ABSTRACT

Title of Document: The Cortical Representations of Speech in Reverberant Conditions

Marisel Villafañe-Delgado, Master of Science, 2013

Directed By: Professor Jonathan Z. Simon, Department of Electrical and Computer Engineering

Speech intelligibility in adverse situations, such as reverberation and noise, is conserved until the degradations reach certain thresholds. Psychoacoustic studies have described the properties of speech that lead to the conservation of its intelligibility under those circumstances. The neural mechanisms that underlie the robustness of intelligibility in these situations, however, are not yet well understood. Here, the cortical representations of speech in reverberation and speech plus noise in reverberation are studied by measuring the cortical responses of human subjects using magnetoencephalography (MEG) while they listened to continuous speech narratives. It was hypothesized that the neural processing of speech in reverberation and speech plus noise in reverberation would follow a lack of cortical synchronization as function of the degradations. Encoding models show, however, that the neural encoding of speech in reverberation follow a different mechanism than that of speech in noise. On the other hand, in the absence of noise, it is possible to reconstruct with high accuracy the envelope of reverberant speech, thus demonstrating that the reverberant speech is well encoded by the brain.

THE CORTICAL REPRESENTATIONS OF SPEECH IN REVERBERANT CONDITIONS

By

Marisel Villafañe-Delgado

Thesis 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
Master of Science
2013

Advisory Committee:
Professor Jonathan Z. Simon, Chair
Professor Carol Espy-Wilson
Professor Daniel A. Butts

Dedication

To those who suffer my absence and deal with my presence…

Acknowledgements

Firstly I want to say thanks to my advisor, Professor Jonathan Z. Simon. I would have not completed this project without his support, wisdom, and encouragement. Thanks for giving me the privilege of working in your laboratory and helping me along in this wonderful field. I also want to say thanks to current and past members of the Computational Sensorymotor Systems Laboratory, Francisco Cervantes Constantino, Krishna Puvvada and Nai Ding for their valuable opinions and discussions. In the Electrical and Computer Engineering Department I want to say thanks to Melanie Prange for all the support when I needed most and getting me out of trouble so many times.

To my friends that in any or other way have helped me around either with food, couch surfing, or time: Teresa and Nelson, Miguel, Sol, WingYee, Leonard, Krishna and Francisco. Also, people that are far physically but are always here: Nayda Santiago and Jaime A. Arbona-Fazzi. Thanks for always being there.

Last but not least, my beloved family. Thanks for the patience and comprehension.

Table of Contents

List of Figures

Chapter 1

Background

1.1 Speech in Adverse Conditions

Everyday listeners are challenged by interferences in the environment that degrade speech before it reaches them. Common examples of distortion are background noise from electromechanical equipment or other speakers, and reverberant reflections of the speech from the boundaries of a room. Regardless these acoustic interferences, under certain circumstances listeners are able to understand the degraded speech. From psychoacoustic studies it is well known which properties of speech are most susceptible to degradation of intelligibility. The neural mechanisms involved in the robustness of speech intelligibility in adverse conditions, however, are not yet well understood. Understanding the neural mechanisms involved in the robustness of the speech intelligibility in these situations is crucial, in particular for the development of aids for people with listening disabilities and for the enhancement of artificial speech recognition and speaker identification systems. This thesis emphasizes on the study of the cortical representations of speech in reverberation and speech distorted by noise in reverberant environments. Furthermore, taking into account reverb and noise together is indispensable since it reflects, in a more realistic manner, the daily experiences of the listeners.

Speech can be described in terms of its main temporal features: the envelope, periodicity and fine-structure (Rosen, 1992). Robustness of the speech intelligibility has been related to the integrity of the signal's slow temporal components, which constitute the speech envelope (Houtgast & Steeneken, 1985; Rosen, 1992). Essentially, the slow temporal modulations denote the syllabic rate of the speech (Greenberg, Carvey, Hitchcock, & Chang, 2003; Rosen, 1992). The modulation spectrum, or the Fourier transform of the speech envelope, serves to characterize the intelligibility of speech. Specifically, speech intelligibility is sustained when the critical bands from 1-7 Hz of the modulation spectrum are not severely degraded (Elliott & Theunissen, 2009).

1.2 Reverberation

In a room, speech can be severely degraded when it is mixed with its reflections coming from the room's boundaries. Thus, the message arriving to the listener is the result of the direct speech in addition to its often-undesired reflections (Figure 1). In the time domain reflections fill the temporal gaps of the speech envelope, increasing the energy of low-frequency components. Figure 2(a) illustrates this effect and how the clean speech can be distorted in a severe reverberant scenario. This increases the difficulty to segregate words and syllables (Drullman, Festen, & Plomp, 1994). In terms of the time-frequency domain reverb causes a spectral blur degrading the formant's transitions, as seen in Figure 2(b). In reverberation, the peak of the modulation spectrum is shifted from 5 Hz to the low frequency bands, around 1-2 Hz (Greenberg et al., 2003). In addition, reverberation distorts interaural time differences and interaural level differences (Shinn-

Cunningham, 2002), distorting not only the information in the message but also its directionality.

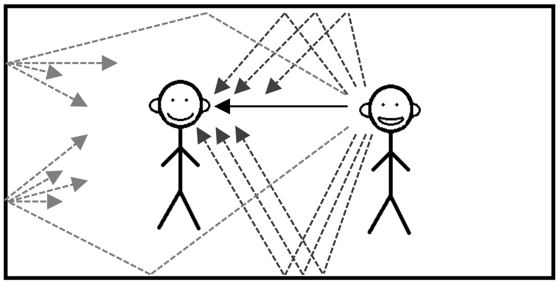
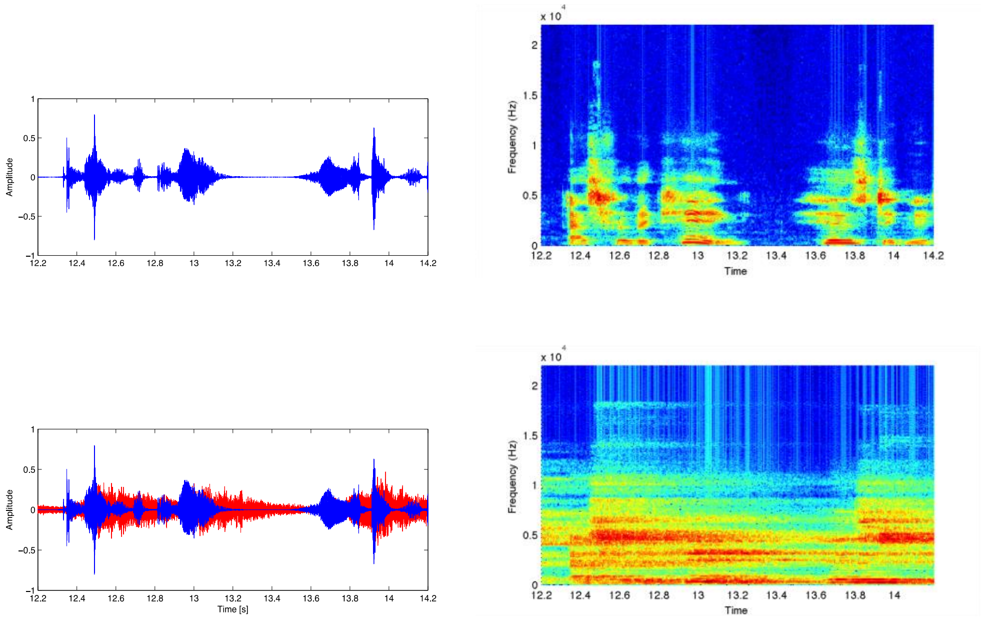


Figure 1. Room reflections with speaker and listener. Solid arrow represents the direct signal, dashed arrows represent the early reflections and the light gray (also dashed) represents the diffuse late reflections.



(a) (b)

Figure 2. Effects of severe reverberation in the time and time-frequency domain. a) Upper figure shows two seconds of clean speech. Lower figure illustrates how reverb distorts the speech envelope when energy spreads after offsets, distorting thus the onsets of subsequent segments. Blue signal corresponds to the clean speech, red signal is speech degraded by reverberation. b) In time-frequency domain the effects of reverberation are seen as spectro-temporal smear.

Reverberation is attributed to two types of masking: overlap masking and self-masking (Nábělek, Letowski, & Tucker, 1989). Overlap masking occurs when a preceding phoneme overlaps with a subsequent segment (Arai, Murakami, Hayashi, Hodoshima, & Kurisu, 2007). This effect is worsened when the preceding segment ends in a vowel and the subsequent segment starts with a consonant, because vowels tend to contain higher energy than consonants (Arai et al., 2007). On the other hand, self-masking refers to cues within consonants that have time-varying characteristics (Watkins, 2005). In terms of speech intelligibility, (Greenberg, 2006) determined that syllables contain three main components: onset, nucleus, coda and from these, onsets are considered more informative than codas since the later can be lost without affecting intelligibility. Since onsets are more relevant for intelligibility than codas, overlap masking has significant consequences in the speech intelligibility. By the nature of the masking types characterizing reverberation, maintaining the speech intelligibility in reverberant conditions might involve different mechanisms than those required when extracting the target speech from noise or multiple speakers. In this regard, speech in noise and simultaneous speakers is affected by energetic masking and informational masking, respectively.

