UNIVERSITY OF CALIFORNIA,
IRVINE

Relative contribution of amplitude and phase spectra to the perception of complex sounds

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Psychology

by

Sierra Noel Broussard

Dissertation Committee:
Professor Kourosh Saberi, Chair
Professor Virginia Richards
Professor Gregory Hickok

2017

DEDICATION

To my mom
who worked at least as hard as I did
to get me to where I am today

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my committee chair and advisor, Professor Kourosh Saberi, for his valuable guidance and cultivation of my independence.

I would also like to thank my committee members, Professor Ginny Richards and Professor Greg Hickok, for all of their constructive input and advice over the last five years.

I am grateful for the guidance of Professor Barbara Sarnecka, who has contributed considerably to the readability of this document.

I would like to express my gratitude to Professor Kay Livesay, who has been a valuable resource for many years and whose enthusiasm for cognitive science sparked my interest in the field.

Finally, I would like to thank the numerous friends and family who have supported me through my graduate career.

Financial support was provided by the University of California, Irvine, and an NIH T32 pre-doctoral fellowship from the UCI Center for Hearing Research.

CURRICULUM VITAE

Sierra Broussard

EDUCATION

Ph.D. in Psychology 2017
Concentration in Cognitive Neuroscience
University of California, Irvine
Advisor: Dr. Kourosh Saberi
Dissertation: Relative contribution of amplitude and phase spectra to the perception of complex sounds

M.S. in Cognitive Neuroscience 2014
University of California, Irvine

B.S. in Psychology and Mathematics 2012
Linfield College
Honors Thesis: Influence of Social Priming on Speech Perception

RESEARCH INTERESTS

My research focuses on using behavioral measures to determine neurological differences between speech and music processing. I am particularly interested in the effects of musical training on hearing in noise and speech intelligibility.

ACADEMIC EMPLOYMENT

Graduate Student Researcher for Dr. Kourosh Saberi 2012-2017
University of California, Irvine

Graduate Student Researcher for Dr. Virginia Richards 2017
University of California, Irvine

Scientific Consultant for Disabilities Services Center 2015-2017
University of California, Irvine
Provide accurate text transcriptions of scientific images and figures for visually impaired students

Student Researcher for Dr. Kay Livesay 2011-2012
Linfield College

POSTERS AND PRESENTATIONS

Broussard, S., Hickok, G., and Saberi, K. Can perceptual oscillations be influenced by attention? Poster presented at the Association for Research in Otolaryngology Midwinter Meeting, February 2017.

Broussard, S., Hickok, G., and Saberi, K. Effects of temporal segmentation depend on smallest meaningful unit size. Poster presented at the Association for Research in Otolaryngology Midwinter Meeting, February 2017.

Broussard, S. (2016, May 28). Envelope and Phase Decorrelation. Presentation at the 11th Annual Center for Hearing Research Symposium, University of California, Irvine.

Broussard, S., Hickok, G., and Saberi, K. Speech intelligibility and melody recognition are differentially affected by degraded amplitude and phase spectra information. Poster presented at the Association for Research in Otolaryngology Midwinter Meeting, February 2016.

Broussard, S., Hickok, G., and Saberi, K. Amplitude and phase spectra information contribute to speech intelligibility and melody recognition differently. Poster presented at the Society for Neurobiology of Language Annual Meeting, October 2015.

Broussard, S., Hickok, G., and Saberi, K. Effects of phase- and amplitude-spectrum decorrelation on speech intelligibility. Poster presented at the Southern California Hearing Conference, August 2015.

Broussard, S., Hickok, G., and Saberi, K. Effects of phase- and amplitude-spectrum decorrelation on speech intelligibility. Poster presented at the Society for Neurobiology of Language Annual Meeting, November 2013.

Livesay, K. and **Broussard, S**. Social Priming and Speech Perception. Poster presented at the Psychonomic Society Annual Meeting, November 2013.

Broussard, S. and Livesay, K. The Influence of Social Priming on Speech Perception. Poster presented at the Society for Computers in Psychology Annual Meeting, November 2012.

Broussard, S. (2012, May 5). The Influence of Social Priming on Speech Perception. Presentation at the Linfield College Psychology Department, McMinnville, OR.

PUBLICATIONS

Broussard, S., Hickok G., and Saberi, K. (in preparation). Relative role of amplitude- and phase-spectrum cues in music perception. Music Perception.

Broussard S., Hickok G., Saberi K. (2017). Robustness of speech intelligibility at moderate levels of spectral degradation. PLOS ONE 12(7): e0180734. https://doi.org/10.1371/journal. pone.0180734

FELLOWSHIPS AND AWARDS

Center For Hearing Research NIH Predoctoral Trainee Fellowship 2015-2017
University of California, Irvine
Tuition covered and $17,000 per year.
A competitive interdisciplinary fellowship awarded to students who demonstrate a potential and motivation for a career in hearing research. Applied two consecutive years and received both times.

Associate Dean's Fellowship, School of Social Sciences 2015
University of California, Irvine
Tuition covered and $6000.

Ploog-Teilmann Award for Outstanding Research in Psychology 2012
Linfield College

TEACHING EXPERIENCE

All teaching experience gained at University of California, Irvine unless otherwise noted.

Instructor 2017
"Psychology Fundamentals B"

Teaching Assistant to Dr. Ted Wright 2017
"Psychology Fundamentals B "

Guest Lecturer for Dr. Emily Grossman 2017
"Cognitive Neuroscience"

Guest Lecturer for Dr. Kayoko Okada, Whittier College 2014, 2015
"Brain and Behavior" (2015) and "Language and the Brain" (2014)

Teaching Assistant to Dr. Bruce Berg and Dr. Don Hoffman 2013, 2014
"Psychology Fundamentals A"

Teaching Assistant to Dr. Kourosh Saberi 2014
"Experimental Psychology"

Guest Lecturer for Dr. Christine Lofgren 2014
"Adolescent Psychology"

Teaching Assistant to Dr. Lisa Pearl 2014
“Language Acquisition I”

Teaching Assistant to Dr. Alyssa Brewer 2013
“Brain Disorders”

Teaching Assistant to Dr. John Hagedorn and Dr. Christine Lofgren 2012,
“Introduction to Psychology” 2013

SERVICE

Workshop Facilitator for UCI Summer Premed Program 2017
Helped organize and run a half-day workshop that introduced high school students to auditory neuroscience concepts and lab work.

Co-Organizer for the Center for Hearing Research Lecture Series 2015-2017
Collaborated to choose speakers for an interdisciplinary lecture series. Contacted and hosted visiting speakers.

Science Fair Judge for Foothill Ranch Elementary School 2016,2017

Co-Organizer for Graduate Student Academic Writing Workshop 2015
Helped lead weekly meetings for year-long program. Discussed writing strategies and aided in developing positive solutions to common writing problems. Encouraged better writing habits and provided constructive feedback on the writing of fellow group members. Helped draft a funding proposal for the workshop and was granted $500 by the department of Social Sciences.

PROFESSIONAL MEMBERSHIPS

Society for Neurobiology of Language
Association for Research in Otolaryngology

ABSTRACT OF THE DISSERTATION

Relative contribution of amplitude and phase spectra to the
perception of complex sounds

By

Sierra Noel Broussard

Doctor of Philosophy in Psychology

University of California, Irvine, 2017

Professor Kourosh Saberi, Chair

Speech processing involves analysis of complex cues in both spectral and temporal domains. This dissertation describes a set of studies that explore how speech and music, the two most complex and ecologically important types of sound, are affected by spectral degradation using a method that orthogonally and parametrically decorrelates their amplitude and phase spectra. The first study investigates how amplitude and phase information differentially contribute to speech intelligibility. Listeners performed a word-identification task after hearing spectrally degraded sentences that were segmented into temporal units of varying lengths (e.g., phoneme and syllable durations) before the decorrelation process. Results showed that for intermediate spectral correlation values, segment length is generally inconsequential to intelligibility, and that intelligibility overall is more adversely affected by phase-spectrum decorrelation than by amplitude-spectrum decorrelation. The second study investigates how amplitude and phase information differentially contribute to melody discrimination and speech intelligibility to better characterize processing differences between music and speech. Listeners heard spectrally degraded melodies and performed a same-different judgement in a psychophysical discrimination task. Melody recognition was relatively unaffected by partial decorrelation of the amplitude spectrum and more resilient to loss of phase-spectrum cues for both short and long-duration analysis segments. The third study examines the effects of speaking rate and spectral degradation on speech intelligibility. Consistent with prior findings, phase-spectrum cues were most useful to intelligibility at longer temporal windows of analysis, and amplitude spectrum cues at short windows. For normal rate speech, the crossover point between these two cues occurred at an estimated window size of 120 ms; i.e., amplitude-spectrum cues were more useful to intelligibility below this value and phase-spectrum cues were more useful above this window size. Increasing speaking rate to twice normal rate, surprisingly seemed to have little to no effect on this crossover point. However, slowing down speaking rate shifted this crossover point to significantly longer temporal window sizes (~230 ms). Implications of these findings for cues critical to intelligibility of speech at different speaking rates, and in particular, the importance of preserving narrowband temporal envelope cues are discussed.

INTRODUCTION

The human vocal tract is able to produce a wide variety of speech sounds, each defined by a place of articulation, manner of articulation, and voicing. There are over 100 different phonemes currently established, and 44 of these are used by English speakers (Crystal, 2010). Vocal tract resonances can be defined using peaks in the spectral envelope, or formants. Each of these possible changes to the vocal tract defines the way the speech signal is produced and creates a unique acoustic signal that must then be interpreted by the auditory system. Consequently, in order to understand speech perception, it is necessary to understand how these sounds are represented in both time and frequency domains, and what type of cues are necessary in order to preserve them for accurate identification.

Humans appear to have certain kinds of signal processing in the auditory cortex that are specialized to interpret speech signals. However, speech is not the only type of signal that carries meaning; music is another type of complex and ecologically important type of sound (Peretz and Zatorre, 2005). Similar to speech, there is significant evidence that music perception is a human specific ability (Bispham, 2006). Music is mainly described in terms of pitch and time components, rhythm and tempo. In both pitch and component duration, there is considerably greater variance in music than in speech. Considerably less research has been dedicated to music processing, but recent studies have begun to illustrate the similarities and difference in speech and music perceptual processing (Peretz and Zatorre, 2005; Nunes-Silva and Haase, 2011; Koelsch, 2011).

Temporal cues used for speech and music perception

Slow rate (~4Hz) temporal features of speech are considered critical to speech intelligibility (Saberi and Perrott, 1999; Greenberg et al., 2003; Greenburg and Arai, 2004; Ghitza and Greenberg, 2009). Shannon et al. (1995) showed that temporal cues prove sufficient for 90% of word identification, suggesting that minimal fine structure information is required for speech recognition when adequate temporal cues are available. Since then, considerable amounts of behavioral and neurological data have supported this finding.

For example, one study synthesized "auditory chimeras" to use as stimuli (Smith, Delgutte, and Oxenham, 2002). These auditory chimeras were created in a similar manner to vocoding but they used the narrowband envelopes of one sentence to modulate the fine structure of a second sentence. They demonstrated that the types of cues used in speech recognition are highly dependent on the number of bands used in synthesis; envelope contributed most to recognition at larger numbers of bands while fine structure was used more at lower band numbers. Furthermore, envelope cues were shown to be more resistant to conflicting information. A later study (Zeng et al., 2004) responded to these findings, suggesting that the role of fine structure was over interpreted due to the auditory filter's ability to recover temporal envelope from broadband fine structure, meaning envelope remains the essential factor in speech intelligibility regardless of the size of the frequency band.

Unlike speech research, there has been a lack of research exploring the relative importance of cue type in the spectral and temporal domains for musical stimuli. Smith and colleagues (1997) investigated the necessary temporal cues for melody discrimination by creating melody-melody chimera stimuli. Their results for the melody-melody chimeras were the opposite of their results for the speech-speech chimeras. For melody recognition, most listeners required 32 bands before they could reliably identify the envelope melody over the fine structure melody. Participants also often reported hearing both melodies, which was much rarer in the speech studies. Because adequate frequency information is unavailable until band sizes are small, and melody recognition relies so heavily on pitch perception, this finding is unsurprising. Similarly, the double melody perception experienced simply meant that both the envelope and fine structure carried enough frequency information to allow both perceptions.

Cochlear implant (CI) users have helped establish the necessity of temporal information for sound discrimination because CIs only transmit temporal envelope information. For example, while temporal envelope information is adequate for speech perception in quiet, the same is not true for noise degraded signals (Dorman et al. 1998; Friesen et al., 2001). One reason is that CI users are not assisted by temporal fluctuations in the masker (Lorenzi et al. 2006; Gnansia et al. 2008). Intelligibility is reduced when there is modulated background noise and listeners are using envelope only cues, suggesting fine structure cues are necessary for masking release.

CI users also suffer from a loss in music perception ability. However, several studies have shown that CI listeners can use temporal information, particularly at low frequencies, to achieve some form of music perception. Surprisingly, some are even able to determine whether or not a note is out of tune in familiar melodies using low frequency temporal information (Zeng, 2002; Shannon, 1989). However, for complex rhythm tasks, there was a great deal of individual differences in perceptual ability, suggesting a relation to cognitive capacity (Pijl, 1997).

Because of the unique constraints on CI users, this population has been used to establish the relative contribution of temporal and spectral cues in music perception. One study (Kong et al., 2004) tested them across both pitch and rhythm tasks and compared their results to those of a normal hearing population. Both groups performed equally on the tempo discrimination task, requiring a rate difference of 4-6 bpm to perceive a difference. CI users performed slightly worse on pattern identification and significantly worse on melody identification.

Speech-rate specific neural responses

There is evidence that the auditory system preferentially responds to stimuli with envelope fluctuations at these slow rates that are highly correlated with intelligibility. Luo and Poeppel (2007) showed that the phase pattern of the theta-band responses from the auditory cortex reliably discriminates the spoken sentence signals. This discrimination is dependent on intelligibility of the sentence, and theta phase tracking became less robust as intelligibility decreased. Their results imply that continuous speech is processed at a temporal window of ~200 ms (5 hz) that changes according to speech dynamics.

The demonstrated window length is unsurprising, since average syllable length is also ~200 ms, and the syllable has been suggested as a fundamental unit for speech perception and production (Greenburg and Arai, 2004). Information regarding the syllable sequence in continuous speech is critical for language understanding, so a temporal window of this size allows for optimal intelligibility (Greenburg et al., 2003). Having a syllable-based window size suggests that the theta phase patterns observed in for distinct sentence stimuli most likely differ due to variation in timing across sentences.

Luo and Poeppel (2012) showed that the auditory system responds in the same way to non-speech stimuli with similar properties and that high frequency (~50 hz) oscillations also aid in speech processing. Their stimuli consisted of concatenated frequency-modulated segments with means of 25, 80, and 200 ms, aligning with low gamma, high alpha, and theta band frequencies respectively. In the gamma frequency range, the 25 ms stimulus elicited the most reliable phase pattern. Low frequencies displayed the same behavior as the previous studies, but the 80 ms stimulus did not drive phase tracking efficiently at any frequency. A lack of phase locking for the 80 ms stimulus reveals that oscillations are not responsive to just any time window, and suggests that the phase locking in the theta and gamma bands are specifically designed to code for speech sounds. The two preferred time windows, 200 ms and 30 ms, correspond to average lengths of syllables and phonemes respectively (Greenberg et al., 2003; Drullman, 1995).

Spectral cues used for speech and music perception

Evidence seems to clearly support that the auditory system relies primarily on temporal envelopes at least for speech recognition cues, especially when the signal is not compromised in any way. However, humans produce both speech and music at a wide variety of tempos, suggesting that something besides temporal envelope may be playing a significant role in the processing of some meaningful acoustic stimuli. The auditory system may make use of more spectral cues for stimuli outside of the standard 3-8 Hz temporal envelope fluctuation rate. Only considering the time domain fails to include potentially important signal features in the spectral domain that contribute to speech recognition. More specifically, both phase and amplitude spectra should be analyzed for their relative importance in speech intelligibility.

Several studies have shown that signal identification is more dependent on long-term phase spectrum information than amplitude information (Oppenheim and Lim, 1981; Traunmüller and Lacerda, 1987; Drullman, Festen, and Plomp, 1994; Liu, He, and Palm., 1997). One such study created hybrid signal stimuli to examine the relative importance of amplitude versus phase spectra for perceiving intervocalic stop consonant sounds (Liu, He, and Palm., 1997). Amplitude spectrum information tended to be primarily more informative for shorter window sizes and got progressively less useful as the window size increased. Listeners tended to name the vowel corresponding to the amplitude information more frequently when the stimuli had different places of articulation or when they shared the same voicing property, especially for voiced phonemes, suggesting that amplitude spectrum cues are more responsible for perception of place, while phase spectrum cues are more important for transmitting voicing information.

Overall, changing the phase spectrum of a speech sound not only changes the sound quality, but phase information also carries cues that are vital for perceiving certain acoustic features, particularly at larger time window sizes. Furthermore, these findings hold for full sentences as well as individual phonemes (Paliwal and Alsteris, 2005; Alsteris and Paliwal, 2006; Kazama et al. 2010). Some have argued that the usefulness of specific spectral cues depend on which type best preserves the temporal envelopes (Kazama et al., 2010). However, speech-in-noise and music aren't fully characterized by envelope, and speech at non-standard rates does not carry the same type of information at these specific low fluctuation rates (Kong et al., 2004; Saberi and Perrott, 1998; Luo and Poeppel 2012; Ghitza and Greenberg, 2009; Elliot and Theunissen, 2009; Venezia, Hickok, and Richards, 2016). Therefore, understanding the type of spectral cues that are most crucial to all types of music recognition and speech ineligibility will enhance our understanding of how speech and music are processed.

