Audio Expansion Techniques For Real-Time Communication And Intelligibility Enhancement

BY

JOHN S. NOVAK, III
BSEE, Bradley University, 1993
MSEE, Bradley University, 1997
MSCS, DePaul University, 2005

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the University of Illinois at Chicago, 2022

Chicago, Illinois

Defense Committee:

Robert Kenyon, Chair
Steve Jones, Communication
Brian Ziebart
Cornelia Caragea
Angus Forbes, Purdue University

For my father, who struggled with machinist’s hearing loss much longer than he let us believe.

Acknowledgements

It is my great privilege and good fortune to have had three distinct communities of friends and colleagues with me through this undertaking: I could not have made it through this without all of the friends and mentors from the University of Illinois at Chicago, from Northrop Grumman, and from that web of connected friends I hazily think of as "real life."

From the University, I thank especially my advisor Professor Robert Kenyon, without whose firehose-volume of good advice and infinite-seeming patience, little would have come of this. I also thank the rest of my committee (Professors Steve Jones, Brian Ziebart, Angus Forbes, and Cornelia Caragea) for their guidance, advice, and especially their patience. Also, of course, Lance Long for unstinting technical assistance and advice, as well as the herculean task of maintaining all the EVL infrastructure. Thanks to Viktor Mateevitsi and Jason Archer, two other veterans of the first Human Augmentics seminar (out of which this dissertation grew) for some of the best and most challenging conversations I've had on that or any other topic. And finally, thanks to extended EVL community, comprised of students and faculty too numerous to mention.

From Northrop Grumman, thanks to Brad Bourgeois and Ken Eakes for their decades long friendship and support. And thanks especially to all those who have had to supervise me during this process, away from work either 50% or 100% of the time: Brad Bourgeois (again), Matt Clark, Sheri Sepeczi, and Eric Lomonaco—I'll probably never know what kind of interference you've had to run for me, and actually, I probably don't even *want* to know. And most especially, thanks to Stu Collar for that chance conversation in the hallway ("Hey, I heard you're interested in doing a PhD… have you heard about the new Fellowship Program?") without which none of this would have happened.

Last, but certainly not least, from the mythical "real life" friends: John and Annette Dilick, for their friendship, their unstinting technical advice and consulting (when I had things too embarrassingly trivial-yet-vexing to take to Lance) and also for the loan of their adult children for pilot testing when I had drained the well at EVL.

Pamela Korda for her friendship and faith in my abilities, her advice on statistics, and her advice on navigating PhD-land. And Leigh Butler, for her friendship, faith in me, eternal encouragement, and mutual support during our writing processes—now go finish your book, wouldja? I want to know how it ends.

CONTRIBUTIONS OF AUTHORS

Chapters One and Two comprise a combined literature survey of the fields of clear speech and temporal scaling of audio signals, respectively, and are not based primarily on existing published work.

Chapter Three details the requirements and implementation details of working, real time temporal scaling systems and is based on two prior publications. The first, [1] describes a laptop-based single-talker system, for which I was the primary author and software developer; Jason Archer provided feedback for the application itself with an eye toward future studies, Professors Robert Kenyon (my advisor), Valeriy Shafiro (Rush University), and Jason Leigh now at the University of Hawai'i) provided mentoring, supervision, and the inspiration for the project. The second [2] describes a smartphone-based wireless dual-talker system, for which I was the primary author and lead software developer and architect; Aashish Tandon provided a considerable amount of software development, and Professors Robert Kenyon and Jason Leigh again provided mentoring and supervision. This chapter also contains the description of a third, unpublished prototype, which was mediated by a permanent server and enabled wireless communication at much greater physical distances. I was the sole software architect and developer for this project.

Chapter Four details a user study [3] which used a Diapix task to elicit spontaneous, conversational speech from subjects, while their voices were modified by our software. I, as primary author, provided the audio manipulation software and technical expertise required to set up the experiment, Jason Archer designed the experiment (which we jointly conducted), Professor Valeriy Shafiro provided detailed analyses of the produced speech, and Professor Robert Kenyon provided considerable mentoring and insight into experimental design and procedures.

Chapter Five describes a user study [4] investigating the effects on speech intelligibility of allowing test subjects to control the rate of received speech in noisy conditions. I am the primary author of this work, and designed user study software, conducted the user study and prepared the results for publication. Professor Robert Kenyon and I jointly designed the study protocol, and Professor Robert Kenyon provided substantial mentoring and oversight of the interpretation of the results.

Chapter Six describes a similar user study [5] investigating the effects of user controlled speech rates on the listening comprehension of foreign language learners. I am the primary author of this publication, and provided software architecture and guidance, as well as the final statistical analyses and write-up; Dan Bunn designed and conducted the user study, and Professor Robert Kenyon provided mentoring, guidance, and assistance with the study design. This work also formed the basis for Dan Bunn's master's degree thesis [6] for which he rightly deserves sole authorship.

Chapter Seven describes a final user study, in preparation for publication, investigating the effects of providing finer-grained control of speech expansion at the phoneme level to experimental subjects. I am the primary author of this work, responsible for software development, study design, and conducting the study, while Professors Robert Kenyon and Steve Jones provided substantial and useful insights into the analysis and interpretation of the data.

Finally, Chapter Eight summarizes and concludes the dissertation, with discussion and useful directions for future research and is not based on any prior publications.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

LIST OF TERMS AND ABBREVIATIONS

Affricate
A voiced or voiceless consonant which begins as a plosive and continues as a fricative; in this dissertation, considered as a Fricative.

Approximant
Voiced phonemes, produced by a narrowing of the vocal tract not sufficient to produce turbulent airflow.

Attention Question
A question in a survey or online experiment designed to test the engagement and attention of test subjects.

Audio Contraction
The act of manipulating audio data such that playback time is decreased, without otherwise altering audio qualities, especially frequency.

Audio Dilation
The act of manipulating audio data such that playback time is increased, without otherwise altering audio qualities, especially frequency. Quantified as the ratio of unmodified to modified playback times.

Audio Expansion
The act of manipulating audio data such that playback time is increased, without otherwise altering audio qualities, especially frequency. Quantified as the ratio of modified to unmodified playback times.

Audio Stretching
Used synonymously with Audio Expansion.

α
See Expansion Factor

Bernoulli Effect
The principle by which an increase in fluid velocity generates a decline in pressure.

Bidirectional Long Short-Term Memory
An LSTM network which collects an entire sequence of data and simultaneously processes it in both the forward and backward directions.

BLSTM
See Bidirectional Long Short-Term Memory

Casual Speech
An umbrella term for speech produced in non-challenging environments.

Clear Speech
An umbrella term for any of several talker-initiated adaptations to speech production intended to make speech more intelligible (i.e., more ‘clear’) for one or more listeners.

Comprehension
A measure of how intelligible human speech is, typically at the length of paragraphs or long utterances.

Conversational Speech
Used synonymously with Casual Speech.

DFT
See Discrete Fourier Transform

Discrete Fourier Transform
A mathematical operation decomposing discrete functions of time into discrete functions of frequency; often refers to an $O(N^2)$ algorithm for calculation of the same.

Expansion Factor
The ratio of modified to unmodified playback time, denoted α

f_0
See Fundamental Pitch.

F1, F2…
See Formant.

Fast Fourier Transform
An algorithm to compute the discrete Fourier Transform in $O(N \log N)$ time rather than $O(N^2)$ time, facilitating the calculation of the DFT on much larger datasets.

FDS
Foreign Directed Speech.

FFT
See Fast Fourier Transform.

Formant
Portions of the frequency response of the Laryngeal Buzz, amplified by the resonances of the vocal tract, responsible for the perception of vowel sounds. Denoted F1, F2…

Fricative
A voiced or voiceless consonant produced by narrowing or constricting the vocal tract to produce turbulent airflow. In this dissertation, also includes Affricate consonants.

Fundamental Pitch
The frequency at which vocal cords vibrate, denoted f_0.

Fourier Transform	A mathematical operation decomposing continuous functions of time into continuous functions of frequency.
Glimpse	A cochleagram cell where the signal exceeds interfering noise by 3 dB or more.
Glimpse Area	The total number of glimpse cells in a cochleagram.
HIDS	Hearing Impaired Directed Speech.
IDS	Infant Directed Speech.
Intelligibility	A measure of how intelligible human speech is, typically at the level of single words, or words in short utterances, with or without grammatical content.
Laryngeal Buzz	The sound resulting from a stream of glottal pulses, unshaped by a vocal tract (i.e., if the vocal cords were exposed directly to the air).
Listener	A participant in an audio communication who is currently listening to and interpreting utterances made of a talker’s words.
Lombard Speech	Speech produced in the presence of noise.
Long Short-Term Memory	A recurrent neural network architecture designed to process sequential data (such as sound) characterized by structures adding short term temporal memory capabilities to the long term memory inherent in the weights of the typical recurrent neural network.
LSTM	See Long Short-Term Memory
MDS	Machine Directed Speech.
Mel Frequency Cepstral Coefficient	A spectral transform related to the short term Fourier transform, which warps the frequency scale to better conform to the non-linear experience of frequency provided by the human ear.
MFCC	See Mel Frequency Cepstral Coefficient

Nasal
Voiced consonants produced by airflow through the noise (and therefore nasal cavity) but not the mouth.

NAT
A Noisy Attention Task, i.e., a transcription task presented in extreme noise conditions, which should result in the use of a transcription pass phrase.

NLAT
A Noiseless Attention Task, i.e., a transcription task presented in very low or no noise, which should result in a perfect transcription.

PER
See Preferred Expansion Rate

Phoneme
A perceptually distinct unit of sound, distinguishing individual words from each other. In this document, see Affricate, Approximant, Fricative, Glide, Liquid, Nasal, Plosive, Stop, Vowel

Pitch Period
The period of time required to complete one glottal cycle, e.g., from vocal cord opening to opening.

Plosive
A phoneme in which the vocal tract is blocked or closed, and subsequently rapidly resumes. Also referred to as a Stop.

Preferred Expansion Factor
See Preferred Expansion Rate

Preferred Expansion Rate
In the experiments of Chapters Five, Six, and Seven, a subjects' final expressed preference for a slowed rate of speech, applied either to utterances or phonemes.

QuickSIN
A fast, proprietary test to determine the ability of a patient to understand a target human voice in the presence of other noise.

Short Term Fourier Transform
The application of a Fourier transform to short, successive segments ('windows') of a longer signal, to better understand the evolution of the signal's spectrum over time.

SNR
Signal to Noise Ratio, typically of speech ('signal') played against noise, and measured in dB. Note that large/positive values connote more signal per noise and are thus easier to correctly extract meaning from.

SNR-50 The SNR at which a patient is expected to be able to correctly recite 50% of the keywords correctly in a QuickSIN test; expected to be 2 dB SNR for young healthy listeners.

SNR-Loss The deviation of one listener’s ability from that of a statistically average young, healthy listener.

Speaker An electromechanical device which produces sound in the atmosphere through its vibrations, e.g., a loudspeaker or a bass speaker. To void ambiguity, in this document, ‘speaker’ never carries the connotation of ‘talker”.

STFT See Short Term Fourier Transform

Stop See Plosive

Talker A participant in audio and verbal communication who is currently talking, i.e., producing utterances made of words. To void ambiguity, in this document, ‘talker’ never carries the connotation of ‘speaker”.

TIMIT A corpus of phonetically balanced sentences, recorded in multiple American dialects, by Texas Instruments (TI) and the Massachusetts Institute of Technology (MIT)

TOEFL The proprietary Test of English as a Foreign Language

VAD Abbreviation for Voice Activation Detection

Voice Activation Detection A software function defined to activate or disactivate hardware or software features based on the presence or absence of a human voice.

Voice Over Internet Protocol Method for streaming sound (especially the human voice) over internet protocol networks.

VOIP Abbreviation for Voice Over Internet Protocol

Vowel Space

A graphical representation of the various vowel sounds, with each sound located as a point in an F1, F2 space. These points are statistical averages of many vowel sounds often across many talkers.

Vocoder

From ‘voice encoder’, a hardware or software device intended to analyze and synthesize human speech.

Vowel

A voiced phoneme produced with an unrestricted vocal tract.

SUMMARY

It has often been observed, informally and in academic literature, that speech directed toward individuals who are hard of hearing is naturally modified by talkers in an attempt to accommodate the listeners. These modifications take numerous forms, from the easily expressed “louder” and/or “slower” to the highly technical “vowel space expansion” and many others. There are also speech modifications made in numerous other circumstances, not only for hard of hearing listeners, but for listeners in noisy backgrounds, second language listeners, and many others, known under the collective (is sometimes too-broad) umbrella as clear speech. There is, however, an asymmetry inherent to all these talker-listener scenarios, which is that talkers have only the broadest guidance at best (“Louder!” “Slower!”) from their listener partners, which they may simply ignore.

With an interest in developing new and practical interventions for the elderly and hard of hearing, this dissertation takes as its research questions: First, the feasibility of the production of a signal processing system which allows real-time modification of audio signals. Second, the tolerance and effectiveness of partially emulating clear speech, by using signal processing techniques to modify (specifically, to slow) the rate of received speech signals. And finally, the effectiveness of placing the control of speech rate directly into the hands of listeners, a capability that has never been possible prior to electronic mediation.

This dissertation is organized as follows:

Chapter One is a survey of both the circumstances which prompt these spontaneous speech modifications, as well as of the details of the modifications which are made.

Chapter Two presents an overview of the various methods of changing the rate of speech including the simple but ineffective technique of re-sampling, as well as several frequency domain techniques. These systems were intended for use on pre-existing, rather than live, signals.

Chapter Three presents the details of several incarnations of this real-time system, including a single-talker system implemented on a laptop computer, presented in [1] and used directly (with two laptops for two talkers) in the study presented in [3]; a two-talker wireless system using Android smartphones and local network connectivity

presented at [2]; and an unpublished system using a central laptop base station, to which Android smart phones could connect even when hundreds of miles apart. This chapter contains details required for the implementation of both live signals, and live control of the playback rate of those signals. Finally, this system includes the design details of systems used in the three user studies of Chapters Five, Six, and Seven.

Chapter Four presents a user study designed to examine the effects of using real time temporal modification software in a conversational setting, previously published in [3]. The key element of the study design was the use of a scored Diapix test [7] which was designed to elicit spontaneous conversational speech between two study participants without the involvement of or prompting by researchers. The main results of this study are (i) that at modest amounts of stretching (40% additional playback time, with lengthy silences not stretched) the speech patterns of subjects are not changed to a statistically significant degree, and (ii) under those same circumstances, there is no statistical evidence of changes in performance on the Diapix test.

Chapter Five presents the first of three user studies focusing on the little-studied subject of the interaction of user choice and user control over received speech rates in adverse conditions. This study was based off the pre-existing QuickSIN test [8], designed to test and characterize a listener's ability to understand short sentences of speech played against a background of noise. In the first part of this study, users were asked to listen to several short sentences each played against multiple levels of background noise. For each level of background noise, the subjects were asked to use a computer interface to set the rate of the speech to whatever rate they believed was most helpful in understanding the foreground speech. (The background babble did not change rate.) These preferred expansion rates were recorded for later use and analysis. In the second part of the study, subjects were asked to listen to and immediately repeat back sentences at those same noise conditions, both with speech modified to their preferred expansion rates, and unmodified. The subjects' repetition of those sentences was also recorded for analysis.

Our results included statistically significant differences in received speech rate at opposite ends of the noise condition, with subjects requesting slower speech in the presence of increase noise; as well as an overall belief that slowing speech was beneficial to subjects' performance. However, analysis of the subjects' ability to repeat was in fact mildly degraded at moderate to high noise conditions, and was not improved at any noise condition. This work was published at Interspeech 2018 [4].

Chapter Six presents the second of the three user control studies. In this study, non-native students with recent TOEFL scores were recruited as test subjects. After familiarization with a graphical interface, test subjects were asked to participate in six comprehension tasks, each of which was a lengthy (several minutes long) audio track of a TOEFL test to which they had not previously been exposed. During three of these tasks, the subjects were given the ability to change the rate of playback as they saw fit, and asked to do so in whatever fashion they believed would best help them understand the audio passage. The other three comprehension tasks were presented unmodified, without a control interface. After each task, the subjects were immediately given a short multiple choice quiz about the contents of the previous audio passage. Subjects' behavior (i.e., the rate changes specified) and quiz responses were collected and analyzed

Our analysis of the data shows that a considerable majority of the subjects used speech slowing in all three of their trials (roughly 2/3) and an overwhelming majority used the technique in at least one of their trials (80%), but did not reveal a statistically significant improvement or degradation in listening comprehension. There is also a weak correlation between lower TOEFL scores and more slowing. This work was published at Interspeech 2019 [5].

In the work described above, the algorithms employed stretch speech uniformly—if set to expand an audio track by 40%, all sounds, including all parts of speech all background noises, and even silences, are expanded by the same rate. For large expansion factors, the speech so produced sounds somewhat unnatural, because this is not how talkers spontaneously produce slow speech. With this in mind, Chapter Seven returns to the topic of speech in noise, but puts even more precise control in the hands of test subjects: A relatively simple neural network was developed to classify individual phonemes into six broad phonemic categories within an audio track. A study was also designed around this tool which asked subjects to listen to short sentences in noise, and then expand or contract the audio only for particular phonemes (e.g., first modify vowels, then modify fricatives, etc.) Following this, all modifications were applied simultaneously, and subjects were asked to listen to modified and unmodified/control sentences in three levels of background noise and asked to transcribe the sentences into a computer interface.

Our analysis found either no statistically significant improvement or degradation (at one noise level) or statistically significant degradation (at two noise levels) of intelligibility while using this technique.

Finally, Chapter Eight presents overall conclusions, with additional discussion of the entire work and of directions for future studies.

1 On Speech and Clear Speech

Communication by speech requires a talker, who produces acoustic words that are assembled into larger utterances according to the grammatical rules of a language; a listener, who receives and interprets those acoustic words into utterances, and utterances into meaning; and a channel or pipeline of channels which carries those acoustic utterances from the talker to the listener. (In this document, 'talker' always carries the meaning of a person who talks by producing words, while 'speaker' is reserved for an electromechanical device which produces speech or non-speech sounds with the sole exception of the phrase 'native speaker', well-established in the literature.)

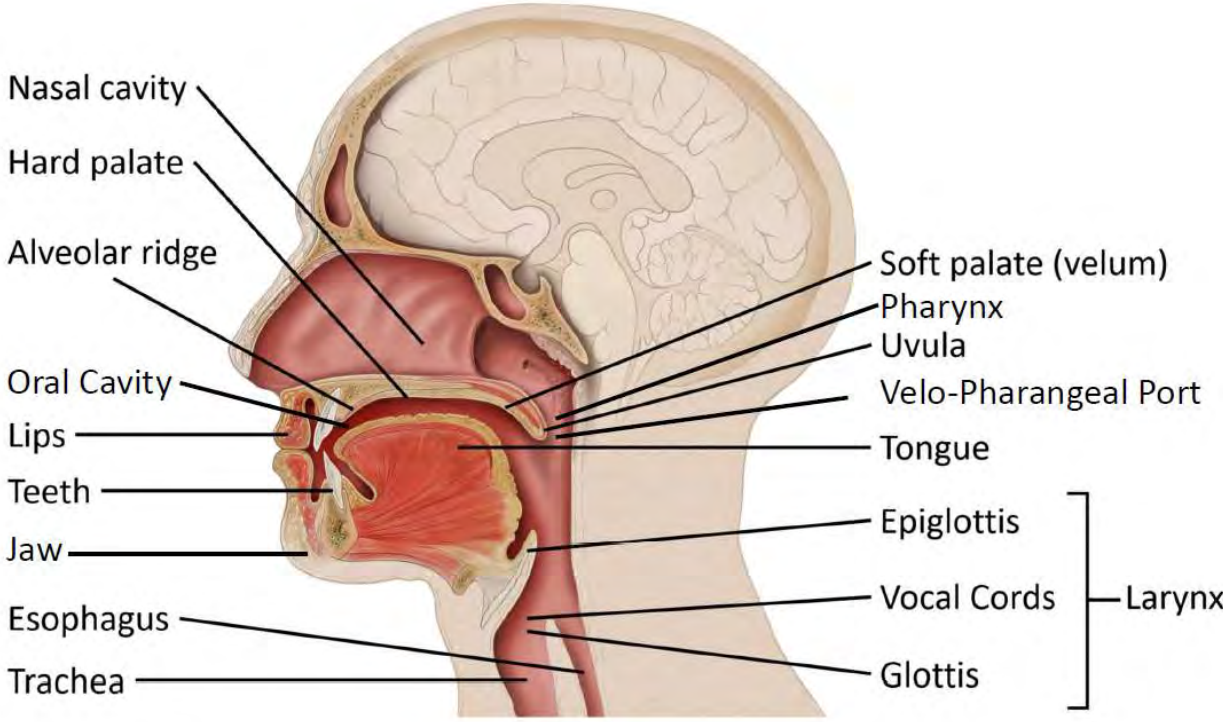
However, difficulties may be introduced at every stage of this simple model: A talker may be unclear, a channel may introduce noise or otherwise degrade the acoustic signals, and a listener may suffer from a wide variety of conditions which make understanding of speech difficult. Talkers, in turn, may attempt to adapt their speech to compensate for or overcome those difficulties, especially those introduced by the channel or the listener. "Clear speech" is an umbrella term for the very wide range of speech modifications made by talkers attempting to so compensate, in contrast to "casual speech" or "conversational speech" which is produced in non-challenging situations. Although this dissertation focuses on narrow questions of listener-control of one narrow type of clear speech adaptation, this chapter serves as a brief introduction to (1) the acts of speech production and reception by talkers and listeners, respectively; (2) representations of and some salient features and metrics of speech; and (3) to the types of modifications talkers may make in various adverse situations.

1.1 <u>Speech Production and Speech Signals</u>

Speech is made of sound, that is, of propagating longitudinal waves of higher and lower pressure in a fluid medium (almost always, and here assumed to be, air). The system of human speech production is comprised of three sub-systems: (1) The lungs (not shown), including the trachea, up to the larynx; (2) the vocal cords, which are housed in the larynx above the trachea; and (3) the vocal tract (i.e., everything above the larynx, Figure 1[1]).

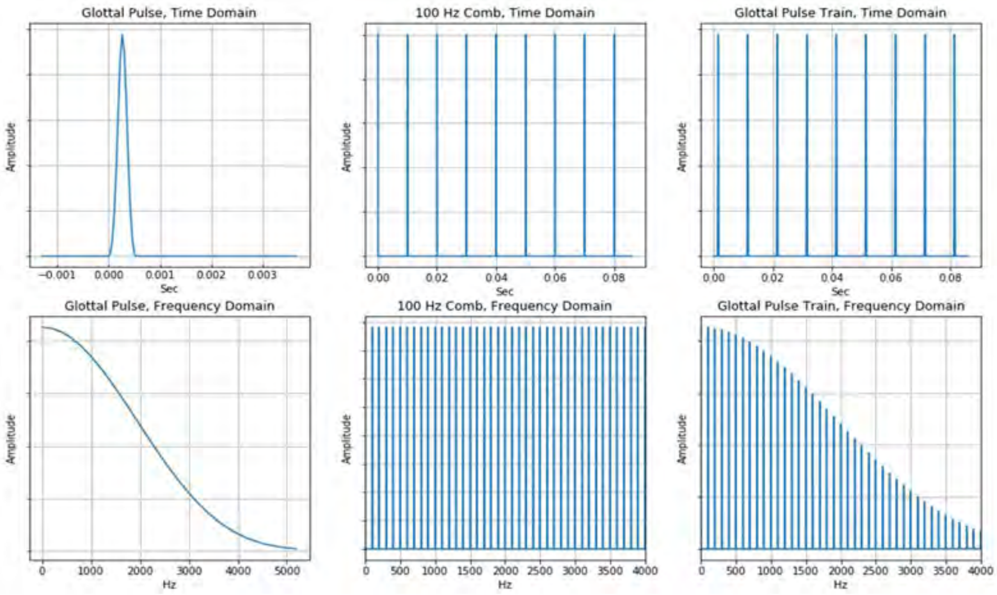
[1] Figure 1 has been adapted from [9] with modest additions and changes to labels in accord with CC YB 2.5

Figure 1 Diagram of Speech Apparatus [Adapted from 9]



The lungs (not shown) serve as a reservoir of air and air pressure during speech, and drive air through the vocal cords. During most unvoiced segments of speech, the vocal cords may remain entirely relaxed allowing laminar air to flow through them, or they may constrict to a greater or lesser degree and allow “noisier” turbulent air to flow through. (In some cases, the vocal cords may constrict tightly and briefly cut off the flow of air entirely, as during a sharply articulated plosive ‘k’ sound.) During voiced segments of speech, however, the vocal cords are closed-- not tightly but nearly in balance with air pressure from the lungs. In this condition, the air pressure from the lungs builds up behind the vocal cords until they are briefly forced open. At the moment of this opening, air flows through the gap in the vocal cords—slowly at first, because air does not accelerate infinitely quickly, but accelerating rapidly. As the velocity of the air through the vocal cords increases, the Bernoulli effect causes a drop in air pressure between the vocal cords, which eventually brings the vocal cords back together in a closed position, blocking airflow once more [10]. This single phonation cycle results in a single pulse of air driven through the vocal cords. As long as air pressure

Figure 2   Production of Laryngeal Buzz



remains from the lungs, this process will continue, generating a train of air pulses. The length of this cycle varies according to the stiffness of the individual vocal cords, ranging from 85 Hz (about 12 msec between pulses) to 255 Hz (about 4 msec between pulses) for a low-pitched adult male and high-pitched female voice, respectively. This measure, the time between pulses, is referred to as the pitch period of a voice; the measure of cycles/sec is referred to as the fundamental pitch of a voice, denoted f_0. Note, however, that the tension and stiffness of the vocal cords can be adjusted by moving the position of the larynx slightly. The fundamental pitch therefore varies in time for individual speakers, and is an instantaneous measure. It should also be noted that this is a biological process, and the resultant pulses are decidedly not square, or triangular, or anything so geometrically regular.

This sequence of pulses (for voiced speech segments) is referred to as "laryngeal buzz" and is not heard directly unless synthesized, because the pulses are affected significantly by the third stage in vocalization. The behavior of laryngeal buzz is best understood in the frequency domain. A single glottal pulse with a duration of roughly 0.005 sec (Figure 2, upper left [2]) has a bandwidth of approximately 4000 Hz in the frequency domain, with

[2] The single pulse shown in Figure 2 is highly idealized.

most of the power concentrated at the lower frequencies (Figure 2, lower left.) A sequence of pulses can be simulated by convolving a single glottal pulse with a time series comb function (Figure 2, upper middle) whose frequency domain representation is also a comb function. (Figure 2, lower middle) Note that the time domain impulses at 100 Hz generate a frequency domain response with spectral lines separated by 100 Hz. The result of the convolution (Figure 2, upper right) in the time domain is a 100 Hz sequence of glottal pulses. To understand the frequency response (Figure 2, lower right) recall that convolution in the time domain is multiplication in the frequency domain—the act of turning a glottal pulse into a glottal pulse *train* at 100 Hz effectively samples the frequency response of a single pulse at 100 Hz intervals in the frequency domain.

These pulses (for voiced speech segments) and periods of laminar or turbulent flow (for unvoiced speech segments) are the direct source of sound from the human system. In the absence of the vocal tract (i.e., if the vocal cords were open to the atmosphere directly), a pressure sensor placed directly outside the vocal cords would detect rapidly alternating periods of high and low pressure, which propagate through the atmosphere in all directions.

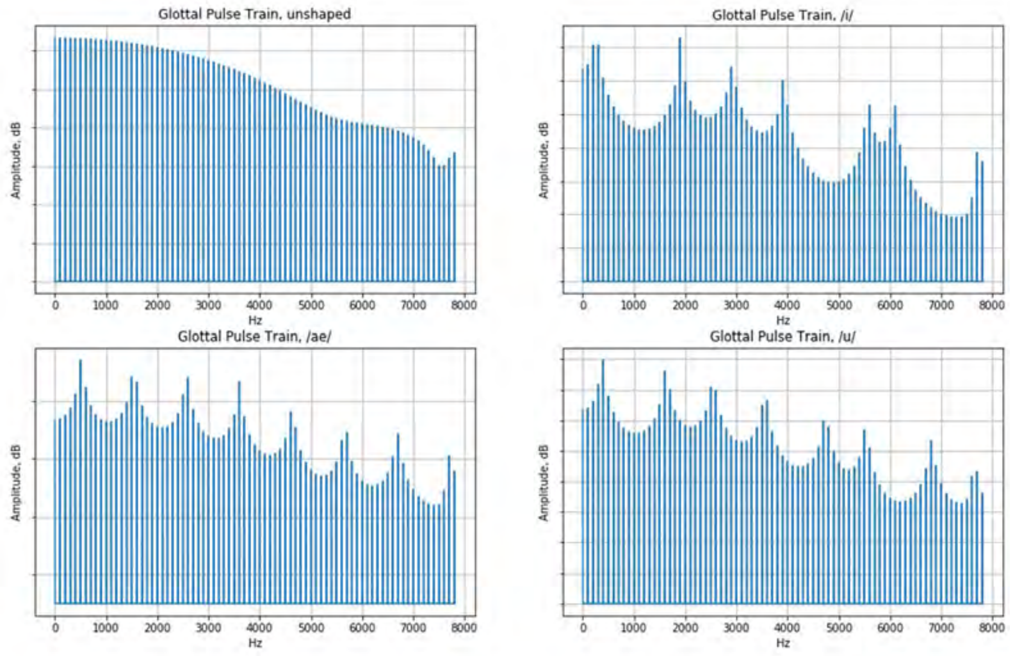
However, in the human speech apparatus, the airflow through the vocal cords next enters the vocal tract, which consists of a set of branching adjustable tubes or cavities. These include the pharynx, directly above the larynx, which can be constricted to adjust its diameter; the nasal cavity, which allows air to escape through the nose and which can be closed by the velo-pharyngeal port; and the oral cavity which can be closed and/or extensively modified by changes in position of the jaw, changes in position of the tongue (including interactions with teeth or other parts of the mouth), and changes in position of the lips (again including interactions with the teeth) in various combinations. The vocal tract acts as a complex and reconfigurable resonant cavity whose resonances and anti-resonances further shape the spectral response of the glottal pulse train. In particular, the first three resonances, F1, F2, and F3 (referred to as formants) each lift up one or (usually) several harmonics of the fundamental pitch, and form the basis for the perception of vowels and other strongly vocalized sounds. Examples of this process can be seen in Figure 3, as the frequency response of a glottal pulse train (Figure 3, upper left) is spectrally shaped into an /i/ (long ee, as in 'feet', Figure 3 upper right), an /ae/ (short a, as in 'fat', Figure 3 lower left), and an /u/ (long u, as in 'food', Figure 3 lower right.) Note that the vocal tract changes on the time-scale of the syllable rate of a language, where even rapidly spoken languages may have syllable lengths of up to 100 ms [11]. This is very slow compared to the time between glottal pulses, which are on the order of 10 ms apart or less. As such, the spectral slices of Figure 3 are, again, instantaneous representations.

Finally, constrictions of the vocal tract can add additional turbulence to the airflow producing additional voiced or unvoiced consonants as the air leaves the mouth or nose.

1.2 Speech Reception And Psychoacoustics

Once speech sounds are generated and leave the vocal tract, they propagate as longitudinal pressure waves, losing strength with distance travelled, until they enter a human ear which converts them from the mechanical energy of pressure waves into neural signals in the human brain. This is done in three stages as signals travel from the outer ear (Figure 4[3]), through the middle ear (Figure 4), into the inner ear, or cochlea (Figure 4 and Figure 5[4]), eventually exiting the ear and are carried to the brain via the cochlear nerve.

