithm by [68] in which input feature vectors are gradually clustered (or routed) into distinct streams. In both the coherence and clustering algorithms, cortical dynamics play a crucial role in integrating incoming data into the appropriate streams, and therefore are expected to exhibit for the most part similar results. In some sense, the distinction between the two approaches is one of implementation rather than fundamental concepts. Clustering patterns and reducing their features are often (but not always) two sides of the same coin, and can be shown under certain conditions to be largely equivalent and yield similar clusters [69]. Nevertheless, from a biological perspective, it is important to adopt the correlation view as it suggests concrete mechanisms to explore.

Our emphasis thus far has been on demonstrating the ability of the model to perform unsupervised (automatic) source segregation, much like a listener that has no specific objectives. In reality, of course, humans and animals utilize intentions and attention to selectively segregate one source as the foreground against the remaining background. This operational mode would similarly apply in applications in which the user of a technology identifies a target voice to enhance and isolate from among several based on the pitch, timbre, location, or other attributes. The temporal coherence algorithm can be readily and gracefully adapted to incorporate such information and task objectives, as when specific subsets of the C-matrix columns are used to segregate a targeted stream (e.g., Fig.3 and Fig.4). In fact, our experience with the model suggests that segregation is usually of better quality and faster to compute with attentional priors.

2.4 Methods

The auditory representation

Sound is first transformed into its auditory spectrogram, followed by a cortical spectrotemporal analysis of the modulations of the spectrogram (Fig.2.1A) [54]. *Pitch* is an additional perceptual attribute that is derived from the resolved (low-order) harmonics and used in the model [62]. It is represented as a ‘pitch-gram’ of additional channels that are simply augmented to the cortical spectral channels prior to subsequent rate analysis (see below). Other perceptual attributes such as location and unresolved harmonic pitch can also be computed and represented by an array of channels analogously to the pitch estimates.

The auditory spectrogram, denoted by $y(t, f)$, is generated by a model of early auditory processing [70], which begins with an affine wavelet transform of the acoustic signal, followed by nonlinear rectification and compression, and lateral inhibition to sharpen features. This results in $F = 128$ frequency channels that are equally spaced on a logarithmic frequency axis over 5.2 octaves.

Cortical spectro-temporal analysis of the spectrogram is effectively performed in two steps [54]: a spectral wavelet decomposition followed by a temporal wavelet decomposition, as depicted in Fig.1A. The first analysis provides multi-scale (multi-bandwidth) views of each spectral slice $y(t,:)$, resulting in a 2D *frequency-scale* representation $s(t,f;\Omega)$. It is implemented by convolving the spectral slice with S complex-valued spectral receptive fields h_i similar to Gabor functions, parametrized by spectral tuning Ω_i, i.e., $s(t,f,\Omega_i) = h(t,f,\Omega_i) *_f y(t,f)$.

The outcome of this step is an array of FxS frequency-scale channels indexed by frequency f and local spectral bandwidth Ω_i at each time instant t. We typically used $S = 2$ to 5 scales in our simulations (e.g., $\Omega_i = 1, 2, 4, ...$ cyc/oct), producing S copies of the spectrogram channels with different degrees of spectral smoothing. In addition, the pitch of each spectrogram frame is also computed (if desired) using a harmonic template-matching algorithm [62]. Pitch values and saliency were then expressed as a *pitch-gram* (P) channels that are appended to the frequency-scale channels (Fig.1B).

The cortical rate-analysis is then applied to the modulus of each of the channel outputs in the freq-scale-pitch array by passing them through an array R of modulation-selective filters ($Q = 1$), each indexed by its center rate ω_i which range over $2 - 32$ Hz in 1/2 octave steps (Fig.1B). This temporal wavelet analysis of the response of each channel is described in detail in [54]. Therefore, the final representation of the cortical outputs (features) is along four axes denoted by $X(t,f,\Omega,\omega)$. It consists of R coincidence matrices per time frame, each of size $(FS+P)$x$(FS+P)$ (Fig.1B).

Coherence computations and nonlinear principal component analysis

The decomposition of the C-matrices is carried out as described earlier in Fig.1B. The iterative procedure to learn the auto-encoder weights employs Limited-memory Broyden-Fletcher-Goldfarb-Shannon (L-BFGS) method as implemented in [71]. The *output* weight vectors (Fig.1B) thus computed are subsequently applied as masks on the input channels $X(t, f, \Omega, \omega)$. This procedure that is repeated every time step using the weights learned in the previous time step as initial conditions to ensure that the assignment of the learned eigenvectors remains consistent over time. Note that the C matrices do not change rapidly, but rather slowly, as fast as the time-constants of their corresponding rate analyses allow ($\approx 1/\omega_i$). For example, for the $\omega_i = 4$ Hz filters, the cortical outputs change slowly reflecting a time-constant of approximately 250 ms. More often, however, the C-matrix entries change much slower reflecting the sustained coincidence patterns between different channels. For example, in the simple case of two alternating tones (Fig.2A), the C-matrix entries reach a steady state after a fraction of a second, and then remain constant reflecting the unchanging coincidence pattern between the two tones. Similarly, if the pitch of a speaker remains relatively constant, then the correlation between the harmonic channels remains approximately constant since the partials are modulated similarly in time. This aspect of the model explains the source of the continuity in the streams. The final step in the model is to invert the *masked* cortical outputs $X_m(t, f, \Omega, \omega)$ back to the sound [54].

2.5 Conclusions

In summary, we have described a model for segregating complex sound mixtures based on the temporal coherence principle. The model computes the coincidence of multi-scale cortical features and clusters the coherent responses as emanating from one source. It requires no prior information, statistics, or knowledge of source properties, but can gracefully incorporate them along with cognitive influences such as attention to, or memory of specific attributes of a target source to segregate it from its background. The model provides a testable framework of the physiological bases and psychophysical manifestations of this remarkable ability. Finally, the relevance of these ideas transcends the auditory modality to elucidate the robust visual perception of cluttered scenes [72, 73].

Chapter 3: Role of temporal coherence, attention and feature binding in parsing a complex auditory scene.

3.1 Overview

Humans and many animals can selectively attend to a target source and segregate it from competing sounds with remarkable ease [74–76]. The mechanism underlying this perceptual feat is not yet clearly understood. Recent studies have proposed that the principle of *temporal coherence* governs this process and organizes its many interrelated components [77, 78]. The *temporal coherence* hypothesis postulates that all features emanating from a single source fluctuate coherently in power over time, and that the listener tracks and utilizes this coherence to extract them from others that are temporally incoherent with them.

Several recent psychoacoustic, neuro-imaging and computational studies have demonstrated the relevance of *temporal coherence* in sound segregation and streaming. A series of psychoacoustic experiments demonstrated that it is far easier to segregate sequences of alternating tones than sequences of synchronous tones, even with large frequency separations [77, 79] or when one of the sequences is stationary while the other fluctuated slightly in frequencies [80]. The effects of temporal coherence and harmonicity in grouping auditory streams was studied in [81]. Temporal coherence also explained why a few synchronous tone sequences perceptually pop-out even in the midst of a dense background of random tones [82]. More recently, temporal coherence has also been demonstrated to play a role in co-modulation masking release [83].

The model presented in the previous chapter [84] has demonstrated in detail how temporal coherence can be exploited to segregate tone sequences as well as complex sound mixtures such as speech and music. Two key ingredients of the model provide the rationale for the experiments described in this chapter. The first is the coincidence measurements between pairs of neural channels encoding various acoustic features. The second basic ingredient of temporal coherence is the binding of coincident channels into one group representing a source. One conception of this associative process is inspired by Hebb's principle of *fire together, wire together* , illustrated by the schematic of Figure 3.1. It postulates that neurons with highly correlated responses form cooperative connectivity that can mutually enhance their responses. By contrast, highly uncorrelated activity leads to competitive (inhibitory) connectivity that suppresses the overall responses while emphasizing the differences between them. Since natural sounds are non-stationary in character, their perceptual features (pitch, timbre, location, and loudness) constantly evolve over time, and hence any correlative interactions (enhancement or suppressive) must be highly and rapidly adaptive in character so as to track the ongoing properties of stimuli.

In summary, the hypothesis behind the design of the experiments in this chap-

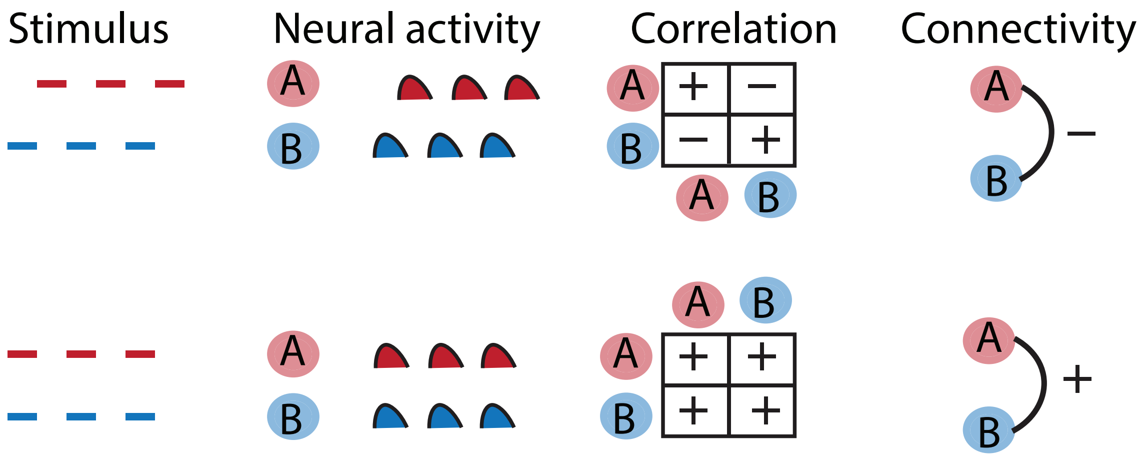

Figure 3.1: **Hebbian principle**. Schematic of coincidence computations between pairs of channels encoding acoustic features.

ter is that, when populations of neurons are driven synchronously or asynchronously, correspondingly facilitative or suppressive connectivity rapidly forms, and that temporarily persists providing the effective context for subsequent stimuli. It is further postulated that such adaptive effects require attention, a conjecture based on previous findings that cortical responses and tuning properties remain unchanged during passive listening [85, 86].

We design a series of EEG experiments to test this hypothesis. EEG provides for a high temporal resolution measurement of the neural activity. In the first experiment, we test this hypothesis in its simplest form, using alternating and synchronous tone sequences. The neural populations encoding the frequency of these tones fire coherently(anti-coherent) when the tones are synchronous (alternating). To facilitate stream segregation and to improve task understanding, one of these tones is amplitude modulated at 40 Hz. The 40 Hz steady state response power is compared across attention conditions for alternating and synchronous tone sequences. When the stimulus is alternating, we expect to see a modulation in the 40Hz steady state response power reflecting the attentional state of the subject. Thus, if the subject is selectively attending to the amplitude modulated sequence, the 40Hz aSSR should be enhanced, while if the subject is attending to the pure tone sequence, the 40Hz aSSR should be suppressed. More importantly, during the global attention condition, when the subject is attending to the auditory scene as one whole, the 40 Hz aSSR should still be suppressed since the two tone sequences are anti-coherent. On the contrary, if the tones are synchronous, we predict a weaker modulation in power in the selective attention condition since synchrony limits the ability of the listeners to segregate the tone sequences. The aSSR power in the global attention condition should be comparable to the selective attention condition because the two tone sequences perceptually fuse into one stream due to their coherence.

To directly contrast the alternating and synchronous conditions, in the second experiment we compare the response to a probe tone (AM tone) played at the end of each trial, while the subject is performing the global attention task. This ensures that stimulus induced differences in the response power are eliminated. Since the alternating stimulus facilitates stream segregation through competitive interactions between neural populations, we expect to see a reduction in the 40Hz aSSR of the probe tone, as these interactions leave the neural population that encodes the amplitude modulated tone in a suppressed state. In contrast, during the synchronous trials, facilitative interactions between the neural populations should produce a strong 40Hz aSSR of the probe tones. The timing of the probe tone was jittered (delay after end of trial - uniformly distributed between 200-250 ms) to avoid formation of an expectation of end of trial. Three different frequency separations between the tones were tested. Thus, in this experiment we seek to understand the effects of temporal coherence in a global attention task.