This dissertation explores how speech and music, the two most complex and ecologically important types of sound, are affected by spectral degradation using a method that independently decorrelates their amplitude and phase spectra. The first chapter investigates how amplitude and phase information differentially contribute to speech intelligibility. The second chapter investigates how amplitude and phase information differentially contribute to melody discrimination and speech intelligibility to better characterize processing differences between music and speech. The third chapter examines the effects of speaking rate and spectral degradation on speech intelligibility. In each of these chapters, we discuss the implications of these findings and address the relative importance of preserving narrowband temporal envelope cues for each stimulus type.

CHAPTER 1: Speech intelligibility at moderate levels of spectral degradation

This study was published in PLOS ONE at https://doi.org/10.1371/journal. pone.0180734.

1.1 Introduction

Phase spectrum analysis is often ignored in models of auditory spectral processing in humans despite our knowledge that humans are not phase deaf when listening to complex sounds. Phonemes, for example, are most often represented as a structural component of the amplitude spectrum (Kazama et al., 2010; Liu, He, and Palm, 1997). However, a number of studies have found that phase plays a major role in speech analysis and recognition. Oppenheim and Lim (1981) found evidence through informal experiments that phase information could be useful in speech-signal reconstruction for long signal times, concluding that changing the phase spectrum of a speech sound can alter its phonetic value.

Humans are able to identify vowels using only phase spectrum information at low fundamental frequencies, and speech comprehension has been shown to be more dependent on long-term phase spectrum than amplitude-spectrum information (Liu, He, and Palm, 1997; Traunmüller and Lacerda, 1987; Drullman, Festen, and Plomp, 1994). Liu and colleagues (1997), for example, investigated the impact of the phase spectrum on stop consonants and found that it is used to determine voicing properties and is critical for setting the structure of formant transitions. Phase information is also more important for consonants with strong burst releases than weak burst releases. Another study found similar results using full sentence stimuli (Kazama et al., 2010). Phase degradation has also been reported to make speech in noise recognition more difficult (Shi, Shanechi, and Arabi, 2006), however the interpretation of this finding is confounded by the methods employed, as adding noise to speech whose phase spectrum has been degraded by a preset value, will further degrade the phase spectrum, resulting in inaccurate measures of the effects of phase-spectrum degradation on intelligibility.

A critical question is the effect of the temporal window of spectral analysis on the relative contribution of amplitude and phase spectra to speech intelligibility. Several studies have shown that the type of spectral information that best maintains intelligibility varies by window length (Kazama et al., 2010; Liu, He, and Palm, 1997). It has been shown that for phoneme length (<128 ms) time windows, amplitude information is most useful to intelligibility. However, at longer (>128 ms) window lengths, phase-spectrum information is more important. This 128 ms crossover point falls almost exactly between the average durations of phonemes and syllables, which have been suggested as basic segments of analysis in speech processing (Giraud and Poeppel, 2012). The average lengths of these speech units are ~30 ms and ~250 ms, respectively, and recent EEG and MEG research has presented evidence of a neural basis for these two window sizes in speech perception (Luo and Poeppel, 2012; Giraud et al., 2007; Howard and Poeppel, 2010; Peelle and Davis, 2012; Gilbert and Lorenzi, 2006). These studies have shown that the auditory cortex prefers stimuli with temporal modulations at gamma-band (~20-80 ms) and theta-band (~150-300 ms) rates, suggesting that these may represent some form of neural parsing or temporal integration (Luo and Poeppel, 2012).

Temporal envelope, fine structure, and periodicity each contribute different types of cues to speech intelligibility (Rosen, 1992). Phonemes are identified by a combination of voicing, manner, and place of articulation. Information about voicing and manner of articulation appear in all three of the previously mentioned signal components. Manner and voicing cues appear in envelope information as differences in rise times (as in 'chip' and 'ship'), long periods of high amplitude for vowels, or as brief silent gaps to indicate a voiceless plosive (Raphael and Isenberg, 1980; Repp et al., 1976; Summerfield et al., 1981). Aperiodicity and high-frequency fine-structure cues can signal that a sound is either voiceless or a fricative (Soli, 1983). Place of articulation is determined by the frequency spectrum of initial release bursts and consecutive formants, which is information found in fine structure (Hazan and Rosen, 1991; Harris, 1958). Tempo and stress help to parse sentences and distinguish between certain types of words (such as rebel and rebel). These parsing cues are only found in periodicity and temporal envelope information. While gaps of silence in the temporal envelope do not necessarily demarcate word boundaries, tempo is still a helpful envelope cue for segmenting words. Similarly, tempo can provide weak cues for vowel identity due to the covariance of vowel length and vowel quality (Lehiste, 1970). Periodicity is the prime correlate of vocal pitch because it represents the rate of vocal fold vibration. Patterns of vocal pitch provide the primary cues used to indicate which words and syllables are stressed; these are extremely important cues to word identity in tonal languages such as Chinese. However, increases in the amplitude of temporal envelope also play a small role in marking stress (Fry, 1968).

Most recent studies on speech intelligibility have focused on the temporal envelope modulations of speech signals. Several studies have demonstrated that, as long as the signal's narrowband temporal envelopes are adequately preserved, a speech signal will be intelligible regardless of how the speech spectrum information is altered (Shannon et al., 1995; Smith, Delgutte, and Oxenham, 2002; Zeng et al., 2004). It is argued that speech is made less intelligible by degrading information in one or both spectral domains (amplitude or phase), mainly because the temporal envelope is also degraded by these manipulations. By modelling the outputs of peripheral filters, one group of researchers determined that the intelligibility of spectrally degraded stimuli was highly correlated with narrowband envelope preservation (Kazama et al., 2010). These findings suggest that the necessary spectral information for intelligibility is ultimately dependent on the type of information that best preserves the temporal envelope.

Naturalistic speech environments, however, are best represented by intermediate spectral correlation values since amplitude and phase spectra of a signal will both be partially degraded in a noisy or reverberant environment. All prior findings in this area of research are based on stimuli with only one type of spectral component preserved, usually achieved by separately decorrelating to zero either the amplitude or phase spectrum relative to the original waveform. Thus, the resulting stimuli maintain either the original amplitude or phase spectrum only, while the other spectral component is usually replaced with noise (Kazama et al., 2010; Liu, He, and Palm, 1997; Paliwal and Alsteris, 2005; Alsteris and Paliwal, 2006).

The purpose of this study is to investigate the relative contributions of phase and amplitude spectra on sentence intelligibility by independently decorrelating, to various degrees, their amplitude and phase spectra relative to those of the original sentence across several time-window sizes. Investigating intelligibility using intermediate phase and amplitude correlation values (between 0 and 1) will allow a better understanding of their individual and joint influence on speech perception. Furthermore, these results will provide intelligibility scores for a larger variety of degraded temporal envelopes, allowing an in-depth analysis of the relationship between spectral and temporal representations of speech stimulus.

1.2 Methods

1.2.1 Participants

Informed written consent was obtained from all participants. Fifteen adult listeners participated in the study (6 females, Mean age = 25 years, σ = 2.2). All participants had normal hearing and were native English speakers. None were familiar with the sentences in the Hearing in Noise Test (HINT) database (Nilsson, Soli, and Sullivan, 2006). Subjects were recruited through IRB-approved postings on campus and through word of mouth starting in 2013 and continuing through 2016. Some had participated in prior experiments and had indicated an interest to participate in the current study. No subjects dropped out of the study or were excluded from data analysis. This study was approved by the IRB of the University of California, Irvine (HS# 2010-7679).

1.2.2 Stimuli

Each stimulus was created by taking a sentence from the HINT database and adding noise through a decorrelation process (Fig 1.1). First, the sentence was divided into one of three time-window sizes: 30 ms, 250 ms, or equal to the duration of the sentence. Each segment was then Fourier transformed, yielding separate amplitude and phase spectra. These spectra were then separately decorrelated relative to the original by a specific amount. The decorrelation process had several stages. First, for amplitude-spectrum decorrelation, we added to each amplitude component in the frequency domain, a random number selected from a Rayleigh distribution. A Rayleigh distribution was selected because the amplitude components of Gaussian noise in the frequency domain are Rayleigh distributed. The vector containing the amplitude-spectrum values of the speech sound was added, on a point-by-point (bin by bin) bases, to a vector of the same size containing the random numbers from the noise distribution (with appropriate adjustments for negative frequency components):

$$a'(f) = k * n(f) + (1 - k) * a(f) \qquad (1)$$

where a(f) is the amplitude-spectrum vector as a function of frequency, n(f) is the noise vector, a'(f) is the new, decorrelated amplitude spectrum, and k is a scalar. We then measured the Pearson product-moment correlation value (r) between a(f) and a'(f). When k=0, the correlation between the new and original amplitude spectrum of speech is 1 (full correlation). When k=1, the amplitude spectrum of speech is fully replaced with that of Gaussian noise, and the correlation is zero. For in-between values (moderate correlation values), we first generated a k-to-r transfer function that provided an initial estimate of how the values of k are associated with specific correlation values (between original and degraded amplitude spectrum of speech). This was done by incrementally adding noise (i.e., increasing value of k) to the amplitude-spectrum of several speech sentence and measuring the resulting correlation. The transfer function was saved and served as an initial starting point for determining the relation between k and r on each trial. On any given trial of the experiment, a speech segment was decorrelated by adding noise to the amplitude spectrum as described above, and fine tuning the value of k iteratively in a loop till the desired correlation between a'(f) and a(f) was achieved within a tolerance limit of smaller than 0.01. This was done for every segment of every speech sentence independently and on every presentation of a new sentence. A similar procedure was used for decorrelating the phase spectrum with the following differences: 1) phase noise was selected from a 0-2pi uniform distribution; 2) the correlation measured was not the linear Pearson value, but a circular statistical correlation value that has the same properties as a linear Pearson, but takes into account the circular nature of phase wrapping (Fisher, 1995; Berens, 2009). Each segment was then inverse Fourier transformed to the time domain, it's RMS level matched to the original segment's RMS, its start and end points smoothed with a ~4 ms linear rise-decay ramp (100 samples at 22.05 kHz) to reduce spectral splatter at transition points between segments in a sentence, and then concatenated with other segments in their original order to generate the degraded sentence. The entire decorrelation process took less than 1 second and done between trials of a run.

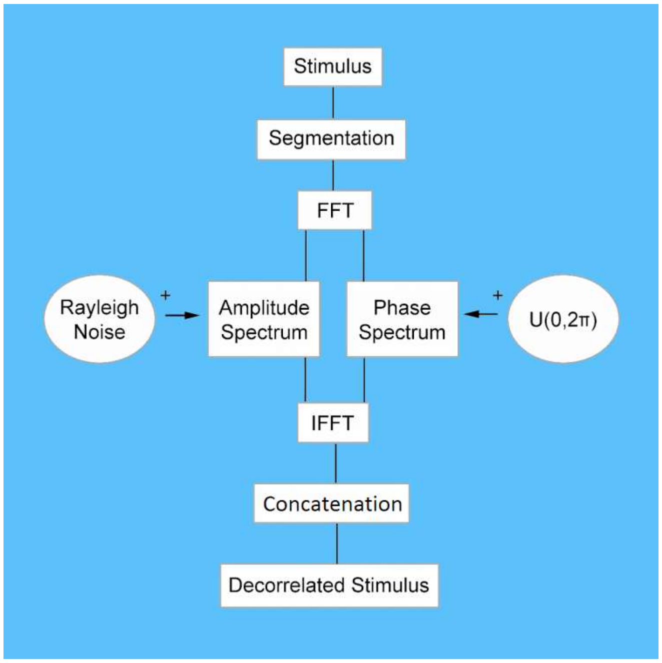

Figure 1.1 Decorrelation Method. Diagram of the method used to decorrelate speech stimuli. Each sentence was divided into segments of equal duration. Each segment was then Fourier transformed, yielding separate amplitude and phase spectra. The phase and amplitude spectra were then independently decorrelated relative to the original by a specific amount. Segments were then inverse Fourier transformed, and concatenated in their original temporal order to form a degraded sentence.

We paired each of the 3 amplitude-spectrum correlation values (0, 0.5, 1) with each phase-spectrum correlation value (0.4, 0.6. 0.8, 1), creating 12 unique (amplitude x phase) conditions. Based on pilot data we determined that these values would be most informative for investigating intelligibility as they provided a wide range of performance levels. Because we were particularly interested in looking at the effects of the phase spectrum, as it has not been studied as extensively as the effects of the amplitude spectrum on intelligibility, we selected a greater number of phase spectrum values. Our pilot study showed that the lower bound of 0.4 for phase-spectrum correlation is adequate since participants were unable to identify any words when the phase-spectrum correlation was below this value. All stimuli were played through HD380 Pro Sennheiser headphones at a sampling rate of 22.05 kHz at an average level of approximately 70 dB SPL (A weighted) measured using a 6-cc coupler, 0.5-inch microphone, and a Precision Sound Analyzer (Brüel & Kjær, Model 2260).

1.2.3 Procedure

Sentences from the HINT database were randomly assigned to each condition and presented to participants in a random order. No sentence was presented more than once per participant. Each subject participated in only one of the three temporal window condition (30 ms, 250 ms, or full length sentence), resulting in a 3 (amplitude correlation) x 4 (phase correlation) x 3 (time window size) mixed-measures experimental design. Five subjects were assigned to each of the three temporal-window conditions, and each subject participated in one experimental session which comprised two blocks of 60 trials that lasted approximately 30 minutes. This resulted in 10 sentences (~40 words) per condition. The experiment was conducted in a double-walled anechoic chamber (Industrial Acoustics Company). Participants were seated at a computer and instructed to listen to each sentence and type as many words as they could understand, ignoring punctuation. Because sentences are semantically meaningful, it is possible that context may provide some cue to word identification. However, use of sentence material to study intelligibility under acoustically degraded conditions is standard practice as such sentences (instead of isolated words) are the type of stimuli most encountered in natural settings. The HINT corpus for example has been used in hundreds of speech intelligibility studies. In addition, subjects were instructed to report words that they were confident about even if it did not make sense semantically because a participant may have misheard an earlier word in the sentence.

There was no time limit for each trial, so participants' typing speed did not affect their ability to perform the task. An experimental run began with 10 practice sentences which were repeated until the subject reported feeling comfortable with the interface and task. The sentences were scored based on individual correct keywords. Potentially confusing verbs ("are/were"), pronouns ("he/she"), prepositions ("in"), conjunctions ("or"), and articles ("the") were excluded from scoring. Average sentence length including non-keywords was 5.3 words, which dropped to 4.1 after exclusions. Total number of correct keywords was compared to total number of keywords for each condition to determine the percent correct for each run. This number represented the degree of intelligibility.

1.3 Results

Fig 1.2 shows average intelligibility scores for each window size as a function of amplitude- and phase-spectrum correlations. Each point is based on 10 sentences (~40 words) per listener (~200 words per point). An intelligibility score of 1 indicates that every subject correctly identified all keywords in all sentences for that condition.

A 3 (amplitude correlation) x 4 (phase correlation) x 3 (time window size) mixed-measures ANOVA showed a significant main effect of amplitude-spectrum correlation ($F(2,24) = 349.21$, $p < .01$) and a significant main effect of phase-spectrum correlation ($F(3,36) = 1231.61$, $p < .01$). No main effect of window size was found ($F(2,12) = .92$, $p = .42$), but there were significant interaction effect between amplitude-spectrum correlation and window size ($F(4,24) = 67.94$, $p < .01$), as well as between phase-spectrum correlation and window size (F(6,36) = 110.69, p < .01). These results suggest that both the effect of amplitude and phase spectrum correlations on speech intelligibility varied by window size. Finally, there was a significant three-way interaction (F(12,72) = 9.28, p < .01), suggesting that the interaction between phase and amplitude correlations was different at different window sizes.

1.3.1 Effects of decorrelation on non-segmented conditions

A 3 (amplitude correlation) x 3 (phase correlation) mixed-measures ANOVA was used to compare the effects of decorrelations on this window size. Note that one of the phase conditions (0.4) was removed from analysis because as shown in Fig 1.2A, intelligibility scores converged to zero at this correlation value even for an amplitude-spectrum correlation of 1. We therefore removed this point from the ANOVA to avoid a misleading significant interaction effect. Both a main effect of amplitude and phase correlation was found (F(2,8) = 59.11, p < .05; F(2,8) = 352.69, p < .01, respectively). A significant interaction was not observed (F(4,16) = X=2.64, p = .07), suggesting that adding phase information did not improve intelligibility more for one level of amplitude correlation than another.

1.3.2 Effects of decorrelation on the 250-ms (syllable length) conditions

A second 3 x 3 mixed-measures ANOVA was calculated to determine the effects of decorrelations on intelligibility specifically for the 250-ms time-window conditions. Similar to the full-length window, there were main effects of both amplitude and phase correlations (F(2,8) = 751.13, p < .05; F(2,8) = 574.87, p < .01, respectively). Unlike the full-length time-window condition, there was a significant interaction effect between amplitude and phase correlations (F(4,16) = 14.44, p < .01). As seen in Fig 1.2B, when amplitude information is partially corrupted (r_α = 0.5), increasing phase-spectrum correlation from 0.6 to 0.8 improves intelligibility scores considerably more than that at other amplitude-spectrum correlations (0 and 1).