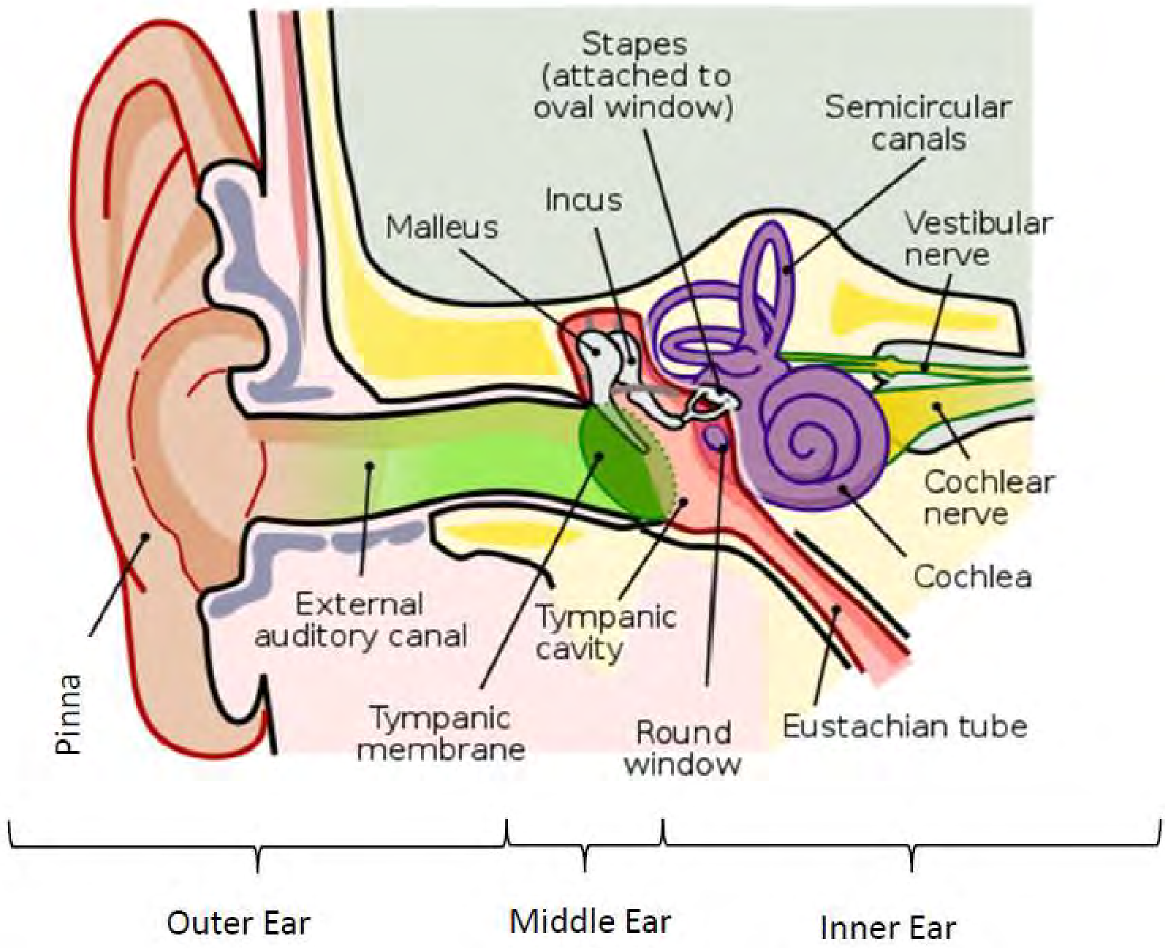
Figure 3 Glottal pulse frequency response: Unshaped, /i/, /ae/, and /u/



[3] Figure 4 has been adapted from [12] with modest additions and changes to labels in accord with CC BY 2.5
[4] Figure 5 has been adapted from [13] with modest additions and changes to labels in accord with CC BY 2.5

The outer (or external) ear is a passive, reactionless system. Sound enters the open end of the ear at the pinna, and travels through the external auditory canal, or ear canal until it strikes the flexible tympanic membrane, or the ear drum. The irregular shapes of the pinna introduce slight top-bottom and front-back asymmetries which assist in directional hearing. The ear canal, with a length of roughly 20-25 mm, acts as a quarter wave transmission line resonating at (and thus emphasizing) frequencies of roughly 3-4 KHz. (Several systems in the ear, including the ear canal, impose their own frequency responses on sound waves travelling through them. Unlike the vocal tract, these resonances cannot be consciously modified, and are omitted for brevity.) As sound waves strike the tympanic membrane, it vibrates in sympathy much like the skin of a musical drum and transmits the vibrations into the middle ear.

Figure 4 Cross Section of Outer, Middle, and Inner Ear [adapted from 12]

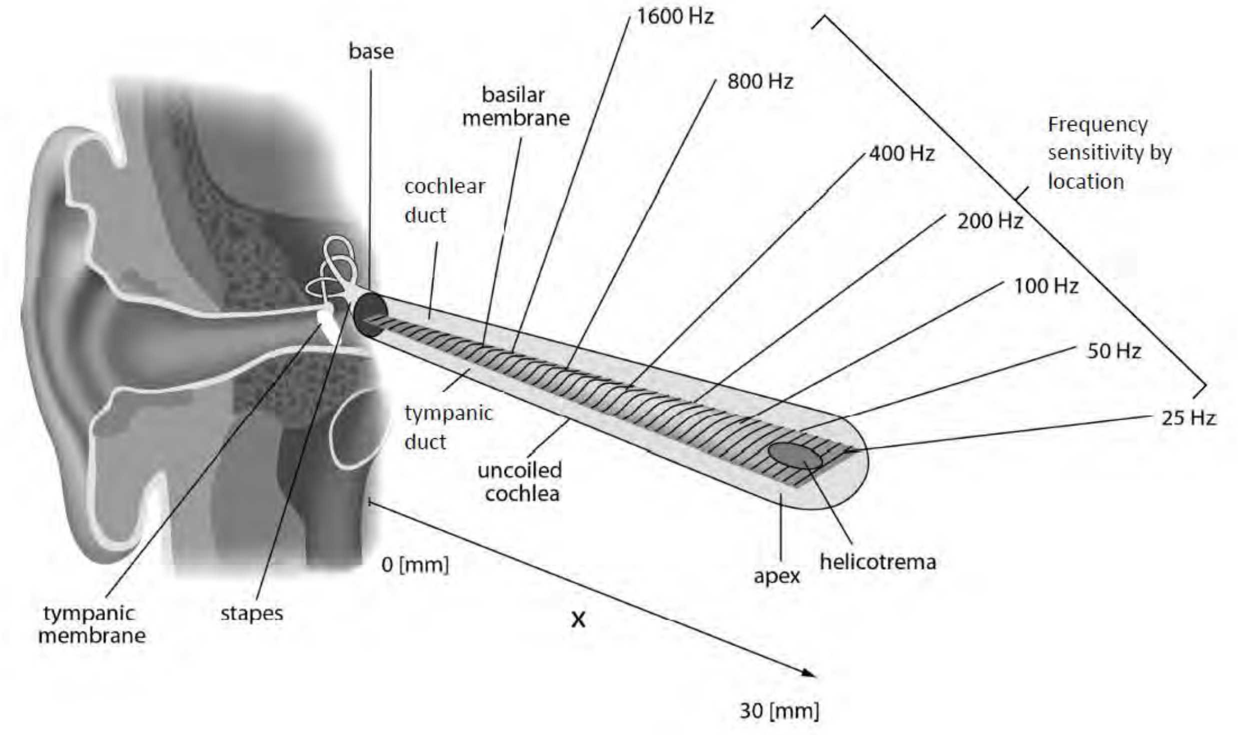


The outer (or external) ear is a passive, reactionless system. Sound enters the open end of the ear at the pinna, and travels through the external auditory canal, or ear canal until it strikes the flexible tympanic membrane, or the ear drum. The irregular shapes of the pinna introduce slight top-bottom and front-back asymmetries which assist in directional hearing. The ear canal, with a length of roughly 20-25 mm, acts as a quarter wave transmission line resonating at (and thus emphasizing) frequencies of roughly 3-4 KHz. (Several systems in the ear, including the ear canal, impose their own frequency responses on sound waves travelling through them. Unlike the vocal tract, these resonances cannot be consciously modified, and are omitted for brevity.) As sound waves strike the tympanic membrane, it vibrates in sympathy much like the skin of a musical drum and transmits the vibrations into the middle ear.

The middle ear is a solid, mechanical linkage between the air-filled and air-driven system of the outer ear and the liquid-filled and liquid-driven system of the inner ear. This solid linkage consists of three tiny bones: the malleus, incus, and the stapes. The malleus is attached directly to the ear drum and transmits vibrations through the incus and into the stapes, which drives like a piston into a flexible portion of the surface of the inner ear called the oval window. The significance of this mechanical linkage is profound: The acoustic impedance (i.e., the mechanical resistance to acoustic vibrations) of the fluids in the inner ear is approximately 3000 times greater than that of the air in the outer ear. This means that if the ear drum directly separated the outer and inner ear systems, in contact with air on one side and the lymphatic fluid of the inner ear on the other, most of the sound would simply be reflected back out of the ear. This middle-ear-less system would impose a direct loss of roughly 30 dB at the tympanic membrane boundary. However, the structure of the bones counteracts and overcomes this loss through two mechanisms. First, the area of the stapes striking the oval window is roughly 20 times smaller than the area of the tympanic membrane, which increases the pressure into the inner ear by the same factor. Additionally, the configuration of the bones acts like a lever providing an additional force multiplication which varies over frequency. The overall effect, then, of the middle ear, is to act as an air-to-lymphatic fluid transducer, overcoming as much of the loss due to acoustic impedance mismatch as possible.

Once transduced into the cochlea of the inner ear, fluid pressure waves travel the distance of two lengthwise channels, the cochlear and tympanic ducts. These chambers are separated by the basilar membrane, whose physical

Figure 5 Unwrapped Schematic of Cochlea [adapted from 13]



characteristics (width, stiffness, etc.) vary such that its resonant frequency changes along its own length. Thus, vibrations in the liquids of these adjacent chambers induce travelling wave vibrations in the basilar membrane with a frequency selectivity according to position: the basilar membrane resonates at the highest frequencies near its base and the lowest frequencies near the apex. In the passive function of the cochlea, the basilar membrane anchors a multitude of small hairs (stereocilia, not shown) which bend with the passing vibrations in the basilar membrane, causing cellular processes with the cilia to generate electrical signals that are captured by the auditory nerve and transmitted to the brain for further processing and interpretation. That is to say, the arrangement of fluid filled chambers and embedded hairs passively conspire to cause parts of the cochlear nerve to fire in response certain frequencies of sound. However, later studies have shown that the hair cells are part of a mechanically active system which serves to amplify the vibrations in the basilar membrane. (See Figure 5 for a schematic of the inner ear featuring an "unwrapped" or "unwound" cochlea for ease in visualizing the cochlear positional frequency mapping; compare and contrast to the more biologically accurate cochlea drawn in Figure 4.)

The human ear, then, is a stack of signal transducers (air vibrations to liquid vibrations to membrane travelling waves to flexing cilia and finally to neural impulses in the auditory nerve) which also perform a tonotopic mapping. That is, they effect a biological spectral transform prior to presentation to the brain, mapping different frequencies of sound to various positions in the cochlea, and ultimately to different nerve fibers into the brain. However, while this spectral transform is similar *in principle* to a short-term Fourier analysis (and provides intuitive justification for psychoacoustic investigations using Fourier analyses as their basis), it differs in several important respects, chiefly that unlike a mathematically perfect Fourier transform or similar operation, the tonotopic mapping is *not* linear with respect to frequency, possessing both a linear region and a logarithmic region, and that the physical characteristics of the basilar membrane permit high frequency sensitivity, but also enforce frequency masking [14] whereby one frequency can perceptually mask another tonally nearby frequency. Thus, while it is possible to distinguish between tones less than 5 Hz apart when played separately [15], those frequencies played simultaneously might not be separately perceived.

1.3 Highly Intelligible "Clear" Speech

Not all speech is created equal. Multiple rigorous and detailed studies have demonstrated that some speech is more or less intelligible than other speech, including but not limited to [16], [17], and [18]. Further, the changes in speaking style which lead to this increased intelligibility spontaneously arise in a variety of contexts and can (with greater or lesser degrees of success) be consciously emulated by talkers. "Clear speech" is a common umbrella term for speech produced in any of these contexts or which successfully emulates speech produced in these contexts. In the sections that follow, I briefly review these contexts, describe some of the more common modifications to speech, and conclude with recent research on the efficacy and universality of the most common.

1.3.1 Contexts: Persons and Situations

The most easily understood context in which a talker may modify their speech to improve intelligibility depends on characteristics, or perceived characteristics, of the listener. These include:

- Hearing Impaired Directed Speech (HIDS), which is marked by a large number of adaptations, including: A slower speech rate including more and longer pauses during speech [19]; increased intensity overall [19];

increased intensity of consonants relative to vowels, especially unvoiced consonants [17]; vowel space expansion [20], and longer transitions in some voiced sounds [19].

- Foreign Directed Speech (FDS) or speech directed to non-native speakers of a language, which is marked by slower speech rate [21], exaggeration of fundamental pitch (f_0) excursions [22], and vowel space expansion [23, 24].

- Infant Directed Speech (IDS), which is marked by significant rise in fundamental pitch, large excursions of fundamental pitch [24], a change in relative intensity favoring vowels over consonants, in contrast to HIDS [25], vowel space expansion [27], exaggerated pauses [28], and slower speech rate [29]. Note that in every case, the degree or severity of these changes varies with the age of the infant.

- Machine Directed Speech (MDS) which is characterized by a reduced range of fundamental pitch [30], vowel space expansion [31], and slower speech rate [31].

In addition to speaker-dependent modifications, some types of clear speech are produced in reaction to the environment. These include:

- Lombard speech, which is speech produced in the presence of noise, characterized by increased intensity, as well as increased intensity of vowels compared to consonants [33], larger excursions of fundamental pitch and slower speech rate [34].

- Speech produced in reverberant environments, marked by increased intensity [35].

- Speech addressed to a distant person, marked by increase intensity [36], increased intensity of vowels over consonants [37], and slower speech rate [37].

There are a number of adaptations which are common to multiple situations, the most common of which are slower speech rates (six environments or contexts), followed by increases in intensity (five), vowel space expansions (four), change of modulation or excursions of fundamental pitch, although some with more or exaggerated excursions, and one with less (four), and changes in intensity of consonants relative to vowels, although with some favoring vowels and others favoring consonants (four).

1.3.2 Common Modifications

The previous section discussed a number of various contexts in which talkers adapt their speech to afford greater clarity to their listeners; here, those adaptations are themselves discussed.

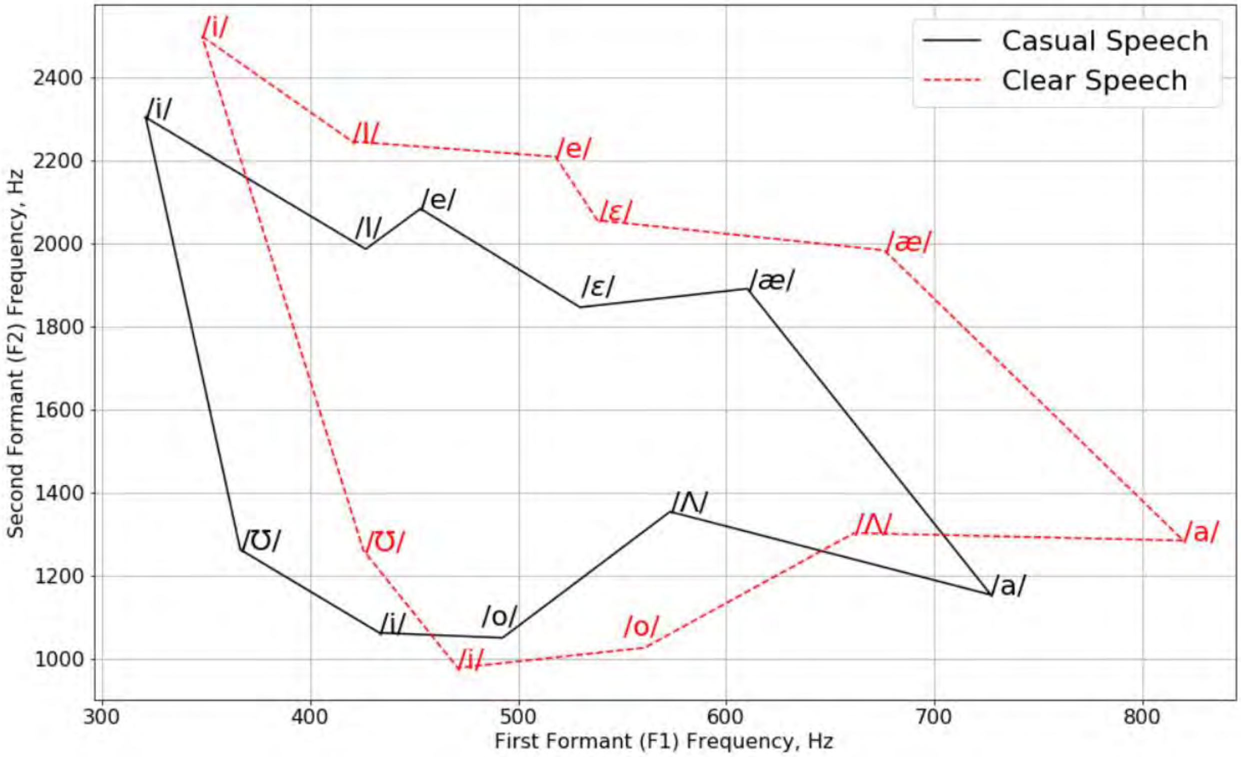
A reduction in speaking rate is one of the most common speech adaptations to increase clarity and is known to factor in three of the above four listener-directed adaptations and two of the three environmental-based adaptations. Furthermore, the rate reduction has been shown to be of benefit for four of the listener-types mentioned above (HIDS, FDS, IDS, MDS), as well as when directed to elderly listeners, and in noisy environments (LS). It is also, with increased volume or intensity, the most easily described, requiring little or no advanced knowledge of signal processing to understand. However, the change in speech rate is not uniform; some of the reduction in speech rate is due to increased pauses or pause durations. Additionally, part of this speech rate modification is part of an overall strategy to enhance voiced sounds, with the duration of vowel sounds increasing more than consonants, which has been observed in Lombard speech, speech at distance, and IDS.

Intensity, or volume increase, is intuitively easy to understand—it is the attempt to overwhelm background noise with increased power of speech, or the attempt to overcome hearing difficulties by adding additional power. However, this intuition is deceptive and incomplete; when, e.g., Lombard speech is normalized to the equivalent volume of a speaking voice, the normalized Lombard speech is often more intelligible. One reason for this is that higher intensity speech is produced by building up higher pressure in the lungs behind the vocal cords. This results in the cords opening faster and, due to higher Bernoulli pressure, closing faster. The overall result on the shape of the glottal pulses is a faster attack and decay of the pulse, which in the frequency domain results in more energy in the high frequency region of the spectrum, particularly (but not exclusively) the region near 3000 Hz where the human ear is most sensitive.

Vowel space expansions refers to a characteristic change in the formants from casual to clear speech modes. The perception of vowels is governed by the presence and position of formants in the spectrum of a voiced signal, especially the first and second formants. The frequency values of the two formants for each vowel can be treated as coordinates in a space of formant frequencies and plotted as such. (Strictly speaking, the points are centroid locations for each formant, extracted carefully from recorded data, analyzed by trained phoneticists; there is considerable variation within and between individual talkers even for individual vowel sounds.) The polygon formed by these frequency points is referred to as a vowel space. It has been observed the vowel space of clear speech is larger (i.e., the polygon has a larger area) than that of casual speech [38], with an overall upward pressure on the frequency of F1, and a general expansion of F2 values; high values become higher, while low values become lower. The overall effect of this formant expansion is to make vowel categories more perceptually distant from one another and thus easier to discriminate. See Figure 6, reconstructed with data from [20], noting that the red 'clear speech' borders extend over a visibly greater range in the second formant frequency without compression in the first formant frequency.

The fundamental pitch of a talker varies naturally over the course of speech production, and can be consciously raised or lowered by adjusting the stress and position of the vocal cords. In many forms of clear speech (including at least IDS, FDS, and LS) the natural swings of fundamental pitch are exaggerated with higher highs and

Figure 6 Vowel Space Expansion



higher lows. Exaggerated pitch excursions have also been observed in MDS in some studies such as (31, 32) but in others a reduction in the range of pitch has been observed [30]. Pitch excursions often happen within a vowel or a voiced consonant and help determine the boundaries of individual phones; this, exaggerating these pitch excursions serves to clarify and accentuate the boundaries between phonemes.

Finally, there are changes of relative intensity, especially between vowels and consonants. In Lombard speech, and speech at a distance, talkers increase the intensity as well as the duration of vowels as compared to consonants. There is strong experimental evidence to show that this aids clarity by selectively boosting the information-bearing vowels above the noise as well as allowing more time for the auditory system to "glimpse" the vowels through momentary apertures of time-varying noise. This is also seen in infant-directed speech, however, where this explanation does not hold. However, speech directed at the hearing impaired is characterized by greater intensity for consonants over vowels.

1.3.3 Efficacy

It is natural and necessary after this introduction to speech and clear speech production to ask a simple question: Do these alterations to speech cause measurable improvements in the intelligibility of the speech? There is strong evidence that they do. In the case of HIDS especially, there is ample evidence that they improve intelligibility for hearing-impaired individuals. A brief summary of important studies of the clear speech effect include:

- The seminal works on clear speech [38, 39] observed increases of word intelligibility in small sentences of 13 – 23% and extending across all phoneme categories. These studies, however, used small numbers of talkers: four talkers and five listeners.

- The study described in [40] uses 10 healthy listeners and two hearing impaired listeners, across a total of nine (for healthy listeners) or six (for hearing impaired) separate combinations of speech-shaped noise, white noise, and simulated room reverberance (including no noise, no reverberation.) The average intelligibility increase across all healthy listeners and all noise conditions was 21%, with every type of noise (including no noise, no reverberation) demonstrating increases. Hearing impaired listeners displayed an average intelligibility increase of 26% across all noise conditions.

- In a comprehensive study [41] on the effects of clear speech produced by (10 young adult healthy talkers, 10 elderly adult healthy talkers, and 10 elderly adult hearing talkers) and heard by three groups (21 young adult healthy listeners, 16 elderly adult healthy listeners, and 19 elderly adult hearing impaired), all three listener groups experienced statistically significant increases in intelligibility; unfortunately, the paper does not include a tabulation of increases for each of the nine test conditions.

- In a detailed study [42] of Lombard Speech, eight talkers produced speech in the presence of various conditions of babble noise and stationary, as well as with no noise. These recordings were normalized to the same sound level and presented to 12 listeners in speech shaped noise; in all conditions, the normalized Lombard Speech produced increases of intelligibility (from 10% to 25%) with the gains increasing as did the noise condition under which the Lombard speech was produced.

- Finally, a study by [43] using two talkers, 32 native speakers and 32 non-native speakers of English, using two levels of white noise as background masking, demonstrates that the effectiveness of clear speech in noise extends to non-native speakers as well. However, the effects were more limited for non-native speakers: In -4 dB SNR white noise, native speakers' proportional intelligibility scores improved 22% to non-native speakers' 17%. In -8 dB SNR noise, native speakers improved by 44%, and non-natives by 24%. (Note: These are proportional changes, rather than a simple subtraction or comparison, to account for the different baselines native vs non-native speaker.)

A large number of other studies have confirmed the benefits of clear speech in a variety of circumstances; two excellent surveys are [44] and [45].

1.3.4 Variability of Clear Speech Benefits

Although the benefits of clear speech are not in dispute, it is important to note that the studies referenced above draw statistical conclusions from increasing numbers of test subjects, usually as listeners but sometimes as talkers. There is evidence that these statistical measures may be hiding complex and idiosyncratic talker adaptations and listener benefits. In [20] the authors conclude that the clear speech strategies of a talker can be mismatched, and therefore ineffective or of reduced effectivity, to the hearing perceptions of a particular listener. They also speculate (but do not demonstrate) that the talker of that study differs in clear speech strategy from the talkers of [19] and [24]. Two subsequent studies provide detailed evidence for this line of thought.

First, [46] used recordings of forty-one talkers, in both clear and casual speech modes. These recordings were used for vowel intelligibility in noise tests for seven young adults with healthy ears. The results were analyzed by talker, averaged across all listeners. While the overall trend is toward increased intelligibility for clear speech, the details show a wide variability from talker to talker: The average increase in intelligibility was 8.5 rationalized arcsine units (RAUs)[5] but with many outliers more than twice that, as well as outliers showing *loss* of intelligibility for some talkers. Moreover, this study contains a striking graph of average vowel intelligibility for each of the forty-one talkers, organized with intelligibility of casual speech increasing from left to right, but with intelligibility of clear speech shown on the same graph. (See [46], specifically, Figure 2). Both talkers with large increases of intelligibility and talkers with nearly none (or even slight degradations) are clearly visible. Since these intelligibility scores are averaged across listeners, this is strong evidence either for different talker adaptations or different talker anatomies affecting the benefits of their clear speech. Although this was a study of vowel intelligibility only, it is possible that the additional context of full sentences would mitigate this effect somewhat; nevertheless, the effect is clear. Unfortunately, this study does not present the "inverse" of this graph or the data to construct it, e.g., a graph showing the effect clear speech benefits by *listener*, averaged across *talkers*.

Second, however, [43], referenced in the previous section, investigates clear speech benefits for non-native listeners, and provides almost exactly that "inverse" view of a different dataset. With two talkers (one male, one female), thirty-two native and thirty-two non-native speakers (all young, healthy listeners) there is a clear difference in the effectiveness of the male talker vs the female talker; the female produces more and greater increases in intelligibility than the male, again demonstrating that clear speech benefits vary with the talker. In addition, there is a detailed breakdown of intelligibility increases for clear vs casual speech by listener. Although the data is not

[5] RAUs are the result of using the Rationalized Arcsine Transform, a statistical transform used in audiological contexts to normalize relative or percentage data, while retaining proportionality to the original percentages over most of the transform's range (47, 48).

averaged across the talkers, the very wide variability of benefits suggests that the benefits of clear speech vary highly with the individual listener. (See [43], specifically the top two panels of Figure 2).

Taken together, these two studies provide strong support for the hypothesis that the benefits of clear speech are dependent both on the particular speech modifications made by individual talkers, and by the cognitive (based on language proficiency, age, or other similar factors) and/or listening apparatuses. Moreover, since it is impossible for a talker to have direct knowledge of the efficacy of their own attempts to provide more intelligible clear speech for the benefit of their conversational partner, the central question of this dissertation arises: Can placing control of some aspects of clear speech techniques (in the case of this dissertation, speech rate) provide benefits to the listener?

2 On Time Scale Modification

As discussed in Chapter One, one of the most salient and easily described features of naturally produced clear speech is its slower cadence, with fewer words per minute and a reduction in articulation rate as measured in syllables per second [19]. This thesis explores the effects of computer intermediation, and especially placing the control of that intermediation in the hands of the listeners, who may directly benefit from it. This chapter presents a brief historical survey of methods and algorithms used to change the rate of speech or playback of recorded audio signals.

2.1 Time Domain Techniques

The earliest historical audio recording technologies, the phonograph and magnetic tape, were intrinsically analog devices. The phonograph, invented by Thomas Edison in 1877 [49] and rapidly developed in the following decades, worked with variations along the following scheme: A sound signal of interest (typically human speech or live music) present in the atmosphere encountered a needle which was pressed against a groove in a moving and impressionable surface such as foil, wax paper, or wax. The needle, when acted on by the vibrations of atmospheric sound (often with mechanical or electromechanical mediation) would press a time series representation of the vibrations in the atmosphere directly onto the physical recording medium. To play back the recorded sound, the needle was again dragged, at the same rate, lightly over the impressions formed previously and would vibrate just as it had during the recording session; now, however, the needle imparted its vibration to a stretched diaphragm (i.e., a speaker, again, typically with mechanical or electromechanical mediation) which re-produced vibrations in the atmosphere similar to those originally played. Decades of experimentation and refinement led to a standardization of this technology in the form of a modern phonograph using electrical mediation to record and play back vibrations on the familiar vinyl disk.

In modern mathematical terminology, the phonograph needle encountered a mechanical signal of atmospheric pressure waves, $f_{in}(t)$, and pressed a mechanical copy of them, $f_{mech}(t)$, into the rotating media. During playback, the needle encountered, $f_{mech}(t)$, on the recorded media and converted it back into atmospheric pressure waves $f_{out}(t) \approx f_{in}(t)$.

Magnetic audio tape, developed in Germany in 1928 [50], itself an elaboration of the earlier, less practical magnetic wire recording systems [50] works in a similar fashion, although less reliant on mechanical mediation and physical deformations. Here, the recording media is a thin plastic ribbon coated with magnetic particles which are initially oriented randomly. To record, atmospheric sound waves are converted into magnetic fields inside an electromagnet. The magnetic tape is passed through a gap in this electromagnet, and the time-varying fields align the particles on the tape to a greater or lesser degree, depending on the field strength. During playback, the process is reversed, the magnetic fields on the tape induce electric currents in a coil of wire, and these currents drive a speaker, reproducing sound waves in the air. Again in modern terms, sound waves, $f_{in}(t)$, are converted to a signal, $f_{elec}(t)$, on the tape, and reconverted to $f_{out}(t) \approx f_{in}(t)$ during playback.

In both cases, the recordings are intrinsically analog. Faithful reproduction of the sound during playback requires moving the record (or tape) across the needle (or playback head) at precisely the same speed as during the recording process. Indeed, a considerable amount of the effort of refining these systems from experimental prototypes to mass produced commodity devices is in the design of devices which carefully control the speed of the media during recording and playback. A little physical intuition and/or experimentation leads to an obvious idea for slowing recorded speech: Simply record at one speed, say, 12 in/sec for magnetic tape, and play back at another speed, say, 6 in/sec thereby producing $f_{out}(t) \approx f_{in}(2t)$ where the output signal is simply expanded in time-- slowed, or dilated-- by a factor of two. Unfortunately, this results in an unacceptable form of distortion: the output signal is reduced in frequency by exactly the same factor by which it is expanded in time. This distortion is easily detectable at even modest expansion (or contraction) factors.

This effect can be explained informally and proven rigorously. Informally, consider a simple audio signal consisting of a single tone or a very few tones (i.e., a single sine wave or a few sinusoids.) By definition, the act of playing an audio signal back more slowly as described above means moving the peaks and troughs of those sinusoids farther apart in time, which by definition changes the frequency of those sinusoids. Formally, all audio signals can be represented as sinusoids, and can be decomposed into those constituent sinusoids with a Fourier transform. However, basic Fourier theory also shows us that the price of an expansion or contraction of a signal in time by some factor α is the contraction or expansion in the frequency domain of that same signal:

$$\alpha F(\alpha\omega) = \int_{-\infty}^{\infty} f(t/\alpha) e^{-2\pi i t} \ dt \tag{1}$$

where $f(t/\alpha)$ denotes a signal in the time domain which has been expanded by a factor of α, relative to $f(t)$.

For digitally sampled audio data, a similar argument holds—although an audio clip can be slowed down or stretched out by recording at some sampling rate $f_{s,rec}$, and playing the audio back at a different, lesser sampling rate $f_{s,play}$, the sound waves produced are stretched out (and the frequencies compressed) in exactly the same fashion as above.

2.2 Vocoding/Spectral Techniques

Because naïve methods of audio expansion introduce unacceptable shifts in frequency, more sophisticated techniques must be brought to bear. One effective category of algorithms (used extensively in the software later described in this dissertation) relies on spectral decomposition using Fourier or closely related techniques to remove the frequency modification from time scaled audio.

2.2.1 Flanagan and Golden

The first known apparatus capable of changing the time scaling of audio *without* also introducing frequency modifications was an extension of Dudley's *channel vocoder* (or 'voice encoder') [51, 52] developed by Flanagan and Golden [53] which they called a *phase vocoder,* so-called for its incorporation of phase information. Dudley's device, which was not used for time scaling, is presented here only in outline form. Briefly, the device was composed of three functional blocks:

- First, a frequency discriminator capable of estimating the fundamental pitch of a spoken voice (or, for non-voiced segments, the lack of pitch); a tunable oscillator capable of reproducing that pitch, with harmonics, for use in voiced speech segments, and a noise source, for use in unvoiced speech segments.

- Second, an analysis block consisting of ten bandpass filters of approximately 300 Hz bandwidth, followed by rectifiers. The filters were contiguous and chosen so that they covered the full useful 3 KH bandwidth of contemporary analog telephone channels. The analysis block acted analogously to a modern spectrum analyzer, capturing the separate magnitude envelopes of ten channels of audio.

- Third, a synthesis block consisting of another ten bandpass filters matching those in the analysis block, followed by amplitude modulators. During voiced segments, the output of the tunable modulator and its harmonics are sent into the filters, and modulated according to the magnitude envelopes of the analysis block. During unvoiced segments, noise from the noise source is sent into the filters and modulated. These signals are recombined into a credible approximation of the speech which produced them.

Conceptually, Dudley's vocoder breaks the voiced segments of an audio signal into multiple amplitude modulated (but not phase modulated) signals of the harmonics of a sinusoid at increasing frequencies. One feature of Golden's vocoder is that the magnitude envelopes and pitch signals, although analog, could in principle be recorded and stored for future use, or future manipulation. Another feature is the beginning of the separation of the frequency components of audio from the time domain components.

Flanagan and Golden detail a substantial improvement on this scheme, which decomposes an audio signal into multiple amplitude and approximate phase modulated sinusoids. (Hence the term 'phase vocoder'.) Their purpose in this design was the transmission of approximate signals (especially intelligible speech signals) with reduced bandwidth. However, they also noted and demonstrated that the separation of modulation into phase and amplitude components, and the use of independent oscillators, facilitates changes in frequency and time scaling. Their scheme is described below; the notation used describes their digital simulation of an otherwise analog technique.