Reverberation introduced by a room is characterized by its room impulse response (RIR). When the room is viewed as a linear system the RIR characterizes the reverberations between source(s) and receiver(s) located in a room. Thus, sounds can be simulated as played in a room with particular acoustic characteristics by the linear convolution of the sound with the RIR. The reverberant impulse response (Figure 3) contains three main components: direct response, the early reflections and the late reflections. The early reflections, occurring in the first 80-100 ms (Gold, Morgan, & Ellis, 2011), are sparse and contribute positively to the speech intelligibility. Late reflections, in the other hand, degrade the speech intelligibility (Haas, 1972; Nabelek & Pickett, 1974).

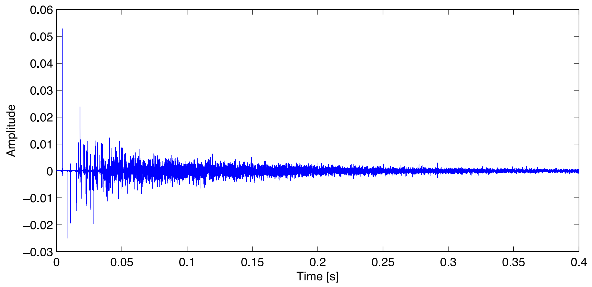


Figure 3. Room impulse response in intermediate reverberation

As time goes by, reflections are absorbed by the room’s surfaces. Thus, surfaces composed of less absorptive materials lead to higher reverberation and longer RIRs. The reverberation time, expressed in seconds, is denoted as RT_{60} and provides a measure of how long does it take for the reflections‘ energy to decay 60 dB below its original level. Comfortable reverberation times vary per room, since it is dependent on the room’s surface materials, size, and localization of the speaker(s) and receiver(s). It is worth mentioning that the RT_{60} should not be taken as a measure for speech intelligibility since two rooms with the same RT_{60} may lead to different intelligibility (Arai et al., 2007). The most popular equation for the RT_{60} was provided by Sabine (Sabine, 1922) and is given by

$$RT_{60} = \frac{4 \ln 10^6}{c} \cdot \frac{V}{Sa}$$

where c is the speed of sound in the room in m/s, V is the volume of the room in m^3, S is the total surface area of the room in m^3, and a is the average absorption coefficient of the room surfaces.

From psychoacoustics, one proposed explanation for the mechanisms involved in the robustness of speech intelligibility in reverberant environments is based on perceptual compensation (Watkins & Makin, 2007; Watkins, 2005). Under this idea it is hypothesized that the additional energy at spectral transitions provide valuable information to the compensation mechanism about how much energy in the signal belongs to reflections (Watkins, 2005). Neurophysiological studies in reverberation have mostly focused in the study of sound localization. Studies in animals have reported directional sensitivity of single neurons in the auditory midbrain (Devore & Delgutte, 2010; Devore, Ihlefeld, Hancock, Shinn-Cunningham, & Delgutte, 2009). In humans, magnetoencephalography (MEG) studies while subjects listen to noise have shown hemispheric lateralization and directional tuning to sound localization (Palomäki, Tiitinen, Mäkinen, May, & Alku, 2005) and a neural code formed by two groups of neural populations corresponding to each hemifield (Salminen, Tiitinen, Yrttiaho, & May, 2010). This study is emphasized in the neural mechanisms of continuous speech reverberation rather than the processing of localization cues.

Here it is hypothesized that reverberation causes the loss of synchronization to the slow temporal modulations of speech as function of the degradation. Furthermore, it is expected to observe that in the presence of reverberation the neural processing of speech in noise is worsen.

1.3 Spectrally matched noise

Spectrally matched noise provides maximum acoustical overlap with speech, reduces its intensity contrast and distorts its spectro-temporal modulations (Ding & Simon, 2013). The modulation spectrum is attenuated by noise uniformly. Insensitivity to the intensity contrast of speech is suggested by psychoacoustic studies as a mechanism for the robustness of speech intelligibility (Stone, Füllgrabe, Mackinnon, & Moore, 2011). The theories for noise-robust speech encoding rely on the stable neural synchronization to the speech envelope, insensitivity to the intensity contrast of speech and selectively processing temporal modulations less corrupted by noise. The cortical representations of speech in noise have been previously studied (Ding & Simon, 2013). They found that intensity contrast grain control and adaptive processing of temporal modulations in the delta and theta band serve as mechanism for sustained neural synchronization to the slow temporal modulations of speech.

Psychoacoustic studies of speech in reverberation and noise combined have found that noise degrades the intelligibility further than reverb alone (Harris & Swenson, 1990). In cochlear implant users it has been observed that the effect of both combined degraded the intelligibility more than any alone (Hazrati & Loizou, 2012).

1.4 The Human Auditory System

In this section, aspects of the human auditory system critical for this work will be discussed. In particular, emphasis will be placed on the peripheral auditory system and the spectro-temporal representation that describe how sounds are transformed through it. Also, although not in detail, it will be discussed how phase-locking to the temporal modulations decays through the central auditory system until the auditory cortex only follow the slow components.

The auditory system is composed of two main areas: the peripheral auditory system and the central auditory system. The peripheral auditory system consists of the outer ear, middle ear and inner ear. The outer ear is composed by the pinna and the ear canal, where the pinna is responsible for collecting sounds, amplifying frequencies relevant for human speech and it also provides information about the directionality of the sound. Once collected, sounds travel through the ear canal until they reach the tympanic membrane in the middle ear. At the middle ear, the ossicles serve as pressure transformers. The later stage in the peripheral auditory system is the inner ear. In the inner ear, the cochlea acts as transducer by converting mechanical vibrations produced by changes in pressure pattern into neural information to the auditory nerve. The basilar membrane performs spectral analysis by the cochlear filter bank. At the hair cell stages, phase-locking decreases beyond 2 kHz. At the end, there is a lateral inhibitory network, which detects discontinuities in the responses across the tonotopic axis and also performs a frequency selectivity enhancement of cochlear filter bank. The auditory spectrogram

(Figure **4**) models the transformation of the acoustic signal into an internal representation performed by the early stage (Yang X, Wang K & Shamma SA., 1992). As opposed to the most common spectrogram, computed by the Short-Time Fourier transform, the auditory spectrogram does not follow the speech formant transitions explicitly.

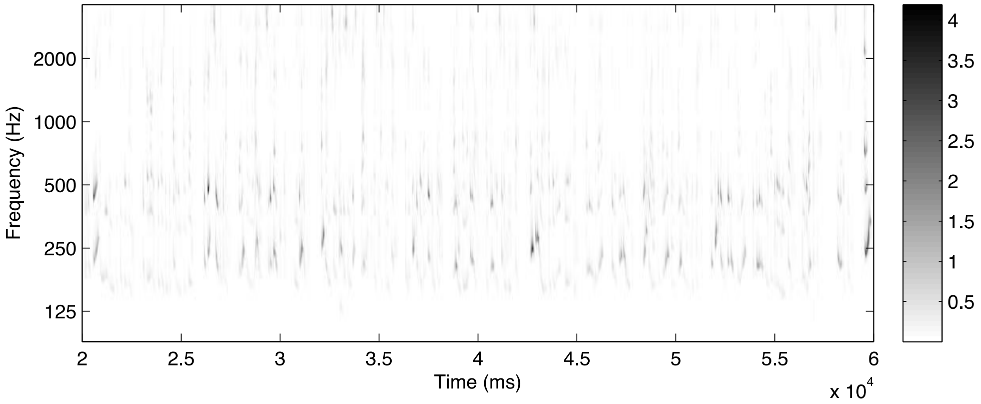


Figure 4. Auditory Spectrogram

The central auditory system receives neural patterns from the auditory nerve carrying information about the sound. It is formed by a series of sub-cortical nuclei that extract information regarding to the directionality (Masterton, 1992) and reduce the synchronization to temporal modulations. At the level of the auditory cortex most neurons in general synchronize to modulations around 10 Hz.

1.5 Magnetoencephalography

The cortical processing of speech in reverberant conditions is studied via MEG. MEG is a noninvasive neuroimaging technique that measures the magnetic fields of neural currents generated by populations of neurons synchronously active in the cerebral cortex.

Specifically, pyramidal neurons in the cortex are responsible of the generation of magnetic fields measured by MEG when they are activated synchronously (Hansen, Kringelbach, & Salmelin, 2010). The exact amount of synchronously active neurons to generate the magnetic fields measured by MEG is unknown, but as suggested by (Hämäläinen, 1993) the weakest measurable cortical signals are in the order of 10nAm, which can be generated by about 50,000 neurons synchronously active.

Neural currents generate weak magnetic fields in the order of femtoTeslas (fT), which are several orders of magnitude weaker than the Earth's magnetic field (~ 10^{-4} T). Superconducting Quantum Interference Device (SQUID) sensors are employed in the recording of these magnetic fields. SQUID sensors are sensitive amplifiers that can detect and amplify relatively small changes in magnetic flux (Baillet, Mosher, & Leahy, 2001). Typical MEG whole-head sensor arrays consist from 100 to 300 sensors distributed around the head. In order to reduce the interference of external magnetic fields the MEG system is located in a magnetically shielded room.