1.3.3 Effects of decorrelation on the 30-ms (phoneme length) conditions

Unlike in the previous two window sizes, there was no point of convergence for the 30 ms time-window conditions. Because of this, the 0.4 phase correlation value, which was excluded from analysis as a floor performance level in the prior two conditions (syllable and full length windows), was included in the statistical analysis of the phoneme-length conditions. A 3 x 4 mixed measures ANOVA showed a main effect of both amplitude- and phase-spectrum correlations ($F(2,8) = 167.26$, $p < .01$; $F(8,3) = 61.12$, $p < .01$, respectively). A significant interaction effect was also observed $F(6,24) = 19.54$, $p < .01$) but the form of this interaction is dissimilar to that seen for the 250 ms condition (compare panels B and C of Fig 1.2).

1.4 Discussion

1.4.1 Speech intelligibility for intermediate correlation values

At the most extreme correlation values (0 and 1) our results are consistent with previous studies that have investigated the effects of spectral decorrelation(Kazama et al., 2010; Liu, He, and Palm, 1997; Paliwal and Alsteris, 2005; Alsteris and Paliwal, 2006). However, real speech rarely occurs under perfect conditions, and it is implausible for only one type of spectral component to be degraded outside of laboratory conditions. Therefore, partially degraded amplitude and phase conditions may more accurately represent naturalistic speech environments.

In general, collapsing across window sizes, intelligibility was more adversely affected by phase-spectrum decorrelation than by amplitude-spectrum decorrelation even though both affected intelligibility to some degree. For longer window conditions, when the phase-spectrum was decorreled to 0.4, speech became unintelligible (Fig 1.2 panels A and B). The one phase-condition under which intelligibility seemed unaffected was for $r_\alpha = 1$ at the shortest time window of 30ms (red square symbols of Fig 1.2C). Conversely, when phase-spectrum information is left intact ($r_\theta = 1$) amplitude-spectrum decorrelation has little impact on intelligibility, except for one case, the shortest time window when $r_\alpha = 0$ (blue circles in the Fig 1.2C). If the phase information is left intact, decorrelating the amplitude spectrum to intermediate values has no effect on intelligibility. If the amplitude information is left intact, decorrelating the phase spectrum to intermediate values significantly degrades intelligibility for the longer time windows.

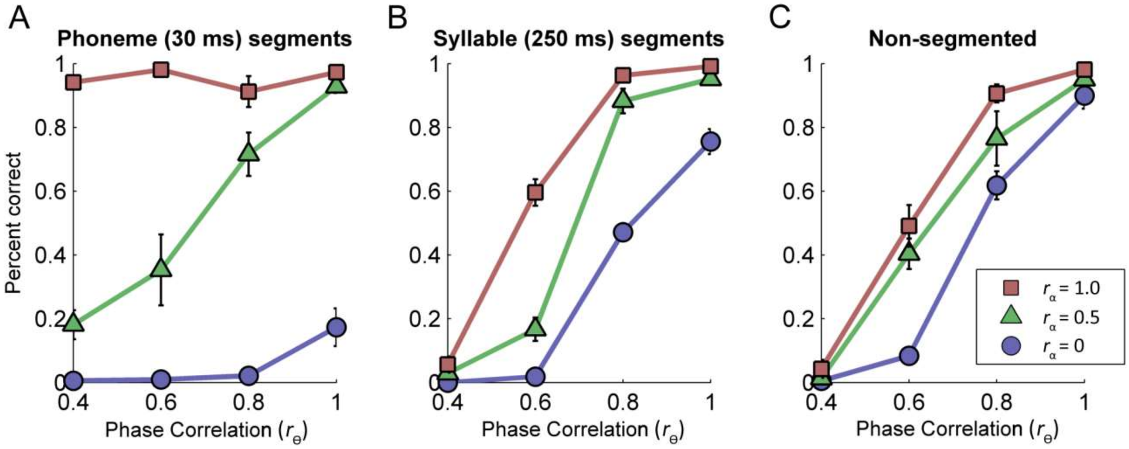

Figure 1.2. Contribution of Phase and Amplitude Spectra to Intelligibility. Speech intelligibility as a function of phase- and amplitude-spectrum decorrelation relative to those of the original unaltered sentence. Each panel depicts results for one of the three temporal window sizes. Each point is calculated from ~40 words per subject. Error bars represent +/- 1 standard error of the mean.

Interestingly, at the short time window (30 ms), phase cues clearly have a major impact on performance at the intermediate amplitude-spectrum correlation (green line, Fig 1.2C). This novel finding is contrary to predictions of prior work that suggests little effect of the phase spectrum at short (phoneme length) time windows. Overall, intermediate correlation values show a significant monotonic effect of phase-spectrum correlation on intelligibility at all time windows (i.e., window size does not matter), a small monotonic effect of amplitude-spectrum correlation for the long time windows and a non-monotonic (interaction) effect of amplitude-spectrum correlation for the short time window.

1.4.2 Equal intelligibility contours

As noted above, in general, the effect of amplitude-spectrum decorrelation increases as window size decreases. Conversely, the effects of phase-spectrum correlation increase as window size increases, but only for extreme correlation values (0 and 1). At an intermediate amplitude-spectrum correlation ($r_\alpha = 0.5$), phase effects seem to be relatively independent of window size (green lines). Our findings suggest, that at least in some cases, there is a tradeoff between the importance of the two cues as a function of temporal window size, though this tradeoff is not necessarily linear. These findings further suggest that there are various combinations of r_θ and r_α that give rise to sets of equal intelligibility contours. Top row of Fig 1.3 shows these contours for the three time windows. A score of 1.0 (dark red) represents perfect intelligibility while dark blue represents an intelligibility score of zero. Note how the slopes of the equal-intelligibility contours increase with window size. The bottom panels of Fig 1.3 show equal-correlation contours between the temporal envelopes of two types of stimuli: 1) the original unaltered sentences and, 2) the same sentences whose phase and amplitude spectra were decorrelated by the values shown along the x-y axes.

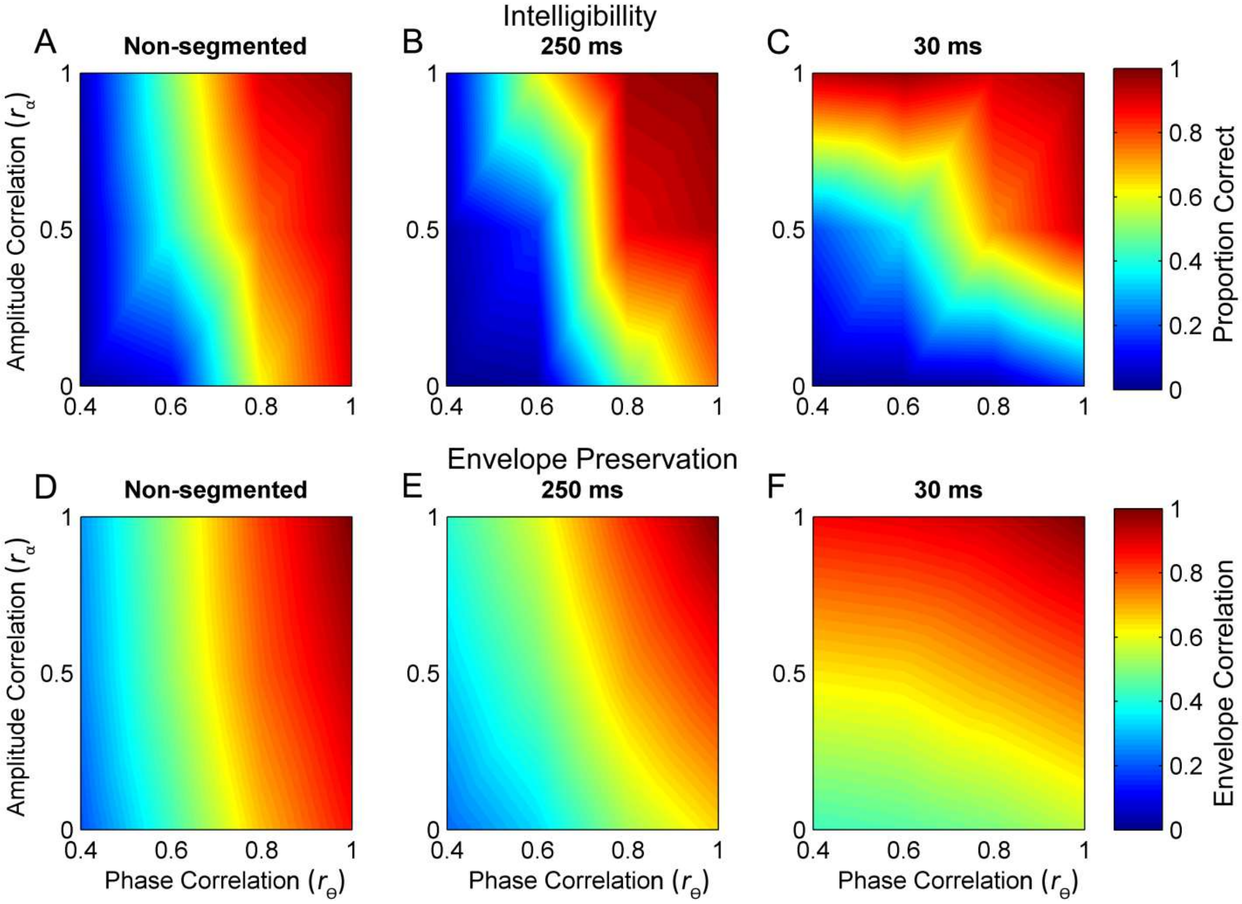

Figure 1.3. Comparison of Equal Intelligibility Contours with Envelope Correlations. Top row shows equal intelligibility contours as a function of phase- and amplitude-spectrum decorrelation. A score of 1.0 (dark red) represents perfect intelligibility while dark blue represents an intelligibility score of zero. (A-C) Equal-correlation contours shown for each of the three window sizes. (D-F) These are the correlations between the temporal envelopes of two types of stimuli centered at 1 kHz: the original unaltered sentences and the same sentences whose phase and amplitude spectra were decorrelated by the values shown along the x-y axes. A score of 1.0 (dark red) represents perfect correlation between the altered and unaltered envelopes while 0 correlation is representes by dark blue.

Note that the bottom panels do not show intelligibility scores (or any other behavioral measure). Rather they show the correlation between the narrowband envelopes of the unaltered and decorrelated sentences, at the output of a filter centered at

1 kHz (simulating the output of a cochlear filter). The reason for filtering at 1 kHz is that, first, the auditory system processes these waveforms not as broadband sounds, but through cochlear filters, and second, because our analysis below (Fig 1.4) demonstrates that the intelligibility performance is best predicted by examining information near the 1-kHz band.

These envelope correlations were calculated using the average values of all sentences in the HINT database. The similarity between equal intelligibility contours (top panels) and equal envelope-correlation contours (bottom panels) suggests that one major cue to intelligibility may be the narrowband temporal envelopes which are degraded more precipitously with phase-spectrum decorrelation than with amplitude-spectrum decorrelation.

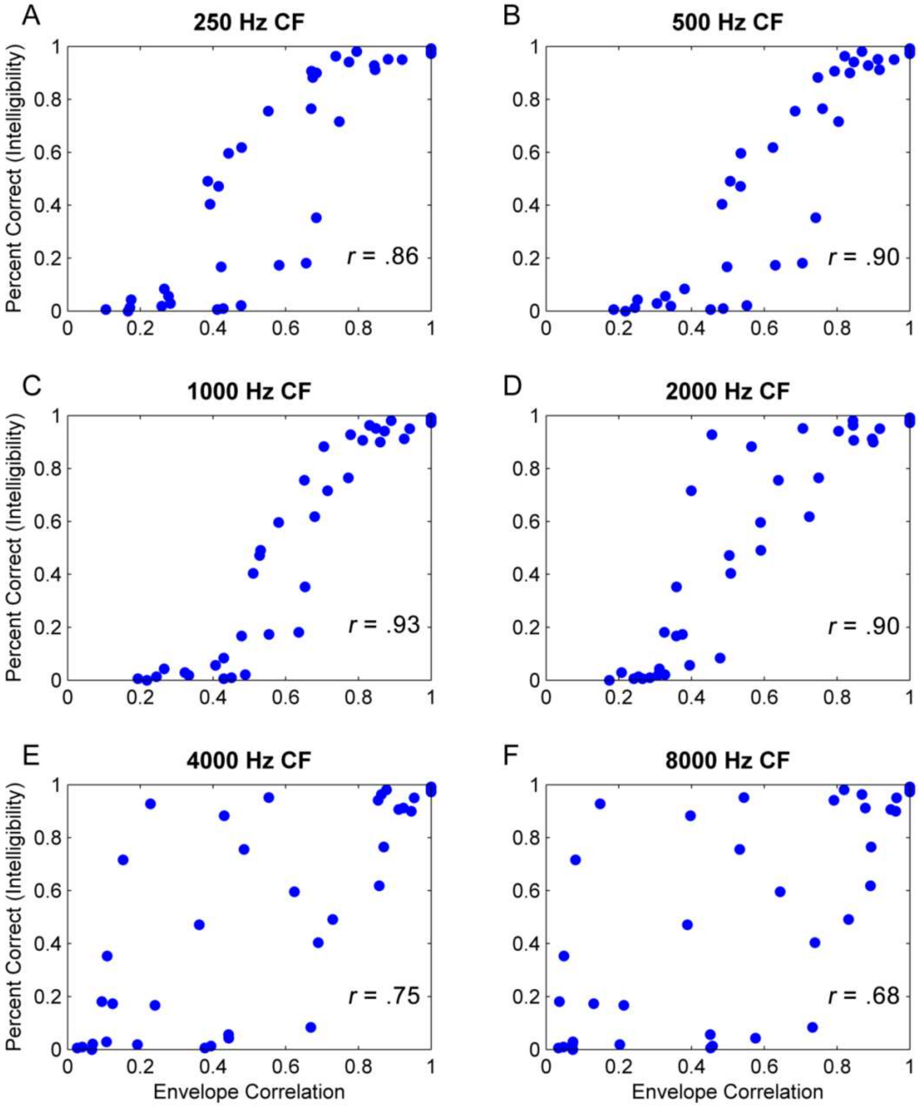

Figure 1.4. Narrow-band Envelope Correlations. Envelope correlations are calculated by comparing the narrow-band envelopes of normal (unaltered) stimuli and the corresponding decorrelated envelopes. Each frequency band determined by a 1/3 octave narrowband Gammatone filter. These correlations were calculated using the average values of all sentences in the HINT database. The correlation value between each frequency band envelope and intelligibility is depicted on the corresponding panel.

Fig 1.4 makes this point clearer by plotting intelligibility scores, collapsed across window sizes, as a function of temporal envelope correlations (i.e., the correlation between the temporal envelopes of the altered and unaltered waveforms at the output of narrowband filters). Each panel shows this analysis for a different filter center frequency:

250, 500, 1000, 2000, 4000, and 8000 Hz. There is a clear relationship between intelligibility and temporal envelope correlation, but only within the lower frequency bands, with virtually no correlation between temporal envelope information and intelligibility at 4 and 8 kHz (Fig 1.4 panels E and F). However, we should qualify that this finding does not mean that speech information may not be extracted from envelopes of filtered waveforms at these higher frequencies, but that given the availability of temporal envelope information at low frequencies, subjects rely primarily on low-frequency cues. The finding that the highest correlation between temporal envelope cues and intelligibility occurs for the 1 kHz band, aligns well with the results of a study by Greenberg and collegues (1998). They suggest that bands in the 750–2350 Hz frequency range carry the most useful intelligibility information despite not containing the most spectral energy. It should be noted that speech is unintelligible when strictly limited to this frequency region, but its intelligibility greatly improves when speech in this band is presented simultaneously with one or more other frequency bands. Furthermore, there is neurological evidence that cortical entrainment to speech occurs primarily at bands in this frequency region (Baltzell et al., 2016).

1.4.3 Spectral and temporal smearing

Spectrograms can be used to visualize the effects of amplitude and phase spectrum decorrelation and help clarify how the decorrelation process degrades temporal and spectral modulations. Fig 1.5 shows one speech sentence at different levels of decorrelation at two window sizes. We can see that amplitude decorrelation (panels B and C) can be thought of as smearing the energy vertically across frequencies, while phase decorrelation (panels D and E) smears the energy horizontally across time.