A signal of interest is analyzed by sampling a live or pre-recorded audio signal, passing that sampled signal through some number N digital bandpass filters, and extracting the real and imaginary components of the discrete Fourier transform as follows. Using standard signal processing techniques at the time (i.e., multiplication of the original signal by sine and cosine tables prior to passage through a digital filter), filtering and Fourier transforms were accomplished in one step as follows, using their notation:

$$a(\omega_n, mT) = T\sum_{l=0}^{m} f(lT)[\cos \omega_n lT]h(mT - lT) \tag{2}$$

$$b(\omega_n, mT) = T \sum_{l=0}^{m} f(lT)[\sin \omega_n lT] h(mT - lT) \tag{3}$$

where ω_n denotes the frequency of channel n where $n \in [1 \ldots N]$, T is the sampling interval (in Flanagan and Golden, 0.1 msec), f(lT) is the lth sample of the audio signal f(t), and h(T) is the impulse response of an appropriate bandpass filter. The authors used N = 30 channels, with $\omega_n = 2\pi n\ 100\ rad/sec$, and 6th order Bessel filters of approximately 100 Hz bandwidth to provide coverage of the typical 3 KHz telephone channel bandwidth. Then, $a(\omega_n, mT)$ and $b(\omega_n, mT)$ represent the real and imaginary portions of the Fourier spectrum at ω_n and time t = m, i.e.:

$$F(\omega_n, mT) = a(\omega_n, mT) - jb(\omega_n, mT) \tag{4}$$

From this format, magnitude and phase derivative can easily be extracted:

$$|F(\omega_n, mT)| = \sqrt{a^2 + b^2} \tag{5}$$

$$\Delta\varphi[\omega_n, mT] = \frac{(b\Delta a - a\Delta b)}{a^2 + b^2} \tag{6}$$

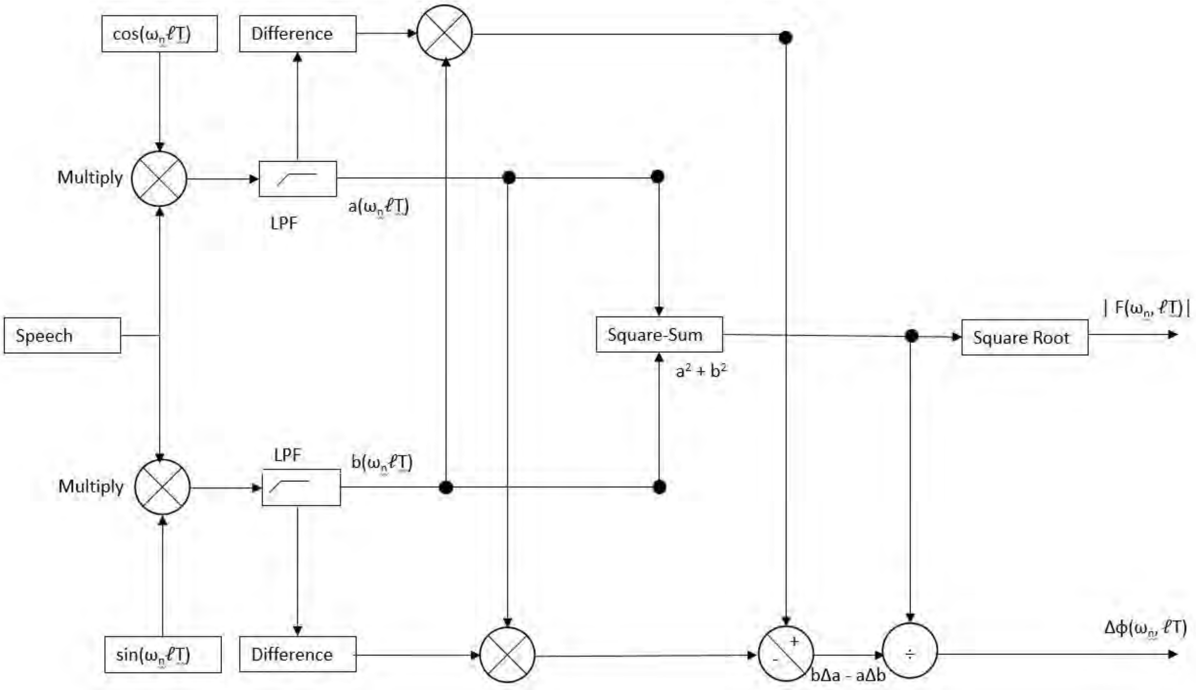
This process is depicted, for a single channel, in Figure 7.

Once this information exists, the magnitude and phase derivative signals can be transmitted directly, and each channel can be synthesized as follows:

$$\tilde{f}(mT) = |F[\omega_n, mT]| \cos\left(\omega_n mT + T \sum_{l=0}^{m} \frac{\Delta\varphi(\omega_n, lT)}{T}\right) \tag{7}$$

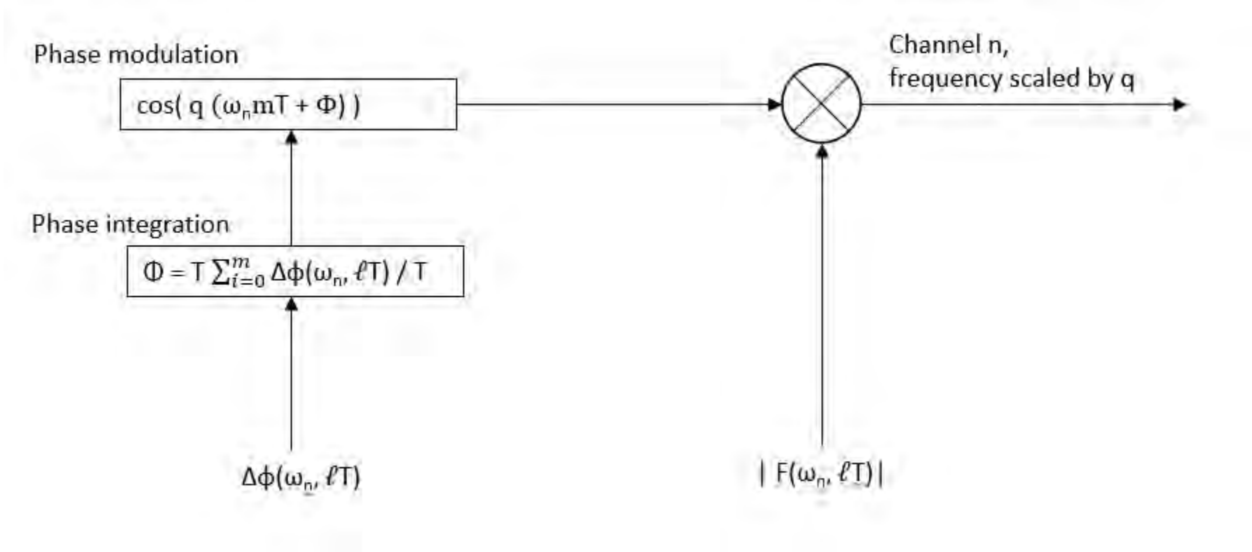
and the resultant synthesized signals recombined. Although this system is implemented and initially described as a set of bandpass filters acting on a signal, and could be implemented with analog or digital components, each filter can also be viewed as (and the authors describe them as) frequency channels in a short time Fourier transform.

Figure 7   Phase Vocoder Analysis, Single Channel



The key insight of Flanagan and Golden as regards frequency- or time-scaling is this: Since the channels of this short time Fourier transform are characterized completely by amplitude and phase modulations, the frequency of each of the N = 30 channels can be changed directly by multiplying the argument of the sinusoid function by a desired factor α, prior to resynthesis. In practice, this can be done by changing the frequency ω_n directly, and scaling the phase derivative signal separately. The magnitude modulation is not scaled. This operation, carried out on all channels, effectively scales the frequencies of the signal without scaling the signal in the time domain. In order to effect the opposite—scaling of the time domain without scaling of the frequency domain—the authors describe a two-phase process. To stretch a signal in time by a factor of, e.g., two without changing the frequency of the signal, first analyze a signal played at half its normal speed—this signal will be stretched by a factor of two in the time domain, but compressed by a factor of two in frequency domain, as described in the previous section. Then, during resynthesis, multiply the phase information by a factor of two—the signal will then be de-compressed to approximately its original frequencies without affecting the already-stretched time domain. The net result is a stretching in the time domain only. This process is depicted, again for a single channel only, in Figure 8.

Figure 8   Phase Vocoder Synthesis and Scaling, Single Channel



2.2.2 Portnoff

Although Flanagan and Golden implemented their approach on a digital computer, it was still a digital simulation of an inherently analog approach, and could easily have been implemented with equivalent analog equipment. Further, while the manuscript was received in 1966, and the authors clearly viewed their technique as a short-time Fourier based approach, it is not clear that the authors were aware of the development of the Fast Fourier Transform (FFT) in 1965 or its applicability to this work in general. However, Michael Portnoff proposed, in a series of papers, [54, 55, 56] a general framework for the representation of discrete, uniformly sampled signals based on Short-Term Fourier Analysis, its rough equivalence to the analog phase vocoder proposed by Flanagan and Golden, and its similar applicability to frequency and time scale modification. This section does not replicate the development in full, as it spans three lengthy papers and is geared to the signal processing concerns of the day; rather, it sketches the highlights using modern notation [57] and points out their applicability to the task of time scale modification.

Portnoff's interest was in the analysis and synthesis of slowly time-varying signals, such as speech. The standard analytic tool for digital signals, the Discrete Fourier Transform (DFT), if applied to such a signal x(t) as below, is not sufficient for the task. Although it can analyze a signal in the frequency or spectral domain, and resynthesize it back into the time domain faithfully, the DFT can only do so for a complete function; it gives no spectral insight to the function at finer time scales, and no insight into how the spectrum varies from time to time. This is because the inner products of the function with sinusoids of various frequencies (the fundamental part of the DFT itself) operate on the entire function, rather than individual slices thereof:

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j\omega_k n};\, \omega_k = \frac{2\pi k}{K} = \frac{2\pi k}{N} \tag{8}$$

Here, K is the size (i.e., the number of frequency bins) of the DFT, which in this dissertation is always equal to the number of samples, N.

To rectify this, a real time signal x(t) of interest is sampled at frequency F_s resulting in the sampled signal x[n], and then recast as a "short-time function" equivalent with two time indexes:

$$\mathrm{x}_{\ell}[\mathrm{n}] = \mathrm{x}[\mathrm{n} + \ell \mathrm{L}_a]\mathrm{w}_a[\mathrm{n}] \tag{9}$$

Here, $w_a[n]$ is an analysis window or analysis filter of length N, which can be slid index by index along the original time dimension x[n]. This window function is understood to have finite support, having non-zero values only at $0 \leq n < N$. In this context, n is a local time index, referenced to the beginning of the analysis window, while ℓ is a frame index, indicating the frame number, and L_a is the analysis hop or analysis stride, i.e., the number of samples that the window shifts between successive frames. This effectively recasts the single dimensional sampled signal x[n] to a two dimensional signal where the n-axis represents advancing time, and the ℓ-axis represents advancing frames. Note that this signal is non-zero only when both $0 \leq n < N$ and $0 \leq l < L$ are true. Note also that when $L_a = 1$, each frame is shifted by one sample and if frames are zero-padded, the total number of frames L in $x_l[n]$ will be N. However, when $L_a > 1$, the resulting $x_l[n]$ is decimated in the L axis by that L_a. Discussions of the analysis window and constraints on it will be deferred to later in this section.

Although this two-dimensional signal can be analyzed with a conventional two-dimensional Fourier analysis, the "partial" digital Fourier transform of the single variable n is more useful:

$$X[k,\ell] = \sum_{n=-\infty}^{\infty} x_{\ell}[n] \qquad (10)$$

$$X[k,\ell] = \sum_{n=0}^{N-1} \mathrm{x}[\mathrm{n} + \ell \mathrm{L}_a]\mathrm{w}_a[\mathrm{n}]\, e^{-j\omega_k n} \qquad (11)$$

Note, the summation index is written to only apply to specified window but must be repeated for each frame ℓ. This application of the partial DFT with respect to n, to the short time version of the signal, is the Short Term Fourier Transform or STFT of the original signal. This transform can be viewed in two ways.

The first view, with a slight re-writing of terms, the STFT can be viewed as a collection of individual DFTs, indexed by ℓ, each corresponding to an individual frame, which in turn corresponds to a small slice of the original complete signal, beginning at $\ell\mathrm{L}_a$ and ending at $\ell\mathrm{L}_a + \mathrm{N} - 1$. This provides the desired property, which is an operation on a signal which provides spectral analysis over small slices of signal which can be used to investigate the frequency behavior over small time slices. To see this, group the original signal and its window in brackets to see that summation is indexed by a single value of ℓ, and constitutes the DFT of a window function (also index by ℓ) multiplied element-wise by the total signal of interest:

$$X[k,\ell] = \sum_{n=0}^{N-1} ([\mathrm{x}(\mathrm{n} + \ell \mathrm{L}_a)\mathrm{w}_a(n)]\, e^{-j\omega_k n}) \qquad (12)$$

The inverse STFT in this framework can be understood as an inverse discrete transform of $X[k,\ell]$. However, there are some complications: First, note that this inverse transform does not convert back to the initial function indexed by time only, but to the intermediate short term function indexed by both time and frame. To see this, recall that $X[k,\ell]$ is a two dimensional function indexed on k and ℓ, and that the direct application of the inverse DFT formula applies an inverse DFT to each frame:

$$\hat{x}_l = [\mathrm{x}(\mathrm{n} + \ell \mathrm{L}_s)\mathrm{w}_s(\mathrm{n})] = \sum_{n=0}^{N-1} X[k,\ell]e^{j\omega_k n} \qquad (13)$$

It is also important to note that in general, the short term function contains redundancies relative to the original time domain function; the exception is when the window functions do not overlap, i.e., when N ≤ L_a. These redundancies, if any, can be accounted for by the overlap-add procedure:

$$\widehat{x_\ell}[n] = \sum_{\ell=0}^{L} x_\ell[n - \ell L] \tag{14}$$

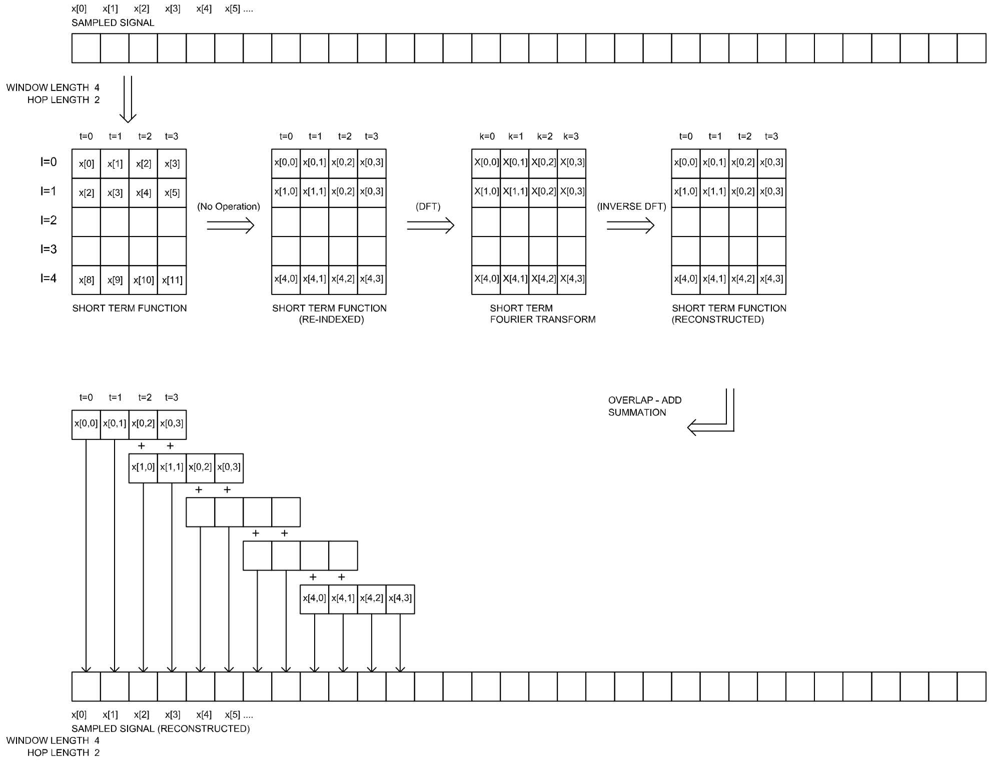
Finally, note that in order for the overlap-add procedure to result in perfect reconstruction, a condition is imposed on the analysis and synthesis windows, although by convention the synthesis window is often a uniform and flat window function. The Constant Overlap Add (COLA) condition of Equation 15 below ensures that components of each addition in Equation 14 are pre-scaled back to unity for all but the first very small number of frames. The easiest way to satisfy these constraints is with appropriately scaled uniform windows. However, there are a family of COLA windows, including Hann, Hamming, Blackman, and other windows, which (with appropriate scaling) can also satisfy this constraint.

$$\sum_{l} \mathrm{w}_a[\ell \mathrm{L} - \mathrm{n}]\mathrm{w}_s[\ell \mathrm{L} - \mathrm{n}] = 1 \tag{15}$$

This process or pipeline, from time function, to short term function, to short term Fourier transform, back to short term function and back to time function, is shown in Figure 9.

In the second but equally vlid view, the STFT can be viewed as the output of a series of digital filter banks. This is no longer interpreted as a collection of DFTs; rather, the term in brackets is considered as the element-wise multiplication of the original signal with sinusoids of frequencies of k/N times the sampling, equally spaced from 0 to 2π, indexed by n. These operations take place inside the summation, and act to shift the original signal x[n] down in frequency by the frequency of the sinusoid. In other words, the original signal is frequency-mixed, or modulated, with each sinusoid creating a new signal, whose baseband frequency contains the spectral content of the original signal near frequency k/N shifted to baseband. Call this signal $x_k[n]$ and note that now the summation can be interpreted as the convolution of that newly defined signal with the analysis window. If the analysis window is properly designed

Figure 9 STFT Frequency Domain Interpretation


x[0] x[1] x[2] x[3] x[4] x[5] ....
SAMPLED SIGNAL
WINDOW LENGTH 4
HOP LENGTH 2
t=0 t=1 t=2 t=3
l=0
l=1
l=2
l=3
l=4
x[0] x[1] x[2] x[3]
x[2] x[3] x[4] x[5]
x[8] x[9] x[10] x[11]
SHORT TERM FUNCTION
(No Operation)
t=0 t=1 t=2 t=3
x[0,0] x[0,1] x[0,2] x[0,3]
x[1,0] x[1,1] x[0,2] x[0,3]
x[4,0] x[4,1] x[4,2] x[4,3]
SHORT TERM FUNCTION
(RE-INDEXED)
(DFT)
k=0 k=1 k=2 k=3
X[0,0] X[0,1] X[0,2] X[0,3]
X[1,0] X[1,1] X[0,2] X[0,3]
X[4,0] X[4,1] X[4,2] X[4,3]
SHORT TERM
FOURIER TRANSFORM
(INVERSE DFT)
t=0 t=1 t=2 t=3
x[0,0] x[0,1] x[0,2] x[0,3]
x[1,0] x[1,1] x[0,2] x[0,3]
x[4,0] x[4,1] x[4,2] x[4,3]
SHORT TERM FUNCTION
(RECONSTRUCTED)
OVERLAP - ADD
SUMMATION
t=0 t=1 t=2 t=3
x[0,0] x[0,1] x[0,2] x[0,3]
+ +
x[1,0] x[1,1] x[0,2] x[0,3]
+ +
+ +
+ +
x[4,0] x[4,1] x[4,2] x[4,3]
x[0] x[1] x[2] x[3] x[4] x[5] ....
SAMPLED SIGNAL (RECONSTRUCTED)
WINDOW LENGTH 4
HOP LENGTH 2

(i.e., if the frequency response of the analysis window is a low-pass filter) then this convolution in the time domain is equivalent to low-pass filtering. The simplest possible example is if the analysis window is a rectangular function with $\mathrm{w_a}[n] = 1$ for $0 \leq n < N$, and $\mathrm{w_a}[n] = 0$ elsewhere; the frequency response of this rectangle function has a characteristic sinc-function shape which functions as a crude low pass filter. Other windows may provide superior filtering.

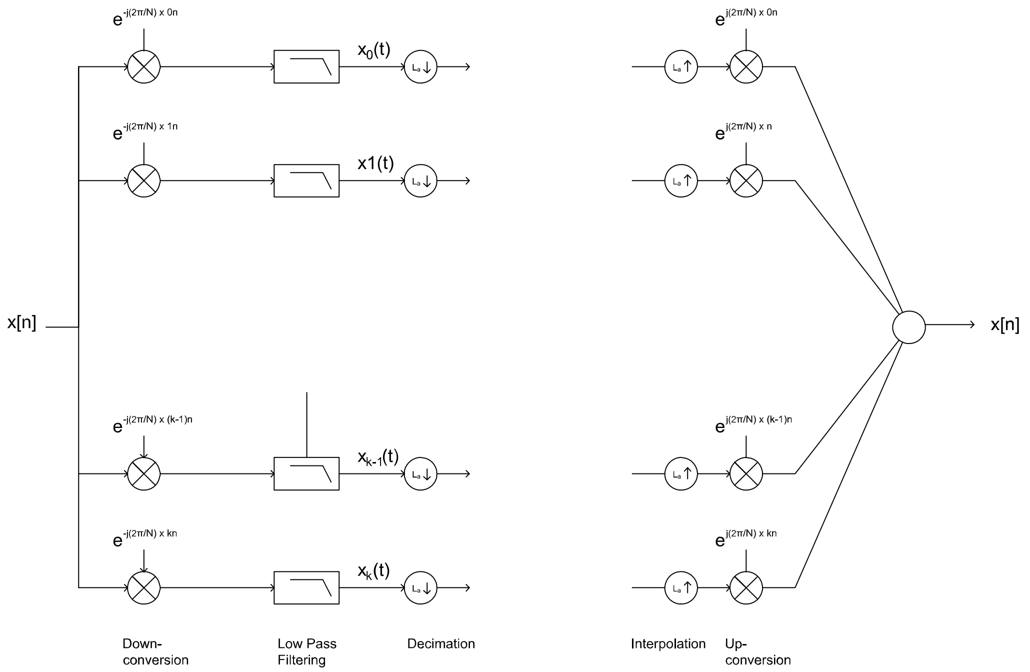
$$X(k,\ell) = \sum_{n=0}^{N-1} ([\mathrm{x}[\mathrm{n} + \ell \mathrm{L}_a] e^{-j\omega_k n}] \mathrm{w}_a[\mathrm{n}]) \tag{16}$$

$$X[k,\ell] = \sum_{n=0}^{N-1} x_k[n + \ell \mathrm{L}_a] \mathrm{w}_a[\mathrm{n}] \tag{17}$$

This filter-bank interpretation of the STFT also has an interpretation of the inverse operation, also cast as a filter-bank operation. In this case, simply recall that each channel of the filter bank n represents a signal whose baseband frequency spectrum has been down-converted from its center frequency about n/N. To undo this, simply upconvert each channel back to its original frequency and sum the outputs of the filter bank at each time sample. If the short term function has not been decimated (i.e., if the analysis stride $L_a = 1$), this is straightforward. If the short term function *has* been decimated, however, then the samples generated by this approach are decimated by the same amount. To fully restore the function (in this case, an estimate of the function), the samples must be interpolated. This can be accomplished by stretching the channel signals (i.e., adding $L_a - 1$ zero samples between each filterbank sample) and applying another lowpass filter prior to up-conversion. Both sides of this process are illustrated in Figure 10.

These two interpretations of the STFT are equivalent and best visualized as a rectangular array of rectangular cells, as shown Figure 11. In the overlapping Fourier interpretation, X(k,i) is interpretated as a series of columns at

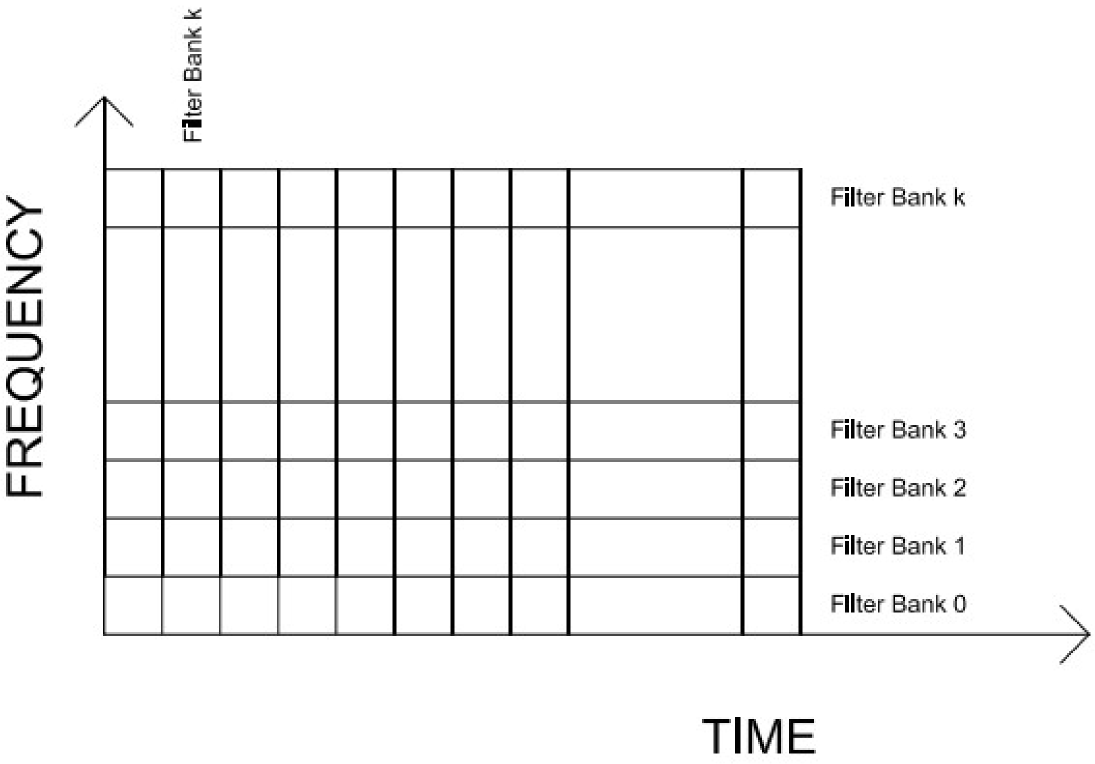
Figure 10 STFT Filter Bank Interpretation


$e^{-j(2\pi/N) \times 0n}$
$x_0(t)$
$e^{-j(2\pi/N) \times 1n}$
x1(t)
x[n]
$e^{-j(2\pi/N) \times (k-1)n}$
$x_{k-1}(t)$
$e^{-j(2\pi/N) \times kn}$
$x_k(t)$
$L_a \downarrow$
$L_a \uparrow$
$e^{j(2\pi/N) \times 0n}$
$e^{j(2\pi/N) \times n}$
$e^{j(2\pi/N) \times (k-1)n}$
$e^{j(2\pi/N) \times kn}$
x[n]
Down-conversion
Low Pass Filtering
Decimation
Interpolation
Up-conversion

separate times, where each column l is the DFT of the signal as windowed by w_a [lL – n]. In the filter bank interpretation, X[k,ℓ] is interpreted as a set of rows at separate frequencies k, where each row is a time-sequence of outputs from a digital filter at digital frequency k/N.

Finally, with this mathematical framework developed, we turn to Portnoff's central insight, as viewed through the DFT interpretation. Regardless the precise implementations of the window functions, time scaling can be achieved by varying L_a and L_s: if $L_a < L_s$, the signal will be slowed down or stretched out; if $L_a > L_s$, the signal will be sped up or compressed. However, this change of window strides introduces an additional complication, namely, phase discontinuities. This technique is illustrated in Figure 12, which shows the simplest possible case—that of time-stretching a single sinusoid. In this case, $L_a = 2$, $L_s = 4$, and the time stretching will be a factor of two. Note that the naïve application of the ISTFT results in frames with severe phase discontinuities at their edges; this is not the desired behavior, and would result in severe degradation of audio quality. To remedy this, and therefore maintain good audio quality, the phase of the value in each frequency bin in the Fourier interpretation must be updated without disturbing

Figure 11 STFT Time-Frequency View

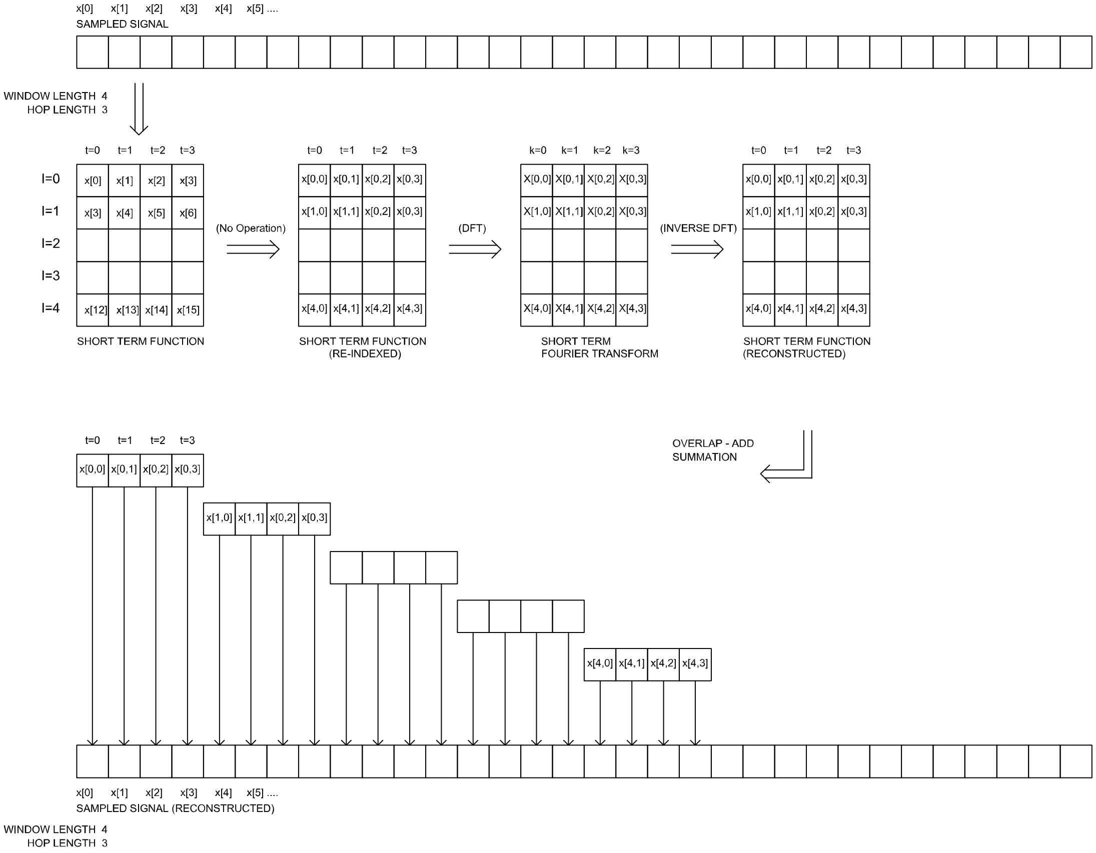


the amplitude of that value, after which ISTFT can be applied as normal. For each bin, k, a phase increment must be determined, which depends on the phase of the current and previous analysis frames, as well as :

$$\Delta_p \Phi_k^l = \angle X(l_a, k) - \angle X((l-1)_a, k) - \frac{L_a 2\pi k}{N} \tag{18}$$

Here, $\angle X(l_a, k)$ and $\angle X((l-1)_a, k)$ are the angles of the ℓth and previous analysis frames, recalling that the contents of X(l,k) are complex numbers, and the remaining term is the expected phase, in radians, of an analysis frame hop. Thus, $\Delta\Phi_k^l$ can be interpreted as a deviation between the expected phase evolution of the bin (the final subtracted term) and the observed phase evolution of the bin (the subtracted phase measurements) expressed in radians. Finally, $\Delta_p\Phi_k^l$, with the p-subscript on Δ, indicates that the phase is expressed as its principal argument, i.e., $-\pi \leq$

Figure 12 Time Stretch By Window Hopping


x[0] x[1] x[2] x[3] x[4] x[5] ....
SAMPLED SIGNAL
WINDOW LENGTH 4
HOP LENGTH 3
t=0 t=1 t=2 t=3
l=0
l=1
l=2
l=3
l=4
x[0] x[1] x[2] x[3]
x[3] x[4] x[5] x[6]
x[12] x[13] x[14] x[15]
SHORT TERM FUNCTION
(No Operation)
x[0,0] x[0,1] x[0,2] x[0,3]
x[1,0] x[1,1] x[0,2] x[0,3]
x[4,0] x[4,1] x[4,2] x[4,3]
SHORT TERM FUNCTION
(RE-INDEXED)
(DFT)
k=0 k=1 k=2 k=3
X[0,0] X[0,1] X[0,2] X[0,3]
X[1,0] X[1,1] X[0,2] X[0,3]
X[4,0] X[4,1] X[4,2] X[4,3]
SHORT TERM
FOURIER TRANSFORM
(INVERSE DFT)
x[0,0] x[0,1] x[0,2] x[0,3]
x[1,0] x[1,1] x[0,2] x[0,3]
x[4,0] x[4,1] x[4,2] x[4,3]
SHORT TERM FUNCTION
(RECONSTRUCTED)
OVERLAP - ADD
SUMMATION
t=0 t=1 t=2 t=3
x[0,0] x[0,1] x[0,2] x[0,3]
x[1,0] x[1,1] x[0,2] x[0,3]
x[4,0] x[4,1] x[4,2] x[4,3]
x[0] x[1] x[2] x[3] x[4] x[5] ....
SAMPLED SIGNAL (RECONSTRUCTED)
WINDOW LENGTH 4
HOP LENGTH 3

$\varDelta_p \varPhi_k^l \leq \pi$. This value can be used to compute an appropriate phase advance for the synthesis frames, by converting it to an instantaneous frequency as follows:

$$\widehat{\omega}_k = \frac{2\pi k}{N} + \frac{\varDelta_p \varPhi_k^l}{L_a} \tag{19}$$

Here, $\widehat{\omega}_k$ indicates the phase evolution of bin k in radians per sample, thus the phases of the synthesis frames can be adjusted as follows:

$$\angle X(l_s, k) = \angle X((l-1)_s, k) + L_s \widehat{\omega}_k \tag{20}$$

2.2.3 <u>Ellis</u>

Although the above scheme, developed by Portnoff and refined by several others is sufficient for both time expansion and time compression of audio signals, Ellis [58] has developed a conceptually simpler synthesis scheme based on the same principles. This scheme relies on the insight that the spectrum of the audio signal varies quite slowly, when compared to the length of a typical window for such analysis (which, as will be discussed in the next chapter, is typically 5 to 10 milliseconds.) As such, it is possible to set the analysis and synthesis frame hops to the same number. Ordinarily this would result in perfect resynthesis with no time-scaling. However, in Ellis' re-conception of the synthesis stage, synthesis frames are *created* by interpolating new ones from adjacent analysis frames, allowing both L_a and L_s to remain equal.