MEG presents great advantages for recording neural activity in the auditory cortex. It provides a good temporal resolution (~1 ms), performs silent recordings and does not require major time for preparation. In particular, MEG sensors are sensitive to currents tangential to the skull or those generated in the cortical sulci. Thus, due to the localization of the auditory cortex, MEG is well suited for studies in auditory neuroscience. It has been shown that MEG is well suited for the study of cortical processing of speech (Ding & Simon, 2012b).

Chapter 2

Cortical Representations of Speech in Reverberant Conditions

2.1 Introduction

Despite the acoustic interference encountered by speech in daily environments its intelligibility is preserved until degradations reach certain thresholds. Among the different temporal scales of the speech signal, those encompassing the slow temporal modulations are critical for the robustness of speech intelligibility (Houtgast & Steeneken, 1985; Rosen, 1992). In terms of the modulation spectrum, the integrity of bands from 1-7 Hz guarantees high intelligibility (Elliott & Theunissen, 2009). It has been suggest that the brain can decode degraded speech as long as the structure of these modulations is preserved (Ghitza & Greenberg, 2009). Studies have demonstrated, through MEG, that populations of neurons synchronize to the slow modulations of continuous speech (Ding & Simon, 2012b). Furthermore, in (Ding & Simon, 2013) it was demonstrated that cortical modulations are robustly synchronized to the slow temporal modulations of speech until noise is 9 dB stronger than speech.

This study emphasizes on the analysis of the cortical representations of speech in reverberation and speech distorted by noise in reverberant environments. Reverberation is characterized by overlap masking, as opposed to the multiple speakers and the background noise scenarios, which are described by informational and energetic masking, respectively. Here, MEG is used to study the neural synchronization to the slow temporal modulations of speech in these adverse conditions.   Subjects listened to spoken narratives in reverberation and noise in reverberation while their cortical responses to these stimuli were recorded.

2.2 Materials and Methods

Subjects

Thirteen subjects (eight females) between 20-30 years old (mean age = 23.8) participated in the study. One subject was not included in the analysis due to problems with the equipment while the experiment was performed.   Experimental procedures were approved by the University of Maryland Institutional Review Board. Written informed consent was obtained from all subjects prior their participation, after the experiment was fully explained. All subjects were right-handed (Oldfield, 1971) native English speakers, reported normal hearing, and were paid for their participation.

Stimuli

The stimuli were retrieved from a single speaker narration of the story *The Light Princess* by George MacDonald, from (“The Light Princess,” n.d.). Speech in reverberant conditions was generated by the convolution of the clean speech with the corresponding RIR for the desired reverberation level. The RIRs were generated using the image-source method (Allen, Berkley, & Hill, 1979) as implemented in a fast manner by (Lehmann & Johansson, 2010). The MATLAB package was retrieved from (http://www.eric-lehmann.com/) under the GNU general public license. This implementation is based on the fact that late reflections of the room impulse response have a decaying noise-like behavior and a decaying random noise model can generate it adequately (Lehmann & Johansson, 2008). The early reflections of the room impulse response are modeled as the original image-source method implemented as described by (Lehmann, 2007).

The simulated room has dimensions 7 x 5 x 3 m (x, y, z). The source is placed at the point 3 m x 2.5 m x 1.7 m, at 1.5 m from the receiver. The receiver is also placed to a height of 1.7 m and is simulated by two microphones placed 0.1 m apart (Figure **5**). The absorption coefficients are uniform for all walls, ceiling and floor, and were adjusted to achieve the desired reverberation time for each room impulse response accordingly. Reverberation times were characterized by means of the RT_{60}. A total of 4 reverberation conditions are considered: anechoic, low reverb, intermediate reverb and severe reverb. Anechoic reverb was simulated by setting the RT_{60} equal to 0 s (Ruggles & Shinn-Cunningham, 2010)

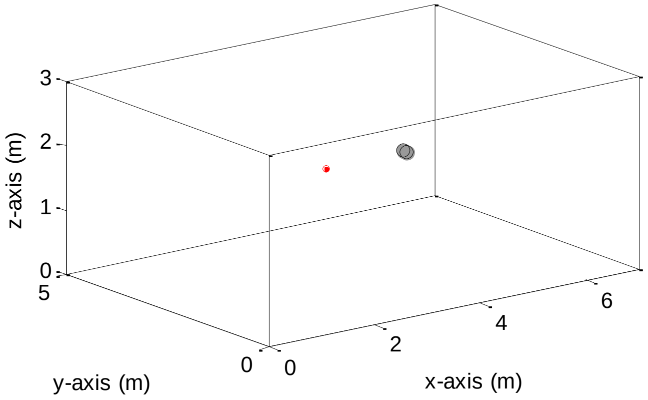


Figure 5. Simulated room. Dimensions of the room and location of the source (red square) and receiver (two ears).

Reverberation time for a RIR $h(t)$ was measured by computing the energy decay $E(t)$ via the Schroeder's integration method as given by

$$E(t) = 10\log_{10}\left(\frac{\int_t^{\infty} h^2(\sigma)d\sigma}{\int_0^{\infty} h^2(\sigma)d\sigma}\right).$$

Measured RT$_{60}$ values found are 0.01 s, 0.29 s, 1.23 s, and 2.15 s for the anechoic, low, intermediate and severe reverb, respectively. The corresponding Schroeder's energy decay plots are shown in Figure 6.

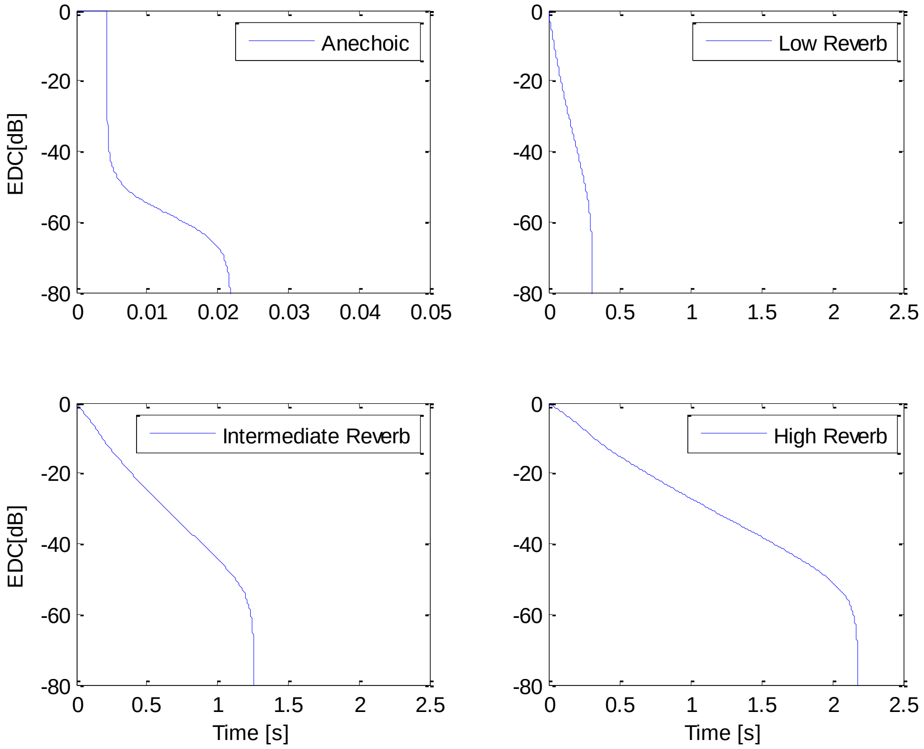


Figure 6. Energy decay plots. Four reverb conditions are considered. Notice the difference in the time scale for the anechoic condition.

In order to simulate the head at the receiver and provide binaural sound, the room impulse response was convolved with a head-related impulse response (HRIR), generating thus a binaural room impulse response (BRIR). HRIRs database was retrieved from ("NSL software: HRTF," n.d.). HRIR are more commonly referred to as Head-Related Transfer function (HRTF) which is the Fourier transform of the HRIR. A total of 7 different HRTF (Grassi, Tulsi, & Shamma, 2003) were considered for this study.

Spectrally matched noise was added in order to study the reverberant noise conditions (Figure 7). Three signal-to-noise ratios were considered: infinite (no noise), -3 dB and -6 dB. Spectrally matched noise was generated by randomizing the phase of the reverberant speech signal and was normalized to have the same energy as the speech. All stimuli were low-pass at 4 kHz and the duration was 60 seconds, each presented for three times.