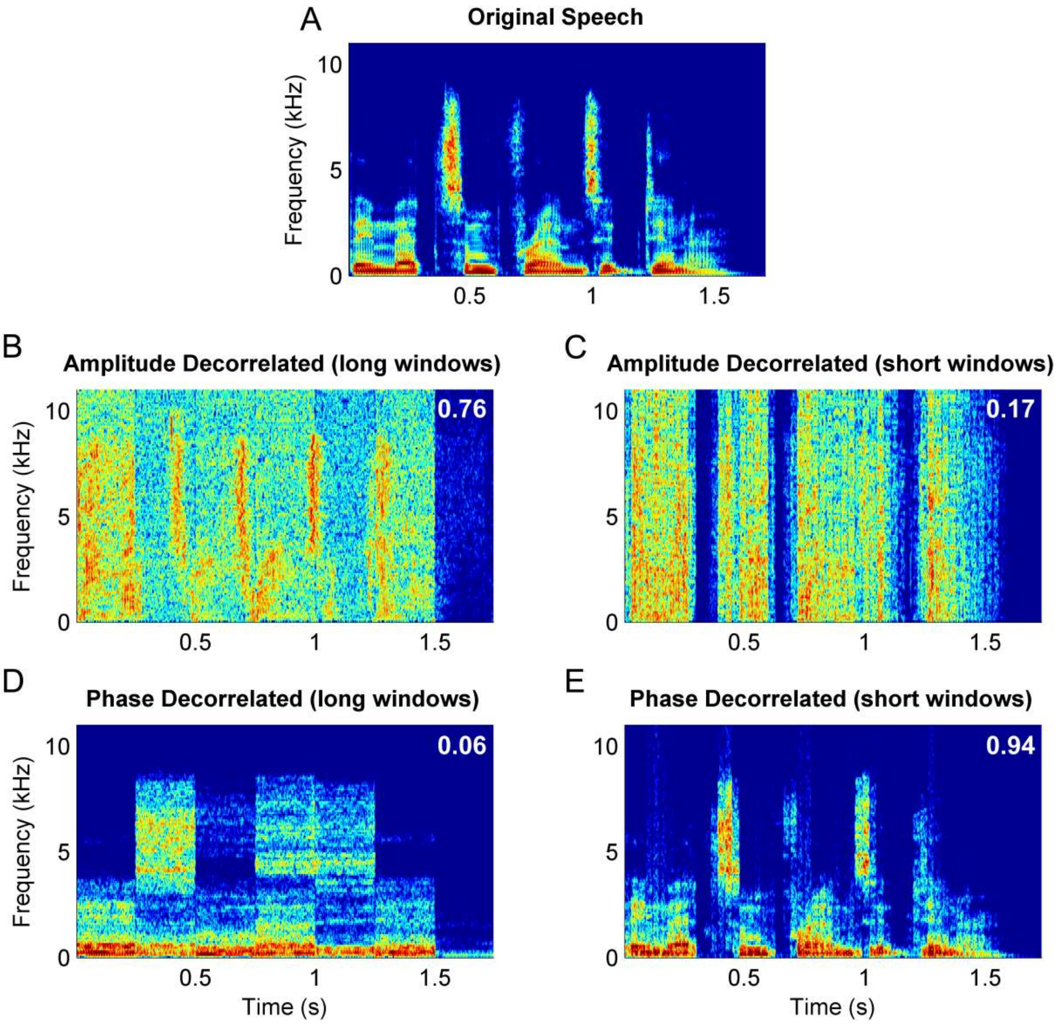

Figure 1.5. Decorrelated speech spectrograms. Spectrograms for the sentence "They met some friends at dinner." (A) Original sentence. (B-C) Amplitude spectrum decorrelated with fully correlated phase spectrum (rα = 0, rθ = 1). (D-E): Phase spectrum decorrelated with unaltered amplitude spectrum (rα = 1, rθ = 0.4). Left panels show spectrograms for 250 ms (syllable length) windows of analysis, and right panels for 30 ms (phoneme length) windows. The average proportion correct for these parameters are listed on each of the panels.

With this in mind, it is clear why phase decorrelation significantly affects the intelligibility of sentences segmented into larger (250 ms) windows but less so the shorter ones (30 ms). Phonemes have a roughly 30 ms duration, and therefore when the energy within a 30 ms window is smeared horizontally, the overall change in the phoneme's energy pattern will be small because it cannot smear as far (it is confined to a brief time window). However, for a 250 ms window length, often encompassing periods of silence as well as several phonemes, smearing along the time axis (horizontally), averages out the energy patterns of several phonemes across time, rendering the speech unintelligible (Fig 1.5D).

Similarly, when the amplitude spectrum is decorrelated in large time windows, it smears energy across frequencies but allows energy fluctuations across time (such as vowel formants or consonant markers) to remain intact. These intact temporal cues preserve formant information, particularly when processed through cochlear filters, and provide sufficient cues to intelligibility. However, when the analysis window becomes too small (30 ms), formants frequency sweeps will become obscured because the sweep is spread across several windows, allowing sections to be averaged to different levels across time (Fig 1.5C).

In summary, the current study investigated how amplitude and phase information differentially contribute to speech intelligibility. We found that intelligibility was more adversely affected by phase-spectrum decorrelation than by amplitude-spectrum decorrelation. If the phase information was left intact, decorrelating the amplitude spectrum to intermediate values had no effect on intelligibility. If the amplitude information was left intact, decorrelating the phase spectrum to intermediate values significantly degraded intelligibility. Interestingly, for intermediate amplitude-spectrum correlation values, segment length was generally inconsequential to intelligibility. These findings provide new insights into how spectral degradation in the phase and amplitude domains affects intelligibility, and demonstrate robustness of the processes that code for speech information in environments that acoustically degrade cues to intelligibility.

CHAPTER 2: Relative role of amplitude- and phase-spectrum cues in music perception

2.1 Introduction

Critical differences in processing of music and speech have become increasingly evident through neuroimaging, behavioral, and clinical population studies (Peretz, et al., 2015; Norman-Haignere, Kanwisher, McDermott, 2015). Electric hearing provides a clear example of these processing differences. Cochlear implants transmit impoverished signals to the auditory cortex, allowing speech to retain most of its intelligibility but rendering music nearly unrecognizable (Limb and Roy, 2014, Zeng; Tang, and Lu, 2014; Gfeller et al., 2007). In order to better understand these processing differences, it is useful to investigate how features of their complex spectra (amplitude and phase) affect the way these two basic types of auditory stimuli are processed.

Phonemes and syllables are considered to be fundamental units of speech. Recent EEG and MEG research has presented evidence of a neural basis for processing ~ 30 and ~250 ms duration sounds, which are typical durations associated with phonemes and syllables respectively (Giraud et al., 2007; Howard and Poeppel, 2010). These studies have demonstrated that the auditory cortex preferentially responds to stimuli with temporal modulations at gamma-band (~20-80 ms) and theta-band (~150-300 ms) rates, suggesting that these rates may represent some basic form of neural parsing or temporal integration

(Luo and Poeppel, 2012). These studies have further suggested that cortical preferences for these temporal segment sizes evolved either around mechanically convenient speech-segment lengths, or because they are optimal oscillatory patterns for processing of complex sounds.

Despite some similarities to the hierarchical structure of language, musical stimuli lack the clear and consistent time-segment lengths observed in speech stimuli. Tempo is usually measured as the number of beats per minute. Most commonly, quarter-notes are designated to define one beat—making quarter notes the most frequently used note—and other durations are specified with respect to the quarter note. Although the durations of phonemes and syllables remain relatively stable, music tempos frequently vary between 80 and 200 bpm, yielding typical quarter-note durations between 300 and 750 ms. While quarter-notes may not be the most frequent notes in some pieces of music, a majority of note durations fall within these temporal limits. Beyond rhythm and tempo, music is also defined by pitch sequences. While pitch is a fundamentally subjective measure, musical pitch is considered to be a measureable metric. It is usually defined using the standard Western tuning system which divides octaves into 12 units—evenly spaced on a logarithmic scale—and defines this distance as a semitone, the smallest musical pitch interval. Speech fundamental frequencies are typically confined to below 300 Hz, but music can contain fundamental frequencies from approximately 16 to 5000 Hz. Both, however, have harmonic structures that extend to higher regions of the spectrum. Musical notes are also much narrower in bandwidth than speech signals, evidenced in their more tonal nature.

There is evidence of music-specific areas of the brain that perform higher-level analysis unrelated to language. For example, people with amusia maintain language abilities but lose music perception abilities such as melody recognition and pitch discrimination (Fedorenko, 2012; Nunes-Silva and Haase, 2011; Peretz, Champod, and Hyde, 2003). However, at a basic acoustic level, speech and melodies are assumed to be processed similarly and it is unclear at what point speech and music processing diverge (Okada et al., 2010; Luo and Poeppel, 2012). If our auditory system is specifically designed for speech processing, as some neuroimaging studies suggest, then it is likely that music is processed in speech unit segments before higher-level analysis.

Behaviorally, the effects of cortical time segmentation on melody recognition or speech intelligibility can be investigated through spectral decorrelation, in which the amplitude and phase components of a stimulus are decorrelated independently relative to the original waveform. To decorrelate spectral components of a signal, it may first be divided into segments of predefined duration (e.g., 30 or 250 ms) and each segment then decorrelated separately prior to concatenation back into the full stimulus in their original sequence. Several speech-intelligibility studies have demonstrated that the length of these segments changes what type of information carries the requisite intelligibility cues (Paliwal and Alsteris, 2005; Liu, He, and Palm, 2007; Kazama et al., 2010; Broussard, Hickok, and Saberi, 2017). When a sentence stimulus is processed in phoneme-length segments, amplitude-spectrum information alone is adequate for intelligibility. For syllable-length segments, phase-spectrum information is necessary for understanding speech. However, these findings are based only on an extreme all-or-none design in which either the phase or amplitude spectrum cues are fully present or absent, i.e., maximum and minimum spectral correlation values (0 and 1). It is likely that at intermediate values, which are more representative of naturalistic environments, the effectiveness of phase- and amplitude-spectrum cues at different temporal segment sizes would be less extreme.

Compared to the large number of studies that have investigated spectral effects on speech intelligibility, there has been little psychophysical exploration into which spectral and temporal cues are necessary for melody recognition. There is limited research on melody processing in general because pitch-based and time-based relations tend to be examined separately (Peretz and Zatorre, 2005). Prior studies have shown that music perception requires temporal fine-structure cues that give rise to pitch, and temporal envelope cues, affected largely by the phase spectrum, that carry information for rhythm identification (Kong et al., 2004). There is currently no research specifically addressing the joint effects of amplitude and phase spectrum cues on melody recognition.

The current study investigates melody recognition ability by presenting listeners with music stimuli whose amplitude and phase spectra have been orthogonally degraded (decorrelated) to various degrees. These results are then compared to prior findings from our laboratory on the effects of phase/amplitude decorrelation on intelligibility of speech (Broussard, Hickok, and Saberi, 2017). Comparison of how perception of melodies and sentences are differentially affected by spectral decorrelation provides important insight into low-level processing differences between these two basic types of complex auditory signals.

2.2 Methods

2.2.1 Participants

Seventeen adult listeners participated in the melody recognition experiment (6 females, M = 24.8 years of age). All participants had normal hearing and were native English speakers. None reported familiarity with the melodies used in this experiment. Six listeners were considered expert musicians (> 10 years of musical training), five identified as amateur musicians (between 3 and 10 years of musical training), and six reported having no formal music training.

2.2.2 Stimuli

Music stimuli were created by taking a melody selected from the Montreal Battery of Evaluation of Amusia (Peretz, Champod, and Hyde, 2003) and adding noise through a decorrelation process (Fig 1.1). First, each stimulus (melody) was divided into one of three time-window sizes: 30 ms, 250 ms, or equal to the duration of the stimulus (i.e., unsegmented). Each segment was then Fourier transformed, yielding separate amplitude and phase spectra. These spectra were then separately decorrelated relative to the original by a specific amount. Decorrelation was achieved by proportionately adding either Rayleigh noise (with scale parameter 1) to the amplitude spectrum or uniform distributed noise (from a range of 0 to 2π) to the phase spectrum. Rayleigh noise was added because the amplitude spectrum of Gaussian noise is Rayleigh distributed. Uniform (0, 2π) noise was added to the phase spectrum using circular statistics methods to achieve a desired degree of phase-spectrum correlation relative to the original unaltered phase spectrum (Fisher, 1996; Berens, 2012). Each windowed stimulus section was then reconstructed as a temporal signal using inverse Fourier transform that combined the new amplitude and phase spectra, and the resultant temporal waveform was normalized to its original segment RMS level. A linear ~2.25 ms rise-decay ramp (100 samples at 44.1 kHz)was imposed on each segment to reduce spectral splatter. The segments were then concatenated in their original order so that each modified stimulus was the same duration as the original

We selected correlation values of 0, 0.5 and 1 for the amplitude spectra and 0.2, 0.4, 0.6, and 1 for phase spectra based on preliminary pilot data, and to allow comparison with prior data from our speech intelligibility study (Broussard, Hickok, and Saberi, 2017). All stimuli were played through HD380 Pro Sennheiser headphones at a sampling rate of 44.1 kHz at an average level of approximately 70 dB SPL (A weighted).

We paired each of the 3 amplitude-spectrum correlation values (0, 0.5, 1) with each phase-spectrum correlation value (0.2, 0.4. 0.6, 1), creating 12 unique (amplitude x phase) conditions. Melodies consisted of short (<5 sec) note sequences played on a piano. The melodies were played in legato style, meaning each note was held until the following note was played but never overlapped. The average note duration was ~450 ms and roughly half of the notes had between 200-300 ms durations. Melodies were played in sequential pairs. The first melody of each pair was one of nine "standard" melodies, and the second was either identical or had one of the following features altered: contour, interval, scale, or rhythm. This resulted in a total of 45 distinct musical melodies. The types of alterations are described below.

A melody's scale describes the notes that can be used in the formation of that melody, which is defined by its characteristic interval pattern and the starting note. A note outside of a melody's scale is noticeable and jarring, even to the untrained ear. Melody contour is defined by the directional relationship of each note to the previous note. A contour is changed by increasing or decreasing the musical pitch of a note (i.e., its fundamental frequency) until its directional relationship with a temporally adjacent note is changed. For example, if Note 1 is a semitone higher than Note 2, and if the frequency of Note 1 is increased, the contour will not change because the directional relationship between Note 1 and Note 2 is the same. However, if Note 1 is instead lowered to a frequency equal to or below Note 2, then the contour is altered because the directional relationship has changed. An interval change is defined as changing the pitch of one note to another pitch within the melody's scale without changing the melody's contour. Scale changes are similar to interval changes, except that the pitch is changed to one outside of the melody's scale. Finally, the rhythm of a melody can be changed by altering the rhythmic value of two or more notes, which maintains the underlying tempo. For example, consider a melody with a rhythm consisting of four notes with the same duration. If the duration of the third note is halved, then one of the adjacent notes will have a duration 1.5 times the original, so that the total time will remain the same.

2.2.3 Procedure

Participants were seated at a computer station in a steel double-walled acoustically isolated chamber (Industrial Acoustics Company) and instructed to listen to pairs of melodies and perform a same-different judgement in a two-interval forced-choice (2IFC) task. Participants were required to determine whether a spectrally decorrelated melody was the same as the preceding sample melody (not spectrally degraded). Each subject participated in one experimental session which comprised six runs of 80 trials. Each run lasted an average of 20 minutes. Participants heard each standard melody approximately

50 times throughout the experiment as the base sample (i.e., interval 1 in the 2IFC task). Half of the trials consisted of two identical melodies, but the second melody of each pair had some degree of spectral decorrelation depending on the selected condition (the correct response on these trials would be "same"). The second melody in the other half of the trials was also spectrally decorrelated and had either a scale, contour, interval, or rhythm alternate (the correct response on these trials would be "different").

Two of the non-musician subjects were unable to perform the melody task with 100% accuracy even when there was no spectral degradation. These subjects were kept in the analysis to rule out musical ability and training as confounds.

2.3 Results

Left panels of Fig. 2.1 show melody discrimination performance as a function of phase- and amplitude-spectrum decorrelation. For comparison, results from (Broussard, Hickok, and Saberi, 2017) on speech intelligibility are plotted in the right panels of Fig. 2.1. Each row of panels shows results for a different time-segment condition (30 ms, 250 ms, non-segmented). To clarify, the lower bound on performance in melody discrimination is 0.5 (proportion correct) because of the same-different 2IFC design used. The lower-bound on performance in the speech intelligibility task is 0 as the dependent measure was the number of keywords correctly identified in a spoken sentence. We therefore emphasize differences in the patterns of performance, instead of absolute values measured, in comparing performance across music and speech tasks. In addition, in reporting these results, we have not distinguished between different types of melody alterations (e.g., contour, scale…) as our main interest was a general comparison between music and speech, and because an analysis of different types of melody change would require a significantly larger number of conditions beyond the scope of the current study.