Mathematically, we begin with the typical Fourier interpretation of the STFT, where X_a represents the analysis STFT, constructed out of DFTs, the integer k indexes frequency bins, and the integer l indexes analysis frames. Each frame l represents a DFT, and each cell of X_a is a complex number, which can also be represented as a magnitude and phase.

$$X_a[k, \ell] = \sum_{n=0}^{N-1} ([\mathrm{x}(\mathrm{n} + \ell \mathrm{L}_s)\mathrm{w}_a(\mathrm{n})]\, e^{-j\omega_k n}) \tag{21}$$

To construct a synthesis STFT, $X_s[k, \ell]$, allow ℓ to range over the field of positive real numbers, rather than integers, as a mathematical convenience. A value of l = 4.5 will connote a temporal position exactly between frames 4 and 5; a value of ℓ = 4.2 will connote a temporal position one fifth of the way from frame 4 to frame 5, etc. The meaning of the k-index is unchanged and remains bound to integer values. With this understanding, the new frames are constructed as follows: First, ℓ_s serves as a cursor, marking progress through the STFT array. For an expansion factor of β (where β = 2 implies stretching the audio out to twice its length), each successive frame adds 1/ β to the ℓ_s index. In other words, to stretch speech by a factor of two, new frames are interpolated exactly between each existing analysis frame. Then, the magnitudes of the new frame are calculated by direct linear interpolation, as below:

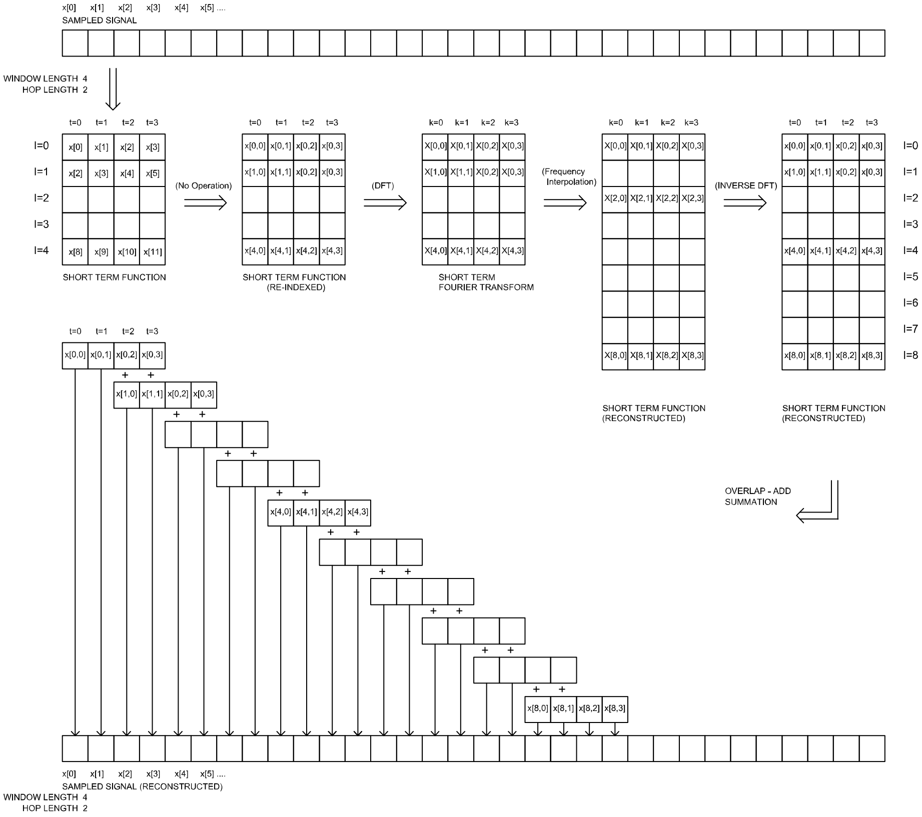
$$MAG\big(X(l_s,k)\big) = Mag\big(X(\lfloor l_s \rfloor, k)\big) + \frac{l_s - l_{\lfloor s \rfloor}}{l_{\lceil s \rceil} - l_{\lfloor s \rfloor}}\Big(Mag\big(X(\lceil l_s \rceil, k)\big) - Mag\big(X(\lfloor l_s \rfloor, k)\big)\Big) \tag{22}$$

The phases are determined by advancing phase from frame to frame similar to in Portnoff's scheme. The instantaneous frequency, $\widehat{\omega}_k$, is determined as normal, from the inter-indexed analysis frames, but advanced by a fractional value of that according to the real-valued synthesis frame:

$$\angle X(l_s,k) = \angle X\big(l_{\lfloor s \rfloor}, k\big) + L_s\big(l_s - l_{\lfloor s \rfloor}\big)\widehat{\omega_k} \tag{23}$$

This process is illustrated in Figure 13.

Figure 13 Time Stretch by Interpolation


x[0] x[1] x[2] x[3] x[4] x[5] ....
SAMPLED SIGNAL
WINDOW LENGTH 4
HOP LENGTH 2
t=0 t=1 t=2 t=3
l=0 x[0] x[1] x[2] x[3]
l=1 x[2] x[3] x[4] x[5]
l=2
l=3
l=4 x[8] x[9] x[10] x[11]
SHORT TERM FUNCTION
(No Operation)
t=0 t=1 t=2 t=3
x[0,0] x[0,1] x[0,2] x[0,3]
x[1,0] x[1,1] x[0,2] x[0,3]
x[4,0] x[4,1] x[4,2] x[4,3]
SHORT TERM FUNCTION
(RE-INDEXED)
(DFT)
k=0 k=1 k=2 k=3
X[0,0] X[0,1] X[0,2] X[0,3]
X[1,0] X[1,1] X[0,2] X[0,3]
X[4,0] X[4,1] X[4,2] X[4,3]
SHORT TERM
FOURIER TRANSFORM
(Frequency
Interpolation)
k=0 k=1 k=2 k=3
X[0,0] X[0,1] X[0,2] X[0,3]
X[2,0] X[2,1] X[2,2] X[2,3]
X[8,0] X[8,1] X[8,2] X[8,3]
SHORT TERM FUNCTION
(RECONSTRUCTED)
(INVERSE DFT)
t=0 t=1 t=2 t=3
x[0,0] x[0,1] x[0,2] x[0,3] l=0
x[1,0] x[1,1] x[0,2] x[0,3] l=1
l=2
l=3
x[4,0] x[4,1] x[4,2] x[4,3] l=4
l=5
l=6
l=7
x[8,0] x[8,1] x[8,2] x[8,3] l=8
SHORT TERM FUNCTION
(RECONSTRUCTED)
OVERLAP - ADD
SUMMATION
t=0 t=1 t=2 t=3
x[0,0] x[0,1] x[0,2] x[0,3]
+ +
x[1,0] x[1,1] x[0,2] x[0,3]
+ +
+ +
+ +
x[4,0] x[4,1] x[4,2] x[4,3]
+ +
+ +
+ +
+ +
+ +
x[8,0] x[8,1] x[8,2] x[8,3]
x[0] x[1] x[2] x[3] x[4] x[5] ....
SAMPLED SIGNAL (RECONSTRUCTED)
WINDOW LENGTH 4
HOP LENGTH 2

3 On System Implementations

The previous chapter gave a theoretical and mathematical grounding for the construction of audio dilation or expansion software; this chapter describes the incremental development and realization of audio dilation software as hardware and software systems. The first two sections of this chapter, respectively, are based on previously published papers, while the final section details unpublished development work:

Novak III, John S., Jason Archer, Valeriy Shafiro, Robert V. Kenyon, and Jason Leigh. "On-line audio dilation for human interaction." In *INTERSPEECH*, pp. 1869-1871. 2013.

Novak III, John S., Aashish Tandon, Jason Leigh, and Robert V. Kenyon. "Networked on-line audio dilation." In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pp. 255-258. 2014.

3.1 Single-Talker Half Prototype

3.1.1 Requirements

The first audio expansion software developed for this dissertation grew out of the first CS94 "Human Augmentics" seminar, and was implemented primarily as an exploration and a prototype with three basic goals: To determine a viable approach; to allow enough real time support and control for experts to "tinker" with it and conduct informal experiment with it; and to determine where any unexpected difficulties might lie, be they related to software, user interface, or audio quality. Informally, the software was conceived as only the core algorithm with a minimal user interface wrapped around it—the application would take in speech, expand it in real time, and play it back in real time. Because all of this was to be implemented on a single laptop without the capability to send or receive remotely, the prototype was best used with a headset with boom microphone to avoid immediate audio feedback. As such the first four requirements were determined in collaboration with faculty members. First, and most importantly, the application must employ an on-line algorithm, here understood not as a networking requirement but as a requirement that the application must be able to proceed with and process partial inputs as they are received. In other words, the application must not wait for a complete utterance or other unit of speech, but must begin the processing and output whenever data is present. This requirement is fundamental to the liveness of the operation—without this condition, the application could not (in later versions) emulate real-time conversations or live speech, and would be unable to change expansion factors in real time. Second, the application must be capable of receiving speech from a microphone and transmitting it from a speaker simultaneously (although at any instant, what is transmitted may be a delayed instance of what is being received, due to the inherent time-lag of audio expansion.) Third, the user interface must allow for both push-to-talk and push-to-listen capabilities. Fourth, the user interface must allow the user to change over a wide range of expansion factors (including no expansion, but not including contraction.)

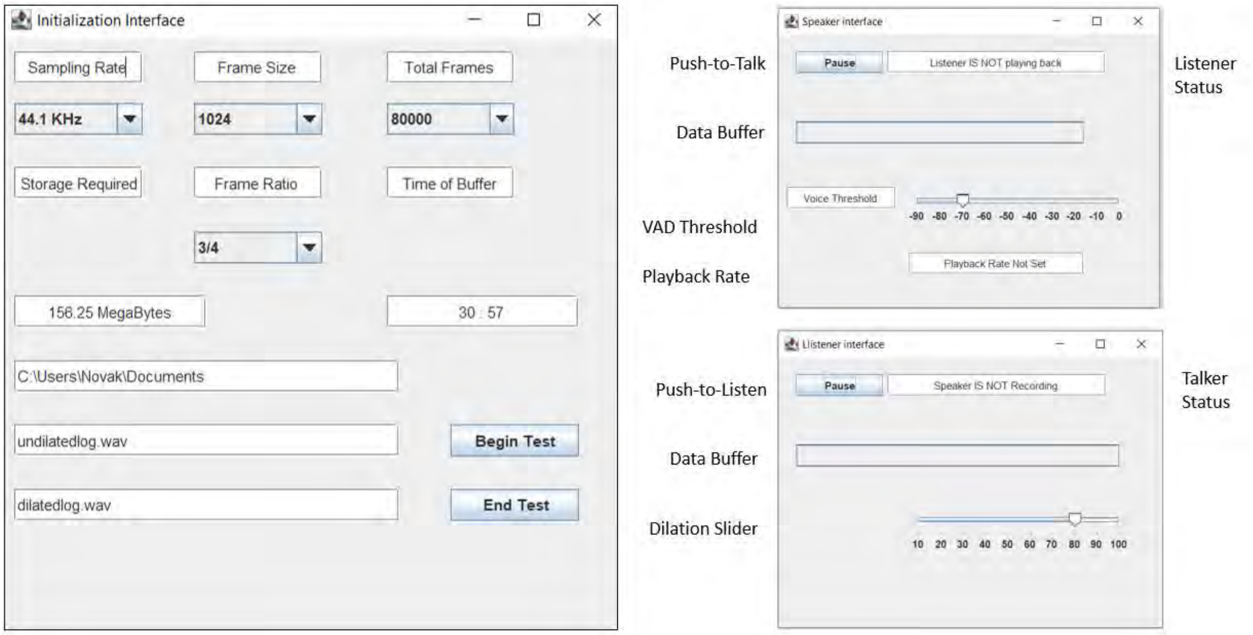
As the system was developed and the platform was considered for early formal user studies, an additional three requirements were added: fifth, the system must track and display memory consumption; *sixth*, the system must record both the speech received and the speech transmitted for later analysis; and *seventh*, the system must incorporate a simple, proof of concept Voice Activation Detection (VAD) system, so that pauses in speech would not be expanded as speech would be.

The system as designed during that seminar and the summer afterward met all these goals.

3.1.2 User Interface

The user interface was designed with Java Swing, and (aside from a brief start-up screen asking where files should be stored) was factored into three panels: An initialization screen, a talker-oriented screen, and a listener-oriented screen. This separation of the user interface was not strictly required for this prototype application but was designed with an eye forward to more complete, networked-based implementations and for early user studies. On the initialization screen (Figure 14, left) technical information (sampling rate, frame size, frame hop, filenames, etc.) could be controlled at start-up, and would typically not be used again during a session. In practice, the defaults shown, other than filenames, were generally correct as populated.

Figure 14 Audio Dilation Single-Talker Prototype UI



Initialization Interface

Sampling Rate	Frame Size	Total Frames
44.1 KHz	1024	80000
Storage Required	Frame Ratio	Time of Buffer
	3/4	
156.25 MegaBytes		30 : 57

C:\Users\Novak\Documents

undilatedlog.wav

dilatedlog.wav

Speaker interface

Listener IS NOT playing back

Playback Rate Not Set

Listener interface

Speaker IS NOT Recording

The talker-oriented screen (Figure 14, upper right, referred to then as ‘Speaker Interface’) contained a push-to-talk button and a voice threshold slider with thresholds ranging from 0 to 90 dB below the top of the input device’s dynamic range, in 10 dB increments. The VAD function is described below, but this slider allowed the talker to determine a threshold of silence below which the application would not record even with push-to-talk engaged. This screen also contained three indicators: one showing whether the listener function (controlled by a different window) was active, one showing what the dilation factor the listener had chosen, and one showing how much of an internal buffer of raw audio data had been filled. This last indicator would show the buffer filing up as the talker spoke into it, and emptying out as the listener listened to it, and was intended to serve as an informal indication of how far behind a conversation a dilated listener might be.

The listener-oriented (Figure 14, lower right, referred to then as ‘Listener Interface’) screen was similar, with a push-to-listen control and a dilation factor slider. With no foreknowledge of what ranges of dilation would be considered reasonable or necessary, I allowed for a dilation range of 10% (an expansion factor of 10) through 100% (no expansion) in increments of 10%. No provision was made for contraction rather than expansion. This slider locked in its value only when the user adjusted it and ended the adjustment with a mouse-up; “chirping” the rate of dilation was not allowed. It also had an indicator showing whether the talker function was active, and the same buffer indicator as the previous talker-oriented screen.

(Note that the interfaces of early prototypes were designed around a notion of “audio dilation” rather than the “audio expansion” more typically used in other literature. Audio expansion, α, is defined as the ratio of playback time to originally recorded time; large values-- greater than one-- indicate increased expansion or stretching. Audio dilation is defined as the ratio of originally recoded time to playback time, or 1/α; small values-- less than one-- indicate increased expansion or time-stretching. Experimental results where necessary are converted to the more conventional notation. In no circumstance were experimental subjects shown labeled interfaces.)

3.1.3 Implementation

The requirements of simultaneous receive and transmit with a live user interface required a programming language with robust threading and concurrency support. On that basis, and the informal idea that this software might one day run on different platforms, Java 6 was chosen as the implementation language. In the following discussion, *and everywhere in this dissertation unless otherwise noted* (although the software allowed for slightly more flexibility), we assume that the sampling rate was 44.1 KHz (i.e., CD quality sampling, easily received and transmitted by a laptop computer), the sampling resolution was 16 bits, the frame length as described in the previous chapter was 1024 samples, and the overlap was ¼ corresponding to a hop length of 256 samples (23.2 msec and 5.8 msec, respectively.) This frame length was chosen to ensure that each frame of voiced speech would contain at least one full glottal cycle of a low-pitched voice; lower end of human speech fundamental frequency is approximately 85 Hz [59] with a corresponding glottal cycle of 11.8 msec. The application receives and transmits sound stored in the WAV format, and converted from and to it for numerical processing; this dissertation does not discuss the low level concerns of these conversion processes.

The basic architecture of the software was a sequence of two producer-consumer pipelines: The first linking the physical microphone input (the first producer) to a processor or dilator of frames (the first consumer), and the second linking that processor/dilator of frames (the second producer) to the physical speaker output. Java’s concurrency libraries provide a simple tool for managing these pipelines—the thread-safe Linked Blocking Queue

(LBQ), which allows read actions only when not empty and allows write actions only when not full. One Linked Blocking Queue was used as part of each Producer-Consumer pair.

At the receiver/microphone, incoming microphone samples are placed into a Receiver LBQ only when three conditions were satisfied: The user interface push-to-talk was engaged, the Receiver LBQ was not full, and the VAD (to be described later) was not throttling the record function. When active, data is recorded in 256-sample "sub-frame" chunks, which can easily be assembled to 1024-sample full frames. The Dilator acts as both the first consumer and the second producer, and was implemented as follows: During start-up, as soon as five subframes of data are present, they are used to create two complete frames of data, a "past" frame comprising subframes 0 through 3, and a "future" frame comprising subframes 1 through 4. From these past and future time domain frames, past and future frequency domain frames are created.

At this point, the Dilator function enters a loop, during which the function of audio dilation is realized: On each pass through the loop, one interpolated frequency domain frame *between* the past and future frames is calculated from them. This happens exactly as discussed in the previous chapter, with interpolation of magnitude components, and appropriate phase advances of the angle components. Each time through the loop, the dilation value is updated to reflect the current slider position on the Listener interface, to ensure liveness. The position or status of the interpolation is maintained in the software and updated after every pass through the loop. If, for example, the dilation factor is currently set to 30% and the position tracker is at 0.5, the tracker advances to 0.8 and interpolates the next frame accordingly. This newly interpolated frame, still in the frequency domain, is then transformed back into the time domain by inverse fast Fourier Transform and re-divided into subframes. Output subframes are held in a four subframe ring buffer. Each time through this loop, each subframe is then added sample-wise to the ring buffer as appropriate. Each time, this process completes one subframe in the ring buffer, and the newly completed subframe is sent to the Transmitter LBQ to be converted into microphone output, erased, and the ring buffer header position is updated.

If at this point the position tracker is less than or equal to 1.0, the loop continues as before. If, however, the position tracker is greater than 1.0, a new interpolation frame cannot be computed—said frame would no longer lie between the past and future frames, so those frames must be updated. At this point, the future frame is copied to the past frame, a new future frame is constructed from the final three subframes of the new past frame, and a new subframe read from the LBQ of raw audio data; this can only happen when the push-to-listen function is engaged and when there is data in the LBQ of raw data (whether or not the system is accumulating more data into the receiver LBQ), so the system halts until those two conditions are met. Once they are met, the loop continues, and additional frames are interpolated and sent to the output queue.

The Voice Activation Detection (VAD) was added to the list of requirements as soon as the prototype was robust enough for informal experimentation; it quickly became apparent that while the push-to-talk feature was a necessary part of an interface, it was not sufficient, because it was too easy to forget to use. As a result, long stretches of talker silence would be dilated into even longer stretches of listener silence. The VAD function was implemented very simply, as befitting a prototype application: Prior to entry into the Receiver LBQ, an RMS value relative to the maximum possible input volume was calculated for each frame. If more than 150 consecutive frames were below the user-specified threshold, no further frames would be added to the LBQ until at least one frame exceeded that threshold. Due to the overlap of frames, this amounted to a somewhat arbitrary span of 0.88 seconds, which was well in excess of most inter- or intra-word pauses, but much shorter than the pauses encountered in turn-taking dialogue. Although strictly speaking this was not a Voice Activation Detection system but a Sound Activation Detection system, it was sufficient for a prototype; when the laptop software was used with headphones and boom microphones, it was more than capable of discriminating between typical laboratory background noise vs intentional speech.

The system as described above meets all criteria: Java's concurrency library operating on a typical Windows laptop was more than sufficient to manage all the various threads, including the mathematically intensive Dilator class with its constant FFT and IFFT operations, in real time. The definition of "in real time" is necessarily somewhat slippery, however. As this prototype application allowed only audio expansion and not contraction, the Listener side of the audio was nearly always somewhat behind the Talker audio. However, when playing and dilating, the necessary Fourier operations were performed much more quickly than it took to play them, so the audio was smooth with no perceptible lag between frames. (In practice, the Transmitter LBQ was found to be unnecessary as a result.) Likewise, when the slider was used to change the rate of play, this also took effect without perceptible delay. The range of expansion from no dilation to 10x dilation was in fact much wider than practical; after a few initial tests, most explorations limited the range from 3x to 1x. The user interface incorporated push-to-talk, push-to-listen, displayed memory consumption, and arranged for recording true input and modified output if required. Finally, VAD was added afterward into the talker end of the system.

3.2 Two-Talker Wireless Full Prototype

The initial prototype, while it successfully met its goals, was also severely limited by its design as a local-only application without the ability to send to or receive from distant locations, except through audio cables. (Chapter Four describes a user study performed in this way, with long audio cables strung between two laptops.) The second major version of this software, while still a prototype, was a more complete prototype designed to run on pairs of Android smart phones, each with the capability to transmit and receive. A consequence of writing the software of the initial Half Prototype in Java was significant code re-use; specifically, the data structures and the core dilation algorithm are effectively unchanged from the Half Prototype, and are not recapitulated in this section.

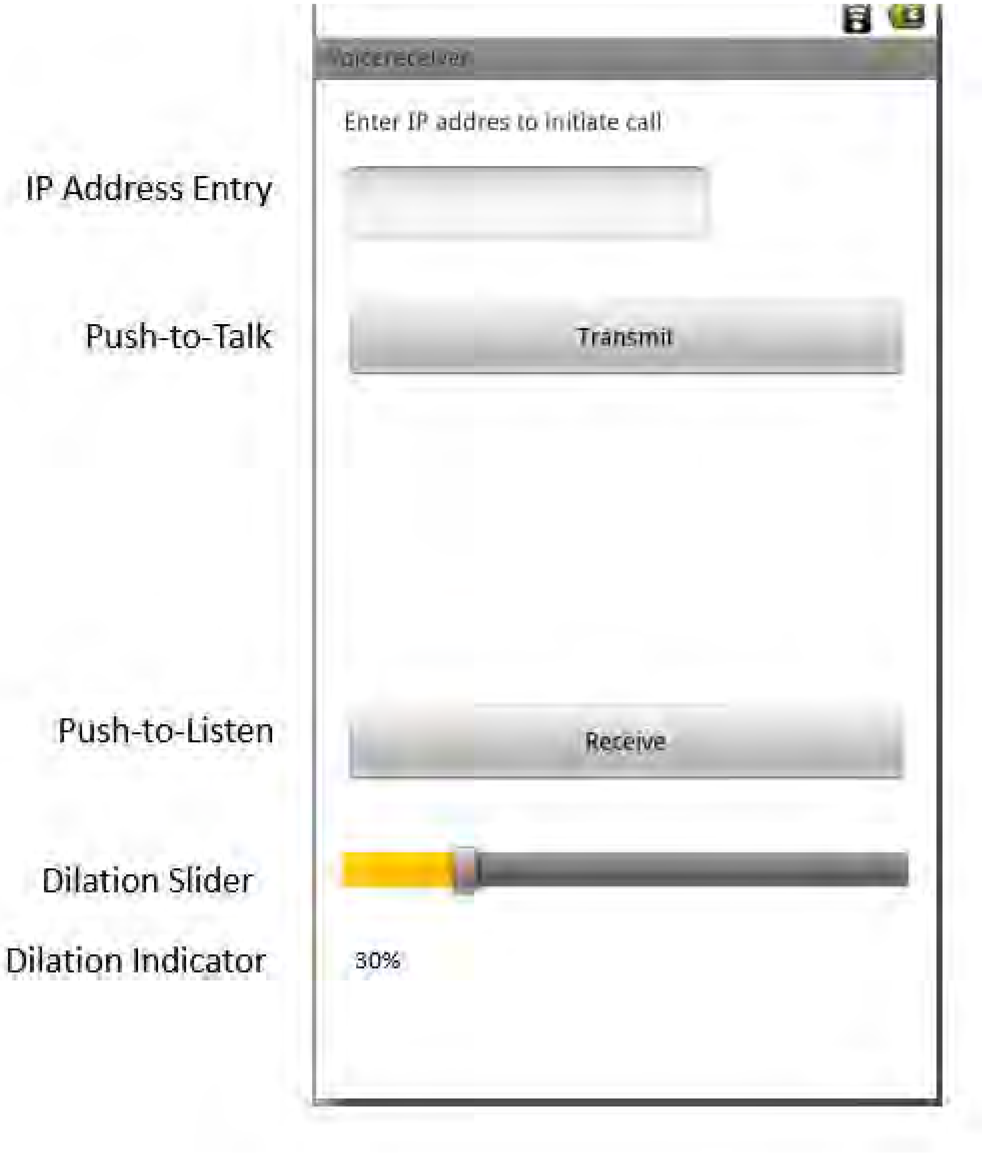
3.2.1 Requirements

The Wireless Prototype retains most of the requirements of the Half Prototype, with the exception of a requirement to log raw input and dilated output; the prototype was not intended to support user studies, but only to investigate the potential of an application supporting full, two-way conversations on handheld devices. As such, most requirements from the laptop prototype system were carried over (with the exception of data-logging and memory consumption updates) but with these additional requirements and clarifications. Now, rather than a single-person standalone application, two android devices must connect in a controlled wireless environment where IP addresses are known (i.e., a controlled laboratory environment, in a single home with wireless networking; the demonstration floor of a convention.) When connected, each handset must be able to simultaneously transmit and receive. The user interface must allow the user to control the dilation rate of *incoming* speech only, received from the other paired handset. And ideally, this should be accomplished with identical software installations on each of the paired handsets.

3.2.2 User Interface

The user interface for this prototype was developed in Android 4.0 and presents a simple interface, which is identical on both pairs of devices. The UI components, starting from the top of a portrait-oriented smart phone, are: (1) A text field for entering the IP address of the phone to connect with, (2) a transmit (i.e., push-to-talk) button, (3) a receive (i.e., push-to-listen) button, (4) a slider to control the rate of dilation, and (5) a text view displaying the current dilation setting. To initiate the software, operators of each Android device would enter the known IP address of the other device into the text field. After this, the Transmit and Receive buttons would act as push-to-talk and push-to-listen buttons, respectively; each button would change its text label and color to indicate the change of status (e.g., transmitting vs not transmitting). The slider bar at the bottom serves to control the dilation rate of incoming speech, while the text field at the bottom would change to report the precise dilation value setting. (Numeric tick-marks on the slider were not employed, as they would have been too small to conveniently read.) This interface is shown in Figure 15, below.

Figure 15 Audio Dilation Wireless Two-Talker UI


Voicereceiver

Enter IP addres to initiate call

30%

It should be noted that, as with the initial laptop prototype, this application is best used with headphones. First, because the push-to-talk and push-to-listen buttons and sliders are on the screens of the devices, operation is difficult if the devices were cradled against the head as telephones often are; also, this software could also be used on Android *tablets*, where such phone-like cradling is simply not feasible. With the device operating while held in the users' hands but without headphones or other listening attachments, the only way the audio can be heard is if it was loud enough to be picked up by the device's own microphones, which would send the output of one handset directly back to the other handset. In the worst case, with both handsets in this mode, audio feedback would result. Headsets eliminated this unwanted feature.

3.2.3 Implementation

This system is implemented in three main threads, a Receiver, a Transmitter, and Dilator, as follows. Each handset manages communication with its counterpart via a pair of datagram sockets, one each for sending and receiving. Datagram sockets were chosen over streaming sockets due to their lower overhead, faster transmission, and overall finer level of control; as is typical in VOIP applications, these factors typically outweigh the lack of a guarantee of delivery or correct order in robust networking environments. Each packet sent contained a single subframe.

The Transmit button toggles between transmit and non-transmit modes by creating and destroying the transmit socket as well as the Android-native resources for recording samples from microphone. In transmit mode these resources are created, activated, and send datagram packets containing sub-frames of audio data subject to the VAD algorithm described above. In non-transmit mode, these resources are shut down and removed. These activities are coordinated by the Receiver thread.

The Receive button similarly toggles between receive and non-receive states by creating and destroying a receiver socket, as well as Android-native resources for playing samples through the speaker. These activities are managed by a Receiver Thread. However, when receiving packets (i.e., sub-frames) from the paired handset, these packets were not played directly, but instead stored in a Linked Blocking Queue. This LBQ would be read by the third and final thread (the Dilator Thread), which consumed them, assembled them into adjacent frames, and performed frequency-domain magnitude interpolation and phase advances in a loop just as described for the laptop prototype. The liveness of the dilation slider was maintained in this loop by polling the slider value each time through the loop and adjusting the interpolation and phase advance parameters as needed. The resultant interpolated frames were converted back to the time domain and sent directly to the speakers.

This implementation also met all its requirements and, in controlled settings where IP addresses were known, stable, and on the same network, was able to provide audio dilation in real time with live and responsive controls—in part because the requirement to be on the same wireless network reduced network delays to single digits of microseconds. Such settings include controlled laboratories, individual homes, and the demonstration floors of a conference. However, the interface, especially the initialization procedure which required knowing and manually entering the IP address and waiting for the partner handset to do the same, was extremely cumbersome.

3.3 Full Transmit-Receive Server Prototype

The final prototype implemented was a server-mediated system intended primarily to demonstrate the possibility of long distance communications with this technique. The simplest conceptual way to do this was to acquire and use a Windows desktop-class machine to use as a server. This server had a known, stable IP address exposed to the outside world, which could then be connected to multiple pairs of handsets, and intermediate between those pairs of handsets to provide simple Voice Over Internet Protocol (VOIP) channels with dilation functions built in. A secondary goal was a cleaner, more unified user interface.