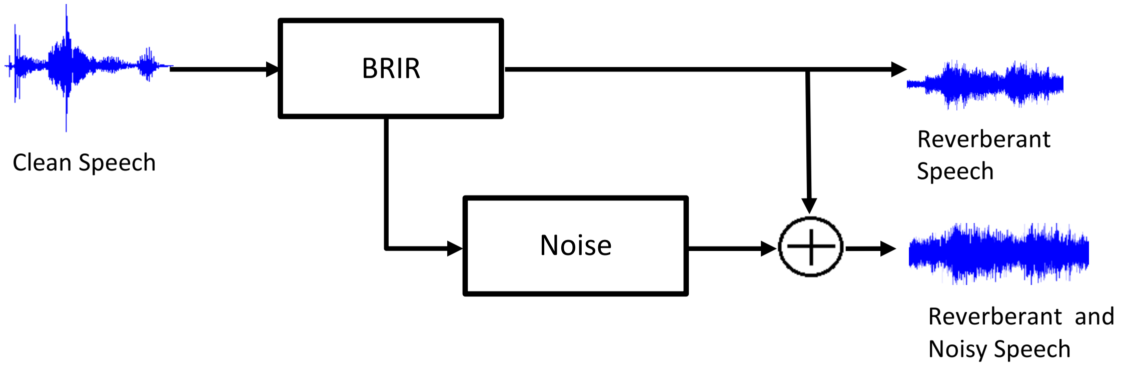


Figure 7. Block diagram for stimulus generation

Procedure

Before the experiment 100 repetitions of a 500 Hz tone pip were presented.   Each tone produces a neural response ~100 ms after the stimulus onset, referred to as the M100, and is an indicator of good auditory response. During the experiment, subjects were required to close their eyes. In order to assure they were paying attention to the excerpts subjects were required to answer comprehensive questions and rate the intelligibility of the passage. The stimuli consisted of three blocks corresponding to each noise condition and within each noise condition the four reverberation conditions were presented. A total of 12 conditions were presented (4 reverb x 3 noise). Before the experiments subjects were asked to test 7 versions of the same stimuli computed by different HRTFs. During a pre-experiment training, subjects also listened to sample stimuli in order to provide an insight of the intelligibility at each condition.

The order in which the stimulus degradations were presented was different for half of the subjects. For one group of subjects the degradations were presented in ascending order. That is, the no noise block was presented first and the -6 dB was the last block. In this scenario the anechoic reverb condition was presented first and the severe reverb condition was the last. For the remaining group, the -6 dB noise block was presented first and the no noise block at the end. Within each noise condition, the severe reverb was presented first and the reverb was decreasing until the last condition was the anechoic reverb.

MEG recordings and data processing

All the experiments were conducted at the University of Maryland, College Park and the MEG recordings were performed on the University of Maryland – Kanazawa Institute of Technology (UMD-KIT) MEG system. This system is placed in a magnetically shielded room and has 157 sensors for recording neural activity. Signals were acquired at 1 kHz sampling rate. A 200 Hz low-pass filter and a 60 Hz band-reject were applied online. Environmental and biological noise was removed offline. Three reference sensors measured the environmental noise and it was suppressed by Time-Shifted Principal Components Analysis (TS-PCA) (De Cheveigné & Simon, 2007). Biological noise arises from stimulus-irrelevant neural activity and, as opposed to responses directly related to the stimulus, these responses are not consistent across trials. Based on this principle, a blind-source separation technique known as Denoise Source Separation (DSS) (De Cheveigné & Simon, 2008) extract the stimulus-relevant components from the neural responses that are consistent across trials. As result, this technique projects the multidimensional neural time series acquired by the MEG system (157) into uncorrelated time series in descending order of reliability. For purposes of this work, only the first component, the most reliable, is considered for all analysis. All responses were down-sampled to 100 Hz and filtered between 1-10 Hz.

Stimulus characterization

Stimuli were characterized by means of the auditory spectrogram. The auditory spectrogram is based on a subcortical model, is computed with 5 ms windows and the frequency is logarithmically spaced. The broadband envelope is defined as the sum of the auditory spectrogram over frequency.

Temporal Response Function

The cortical representations of speech in adverse conditions are modeled by a temporal response function (TRF). The TRF is a model that describes the relationship between the sub-cortical representation of speech and the evoked neural response (

Figure 8). It is obtained by the de-convolution of the cortical response with the speech narrative.

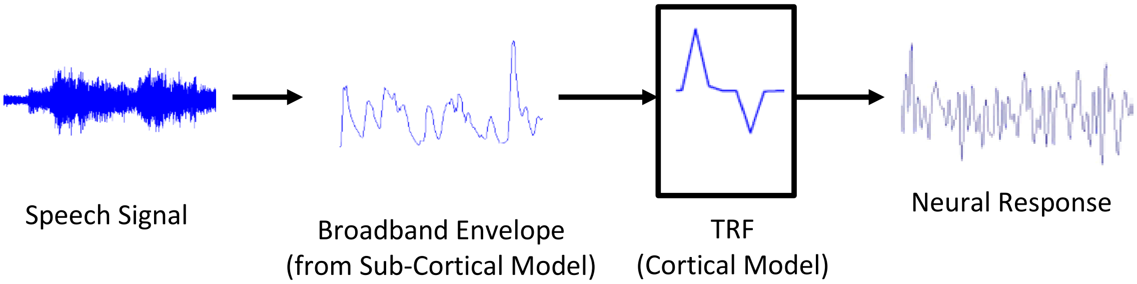


Figure 8. Systems approach of auditory processing

When seen as a linear time-invariant system, the input-output relationship of a system is characterized by the linear convolution. Let $x(t), t \in \{0, 1, \ldots, T\}$ denote the temporal modulations of the stimulus, $y(t), t \in \{0, 1, \ldots, T\}$ be the neural response to that stimulus, $h(t), t \in \{0, 1, \ldots, L\}$ be the impulse response of the system or TRF, where $L < T$, and $\varepsilon(t)$ is the residual error not explained by the linear model. By linear convolution, $y(t)$ and $x(t)$ are related as

$$y(t) = \sum_{\tau=-\infty}^{\infty} x(t-\tau)h(\tau) + \varepsilon(t)$$

In this work, the TRF is estimated by boosting with 10-fold cross validation (David, Mesgarani, & Shamma, 2007). To avoid overfitting, the best model is the one that provides the global minimum for the validation error (Duda, Hart, & Stork, 2001).

The encoding of continuous speech in noise has been previously studied via boosting (Ding & Simon, 2013). In this work the TRF is estimated from the first DSS component.

In general, the TRF for speech contains two dominant peaks at around 50 ms and 100 ms, denoted as M50$_{\text{TRF}}$ and M100$_{\text{TRF}}$ respectively. The M50$_{\text{TRF}}$ is determined by the peak between 0 to 80 ms and M100$_{\text{TRF}}$ is found from 80 to 180 ms.

Neural Decoding

The broadband envelope is reconstructed linearly from the neural responses based on the first DSS component. The estimated envelope computed as

$$\hat{E}(t) = \sum_{\tau=0}^{0.3} M(t + \tau)D(\tau)$$

where $M(t)$ is the cortical response acquired by MEG and $D(t)$ is the linear decoder. The decoder is optimized via boosting, using an integration window of 300 ms.

Coherence Analysis

Inter-trial correlation

The inter-trial correlation is computed as a measure of phase-locking of the neural activity across trials. Here, within each condition, the cross correlation between the neural responses in 1 Hz bands for each trial is computed for the first DSS component. The phase-locking spectrum was also computed for the delta (1-4 Hz), theta (4-8 Hz) and alpha (8-12 Hz) bands.

Phase-coherence

The phase coherence was computed to investigate if it tracked the amplitude-based phase-locking spectrum in different bands (delta, theta and alpha). The first DSS component was band-pass filtered in the frequency band of interest and was converted into its analytical form by the Hilbert transform. The instantaneous phase $\theta(t)$ is extracted by the modulus-argument decomposition. Thus, the phase coherence was computed as

$$C_i = \frac{1}{J}\sum_{j=1}^{J}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\cos(\theta_{nij})\right)^2 + \left(\frac{1}{N}\sum_{n=1}^{N}\sin(\theta_{nij})\right)^2\right]$$

where θ_{nij} is the phase for the frequency bin i, time bin j and trial n (Luo & Poeppel, 2012).

2.3. Results

All subjects showed auditory response as measured by the M100 response. Ten subjects selected the HRTF #1 (ITD = 4.3 ms); two subjects selected HRTF #2 (ITD = 4.3 ms) and one subject chose HRTF#3 (ITD = 4.2 ms).

Intelligibility Assessment

During the experiment subjects were asked to rate the intelligibility under each condition. Figure **9** shows the subjective intelligibility scores, where each plot corresponds to a noise level and each group of bars belong to a reverberation degree. It was found that, regardless the order in which the degradations were presented, the intelligibility decreased monotonically as function of the degradation. In general, subjects for whom degradations were presented in descending order reported higher scores. In order to investigate if the order in which the degradations were presented affected the intelligibility scores significantly, 3-way ANOVA tests were performed (SNR x Reverb x Intelligibility). Results showed no interaction between intelligibility and noise level nor reverb degree.

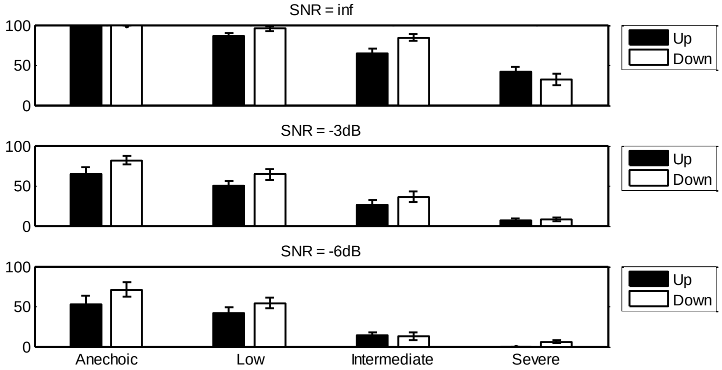


Figure 9. Intelligibility scores reported by subjects. Stimulus degradations were presented in ascending order for half of the group (filled bars) and in descending order for the remaining half. In general, those who listened to the degradations in descending order tend to rate intelligibility higher. No statistical significance of difference between those (3-way ANOVA). As observed, in both cases intelligibility decreased monotonically as function of degradation.