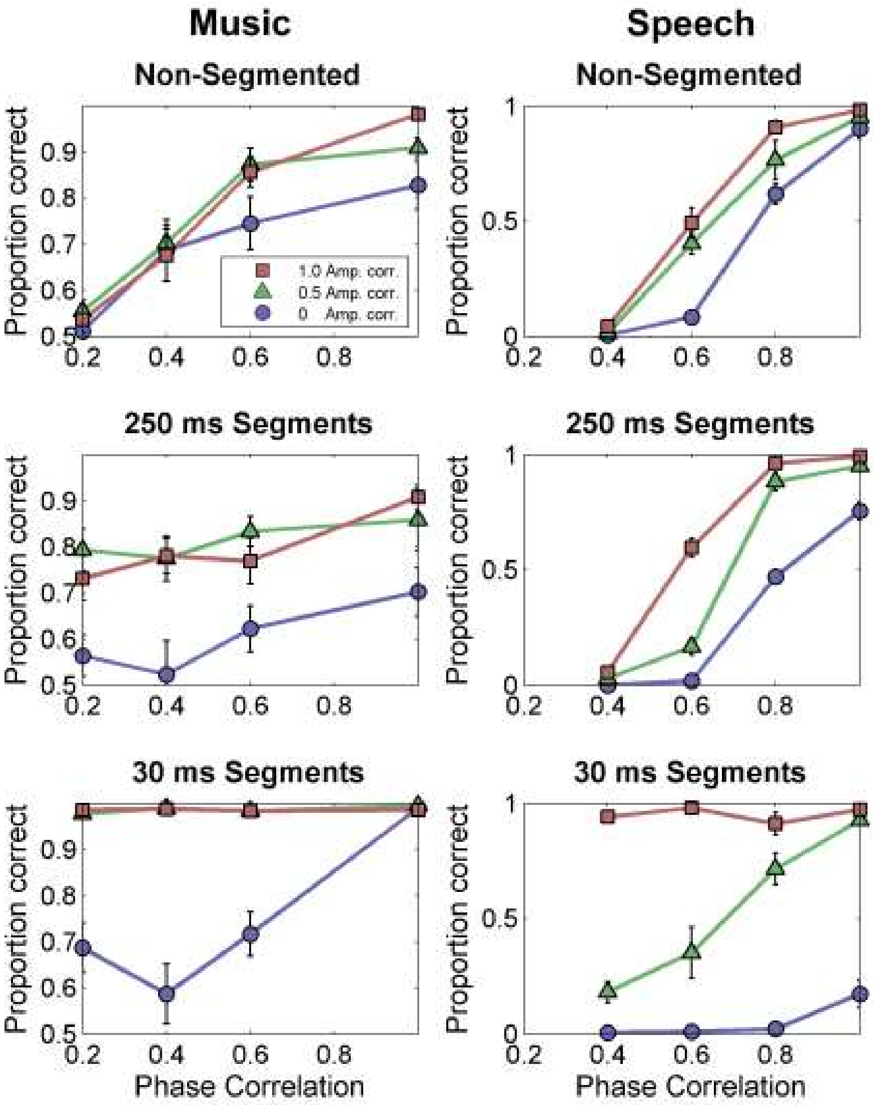

Figure 2.1. Melody discrimination (left panels) and speech intelligibility performance (right panels) as a function of phase- and amplitude-spectrum decorrelation at each of 3 time-windows of analysis (rows of panels). Error bars are +/- one 1 standard error.

A 3 (amplitude correlation) x 4 (phase correlation) x 3 (time window size) mixed measures factorial ANOVA was performed to compare melody identification across conditions. A similar analysis had been conducted on the speech intelligibility data and those statistics are described here for comparison. The main effect of amplitude-spectrum decorrelation was significant for both music and speech ($F(2,28) = 59.33$, $p < .05$, and $F(2,24) = 349.21$, $p < .05$, respectively). The main effect of phase-spectrum decorrelation was also significant for both music and speech ($F(3,42) = 36.27$, $p < .05$, and $F(3,36) = 1231.61$, $p < .05$). No overall main effect of window size was found in the speech task ($F(2,12) = .92$, $p = .42$), indicating that in general, participants found speech at all time windows equally intelligible when collapsing across all conditions (correlation values). However, for music we observed a significant effect of window size ($F(2,14) = 17.82$, $p < .05$). We also observed significant interaction effects between amplitude-spectrum correlation and window size for both music ($F(4,28) = 5.24$, $p < .05$) and speech ($F(4,24) = 67.94$, $p < .05$). Similarly, there was an interaction effect between phase-spectrum correlation and window size for both music and speech ($F(6,42) = 7.24$, $p < .05$ and $F(6,36) = 110.69$, $p < .05$, respectively). These results show that, overall, the effects of amplitude- and phase-spectrum correlation varied by window size for both speech and music.

An interaction between amplitude-spectrum correlation and phase-spectrum correlation was also observed for music ($F(6,84) = 2.23$, $p = .04$) and speech ($F(6,72) = 21.5$, $p < .05$), suggesting that the effect of amplitude decorrelation changed depending on the phase correlation value. Finally, for both music and speech, we observed a three-way interaction effect between phase, amplitude, and window size ($F(12,84) = 2.99$, $p < .05$ and $F(12,72) = 9.28$, $p < .05$, respectively).

2.4 Discussion

Comparison across conditions reveals a number of interesting patterns. Melody discrimination shows a greater overall resistance to adverse effects of spectral decorrelation, both in phase and amplitude domains. The left panels of Fig. 2.1 show that at a phase-correlation value of 0.4, melodies are nearly always identified at above chance levels. Even at a phase-correlation value of 0.2, and window sizes of 30 and 250 ms, melodies remain largely discriminable, especially when some amplitude-spectrum information is available. Another clear example of this is observed in the data shown in the two bottom panels of Fig. 2.1 (30-ms conditions); note that the two green lines which represent an amplitude-spectrum correlation of 0.5, are very differently affected by phase decorrelation. For this example, as noted earlier, there is no adverse effect of phase correlation on melody recognition (i.e., near-perfect performance) while speech intelligibility is strongly impacted by phase decorrelation.

For speech, reducing phase-spectrum correlation for the two longer-segment conditions degrades intelligibility more dramatically than reducing amplitude-spectrum correlation. The opposite is true for the 30 ms window size, where having a highly correlated-amplitude spectrum is more important to preserving intelligibility than having an intact phase spectrum. These window-size effects on performance are possibly due to how temporal and frequency resolution are affected by window size. Decorrelation of speech segments in long temporal windows will result in poor frequency resolution because of the spectrally dynamic nature of speech (i.e., smearing of spectral details across time), while small windows will preserve precise frequency resolution as phonemes, formants, their transitions, and other frequency cues are not averaged (or smeared) across time. Therefore, amplitude-spectrum cues are more useful at short temporal windows. Phase-spectrum information, however, preserves temporal envelopes within frequency bands, which have been identified as critical to accurate speech intelligibility (Shannon et al., 1995).

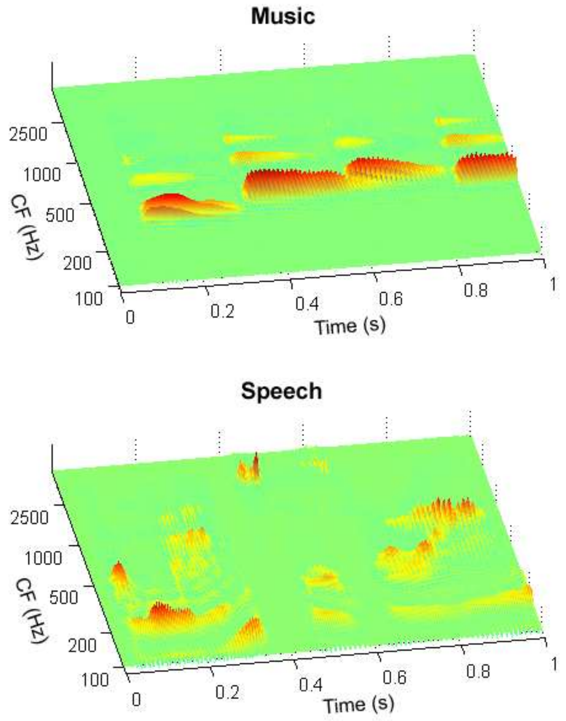

Figure 2.2. Output of a GammaTone filterbank model in response to one- second samples of music and speech stimuli used in the current study.

To better demonstrate these differences, Fig. 2.2 shows the output of a GammaTone filterbank (Holdsworth et al., 1988) which simulates the response of the auditory periphery in response to two brief segments of melody and speech sounds used in our study (top and bottom panels respectively). We selected brief 1-second segments to better observe the differences in spectro-temporal patterns of speech and music. The model comprised 30 bandpass filters with center-frequencies that were logarithmically spaced from 100 to 5000 Hz. Filter bandwidths were based on human auditory filter estimates measured in notched noise (Glasberg and Moore, 1990). Several distinct patterns are observed. Musical notes are confined to a much narrower frequency band (compared to speech) with redundant cues to note identity at harmonics of the musical note's fundamental frequency. There is, however, far less redundancy in the spectral pattern of speech and a more complex pattern across frequency channels. Furthermore, the pattern of activity across time is more complex and detailed for speech than for music. How are these patterns affected by either phase or amplitude decorrelation as a function of window size?

First, let's consider phase-spectrum decorrelation which affects the waveform's temporal envelope, smearing the pattern of activity across time within a frequency channel, but which does not smear spectral energy across frequency channels. If the analysis window is long (250 ms), smearing across time but not frequency, will not substantially affect note identity or tempo for music stimuli because of the note's narrowband nature and because internote intervals within a frequency channel are quite long (for example, in Fig. 2.2, the first note does not repeat within the entire 1s duration shown). For speech, however, smearing of details within 250 ms windows, even within a frequency channel, results in significant loss of information as important transient details to speech identity are confined to brief periods whose temporal relationship to other proximate parts of speech are critical to intelligibility. This may clearly be seen in the lower panel of Fig. 2.2 where complex temporal patterns are observed within 250 ms windows. If, however, the analysis window is short (30 ms), smearing across time (but not frequency) does not have as much impact on either speech intelligibility or melody recognition because the complex pattern of activity across time is better preserved.

Second, let's consider amplitude-spectrum decorrelation which smears information across frequency channels, but largely (though not completely) leaves temporal envelope information intact within frequency channels. When the analysis window is long (250 ms), modest degradation of amplitude-spectrum cues has virtually no impact on melody recognition because much of the spectral energy at the fundamental frequency is preserved. However, if amplitude-spectrum is fully decorrelated, leaving only phase-spectrum cues intact, then note identity is lost, but melody discrimination performance does not decline to chance because the phase-spectrum still provides cues to rhythm. Speech intelligibility is also mostly unaffected by amplitude-spectrum decorrelation at long windows. If, however, the analysis window is short (30 ms), loss of amplitude-spectrum cues critically degrades speech intelligibility (blue circles in bottom-right panel of Fig. 2.2), whereas loss of phase-spectrum cues has no effect on intelligibility as long as amplitude-spectrum cues are unaltered (red squares, same panel). For melody discrimination, partial loss of amplitude-spectrum cues has no effect on performance, even when phase cues are severely degraded (green triangles, bottom-left panel of Fig. 2.2). Furthermore, if phase-spectrum cues are unaltered, complete loss of amplitude-spectrum cues has no effect on melody discrimination at this short window of analysis (same panel, blue circle at a phase correlation of 1). This is likely because rhythm is preserved, providing sufficient cues to melody discrimination.

2.4.1 Comparison of music and speech processing for moderately degraded stimuli

Normal sounds are rarely heard in perfectly quiet conditions. Rather, they are often altered spectrally and temporally by extraneous environmental sounds and competing speech (or other) signals that decorrelate the amplitude and phase spectra of the target sound. Degrading only one type of spectral component entirely (phase spectrum) while leaving the other intact (amplitude spectrum), as has typically been done in prior research, creates an unusual sound that does not exist outside of a laboratory (Liu, He, and Palm, 2007; Kazama et al., 2010). One of our main motivations behind using a greater range of amplitude and phase correlations was to test stimuli that better reflect those heard in naturalistic environments. By considering midlevel correlation values, we can better understand the types of cues most important to processing sounds typically encountered outside the laboratory.

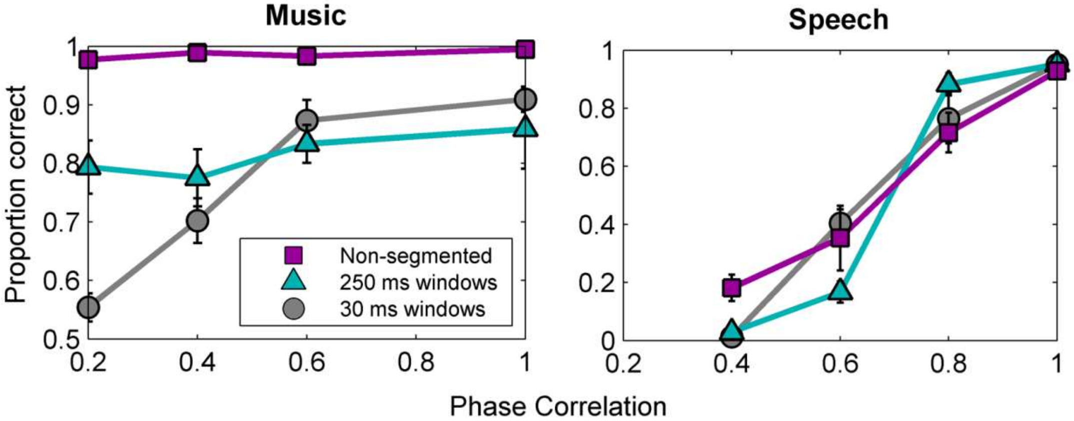

Figure 2.3. Effects of moderate levels of spectral decorrelation on melody discrimination and speech intelligibility. Data are shown only for an amplitude-spectrum correlation of 0.5. The parameter is window size. Error bars are +/- one standard error.

Figure 2.3 shows the effects of moderate levels of spectral decorrelation on melody discrimination and speech intelligibility over different window sizes. A significant effect of window size on melody recognition is observed, whereas very little variation in speech intelligibly across window sizes is evident. This suggests that for sounds whose spectra are partially corrupted, as one may expect to encounter in natural environments, the size of the temporal integration window does not greatly affect speech processing, whereas melody recognition is significantly affect by window size, and that this disparity intensifies as phase correlation is reduced.

2.4.2 Narrow-band envelope preservation

Spectral decorrelation affects both the fine structure and temporal envelope of a waveform. Speech intelligibility has been shown to be affected largely by the degree of narrowband envelope preservation (Kazama et al., 2010; Broussard et al., 2017; Shannon et al., 1995; Alsteris and Paliwal, 2006; Drullman, Festen, and Plomp, 1994). Music discrimination, on the other hand, is unlikely to be strongly correlated with temporal envelope preservation, as evidenced by the difficulties cochlear implant users experience when listening to music (Kong et al., 2004; Smith, Delgutte, and Oxenham, 2002; Jung et al. 2012).

We analyzed narrowband envelope preservation for both melody and speech stimuli to determine the degree to which narrowband envelope cues relate to performance. We calculated the narrowband temporal envelopes of every stimulus used in the current study at each of 12 decorrelation values (3 amplitude by 4 phase values). Each stimulus was first filtered through 1/3 octave wide filters at six center frequencies (0.25, 0.5, 1, 2, 4, and 8 kHz). The temporal envelope at each band was then extracted using the Hilbert transform, and compared to its corresponding unaltered narrowband envelope. Specifically, we measured the Pearson's r correlation between the altered and unaltered narrowband envelopes, resulting in one correlation value per each stimulus at each of the 12 decorrelation conditions. We then calculated the average correlation for each of the six bands across all stimuli for each decorrelation condition. This resulted in 216 correlation values for both the melodies and speech sentences (6 bands x 36 decorrelation conditions).

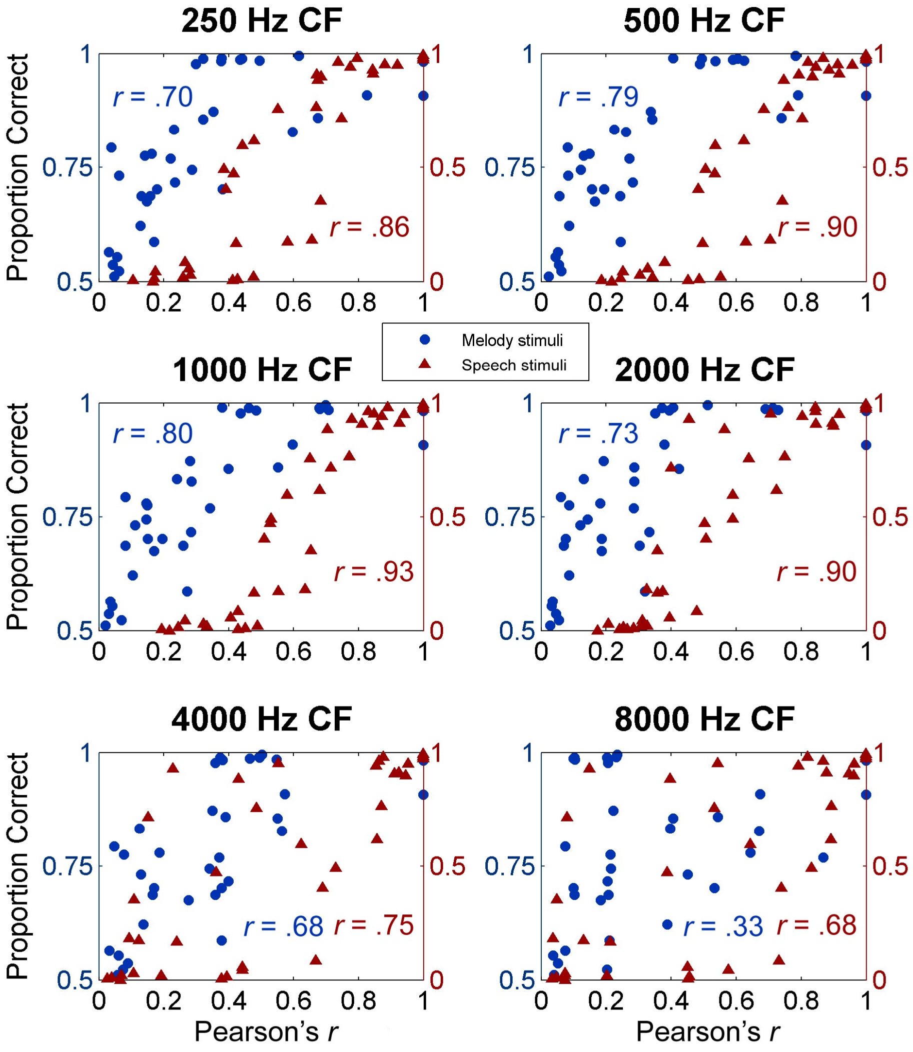

Figure 2.4. Relation between performance and the degree to which narrowband temporal envelopes are preserved within each of 6 different frequency bands (6 panels). The abscissa represents the Pearson's correlation (r) between temporal envelopes of original and altered (spectrally decorrelated) stimuli (see text for details). The abscissa values should not be confused with the r values shown within panels which represent the strength of the relationship between the variables on the x and y axes.