3.3.1 Requirements

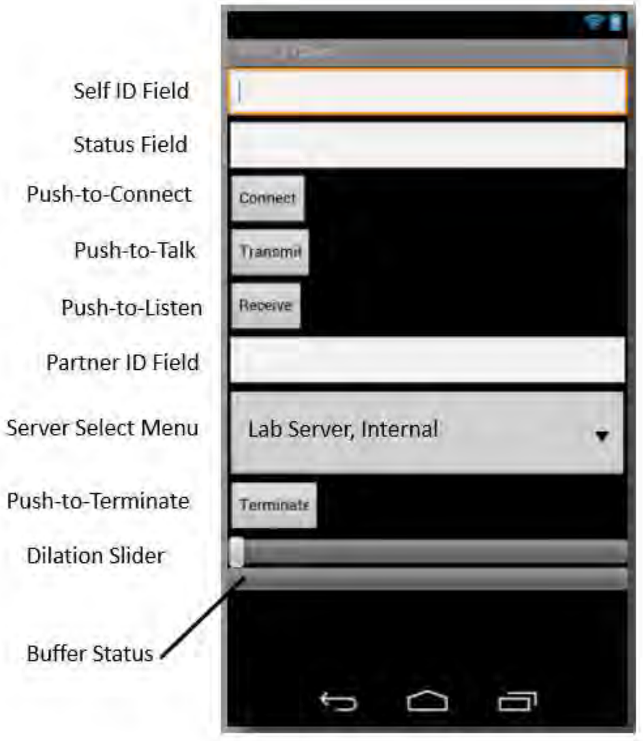
The server mediated prototype is a continuation of the previous Half Prototype and the Wireless Full Prototype, and shares many requirements with them. However, due to the growing complexity of the system, the complete set of requirements are presented: Each handset must allow the selection of a server from a simple, drop-down menu (in contrast to the Wireless Protocol, which required the users to remember and type out IP addresses by hand.) Each handset must present a clean button interface for connecting to and disengaging from the server, and for push-to-talk and push-to-listen functionality; talking/transmitting and listening/receiving must both be possible at the same time. Each handset must also allow for on-line dilation of incoming speech (similar to the Wireless Prototype) and must display an estimate of the amount of conversation still in queue at the current rate of dilation. Finally, the server itself must listen for handset connections in pairs and connect them to each other as they come in.

3.3.2 User Interface

The user interface of the handset is similar to, but more elaborate than, the user interface of the previous Wireless Prototype. From top to bottom (in a portrait-oriented device) the interface consists of a column of: (1) An identity text field (unused); (2) a status text field (to indicate whether the device was paired with another device or not), (3) connect, transmit, and receive buttons (to initiate a connection to server, to initiate push-to-talk, and to initiate push-to-listen, respectively); (4) a partner text field (unused); (5) a dropdown list (to select a server); (6) a terminate button (to unpair the devices); (7) a slider (to set dilation); (8) a buffer status bar (to indicate the amount of audio remaining at the server, at the present speed of playback.) This interface is shown in Figure 16 below.

To pair the handset to another, a user first selects a server from the dropdown list and presses the Connect button. If this is the first handset connecting, the user must wait until a second handset also completes this process; if the second, then the server would connect the two handsets. Note that this means handset pair occurs in the order that the pairing requests come in-- the first two handsets, the second two handsets, etc. This implementation was made to

Figure 16 Audio Dilation Server Mediated Prototype



simplify initial programming of both the handsets and the server. An additional feature, assigning names to handsets and allowing users to request specific other users to pair with, was never implemented, although vestiges of it can be seen in the unused "Identity" and "Partner" fields. Once paired, the interface is straightforward and similar to the Wireless Prototype: Transmit and Receive buttons act as push-to-talk and push-to-receive, the slider acts to change the dilation factor for the incoming speech, and the terminate button ends the pairing, while the progress bar is an indicator of how much conversation is queued up and yet to be played in the server. Note that the server, has no user interface beyond logging and debugging outputs designed for development work. Users were not intended to interact with the server in any way other than by pairing handsets to it.

3.3.3 Implementation

The separation of the software into a server and pairs of clients has significant ramifications on the system implementation. Simply stated, the handsets and handset software are now truly clients of the server and are responsibly primarily for taking sound from a microwave, converting it to sub-frames and sending those sub-frames to the server; as well as receiving sub-frames from the server, converting them to samples, and sending them to a speaker output. The only remaining responsibilities of the handset are to send and receive commands and status messages to and from the server (i.e., "Toggle push-to-talk" or "Update remaining playback time"). As such, the implementation of the handset software is greatly simplified, consisting of the following threads: First, a Connection thread engages during the handset pairing to the server, and sets up the resources required including sockets for transmit, receive, and UI message passing. After this (and after another handset is paired to it) a Parser thread monitors the message socket for incoming messages from the server and manages appropriate updates to the UI on that basis. (Outgoing messages do not need to be parsed, and are sent to the message socket directly from the local UI.) Finally, a Transmit thread manages local talk-related resources, collecting samples from the microphone, converting them into subframe-sized packets and sends them through the transmit socket; at the same time a Receive thread manages local listen-related resources, taking subframe-containing packets and placing them into a Linked Blocking Queue, which is read by a Playback thread before sending audio to a speaker resource. Note that unlike the Wireless Prototype, the numerical processing of the audio dilation is not handled locally on a handset, but remotely on the server.

The server implementation is considerably more complex. Conceptually, the server treats each pair of handset as a *pair* of software-implemented transceivers: one to receive data from handset A, dilate or pause appropriately per messages from the handsets, and transmit the results to handset B; one to perform the same functions from handset B to handset A. Each transceiver object employs multiple threads in an architecture similar to the original laptop prototype, except wrapped in sockets. A Parser listens on message sockets for messages incoming from the handsets, the most important of which were the push-to-listen (so that the server could change the status of the thread transmitting to that handset), and the dilation slider (so that the server could adjust necessary constants in the dilation algorithm). Very similarly to the laptop implementation, a Receiver thread takes subframes from the Transciever's receiver socket and places them into a Linked Blocking Queue for consumption by a Dilator thread. The Dilator thread performs all of the Fourier/inverse Fourier transforms, and all the interpolations and phase advances necessary for the created of new audio subframes, and passes them to another Linked Blocking Queue which are consumed by the Transmitter thread and sent by socket to the receiving handset. At any moment, Transceiver's receiver may or may not be receiving data, its transmitter may or may not be sending data, and the dilator may change its dilation factor at any moment based on commands from the receiving handset; the multiple LBQs in each transceiver are necessary to effect the necessary "liveness" of the application. Most importantly, the Dilator thread should be creating new frames only when needed; if they are processed as soon as received, the dilation slider control would have considerable lag.

This implementation was never published. Due to hardware limitations (i.e., not enough handsets) I was unable to determine how many paired phones the Windows server could support. However, two pairs of handsets were not challenging—audio quality remained high and responsivity was good. Additionally, the application was tested with one pair of handsets separated from the server by several hundred miles, with similar results: good audio quality, no lagging or stuttering, and no perceptible lag in the UI.

4 Audio Expansion in Real Time Conversation

The prior two chapters addressed the mathematical and software foundations of real-time audio expansion software. The present chapter uses the first prototype described in Chapter Three to conduct a truly interactive, two-participant study, investigating the effects of real-time expanded speech on the dynamics of spontaneously produced conversation.

This chapter is adapted from the previously published paper in the *Proceedings of Meetings on Acoustics*:

Novak, J. S., Archer, J., Kenyon, R. V., & Shafiro, V. (2015, May). Audio dilation in real time speech communication. In *Proceedings of Meetings on Acoustics 169 ASA* (Vol. 23, No. 1, p. 050008). Acoustical Society of America, with the permission of AIP publishing.

4.1 <u>Introduction</u>

While there is broad agreement that naturally produced clear speech provides real, measurable benefits to listeners who are elderly, have impaired hearing, or are in challenging auditory environments, these benefits are often studied in very artificial laboratory settings. Studies typically use pre-recorded materials, and the unit of analysis is usually intelligibility (i.e, number of words correct) in sentences, or word recognition for single words, or sometimes consonant or vowel recognition within words. This is not how clear speech is used outside the laboratory in natural settings; rather, one person speaks to other people, or two or more people speak to each other in a conversation. In either case, there should be no mismatch between the talker's perception of their own speech rate and the listeners' perception of that speech rate; unless the listener's hearing is overwhelmed with noise or hearing impairment and they are unable to see the talker's lips, they can perceive the rate of speech without understanding the speech. This is true whether the talker is speaking clearly or casually.

However, if the listener's experience is manipulated through the techniques developed in prior chapters (either by their own choice or without their knowledge) then their own experience of the talker's conversational rate is private information, and no longer matched to the talker's experience of their own speech. Even in an idealized two-person conversation where each conversant utters a few sentences while the other listens in patient silence, allowing one or both to temporally expand incoming speech presents hypothetical problems, even if it may solve others. The first and most obvious of these is that a type of "desynchronization" may occur (especially if listening pauses are expanded as well) where one is somewhat "behind" in the conversation. However, other more subtle effects may obtain also, such as changes in speaking rate, changes in turn-taking, changes in the talker to talker ratio of words or time spoken, etc.

This chapter presents a study designed to evaluate these effects directly, by analyzing speech rates and effectiveness of pairs of conversants whose speech rates are artificially expanded without their knowledge. This study was presented as a poster at The Acoustical Society of America in 2015 and expanded into a full publication for the Proceeding of Meetings on Acoustics in 2022 [3].

4.2 Related Work

4.2.1 Technical Work

The software used in this study is the earliest of three experimental versions of audio expansion software described in Chapter Three, the Single-Talker Half Prototype, implementing only single-talker functionality, but providing a live/on-line audio expansion experience using a laptop and headset. Specifically, with a headset attached, it would take in audio from a talker, temporally expand the spoken (and any other) audio without modifying the pitch, and send the manipulated audio elsewhere.

Although other applications for time compression and time expansion of audio exist, including such applications as Professional DJ software such as Ableton Live; media players, such as the VLC player; professional music studio software such as Adobe Audition, Avid Pro Tools; streaming video sites, such as YouTube, and online academies such as Coursera, I was then and am now unaware of any that are intended to act "live", especially in a conversational fashion, rather than on pre-recorded audio. All of the applications mentioned are intended to work with pre-recorded sounds, without any two-way interactivity. Even in the case of interactive communication software such as teleconferencing applications, I am unaware of any which support rate changes in the fashion discussed below. The novelty of this work lies in the expansion from pre-recorded single-point broadcast applications, such as the above, to live multi-point communications—in this case, two-person conversations.

4.2.2 Experimental Work

Prior experimental work supports the idea that there are several potential benefits of being able to manipulate speech in real time, to personal preferences, including increased intelligibility, and increased comprehension. Further, as people age, both normal and hearing-impaired adults experience greater difficulty with speech discrimination and comprehension in complicated multi-talker environments. Although the primary cause of these difficulties is hearing loss, several studies have shown the influence of cognitive factors such as memory and attention [60, 61, 62, 63]. Audio expansion is often suggested as a mechanism to improve the efficacy of spoken communication.

Studies also support the idea of a connection between slowed speech and increased comprehension (rather than the typical intelligibility measures), especially for non-native speakers [64]. However, other studies suggested that elongation of gaps between words may play a greater role than slowing of speech sounds, per se [65]. A study by [66] does not use technical intervention, but instead studies the effects of naturally changing speech rates on communication dynamics.

Finally, audio expansion has been suggested as a future feature for advanced hearing aids [67], and as the technology becomes simpler and better understood, and working conditions change in the wake of COVID-19, the technology may become a standard feature in other everyday settings including, e.g., cell phone communications (as per the prototype described in Chapter Three), learning and practicing foreign languages, and conference calling software.

4.3 Motivation

Aside from the potential applications noted above, there are potential problems or stumbling blocks to using this technique in live, multi-person communication that do not hold for more conventional techniques such as hearing aids. These objects relate specifically to changing the rate of speech and the necessary introduction of time delays which follows. In conventional hearing aids, delays as short as 10 ms can cause irritation due to a talker's new perception of their own altered speech [68]. Delays of 20 ms or more may interfere with speech production [69]. Finally, longer delays of 200 ms or more may also interfere with audio-visual processing and reduce the intelligibility of received speech [70].

The very nature of audio expansion techniques will introduce delays much larger than—and much more variable than!— the simple processing delays and group delay effects of hearing aid amplification. This study is intended to assess the effects of audio expansion on real time spoken communications in carefully controlled audio-only environment, divorced from the contexts of audio-visual integration or self-interference.

4.3.1 Communication Dynamics

Prior studies [71] have shown that the acoustic characteristics of two talkers' speech, when engaged in conversation, tend to converge toward common values, e.g., duration of stop consonant voicing, or intonation patterns. Similarly, there is evidence that listening to slower rate speech may result in a sympathetic decreased speech rate on the part of the listeners [66]. It is possible that employing audio expansion techniques in conversational settings may have similar effects; since it is possible for both conversants to employ audio expansion, it is further possible that this may result in a feedback loop with extremely deleterious effects on the communication of both parties. Some of these effects may inadvertently arise in VoIP interactions as artifacts of network delays. However, these unpredictable communication channel effects are typically seen as a nuisance. Audio expansion, however, is a richer phenomenon than simple network-induced delay-- it is an affine auditory delay involving both scale transforms and cascading delays. The ability to control the nature of the expansion will allow for a systematic exploration of their constructive and distractive aspects in group interactions.

While a formal model of behavioral dynamics that may arise due to even single-speaker expansion is beyond the scope of this work, this study examines the effects of single- and dual-speaker audio expansion on speech interaction during cooperative problem-solving tasks with two or more members, as a necessary prelude to developing useful devices and interventions.

4.3.2 Research Questions

Although prior research indicates that communication dynamics can be changed by slowing pre-recorded speech or by the presentation of naturally produced slower speech, this study tests whether real-time expansion techniques might produce similar results. Since this study investigates a novel technique for slowing down speech no hypotheses are posed. Rather, this experiment is exploratory. In that vein three research questions guided the design of this experiment:

- First, how does audio expansion affect performance on an audio-only coordination task?

- Second, how does audio expansion impact communication dynamics in an audio-only communication task? Specifically, does listening to expanded speech, or having one's speech expanded, affect speech rate?

- Third, what attitudes are reported about interacting through the audio-expansion software?

4.4 Methods

To test audio expansion in real time speech communication, this study is designed as a collaborative experiment involving pairs of conversants assigned to work on a collaborative task. The experiment investigates the effects of audio expansion on real time interaction, using previously designed prototype software and a Diapix task [72] to elicit spontaneous conversation between participants without experimenter intervention.

4.4.1 Participants

The University of Illinois at Chicago Office for the Protection of Research Subjects approved this research prior to any recruiting or experimentation, as protocol 2012-0867; see Appendix A. Ten participants (seven male, three female) between the ages of 18 and 57, with no history of hearing loss were recruited to take part in this study. Seven subjects were native speakers of English. No compensation or incentive was provided. See Appendix A for recruitment and consent documents.

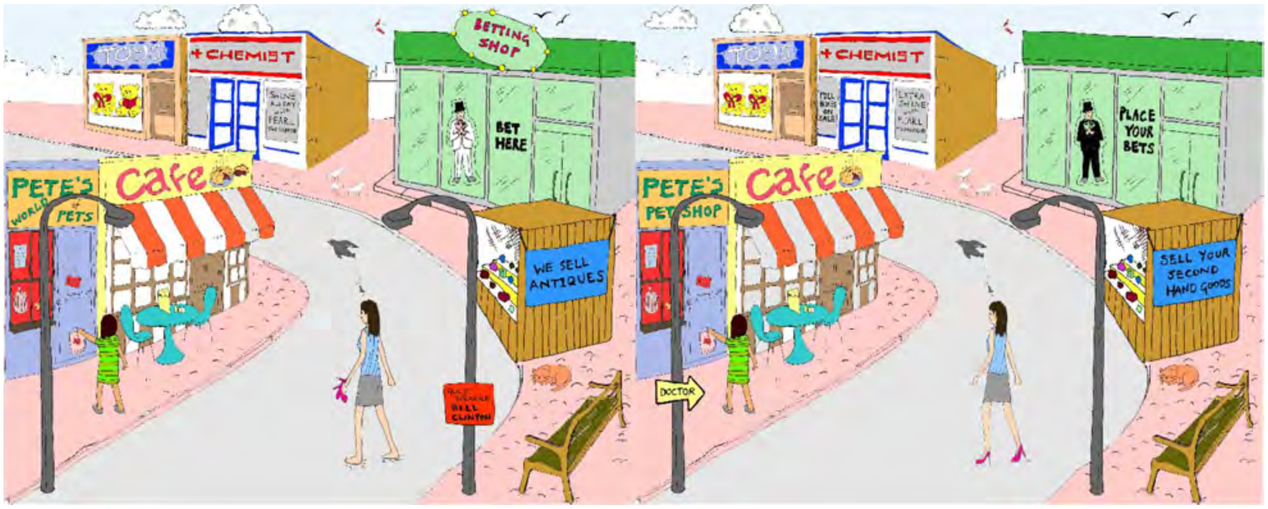
4.4.2 Materials

4.4.2.1 Diapix Tasks

A Diapix task is a collaborative two-person game designed to elicit spontaneous, natural, conversational speech from two participants. Each task consists of two matching images of the same scene, but with twelve subtle differences between them. The participants must, through conversation alone (i.e., without showing each other their respective images) determine what those twelve differences are, and clearly mark them on the images with pens. An example is shown in Figure 17 [6] below. A total of four Diapix tasks were used in this study.

[6] Figure 19 has been adapted from [73] unaltered, in accord with CC 4.0

Figure 17   Diapix Examples [Adapted from 73]



4.4.2.2 On-Line Audio Expansion Software

Prototype audio expansion software, as described in Chapter Three, was used for this experiment. This software provides real-time audio expansion, using a magnitude-interpolated, phase -advanced buffered phase vocoder as described previously, specifically, the Single-Talker Half-Prototype. To briefly reiterate: Live audio input is received from a microphone, divided into windowed sub frames, and each frame is converted to the frequency domain by means of an FFT, effectively constituting an online Short Term Fourier Transform (STFT), updated as audio data arrives. A separate output STFT is calculated from the input STFT-- an expansion factor of, e.g., two requires the creation of a single additional frame at the midpoint of every original frame. These output STFT frames are held in a separate buffer, converted back to the time domain and sent to speakers for playback as required. Pilot studies and informal experiments indicated that expanding lengthy silences and pauses would cause rapid

4.4.2.3 On-Line Audio Expansion Software

Prototype audio expansion software, as described in Chapter Three, was used for this experiment. This software provides real-time audio expansion, using a magnitude-interpolated, phase -advanced buffered phase vocoder after the fashion of [58]. In brief, this software functions as follows: Live audio input is received from a microphone, divided into windowed sub frames, and each frame is converted to the frequency domain by means of an FFT, effectively constituting an online Short Term Fourier Transform (STFT), updated as audio data arrives. A separate output STFT is calculated from the input STFT-- an expansion factor of, e.g., two requires the creation of a single additional frame at the midpoint of every original frame. These output STFT frames are held in a separate buffer, converted back to the time domain and sent to speakers for playback as required. Pilot studies and informal experiments indicated that expanding lengthy silences and pauses would cause rapid desynchronization of conversations. Therefore, this application also incorporated a Voice Activation Detection (VAD) function to prevent the expansion of lengthy silences or pauses, with the intent of minimizing or eliminating this desynchronization effect. These processes (both the time expansion and the VAD) are described at greater length in Chapter Three and in [1] and [2].

4.4.2.4 Hardware And Software Set-up

Note that the software described above effects time expansion on one audio stream; therefore, this experiment with two conversants at a time, requires two laptops to perform audio expansion on two simultaneous audio streams. Prototype software as described in the previous section was installed on two laptops with sufficient processing power to smoothly expand audio input. One laptop ("A") was situated in a soundproof anechoic chamber. Another laptop ("B") was placed outside that chamber in a separate but quiet and isolated room.

Headset and audio cables were used to connect the laptops through a small cable portal through the anechoic chamber wall. The equipment was connected so that:

- Conversant A's speech passed into the boom microphone of Headset A and into Laptop A's audio input
- Laptop A expanded Conversant A's audio, and sent it to the audio input of Conversant B's headset
- Conversant B's speech passed into the boom microphone of Headset B and into Laptop B's audio input
- Laptop B expanded Conversant B's audio, and sent it to the audio input of Conversant A's headset.

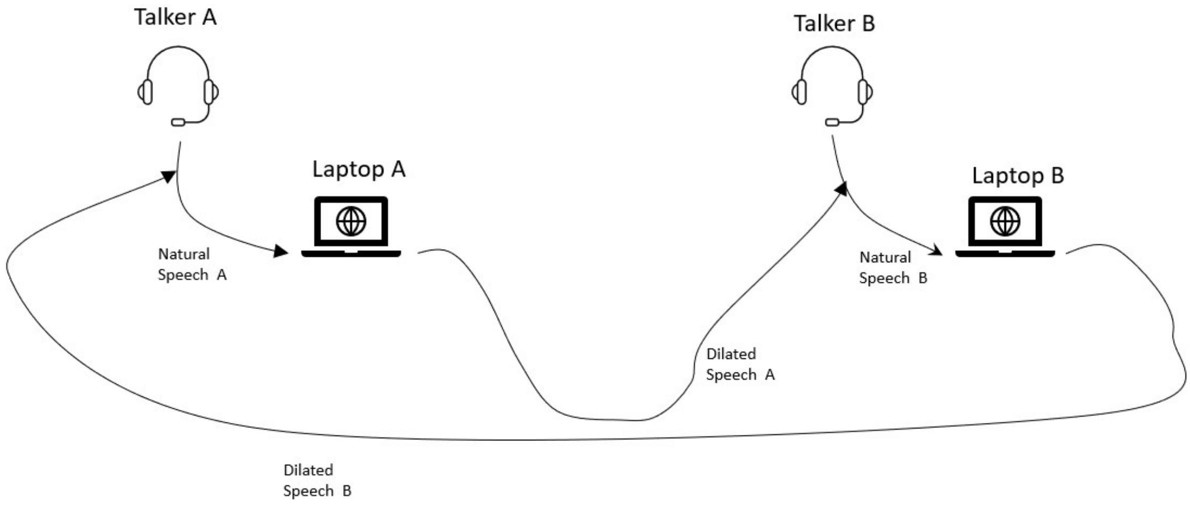
Note that neither the laptops nor the headsets had separate audio input and output jacks, therefore, a system of audio headphone splitters and supplement cables was used to make these connections. See Figure 18. The end result of this arrangement was that each conversant could speak into the boom microphone of their own headset, have their audio expanded by their own laptop, but hear the expanded audio of their conversational partner.

The software on both laptops was controlled by two experimenters who selected the initial expansion and VAD parameters. These experimenters also simultaneously initiated the laptops (to preserve initial conversational synchrony) but did not otherwise intervene or interfere with the subsequent communications.

4.4.3 Procedure

Once hardware and software were in place as described above, participants were divided into five pairs of two conversants each. From each pair, one participant was designated Conversant A and situated at Laptop A in the anechoic chamber; the other, designated Conversant B, remained outside the chamber and was situated at Laptop B. Two researchers instructed the conversants as to the nature of the task, helped them equip the headphones and situated them at their respective laptops, and synchronously activated the laptops to begin each task. They also distributed and collected matched Diapix images as appropriate.

Figure 18 Diapix Wiring Diagram



Critically, the conversants were not informed that their audio was being manipulated in any way.

Each pair of conversants completed a sequence of four Diapix tests, with all communication intermediated by the pair of laptops. The door to the anechoic chamber was shut and neither conversant could see the Diapix image of the other. Different Diapix pairs were used for each of the four tasks. Each task was ended either after the conversants reported that they had identified all twelve differences, or after ten minutes if they failed to do so. The full experimental protocol of four tasks was executed in the following order:

- Audio of Conversants A and B both expanded by 140%;
- Only Conversant A expanded by 140%;
- Only Conversant B expanded by 140%;
- Neither Conversant A nor B expanded.

During any trial in which a conversant's speech was expanded, that conversant's VAD was activated, as described above.

During each trial, the prototype software on each laptop made two separate audio recordings: (1) Unmodified audio (not expanded and thus not subject to VAD algorithms) as the conversant spoke it, and (2) Modified audio (if applicable) subject to both expansion and VAD. Each task therefore generated four audio recordings. (Some of these recordings might be identical, if a conversant's speech was not expanded during a given task.)

The expansion value of 140% was chosen after a period of informal experimentation with the prototype software, and later by pilot studies. Values below approximately 125% were at times too subtle to be noticed, while more than 160% began to sound distorted or unnatural even though the pitch of the audio was not shifting. There was also some concern that even with the VAD feature engaged, conversations might experience obvious desynchronization at larger expansion values.

After the experiment, each subject was invited to answer a brief survey, after which the researchers conducted informal interviews.

4.5 Results

4.5.1 Speech Analysis

The basic units of analysis for this study were the recordings of the unmodified speech, described in the previous section. These recordings were analyzed using a Praat script [74] with two prosodic measures: Articulation rates (syllables/second, not including pauses) and number of pauses longer than 300 milliseconds. A one-way Anova was performed with Tukey post-hoc analysis. No articulation rate is statistically different than another (all $p > 0.95$). No pause count is statistically different than another (all $p > 0.1$). The results are presented in Table I below:

Table I Summary of Speech Analysis

Condition	UU	UD	DD	DU
Articulation Rate, syl/sec Mean (Std Dev)	3.94 (0.32)	3.98 (0.23)	3.96 (0.22)	3.99 (0.23)
Number of Pauses Mean (Std Dev)	122.6 (40.0)	109.9 (29.4)	145.6 (48.7)	115.3 (34.3)

Diapix test scores and test times were also collected and analyzed, as shown in Table II below. One-way Anova tests did not reveal significant differences in test scores (all $p > 0.7$) or test times (all $p > 0.7$).

Table II Summary of Test Scores and Times

CONDITION	NEITHER EXPANDED	BOTH EXPANDED	MIXED
TEST SCORES MEAN (STD DEV)	10 (0)	9.2 (1.3)	9.3 (1.16)
TEST TIMES, SEC MEAN (STD DEV)	488.4 (106.7)	517.6 (145.5)	473.3 (119.7)

4.5.2 Survey Data

At the conclusion of the experiment, subjects were asked to answer three survey questions, each on a Likert scale from 1 to 10.

First, “Communicating via the dilation software was…” where 1 was “distracting” and 10 was “intuitive.” Scores ranged from 3 to 10, with a mean of 5.9 (SD = 2.64) across all participants.

Second, “Getting used to the dilated audio was…” where 1 was “easy” and 10 was “difficult.” Scores ranged from 1 to 7 with a mean score of 4.0 (SD = 2.21) across all participants.

Third, “The delays introduced by the software were…” where 1 was “frustrating” and 10 was “helpful.” Scores ranged from 2 to 8 with a mean score of 3.9 (SD = 2.08) across all participants.

4.6 Discussion

The analysis of the speech as spoken (as opposed to speech which may have been expanded, which was collected but not analyzed) does not indicate a change in articulation in any of the four possible combinations of expanded and unexpanded speech. This indicates that, at modest rates of expansion (140% expansion, or an expansion factor of 1.4, in this experiment), there are negligible risks of users unknowingly or unintentionally slowing or speeding their own speech, either when they hear expanded speech or when their own speech is expanded (or both.) However, this is at odds with the only similar study of which I am aware. [66] performed a similar user study, but rather than using computer intermediation to artificially slow down speech, used trained talkers to naturally produce speech at slower articulation rates, specifically 45% fewer syllables/second, which is equivalent to a 180% expansion factor. This resulted in a small but statistically significant rate reduction of 6% fewer syllables/second for the participants in this study. There are three possible reasons for the different results of these two studies: First, a 180% expansion factor is audibly different from a 140% expansion factor; it may be that if the expansion experiment had chosen 180% expansion, it would have produced a statistically significant effect as well. Second, also in [66] one participant per experiment adapted to the speech rates of trained talker who were maintaining a particular rate of speech; in this study, two participants per example co-adapted (or not) to each other. Third and finally, the computer intermediation of this study had the potential for desynchronization as described above, where in [66] this was not an issue.

In addition, the expansion factor of this study was chosen based on pilot studies and informal experimentation. However, other studies which have allowed users to set their own expansion factors in various adverse conditions are comparable or use less expansion. For speech in a background of four-person babble noise (4), users selected expansion factors from 105% (20 dB SNR) to 150% (0 dB SNR) in a study of sentence-length intelligibility. Non-native speakers of English with recent TOEFL scores ranging from 60 to 110 selected expansion ranges from 100% to 143%, with an average of 128%, in a study of long pass comprehension. In contrast, however, (75) shows preferences for faster speech, even in various conditions of noise. Overall, the rates selected by users in practice tend to use less expansion than used in this study and therefore overall there is little or no risk of audio expansion techniques changing the articulation rates of users.

Surveys and interview revealed occasional frustration with the delays introduced by the audio expansion due to the timed nature of the tasks. However, the changed introduced were reported to be easy to adapt to, and communication remained stable. Participants were not confused the audio expansion, and did not report and conversational desynchronization or other disruptions in their communications. Many participants mentioned an awareness that their audio was being manipulated, but none correctly identified that their or their partners' rates of speech had been altered.

Analysis of the test scores and test times shows no significant change in either, and does not support the idea that slowed speech might aid in increasing effectiveness of coordinated tasks. However, Diapix tasks are not intended to be difficult to solve, per se, merely to elicit spontaneous speech.

This study, which uses audio expansion in an experimental setting to alter speech between a pair of talkers working to complete a coordinated task in separate visual and acoustic environments adds evidence to the efficacy of the audio expansion software, in prototype form, as a program which can work to slow speech in real time without meaningfully disrupting communicators. While the experiment does not suggest that audio expansion alters speech communication patterns, nor does it suggest that it makes it easier to complete coordinated audio-only tasks, this could be due to many factors stemming from the design of this experiment and due to external factors; recall, e.g., that subjects were asked to complete the task as quickly as possible and tests were halted after ten minutes, creating a sense of time pressure.

4.7 Conclusions

Taken as a whole, these results indicate that the technique of slowing the rate of speech is well-tolerated, causes no change to natural produced rates of speech, causes no disruptions in audio-only communication, and does not adversely affect collaborative problem solving skills. This in turn suggests that the technique of slowing speech in real time does not interfere with the basic transmission of spoken information even though it also does not necessarily improve audio-only coordination on cooperative tasks.

However, the findings of this study are necessarily limited by the small size of the participant pool, its demographics, and pre-determined audio expansion factors. The number of participants offers suggestive evidence, but more participants are required to confirm statistical significance (or lack thereof) in the results. The experiment is also limited by the demographic makeup of the participant pool, which was limited in age, location, gender, language diversity, and educational background. In particular, the results of this study may have been more significant with better control for the native and nonnative language pairing of the participants. Prior research suggests that slowing speech for non-native speakers can help increase general comprehension [64], and that focusing on an older population may have produced more significant results [60, 61]. Finally, prior literature [76] has indicated that allowing non-native listeners to determine the speed of received speech increases their comprehension more than pre-determined speeds, which was used in this experiment to maintain greater control. These limitations suggest a series of technical improvements and experimental design suggestions that future studies may consider implementing.

Increasing the number of participants, with control over the paring of native to non-native speakers would improve the experimental design. Other improvements may include allowing users control over incoming speech rate, as prior research [76] indicates that comprehension increases when speakers can control the speed of audio they are receiving. Specifically, allowing participants to manually control the speaking rate of their partner may yield insights about preferred rates of speech and offered insights about whether those preferences led to better outcomes on the Diapix task or not. A counterpart of [66] using fast talkers rather than slow talkers might be particularly enlightening.

Additional targets of analysis may also be useful. This study analyzed articulation rate and number of pauses; it may also be useful to analyze the overall amounts of time that each Conversant speaks (i.e., does audio expansion tend to cause the affected conversant to speak for greater or lesser amounts of time?), the average lengths of spoken segments (i.e., does audio expansion tend to cause the conversant to speak in longer or short chunks, or with longer or shorter listening pauses), and the dynamics of turn-taking and simultaneous speech. This last analysis in particular may prove both novel and difficult due to the differing perceptions of the conversants. All of these analyses, however, may prove useful in understanding if and how the use of audio expansion shifts the overall dynamics of a conversation—in particular, does expansion privilege or advantage one conversant over another? If so, there may be ethical considerations to the use of this technology as well.

5 On Speech in Noise

Previous chapters of this dissertation have focused on the development of software systems capable of real-time stretching of audio signals, as well as a user study which show that in at least one situation (a dual participant study of interaction and communication), the technique of audio stretching is well tolerated and does no harm. The remainder of this dissertation changes focus from real time systems without user control, to systems which afford test subjects control over their own listening rates to investigate the role of user choice in intelligibility in various adverse situations.

This chapter is adapted from the previously published paper:

Novak III, J.S. and Kenyon, R.V., 2018. Effects of User Controlled Speech Rate on Intelligibility in Noisy Environments. In *INTERSPEECH* (pp. 1853-1857).

5.1 Introduction

Although it is the listener who needs and derives the benefits of highly intelligible speech, improving that intelligibility typically remains the responsibility of the talker, due to their active roles in speech production, while the listener's role remains largely passive. A great deal of work has examined how listeners process different characteristics of a talker's speech behavior and how such changes impact intelligibility and other factors. For example, in noisy or adverse listening environments, talkers adapt their speech (ostensibly to improve intelligibility) in a number of ways: reducing speech rate, changing pitch, changing formant patterns, and increasing consonant-vowel energy ratios [38, 34]. Such speech adaptations are also found when speaking to non-native listeners [21], listeners with a hearing-impairment [38], and when talking to infants [77, 78]. However, little has been done to give control to *listeners* to improve intelligibility other than amplification and, more recently, noise cancellation technologies. Hearing aids, e.g., allow listeners to change filter and amplifier settings, and some modern hearing aids allow for remote programming via smart phone apps; however, while these tuning settings are sometimes exposed to the wearer, programming of a hearing aid is typically regarded as an expert task for audiologists rather than patients. However, more may be possible given advances in electronic processing power [2]. Giving listeners control of the temporal characteristics of incoming audio signals may open a new avenue to improve intelligibility in noisy and other adverse conditions. But to provide such affordances it is necessary to know what changes in the auditory signal can assist a listener.