Neural Encoding of Speech

The processing of the spectro-temporal features of the stimulus in the cortex is investigated by the estimation of the TRF. The TRF can be seen as the characterization of the time course of neural activity evoked by a unit power increase of the stimulus

(Ding & Simon, 2013). Here, the stimulus and the response are normalized and thus the estimated TRF is independent of the stimulus statistics.

Only as an illustration and not intended to be considered for final conclusions, TRFs averaged across subjects for the clean and actual stimuli are presented in Figure 10 and Figure 11, respectively. These models show two prominent peaks of opposite polarity at ~50 ms and ~100 ms. The latencies of these peaks provide information regarding to the cortical area that is involved in the processing of a given feature. Specifically, (Ding & Simon, 2012a) relate the $M50_{TRF}$ to be processed about 10 mm anterior to the $M100_{TRF}$, consistent from originating from Heschl's gyrus and planum temporale, respectively.

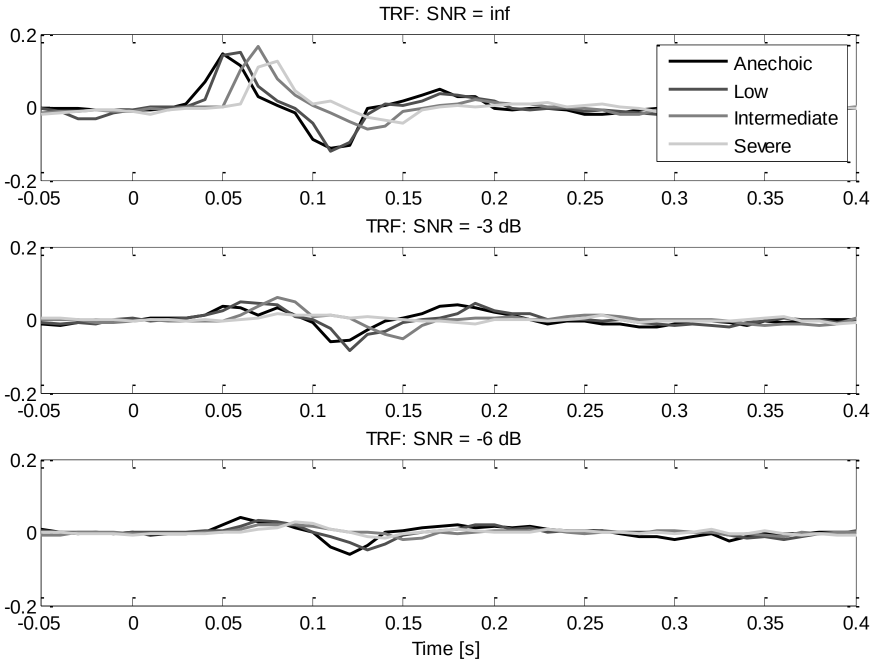


Figure 10. TRFs when clean speech is considered as input. TRFs are shown for each noise level as function of reverb. This is a model that relates the clean speech with the cortical response to the actual stimuli.

From Figure **10**, a filter that relates the clean speech (without noise or reverb added, even when the speech listened to did have noise or reverb) to the measured cortical response evoked by the actual stimuli should have M50$_{TRF}$ and M100$_{TRF}$ peaks delayed and slightly attenuated as effect of reverberation and should be modulated by noise. From the model relating the actual stimuli with the cortical response evoked by it (Figure **11**) no effects of reverberation are observed (at least not until the severe level under no noise), whereas it is clear that noise attenuates both the M50$_{TRF}$ and M100$_{TRF}$.

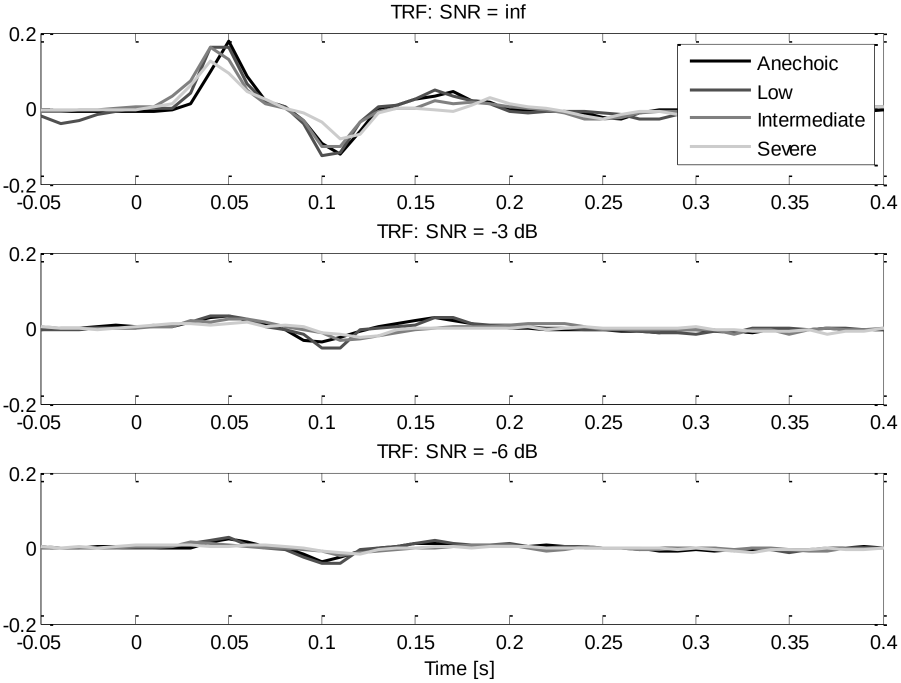
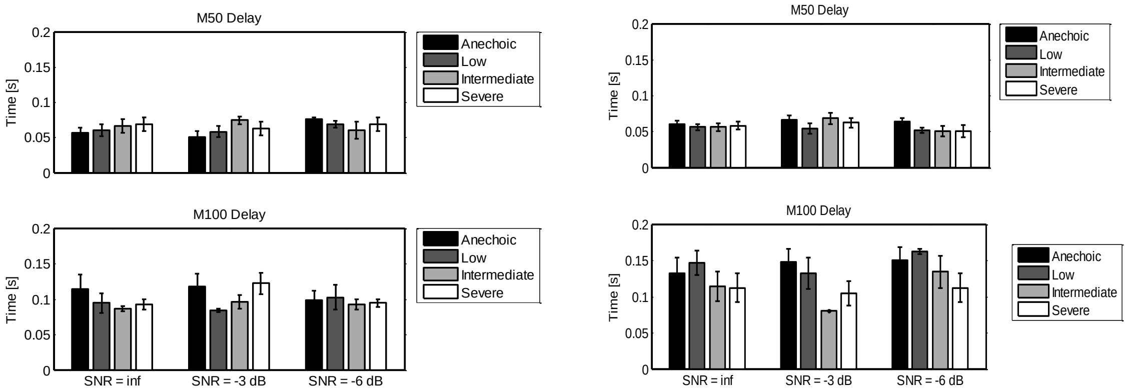


Figure 11. TRFs when actual stimuli is considered as input. TRFs for each noise level as function of reverb. This model describes the relationship between the actual stimuli and the measured cortical response.

To investigate the effects of reverb and noise in reverb on the models the delays and the amplitude of the M50$_{TRF}$ and M100$_{TRF}$ are studied. These are obtained by extracting the peaks from each subject's TRF and then averaging across subjects. From Figure 12, in the model for clean speech (Figure 12 (a)) it is clear that for speech, the cleaning of its noise and reverb delays the $M50_{TRF}$, whereas this is not observed for the $M50_{TRF}$ in the model with the actual stimuli as input. Although no clear pattern is observed in delays on the $M100_{TRF}$, it is clear that the model for the actual stimuli differs from the clean speech model.