Figure 2.4 shows results of this analysis. As expected, there is a strong correlation between narrowband envelope preservation and speech intelligibility, especially at the 0.5, 1, and 2-kHz bands, with the strength of the correlation decreasing as the center frequency increases. For melody discrimination, there is also a significant contribution of narrowband temporal envelop information to performance, but a major difference is that performance for melody discrimination reaches peak levels (perfect performance) at much lower Pearson r values, i.e., the blue symbols are generally shifted toward the left of the red symbols. This may suggest that other cues, such as fine-structure pitch cues contribute more heavily to melody discrimination than speech intelligibility.

2.4.3 Conclusions

Music and speech are the two most important types of complex sounds with distinct spectro-temporal dynamics processed both by separate and overlapping cortical networks (Stewart et al., 2001; Koelsch et al., 2002). We found significant differences between how phase- and amplitude-spectrum cues contribute to processing of music and speech. Melody discrimination is generally more resilient than speech to degradation of both types of spectra, likely due to the narrow bandwidths of notes and long internote intervals within frequency channels. The effects of spectral degradation, however, is dependent on the temporal window-size of analysis. Prior neuroimaging studies have shown two distinct types of time windows (~30 and ~250 ms) critical to cortical processing of auditory signals (Giraud et al., 2007; Howard and Poeppel, 2010). For long time windows, degradation of phase-spectrum cues has little effect on melody recognition but a significant adverse impact on speech intelligibility. For short windows, phase degradation has no impact on melody discrimination or speech intelligibility as long as the amplitude spectrum remains intact. However, partial degradation of the amplitude-spectrum results in very different patterns of performance for speech and music stimuli at short windows; speech intelligibility is significantly affected by phase-spectrum degradation (i.e., an interaction effect between phase and amplitude spectrum cues), whereas melody discrimination is unaffected by phase degradation (no interaction). In spite of these very different patterns of psychophysical performance for speech and music, it would be of interest to examine in future studies the effects of spectral decorrelation on combined speech-melody stimuli, as in singing, where their joint spectral features may mitigate effects of spectral degradation.

CHAPTER 3: Effects of spectral decorrelation at a variety of speech rates

3.1 Introduction

People have the ability to produce speech over a wide range of rates. Speaking rates have been shown to vary between languages, between speakers, within speakers, and even within a sentence (Pellegrino, Coupé, & Marsico, 2011; Quené, 2008; Quené, 2013; Miller, Grosjean, and Lomanto, 1984; Adank and Janse 2009; Bosker 2016). In an average conversation, speaking rate may vary between 140-180 wpm, and those changes may not be linear across speech components. For example, in natural speech, an increase in speech rate is more likely to affect vowels than consonants (Adank and Janse 2009; Lehiste, 1970; Max and Caruso, 1997). All of these rate fluctuations create challenges for the listener, which become more difficult when the listener is older, has a hearing impairment, or is a non-native speaker of the language (Banai and Lavner 2012; Zhao, 1997; Schneider and Pichora-Fuller, 2001).

Listeners must—and do—have the ability to adapt quickly to unusual speech rates to a large degree. We are able to adjust to small rate changes on the fly, as well as able to adapt to large changes (such as doubling the speaking rate) after only 10 to 20 sentences (Peele and Wingfield, 2005; Adank and Janse, 2009; Dupoux and Green, 1997). It seems counter-intuitive then, that precise temporal information is crucial for optimal speech intelligibility (Shannon et al. 1995; Saberi and Perrott, 1999). There is considerable evidence that the neural tracking of slow (~ 4-8 Hz) temporal envelopes is essential for speech processing and speech intelligibility (Luo and Poeppel, 2012; Howard and Poeppel, 2010) even though the exact function of this neural tracking is unclear (Peele and Davis, 2012). Therefore, adapting to different speech rates requires considerable flexibility in the speech tracking mechanism.

Phonemes are defined by a combination of voicing, manner, and place of articulation. Temporal envelopes carry many of the cues for voice onset time and manner of articulation, which means that these features will be the most affected by changes in speaking rate (Rosen, 1992). Both manner and voicing cues appear in envelope information as differences in rise times (as in 'chip' and 'ship'), long periods of high amplitude for vowels, or as brief silent gaps to indicate a voiceless plosive (Raphael and Isenberg, 1980; Repp et al., 1976; Summerfield et al., 1981). Tempo, which is also primarily transmitted through temporal envelope information, helps to parse sentences. It is also helpful for segmenting word boundaries, even though analysis of a speech signals shows that gaps of silence in the temporal envelope do not always demarcate word boundaries. Similarly, tempo can only provide weak cues to vowel identity due to the covariance of vowel length and vowel identity (Lehiste, 1970).

Most recent studies on speech intelligibility have focused on temporal envelope modulations of speech signals (Shannon et al., 1995; Saberi and Perrott, 1999; Greenberg et al., 2003; Greenburg and Arai, 2004; Ghitza and Greenberg, 2009). However, other studies have focused on how degrading amplitude- and phase-spectrum information affects intelligibility (Broussard, Hickok, and Saberi, 2017, Oppenheim and Lim, 1981; Traunmüller and Lacerda, 1987; Drullman, Festen, and Plomp, 1994; Liu, He, and Palm., 1997; Kazama et al., 2010; Paliwal and Alsteris, 2005; Alsteris and Paliwal, 2006). For example, humans are able to identify vowels using only phase spectrum information at low fundamental frequencies, and speech comprehension has been shown to be more dependent on long-term phase spectrum than amplitude-spectrum information (Kazama et al., 2010; Paliwal and Alsteris, 2005; Alsteris and Paliwal, 2006; Broussard, Hickok, and Saberi, 2017). Liu and colleagues (1997) investigated the impact of the phase spectrum on stop consonants and found that phase cues are required to determine the shape of formant transitions and to help determine voicing properties. Phase degradation has also been reported to make speech in noise recognition more difficult (Shi, Shanechi, and Arabi, 2006), though the interpretation of this finding is confounded by the methods employed since changing SNR (adding noise to speech) will itself degrade the phase spectrum beyond the intentional direct phase degradation, resulting in inaccurate measures of the effects of phase-spectrum degradation on intelligibility.

Several studies have demonstrated that a speech signal will be intelligible regardless of how its spectrum is altered as long as the signal's narrowband temporal envelopes are adequately preserved (Shannon et al., 1995; Saberi and Perrott, 1999; Greenberg et al., 2003; Greenburg and Arai, 2004; Ghitza and Greenberg, 2009; Kazama et al., 2010). These studies suggest that speech is made less intelligible by degrading information in one or both spectral domains (amplitude or phase) primarily because the temporal envelope is consequently degraded by these spectral manipulations. Kazama and collegues (2010) modeled the outputs of peripheral filters in response to speech and determined that the intelligibility of spectrally degraded stimuli was highly correlated with narrowband envelope preservation. These findings suggest that the necessary spectral information for intelligibility is ultimately dependent on the type of information that best preserves the temporal envelope.

Because of the interdependency of spectral and temporal information, it is necessary to consider the effects of the temporal window of analysis on the relative contribution of amplitude and phase spectra to speech intelligibility. Several studies have shown that the type of spectral information that best maintains intelligibility varies by window length (Liu, He, and Palm., 1997; Kazama et al., 2010; Paliwal and Alsteris, 2005; Alsteris and Paliwal, 2006). It has been shown that for time windows shorter than 128 ms, amplitude-spectrum information is most useful to intelligibility. However, for window lengths longer than 128 ms, phase-spectrum information is more critical. This 128 ms crossover point falls almost exactly between the average durations of phonemes and syllables, which have been suggested as basic segments of analysis in speech processing (Greenberg et al., 2003; Harris, 1958). The average lengths of these speech units are ~30 ms and ~250 ms, respectively, and recent EEG and MEG research has presented evidence of a neural basis for these two window sizes in speech perception (Luo and Poeppel, 2012; Giraud et al., 2007; Howard and Poeppel, 2010; Peelle and Davis, 2012; Gilbert and Lorenzi, 2006). These studies have shown that the auditory cortex prefers stimuli with temporal modulations at gamma-band and theta-band rates. Gamma-band frequencies are typically around 40 Hz (~25 ms periods) and theta-band frequencies are typically around 4 Hz (~250 ms periods), which suggests that these may represent some form of neural parsing or temporal integration (Howard and Poeppel, 2010;Luo and Poeppel, 2012).

Clearly, fluctuations in speech rate—whether between or within speakers—will alter the amount of information processed within a particular temporal window. For example, if speech rate is doubled (making it half the original length), then a 100 ms temporal window will provide double the information that it would for a normal rate sentence. The opposite will be true of a sentence that has been expanded in time. Note that in the first example, we would expect the amplitude spectrum of a normal-rate sentence to carry most of the intelligibility cues because the window size is less than the 128 ms crossover point discussed above. However, that same 100 ms window should be primarily affected by phase-spectrum cues for the double-rate sentence. Assuming that the spectral information is identical between the two sentence rates (normal and fast), the spectral analysis of double-rate speech for a 100 ms window will effectively be the same as a spectral analysis of normal-rate speech for a 200 ms window. There are, however, a few important caveats. First, increasing the rate of occurrence of speech cues also doubles the average envelope modulation rate, which at some point will fall outside the range identified as optimum for intelligibility (3 to 8 Hz). Conversely, showing down the speech rate may bring the average envelope modulation rate to below the lower cutoff (3 Hz) and hence degrade intelligibility to below what one may expect from simple linear analysis (Saberi and Perrott, 1999; Peele and Davis, 2012, Luo and Poeppel, 2012). Furthermore, cognitive factors should also be considered in processing intelligibility of speech that has been increased significantly above normal rates, since a potential cognitive bottleneck in information processing could further limit recognition and identification of rapidly presented words.

For the purposes of the current study, we predict that speech rate would be a determining factor of the type of spectral information (amplitude or phase) that carries the most intelligibility cues at a given temporal window size. Time compressing (and expanding) speech provides a useful method for studying the effects of speech rate while controlling for the non-linear compressions that occur naturally when people increase their speaking rate, such as the previous example of shortening consonants more than vowels. The purpose of this study is to investigate the relative contributions of phase and amplitude spectra on sentence intelligibility at three speech rates by independently decorrelating, to various degrees, amplitude and phase spectra relative to those of the original sentence across several time-window sizes. This study will further provide intelligibility scores for a large variety of degraded temporal envelopes, allowing an in-depth analysis of the relationship between spectral and temporal representations of speech stimuli at a wide range of rates.

3.2 Methods

3.2.1 Participants

Five adult listeners participated in each condition of the experiment (3 females, M = 25.6 years of age). All participants had normal hearing and were native English speakers. None of the participants were familiar with the spoken sentences in the HINT (Hearing in Noise Test) database.

3.2.2 Stimuli

We modified speaking rate by artificially lengthening or shortening the original HINT sentences using the overlap-add algorithm implemented in PRAAT (Moulines and Charpentier, 1990; Boersma and Weenink, 2015). This algorithm alters duration uniformly across time within complex stimuli and mimics changes in speaking rate without affecting the pitch contour or the speaker's fundamental frequency. We used a lengthening factor of 2 for the slow-rate speech, which doubled length of the original speech sentence, and used a factor of 0.5 for the fast-rate speech, which compressed the speech to half its original length.

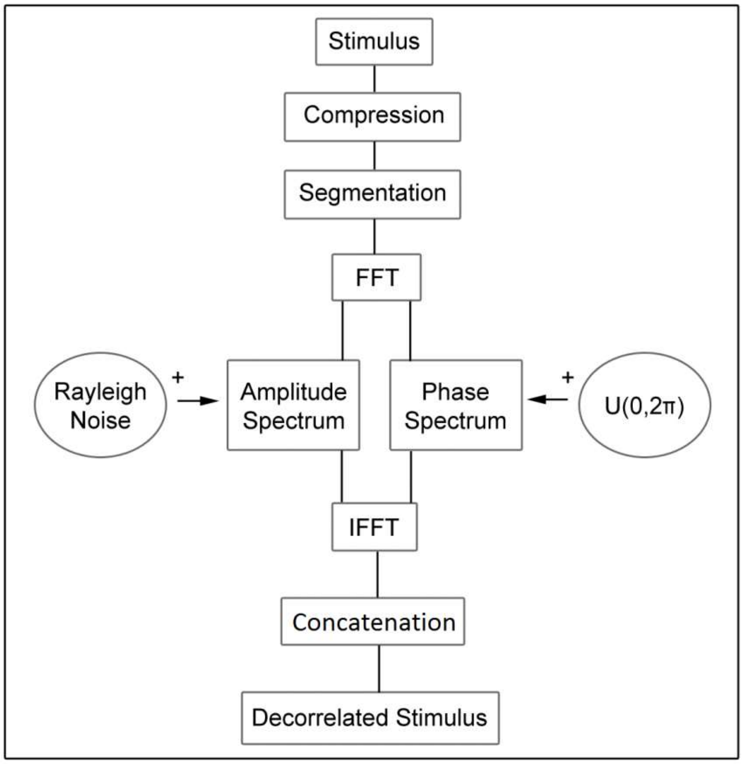

Figure 3.1. Schematic diagram of the method used to generate rate-altered and spectrally decorrelated speech.

Each sentence was then segmented into one of five time-window sizes: 32 ms, 0.62 ms, 128 ms, 192 ms, and 256 ms. These durations were chosen to replicate the window sizes used in a previous experiment by Kazama et al. (2010). Each segment was then Fourier transformed, yielding separate amplitude and phase spectra. These spectra were then separately decorrelated relative to the original by a specific amount. Decorrelation was achieved by proportionately adding either Rayleigh noise to the amplitude spectrum or uniform distributed noise (from a range of 0 to 2π) to the phase spectrum. Rayleigh noise was added because the amplitude spectrum of Gaussian noise is Rayleigh distributed. Uniform (0, 2π) noise was added to the phase spectrum using circular statistics methods to achieve a desired degree of phase-spectrum correlation relative to the original unaltered phase spectrum (Berens, 2009; Fisher, 1995). Each windowed stimulus section was then reconstructed as a temporal signal using inverse Fourier transform that combined the new amplitude and phase spectra, and the resultant temporal waveform was normalized to its original segment RMS level. A linear 4.5 ms rise-decay ramp (100 samples at 22.05 kHz) was imposed on each segment to reduce spectral splatter. The segments were then concatenated in their original order so that each modified stimulus was the same duration as the original (Fig 3.1). See (Broussard, Hickok, and Saberi, 2017) for additional technical details. Note that this exact same procedure was applied to all stimuli, including the faster and slower-rate sentences.

We selected two sets of amplitude- and phase- spectrum correlations. In one condition, the amplitude spectra were completely preserved (r_α= 1) while the phase spectra were completely decorrelated (r_θ = 0). The other set of values were the opposite (r_α

= 0, r_θ = 1). All stimuli were played through HD380 Pro Sennheiser headphones at a sampling rate of 22.05 kHz at an average level of approximately 70 dB SPL (A weighted).

3.2.3 Procedure

HINT sentences were randomly assigned to each condition. No sentence was presented more than once per participant. Each subject participated in all three speed conditions, resulting in a 2 (spectrum type) x 3 (speed) x 5 (window size) design comprising three blocks of 80 trials that lasted approximately 30 minutes with each block consisting of a single speech speed. This resulted in 8 sentences per condition for each subject (240 trials divided by 30 conditions).

The experiment was conducted in a double-walled anechoic chamber (Industrial Acoustics Company). Participants were seated at a computer and instructed to listen to each sentence and type as many words as they could understand, ignoring punctuation. Because sentences are semantically meaningful, it is possible that context may provide some cue to word identification. However, use of sentence material to study intelligibility under acoustically degraded conditions is standard practice as such sentences (instead of isolated words) are the type of stimuli most encountered in natural settings. For example, the HINT corpus has been used in hundreds of speech intelligibility studies. In addition, subjects were instructed to report words that they were confident of even if it did not make sense semantically because a participant may have misheard an earlier word in the sentence.

3.3 Results

Fig 3.2 shows average intelligibility scores for each speaking rate (panels) as a function of temporal window size (abscissa), and amplitude- and phase-spectrum correlations (parameter). Each point is based on 8 sentences (~32 words) per listener (~160 words per point). An intelligibility score of 1 indicates that every subject correctly identified all keywords in all sentences for that condition.