In the next experiment, we study the effects of temporal coherence and selective attention in a stream segregation task on alternating pure tones. We use an amplitude modulated sequence concurrent with the alternating pure tone sequence and compare the response of the AM tone during selective attention to one of the pure tone sequences. In particular, we study if the AM tone response is modulated during the task. Since the AM tone is only partially coherent with each of the pure tone sequences, we examine whether the aSSR at 40Hz is modulated based on attention to the pure tone sequence or unmodulated because of its partial coherence to the attended stream.

In the final experiment, we use alternating harmonic sequences to study the influence of temporal coherence in feature binding. Subjects were asked to pay attention to a target pitch in an alternating harmonic sequence. At the end of each trial, we measure the response to probe tones centered at harmonics of the attended pitch and the unattended pitch. We then compare the response to the probe tones under the two attention conditions (attend to pitch A and attend to pitch B).

3.2 Participants

A total of 45 subjects participated in the study. The study was divided into a series of four experiments. Ten subjects participated in each of the first three experiments involving pure tones and AM tones. Fifteen subjects participated in the fourth experiment involving harmonic complexes. All subjects reported normal hearing. Written informed consent was obtained from each participant and subjects were paid for their participation. The experimental protocol was approved by the Institutional Review Board of the University of Maryland.

3.3 Experiment 1: Selective attention vs. global attention

3.3.1 Stimulus design

Attention plays an important role in the parsing of complex auditory scenes. In this experiment, the neural signature of selective attention on a target stream is contrasted against the neural signature of global attention on the same auditory scene. During the selective attention task, the subject is focusing on a part of the auditory scene (target), ignoring all other competing sources (distractors). During the global attention task, the subject is experiencing the entire auditory scene as one whole, actively listening to the auditory scene. The observed neural signature is hypothesized to depend upon the degree of temporal coherence in the stimulus and the attentive state of the subject. The stimulus (Figure 3.2) consists of a sequence of pure tones (tone A at 420 Hz) and AM tones (tone B centered at 1000

Hz, amplitude modulated at 40 Hz), either alternating or synchronous with each other. Each tone is 225 msecs long. The inter-tone interval in both the streams is 225 msecs, with 20 tone repetitions in each stream resulting in 11 sec long trials. During the selective attention condition, subjects were asked to pay attention to either the pure tone sequence or AM tone sequence and report if they heard a 4dB intensity deviant in the attended tone stream. To ensure that the subject was not reporting overall changes in intensity level during a trial, distractor deviants were introduced in the competing stream. The deviants were randomly spaced in the 4th-20th tone positions. Approximately 50% of the trials had deviants in the attended stream. During the global attention condition, subjects were asked to report if they heard a deviant in both tone A and tone B. For each stimulus, three different attention conditions were tested namely, attend to pure tone sequence, attend to AM tone sequence and attend to both the tone sequences (global attention condition). Overall, this resulted in six different conditions (2 types of stimulus X 3 attention conditions).

3.3.2 Experimental procedure

Subjects were seated in a sound proof room and EEG data was acquired using a 64 channel Brainvision acti-Champ system at 1000Hz. Stimuli were presented through Etymotics Research ER-2 insert earphones at a comfortable loudness level(70dB). A short training module was presented before each experimental block. Subjects received feedback after each trial. The main experiment was split

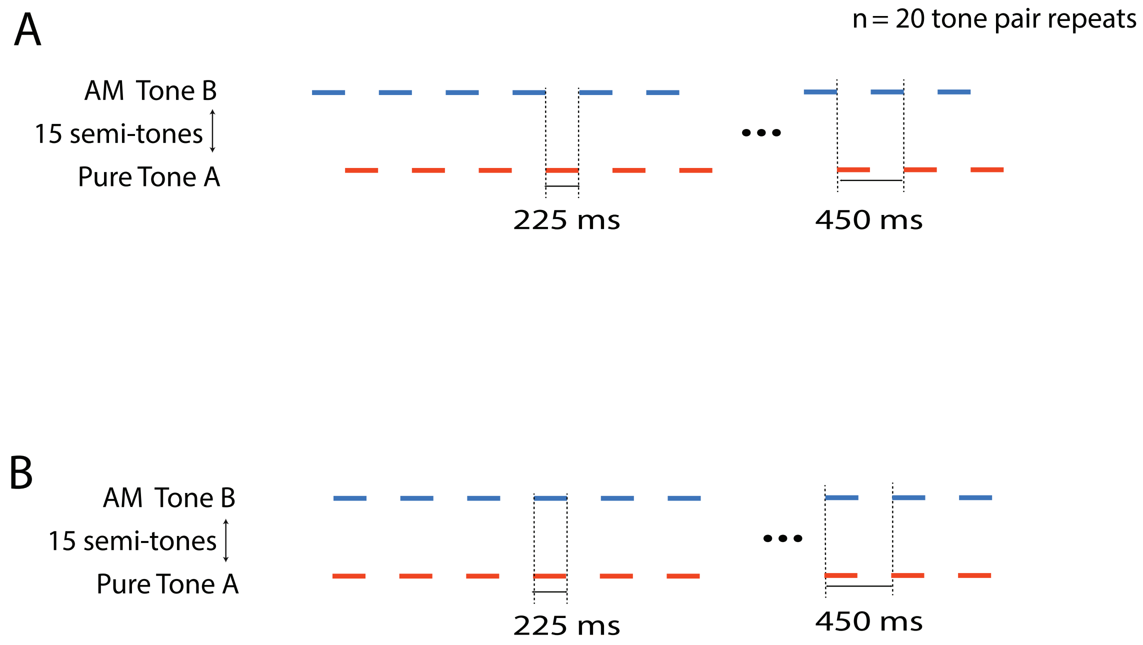

Figure 3.2: (A) **Alternating stimulus**. Two tones 15 semi-tones apart(tone B at 1000Hz and tone A at 420Hz), alternating with each other. Tone B is amplitude modulated at 40 Hz.(B) **Synchronous stimulus**. The same two tones, concurrent with each other. During the selective attention condition, subjects were asked to pay attention to either pure tone A or AM tone B and report if they heard an intensity deviant (4dB) in the attended tone stream. During the global attention condition subjects were asked to report if they heard a deviant in both tone A and tone B.

into three blocks. In the first block, subjects were asked to pay attention to the AM tone. During this block, the first 25 trials consisted of only the target stream to help familiarize the subject with the target stream. After 25 trials, a distractor stream (pure tone stream, in this case), alternating with the AM tone sequence was introduced. Subjects were still instructed to selectively attend to the AM stream. In the last 25 trials of the first block, the two tone streams were presented synchronously and the subject was asked to selectively attend to the target AM tone sequence and report if they heard an intensity deviant in the AM tone sequence. In the next block, the first 25 trials consisted of only the pure tone sequence followed by 25 trials of alternating tones and then 25 trials of synchronous tones, with the subject paying attention to the pure tone sequence throughout the block and report if they heard an intensity deviant in the pure tone sequence. The last block consisted of 25 trials of the alternating tones followed by 25 trials of synchronous tones during which the subject was asked to pay attention to both the streams and report if they heard an intensity deviant in both the streams. The stimuli were designed in Matlab and presented using the Psychtoolbox extension [87].

3.3.3 Data Analysis

Noisy channels were removed from the raw EEG data and the data was re-referenced to the average of the remaining channels. Data was then band-pass filtered between 1-70 Hz. The data was epoched into 9 second long trials and the leading and trailing 1.35 secs of the trial were chopped from these epochs. After removing outlier epochs, denoising source separation [88] was used to extract the most repeatable EEG component in each block. Each trial was folded into mini-epochs comprising of the response to a single AM tone, pure tone pair. The 40 Hz aSSR power was estimated from the power in the 125-200 msec region of averaged (evoked) AM tone mini-epoch.

3.3.4 Results

Subjects are given three separate tasks of attending to the AM tone sequence, attending to the pure tone sequence and globally attending to both the sequences, for the same stimuli comprising of either alternating or synchronous tone sequences. In each task, subjects are instructed to detect a 4dB intensity deviant in the attended stream. After an initial training block, all subjects were able to perform the task at a hit rate of more than 80%. The neural representation of the two tone sequence depends on the attentional focus of the subject. The 40Hz aSSR power is compared across attention conditions for the alternating and synchronous stimuli. The 40 Hz aSSR power is strongest during the attend AM task (Figure 3.4). For the alternating trials, the power at 40Hz is considerably weakened when the subject is paying attention to the pure tone reflecting the change in attentional focus. However, the most striking contrast between the alternating and synchronous trials is when the subject is paying global attention. During the global attention task, the power at 40Hz is considerably lower than the attend to AM condition if the stimulus is alternating. However, for the synchronous trials, the global attention power is not

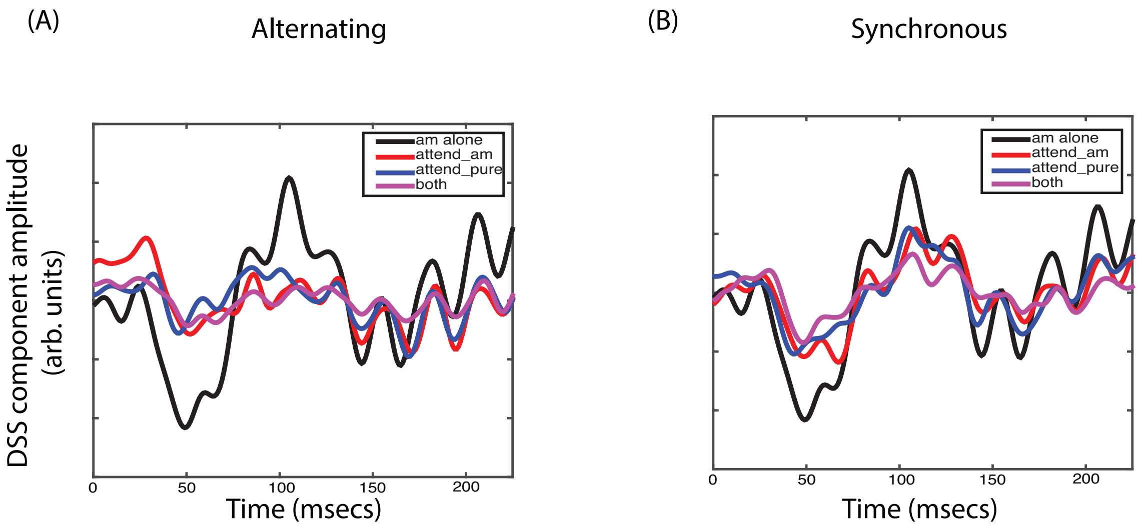

Figure 3.3: **DSS component for one exemplar subject**. (A) The most repeatable DSS component is plotted for the various attention conditions on alternating stimulus. Between 125-200 msecs the 40Hz aSSR is prominent. (B) The most repeatable DSS component is plotted for the various attention conditions on synchronous stimulus.

significantly different from the power during attend to AM tone condition.