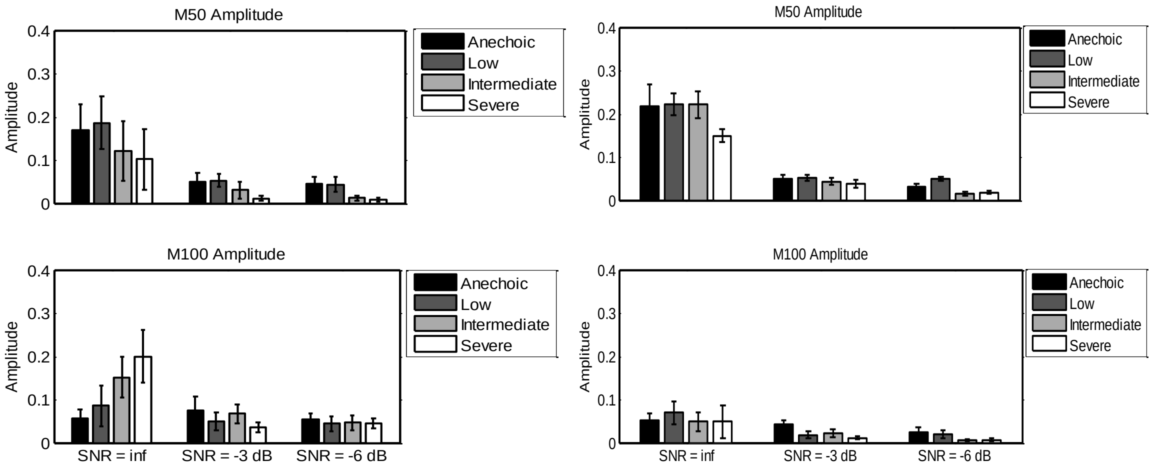


(a)

(b)

Figure 12. Delays of the $M50_{TRF}$ and $M100_{TRF.}$ (a) clean speech model and (b) actual stimuli model

Amplitudes of the models also suggest that there is no cortical evidence of any removal of reverberation, as seen for the $M50_{TRF}$ and $M100_{TRF}$ amplitude peaks (Figure 13). In the absence of noise, a model that relate the response to reverberation should increase the amplitude of the $M100_{TRF}$ peak as a function of reverb, but this is not observed in the model for actual stimuli.



(a) (b)

Figure 13. Amplitudes of the M50$_{\text{TRF}}$ and M100$_{\text{TRF.}}$ (a) clean speech model and (b) actual stimuli model

Cortical reconstruction of speech

In order to study the precision with which the broadband envelope is locked to the synchronized cortical activity it was attempted to reconstruct the clean speech envelope from the neural responses to the reverberant speech. This accuracy is described by the Pearson's correlation coefficient between the clean speech and the neurally-based reconstruction. As Figure 14 top shows, across all reverb conditions it is observed that noise degraded the reconstruction. This is opposed to what was previously found for noise (without reverb) by (Ding & Simon, 2013) where the reconstruction accuracy remained constant until the noise was 9 dB stronger than the clean speech. Thus, reverb worsens the effects of noise. In the same figure, lower plot shows the reconstructions for all reverb conditions as a function of SNR. In the absence of noise (Figure 14 bottom) the reconstruction accuracy remains high across all reverb conditions. This indicates that under pure reverb there must exist a mechanism compensating for the degradations outside the core-auditory cortex.

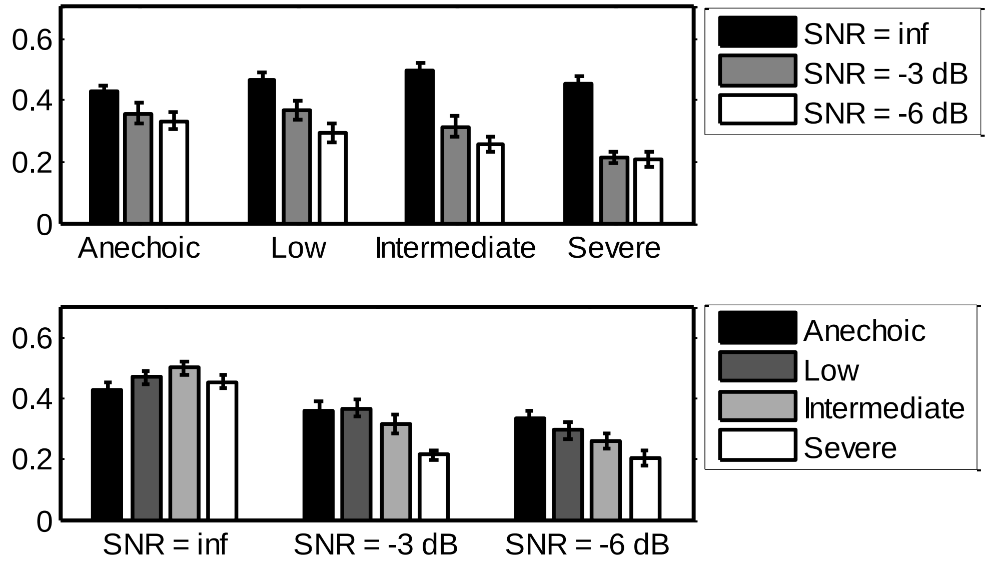


Figure 14. Reconstruction of clean speech from the cortical responses to reverb and noise in reverb (left ear). Top figure illustrates the reconstruction accuracy for each reverb condition as function of noise. In lower figure the reconstruction accuracy for each noise level is presented as function of reverb. In this figure it is better appreciated that in the absence of noise regardless the degree of reverb the reconstruction remains remarkably high.

To investigate further how the cortical activity is synchronized to noise or reverb, the purely noisy or purely reverberant speech was reconstructed from cortical measurements. Figure 15 shows the reconstruction accuracy for the four models considered in this work. Each subplot corresponds to a reverb condition. Reverb and clean are almost equally reconstructed, except at the most severe reverberant condition under no noise, where clean speech is better reconstructed. Across all reverb conditions, clean speech is better decoded than speech in the presence of noise, as previously reported in (Ding & Simon, 2013).

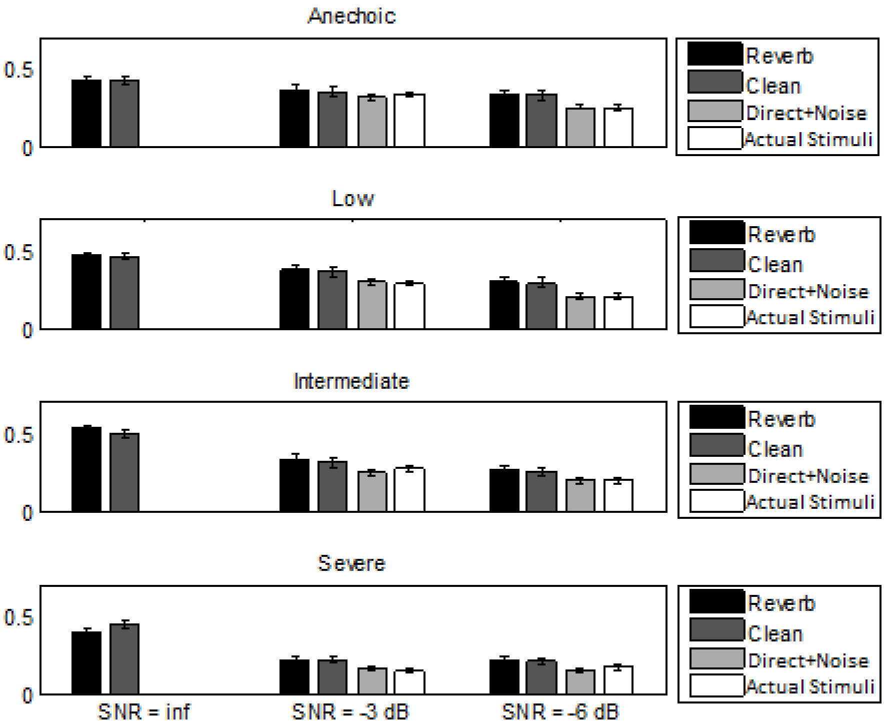
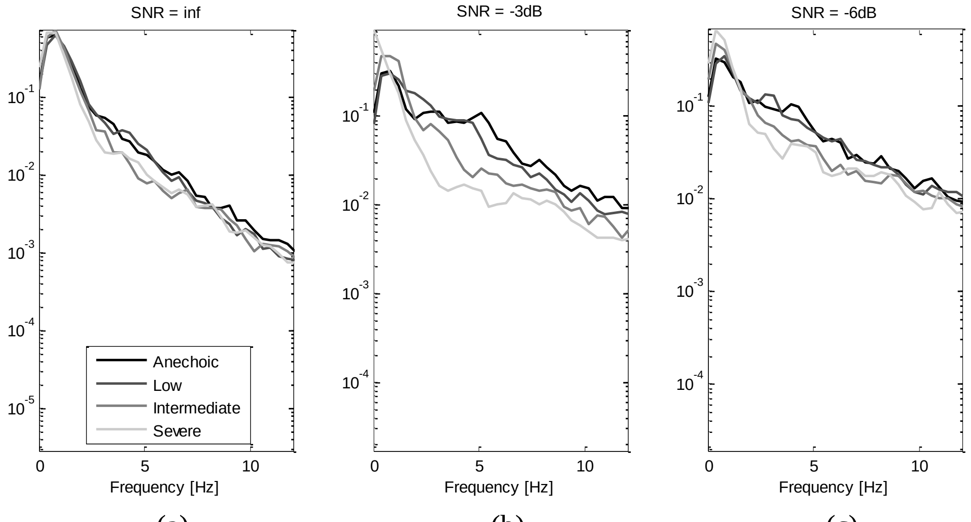


Figure 15. Reconstruction accuracy across models. Each subplot corresponds to a reverb condition.