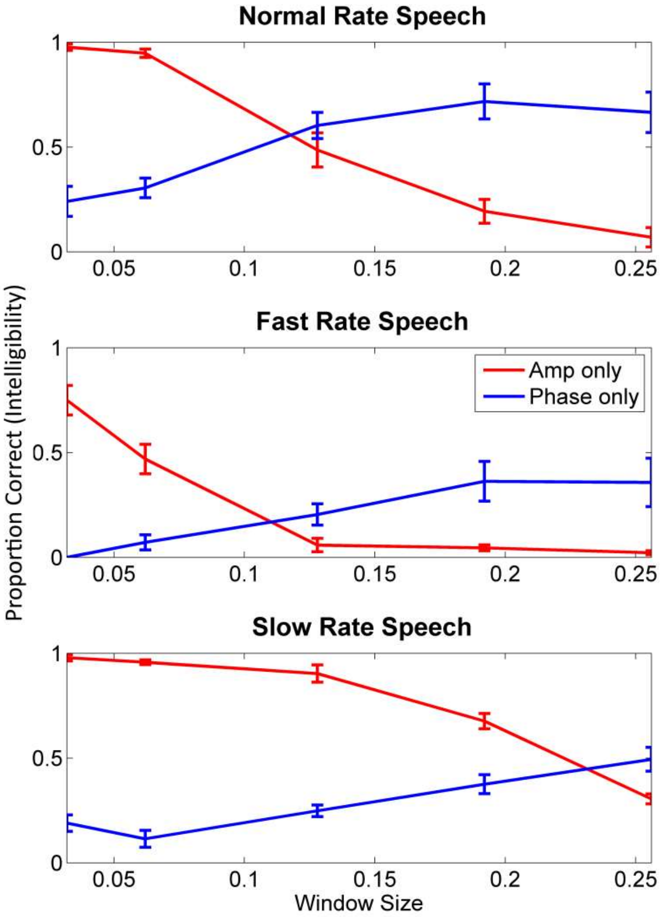

Figure 3.2. Proportion intelligibility scores as a function of speech rate, spectral cue (phase or amplitude), and temporal window size of analysis. Each panel shows results from one speech rate. Parameter is spectral-cue condition (Blue: rα = 0, rθ = 1; Red: rα = 1, rθ = 0). Data are averaged across five listeners. Error bars are +/- 1 standard error.

A 2 (spectra type) x 3 (speed) x 5 (window size) repeated-measures ANOVA showed a significant main effect of spectrum type (F(1,4) = 21.06, p < .01), speaking rate (F(2,8) =

83.68, p < .01), and window size (F(4,16) = 52.00, p < .01). Furthermore, there were significant interaction effects between spectrum type and speaking rate (F(2,8) = 131.39, p < .01), spectrum type and window size (F(4,16) = 132.35, p < .01), as well as speaking rate and window size (F(8,32) = 4.59, p < .01). Finally, there was a significant three-way interaction (F(8,12) = 16.19, p < .01), suggesting that the intelligibility is affected by interaction between spectrum type, window size, and speaking rate.

3.3.1 Effects of speed on the amplitude-phase crossover point

In order to determine the window size at which amplitude-spectrum cues become less useful than phase-spectrum cues for each speech rate condition, we further examined the relationship between window size and rate from a different viewpoint. We graphically estimated the crossover point for each subject and speaking rate, i.e., the point at which the two curves cross each other in each panel of Fig. 3.2. These 3 crossover points are shown in Fig. 3.3, with the error bars represented plus and minus 1 standard error across five subjects. A one-way repeated-measures ANOVA showed a main effect of speech rate (F(2,8) = 120.2, p < .01), which may primarily be attributed to the shift in crossover point in the slow rate condition relative to the other two rates. Post-hoc t-test showed no significant different between the crossover points for the normal and fast rate conditions, though the average fast-rate crossover point is marginally shifted toward smaller temporal window conditions.

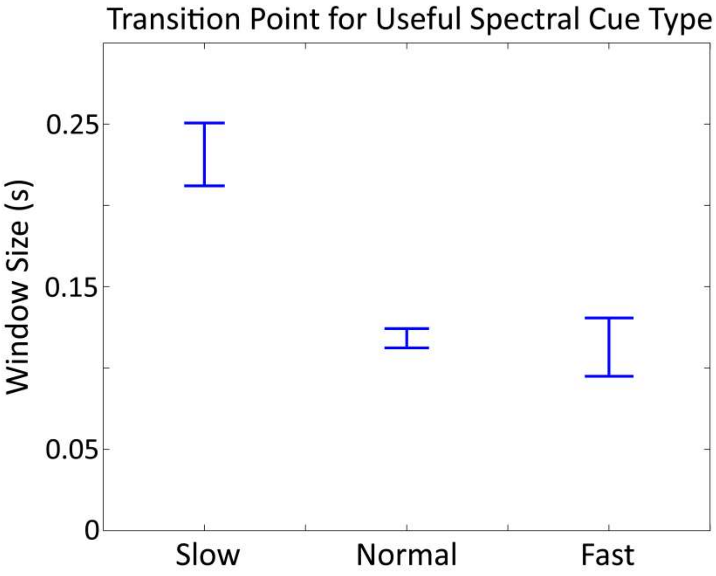

Figure 3.3. Crossover points from Fig. 2. Each point represents the estimated temporal window size at which the two curves in each panel of Fig. 2 cross each other. This is the point at which one type of spectral information becomes more significant than the other in contributing to intelligibility. Each crossover point was estimated separately for each of the five listeners. Error bars are +/-1 standard error across listeners.

3.3.2 Narrow-band envelope correlations

We calculated the average narrow-band temporal envelopes of our stimuli to model the stimulus post-peripheral processing. For every sentence in each condition, we used a Gammatone filterbank to filter the stimuli into six 1/3 octave bands at each of the following center frequencies: 0.25, 0.5, 1, 2, 4, and 8 kHz. We then extracted the temporal envelope of each band using the Hilbert Transfrom, followed by lowpass filtering each envelope at 10 Hz consistent with psychophysical studies that have shown that the most critical cues to speech intelligibility are carried by temporal envelope rates below this cutoff point.

Because intelligibility of normal-rate speech is highly correlated with preservation of temporal envelopes, we compared each degraded sentence envelope with the original sentence envelopes to get a Pearson's *r* correlation value at each frequency band and for every sentence in all of the degraded conditions. We then averaged the *r* values for each condition, resulting in six correlation values (from 6 frequency bands) for each of the 30 conditions (2 spectrum types x 5 temporal windows x 3 speaking rates). This gave us a quantifiable "degree of envelope preservation" for each decorrelation condition at each of the six bands. Finally, we compared the amount of narrowband envelope preservation for each condition with the average proportion correct intelligibility for that condition using a Pearson's *r* to calculate the overall correlation between these two factors (Fig 3.4). Note that this latter measure of Pearson's *r* should not be confused with the correlation value (*r*) noted earlier in comparing temporal envelopes of degraded and original sentences. For the normal-rate condition, the *r* values are 0.82, 0.87, 0.94, 0.93, 0.69, and 0.71 in ascending center frequency order. The values for the slow and fast conditions are 0.65, 0.58, 0.82, 0.95, 0.96, 0.96, and 0.76, 0.75, 0.93, 0.94, 0.82, 0.82, respectively. (These values are shown graphically in Figure 3.5.) Note that the high correlation values are shifted to higher center frequencies for the slow speaking rate compared to normal and fast rates.

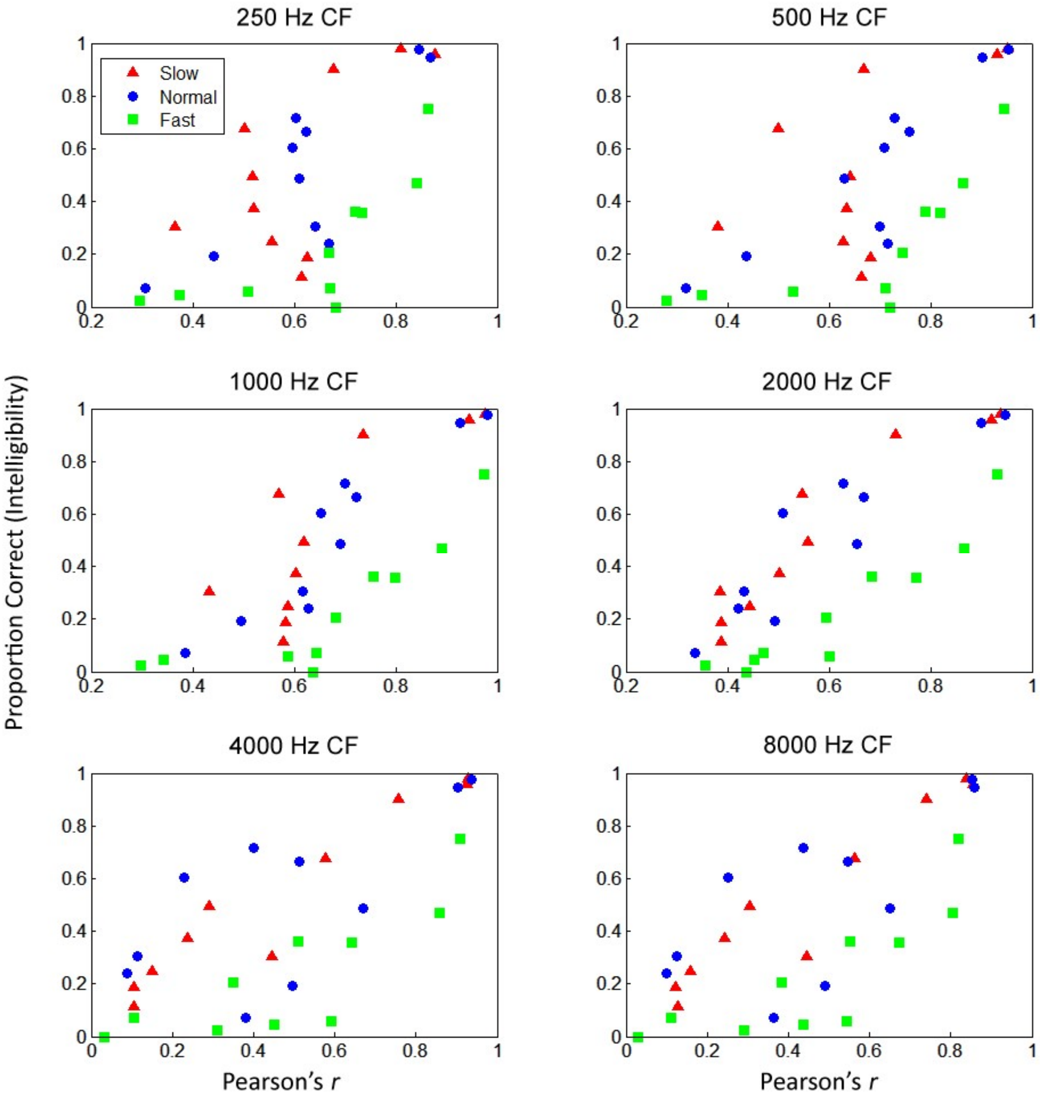

Figure 3.4. Relationship between the amount of narrowband envelope preservation (abscissa) and intelligibility (ordinate). Each panels shows this analysis at a different narrow frequency band. Different symbols (colors) represent different speech rates. Note that within each panel, there are 10 symbols per color. These correspond to the 5 time windows by 2 spectral conditions (see methods). At each of 6 frequency bands, the temporal envelope of a degraded speech sentence was extracted and its correlation with the envelope of the same unaltered sentence at that same frequency band measured (abscissa value). This correlation value, averaged across all sentences of that condition, was then plotted against the average intelligibility score from 5 subjects.

3.4 Discussion

3.4.1 Effects of window size on spectral cues used for intelligibility

Our results for the normal-rate speech closely matched those of Kazama et al. (2010). Consistent with their findings, we observed that listeners use amplitude-spectrum cues for intelligibility when the window size is small, and then transition to phase-spectrum cues for window sizes longer than 120 ms. At short (near phoneme size) time windows, amplitude cues provide enough information for listeners to reach perfect performance. While performance in the phase-only condition never reaches 100% accuracy at any time window size (Fig. 3.2), the results from both Kazama et al. (2010) and our previous work (Broussard et al., 2017) suggest that phase-spectrum cues alone would have been adequate for perfect intelligibility at window sizes longer than the maximum duration tested here. Based on our previous findings (Broussard et al., 2017), we expected phase-spectrum information to be more useful than amplitude-spectrum information over a larger number of window sizes for all speech speeds. However, this was not evident in our current findings since performance never asymptoted to near-perfect levels, possibly because we did not measure intelligibility for window sizes exceeding 256 ms.

The interactions we found further support these findings. If we collapse across window sizes for each speed condition, performance at longer window sizes for fast-rate speech is considerably worse than it is for either slow- or normal-rate speech. Collapsing across spectrum type for each window size, we can see that phase is considerably less useful overall for both of the non-standard rate conditions. The three plots in Fig. 3.2 show how the three-way interaction confirms that the effect of window size on the type of spectral cues that carry the most useful information depends on speech rate.

3.4.2 Phase usefulness for non-standard speech speeds

As noted earlier, we found a main effect of spectrum type for normal speech rate, which may largely be attributed to differences in performance in the phase-only condition at long window sizes. At both faster and slower speech rates, we see a similar pattern to the one seen for the normal rate: amplitude cues are more helpful at shorter window sizes and phase cues become more helpful as window size increases (Fig. 3.2). In both of these cases, however, phase cues never become fully adequate for perfect word identification even at the longest window sizes used. In the slow-rate case, this is not surprising because we expected to see amplitude cues remaining useful over more window sizes. Most likely, phase-spectrum information will become more useful at longer window sizes in the slow condition if we had used a larger range of window sizes.

Interestingly, we also see a decline in usefulness of phase information in the fast-rate speech condition (compared to normal rate), despite expecting phase cues to become more useful at shorter window sizes. In fact, we see that not only does amplitude remain more useful until ~120 ms time windows, listeners were unable to use phase-spectrum-only information even at window sizes where phase information alone had been adequate for normal-rate speech intelligibility. This suggests that amplitude spectrum cues are more robust than phase cues across different speech rates, which is supported by the fact that phase spectrum information is inherently more severely affected by temporal changes.

3.4.3 Failure to shift phase-amplitude crossover point in fast speech

If the importance of amplitude and phase cues for intelligibility is primarily dependent on which cue type best preserves the narrowband temporal envelopes, as Kazama et al. (2010) have suggested, then we would expect intelligibility to continue to track with envelope preservation even for speech with unusual rates. As noted earlier, when normal-rate speech is decorrelated in 100-ms window segments, this should be equivalent to decorrelating half-rate speech in 200-ms windows or decorrelating double-rate speech in 50-ms windows. In other words, the amount of narrowband envelope preservation for a normal-rate sentence decorrelated in 100-ms window segments should be roughly equal to the amount of envelope preservation for a half-rate sentence decorrelated in 200-ms windows and a fast-rate sentence decorrelated in 50-ms windows. Thus, we predicted that the transition point (see Fig. 3.2) where phase-spectrum information becomes more useful than amplitude spectrum in normal-rate speech (~120 ms) would double for slower speech rate (240 ms) and halve for the faster speech rates (60 ms).

Surprisingly, our results were inconsistent with this prediction. Figure 3.3 shows that while we did find a significant difference between the normal- and slow-rate speech, we failed to find a difference between the fast- and normal-rate conditions. Furthermore, although the crossover point for the slow condition shifts significantly relative to that for the normal condition, it does not shift far enough to result in a perfectly doubled window size. While this initially suggested to us that intelligibility was not tracking with envelope preservation as well as in the non-standard speech rates, after a closer analysis of the stimuli's narrowband envelopes, we determined that this was not true. There were still high levels of correlation between envelope preservation and intelligibility (Fig. 4).

3.4.4 Differences in narrow-band envelope preservation

The results from our previous study (Broussard et al. 2017) suggest that the strongest predictor of the most useful type of spectral information (amplitude or phase) for intelligibility is the amount of temporal envelope information it carries. If the narrowband temporal envelopes are intact, specifically bands in the 700 to 2000 Hz region, then the stimulus will be highly intelligible. This is consistent with several prior behavioral findings (Greenburg et al., 1998; Kazama et al., 2010) even though speech energy peaks at lower frequencies. Furthermore, there is evidence that neural entrainment is greater in these middle-frequency bands when a listener is attending to a sentence (Baltzell et al. 2015). Increased neural entrainment is usually correlated with better intelligibility (Luo and Poeppel, 2007; Ahissar et al., 2001) and it has been shown that degraded temporal envelopes result in decreased neural entrainment, specifically as seen in auditory evoked potentials (Nourski et al., 2009). Therefore, it is reasonable for us to assume that if narrowband envelopes in our stimuli are preserved, this likely indicates a greater amount of neural entrainment to that particular band.

Figures 3.4 and 3.5 depicts the relationship between the amount of narrowband envelope preservation for each of the 30 conditions and that condition's average intelligibility score at each of six frequency bands. Interestingly, while this relationship for the normal-speed speech stimuli is consistent with our previous findings: relatively high correlations across all frequency bands, with the greatest correlations occurring between 500 to 2000 Hz, both the slow- and fast-rate speech conditions behave differently. The fast condition peaks between 1000 and 2000 Hz, similar to the normal condition, but the higher bands appear to carry more useful information in the fast condition. In the slow condition, the change is more obvious: the peak in the correlations occurs at the highest frequency bands (>2000). These results suggest that listeners may be relying on information in (and likely entraining to) these higher frequency bands more when the speech rate is unusually fast or slow.