3.4 Experiment 2: Effects of temporal coherence in the stimulus during a global attention task

3.4.1 Stimulus design

The stimulus (Figure 3.5) consists of the same alternating and synchronous tones as Experiment 1 with an additional test tone B at the end of each trial. The timing of the test tone was jittered to avoid formation of an expectation of end of trial. Subjects were asked to pay attention to both the tones. The test tone response was compared when the tones were preceded by alternating vs. synchronous tones

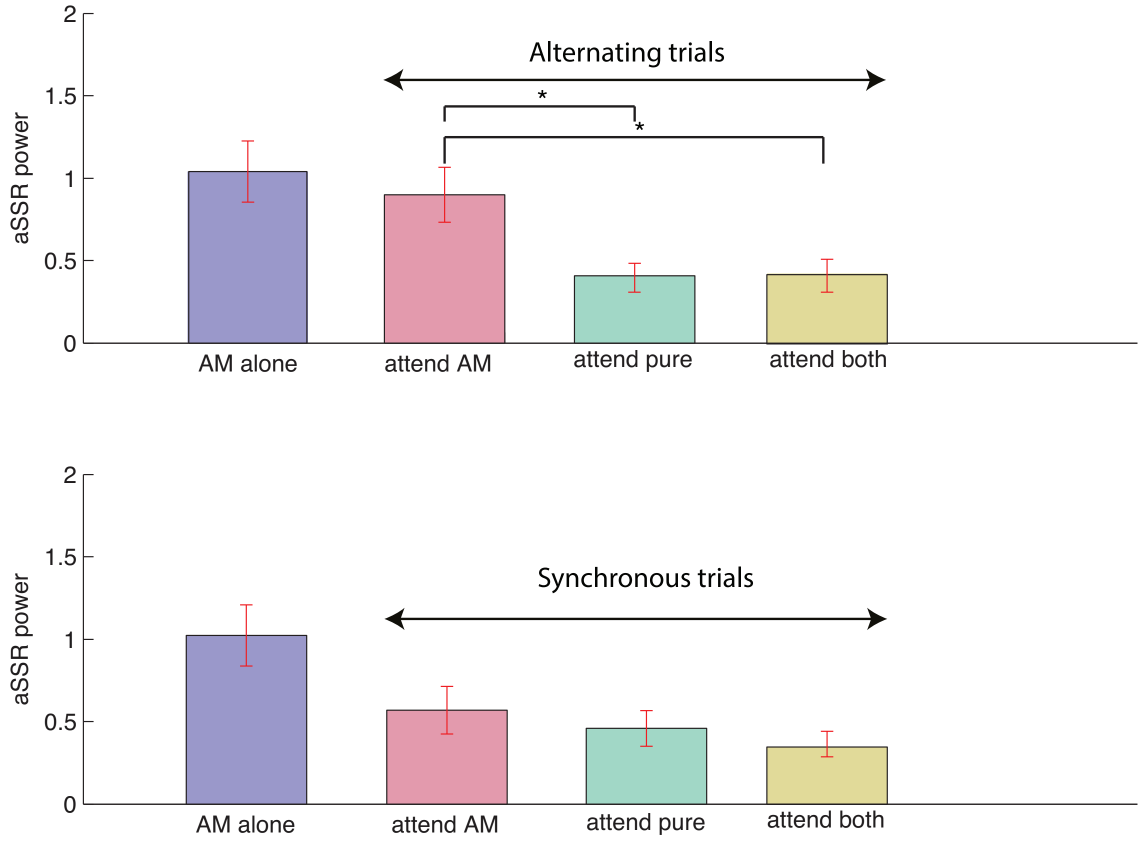

Figure 3.4: **Auditory steady state response power**. (Top) The aSSR power is compared under different attention conditions on the alternating stimulus. The aSSR power is strongest in the absence of any distractor tones. When a competing pure-tone sequence is alternating with the AM tone sequence, the aSSR is strongest when the subject is attending to the AM tone sequence. When the subject is attending to the pure tone sequence, the aSSR is suppressed, reflecting competing interactions between the two streams. Even when the subject is attending to both the streams, the aSSR at 40 Hz is suppressed and is comparable to the attend pure power. (Bottom) During the synchronous trials, the 40Hz aSSR is comparable across attention conditions.

for three different frequency separations between the two tones corresponding to 15, 9 and 6 semi-tones.

3.4.2 Experimental procedure

Subjects were seated in a sound proof room and EEG data was acquired using a 64 channel Brainvision acti-Champ system at 1000Hz. Stimuli were presented through Etymotics Research ER-2 insert earphones at a comfortable loudness level(70dB). A short training module was presented before each experimental block. Subjects received feedback after each trial. The main experiment was split into two repetitions of three blocks corresponding to three different frequency separation (15, 9, 6 semi-tones). In all the blocks, subjects were asked to pay global attention to both streams. During each block, the first 15 trials consisted of alternating trials and the next 15 trials consisted of synchronous trials. Subjects were instructed to report if they heard intensity deviants in both streams. The stimuli were designed in Matlab and presented using the Psychtoolbox extension [87].

3.4.3 Data Analysis

The raw EEG waveform was filtered between 0.1-70 Hz using a zero-phase 8th order Butterworth filter. The EEG waveform was then epoched into 9.475 sec long trials (9 sec long stimulus followed by the test tone). Outlier channels and outlier trials were rejected using a threshold criterion on total power. Power line noise was removed using DSS. The de-noised data was then analyzed using DSS to obtain the

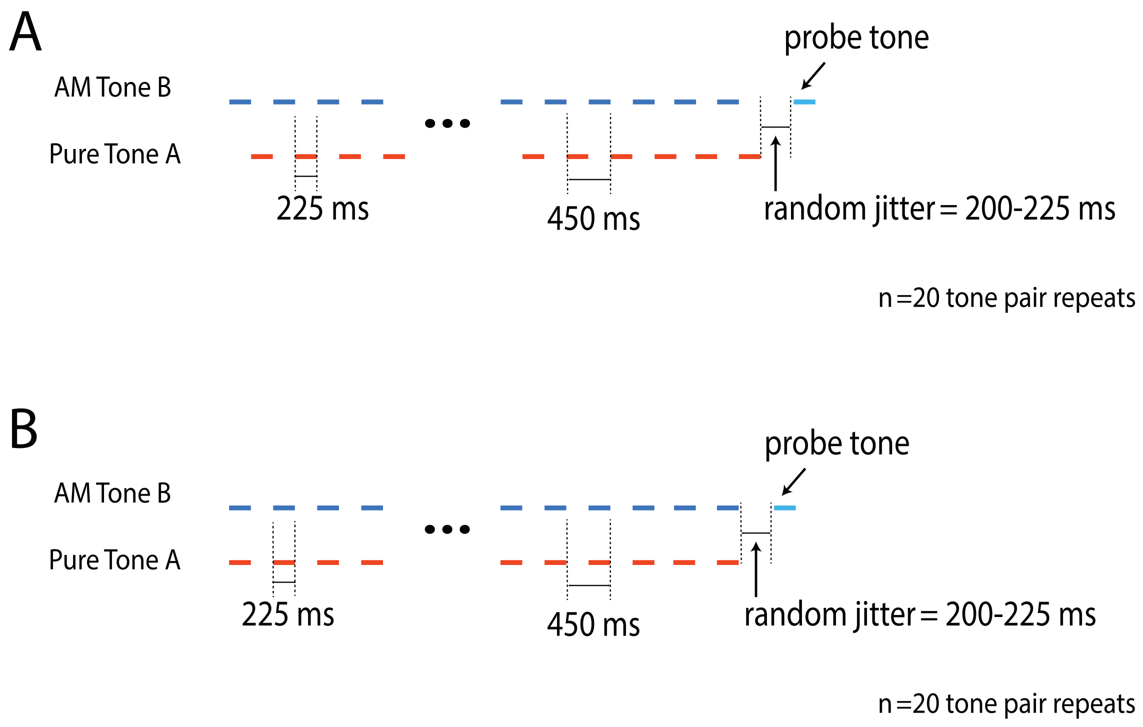

Figure 3.5: **Stimulus for experiment #2**. (A) **Alternating stimulus**. Two tones 15, 9 or 6 semi-tones apart(tone B at 1000Hz and tone A at 420 Hz, 595 Hz and 707Hz), alternating with each other. Tone B is amplitude modulated at 40 Hz. Each tone is 225 ms long and the inter-tone-interval for both the tones is 225 ms. After 20 tone-pair repeats, a test tone at tone B is played after a random jitter between 220-225 ms.(B) **Synchronous stimulus**. The same two tones, concurrent with each other. A test tone B is introduced at the end of each trail. Subjects were asked to report if they heard a deviant in both tone A and tone B and the test tone response was compared under different preceding context of alternating or synchronous stimuli.

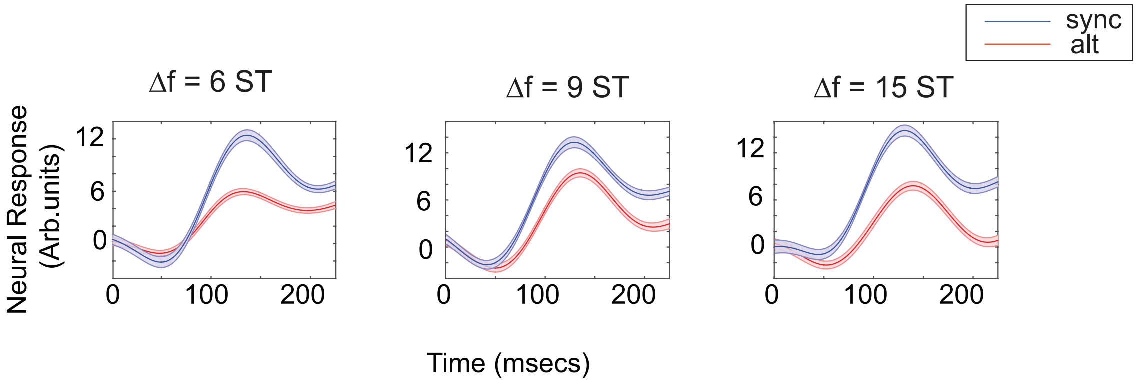

Figure 3.6: **Probe tone response.** The strength of the probe tone response is compared when the trials are alternating vs. synchronous. The probe tone response is much stronger when the preceding stimulus is synchronous than when the preceding stimulus is alternating.

most repeatable auditory component across trials.

3.4.4 Results

The test tone response was compared between alternating and synchronous trials, for three different frequency separations between the AM tone and pure tone. For all the frequency separation conditions, the test tone response is stronger when the preceding stimulus consists of synchronous tones Figure 3.6. Further, a latency trend is observed across frequency separations Figure 3.6. When the tones are 6 semi-tones apart, the test tone peak response is delayed more when the preceding stimulus is alternating, than when the preceding stimulus is synchronous. This delay is reduced as the frequency separation between the tones is increased. The probe tone neural response difference between the alternating and synchronous trials were averaged across 10 subjects.

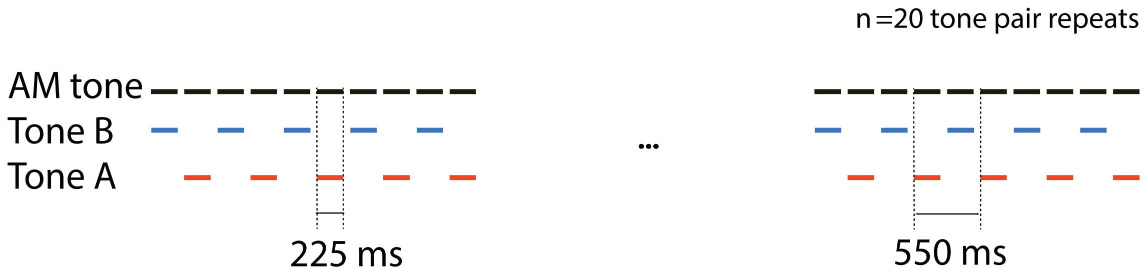

Figure 3.7: **Stimulus for experiment #3.** A sequence of two alternating pure tones 9 semi-tones apart(tone B at 600Hz and tone A at 357Hz) and an amplitude modulated tone concurrent with both the pure tones(centered at 1009 Hz, am rate = 40Hz).Subjects were asked to pay attention to either tone A or tone B and report if they heard an intensity deviant (4dB) in the attended tone stream. Each tone is 225 ms long and the inter-tone-interval between the A tones is 225 ms and B tones is 225 ms.. Modulation in the 40Hz aSSR was studied under different attention conditions.

3.5 Experiment 3: Effects of temporal coherence in the stimulus during a selective attention task

3.5.1 Stimulus design

The stimulus (Figure 3.7) consists of a sequence of two alternating pure tones 9 semi-tones apart(tone B at 600Hz and tone A at 357Hz) and an amplitude modulated tone concurrent with both the pure tones(centered at 1009 Hz, am rate = 40Hz). Subjects were asked to pay attention to either tone A or tone B and report if they heard an intensity deviant (4dB) in the attended tone stream. Modulation in the 40Hz aSSR was studied under different attention conditions.

3.5.2 Experimental procedure

Subjects were seated in a sound proof room and sounds were played through Etymotics research ER-2 insert earphones at a comfortable loudness level. EEG data was acquired using a 64-channel Brainvision acti-Champ system at 1000 Hz. The experiment was conducted in two blocks of 30 trials each. In the first block subjects were asked to pay attention to the tone at 600Hz and in the second block subjects were asked to pay attention to the tone at 357 Hz. To improve task understanding, first 15 trials of each block consisted only of the target frequency tone and the AM tone sequence. Before neural data collection, a training module was provided. Subjects received feedback after each trial during the experiment.