Due to the prevalence of talkers adapting their *speech rate* in challenging environments, much prior work has examined the effects of artificially modified speech rates on intelligibility [19, 79]. However, these experiments typically use a small number of expansion factors chosen directly by the experimenter applied uniformly across all utterances, all parts of utterances, and all test subjects. Both experiments cited above showed decreases in intelligibility with time expansion.

However, this experimental paradigm rests on several assumptions, any or all of which may be invalid: First, that the experimenter (based on surveys of the literature) has a special insight into the necessary amounts of expansion; second, that the listener requirements are constant across a group of subjects; and third, that a uniform expansion across a given utterance may be a useful solution.

Some related experiments which do feature user control or pacing in noiseless environments have shown improvement in related measures: In [76], a group of 15 participants, all non-native speakers of English (although of unknown TOEFL scores) were allowed to choose from a small number of pre-determined speech rates from 75% to 200% playback rate of passages originally delivered in at 194 words per minute. After selecting an initial speed, however, the subjects were unable to change the speed afterword. Their goal was to improve comprehension (not intelligibility), and some improvement, from 30% correct answers (unmodified) to 50% correct answers (modified) was seen. In [80], participants with both clinically normal and mild to moderate hearing loss (12 each) were allowed to self-pace audio passages. This self-pacing was not achieved by changing speech rate, per se, but by advancing through the passages segment by segment. Segments, in this case, were pre-defined by the investigators as sentence boundaries and major syntactic clauses within sentences. The subjects' goal was to improve recall (not intelligibility), and the study provides strong evidence that both hearing-normal and hearing-impaired subjects improved, with the hearing-impaired subjects improving more.

Given these improvements when listeners have some control over speech rates, this study is designed to provide additional and finer-grained control to participants in a challenging, noisy environment for the purpose of improving intelligibility.

5.2 Materials

This study adapted QuickSIN [81, 82] speech tests, devised by Etymotic Research, as its basis. The purpose of the QuickSin test is to quickly provide a single figure-of-merit indicating the difficulty a listener experiences in noisy environments, and is often administered to elderly or hard of hearing patients. Each QuickSIN test consists of a single list of six recorded Harvard sentences [83], spoken in a conversational manner by a single female talker ("signal"), in the presence of varying amounts of four-person babble noise ("noise"). Each sentence contains five pre-determined keywords. During a QuickSIN test, these sentences are played for a subject against background noise, after which the subject recites as much of the sentence as he or she was able to extract from the noise. An audiologist or research makes note of which and how many keywords were recited correctly for each sentence. The sentences are played against noise in the following order: 25 dB, 20 dB, 15 dB, 10 dB, 5 dB and 0 dB SNR ("Signal to noise ratio"), where 25 dB SNR is considered to be very easy for young healthy ears to understand and 0 dB SNR is considered to be very difficult even for those same young healthy ears. Once the subject has listened to and attempted to recite all six sentences, the test scorer counts the number of correct words overall, N, and converts it into one or both of the following metrics:

$$\mathrm{SNR}_{50} = 27.5\ \mathrm{dB} - \mathrm{N} \tag{24}$$

$$\mathrm{SNR}_{loss} = 25.5\ \mathrm{dB} - \mathrm{N} \tag{25}$$

Here SNR_{50} represents the SNR at which the subject is expected to be able to correctly recite 50% of the keywords correctly; young healthy listeners are expected to have an SNR-50 of approximately 2 dB. The related metric, $\mathrm{SNR}_{\mathrm{loss}}$, represents the deviation of one listener's ability from that of a statistically average young, healthy listener. For comparison, a listener with $\mathrm{SNR}_{50} = 5$ dB is expected to be able to correctly recite 50% of words in 5 dB noise. This same listener has an $\mathrm{SNR}_{\mathrm{loss}}$ of +3 dB meaning that the listener extracts at 50% of the keywords at 3 dB lower noise than would a typical healthy listener. Positive values of $\mathrm{SNR}_{\mathrm{loss}}$ represent additional challenge or degradation relative to young healthy listeners. (The constant 25.5 dB is determined from [84], and the constant 27.5 dB is an adjustment based on statistical observations of the SNR-50 scores of many young, healthy listeners.)

The QuickSIN test was chosen as a basis not only for its simplicity in administration and scoring, but also because it employs the widely used Harvard sentences, which are 72 lists of ten sentences each, with each list designed to be phonetically balanced. Further, the particular lists employed in the QuickSIN test have been further analyzed over time and are strongly believed to be of equivalent difficulty in the context of intelligibility.

This study employed the same basic structure, but with the noise levels altered as follows to include: 20, 15, 12.5, 10, 8.3, 6.7, 5, 2.5, and 0 dB. New/non-standard values are underlined for clarity, and the 25 dB noise condition is removed.

The QuickSIN tracks purchased from Etymotic Research were separated into twelve signal and twelve noise tracks, and converted to 16 bit, 44.1 KHz sampled WAV format sound files, and are referred to as Lists 1 through 12. One track (List 13) contained both signal and noise on the same track and could not be separated. This non-separable track was used as a “practice” list. Finally, one audio segment of an actor reciting the Gettysburg Address [85] was converted to the same format. All WAV files were normalized with a sound level meter to the same sound intensity.

The software for this study was built around the engine described in Chapter Three of this dissertation. This software allowed subjects to manually control the rate of audio playback in real time, using a frame-based, magnitude-interpolating phase vocoder [58] as described in [2]. Subjects controlled audio playback rate with an on-screen slider bar, which could be dragged quickly with a computer mouse. Note: In this experiment, the slider control was improved relative to the systems presented in Chapter Three, such that the subjects could smoothly vary or “chirp” the audio dilation rate. The subjects controlled the dilation ratio of the audio signal (defined as the ratio of the unmodified to modified length of an audio track) which was linearly related to the movement of the on-screen slider bar. Allowable values of dilation ranged from 1 (at the left) to 0.4 (at the right) inclusive, with a resolution of 0.01. This level of granularity is impossible to distinguish by ear. Subsequently these data were transformed into expansion ratio (1/dilation ratio) to maintain consistency with conventional analyses in audio research. Therefore, subjects adjusted the expansion factor from 1.0 to 2.5 in non-uniform increments of not larger than 0.06. Note that the subjects’ interface was not labeled with any numerical values, to avoid consciously or subconsciously biasing their responses. Audio playback rate responded smoothly, in real time. The time stretching technique did not change the pitch or tonal qualities of the audio. In all cases, the slider controlled the playback speed of the *signal tracks*, while in all but one case, the speed of the *noise tracks* was unchanged. The software set SNR values by keeping the signal constant and changing the intensity of the noise tracks.

5.3 Methods

5.3.1 Participants

Twenty-eight young (age 18 - 30), healthy (self-reporting no hearing problems), native English speaking adults took part in this study. Participants were entered into a raffle for a $50 Amazon Gift Card, which has since been disbursed. This study was approved by the UIC Office for the Protection of Research Subjects as protocol 2015-1164; see Appendix B. Participants were recruited by announcements to UIC student mailing lists and in lectures, and provided written statements of informed consent prior to participation. Participants were informed of their right to halt participation at any time, without removal from the raffle.

5.3.2 Procedures

The custom designed software was installed on a Windows laptop, connected to Sennheiser HD 598SE over-ear headphones, with audio track sound levels calibrated to present signal audio at 65 dB SPL. Participants then engaged in the four-phase experiment described below, followed by subject interviews. The experiment took approximately 45 minutes per subject.

Training Phase: Subjects trained with the interface, controlling the expansion of a 183 second (unmodified) clip of a recitation of the Gettysburg Address with a horizontal slider. Note that during this phase, the slider was "live," and subjects could adjust the speed in real time, as audio was playing; the purpose of this phase was to familiarize the subjects with the "liveness" of the interface so that they could exploit it if they chose.

Practice Phase: Subjects were presented with sentences from List 13 with audible four-person babble noise, with a fixed signal to noise ratio of 10 dB. The purpose of this phase was to familiarize the subjects with a "target" voice (i.e., the voice to pay attention to in background noise) before the start of the next phase and to understand the sentence lengths (three to five seconds, and approximately eight words in length.) Subjects listened to as many sentences as necessary to be able to remember and distinguish the target voice, and were not allowed to adjust the speaking rate during this phase. This vocal familiarization was necessary because, due to the randomization in later phases, it would be possible to begin in a condition of such high noise that the target voice might not be distinguishable.

Personalization Phase: Subjects listened to QuickSIN Lists 1 through 9, each containing six sentences. Each list was presented at a different SNR level, as described above. List order, noise condition order, and sentence orders within each list were randomized. (The prior *Practice Phase* mitigates against the possibility of randomly drawing a very high noise signal as the first part of this phase.) At the beginning of each list, the initial position of the slider was randomized to prevent historicity. Note that, as discussed above, each noise condition is confined to a single list, previously shown to be of equivalent difficulty.

During this phase, the slider was "live", and subjects were asked to set the expansion to the speed they believed was most useful for understanding the target speech, with no guidance as to what settings might be "useful" other than their own experimentation and exploration. However, in order to prevent the subjects from listening to the sentences often enough to learn or memorize them (which might skew the results), subjects were limited to listening to each sentence only once. Subjects were asked to leave the slider in its most useful location before proceeding to the next group of sentences. These final settings were recorded as the subjects' personalized Preferred Expansion Rates (PER).

Evaluation Phase: Eighteen sentences in noise were drawn from QuickSIN lists as follows: Five each from Lists 10 and 11 were used for the five standard QuickSIN SNR values, in modified and unmodified conditions respectively. The remaining sentences from those lists were combined with those of List 12 to extend the QuickSIN test to four non-standard SNR values. This is summarized in Table III, below.

Table III Speech in Noise Sentence Lists

SNR, dB	**Modified**	**Unmodified**	**Notes**
0, 5, 10, 15, 20	List 10	List 11	Standard
2.5, 6.7, 8.3, 12.5	List 10, 12	List 11, 12	Non-standard

These sentences were presented in random order, and after each sentence the subject immediately repeated the sentence back to the researcher. All keyword responses were transcribed as they were spoken or noted if not spoken.

Subject Interviews: After each experiment, subjects were interviewed and asked whether they believed the overall technique of time expansion helpful, harmful, both or neither; whether they had adopted any strategies or patterns of use; and whether they had specific improvements to suggest.

5.3.3 Experimental Records

Electronic Records: In all phases, the software logged subjects' activity, including wherever applicable the identities of signal and noise tracks, the SNR of combined audio, and all expansion factors. Although it proved not to be necessary, subject activity was recorded in sufficient detail to reproduce what they heard exactly, even during audio tracks with "live" sliders.

Scoring Records: QuickSIN test scores are calculated based on the number of keywords correctly recited back to the test administrator. Subjects' recitations were transcribed, except where subjects omitted keywords entirely.

5.3.4 Analysis Techniques

In addition to PSR and keyword error counts, two specialized measures were used: A modified QuickSIN SNR-Loss, measuring overall intelligibility across five SNR settings; and glimpse increase, expressing the amount of hypothetical perceptual benefit provided by time stretching. [20]

SNR-Loss: As noted above, QuickSIN intelligibility is reported as an SNR-Loss score, calculated from the number N of correctly repeated keywords across all six sentences (i.e., SNR-Loss ≡ 25.5 – N) using the Tillman-Olsen method [84].

However, during the Test Phase, subjects performed two modified and interleaved QuickSIN tests. These tests were altered, inserting additional SNR conditions to better probe the more challenging, lower signal to noise ratio region. The complete set of SNR values is as follows: 0, 2.5, 5, 6.7, 8.3, 10, 12.5, 15, and 20 dB. (New values underlined.) The standard value of 25 dB was removed since very little change was expected from the 20 to 25 dB conditions in a young, healthy population [86].

Because this study used an alternate set of noise values, an alternate SNR-Loss formula was derived following the procedures in [87]. Note especially that these procedures require equally spaced steps of SNR, a condition met only by the non-underlined values. The SNR-Loss scores reported herein are calculated only as above, using the restricted, equally spaced five-point data set. Therefore, only keywords corresponding to sentences from this restricted set were used in this calculation (i.e., SNR-Loss ≡ 20.5 dB – N). As described in Table III above, the SNR-Loss calculation for the modified and unmodified conditions uses sentences from Lists 10 and 11 respectively, thus confining calculations to within single lists shown previously to be of equivalent difficulty.

Glimpsing Data: The effects of audio expansion were analyzed from a psychoacoustic standpoint using glimpse patterns [88]. Cochleagrams were generated [88, 89] from unmodified test sentence and noise tracks by dividing these sound files into 58 uneven bands corresponding to cochlear frequency sensitivity from 50 to 7500 Hz with an integration time of 8 ms and frame length of 10 ms. Each pair of signal and noise pattern were aligned against each other before combining, to match the randomized stimuli heard by subjects. Glimpses are defined as those time-frequency cells in the cochleagrams where signal exceeded noise by 3 dB or more; the glimpse area is the total number of such cells in each pair of signal and noise tracks.

Then, each unmodified test sentence was fed to SPPAS automated speech segmentation software [91] with a transcript, which returned segmentation data for the words of each sentence, identifying onset/offset silences. These were used to generate masks for the glimpse patterns, including whole-utterance masks (excluding offset and onset silences) and keyword-segmented masks (including only keywords.)

Finally, audio files for the modified sentences heard by subjects during their Evaluation Phase were generated from the combination of signal and noise tracks, according to their discovered PSRs. Corresponding segmentations and masks for these expanded sentences were generated by directly expanding the SPPAS results for the unmodified sentences. SPPAS was not used directly on expanded audio tracks because SPPAS was not designed for use on such manipulated speech; it is unclear whether SPPAS would properly segment such manipulated speech.

Finally, as above, glimpse patterns and masks were generated by aligning the modified sentences with noise tracks as heard by the subjects and noting time-frequency cells where the signal exceeds noise by 3 dB.

5.4 Results

5.4.1 Preferred Expansion Rates

Preferred Expansion Rates were determined for each SNR. A Shapiro-Wilk test for normality was performed (Table IV) and failed to show normality for any SNR group. Therefore, median values are presented (Figure 19, upper panel) with 95% confidence interval estimates. Medians ranged from 1.05 (20 dB SNR) to 1.50 (0 dB SNR.) There is a trend of increasingly large confidence intervals with increasing noise, with median values increasing nearly linearly ($r^2 = 0.91$) as SNR declines.

Table IV Shapiro-Wilk Normality Test of PSRs

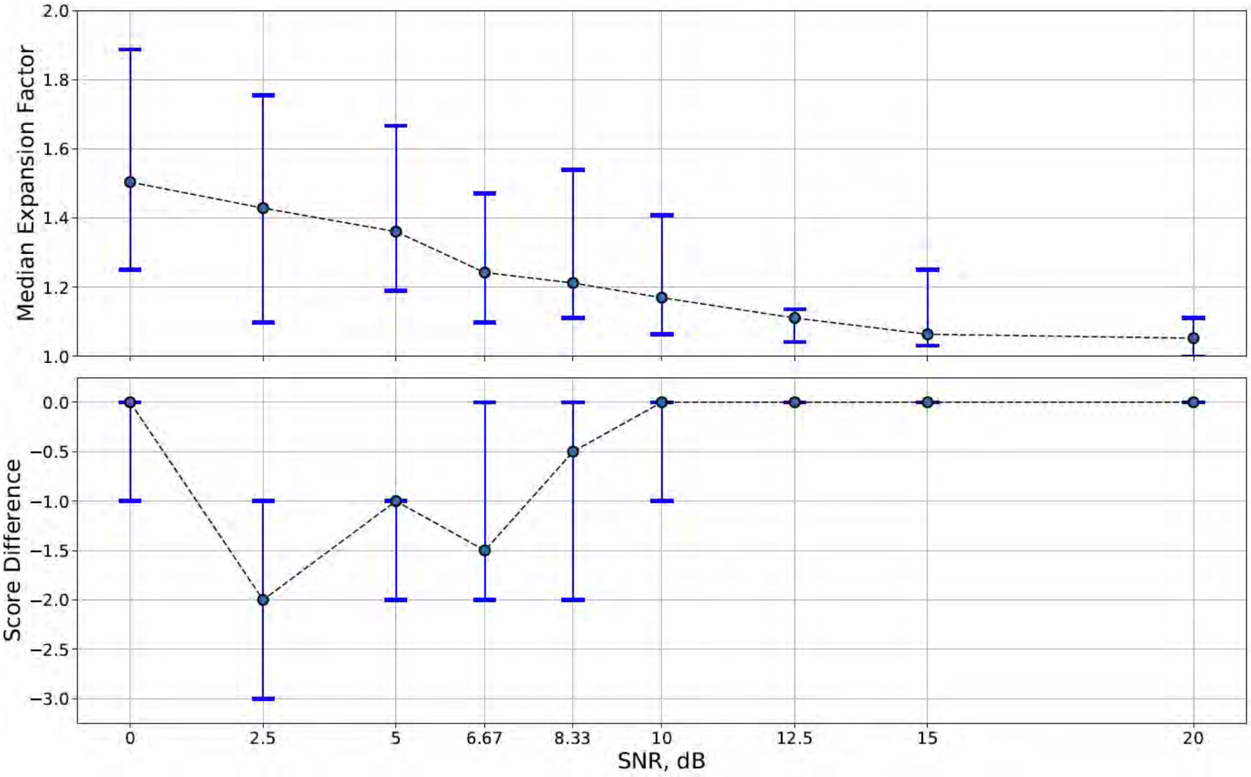
SNR	**0**	**2.5**	**5**	**6.7**	**8.3**	**10**	**12.5**	**15**	**20**
P-value	0.027	.0122	0.003	4.06e-4	0.001	2.13e-5	2.18e-4	8.70e-6	7.90e-7
Normal	False	False	False	False	False	False	False	False	False

Due to the non-normality of the PSR data, standard ANOVA tests cannot be used. Instead, a Kruskal-Wallis non-parametric test was performed (χ^2 = 36.97, df = 8, p = 1.17e-05), followed by Dunn's test with Benjamini-Hochberg post hoc correction, controlling False Discovery Rate to 0.05. The results of the Dunn's test analysis, shown Table V below, indicate that the expansion factors are statistically different at opposite ends of the SNR range, with results for all of the SNRs 0 dB, 2.5 dB and 5 dB differing ($p < 0.05$) from all of the SNRs 12.5 dB, 15 dB and 20 dB.

5.4.2 Intelligibility

Intelligibility degraded significantly from 8.3 dB to 2.5 dB SNR ($p < 0.05$). The median change in score from the unmodified to modified conditions is presented, for each SNR, shown Figure 19, lower panel.

Figure 19 Expansion Factors vs SNR (Above), Score Differences vs SNR (Below)



As before, the Shapiro-Wilk test failed to show normality. A sequence of one-sided Wilcoxon signed rank tests comparing words scored correct in modified vs unmodified conditions shows significant degradation in intelligibility ($p < 0.05$) at all SNRs from 0 dB through 8.3 dB.

Table V   Dunn's Test Of Preferred Expansion Rates (Bold values significant to $p < 0.05$)

SNR	**2.5**	**5**	**6.7**	**8.3**	**10**	**12.5**	**15**	**20**
0	0.683	0.777	0.212	0.261	0.111	**0.009**	**0.004**	**0.001**
2.5	-	0.863	0.399	0.497	0.239	**0.023**	**0.015**	**0.002**
5		-	0.334	0.399	0.197	**0.016**	**0.010**	**0.001**
6.7			-	0.863	0.749	0.197	0.134	**0.021**
8.3				-	0.648	0.150	0.100	**0.015**
10					-	0.334	0.239	0.061
12.5						-	0.842	0.391
15							-	0.497

Overall SNR-Loss scores are also examined, as described above. SNR-Loss scores were calculated from raw data. The Lilliefors test indicated normality for the SNR-Loss scores. SNR-Loss worsened from a loss of 1.7 dB in the unstretched condition to 3.9 dB in the stretched condition—a considerable degradation in the ability to extract intelligible utterances from the background babble noise.

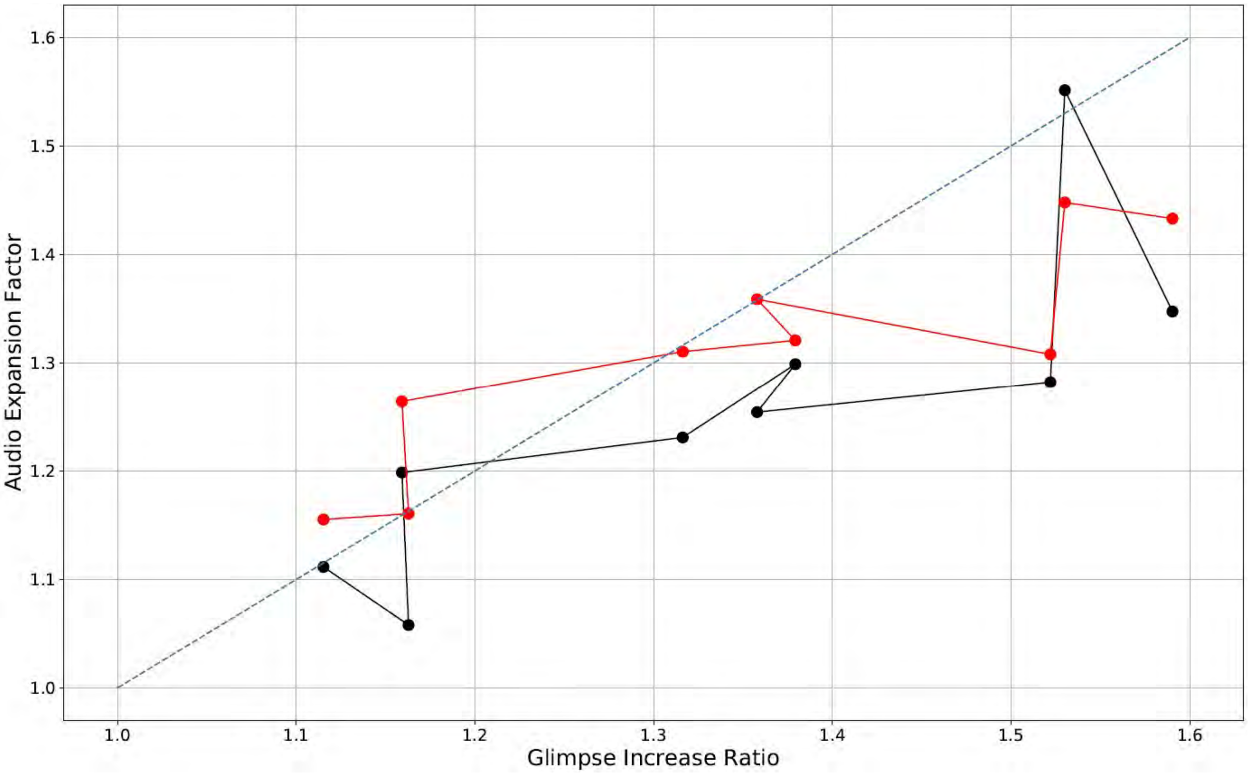
5.4.3 Glimpse Increase

The analysis shows that an expanded sentence yields greater GAs on average, as measured by the glimpse increase ratio (GIR), i.e., ratio of the GA of a time-expanded sentence to a non-expanded sentence, as expected. Expanding an utterance requires the creation of audio frames resulting in a longer track. Therefore, there are more time-frequency signal cells to compare to noise audio in the glimpse analysis, with the same statistical distribution as the original audio. On average this leads to a greater GA, and a greater GIR as temporal expansion increases.

However, this experiment makes direct comparisons between individual sentences difficult. First, to prevent subject learning effects, different sentences of differing lengths were used for test and control conditions. (Evaluation sentences range from 2.05 to 3.39 seconds.) Second, noise track order was randomized. Third, signal tracks contain leading and trailing silences, which are stretched during the Evaluation Phase, but discarded in the glimpse analysis. This changes the position of the voiced portion of a signal track relative to the noise track. Simulations indicate that GA and GIR are sensitive to both the second and third factors.

Since GIRs are highly variable, all GIRs are grouped according to their SNR conditions, and computed the average GIR by condition. This mean GIR is plotted against the mean expansion factor for each SNR along with a reference line of unity slope, shown Figure 20. This process was performed using both the sentence-level mask and the keyword-level mask. Linear regression models were constructed, estimating the slope of the lines as 0.51 ($r^2 = 0.773$) and 0.65 ($r^2 = 0.666$), respectively. Both regressions reject the null hypothesis of zero slope ($p < 0.05$). The sentence-level regression rejects the hypothesis of unity slope ($p < 0.05$) while the keyword-only regression does not ($p > 0.05$). These analyses show that GAs increase with increasing temporal expansion; however, at the level of a whole utterance this increase is less than unity.

Figure 20 Audio vs Glimpse Increase Ratio (GIR)



5.4.4 Interview Data

Post-experiment interviews revealed that 18 respondents (64%) expressed a full or qualified belief that expansion was helpful. “Qualified belief” includes variations (paraphrased) such as “helpful at some noise levels,” or “sometimes helpful, sometimes harmful.” Six subjects (21%) expressed no opinion, or reported that expansion was neither helpful nor harmful. Four subjects (14%) reported that the technique was harmful. Three of the four subjects who reported the technique was harmful set the slider to no expansion for all noise conditions.

When asked about their usage, 15 subjects (54%) volunteered that they employed more stretching when they perceived more noise. Of these, 14 subjects were part of the subset who believed that stretching was helpful; the other thought stretching was neither helpful nor harmful. No subjects reported more expansion with decreasing noise levels.

Three subjects offered that stretched speech, especially highly stretched speech was perceptually odd or deficient (“unnatural,” “not normal speech”, words running together).

Finally, seven subjects referred to the ability to track a particular voice, and/or distinguished between that and the ability to understand the words as spoken by that voice. Three subjects thought slower voices were harder to track (one of whom thought expansion was generally harmful, two of whom thought expansion was neither helpful nor harmful) while three thought slower voices were easier to track (two of whom thought the expansion was helpful, one of whom thought expansion neither helped nor hindered.)

5.5 Discussion

These experimental results are similar to previous investigations [79], namely, that expansion of a speech signal embedded in babble noise does not improve but *degrades* intelligibility. This degradation is concentrated at low SNR conditions where intelligibility is already degraded, but also where expansion was predicted to aid in intelligibility. However, in contrast to previous work where expansion values were imposed on subjects, this degradation occurs despite allowing subjects to choose expansions they feel benefit them most. Thus, given the ability to choose expansions, a majority chose to expand low SNR conditions despite degradation in their performance. It might be that subjects were using a different criteria or performance metric than intelligibility.

One possible explanation lies in the glimpse results, showing that time expansion is accompanied by increased GAs in whole utterances and keyword only portions. Although this glimpse increase did not increase intelligibility, it may enhance ability to track the target voice through the underlying noise as seven subjects (25%) noted. Of those seven, three believed expansion helped to isolate a voice, and was found to be non-harmful to intelligibility. Conversely, three believed expansion harmed the ability to isolate a voice believed that expansion was found to be non-helpful to intelligibility. Perhaps the advantage is that an expanded target voice in noise can be tracked more easily, explaining why most subjects used an expanded audio target signal as the noise level increased. If true, however, this ability to track a voice does not increase intelligibility.

However, the root cause of intelligibility degradation under time expansion is not clear. In [91] several potential explanations are summarized including algorithmically induced artifacts or distortions, which they discount on the strength of modern algorithms; that expansion factors larger than 1.4 may cause degradation by stretching syllables beyond a psychoacoustic “perceptual window”; and that the greatest benefit may be found in situations of cognitive load. Vocoder-based time-stretching algorithms such as the technique employed here are also introduce perceptual artifacts such as “phasiness” [90] and transient smearing which may reduce the psychoacoustic benefits of temporal expansion. Analyses of whole utterances and of the keywords show clear increases in GAs when signal tracks are stretched, indicating an overall increase in receivable signal to the listeners without signal amplification. However, the glimpse increase as measured across complete utterances increases more slowly than does the temporal expansion. This may be the result of the vocoding algorithm, and might be remedied with other more advanced algorithms if implemented in real-time.

Another possibility is that uniform expansion of speech is not sufficiently faithful to natural slowed speech. Various sources note differences in expansion ratios by phoneme [38, 34, 91, 92]. This is tentatively supported by experiments in [94] which test non-uniform time expansion of conversational speech intended to mimic clear speech. While this non-uniform expansion degraded intelligibility, this degradation was much less than that caused by uniform time expansion. In [79], speech signals were expanded based on a local power threshold, with the intent of stretching only vowels. Following this non-uniform stretching, additional distortion was added to simulate hearing loss, with mixed results: time stretching effects on intelligibility were not significant for simulated hearing loss, however, for simulated hearing loss with amplification the effect of time stretching was significant and harmful. However, as the authors note, while this non-uniform stretching tended to stretch vowels rather than consonants, the overall effect was “unnatural” and did not match the cadence of naturally produced slow speech.

The intelligibility degradation may be only one effect of expanded audio. The subjects’ report of ability to track a stretched voice in noise may be of significant value: the three subjects for whom tracking a voice was easier with increased expansion performed below average; however, the three subjects who found tracking more difficult used less expansion performed better. While this small subject pool does not allow statistical analysis, it does highlight a possible trade-off between tracking and intelligibility that may occur with audio expansion. While the intelligibility of an anomalous string of words may be impeded by expansion, tracking a voice in a conversational environment may have additional benefit to the listener. If a natural voice is lost in environmental noise and a stretched voice is not, then even the distortions produced by expansion may be overcome by the subjects’ ability to choose the correct word within the context of the sentence and the conversation. It would be interesting to evaluate this condition in future experiments.

5.6 Conclusions

This study gave listeners control of auditory signals, showing that individual preferences exist; that increased noise results in more expansion; and that listeners perceived intelligibility improvement. There is also a statistically significant degradation in intelligibility even with listener chosen conditions.

I believe these results show the difficulties associated with listener control of personal acoustic experience. Especially, I believe that while some subjects may be conflating the ease of isolating or tracking a voice through noise with the intelligibility of such a tracked voice, this distinction may be a fruitful research direction for related applications such as cognitive load and user comfort.

6 Audio Expansion for English as a Second Language

This is the second of three chapters which design and implement user studies allowing participants to make use of the ability to slow speech. Rather than focusing on the difficulty of the intelligibility of single sentences in noise, this chapter studies the use of speech slowing technology for the comprehension and recall of much longer, multi-paragraph passages by foreign language learners.

This chapter is adapted from the previously published paper:

Novak III, J.S., Bunn, D. and Kenyon, R.V., 2019. The Effects of Time Expansion on English as a Second Language Individuals. In *INTERSPEECH* (pp. 2643-2647).

6.1 Introduction

The ability to actively and successfully listen for meaning and context-- in other words, to understand or comprehend a spoken language-- is crucial to the acquisition of a spoken language [95] and may be the most important skill in second language acquisitions [96]. However, it may also be the most difficult of the four basic language skills (i.e, listening, reading, writing, and speaking) to learn [97]. It is also, of course, important in the daily application of those second language skills.

However, as noted previously, not all speech is created equal. "Clear speech" is an umbrella term for several related speech styles, all of which are adopted by talkers on behalf of their listeners to accommodate some form of adversity, including not only incomplete mastery of a language, but also noisy environments, hearing loss, cognitive decline, or combinations of these and related factors. The distinctions between clear speech and casual speech include (most obviously) a reduction in speech rate, as well as modification of pitch, expansion of vowel space, and an increase in consonant to vowel energy ratio, and several other more subtle effects.

Previous research has shown that naturally produced clear speech does indeed benefit the listeners through increased intelligibility and understanding. One study [98] shows that when clear speech is presented in a noisy background, intelligibility improves for both native and non-native listeners. However, this study also shows that non-native listeners derive a smaller relative benefit than native listeners. A subsequent study [99], also in the context of intelligibility in noise for second language learners, controls the listening test with high-predictability sentences (i.e., sentences whose final words are readily apparent from prior context, allowing language-proficient listeners to predict them) and low-predictability sentences (i.e., sentences whose final words cannot be anticipated.) This study shows that native listeners benefit from clear speech for both types of sentences, while non-native speakers derive benefit almost entirely from high-predictability sentences. One implication of this study is that second language learners may benefit substantially from clear speech techniques in noise; however, it is not clear whether this benefit derives from additional time to comprehend, predict, and compare predictions against sensory information (i.e., if the reduction of speaking rate is at play) or whether it derives from clarity-enhancing features of clear speech such as vowel space enhancement.

However, just as all speech is not created equal, neither are all talkers and listeners created equal. There is evidence in [43] that the benefits of clear vs casual speech vary from individual talker to talker. However, [46] provides evidence of the dual, that the benefits of clear vs casual speech vary from listener to listener. Statistically, clear speech provides intelligibility benefits, but those benefits may be highly idiosyncratic with individual talker-listener pairings.