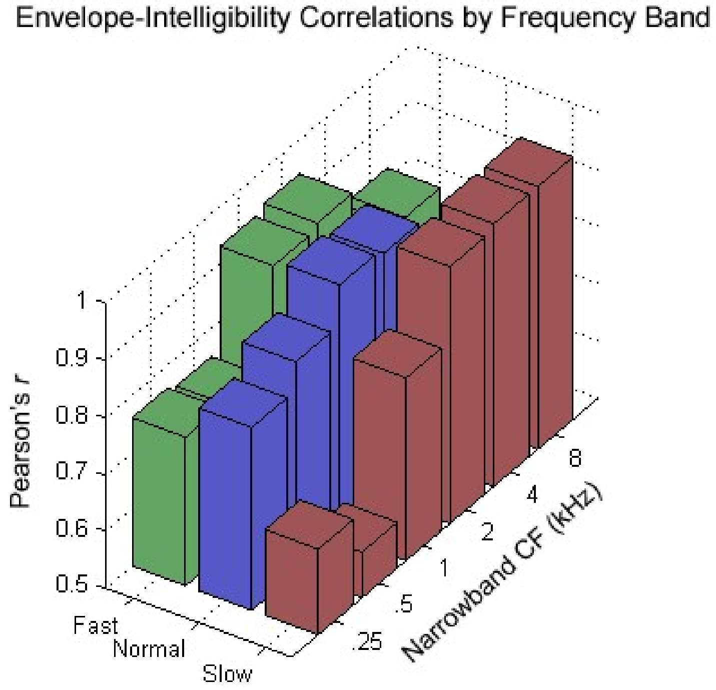

Figure 3.5. Correlations between intelligibility and narrowband envelope for each speech speed. The Pearson's *r* values for each of the six panels of Figure 4 are depicted here on the z-axis for each speech rate.

These findings suggest that participants relied on different cues for speech with unusual rates. Additionally, because older listeners often have poor high-frequency hearing, the necessity of envelope preservation in high-frequency bands may partly explain why older adults have more difficulty understanding unusual speech rates than younger adults.

CONCLUSION

By independently decorrelating the amplitude and phase spectra of multiple types of stimuli, these three experiments evaluated the relative importance of different types of spectral cues to the processing of speech and music. Chapter One investigated how amplitude and phase information differentially contribute to speech intelligibility. Listeners performed a word-identification task after hearing spectrally degraded sentences. Each stimulus was degraded by first dividing it into segments, then the amplitude and phase components of each segment were decorrelated independently to various degrees relative to those of the original segment. Segments were then concatenated into their original sequence to present to the listener. We used three segment lengths: 30 ms (phoneme length), 250 ms (syllable length), and full sentence (non-segmented). We found that for intermediate spectral correlation values, segment length is generally inconsequential to intelligibility. Overall, intelligibility was more adversely affected by phase-spectrum decorrelation than by amplitude-spectrum decorrelation. If the phase information was left intact, decorrelating the amplitude spectrum to intermediate values had no effect on intelligibility. If the amplitude information was left intact, decorrelating the phase spectrum to intermediate values significantly degraded intelligibility, with a few exceptions. These results delineate the range of amplitude- and phase-spectrum correlations necessary for speech processing and its dependency on the temporal window of analysis (phoneme or syllable length). Results further pointed to the robustness of speech information in environments that acoustically degrade cues to intelligibility (e.g., reverberant or noisy environments).

In Chapter Two, we investigated how amplitude and phase information differentially contribute to music recognition and speech intelligibility. Listeners heard degraded melodies and performed a same-different judgement in a melody-discrimination task. Each waveform was first temporally segmented (30 ms, 250 ms, or unsegmented) and the amplitude and phase spectra of each segment independently decorrelated to different degrees relative to their original unaltered spectra to generate decorrelated melodies. We compared findings from this study to the results in Chapter One which investigated the effects of spectral decorrelation on speech intelligibility and found that, compared to speech, melody recognition is more resilient to loss of phase-spectrum cues due to relatively long internote intervals within frequency channels. This robustness was observed both for short and long segment durations. Melody recognition was also relatively unaffected by partial decorrelation of the amplitude spectrum. Conversely, we found a greater decline in speech intelligibility from loss of phase-spectrum cues as speech recognition is heavily reliant on temporal envelope structure. For short-duration segments, partial decorrelation of the amplitude spectrum had a major impact on speech intelligibility but no effect on melody recognition. Further analysis showed that melody discrimination is primarily affected by degradation of the waveform fine structure, whereas speech intelligibility can be largely explained by preservation of temporal envelopes within frequency channels.

The third chapter examined the effects of speaking rate and spectral degradation on speech intelligibility. Five normal-hearing subjects listened to spoken sentences that were either sped up to twice that of normal rate or slowed down to half normal rate. Sentences were first segmented into analysis windows ranging from 32 to 256 ms. Each segment was then spectrally degraded by either replacing its amplitude spectrum with that of Gaussian noise, leaving its phase spectrum intact, or vice versa. Consistent with prior findings, phase-spectrum cues were most useful to intelligibility at longer temporal windows of analysis, and amplitude spectrum cues at short windows. For standard rate speech, the crossover point between these two cues occurred at an estimated window size of 120 ms. Increasing speaking rate to twice normal rate, surprisingly seemed to have little to no effect on this crossover point. However, slowing down speaking rate shifted this crossover point to significantly longer temporal window sizes (~230 ms), slightly smaller than twice the crossover point for normal speaking rates. Analysis of narrowband envelopes show that for non-standard speech rates, higher frequency bands are more useful for intelligibility than they are for standard speech rates.

Collectively, the set of three studies described in this dissertation demonstrate a different approach to investigating how complex auditory stimuli may be processed. While temporal envelope is useful for predicting intelligibility in standard speech stimuli, it is not sufficient in accurately determining intelligibility when considering non-standard speech rates and other complex, meaningful stimuli such as musical phrases. The current studies contributed to our understanding of speech intelligibility and music recognition at different spectral decorrelation values under a variety experimental conditions. Additionally, because this method allows for precisely defined stimulus degradation levels in the spectral domain, it provided a technique for generating a range of stimuli with degraded temporal envelopes such that one could more accurately quantify the contribution of different frequency bands to discriminability of any type of complex auditory stimulus.

REFERENCES

Adank, P., & Devlin, J. T. (2010). On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. Neuroimage, 49(1), 1124-1132.

Alsteris, L. D., & Paliwal, K. K. (2006). Further intelligibility results from human listening tests using the short-time phase spectrum. Speech Communication, 48(6), 727–736. doi:10.1016/j.specom.2005.10.005

Baltzell, L.S., Horton, C., Shen, Y., Richards, V.M., D'Zmura, M., & Srinivasan, R. (2016). Attention selectively modulates cortical entrainment in different regions of the speech spectrum. Brain research. Aug 1;1644:203-12.

Banai, K., & Lavner, Y. (2012). Perceptual learning of time-compressed speech: More than rapid adaptation. PloS one, 7(10), e47099.

Berens, P. CircStat: a MATLAB toolbox for circular statistics. J Stat Software. 2009 Sep 23;31(10):1-21. DOI: 10.18637/jss.v031.i10.

Boersma, P., & Weenink, D. (2015). Praat version 5.4. 08. Doing phonetics by computer.

Bosker, H. R. (2017). Accounting for rate-dependent category boundary shifts in speech perception. Attention, Perception, & Psychophysics, 79(1), 333-343.

Broussard, S., Hickok, G., & Saberi, K. (2017). Robustness of speech intelligibility at moderate levels of spectral degradation. PloS one, 12(7), e0180734.

Crystal, D. (2010), The Cambridge Encyclopedia of Language (3rd ed.), Cambridge.

Dorman, M. F., Loizou, P. C., Fitzke, J., & Tu, Z. (1998). The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. The Journal of the Acoustical Society of America, 104, 3583.

Drullman, R. (1995). Speech intelligibility in noise: relative contribution of speech elements above and below the noise level. The Journal of the Acoustical Society of America, 98(3), 1796-1798.

Drullman, R., Festen, J.M., & Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. The Journal of the Acoustical Society of America. Feb;95(2):1053-64.

Dupoux E. & Green K. (1997) Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. Journal of Experimental Psychology-Human Perception and Performance 23: 914–927

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. PLoS computational biology, 5(3), e1000302.

Fedorenko, E., McDermott, J. H., Norman-Haignere, S., & Kanwisher, N. (2012). Sensitivity to musical structure in the human brain. Journal of neurophysiology, 108(12), 3289-3300

Fisher, N.I. Statistical Analysis of Circular Data. Cambridge, UK: Cambridge University Press; 1995.

Friesen, L. M., Shannon, R. V., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. The Journal of the Acoustical Society of America, 110(2), 1150. doi:10.1121/1.1381538

Fry, D.B. Prosodic phenomena. Manual of phonetics. 1968:365-410.

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), 113-126.

Gilbert, G. & Lorenzi, C. (2006). The ability of listeners to use recovered envelope cues from speech fine structure. The Journal of the Acoustical Society of America. Apr;119(4):2438-44.

Giraud, A.L., Kleinschmidt, A., Poeppel, D., Lund, T.E., Frackowiak, R.S., & Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. Neuron. Dec 20;56(6):1127-34.

Giraud, A.L., Poeppel, D. Cortical oscillations and speech processing: emerging computational principles and operations. Nature neuroscience. 2012 Apr 1;15(4):511-7.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. Hearing research, 47(1), 103-138.

Gnansia, D., Jourdes, V., & Lorenzi, C. (2008). Effect of masker modulation depth on speech masking release. Hearing research, 239(1), 60-68.

Greenberg, S., & Arai, T. (2004). What are the essential cues for understanding spoken language?. IEICE TRANSACTIONS on Information and Systems, 87(5), 1059-1070.

Greenberg, S., Arai, T., & Silipo, R. (1998). Speech intelligibility derived from exceedingly sparse spectral information. InICSLP Dec.

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

Gfeller, K., Turner, C., Oleson, J., Zhang, X., Gantz, B., Froman, R., & Olszewski, C. (2007). Accuracy of cochlear implant recipients on pitch perception, melody recognition, and speech reception in noise. Ear and hearing, 28(3), 412-423.

Harris, K.S. (1958). Cues for the discrimination of American English fricatives in spoken syllables. Language and speech. Jan;1(1):1-7.

Hazan, V. & Rosen, S. Individual variability in the perception of cues to place contrasts in initial stops. Attention, Perception, & Psychophysics. 1991 Mar 1;49(2):187-200.

Holdsworth, J., Nimmo-Smith, I., Patterson, R., & Rice, P. (1988). Implementing a gammatone filter bank. Annex C of the SVOS Final Report: Part A: The Auditory Filterbank, 1, 1-5.

Howard, M.F. & Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. Journal of neurophysiology. Nov 1;104(5):2500-

Jung, K. H., Won, J. H., Drennan, W. R., Jameyson, E., Miyasaki, G., Norton, S. J., & Rubinstein, J. T. (2012). Psychoacoustic performance and music and speech perception in prelingually deafened children with cochlear implants. Audiology and Neurotology, 17(3), 189-197.11.

Kazama, M., Gotoh, S., Tohyama, M., & Houtgast, T. (2010). On the significance of phase in the short term Fourier spectrum for speech intelligibility. The Journal of the Acoustical Society of America, 127(3), 1432. doi:10.1121/1.3294554

Koelsch, S. (2011). Toward a Neural Basis of Music Perception – A review and updated model. Frontier in Psychology, 2. doi:10.3389/fpsyg.2011.00110

Koelsch, S., Gunter, T. C., Cramon, D. Y. V., Zysset, S., Lohmann, G., & Friederici, A. D. (2002). Bach speaks: a cortical “language-network” serves the processing of music. Neuroimage, 17(2), 956-966.

Kong, Y.-Y., Cruz, R., Jones, J. A., & Zeng, F.-G. (2004). Music perception with temporal cues in acoustic and electric hearing. Ear and Hearing, 25(2), 173–185. doi:10.1097/01.AUD.0000120365.97792.2F

Lehiste, I. Suprasegmentals. Cambridge, Massachusetts: MIT Press; 1970.

Limb, C. J., & Roy, A. T. (2014). Technological, biological, and acoustical constraints to music perception in cochlear implant users. Hearing research, 308, 13-26.

Liu, L., He, J., & Palm, G. (1997). Effects of phase on the perception of intervocalic stop consonants. Speech Communication, 22(4), 403–417.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. Proceedings of the National Academy of Sciences, 103(49), 18866–18869.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron, 54(6), 1001–1010. doi:10.1016/j.neuron.2007.06.004

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. doi:10.3389/fpsyg.2012.00170

Max, L., and Caruso, A. J. (1997). “Acoustic measures of temporal intervals across speaking rates: Variability of syllable- and phrase-level relative timing,” J. Speech Lang. Hear. Res. https://doi.org/JSLRFW 40, 1097–1110.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), Perspectives on the study of speech (pp. 39–74). Hillsdale, NJ: Erlbaum.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech communication, 9(5-6), 453-467.

Nilsson, M., Soli, S.D., & Sullivan, J.A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. The Journal of the Acoustical Society of America. Feb;95(2):1085-99.

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron, 88(6), 1281-1296.

Nunes-Silva, M., & Haase, V. (2013). Amusias and modularity of musical cognitive processing. Psychology and Neuroscience, 6(1), 45–56. doi:10.3922/j.psns.2013.1.08

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I. H., Saberi, K., & Hickok, G. (2010). Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cerebral Cortex, 20(10), 2486-2495.

Oppenheim, A. V., & Lim, J. S. (1981). The importance of phase in signals. Proceedings of the IEEE, 69(5), 529-541.

Paliwal, K. K., & Alsteris, L. D. (2005). On the usefulness of STFT phase spectrum in human listening tests. Speech Communication, 45(2), 153–170. doi:10.1016/j.specom.2004.08.00

Peelle J.E. & Davis M.H. Neural oscillations carry speech rhythm through to comprehension. Frontiers in psychology. 2012 Sep 6;3:320.

Peelle J.E. & Wingfield A. (2005) Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech. J Exp Psychol Hum Percept Perform 31: 1315–1330.

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. Language, 87(3), 539–558.

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of musical disorders. Annals of the New York Academy of Sciences, 999(1), 58-75.

Peretz, I., & Zatorre, R. J. (2005). Brain organization for music processing. Annu. Rev. Psychol., 56, 89-114.

Peretz, I., Vuvan, D., Lagrois, M. É., & Armony, J. L. (2015). Neural overlap in processing music and speech. Phil. Trans. R. Soc. B, 370(1664), 20140090.

Pijl, S. (1997). Labeling of musical interval size by cochlear implant patients and normally hearing subjects. Ear and Hearing, 18(5), 364-372.

Raphael, L.J. & Isenberg, D. (1980). Acoustic cues for a fricative-affricate contrast in word-final position. Journal of Phonetics.;8:397-405.

Repp, B.H., Liberman, A.M., Eccardt, T., Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: Human Perception and Performance. Nov;4(4):621.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. The Journal of the Acoustical Society of America, 123(2), 1104–1113.

Quené, H. (2013). Longitudinal trends in speech tempo: The case of Queen Beatrix. The Journal of the Acoustical Society of America, 133(6), EL452–EL457.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. Philosophical Transactions of the Royal Society of London B: Biological Sciences. Jun 29;336(1278):367-73.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. Nature, 398(6730),

Shannon, R. V. (1989). Detection of gaps in sinusoids and pulse trains by patients with cochlear implants. Journal of the Acoustical Society of America, 85, 2587–2592.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. Science, 270(5234), 303–304.

Schneider B. & Pichora-Fuller K. (2001) Age-related changes in temporal processing: Implications for speech perception. Seminars in Hearing 22: 227–240.

Shi, G., Shanechi, M.M., & Aarabi, P. (2006). On the importance of phase in human speech recognition. IEEE transactions on audio, speech, and language processing. Sep;14(5):1867-74.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. Nature, 416(6876), 87–90.

Soli, S.D. (1983). The role of spectral cues in discrimination of voice onset time differences. The Journal of the Acoustical Society of America. Jun;73(6):2150-65.

Stewart, L., Walsh, V., Frith, U., & Rothwell, J. (2001). Transcranial magnetic stimulation produces speech arrest but not song arrest. Annals of New York Academy of Sciences, 433-435.

Summerfield, Q., Bailey, P.J., Seton, J., Dorman, M.F. (1981). Fricative envelope parameters and silent intervals in distinguishing 'slit'and 'split'. Phonetica. Jul 1;38(1-3):181-92.

Traunmüller, H., & Lacerda, F. (1987). Perceptual relativity in identification of two-formant vowels. Speech Communication, 6(2), 143-157.

Venezia, J. H., Hickok, G., & Richards, V. M. (2016). Auditory "bubbles": Efficient classification of the spectrotemporal modulations essential for speech intelligibility. The Journal of the Acoustical Society of America, 140(2), 1072-1088.

Zeng, F. G. (2002). Temporal pitch in electric hearing. Hearing Research, 174, 101–106.

Zeng, F.-G., Nie, K., Liu, S., Stickney, G., Del Rio, E., Kong, Y.-Y., & Chen, H. (2004). On the dichotomy in auditory perception between temporal envelope and fine structure cues. The Journal of the Acoustical Society of America, 116(3), 1351. doi:10.1121/1.1777938

Zeng, F. G., Tang, Q., & Lu, T. (2014). Abnormal pitch perception produced by cochlear implant stimulation. PloS one, 9(2), e88662.

Zhao Y. (1997) The effects of listeners' control of speech rate on second language comprehension. Applied Linguistics 18: 49–68.