3.5.3 Data Analysis

Raw EEG data from the clean sensors was filtered between 1-48 Hz using zero-phase shift fft filters. Eye blink artifacts were removed from the data using the data collected in the HEOG and VEOG sensors as reference. The clean trials were split into mini-epochs consisting of the response to one low frequency tone - high frequency tone pair. The most repeatable EEG response during these mini-epochs was computed using a denoising source separation algorithm with a Ledoit Wolf covariance estimator. Figure 3.8 shows the spatial distribution of the most repeatable neural response for one subject. Figure 3.10 shows the neural response for the same subject. Data from 10 subjects were averaged to compute the grand average response. The 40 Hz power in the mini-epochs was compared under attend

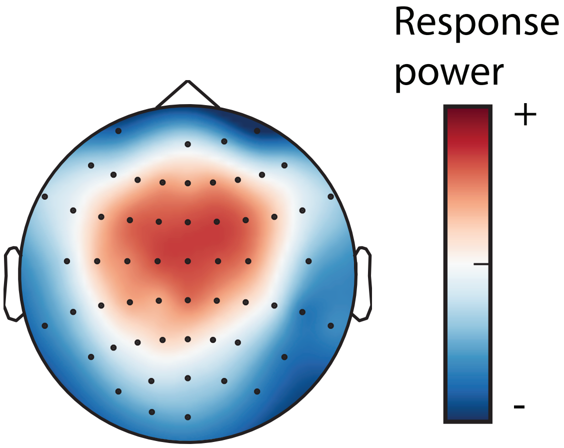

Figure 3.8: Spatial distribution of neural response for one representative subject.

to low frequency and attend to high frequency conditions.

3.5.4 Results

The 40Hz aSSR was compared during the tone A epochs and tone B epochs for two different attention conditions, namely attend to tone A and attend to tone B. The neural representation of the three tone sequence showed modulation in the 40Hz aSSR reflecting the attentional state of the subject. In general, the 40Hz aSSR was stronger during the tone A epochs (ttest, p¡¡0.01). During the attend to tone A task, the difference in the 40Hz aSSR between the tone A epochs and tone B epochs was much larger than during attend to tone B task (Figure 3.10).

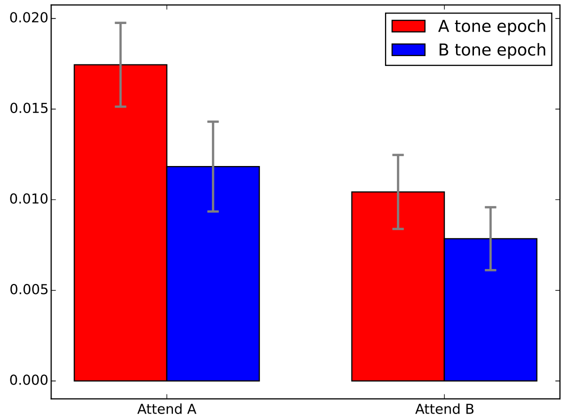

Figure 3.9: The bar plots represent the RMS power at 40Hz during different attention conditions in the A tone epoch and B tone epoch. The 40 Hz power during the A tone epochs is always larger than the 40 Hz power during the B tone epochs. However, when the subject is attending to tone B, the difference between the power in the A tone epochs and B tone epochs is smaller.

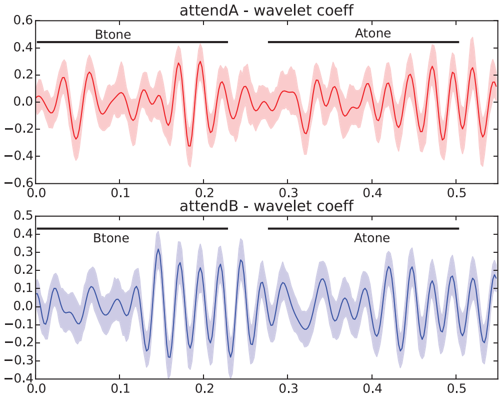

Figure 3.10: The wavelet component at a scale corresponding to 40 Hz is plotted during the attend to tone A and attend to tone B conditions.

3.6 Experiment 4: Temporal coherence and feature binding in the streaming of complex tones

3.6.1 Stimulus Design

To discern the mechanism of binding unique and shared features of two streams during selective attention, we use two recurring complex tone sequences (7 harmonics in each) as stimulus, with pitch at $f_a = 150$Hz and $f_b = 225$ Hz (Figure 3.11). This results in two shared frequency components (450 Hz and 900 Hz) in the competing streams. During the first 1 second of each trial, a prime tone sequence consisting of only the target tone complex is played. During the next two seconds of the trial, both the target and the distractor tone complex are played. In the last one second of the trial, only the probe tone sequence is played.The total stimulus duration is 4.625 secs if the prime tone is at pitch 125 Hz and 4.5 secs if the prime tone pitch is 225 Hz. In each trial, subjects were asked to pay attention to the leading sequence of tone complexes (prime tones) and report if they heard an intensity deviant in the attended stream. After each trial, we measure the neural response to a sequence of pure tones centered at frequencies that are either shared between the two complexes (450 Hz and 900 Hz) or are unique to tone complex A (150 Hz and 300 Hz) or unique to tone complex B (225Hz and 675Hz). The neural response to the probe tones is compared under different attention conditions.

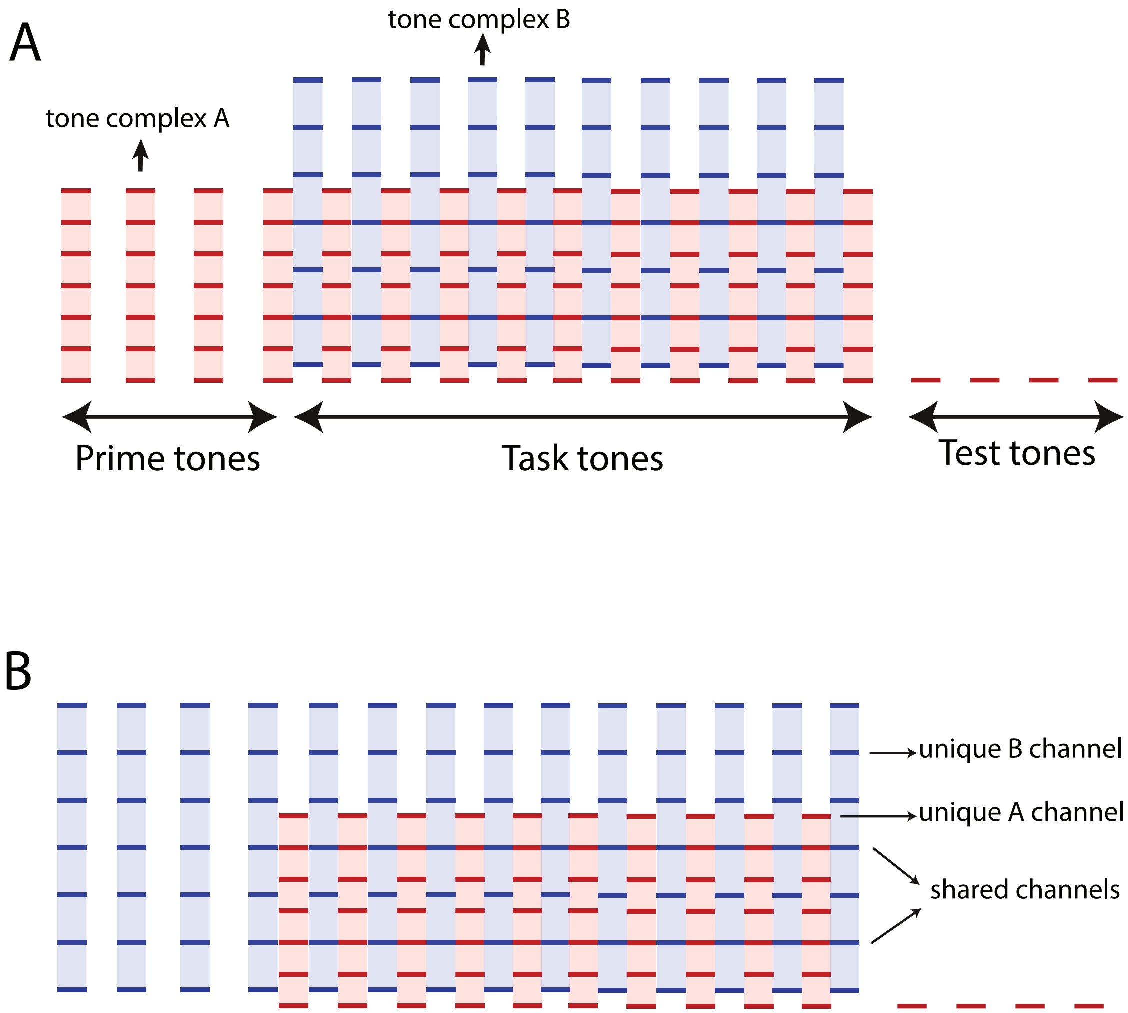

Figure 3.11: The stimulus consists of two harmonic complex sequences with pitches $f_a = 150$Hz and $f_b = 225$ Hz. In each task, subjects are instructed to selectively attend to either tone complex A or tone complex B and detect intensity deviants in the target stream. (A) **Attend to tone complex A, test tone frequency = $\mathbf{a}_0$**. For the first 1 second of the trial, only tone complex A is played (prime tones), followed by 2.5 seconds of both the harmonic complexes (task tones). During the last 1 second of the trial, test tones are played at 4 Hz. The frequency of these test tones were varied to probe the response at frequencies unique to tone complex A, unique to tone complex B and shared between both tone complexes. (B) **Attend to tone complex B, test tone frequency = $\mathbf{a}_0$**. For the first 1 second of the trial, only tone complex B is played (prime tones), followed by 2.5 seconds of both the harmonic complexes (task tones). During the last 1 second of the trial, test tones are played at 4 Hz.

3.6.2 Experimental procedure

Subjects were seated in a sound proof room and sounds were played through Etymotics research ER-2 insert earphones at a comfortable loudness level. EEG data was acquired using a 64-channel Brainvision acti-Champ system at 1000 Hz. The left mastoid was used as the reference for the recording. The experiment was conducted in three blocks of 60 trials each. The prime tone was tone complex B during the first 30 trials of each block, and tone complex A during the last 30 trials. During the first block, the probe tone frequency alternated between 150 Hz and 225 Hz every trial. During the second block, the probe tone frequencies were 450 Hz and 675 Hz, in the last block the probe tone frequencies were 300Hz and 900 Hz. Before neural data collection, a training module was provided. Subjects received feedback after each trial.

3.6.3 Data analysis

Data from corrupt sensors were removed using a threshold criterion on the variance of each sensor. Data from the clean sensors was filtered between 2-10 Hz using a zero-phase fft filter with transition bandwidth of 0.5 Hz. Eye-blink artifacts were regressed out of the data using the data recorded on HEOG and VEOG sensors. Data was then epoched and baseline corrected (using 200 msecs before the start of each trial). Figure 3.12 shows data from one representative subject after the removal of eye-blink artifacts.The most repeatable EEG response across trials was computed using an improved de-noising source separation algorithm, that uses a Ledoit Wolf covariance estimator [89]. The neural response was averaged across 15 subjects. The average neural response across 15 subjects is plotted in Figure 3.13. All the data analysis was done in python using the mne-python package [90, 91].

3.6.4 Results

Figure 3.13 shows the average neural response across 15 subjects. When the subject is paying attention to tone complex A(tone complex B), response to probe tones which are unique harmonics of tone complex A(tone complex B) are enhanced, while the response to unique harmonics of tone complex B(tone complex A) don't show such enhancement. The response to the shared harmonics of tone complex A and tone complex B are comparable in both attention conditions. The average RMS probe tone response across different attention conditions is summarized in Figure 3.14.