Several studies have attempted to replicate or convert casual speech to clear speech by computer intermediation, most often by manipulating speech rate and investigating the effects on intelligibility or comprehension. Numerous studies show a deleterious effect on comprehension when speech rate increases [46, 100, 101]. The results of slowing speech are conflicting: One study [102] shows an increase in intelligibility when non-native speakers are allowed to select from one of a small number of pre-selected speech expansion rates. However, other similar experiments [76,103] show no statistical improvement of speech comprehension.

Motivated by prior contradictory results and by evidence of idiosyncrasy, this study considers the possibility that ideal speech rates may be highly personalized, especially, e.g., due to individual language proficiency. To that end this study employs the tools discussed in Chapter Three, to grant fine grained control of delivered speech rate, with real time responsivity, to study participants. This tool is used in conjunction with TOEFL iBT tests to study non-native speakers' preferred speech rates and the effects of those preferred rates on comprehension.

6.2 Materials

6.2.1 Audio and Quiz Material

Listening skill was tested using the Official TOEFL iBT Tests, Vol. 1, comprising 106 separately prepared audio tracks with complementary multiple-choice quizzes. Nine of these audio tracks were selected, such that all nine of their quizzes are of the same format: Four multiple choice questions each with one single correct selection out of four, followed by a final multiple-choice question with two jointly correct selections out of four. Questions with two answers were clearly labeled as such for the quiz-takers.

Of these nine audio tracks, five were fictitious single-talker lectures and four were fictitious two-talker conversations. Each audio track was copied to Wav audio format, and each quiz was transcribed into a text file.

6.2.2 Experimental Software

To test the hypotheses that (1) research subjects would use audio expansion, and (2) would show increased listening comprehension as a result, custom software was developed to administer pairs of audio clips and quizzes in a way which afforded listening rate control to the research subjects.

Specifically, the phase vocoder detailed in Chapter Three of this dissertation was used. This software transformed audio tracks into the frequency domain with a short term Fourier transform, used frame-wise magnitude interpolation and frequency-wise phase advancement to create new frequency domain frames, and transformed back into time domain audio signals with an inverse Fourier transform. The specifics of the interpolation (and thus, the specifics of the audio expansion) were controlled by a simple on-screen slider: While playing an audio clip, moving the slider left would slow the audio down and the opposite would speed the audio up. Time expansion ranged from 1.0 to 2.5 of the original length (1.0 to 0.4 of original speed, or increasing instantaneous playback time by 0 to 150%) with 61 possible slider settings. Note, for this experiment, the software of Chapter Three was upgraded so that the slider responded instantly, allowing subjects to “chirp” the speech delivery rate. The fine granularity of the intervals and responsiveness combined to provide a smooth interface analogous to a volume control.

This interface was used to present participants with alternating audio tracks and multiple choice comprehension quizzes according to the protocol described below. In addition, subjects were free to adjust the audio playback rate throughout the applicable audio clips; these changes of the speed settings along with appropriate timestamps were stored for analysis.

6.3 Methods

6.3.1 Participants

The participants were twenty-six young (age 18 to 30 years), healthy (no self-reported diagnoses of hearing problems), English as a Second Language (ESL) individuals to participate in this study. These participants all self-reported prior TOEFL scores between 60 and 110 at a time no greater than twelve months prior to their participation date. No incentives were offered for participation.

This study was approved by the UIC Office for the Protection of Research Subjects as protocol 2016-0608; see Appendix C. Participants were recruited by announcements to UIC student mailing lists and in lectures. Participants provided written statements of informed consent prior to participation.

6.3.2 Procedures

The custom software was installed on a Windows laptop, connected to Sennheiser HD 598SE over-ear headphones. Audio track sound levels were calibrated to present audio at approximately 65 dB SPL. Participants then engaged in a four-phase experiment as described below. The researcher interviewed each subject directly after the experiment. The experiment took approximately 45 minutes per subject.

Instruction Phase: In this phase, the researcher seated the subjects in front of the research laptop, explained the idea of audio time expansion to the subjects, and demonstrated the use of the slider interface. The researcher then explained that the subjects would be asked to listen to several audio clips and immediately answer quiz questions about them; and that during several of these clips they would have the ability to control the rate of audio. In these cases, the subjects were asked to use the slider to best increase their ability to understand the clips and answer the quiz questions. No guidance was given as to what slider setting or speed might achieve this.

Training Phase: After instruction, the subjects were allowed to experiment with the interface by listening to an audio clip of an actor reciting the Gettysburg Address [85]. The slider was active during the training phase, but no quiz questions were asked. Subjects were allowed to repeat the training phase as often as desired, to feel fully familiar with the interface.

Experimental Phase: After training, the subjects engaged in a sequence of six trials. In each trial, the software presented an audio clip, randomly drawn from the nine clips described above. However, in all cases the 1st, 3rd, and 5th trials were "treatment trials" and gave subjects the opportunity to control the audio rate, while the 2nd, 4th, and 6th trials were "control trials" and did not allow the participants to change the audio speed. During each treatment trial, the initial position of the control slider was randomized, to prevent historicity. Immediately after each audio clip, as part of each trial, the software administered a written multiple-choice quiz to test comprehension of the passage in question, which the subjects answered using a mouse-based interface. Unlike a true TOEFL test, subjects were not allowed to make written notes during any of the trials.

Survey/Interview Phase: After subjects completed the six trials, the researcher presented a Likert scaled survey, with the following three items, each with five selections (Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree):

Question 1 "In general, audio expansion for listening skill was useful."

Question 2 "Slowing speech for the audio files I listened to was useful."

Question 3 "The ability to control audio expansion was easy to use."

Finally, immediately on completion of the survey, the researcher conducted a free form interview with each participant.

6.3.3 Experimental Records

The test software maintained automatic, anonymized records of user activities and responses through both the experimentation and survey phases, in addition to the self-reported TOEFL scores described above. This includes the identity and order of the audio clips which were selected for each trial, and the corresponding quiz responses. In addition, for the treatment trials the software recorded the initial position of the slider (i.e., the initial rate of speech) and the time of each audio rate adjustment relative to the start of the experiment. This record is sufficiently detailed to reconstruct each audio clip as each subject heard it, to calculate expansion factor statistics, and to facilitate manual inspection of user behavior.

After all six trials, the software administered the three Likert-scale questions described above, and recorded the answers.

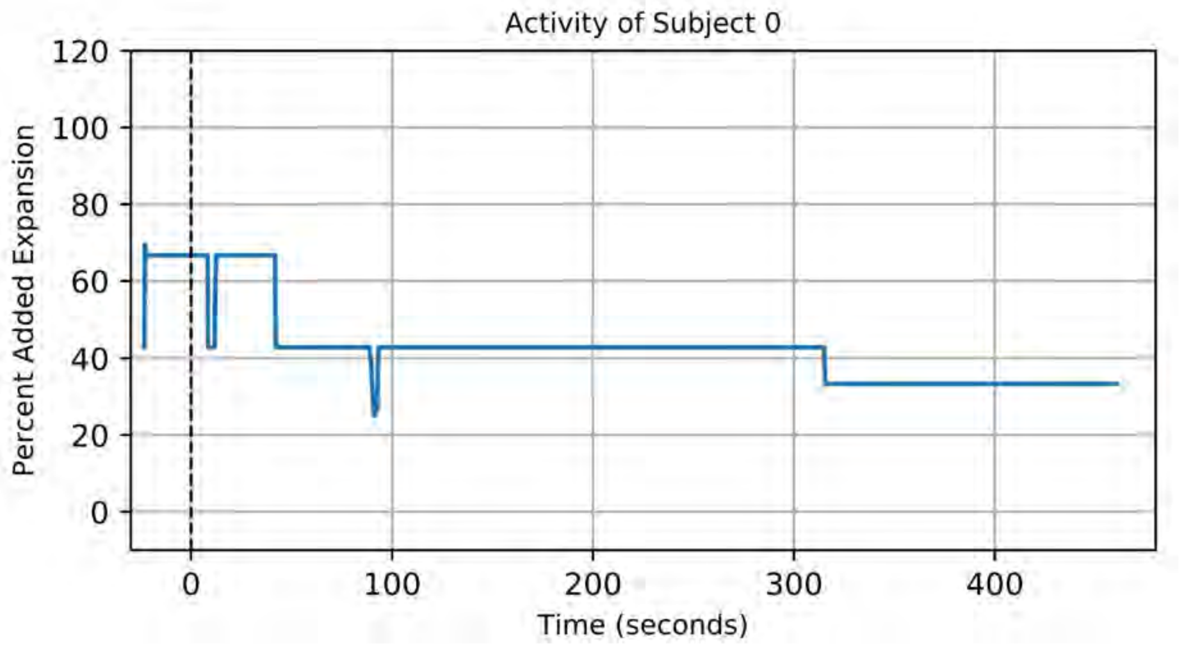
6.3.4 Analysis Techniques

To analyze subject behavior, the recorded data of each treatment trial was inspected, and the following metrics were extracted. First, subject behavior for each treatment trial was plotted with time on the horizontal axis and expansion on the vertical axis. Second, average expansion values were calculated as the ratio of modified to unmodified audio playback times. Third, final expansion values were read directly from the experimental records as the expansion value at the end of the audio clip. Fourth, modal expansion values were determined from the details of the experimental records, where the modal expansion value is defined as the expansion value for each trial that was experienced for the most amount of time. Since each subject experienced three treatment trials, the previously described metrics were averaged into a single score for each subject, i.e., a subject average expansion, a subject final expansion and a subject modal expansion.

Finally, the subjects' quiz scores were calculated automatically. Each passage, treatment or control, simulated lecture or simulated conversation, was followed by a multiple choice quiz in the same format: The first four questions had a single choice correct, which the fifth and final question required two selections out of four. The difference in question styles was clearly highlighted by the software. Each question was weighted equally, and worth 0.2 points, so that each quiz score ran from 0 to 1. Further, as each subject took three quizzes under the treatment condition and three under the control condition, these scores were averaged for each subject into aggregate control and treatment scores, also ranging from 0 to 1.

Figure 21 depicts the first trial of Subject Zero which began at a randomly selected expansion value of 1.43 (i.e., increasing playback time by 43%). The subject changed the expansion factor to 1.67 [+67%], then began the audio track (denoted by the dashed vertical line), moved back and forth between those two values before remaining at

Figure 21 Example of Subject Behavior Over Time



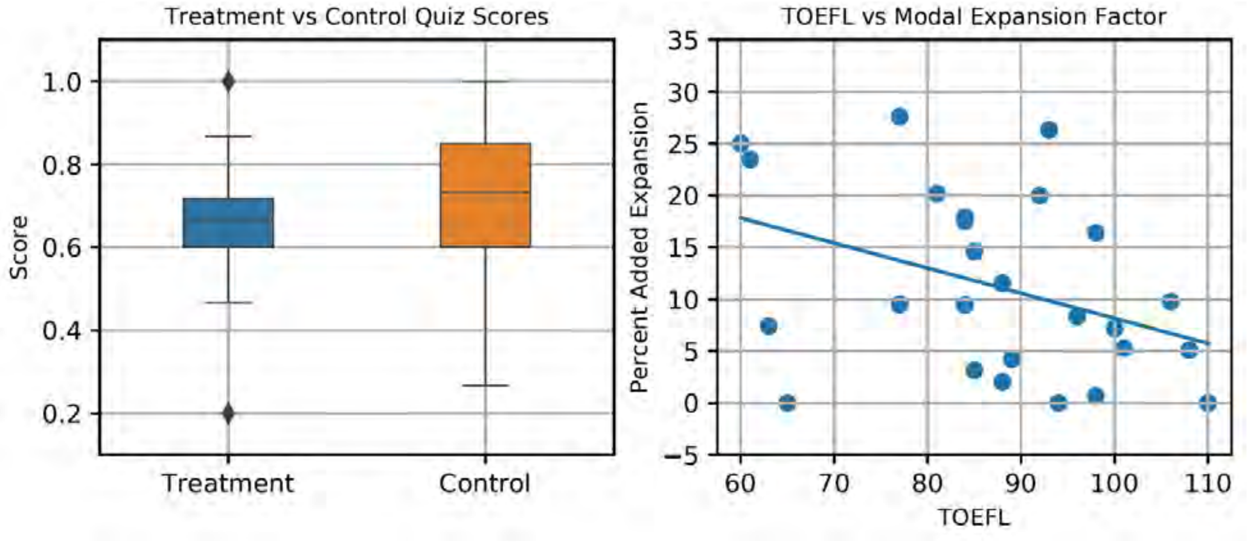
an expansion of 1.43 [+43%] for most of the track, and finally reducing the expansion to 1.33 [+33%] for the remainder of the track. The unmodified length of the track is 324 seconds, which, due to the expansion values applied, played out in 461 seconds. The modal expansion for this trial is 1.43 [+43%], the final expansion is 1.33 [+33%], and the average expansion is 1.42 [+42%]. A complete record of such behavior tracks can be found in Appendix D. The results of this chapter do not vary with the method of calculation (average, final, or modal), and all discussion following uses modal expansion values unless otherwise noted.

6.4 Results

6.4.1 Effects on Comprehension

Control and treatment quiz scores were tested for normality using the Lilliefors test. The control quiz scores (M = 0.70, SD = 0.19) and the treatment quiz scores (M = 0.65, SD = 0.15) both rejected the null hypothesis of normal distribution (p < 0.05). Therefore, a Kruskal Wallis non-parametric test was used to analyze the distributions of control and treatment data, and did not find evidence that the distributions of control and treatment quiz scores were statistically significant different (p > 0.05) when analyzed overall. See Figure 22 (Left).

Figure 22 Left: Effects On Reading Comprehension, Right: TOEFL vs Expansion



We can conclude that in this setting with this audio technique, user control of speech rate neither improved nor degraded subject comprehension.

6.4.2 Subject Behavior

Two aspects of subject behavior were examined during treatment trials. First, whether subjects used audio expansion at all. Utilization or rejection of the technique can be defined in two ways. First, if an individual treatment trial ended with the audio playing at an unmodified rate (i.e., an expansion value of unity) then that trial was considered a rejection of the expansion technique. Any subject who rejected the expansion technique during only one or two of their treatment trials is considered as partially rejecting audio expansion. Any subject who rejected audio expansion during all three trials is considered as fully rejecting audio expansion. The 26 subjects experienced a total of 78 treatment trials. By this definition, audio expansion was rejected in 16 of the trials (20.5%). Three of the subjects (11.5%) fully rejected audio expansion, six of the subjects (23%) partially rejected audio expansion, and 17 (65.5%) of the subjects used audio expansion in all three of their treatment trials. Of those subjects who partially rejected the expansion, there is no evidence that the expansion was used in earlier trials and rejected in later trials.

A similar analysis of subject utilization or rejection can be performed using modal expansion values (i.e., a modal expansion value of unity is considered to be rejecting the technique) gave nearly identical results: rejection in 15 of the trials (19%), five subjects (19%) partially rejecting the technique, and 18 (69%) of the subjects using expansion in all three treatment trials.

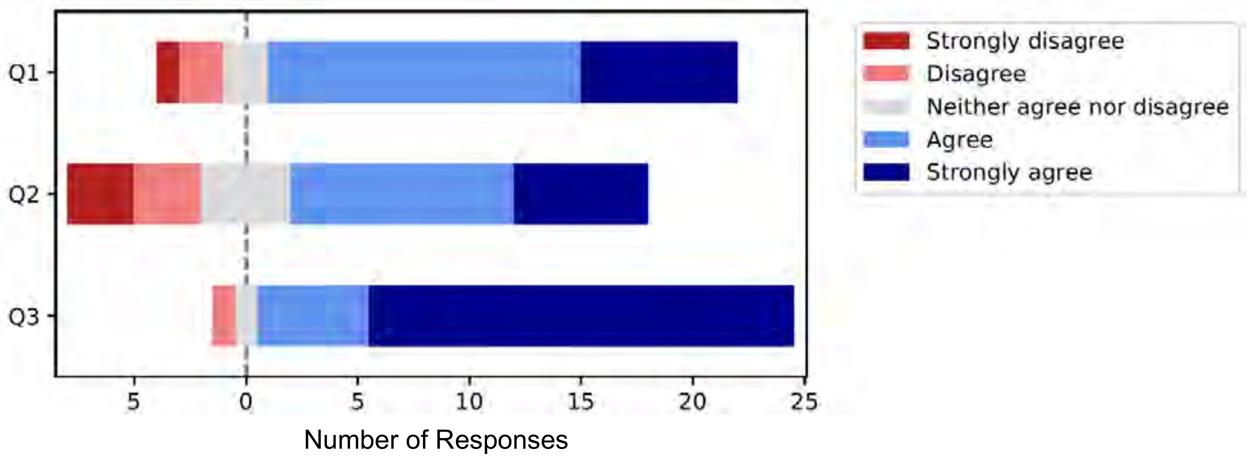
Second, whether and to what degree the subjects' use of audio expansion correlated with their TOEFL scores. To capture the effect of the audio experiences, subject average expansion scores are used, as described above. A linear regression model was constructed of TOEFL score and subject average expansion, which showed a mild reduction in the use of audio expansion as TOEFL score increased: with an R^2-value of 0.15, an increase of 10 points of self-reported TOEFL led to a 2.4 percent reduction in playback time. See Figure 22 (right).

6.4.3 Survey and Interview Data

Finally, survey and interview data collected from the subjects were analyzed. The first two Likert survey questions asked for the subjects' opinion of the usefulness of user-controlled audio expansion for listening skill.

The first item, asking about audio expansion and listening skill in general, received 81% positive responses ("Agree" or "Strongly Agree"), 7% neutral responses, and 12% negative responses ("Disagree" or "Strongly Disagree.") The second item, asking about audio expansion in relation to these specific clips received 61% positive,

Figure 23 Likert Survey, ESL



19% neutral, and 20% negative responses. The third item, asking about the ease of use of the interface, received 88% positive, 8% negative, and only 4% negative responses. These responses are summarized in Figure 23.

The spontaneous interview data revealed three recurring concerns among the subjects. First, 30% of the interviewees remarked at the length of the audio passages, and/or the difficulty of remembering information from the beginning of lengthy passages. Second, 40% of the interviewees noted their difficulties with the vocabulary, and that audio expansion does not help with this aspect of a listening skill test. Third, and possibly related, 45% of the interviewees spontaneously remarked that lecture passages were more difficult than conversational passages, often citing vocabulary as a factor.

6.5 Discussion

Behaviorally, there is strong evidence that subjects will use audio expansion techniques in an attempt to increase their listening comprehension: 88.5% of subjects used audio expansion in at least one of their three treatment trials, and 65.5% used audio expansion in all three of their treatment trials. The average modal expansion across all treatment trials was 1.11 [+11% playback time], and across all trials where audio expansion was not fully rejected was 1.13 [+13%]. These expansion values, while not extreme, are noticeable to an untrained ear.

Note also that the subjects' measured behavior is broadly in line with their survey responses. The second survey item ("Slowing speech for the audio files I listened to was useful") asked directly about audio expansion as regards these audio tracks, with 61% agreeing or strongly agreeing, while 65.5% of the subject did use the technique in all three treatment trials.

There is also evidence of personalization, with subject modal expansion factors ranging from 1.0 (i.e., a total rejection of audio expansion across all three trials, no added playback time) to a maximum average expansion of 1.28 [+28%] across three trials for one subject, and a maximum expansion of 1.43 [+43%] for an individual trial. There is also evidence that some of this personalization correlates with TOEFL scores, with greater language proficiency leading to less expansion. However, the effect is modest, with an R^2 value of only 0.15, and a change in expansion factor of only 0.024 per ten points of TOEFL (i.e., a change of 2.4 percent playback time per ten points of TOEFL score.)

However, when measuring objective performance, there is no statistically significant change to listening comprehension. These results are similar to those of the previous chapter, as reported in [2], where native English speakers' preferences for audio expansion were studied under various adverse noise conditions. In that study, when instructed to use (or not use) audio expansion to obtain the best speech intelligibility in background noise, subjects reliably selected increasing amounts of audio expansion with increasing amount of background noise. In that study, subjects also expressed through post-experimental survey a qualified belief that audio expansion was helpful for understanding speech in noise. However, that experiment demonstrated that user-directed audio expansion resulted in statistically significant *degradation* of performance on intelligibility tests. In both studies, the measured behavior and post-experimental survey data is at odds with subject performance.

Several factors may contribute to this effect. First, as in the previous chapter, the audio expansion is linear; once a subject selects an expansion value (in the absence of further adjustments) all part of speech are expanded equally. A number of sources [2, 38, 34] observe that clear speech is not produced by a process of strictly linear expansion, and that different phoneme classes may experience statistically different values of expansion or even contraction. Other subtle changes are also present, including modifications to vowel pronunciation and vowel space, as well as changing ratios of consonant to vowel energy. This may result in linearly expanded speech sounding somewhat unnatural, as indeed several of the subjects remarked in the interviews. This slightly incorrect cadence may inhibit comprehension.

It may also be the case that audio expansion, by lengthening the playback times of the audio passages, may be making it more difficult to remember information imparted throughout the audio clip. Nearly one third of the subjects expressed a concern about the length of the audio and their ability to recall information. In addition, almost half the subjects noted that while audio expansion may at times make it easier to hear or understand individual words, if the words were unknown to them then no amount of audio expansion would help.

These two factors may combine to create an illusion of improved performance, whereby the expanded audio is "easier" to listen to, which seems will be of some benefit, but these hypothetical benefits fail to materialize due to vocabulary or memory effects. If this is the case, I further speculate that a physiological cognitive load test may show direct evidence of a decreased load with increasing audio expansion.

Finally, note that that the TOEFL scores shown in Figure 22 (right) seem clustered into a low-score and a high score group, with low TOEFL scores characterized by high variance of measured expansion values. This may be similar to results found in [4] which suggest that once difficulties (there, noise-induced difficulties; here, skill-based) pass a certain threshold, the variance of subject behavior becomes very large, possibly because one setting is as good (or bad) as any other. However, this still suggests that what is found in lower difficulty (i.e., lower-noise or higher-skill) regions is a stronger illusion that audio expansion is beneficial.

6.6 Conclusions

This study gives ESL listeners fine, real-time control of the pace of lengthy audio passages. This tool was used to examine ESL subjects' preferences, performance, and perceived utility of the tool in conjunction with listening comprehension tests. There is strong evidence that ESL subjects will use audio expansion for the purpose of increasing their comprehension, and evidence of a small tendency to add more expansion as self-reported TOEFL skills decrease. However, there is no evidence that these subject-directed audio expansions improve listening comprehension.

I believe these results highlight a serious difficulty associated with subject-directed auditory interventions specifically, and subject-directed sensory modification in general. Namely, that users may have preferences and beliefs about the utility of sensory modifications, but that these beliefs may diverge from objective measures of performance.

7 Phoneme Aware Speech in Noise

This is the third and final of three chapters which design and implement user studies, which allow participants to make use of the ability to slow speech according to their own preferences. This is a direct continuation of the work in Chapter Six, but rather than controlling the rate of all speech uniformly, subjects are now given the ability to control the rates of individual parts of speech, by phoneme.

This chapter is adapted from material which is not yet published.

7.1 Introduction

In situations where the intelligibility of spoken words is degraded, listeners have few options to increase that intelligibility; rather, they are forced to rely on whatever adaptations to speech that talkers make. While these adaptations (including changes of pitch, formant space, vowel-consonant contrast, and especially speech rate) have been long studied for noisy environments (Lombard speech) [38, 34], when speaking to non-native speakers, [21], and when speaking to infants [78, 77], the talker must intuit what is helpful, while lacking the listener's direct experience of what is and is not helpful. In a sense, the talker imposes what the talker believes is helpful onto the listener. However, this may not be the optimal intervention for that particular listener.

Prior studies of computer enhancement of intelligibility which focus on changes of speech rate have also tended to follow the following paradigm: researchers often choose one or a few speech rate modifications and impose them on all subjects equally, and also tend to apply these changes uniformly across whole utterances [19, 79].

Modern computers now possess ample processing and networking capabilities and can perform real-time (or near real-time) speech modification, as discussed previously in Chapter Three. These capabilities offer new opportunities to place intelligibility-enhancing adaptations in the hands of listeners. Several studies use these techniques, including two previous chapters of this dissertation: Chapter Five describes and implements a study placing control of speech rate in noise directly into the hands of listeners with the goal of increasing speech intelligibility at the sentence level (also published in [4]), while Chapter Six reports on a study of speech comprehension at the level of paragraphs or longer passages also published in [5]). In such cases, the results have been negative, or at best mixed, failing to demonstrate improved intelligibility or comprehension, but nevertheless being preferred by test listeners. This may be because in such studies, while listeners had control over speech rates, the user-defined interventions were applied uniformly to all speech and parts of speech, resulting in unnatural cadences. Additionally, there is ample evidence that when talkers slow their speech naturally, to produce clear speech, they do so non-uniformly [94].

This chapter presents a study which places the ability to control the expansions or contractions of individual types of phonemes into the hands of test subjects, using this capability to (a) test the hypothesis that subjects will achieve improved sentence intelligibility by setting different expansion factors for different classes of phonemes, and (b) examine the details of the selected expansion/contraction factors.

7.2 Materials

This study is designed with three major components to test the hypothesis that giving listeners control over individual types of phonemes will enable them to increase the intelligibility of sentences received in the presence of noise. Additionally, to facilitate testing during the Covid-19 pandemic, the study was designed to be administered remotely; subjects took the study from their homes using networked computers, during lockdown and suspension of in-person studies. The three components are:

First: A neural network, inspired by [104], that was trained to recognize categories of phonemes in audio tracks. The network was trained on sentences from the TIMIT database [105], which had been divided into 25 ms length non-overlapping frames, each of which were decomposed into 39 mel frequency cepstral components [106] (MFCCs). These MFCCs were used as inputs to a simple Bi-directional Long Short Term Memory (BLSTM) architecture neural network, with three hidden layers of 500 LSTM cells each. Rather than training the network to classify all 39 phonemes found in English speech, the network was trained on five broader categories of phonemes: Vowels, Fricatives and Affricates (henceforth: “Fricatives”), Approximants (comprising Liquids and Glides), Nasals, and Stops (including both the closure and the release of each plosive) as well as a sixth category (Silence) comprising various types of silences, especially pre- and post-utterance silences for completeness. This schema was inspired directly by [107] to yield the best accuracy possible for phonemic classification, while also keeping the number of classifications to a minimum, in order to keep the complexity of the subject engagement to reasonable levels. This classification also naturally yielded a classification with categories easily explained to and understood by non-expert study participants. In this case, the categorization can be understood as roughly corresponding to the degree of obstruction in the vocal tract: Vowels are voiced and without any obstruction in the vocal tract; Approximants are voiced but with very slight obstructions (i.e., the tap of the tongue against the teeth or top of the mouth forming the 'l' sound in 'yellow'); Nasals are such that the obstruction re-routes airflow through the nose causing a changing in resonances and formants; Fricatives are produced by significant obstructions in the vocal tract that cause airflow to become audibly turbulent or noisy; and Stops are obstructions so great that they temporarily halt the airflow, allowing pressure to build up and release with a sudden onset of voiced sound. This is not the only possible grouping that is logical from a phoneticist's point of view and that is easily explained to non-experts—a distinction between vowels, voiced consonants and voiceless consonants is also reasonable. But per [107] those schemes are likely to be less amenable to neural network classification.

The only deviations from that scheme are the merging of Fricatives and Affricates into a single category, and the reclassification of ARPAbet's 'hh' sound into the approximant category, as it is in the ARPAbet classification itself. Pauses, although incorporated in 'Silence' above, are not considered as a separate category because, in the corpus detailed below, the utterances are short enough that intra-utterance pauses are rare. (If developing a similar application for longer passages of speech, especially spontaneous of conversational speech which may be rich in such pauses, it would be wise to consider such pauses as a separate category.) The complete list of phonemes in each category are provided in Table VI below. The performance of this network was comparable to that of [107]. This network reached frame-wise accuracies of 97% or greater on all phoneme categories in a training set of 3696 sentences, and 84% or greater on all categories except approximants (72%, often confused with vowels) on a test set of 1152 sentences. Detailed confusion matrix data can be seen in Appendix F.

Table VI Phonemes By Category

Group	**ARPAbet Phonemes**
Vowels	iy ih eh ae ax uw uh ao ey ay oy aw ow er
Approximants	l r w y hh
Nasals	m n ng
Fricatives	s z zh f v th dh hh jh ch
Stops	b d g k p t dx (and closures)

Second: A web-based interface allowing subjects to (1) specify their preferences for phoneme expansion or contraction via an on-screen control slider, (2) test the intelligibility of modified to unmodified sentences in various levels of background noise via transcription tasks, and (3) answer a brief post-study survey. The control slider ranged from expansion factors of 2.5 at the rightward edge, and 0.4 (i.e., contraction) at the leftward edge, with a value of 1.0 (no modification) at the center. Note that the control slider contained a non-linearity on the left side: Points representing X and 1/X expansion are always equidistant from the center. This prevents the contraction portion of the slider from being much smaller than the expansion portion, which would likely bias the subjects' response toward the expansion portion. Audio tracks were expanded or contracted using a frame-based interpolating phase vocoder [58] and normalized for sound levels per ITU-R BS.1770 with Pyloudnorm [109].

Third: Lists of Harvard sentences [83] drawn from the University of Washington/Northwestern University (UW/NU) Corpus 2.0 [108], Speaker PNF137, and four-person babble tracks drawn from the QuickSin Speech in Noise Test [81], [82]. Each such whole list comprises ten short sentences, typically less than ten words each, with the sentences in each list chosen to be phonetically balanced, i.e., with phonemes having approximately the same frequency as in spoken English. Three whole lists were selected for various practice and familiarization tasks; six whole lists were selected for transcription tasks (described below); and five composite lists were constructed, one for each phoneme group under consideration. The composite lists only were composed of other lists to be rich in the phonemes under consideration; whole lists were selected intact, to maintain their phonetic balance. The reasons for the composition of these lists will be discussed in section 8.3.2 below.

7.3 Methods

7.3.1 Participants

Forty-four healthy (with no self-reported hearing problems), young (age 18 – 30), native speakers of English [110] took part in this study. Participants were recruited from three student mailing lists: University of Illinois at Chicago engineering undergraduate students, University of Illinois at Chicago engineering graduate students, and graduate students from all University of Illinois campuses.

The study took approximately one hour; subjects who completed it received a $12 Amazon gift card, equal to one hour of minimum wage in Illinois during the year of their participation. This study was approved by the UIC

Office for the Protection of Research Subjects as protocol 2012-0112, and subjects gave informed consent through a web interface prior to the start of the study. See Appendix E.

7.3.2 Procedures

The study was administered automatically and remotely in five phases. In all cases, background noise was four-person babble.

Screening and Consent: Subjects were screened for age, hearing problems, native speaker status, consent to the use of "attention questions" (discussed below) and required to commit to the use of over- or in-ear headphones rather than external speakers. With satisfactory screening answers, they were shown a university consent document and asked for consent. Finally, they were instructed to disengage any noise cancellation features on their listening devices.

Introduction: Subjects were introduced to tasks they would perform later in the study, and to a "target voice" that subjects would need to extract from background noise. Although this target voice is always presented at audio levels louder than the background noise, this early introduction mitigates against subjects being unable to identify the voice in later phases.

Audio Adjustment: Subjects were asked to listen to at least three (and as many as desired) sentences, while adjusting their own volume control to a comfortable level.

Slider Familiarization: Subjects were presented with the on-screen slider and told that it would slow down (rightward) or speed up (leftward) the foreground speaker's voice. Subjects listened to at least three and as many as six sentences in light background noise, adjusting the speech rate with the slider interface for each sentence.

Transcription Familiarization: Subjects were asked to listen to exactly three sentences in varying levels of noise and transcribe as many words as possible back into the interface. Each task included a visually displayed, randomly drawn "pass phrase" which the subjects were instructed to enter verbatim *if and only if the sentence was completely unintelligible*. One of the familiarization sentences was a simple transcription task (STT), played against mild 10 dB SNR noise. Another sentence was a Noisy Attention Task (NAT), played against -10 dB SNR noise; this level of noise causes the target sentence to be unintelligible, and subjects were expected to use the pass phrase. One sentence was a NoiseLess Attention Task (NLAT), played against 60 dB SNR noise; this level of noise is inaudible, and subjects were expected to easily extract the correct sentence.

These attention tasks were developed as advised by [111] as objective methods to determine if subjects were not paying attention or otherwise not engaging honestly with the task (e.g., providing minimal effort to finish the study as quickly as possible to proceed to a payment phase.) Subjects not entering an NLAT sentence correctly (up to minor typos) and subjects not entering the correct pass phrase during an NAT were subject to disqualification and non-payment. The precise nature of the attention tasks was not disclosed, although their presence and potential non-payment was noted in the consent document.