Modulation sensitivity

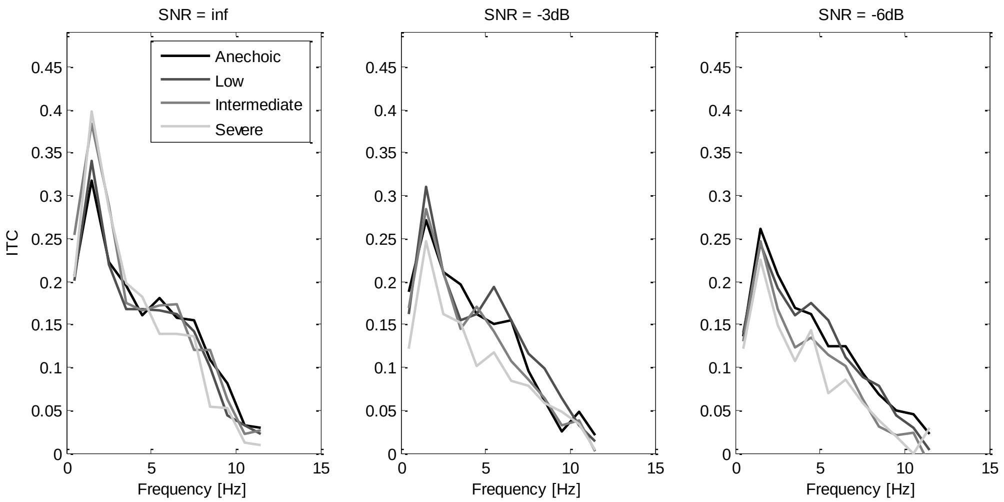
Noise and reverb affect the modulation spectrum in different ways. For the stimuli in this study, noise increases the energy in higher frequencies (>3 Hz) whereas reverb increases the energy in low frequencies (Figure 16 and Figure 18). To investigate if the cortical activity is strictly following the broadband envelope, the phase-locking spectrum is computed in (narrow) bands of 1 Hz width.

The phase-locking spectrum by means of the inter-trial coherence for each noise level as function of reverb is shown in Figure 17. The ITC spectrum exhibits a band-pass like shape in the low frequency range (< 5Hz). The bandwidth in this region was computed, but it remained constant across conditions. Under no noise (Figure 17 (a)) high reverb conditions (intermediate and severe) contain higher energy at frequencies less than 5 Hz. This is not observed in the modulation spectrum (Figure 16 (a)) where under no noise in the band from 1Hz to 5 Hz all stimuli followed the same power distributions. When noise is introduced (Figure 17 (b) and Figure 17 (c)) the coherence decreases and again the ITC spectrum opposes the modulation spectrum between 1 Hz and 5 Hz. Thus, this trend indicate that cortical activity is not strictly following the slow temporal modulations of speech and there must be a neural mechanism involved in the compensation for reverberation.



(a) (b) (c)

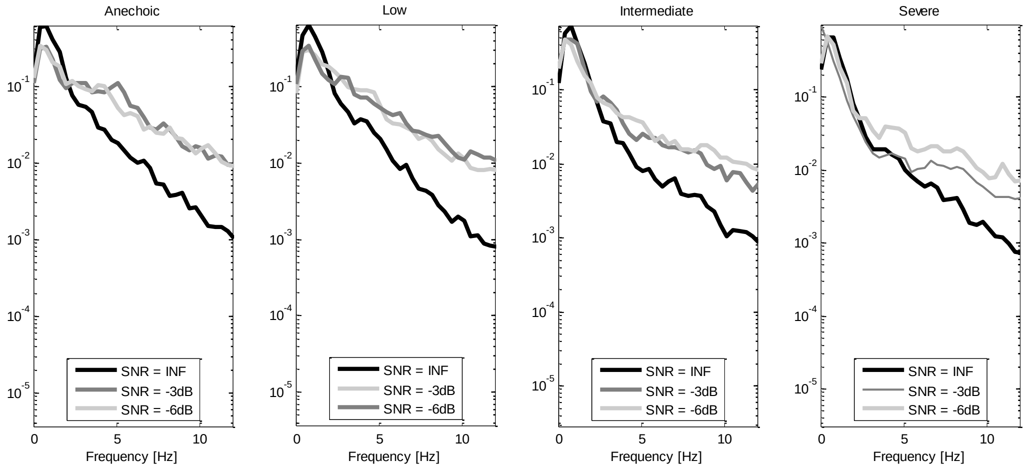
Figure 16. Stimulus power spectral density for each noise level as function of reverb



(a) (b) (c)

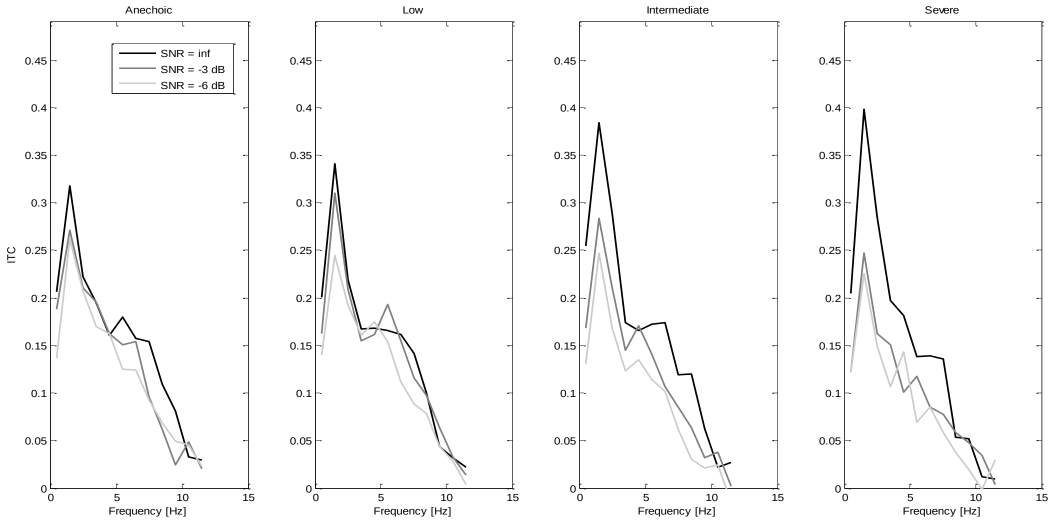
Figure 17. ITC spectrum for each noise level as function of reverb

From Figure 19 as reverb is increased, the power of the ITC spectrum for no noise increases as opposed to the modulation spectrum. In the presence of noise, however, the modulation spectrum for anechoic speech is attenuated but this component is the one containing higher power in the ITC spectrum.



(a) (b) (c) (d)

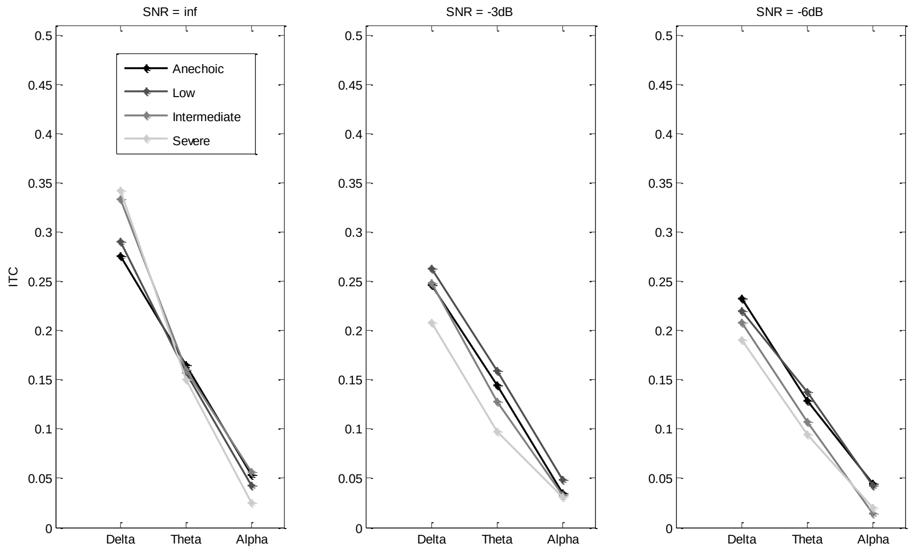
Figure 18. Stimulus power spectral density for each reverb degree as function of noise



(a) (b) (c) (d)

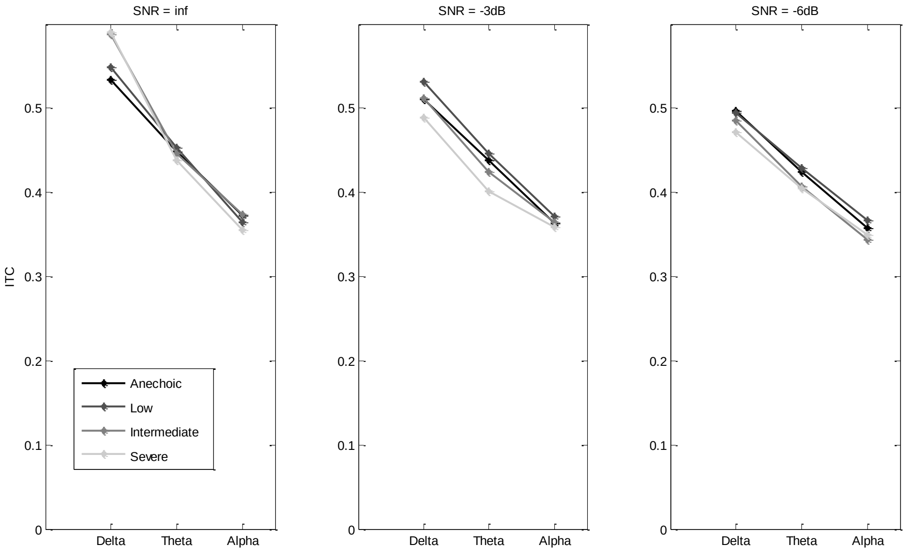
Figure 19. ITC spectrum for each reverb level

In order to investigate whether phase coherence tracks the power-based ITC, both measures are computed for each band: delta, theta and alpha. Comparing ITC and phase coherence for each noise level as function of reverb Figure 20 and 21 it is observed that phase coherence tracks the power-based ITC across the three bands. This is of interest since studies involving stimuli with temporal structures relevant to those encountered in speech have found contradictory tracking between phase coherence and power-based ITC suggesting a dual temporal window mechanism in the human auditory cortex (Luo & Poeppel, 2012).