3.7 Discussion

This study explores the mechanisms of temporal coherence and attention in the binding of features that belong to one auditory source and its segregation from features that belong to other sources in a cluttered auditory scene. Subjects were performing an intensity deviance detection task while their EEG was simultaneously recorded. In the first experiment subjects performed two types of tasks. In the first task, listeners judged whether they heard an oddball intensity deviant at a given target frequency in the presence of an alternating/synchronous distractor (selective

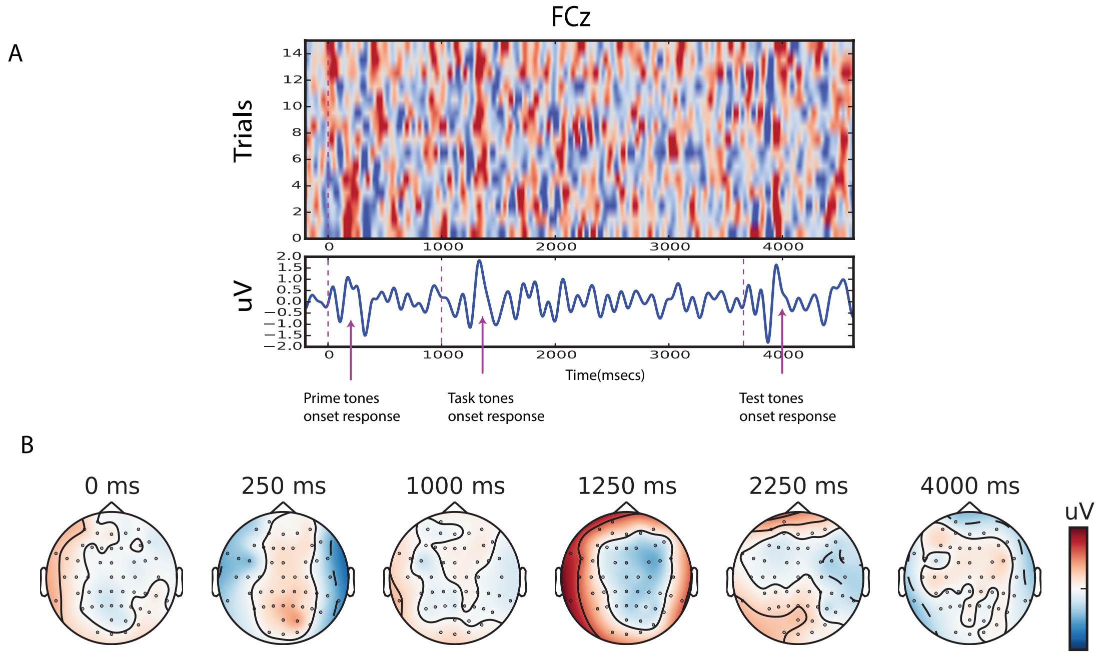

Figure 3.12: **Raw data from one representative subject.** (A) Response across trials on channel FCz. (B) Topoplot of evoked response as function of time.

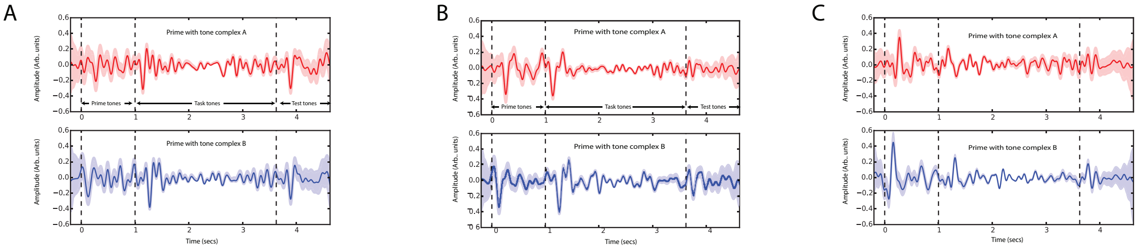

Figure 3.13: **Grand average neural response.** (A) Unique A channel probe tones. Average DSS component when probe tones are unique to tone complex A, under attend to tone complex A (top) and attend to tone complex B (bottom) condition. (B) Unique B channel probe tones. Average DSS component when probe tones are unique to tone complex B, under attend to tone complex A (top) and attend to tone complex B (bottom) condition. (C) Shared probe tones. Average DSS component when probe tones are shared between tone complex A and tone complex B, under attend to tone complex A (top) and attend to tone complex B (bottom) condition.

attention task). In the second task, subjects were asked to report whether they heard oddballs in both the frequencies (global attention task). Thus, the subjects would perform well in the first task if the streams are segregated perceptually while during the global attention task they are at an advantage if the streams are fused perceptually. The neural representation of one of these tone sequences compared across attention and stimulus conditions, revealed that stream segregation is facilitated when the tones alternate, whereas synchronous tones promote the grouping of the tone sequences. These results are consistent with the predictions of the temporal coherence model for stream segregation.

To eliminate stimulus induced differences in the neural response, in the second experiment, response to an isolated probe tone preceded by alternating and synchronous distractor tones was compared, while the subject was attending to the

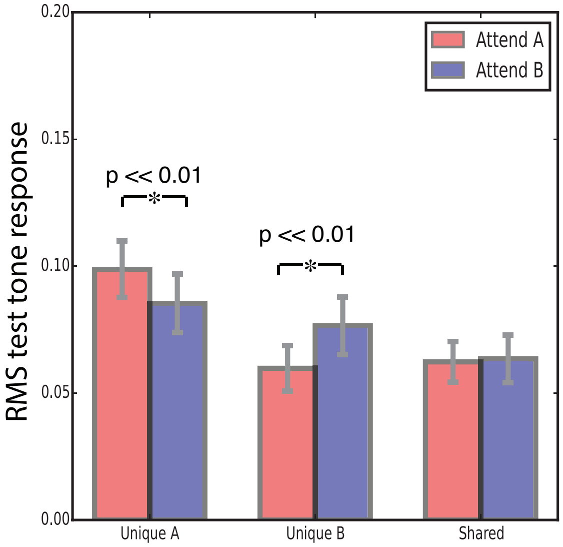

Figure 3.14: **RMS probe tone response.** The root-mean square response strength is compared when the subject is attending to tone complex A vs. tone complex B. The response to the test tones unique to tone complex A is stronger when the subject is attending to tone complex A. Similarly, the response to unique B tones is stronger when subject is attending to complex B than when attending to tone complex A. No significant difference was observed in the response to shared channels when the subject was attending to tone complex A or tone complex B. Results were obtained by averaging data across n = 15 subjects.

entire stimulus (global attention). For the identical probe tone stimulus, the response was stronger when it was preceded by a synchronous sequence, than when it was preceded by an alternating sequence reflecting the facilitative interactions between the tone sequences when they are synchronous. Significant differences between the alternating and synchronous presentation modes were observed even for large frequency separation (15 semi-tones) between the tones. For such large separations, it is unlikely that such an effect can be explained due to peripheral (cochlear) interactions.

While the previous two experiments used alternating or synchronous tone sequences that were anti-coherent and perfectly coherent respectively, in the third experiment, we introduce an additional amplitude modulated tone sequence concurrent with an alternating sequence. Thus, the AM tone sequence is only partially coherent with either tone sequence in the alternating stream. We then investigate modulation in the response to the AM sequence based on the attentive state of the subject. The RMS power in the 40 Hz aSSR showed modulation on a per epoch basis, consistent with the hypothesis of facilitative interactions between the attended sequence(or attended feature) and all features of the auditory scene temporally coherent with it. The findings of this experiment are conforming with results from human studies using speech stimuli [92] that show on individual electrode sites sensitive to a particular high frequency range, the neural response to the same mixture sound in two attention conditions were enhanced only for the target, with responses for similar sounds in the masker speaker suppressed.

The temporal coherence model predicts that interactions between neural populations with coherent activation are facilitative and that these facilitative interactions evolve at a rate dictated by the dynamics of the sound stimulus. These interactions set up an expectation for oncoming sounds and aid their segregation. Previous studies [?, 5] have shown that response to an attended stream is stronger than the response to unattended streams in a mixture. However, it is unclear if individual features that belong to the attended stream are also enhanced. In the final experiment, we use alternating harmonic complexes with overlapping harmonic components and test the response to individual harmonics after the task, using a probe tone. Harmonic components unique to the attended stream remain in an enhanced state at the end of the task and show a stronger response than components that belong to the unattended stream. Even though the subjects were instructed to pay attention to the harmonic complex sequence and had no access to track individual harmonic components, individual features (harmonic components) that belong to the attended sequence also show enhancement supporting the premise of the temporal coherence model. Moreover, the response to the shared components were comparable for both attention conditions, reflecting their contribution to the perception of the attended stream.

The results of these experiments are in agreement with the predictions of the computational model described in the previous chapter - in which temporal coherence between the response of neural populations activated by sounds mediates stream segregation through the formation of facilitative or suppressive connectivity between neural populations.

3.8 Conclusions

The present findings provide insights into the mechanism of temporal coherence for stream segregation. The results establish a strong influence of attention modulated, stimulus driven coherence in the perception of distinct auditory streams. Using simple tone based stimuli we were able to test the temporal coherence model in its most direct form. In the next chapter, we extend the study to include complex natural stimuli such as speech mixtures. While EEG provides a window to observe the overall neural activity of large neural populations, more direct physiological studies are required to probe the existence of coincidence detectors and to test the formation of facilitative/depressive connections.

Chapter 4: Decoding auditory attention to competing speech using EEG spatial patterns.

Recent efforts to decode the attentional focus of human subjects while listening to competing speech stimuli have focussed on exploiting the differences in the temporal structure of the speech sentences. Most efforts have pursued a stimulus reconstruction approach, whereby using the envelope of the clean speech as the output and the time series data on multiple channels as input, a regression model is learnt whose prediction shows a stronger correlation with the attended speech envelope when the model is trained using the attended speech than the unattended speech. In this work, we build a classifier that uses the EEG spatial patterns exclusively, to predict the attentional focus of the subject. The classifier is trained on the EEG data of a group of trials where the subject is paying attention to one of two speakers in a mixture. We then make predictions on the attentional focus of the subject on a separate set of test trials using this classifier. We use common spatial patterns [93] to generate features that can discriminate between the two attentional conditions.

4.1 Introduction

Attention plays a key role in helping humans segregate a target speaker's speech from other competing background speech or noise. Most recent methods to decode the focus of attention of a subject while they listen to competing speech utterances are based on using the changes in cortical activity that track the dynamic temporal modulations in speech [5, 92, 94–98]. The attentional focus is decoded by modeling a single input, multi-output linear system with the speech envelope as input and the measured EEG waveforms as output. It has been shown that attending to one of the speech streams produces a modulation in the measured impulse response around 200 ms after stimulus onset in the left hemisphere [97].

Other MEG studies [5] have shown that there exists separate neural representations for the speech of the two speakers, with each being selectively phase locked to the rhythm of the corresponding speech stream. Another study showed that reconstructed spectrograms from surface ECoG recordings contained spectral and temporal characteristics that resembled the attended speech more than the unattended speech [92]. Another method uses the clean speech waveform and measured neural response to build a statistical model whose parameters are predicted using methods such as the expectation maximization algorithm [99].

Most EEG methods for understanding the neural mechanisms that allow the human brain to solve the cocktail party problem with such remarkable robustness rely on trial averaging. Other methods such as [99,100] rely explicitly on the acoustic waveforms to generate a measure of correlation between the measured EEG waveforms and acoustic stimulus. In some of this work, a repeatable neural response is obtained through some de-noising technique such as DSS and the resulting single waveform is used to build decoders. In this work, the feature extraction stage is combined with the decoding problem to build a model that exploits only the spatial distribution of the EEG response to generate a decoder. Subjects were asked to attend to either a male or female voice in a mixture. Feature vectors were extracted from the measured EEG waveform using the method of common spatial patterns and classified into attend male or attend female conditions using linear discriminant analysis. This method lends itself to BCI applications because of its simplicity and its reliance only on the measured EEG waveforms without utilizing the raw stimulus waveforms.

4.2 Results

Classification accuracy The classification accuracy of the decoder is shown in Figure 4.4 for all subjects. The attentional orientation could be decoded at a mean accuracy of approx. 85%. The accuracy was measured using 4-fold cross-validation across trials.

Spatial distribution of decoder weights Figure 4.2 shows the projection of common spatial patterns on the sensor space for each subject, averaged across trial length. The spatial patterns show a distribution centered at the top of the head and along the temporal areas, consistent with observed spatial patterns in EEG for auditory related tasks [100].

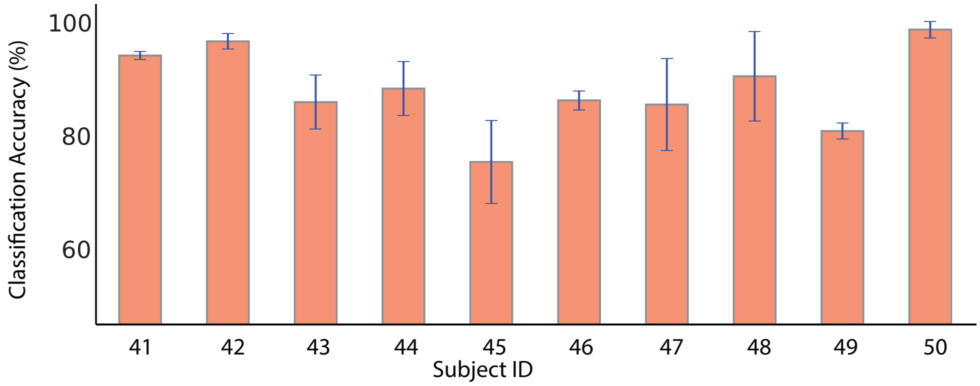

Figure 4.1: **Classification accuracy across individual subjects** .

Decoding attention over time Figure 4.3 depicts the classifier accuracy as a function of time during the trial for one subject. At $t = 0$ the target speech sentence is played. At $t = 0.5$ the distractor speech is introduced. The classification accuracy starts increasing at sound onset and then reduces when the distractor speech is introduced. About 0.5s after distractor speech onset the classification accuracy increases again reflecting the attentional state of the subject.

Correlation of features with attended and unattended speech The extracted common spatial pattern features are correlated with the speech envelope of the two competing speech sentences. Figure 4.4 plots the difference in correlation between female speech and male speech averaged across all subjects. During the trials where the subjects were paying attention to the male speech this difference is negative, while during the attend female trials, this difference is positive. In both

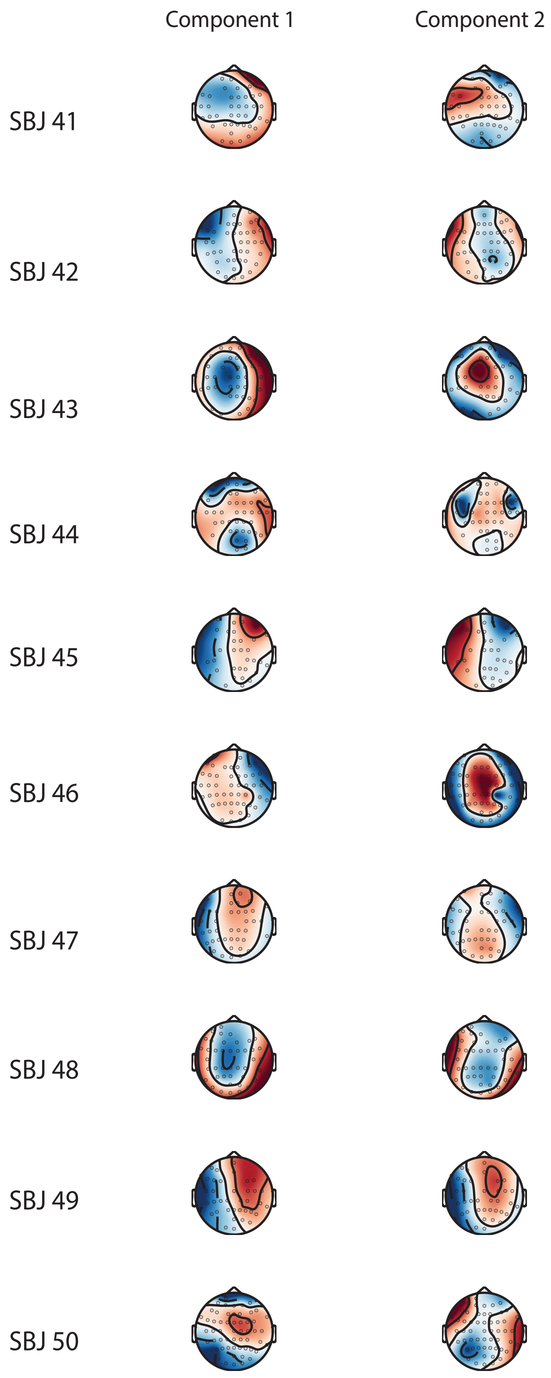

Figure 4.2: **Visualisation of common spatial patterns across subjects .**

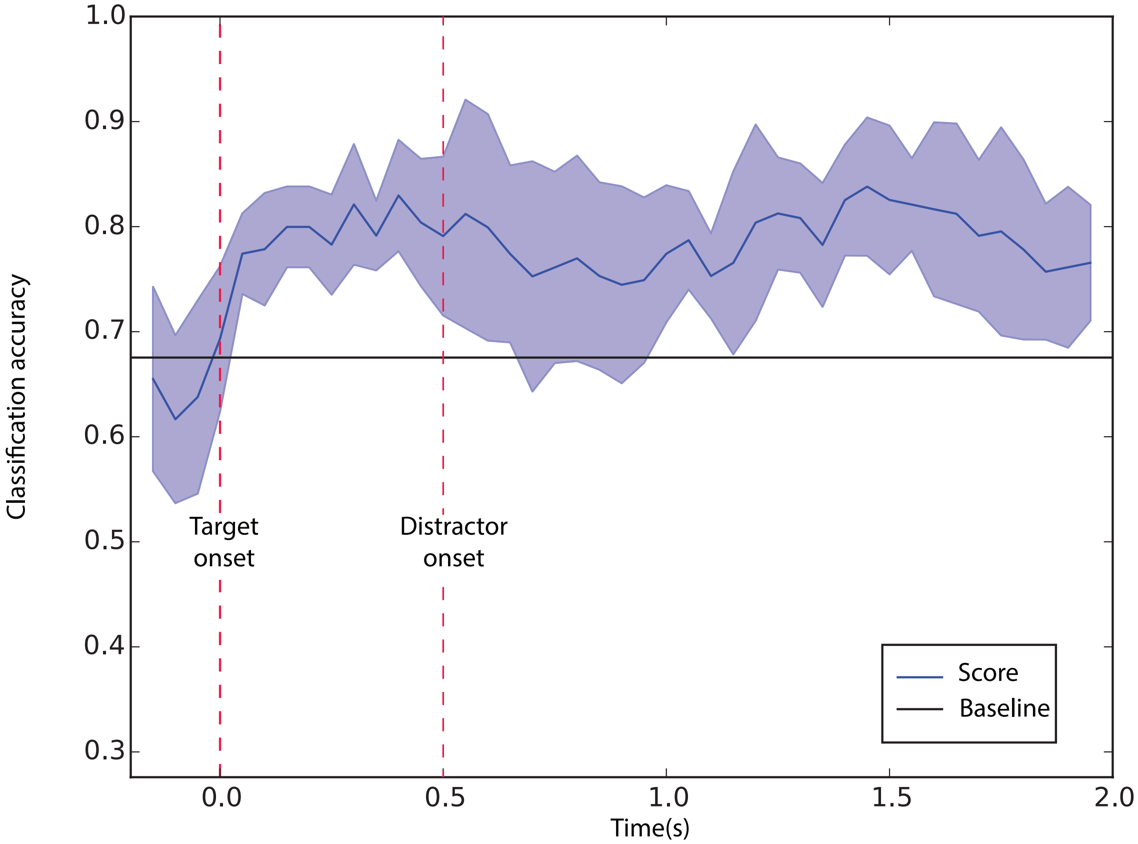

Figure 4.3: **Performance of the decoder over time**. Classifier performance over time for an exemplar subject.

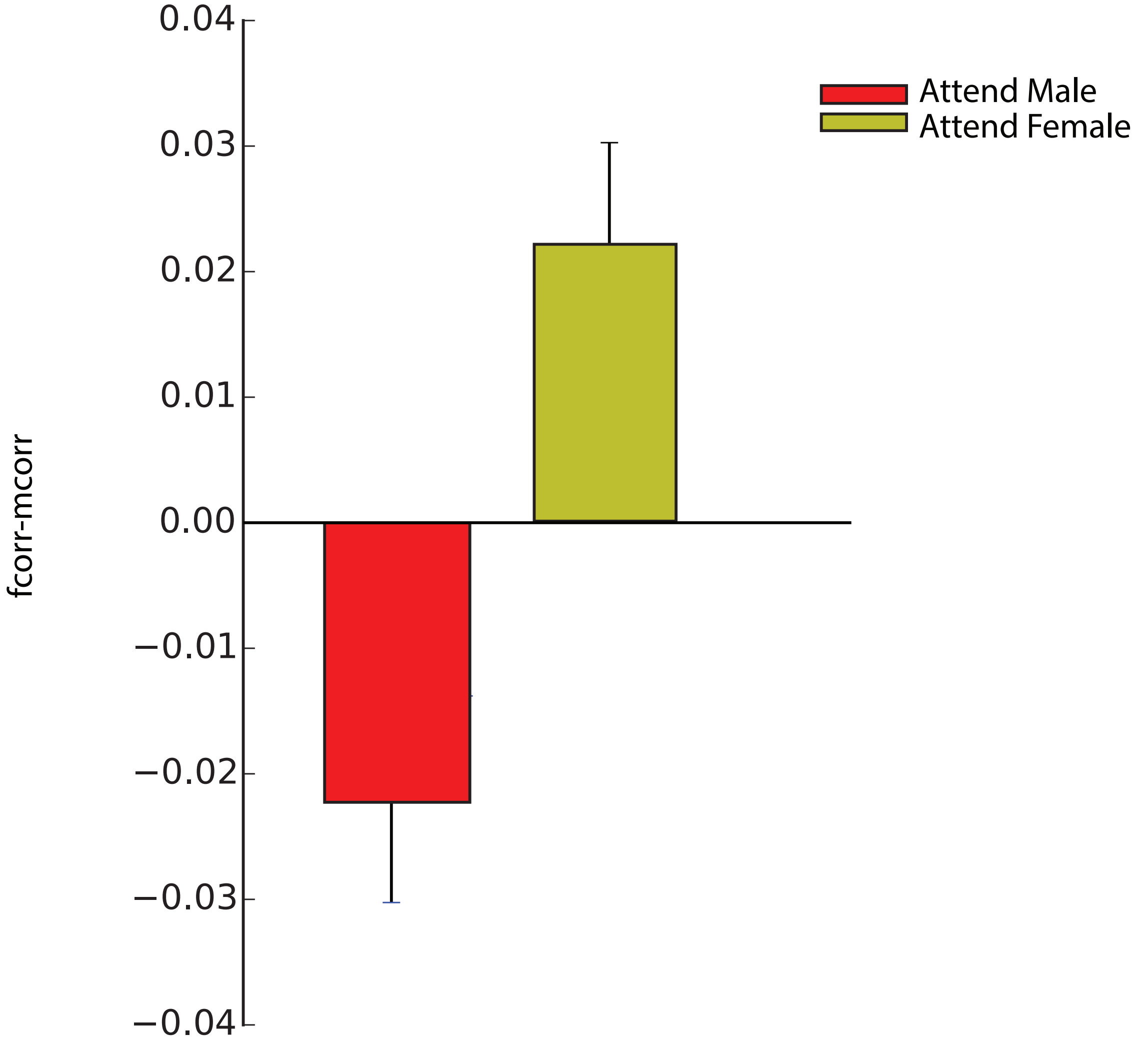

Figure 4.4: **Difference in correlation with the female vs. male speaker - grouped results .**

cases, the magnitude of difference was 0.02, with a $p < 0.05$.

4.3 Discussion

Traditional methods of processing EEG data rely on trial averaging to eliminate the effects of large background neural activity. An auditory evoked response obtained by averaging over several trials is a common analysis technique used to obtain the evoked response to discrete stimulus onset events. [101, 102]. Recently, regression methods have been used to quantify the relationship between the auditory stimulus and the neural response [5, 92, 94, 100]. These methods require access to the clean, un-mixed stimulus waveforms to derive decoders that depend on the second order statistics of the input stimulus, output neural response and their cross-correlations. In this work, we seek to utilize only the information from the spatial sensors to decode auditory attention without relying on any temporal features of the recordings or the stimulus waveforms. We derive spatial filters in sensor space that optimize the classification accuracy of a classifier that predicts whether the subject was paying attention to male or female speech.