(a) (b) (c)

Figure 20. ITC in Delta, Theta and Alpha bands



(a) (b) (c)

Figure 21. Phase coherence in Delta, Theta and Alpha bands

2.4 Discussion

Based on the cortical responses to speech in reverberant conditions acquired with MEG, by means of encoding models and decoding accuracy, this study suggests that the brain might process speech in reverberation differently than speech in noise. By comparing models that characterize how the clean speech is encoded by the brain to models that describe the encoding of reverberant speech (reverb alone and reverb plus noise) it was found that the latter are insensitive to the effects of reverberation. Thus, it can be argued that the brain processes reverberation and noise in a separate manner.

When attempting to reconstruct the speech free from reverberation and noise from cortical responses to the actual stimuli, results show that across all reverb conditions noise degraded the reconstruction accuracy. This result contrasts with results found in previous studies (Ding & Simon, 2013) where the reconstruction accuracy (for speech plus noise alone) was robust until the SNR was – 9dB. Thus, although reverberation is processed differently than noise as suggested from findings in the encoding models it worsens the processing of speech in noise. Furthermore, these reconstructions track the reported intelligibility scores only in the presence of noise. Comparisons of the reconstructions for the clean, clean in presence of noise only, reverberant and actual stimuli demonstrated no difference in the decoding of clean and reverb. Reconstructions of clean speech and clean speech in noise alone show that the neural response represents the clean speech, as found by a previous study (Ding & Simon, 2013).

Interestingly, the ITC spectrum shows that cortical activity is not strictly following the slow temporal modulations of neither speech in reverberation nor speech in reverberation and noise. This implies that there must be a neural mechanism involved in the processing of reverberation that is not perceived by the current encoding/decoding techniques.

References

Allen, J. B., Berkley, D. A., & Hill, M. (1979). Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America, 65(4),* 943-950.

Arai, T., Murakami, Y., Hayashi, N., Hodoshima, N., & Kurisu, K. (2007). Inverse correlation of intelligibility of speech in reverberation with the amount of overlap-masking. *Acoustical Science and Technology, 28*(6), 438–441.

Baillet, S., Mosher, J. C., & Leahy, R. M. (2001). Electromagnetic brain mapping. *IEEE Signal Processing Magazine,* (November).

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America, 118*(2), 887.

David, S. V, Mesgarani, N., & Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network (Bristol, England), 18*(3), 191–212.

De Cheveigné, A., & Simon, J. Z. (2007). Denoising based on time-shift PCA. *Journal of neuroscience methods, 165*(2), 297–305. doi:10.1016/j.jneumeth.2007.06.003

De Cheveigné, A., & Simon, J. Z. (2008). Denoising based on spatial filtering. *Journal of neuroscience methods, 171*(2), 331–9.

Devore, S., & Delgutte, B. (2010). Effects of reverberation on the directional sensitivity of auditory neurons across the tonotopic axis: influences of interaural time and level differences. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, *30*(23), 7826–37.

Devore, S., Ihlefeld, A., Hancock, K., Shinn-Cunningham, B., & Delgutte, B. (2009). Accurate sound localization in reverberant environments is mediated by robust encoding of spatial cues in the auditory midbrain. *Neuron*, *62*(1), 123–34.

Ding, N., & Simon, J. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences, 109(29), 11854-11859. 2012*.

Ding, N., & Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of neurophysiology*, *107*(1), 78–89.

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, *33*(13), 5728–35.

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of temporal envelope smearing on speech reception, *95*(February), 1053–1064.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (p. 654).

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS computational biology*, *5*(3).

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1-2), 113–26.

Gold, B., Morgan, N., & Ellis, D. (2011). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*.

Grassi, E., Tulsi, J., & Shamma, S. A. (2003). Measurement of head-related transfer functions based on the empirical transfer function estimate. *Proceedings of the 2003 International Conference on Auditory Display* (pp. 119–122).

Greenberg, S. (2006). A multi-tier framework for understanding spoken language. *Listening to speech: an auditory perspective,* 411-433.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, *31*(3-4), 465–485.

Haas, H. (1972). The Influence of a Single Echo on the Ability of Speech. *Journal of Audio Engineering Society*, *20*(2), 146–159.

Hämäläinen, M. (1993). Magnetoencephalography-theory, instrumentation, and applicatiokns to noninvasive studies of the working human brain. *Reviews of Modern Physics*.

Hansen, P. C., Kringelbach, M. L., & Salmelin, R. (2010). *MEG: An introduction to methods*.

Harris, R. W., & Swenson, D. W. (1990). Effects of reverberation and noise on speech recognition by adults with various amounts of sensorineural hearing impairment. *Audiology*, *29*, 314–321.

Hazrati, O., & Loizou, P. C. (2012). The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners. *International journal of audiology*, *51*(6), 437–43.

Houtgast, T., & Steeneken, H. J. M. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *Journal of Acoustic Society of America*, *77*(3), 1069–1077.

Lehmann, E. A., & Johansson, A. M. (2010). Diffuse Reverberation Model for Efficient Image-Source Simulation of Room Impulse Responses. *Ieee Transactions On Audio Speech And Language Processing*, *18*(6), 1429–1439.

Lehmann, E. A. (2007). Reverberation-time prediction method for room impulse responses simulated with the image-source model, (1), 159–162.

Lehmann, E. A., & Johansson, A. M. (2008). Prediction of energy decay in room impulse responses simulated with an image-source model. *The Journal of the Acoustical Society of America*, *124*(1), 269–77.

Luo, H., & Poeppel, D. (2012). Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex. *Frontiers in psychology*, *3*(May), 170. doi:10.3389/fpsyg.2012.00170

Masterton, R. B. (1992). Role of central auditory system in hearing: a new direction. *Trends in Neurosciences*, *15*(8), 280–285.

Nabelek, A. K., & Pickett, J. M. (1974). Monaural and Binaural Speech Perception Through Hearing Aids Under Noise and Reverberation with Normal and Hearing-Impaired Listeners. *Journal of Speech, Language, and Hearing Research*, *17*(4), 724–740.

Nábělek, A., Letowski, T., & Tucker, F. (1989). Reverberant overlap-and self-masking in consonant identification. *The Journal of the Acoustical Society of America*, *86*(October), 1259–1265.

NSL software: HRTF. (n.d.). Retrieved from http://www.isr.umd.edu/Labs/NSL/Software.htm

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edingburg inventory. *Neuropsychologia*, *9*(1), 97–113.

Palomäki, K. J., Tiitinen, H., Mäkinen, V., May, P. J. C., & Alku, P. (2005). Spatial processing in human auditory cortex: the effects of 3D, ITD, and ILD stimulation techniques. *Brain research. Cognitive brain research*, *24*(3), 364–79.

Rosen, S. (1992). Rosen. Temporal information in speech-acoustic, auditory and linguistic aspects.pdf. *Philosophical Transactions: Biological Sciences*, *336*(1278), 367–373.

Ruggles, D., & Shinn-Cunningham, B. (2010). Spatial selective auditory attention in the presence of reverberant energy: Individual differences in normal-hearing listeners. *Journal of the Association of Research in Otolaryngology*.

Sabine, W. C. (1922). *Collected papers on Acoustics*.

Salminen, N. H., Tiitinen, H., Yrttiaho, S., & May, P. J. C. (2010). The neural code for interaural time difference in human auditory cortex. *The Journal of the Acoustical Society of America*, *127*(2).

Shinn-Cunningham, B. (2002). Speech Intelligibility, Spatial Unmasking, and Realism in Reverberant Spatial Auditory Displays. *Proceedings of the 2002 International Conference on Auditory Display*.

Stone, M. a, Füllgrabe, C., Mackinnon, R. C., & Moore, B. C. J. (2011). The importance for speech intelligibility of random fluctuations in "steady" background noise. *The Journal of the Acoustical Society of America*, *130*(5), 2874–81.

The Light Princess. (n.d.). Retrieved from http://librivox.org/the-light-princess-by-george-macdonald/).

Watkins, A. J. (2005). Perceptual compensation for effects of reverberation in speech identification. *The Journal of the Acoustical Society of America*, *118*(1), 249.

Watkins, A. J., & Makin, S. J. (2007). Perceptual Compensation for Reverberation in Speech Identification : Effects of Single-Band , Multiple-Band and Wideband Noise Contexts, *93*, 403–410.

Yang X., Wang K., Shamma S.A. (1992). Auditory representations of acoustic signals. IEEE Transactions on Information Theory, *38*, 824–839.