Using the method of common spatial patterns for feature extraction and linear discriminant analysis for classification, we achieve an average accuracy of 85% across ten subjects. Projection of the extracted features back in the sensor space shows a distribution that matches expected EEG spatial activation for auditory tasks. This method also provides a mechanism to perform online decoding, since the decoder does not use the temporal information in the stimulus. Although the decoder was built without explicitly relying on the temporal features of the acoustic stimulus, the extracted common spatial pattern features can be correlated with the acoustic waveforms to obtain an additional metric for decoding the trials.

4.4 Methods

4.4.1 Participants

Fifteen human subjects participated in the experiment. All subjects reported normal hearing. Written informed consent was obtained from each participant. The Institutional Review Board of the University of Maryland approved the experimental protocol.

4.4.2 Stimuli and Experimental Setup

The stimulus consisted of speech sentences from the GRID corpus [103]. Speech from a male speaker and a female speaker were mixed at equal power. To facilitate speech segregation, the distractor speaker's speech was introduced after a 500 ms delay from start of each trial. The experiment consisted of 240 trials split into 4 blocks of 60 trials. In the first and third blocks the subject was asked to pay attention to the female voice, while in the second and fourth blocks the subject was asked to pay attention to the male voice. Each sentence in the GRID corpus consisted of a six word sequence of the form *command color preposition letter digit adverb.*

Subjects were asked to report the color and digit combination that the target speaker uttered. EEG data was acquired at a sampling rate of 500Hz using a

BrainVision actiChamp system of 64 electrodes and 4 peripheral sensors.

4.4.3 Data Analysis

Data from faulty sensors were removed from the raw data and then filtered in the range of 1-10 Hz. Eye blink artifacts were removed from the data through regression. The raw data was segmented into epochs. Epochs were sorted based on attend male and attend female conditions. Outlier trials were removed from the data set using a threshold criterion on maximum power across all channels. Data from the two classes were then segregated into training(∼150) and testing(∼50) trials. Using the method of common spatial patterns, feature vectors were learnt from the training trials to maximize classification accuracy. The learnt features were then classified using linear discriminant analysis. All the data analysis was done in python using mne-python [90].

Decoding in sensor space using common spatial patterns

Common spatial patterns [93] is a linear transformation technique that transforms the data into a basis that maximally separates two classes.

The EEG data is represented by a $N \times T$ matrix X, where N is the number of electrodes and T is the length of time samples in each epoch. For each epoch, an estimate of the spatial covariance in the EEG during the epoch is obtained as:

$$R = XX^T/tr(XX^T) \tag{4.1}$$

The normalization in equation 4.1 eliminates magnitude variations in the EEG between epochs. The covariance matrix was then averaged over epochs per condition to obtain C_a and C_b.

$$\begin{aligned} C_a &= \textstyle\sum_{i \in class_a} R_i \\ C_b &= \textstyle\sum_{i \in class_b} R_i \end{aligned} \tag{4.2}$$

To determine if the two classes possess spatial patterns that allow to be discriminated, a composite covariance matrix is constructed.

$$C = 0.5 * (C_a + C_b) \tag{4.3}$$

The matrix C is then factored as:

$$C = U \lambda_c U^T \tag{4.4}$$

The covariance matrices C_a and C_b are then whitened to form S_a and S_b using P formed as:

$$P = \sqrt{\frac{1}{\lambda_c}} U^T \tag{4.5}$$

$$\begin{aligned} S_a &= P C_a P^T \\ S_b &= P C_b P^T \end{aligned} \tag{4.6}$$

S_a and S_b share the same eigenvectors and the sum of the corresponding eigenvalues will be 1[[104]].

S_a and S_b can be decomposed as:

$$\begin{aligned} S_a &= D \psi_a D^T \\ S_b &= D \psi_b D^T \end{aligned} \tag{4.7}$$

where $\psi_a + \psi_b = I$

Since the eigenvectors which span a measurement space are known to be optimal in the least squares sense, for the amount of variance in the measurements they can account for, projection of whitened EEG epochs on D^T will yield feature vectors which are optimal for discriminating between two populations. Thus, in order to discriminate two classes based on their spatial features, each epoch in the EEG is first normalized the trace of its own covariance matrix.

$$X_{norm} = X/tr(XX^T) \tag{4.8}$$

The normalized data is then whitened using P obtained from the composite covariance matrix and then projected along the eigenvectors of its own whitened population covariance. The overall transformation applied to the data is:

$$\bar{X} = D^T P X_{norm} \tag{4.9}$$

4.5 Conclusions

In this work, we have implemented a fast binary classifier to decode the attentional focus of human subjects while they listen to competing speakers. While being fast and computationally simple, this method lends itself to online decoding of individual trials, since it only exploits the differences in the spatial signatures of neural activation between the two conditions, without relying on the acoustic waveform for guiding the decoding. However, the learnt features can be correlated with the acoustic waveforms to obtain additional metrics for classifying each trial. This method can have potential applications in fast BCI applications.

Chapter 5: Conclusions and Future Work.

In summary, we have described a model for segregating complex sound mixtures based on the temporal coherence principle. The model computes the coincidence of multi-scale cortical features and clusters the coherent responses as emanating from one source. It requires no prior information, statistics, or knowledge of source properties, but can gracefully incorporate them along with cognitive influences such as attention to, or memory of specific attributes of a target source to segregate it from its background. The model provides a testable framework of the physiological bases and psychophysical manifestations of this remarkable ability.

Future work can aim to exploit the power of training multiple hidden layers in the auto-encoder architecture. Our computational algorithm lends itself to easy extension into a deep neural network framework. Better signal reconstruction techniques can be employed to improve the quality of reconstruction.

The findings from EEG experiments provide insights into the mechanism of temporal coherence for stream segregation. The results establish a strong influence of attention modulated, stimulus driven coherence in the perception of distinct auditory streams. Using simple tone based stimuli we were able to test the temporal coherence model in its most direct form. We have also implemented a fast binary classifier to decode the attentional focus of human subjects while they listen to competing speakers. The sensor space decoder exploits the second order statistics of the sensors. While being fast and computationally simple, this method lends itself to online decoding of individual trials, since it only exploits the differences in the spatial signatures of neural activation between the two conditions, without relying on the acoustic waveform for guiding the decoding. However, the learnt features can be correlated with the acoustic waveforms to obtain additional metrics for classifying each trial. This method can have potential applications in fast BCI applications.

Although the results from our EEG experiments strongly support the temporal coherence hypothesis for auditory source segregation, more direct physiological experiments are required to further confirm the exact nature of these neural computations, how correlation can be represented through spiking activity and where these computations are performed.

Bibliography

[1] Istvn Winkler, Susan L Denham, and Israel Nelken. Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in cognitive sciences*, 13(12):532–540, December 2009. PMID: 19828357.

[2] Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

[3] Timothy D. Griffiths and Jason D. Warren. What is an auditory object? *Nature Reviews Neuroscience*, 5(11):887–892, November 2004.

[4] Barbara G. Shinn-Cunningham. Object-based auditory and visual attention. *Trends in Cognitive Sciences*, 12(5):182–186, May 2008.

[5] Nai Ding and Jonathan Z Simon. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29):11854–11859, 2012.

[6] R M Warren, C J Obusek, and J M Ackroff. Auditory induction: perceptual synthesis of absent sounds. *Science (New York, N.Y.)*, 176(4039):1149–1151, June 1972. PMID: 5035477.

[7] A. Bendixen, E. Schroger, and I. Winkler. I heard that coming: Event-related potential evidence for stimulus-driven prediction in the auditory system. *Journal of Neuroscience*, 29(26):8447–8451, July 2009.

[8] Christopher I. Petkov, Kevin N. O'Connor, and Mitchell L. Sutter. Encoding of illusory continuity in primary auditory cortex. *Neuron*, 54(1):153–165, April 2007. PMID: 17408584 PMCID: PMC2628590.

[9] Christophe Micheyl, Robert P. Carlyon, Yury Shtyrov, Olaf Hauk, Tara Dodson, and Friedemann Pullvermller. The neurophysiological basis of the auditory continuity illusion: a mismatch negativity study. *Journal of Cognitive Neuroscience*, 15(5):747758, 2003.

[10] Jennifer K. Bizley and Yale E. Cohen. The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, 14(10):693–707, September 2013.

[11] DeLiang Wang and Brown,Guy. *Computational auditory scene analysis: principles, algorithms, and applications.* IEEE Press ; Wiley Interscience, Piscataway, N.J. : Hoboken, N.J, 2006.

[12] S.J. Rennie, J.R. Hershey, and P.A. Olsen. Efficient model-based speech separation and denoising using non-negative subspace analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, pages 1833–1836, 2008.

[13] M.I. Mandel, R.J. Weiss, and D.P.W. Ellis. Model-based expectation-maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):382–394, 2010.

[14] Michael W. Beauvois and Ray Meddis. Computer simulation of auditory stream segregation in alternating-tone sequences. *The Journal of the Acoustical Society of America*, 99:2270, 1996.

[15] I. Winkler, S. Denham, R. Mill, T. M. Bohm, and A. Bendixen. Multistability in auditory stream segregation: a predictive coding view. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1591):1001–1012, February 2012.

[16] N. Ding and J. Z. Simon. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience*, 33(13):5728–5735, March 2013.

[17] Nai Ding and Jonathan Z. Simon. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the United States of America*, 109(29):11854–11859, July 2012. PMID: 22753470 PMCID: PMC3406818.

[18] Nima Mesgarani and Edward F. Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236, April 2012.

[19] ElanaM. ZionGolumbic, Nai Ding, Stephan Bickel, Peter Lakatos, CatherineA. Schevon, GuyM. McKhann, RobertR. Goodman, Ronald Emerson, AsheshD. Mehta, JonathanZ. Simon, David Poeppel, and CharlesE. Schroeder. Mechanisms underlying selective neuronal tracking of attended speech at a Cocktail party. *Neuron*, 77(5):980–991, March 2013.

[20] J. Xiang, J. Simon, and M. Elhilali. Competing streams at the cocktail party: Exploring the mechanisms of attention and temporal integration. *Journal of Neuroscience*, 30(36):12084–12093, September 2010.

[21] Christophe Micheyl, Coral Hanson, Laurent Demany, Shihab Shamma, and Andrew J Oxenham. Auditory stream segregation for alternating and synchronous tones. *Journal of experimental psychology. Human perception and performance*, April 2013. PMID: 23544676.

[22] E. Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5):975–979, June 1953.

[23] Mark A. Bee and Christophe Micheyl. The "Cocktail party problem": What is it? how can it be solved? and why should animal behaviorists study it? *Journal of comparative psychology*, 122(3):235–251, 2008.

[24] Nandini C. Singh and Frdric E. Theunissen. Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, 114(6):3394–3411, December 2003.

[25] John M. Henderson, Myriam Chanceaux, and Tim J. Smith. The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, 9(1):32–40, 2009.

[26] Gareth Jones. Sensory biology: Listening in the dark for echoes from silent and stationary prey. *Current Biology*, 23(6):R249–R251, 2013.

[27] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, and R. Gopinath. Superhuman multi-talker speech recognition: The IBM 2006 speech separation challenge system. In *in ICSLP*, pages 97–100, 2006.

[28] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications.* Academic Press, 2010.

[29] Paris Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In Carlos G. Puntonet and Alberto Prieto, editors, *Independent Component Analysis and Blind Signal Separation*, number 3195 in Lecture Notes in Computer Science, pages 494–499. Springer Berlin Heidelberg, January 2004.

[30] Daniel P. W. Ellis. Model-based scene analysis. In *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, pages 115–146. Wiley/IEEE Press, 2006.

[31] B. King and L. Atlas. Single-channel source separation using simplified-training complex matrix factorization. In *2010 IEEE International Conference on Acoustics Speech and Signal Processing*, pages 4206–4209, 2010.

[32] Martin Cooke, John R. Hershey, and Steven J. Rennie. Monaural speech separation and recognition challenge. *Computer Speech & Language*, 24(1):1–15, 2010.

[33] Guy J. Brown. Physiological models of auditory scene analysis. In Ray Meddis, Enrique A. Lopez-Poveda, Richard R. Fay, and Arthur N. Popper, editors, *Computational Models of the Auditory System*, number 35 in Springer Handbook of Auditory Research, pages 203–236. Springer US, January 2010.

[34] Yang Shao and DeLiang Wang. Sequential organization of speech in computational auditory scene analysis. *Speech Communication*, 51(8):657–667, 2009.

[35] William Morris Hartmann and Douglas Johnson. Stream segregation and peripheral channeling. *Music Perception: An Interdisciplinary Journal*, 9(2):155–183, 1991.

[36] Susan L. McCabe and Michael J. Denham. A model of auditory streaming. *The Journal of the Acoustical Society of America*, 101(3):1611–1621, March 1997.

[37] M. Stark, M. Wohlmayr, and F. Pernkopf. Source-filter-based single-channel speech separation using pitch information. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):242–255, 2011.

[38] Guoning Hu and DeLiang Wang. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2067–2079, 2010.

[39] P. Clark and L.E. Atlas. Time-frequency coherent modulation filtering of nonstationary signals. *IEEE Transactions on Signal Processing*, 57(11):4323–4332, 2009.

[40] R. Mill, T. Bohm, A. Bendixen, I. Winkler, and S.L. Denham. CHAINS: competition and cooperation between fragmentary event predictors in a model of auditory scene analysis. In *2011 45th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, 2011.

[41] Jean-Michel Hupe and Daniel Pressnitzer. The initial phase of auditory and visual scene analysis. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1591):942–953, April 2012.

[42] Christoph Von Der Malsburg. The correlation theory of brain function. *Models of neural networks*, 2:95119, 1994.

[43] DeLiang Wang, Joachim Buhmann, and Christoph von der Malsburg. Pattern segmentation in associative memory. *Neural Computation*, 2(1):94–106, 1990.

[44] Shihab A. Shamma, Mounya Elhilali, and Christophe Micheyl. Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, 34(3):114–123, March 2011.

[45] Mounya Elhilali, Ling Ma, Christophe Micheyl, Andrew J Oxenham, and Shihab A Shamma. Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron*, 61(2):317–329, January 2009.

[46] Christophe Micheyl, Cynthia Hunter, and Andrew J Oxenham. Auditory stream segregation and the perception of across-frequency synchrony. *Journal of experimental psychology. Human perception and performance*, 36(4):1029–1039, 2010.

[47] Christophe Micheyl and Andrew J Oxenham. Pitch, harmonicity and concurrent sound segregation: psychoacoustical and neurophysiological findings. *Hearing research*, 266(1-2):36–51, July 2010.

[48] Sundeep Teki, Maria Chait, Sukhbinder Kumar, Shihab Shamma, and Timothy D. Griffiths. Segregation of complex acoustic scenes based on temporal coherence. *eLife*, 2, July 2013.

[49] DeLiang Wang. Primitive auditory segregation based on oscillatory correlation. *Cognitive Science*, 20(3):409–456, July 1996.

[50] Edward W. Large and Mari Riess Jones. The dynamics of attending: How people track time-varying events. *Psychological Review*, 106(1):119–159, 1999.

[51] D.L. Wang and Guy J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks*, 10(3):684–697, May 1999.

[52] S.N. Wrigley and Guy J. Brown. A computational model of auditory selective attention. *IEEE Transactions on Neural Networks*, 15(5):1151–1163, September 2004.

[53] Felix Almonte, Viktor K. Jirsa, Edward W. Large, and Betty Tuller. Integration and segregation in auditory streaming. *Physica D: Nonlinear Phenomena*, 212(1):137–159, December 2005.

[54] Taishih Chi, Powen Ru, and Shihab A. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, August 2005.

[55] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, February 1991.

[56] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

[57] Neal F. Viemeister, Mark A. Stellmack, and Andrew J. Byrne. The role of temporal structure in envelope processing. In Daniel Pressnitzer, Alain de Cheveign, Stephen McAdams, and Lionel Collet, editors, *Auditory Signal Processing*, pages 220–228. Springer New York, January 2005.

[58] Chandramouli Chandrasekaran, Andrea Trubanova, Sbastien Stillittano, Alice Caplier, and Asif A. Ghazanfar. The natural statistics of audiovisual speech. *PLoS Comput Biol*, 5(7):e1000436, July 2009.

[59] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2000.

[60] C von der Malsburg and W Schneider. A neural cocktail-party processor. *Biological cybernetics*, 54(1):29–40, 1986.

[61] S.M. Schimmel, L.E. Atlas, and Kaibao Nie. Feasibility of single channel speaker separation based on modulation frequency analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*, volume 4, pages 605–608, 2007.

[62] S Shamma and D Klein. The case of the missing pitch templates: how harmonic templates emerge in the early auditory system. *The Journal of the Acoustical Society of America*, 107(5 Pt 1):2631–2644, May 2000.

[63] Brian C. J Moore. *An introduction to the psychology of hearing.* Academic Press, Amsterdam; Boston, 2003.

[64] Simon Krogholt Christiansen, Morten Lve Jepsen, and Torsten Dau. Effects of tonotopicity, adaptation, modulation tuning, and temporal coherence in primitive auditory stream segregationa). *The Journal of the Acoustical Society of America*, 135(1):323–333, January 2014.

[65] Shihab Shamma, Mounya Elhilali, Ling Ma, Christophe Micheyl, Andrew J Oxenham, Daniel Pressnitzer, Pingbo Yin, and Yanbo Xu. Temporal coherence and the streaming of complex sounds. *Advances in experimental medicine and biology*, 787:535–543, 2013.

[66] Terrence J Sejnowski and Gerald Tesauro. The hebb rule for synaptic plasticity: algorithms and implementations. In *Neural models of plasticity: Experimental and theoretical approaches*, pages 94–103. Academic Press, New York, 1989.

[67] L. F. Abbott and Sacha B. Nelson. Synaptic plasticity: taming the beast. *Nature Neuroscience*, 3:1178–1183, 2000.

[68] Mounya Elhilali and Shihab A. Shamma. A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*, 124(6):3751, 2008.

[69] Richard O Duda and Peter E Hart. *Pattern classification and scene analysis.* Wiley, New York, 1973.

[70] K. Wang and S. Shamma. Self-normalization and noise-robustness in early auditory representations. *IEEE Transactions on Speech and Audio Processing*, 2(3):421–435, 1994.

[71] Mark Schmidt. minFunc - unconstrained differentiable multivariate optimization in matlab, 2012.

[72] Randolph Blake and Sang-Hun Lee. The role of temporal structure in human vision. *Behavioral and Cognitive Neuroscience Reviews*, 4(1):21–42, March 2005.

[73] David Alais, Randolph Blake, and Sang-Hun Lee. Visual features that vary together over time group together over space. *Nature neuroscience*, 1(2):160164, 1998.

[74] Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1990.

[75] Brian CJ Moore and Hedwig Gockel. Factors influencing sequential stream segregation. *Acta Acustica United with Acustica*, 88(3):320–333, 2002.

[76] Robert P Carlyon and Hedwig E Gockel. Effects of harmonicity and regularity on the perception of sound sources. In *Auditory perception of sound sources*, pages 191–213. Springer, 2008.

[77] Mounya Elhilali, Ling Ma, Christophe Micheyl, Andrew J Oxenham, and Shihab A Shamma. Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron*, 61(2):317–329, 2009.

[78] Shihab A Shamma, Mounya Elhilali, and Christophe Micheyl. Temporal coherence and attention in auditory scene analysis. *Trends in neurosciences*, 34(3):114–123, 2011.

[79] Christophe Micheyl, Cynthia Hunter, and Andrew J Oxenham. Auditory stream segregation and the perception of across-frequency synchrony. *Journal of Experimental Psychology: Human Perception and Performance*, 36(4):1029, 2010.

[80] Christophe Micheyl, Coral Hanson, Laurent Demany, Shihab Shamma, and Andrew J Oxenham. Auditory stream segregation for alternating and synchronous tones. *Journal of Experimental Psychology: Human Perception and Performance*, 39(6):1568, 2013.

[81] Christophe Micheyl, Heather Kreft, Shihab Shamma, and Andrew J Oxenham. Temporal coherence versus harmonicity in auditory stream formation. *The Journal of the Acoustical Society of America*, 133(3):EL188–EL194, 2013.

[82] Sundeep Teki, Maria Chait, Sukhbinder Kumar, Shihab Shamma, and Timothy D Griffiths. Segregation of complex acoustic scenes based on temporal coherence. *Elife*, 2:e00699, 2013.

[83] Simon Krogholt Christiansen and Andrew J. Oxenham. Assessing the effects of temporal coherence on auditory stream formation through comodulation masking release. *The Journal of the Acoustical Society of America*, 135(6):3520–3529, 2014.

[84] Lakshmi Krishnan, Mounya Elhilali, and Shihab Shamma. Segregating complex sound sources through temporal coherence. *PLoS computational biology*, 10(12):e1003985, 2014.

[85] Jonathan Fritz, Shihab Shamma, Mounya Elhilali, and David Klein. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature neuroscience*, 6(11):1216–1223, 2003.

[86] Daniel B Polley, Elizabeth E Steinberg, and Michael M Merzenich. Perceptual learning directs auditory cortical map reorganization through top-down influences. *The journal of neuroscience*, 26(18):4970–4982, 2006.

[87] David H Brainard. The psychophysics toolbox. *Spatial vision*, 10:433–436, 1997.

[88] Alain de Cheveigne and Jonathan Z Simon. Denoising based on spatial filtering. *Journal of neuroscience methods*, 171(2):331–339, 2008.

[89] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

[90] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 7, 2013.

[91] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen. Mne software for processing meg and eeg data. *Neuroimage*, 86:446–460, 2014.

[92] Nima Mesgarani and Edward F Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236, 2012.

[93] Zoltan Joseph Koles. The quantitative extraction and topographic mapping of the abnormal components in the clinical eeg. *Electroencephalography and clinical Neurophysiology*, 79(6):440–447, 1991.

[94] Jess R Kerlin, Antoine J Shahin, and Lee M Miller. Attentional gain control of ongoing cortical speech representations in a ?cocktail party? *The Journal of Neuroscience*, 30(2):620–628, 2010.

[95] Miika Koskinen, Jaakko Viinikanoja, Mikko Kurimo, Arto Klami, Samuel Kaski, and Riitta Hari. Identifying fragments of natural speech from the listener's meg signals. *Human brain mapping*, 34(6):1477–1489, 2013.

[96] Edmund C Lalor and John J Foxe. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European journal of neuroscience*, 31(1):189–193, 2010.

[97] Alan J Power, John J Foxe, Emma-Jane Forde, Richard B Reilly, and Edmund C Lalor. At what time is the cocktail party? a late locus of selective attention to natural speech. *European Journal of Neuroscience*, 35(9):1497–1503, 2012.

[98] Elana M Zion Golumbic, Nai Ding, Stephan Bickel, Peter Lakatos, Catherine A Schevon, Guy M McKhann, Robert R Goodman, Ronald Emerson, Ashesh D Mehta, Jonathan Z Simon, et al. Mechanisms underlying selective neuronal tracking of attended speech at a ?cocktail party? *Neuron*, 77(5):980–991, 2013.

[99] Sahar Akram, Jonathan Z Simon, Shihab A Shamma, and Behtash Babadi. A state-space model for decoding auditory attentional modulation from meg in a competing-speaker environment. In *Advances in Neural Information Processing Systems*, pages 460–468, 2014.

[100] James A O'Sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral Cortex*, page bht355, 2014.

[101] Terry W Picton, Steven A Hillyard, Howard I Krausz, and Robert Galambos. Human auditory evoked potentials. i: Evaluation of components. *Electroencephalography and clinical neurophysiology*, 36:179–190, 1974.

[102] TW Picton and SA Hillyard. Human auditory evoked potentials. ii. effects of attention. *Electroencephalography and clinical neurophysiology*, 36(2):191, 1974.

[103] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[104] Keinosuke Fukunaga. *Introduction to statistical pattern recognition.* Academic Press Professional, Inc., 1990.