Personalization Phase: Subjects were presented with a sequence of five personalization tasks, all the same format. On the screen, subjects were shown the control slider, and brief instructions informing them that the slider now expanded only certain parts of speech, e.g., “fricatives and affricates,” along with an exhaustive set of examples, e.g., “The S in sleep”. Subjects were instructed to use the slider to adjust the expansion or contraction of these phonemes so that the sentences (played against 5 dB SNR noise) were easiest for them to understand.

For each of the following categories of phonemes as defined in the previous section, the subjects were required to listen to at least six, and as many as ten, sentences: Vowels, Approximants, Nasals, Fricatives, and Stops. The lists of sentences in this phase, as noted above, were not drawn from individual lists of Harvard Sentences. Instead, sentences in these lists were individually selected from the Harvard Sentences at large to be rich in the targeted phonemes; otherwise, it would be possible to randomly draw sentences which contained none of the target phonemes at all, defeating the purpose of this phase and potentially confusing or frustrating the subjects. The statistics of the target phonemes in these Personalization Lists is detailed in Table VII. No attention tasks were presented in this section of the study.

Table VII Statistics of Target Phonemes in Phoneme-Specific Lists

List	Average	Std Dev
Vowels	9.2	1.0
Approximants	3.5	0.5

Nasals	2.1	0.3
Fricatives	5.1	0.3
Stops	4.6	1.3

Intelligibility Phase: Subjects were presented with six intelligibility *tests* consisting of eight transcription *tasks* each. Each intelligibility test tested a different condition: Three tests where all phonemes were modified as the subject specified during the personalization phase, at 8 dB, 5 dB, and 2 dB SNR background noise; and three tests where no phonemes were modified at all, again at 8 dB, 5 dB, and 2 dB SNR background noise. Each set of eight tasks included six STTs (the basis of subsequent scoring), one NAT, and one NLAT, with the NLAT having no modified phones, to present the sentence in perfect clarity. The phonetic balance of these Transcription Lists is detailed in Table VIII.

Table VIII Phonetic Balance of Transcription Lists

List	**List 1**	**List 2**	**List 3**	**List 4**	**List 5**	**List 6**	**Avg (StdDev)**
SNR	8	8	5	5	2	2	N/A
Modified	Yes	No	Yes	No	Yes	No	N/A
Vowels	61	53	52	56	56	54	55.3 (3.2)
Approximants	20	25	25	19	25	23	22.8 (2.7)
Nasals	12	12	16	9	17	13	13.2 (2.9)
Fricatives	28	24	26	34	34	28	29.0 (4.1)
Stops	33	30	38	37	31	28	32.8 (4.0)

Survey Phase: Finally, subjects were given an exit survey, with each question on a seven-point Likert scale, eliciting opinions about the helpfulness of the phoneme expansion technique in general, for intelligibility, for ease of listening, and the naturalness of the resulting speech.

7.3.3 Experimental Records

All records are electronic, recorded and anonymized by the study software on a secure server, including:

- Consent, and answers to screening/survey questions

- Slider positions for all apt phases and activities
- Copies of all audio played to the subjects
- Subject transcriptions or use of "pass phrases", and the task's status as STT, NAT, or NLAT

7.3.4 Analysis Techniques

The basic objects of analysis in this study are the subject engagement, sentence transcriptions, expansion/contraction factors of each phoneme group, and Likert response data.

Engagement: Engagement analysis was as follows: All NAT/NLAT responses were inspected and compared to the pass phrase (NAT) or target sentence (NLAT). While good faith errors were detected (e.g., interpreting a prominent babble voice as the target voice during an NAT, especially in high noise conditions; various typos) no evidence of non-engagement was found in any subject.

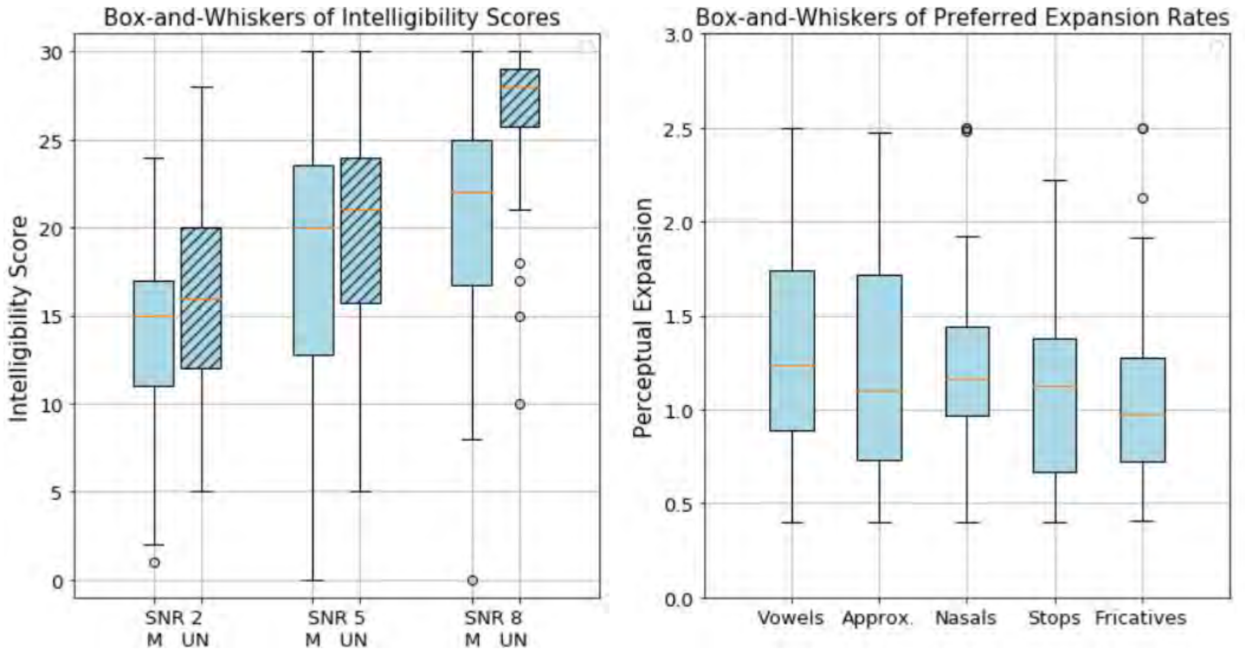
Intelligibility: Intelligibility scores for each intelligibility test were derived as follows. Each STT sentence contains five pre-determined keywords, with each correctly transcribed keyword worth one point. STT scores were summed within each test, resulting in an intelligibility test score ranging from 0 to 30 for each of the six tests. STT sentences were scored by hand.

Preferred Expansion Factors: Each subject determined their own preferred expansion or contraction factor for each of the five phoneme groups. Because of the previously mentioned non-linearity of the slider, these factors are analyzed perceptually, i.e., from expansion factor 0.4 (maximum expansion) through 1.0 (no expansion or contraction) to 2.5 (maximum contraction), and by direct slider position, i.e., from 0, (expansion 0.4) through 50 (expansion 1.0) to 100 (expansion 2.5). This distinction did not change any results or conclusions. Thus, only perceptual data are presented and discussed below.

7.4 Results

Intelligibility: Box and whisker plots of intelligibility scores at tested conditions are shown in Figure 24 (left). Intelligibility scores for all six test conditions (Modified and Unmodified; both at 2 dB, 5 dB and 8 dB SNR each) were analyzed for normality with a Shapiro-Wilk test. Multiple conditions were found to be non-normal; see Table IX (normal values in bold at p <= 0.05 cutoff). As such, a two-way ANOVA test is inapplicable. Instead, the application of a Scheirer-Ray-Hare test indicates that the effects of SNR (H = 59.63, p = 1.13e-13) and modified vs unmodified phonemes (H = 12.71, p = 3.65e-4) are significant, with interactions between these effects (H = 6.50, p = 3.88e-2).

Figure 24 Pairwise Intelligibility Scores by SNR (Left), Preferred Expansions of Phonemes (Right)



Again due again to the non-normality of the data, Tukey's Honest Significant Difference procedure is inappropriate. Instead, pairwise Wilcoxon tests were performed on all fifteen possible pairs followed by a Benjamini-Hochberg correction. Of these fifteen pairs, presented in on the first page of Appendix G, nine have interpretative value: Three pairs which directly compare modified to unmodified scores at the three SNR conditions; three comparing scores at the three SNRs within the Modified condition, and three comparing scores at the three SNRs within the unmodified condition.

Preferred Expansion Rates: Box and whisker plots of the preferred expansion rates of each of the five categories of phonemes are shown in Figure 24 (right). The preferred expansion rates were analyzed with Shapiro-

Wilk for normality and two could not be assumed to be normal. Again, due to non-normal data, a one-way ANOVA test is inapplicable, and instead employed a Kruskal-Wallis test (H = 6.86, p = 0.14) was performed, which indicates that none of the preferred expansions differ significantly from one another; no pairwise tests were necessary. Table X presents the mean and standard deviation of the phoneme groups, as well as the p-values of the Shapiro-Wilk tests (normal values in bold at p <= 0.05 cutoff), and the percentage of subjects expanding, rather than contracting, these phoneme groups.

Expansion Correlations: Linear regression analyses were performed on all possible pairs of phoneme group data. The resulting R^2 values ranged from 7.64e-8 to 1.01e-1, yielding no evidence of correlation between any phoneme group pairs, i.e., the expansion of any one particular phoneme class does not correlate with the expansion or contraction of any other. A detailed set of plots and R^2 values is presented in Appendix H.

Table IX Intelligibility Normality (Shapiro Wilk)

p-values	**SNR**		
	2 dB	5 dB	8 dB
Modified	**.202**	**.104**	.023
Unmodified	**.786**	**.057**	6.08e-7

Table X Phoneme Expansion Statistics

	Vowels	**Approximants**	**Nasals**	**Stops**	**Fricatives**
Mean	1.31	1.20	1.22	1.11	1.07
St Dev	0.54	0.55	0.51	0.49	0.50
% Expand	63%	59%	65%	56%	41%
p-value	**0.066**	0.032	0.015	**0.133**	0.001

Intelligibility Vs Expansion: The effects of individual phoneme expansions on intelligibility were also analyzed. For each test condition (2, 5, and 8 dB SNR) and each phoneme group, a quadratic polynomial was fit to the scatter plot of the phoneme expansion vs intelligibility scores, for both modified and unmodified data. Figure 25 shows the collection of such plots. Note that in all cases, the quadratic coefficient (β_2) is negative, indicating that the best fit polynomials are downward concave, and implying the existence of theoretical optimal expansions, which are called E* at the peaks, with performance dropping off at greater or lesser values of expansion. Table XI, below, summarizes the locations of these peaks, E*, and the values of the quadratic coefficients, β_2.

Figure 25 Phoneme Expansion Estimates vs Intelligibility

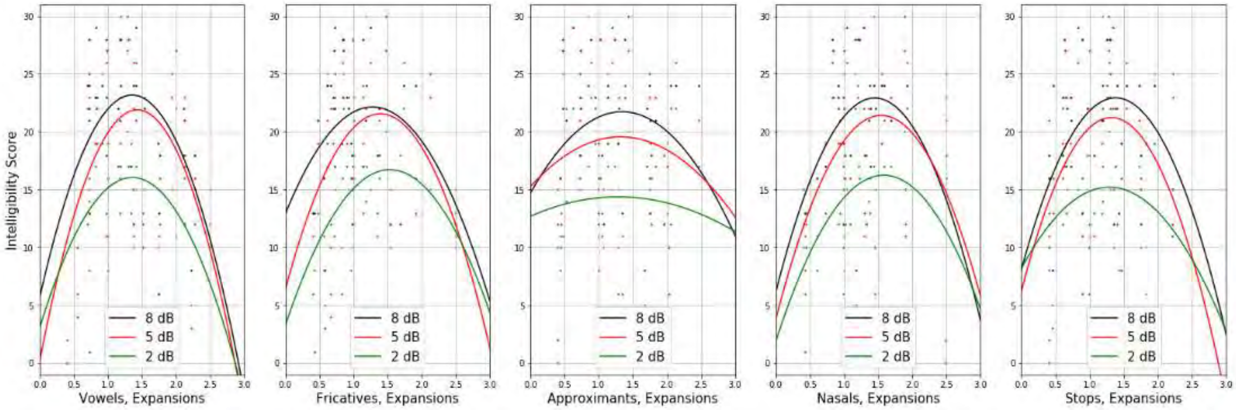


Table XI Predicted Phoneme Expansion Peaks

		Vowels	Approximants	Nasals	Stops	Fricatives
2 dB SNR	E*	1.35	1.34	1.45	1.38	1.28
	β_2	-7.04	-1.02	-5.73	-4.23	-5.73
5 dB SNR	E*	1.43	1.32	1.54	1.32	1.39
	β_2	-10.61	-2.45	-7.37	-8.61	-7.82
8 dB SNR	E*	1.36	1.28	1.58	1.29	1.53
	β_2	-9.51	-3.89	-7.98	-7.76	-5.66

Survey Data: Each subject was asked four questions, each on a 7-point Likert scale. Results are shown for both data sets in Figure 26.

Question 1 "*In general, how helpful or harmful was changing the cadence of the woman's speech*?" (From 1: "Very Harmful" to 7 "Very Helpful")

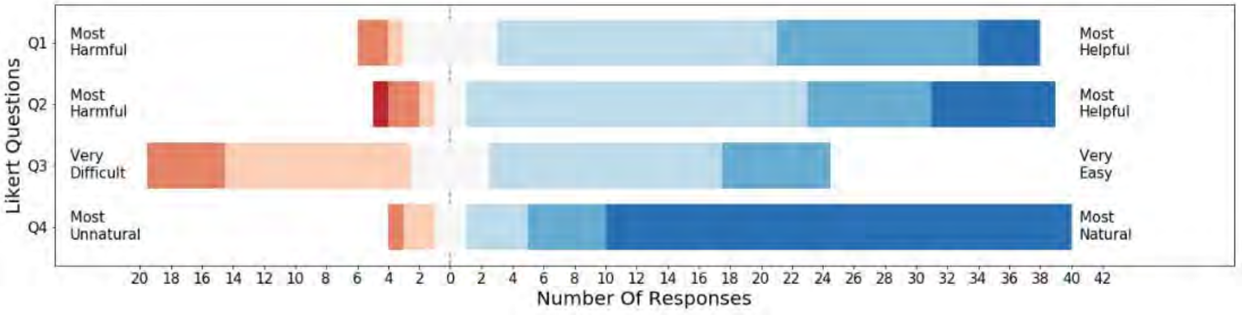
Question 2 "*For understanding the woman's speech in noise, how helpful or harmful was changing the cadence of her speech*?" (From 1: "Very Harmful" to 7 "Very Helpful")

Question 3 “*Compared to normal speech, how easy or difficult was listening to the speech with changed cadence”* (From 1: “Very Difficult” to 7 “Very Easy”)

Question 4 “*How natural or unnatural did you find the speech with modified cadence, relative to unmodified speech?”* (From 1: “Very Unnatural” to 7 “Very Natural”)

Motivated by the strikingly bimodal distribution of answers to the third question (17 subjects answered that the modified speech was more difficult to listen to than unmodified speech; 22 subjects answered that the modified speech was easier to understand; and 5 subject answered neutrally) the original dataset (N = 44) is partitioned into two small sets— “Hard” (N = 17) and “Easy” (N = 22)—while discarding the remaining neutral answers. The box and whisker plots of these new sets are presented, in comparison with the complete dataset, in Figure 27 below. Additional statistics on the pair-wise comparisons for these conditions are also added to the second and third sheets of 0.

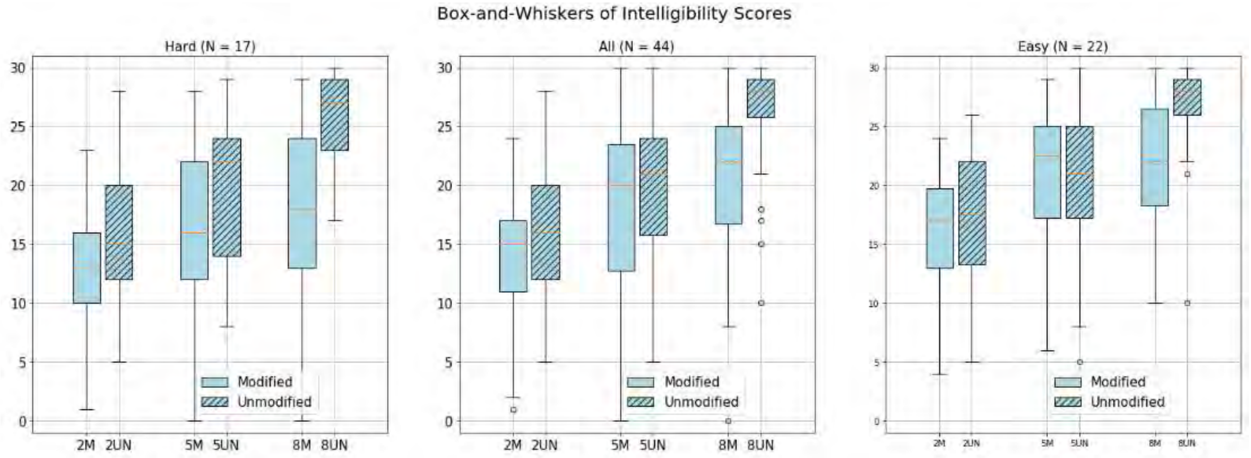
Figure 26 Likert Survey, Phoneme Speech In Noise



7.4.1 Discussion

As discussed previously, no evidence of disengagement with the transcription tasks was found, possibly due to the attention tasks and possibility of non-payment.

Figure 27 Pairwise Intelligibility Grouped By Ease Of Listening



There is broad evidence that subjects overall attempted to use phoneme expansion to increase intelligibility. In Figure 24, above, note that for all phoneme groups, the average behavior of the subjects was to expand all phonemes, if only in some cases by a small amount. Moreover, the voiced phonemes (vowels, approximants, and nasals) experience higher percentages of subjects choosing expansion and (although without statistical significance) experience higher levels of expansion. This is broadly in line with prior detailed studies of Lombard speech. Prior research [112] shows an overall lengthening of voiced relative to unvoiced sounds, ranging from 2% to 18% (average 8%) increase across subjects in Lombard speech conditions; [113] showing (some) shortened stops and (some) lengthened approximants and especially vowels. Finally, the range of expansions selected (i.e., the broad standard deviations on the datasets) shows strong evidence that the subjects select their expansion or contraction values idiosyncratically, rather than clustering sharply around precise values.

The analysis and quadratic fit of phoneme expansions vs test scores suggests that performance on the intelligibility tasks peaks at modest expansions of between 30% and 50% additional length, and degrades rapidly at extremes of either expansion. This observation agrees with previous anecdotal and pilot studies that extreme values of expansion or contraction are not useful. These values are also in line with the upper values used or found in [114], which reports 24-62% increases in vowel and semi-vowel duration, and only a modest 3% expansion of consonants. Of interest is that the quadratic fit predicts somewhat greater expansion values than the direct averaging of the slider positions. These results may prove useful in determining narrower experimental values for future studies, and as far as I am aware represent the first experimental evidence of phoneme-level duration or expansion preferences for *listeners* under noisy conditions.

The results of the intelligibility scores across tests (modified and unmodified, and three different SNR values each) are somewhat puzzling. Statistical analysis shows that the only pairwise comparisons of the complete dataset (N = 44) that are *not* statistically significant are the 5 dB modified vs 5 dB unmodified conditions, and the 8 dB modified vs 5 dB unmodified conditions, the latter of which is not a comparison with any interpretive value. Intelligibility increases with increasing SNR (in both modified and unmodified conditions) as expected, and all of those relevant pairwise comparisons in that sense are statistically significant. However, the interpretation of the modified vs unmodified comparisons is more difficult; only two of them (2 dB modified vs 2 dB unmodified, and 8 dB modified vs 8 dB unmodified) were statistically significant, and in both cases the intelligibility scores *decreased* under modification. However, during the personalization phase when the modifications were chosen, the background noise was set to 5 dB SNR, and the pairwise comparison at that noise level shows no statistical significance. One possible explanation for this is that expansions chosen for a particular SNR are highly attuned to that SNR (in this case, 5 dB), and although listening to expansions at the SNR at which they were obtained does not improve intelligibility, deviations in either direction degrade it.

The survey/Likert data overall shows strong statistical belief that the technique of modifying phoneme durations was helpful, both in general (Q1) and specifically (Q2) for understanding the target voice in noise, which is directly at odds with performance on the intelligibility tasks. These results are similar to the findings in [4, 5] and constitute an ongoing puzzle. This belief explains the subjects' use of the technique, but the reason for the belief itself remains unexplained. Those previous studies speculated that slowing speech down may make the target voice easier to hear or listen to, without making it easier to understand, and that subjects considered one or more of these qualities as "helpful." It may also be the case that although these techniques allow subjects to roughly approximate patterns of clear or Lombard speech, which aids in hearing the voices, the results may not be sufficient to truly increase intelligibility. The effects of relative phoneme positioning [115] cannot be emulated with this technique, nor the non-durational components such as energy ratio, spectral tilt, etc. There is also broad agreement among the subjects that the cadence of the resulting speech is unnatural. The overall effect might be the result of improved ability to hear the voice, but degraded intelligibility due to the unnaturalness of the cadence.

However, as previously noted, the results of Q3 are strongly bimodal, suggesting a partition of the complete (N = 44) dataset into "Hard and "Easy" datasets (N = 17, N = 22, respectively, discarding datasets corresponding to neutral responses), and re-analyzing aspects of those new datasets. This reveals a trend, which, although there is little statistical power behind it, is very suggestive: In general, those who responded that listening to the modified voice was harder tend to perform worse at the 2 dB and 5 dB SNR conditions; likewise, those responding that the modified voice was easier tend to perform better. Note especially that the "Easy" dataset no longer shows a statistically significant (p = 0.34) degradation in intelligibility from modified to unmodified in the 2 dB condition; and that although there is no significance to it, in the 5 dB SNR condition, the median value of the modified condition is now higher than the unmodified condition. This strongly suggests that subjects' responses about ease of *listening* correlate more meaningfully with their performance than their responses about the ease of *understanding*. (It may also be that questions about ease of listening, which are not directly related to the purpose of the study, are also less subject to response bias.)

Overall, these results do not demonstrate that granting subjects the ability to manipulate individual classes of phoneme results in increased intelligibility. However, there are several ways in which this study could be modified based on its findings.

First, there is strong evidence in this study and the study reported in Chapter Five, that different levels of noise prompt different levels of expansion; now, with this study, there is also strong evidence that these differences may be necessary, i.e., that expansion factors chosen at 5 dB SNR may be inapplicable or even cause degradation of intelligibility at SNR levels only a few dB away. The follow-up and potential remedy to this is to redesign the study to also collect phoneme personalization profiles at other SNR levels. However, this may prove difficult in practice—this study took approximately an hour for each subject to complete, and adding two additional sets of five phoneme classifications may cause the study to expand to well over an hour and a half, running the risk of subject fatigue and/or involuntary disengagement from the task. A more practical approach may simply be to triple (or more) the number of subjects, partition them in advance into three groups, and collect phoneme profiles at each of the three noise conditions.

Second, it may be useful to increase the number of subjects without changing the study; the sub-partitioning of dataset into the "Easy" and "Hard" groups resulted in groups of half the size of the original study or less, which may simply be too small to show statistical significance.

Third, it may be useful to introduce a new screening phase into the experiment: Before any detailed instructions are given (i.e., before subjects can infer the goal of intelligibility), present them with a uniformly expanded set of sentences in noise (or a set of sentences with only vocalized phonemes expanded) and screen them on the basis of their answer to a question similar to Question 3 of this study. The expansion factor would be roughly in line with the predicted optimal results of this study, approximately 1.3. Subjects screened out must still be compensated, but at a lower rate, thus extending the funding of any future study.

All the approaches above, however, assume a plentiful supply of what in fact is often a scarce resource: experimental subjects. The initial concept of this study used Amazon Mechanical Turk (AMT) workers, which are as plentiful as required for larger studies of this sort. Unfortunately, repeated pilot tests using friends and/or UIC students yielded considerably different results than pilot tests performed with anonymous AMT workers; AMT workers performed noticeably worse, even though the Attention Questions did not indicate any type of disengagement. The reasons for this are entirely unclear, and for that reason the AMT workers were replaced with a more traditional subject pool of University of Illinois students. However, if the issues with the AMT approach were understood and corrected, this might prove to be an easy and reliable means for acquiring new test subjects, subject to the limits of funding.

Finally, it may be the case that this study was simply too complex for the subjects. This may be due to the large number of factors (five phoneme groups) that the subjects were asked to control, including all of their potential interactions, while also being asked to change those five factors one at a time, in a proscribed order, without the possibility for subsequent readjustment. This may also be due to the expertise required of the subjects; while native speakers of English might be expected to know the difference between a vowel and a consonant, it is not reasonable to expect a detailed understanding of which phonemes fall into which categories. While the instructions were tailored to alleviate this, by providing exhaustive lists and contextual examples of each phoneme under consideration, short text instructions can only do so much. While the expansion and contraction of individual phoneme classes may be useful, it may simply not be feasible to expect subjects to evaluate all of the necessary trade-offs for this approach to bear fruit. If this is true, then modifying or entirely removing the concept of ‘user choice’ might be in order. However, this discussion is deferred to the concluding chapter of the dissertation.

7.4.2 Conclusions

This study gave listeners very detailed control over speech signals, allowing them to lengthen or shorten five different groups of phonemes to differing degrees, and instructed them to use this control to increase their ability to understand speech in background noise. Subjects display a general willingness to use the tool to this purpose; \more subjects used expansion for voiced phonemes vs voiceless phonemes; and extreme values of expansion or contraction correspond to worse performance. A strong majority of subjects also believed this technique was helpful.

However, there is no statistical evidence that this technique increases or improves intelligibility. There is no evidence of help or harm at the 5 dB SNR where subjects determined their preferred expansions, and there is evidence of degradation of intelligibility at higher and lower SNR value. Finally, there is no evidence that subject expand any phoneme group, as this study has grouped them, more or less than any other.

8 Conclusions

8.1 Contributions

This research program was sparked by one simple observation: that speech spontaneously delivered to listeners in a variety of difficult situations—to the hard of hearing, to the elderly, in noisy environments, etc—seems to be delivered more slowly than speech delivered without those complicating features. This observation cascaded into a host of questions, some of which could be answered (in part, if not definitively) through surveys of existing literature: Is this perception that speech delivered to distressed listeners is slower actually true? Yes, the umbrella term for speech modified to assist distressed listeners is clear speech, and in an overall, statistical sense it is indeed slower. [19, 21, 29] Is this change of rate the only modification made? No, it is not, there are a host of other changes made as well. [33 - 38]. Does this modified speech actually increase the intelligibility in listeners? Again, yes, in a statistical sense [38 - 45], but once one moves past the statistical summaries, there is strong evidence that clear speech is not identical from person to person: Some talkers produce unusually strong or weak boosts to intelligibility, and conversely some listeners receive unusually strong or weak boosts to intelligibility[43, 46].

That final literature-fueled observation led to a sharpened series of research questions, which I rephrase from the summary of this dissertation:

First: Is it feasible, using modern off-the-shelf hardware (i.e., laptops or smart phones) and custom signal processing software, to produce a signal processing system which allows real-time modification of the speed of speech signals?

Over the space of several projects and publications, I have demonstrated that this is possible. As part of a project in a research seminar, [1] I developed a working prototype of such software, a so-called half-duplex application which could receive one audio stream through microphones, and transmit a modified (slowed) audio stream through speakers. Critically, this modification of speed did not modify the pitch, but did automatically turn off audio expansion when the input stream was silence. This software also featured a highly responsive, real-time control system, such that a listener could change the expansion factor and hear the output audio stream respond immediately. This software was used directly to investigate the second research question below. However, in [2] I extended this half-duplex laptop application to a full-duplex networked application which ran on Android smart phones. This project, still a prototype, allowed two smart phones on the same network to connect to each other, send and receive audio data to and from one another, while allowing each user to slow down the speech of the received audio only.

Finally, subsequent unpublished work placed a desktop "server" computer with a known, stable IP address between the two smart phones as an intermediating device. This innovation allowed two smart phones to connect even at long distances, with the handsets separated from the server by hundreds of miles; effectively, I designed a proof-of-concept (if small-scale) VOIP system which incorporated live, user-controlled audio rate changes into the design. Although the audio modification algorithms were not novel, this was to my knowledge the first time they had been demonstrated in live systems designed for direct human communication. There is no fundamental barrier to incorporating this technology in larger systems, or even videoconferencing systems, as long as the video is slowed down in lockstep with the audio.

Second: If such a system is created, will the technique of slowing speech in this way be well-tolerated by subjects conversing with it, or will the technique itself directly interfere in the communication process?

To investigate this issue, I and collaborators designed a novel study using two instances of the single-talker half prototype described above, on two laptops, to investigate the effects of this technique on human communication and interaction using multiple Diapix tests. Although I am aware of one other similar study in which researchers intentionally modified their own natural speech rates to study the effects of communication rates on subjects, I am unaware of any other studies of this nature using computer intermediation to enforce rate changes, or any studies which induce rate changes on two conversational partners at once. Although by its nature the study could not be comprehensive, and induced audio expansions only at literature-informed "best guess" values, the results were encouraging. The technique was well-tolerated, in the sense that no conversational aberrations (such as conversational desynchronization) were detected, no distress was reported, and artificially changing the tempo of one or both subjects induced no statistically significant changes in naturally produced speech rate of either subject.

Third: Given the evidence of idiosyncratic production of clear speech, and idiosyncratic benefits of listening to clear speech, is it useful or effective to place the control of speech rate directly into the hands of listeners?

To investigate this, I designed and implemented a suite of three user studies: First, a relatively straightforward study which allowed subjects to adjust (specifically, to slow down) the speech of one particular talker in the presence of four-person babble background noise at various levels of noise. This was followed by intelligibility tests of modified and unmodified speech in all of these noise conditions. This study confirmed several of its hypotheses: Subjects do use this technique to try to improve the intelligibility of received speech in noise; they do so idiosyncratically, and they react by increasing the expansion of their received audio in response to rising levels of noise. They also, based on survey results, believe that this intervention helps them. However, objective measurements indicate that this technique actually degrades intelligibility slightly, and higher noise levels.

Second, another relatively straightforward study which allowed ESL subjects with recently obtained TOEFL scores to adjust the speech rate of noiseless audio during simulated TOEFL tests. Here, the focus of the study (and study instructions) was on overall long-passage *comprehension* rather than short sentence *intelligibility*. Again, however, the results are puzzling: Subjects overwhelmingly used the technique to slow audio: 65.5% of subjects expanded their audio in all of their trials, while 88.5% did so in at least one trial. The survey results show similar enthusiasm for the idea: 81% responded that the technique was useful for listening in general, while 61% responded that the technique was useful for the specific audio clips to which they listened. However, there was no statistically significant effect in either direction on the subjects' listening comprehension, nor was there strong explanatory power in the correlation between TOEFL scores and final expansion rates.

Third, returning to the theme of intelligibility, I hypothesized that the puzzling results of the first study may be due to the unnatural cadence of speech produced by uniform slowing of speech. Although detailed statistics are not available, it is well understood that naturally produced clear speech changes the rate of speech in non-uniform ways: Some parts of speech are expanded more than others, and some parts, in some contexts, may even be compressed. To this end, I designed and trained a neural network to act as a phoneme classifying pre-processor and designed a new study around the expansion of particular phoneme groups under noisy conditions. The results were similar to the intelligibility in noise study using uniform expansion: Subjects used the technique, and did so in individual and personalized way, and reported that they believed this helped them understand sentences in noise better, but the objective results of intelligibility tests did not bear these beliefs out.

8.2 Discussion

The first of my research questions, “Is it feasible to create devices which place real-time audio rate changing effects into the hands of listeners?” is effectively answered in the first several chapters of this dissertation: Yes. The processing power of good laptops then, and even common smart phones now, is more than sufficient to handle all necessary aspects for such applications. Processors have sufficient power to perform the various Fourier transform, inverse Fourier transform, and interpolation operations seamlessly and without perceptible lag; memory is sufficiently plentiful to allow for long buffers of recorded speech; and network speeds are swift enough to enable these applications on networked devices, even hundreds of miles distant. Since those initial experiments, the costs of processing power, memory, and network connectivity have only fallen, and the power of devices has risen accordingly.

The second of my research questions, “Does the technique of audio expansion present any fundamental barriers to communication?” now has at least a preliminary answer: No. Prior to the experiment detailed in Chapter Four, I am not aware of any studies on the effects of interpersonal communications which involved directly changing the rates of speech of one or more conversational partners. In informal discussions during the development of the prototype and proof-of-concept communication software, reactions were often mixed: Cautious optimism at the idea that slowing speech might yield intelligibility, comprehension, or other benefits, but also skepticism that this technique could be applied to true back-and-forth conversations rather than listen-only scenarios such as canned audio or video lectures. The most common objection related to lengthy pauses during a conversation where one conversation partner listened to another: Would these lengthy pauses be expanded as well, and would that not have the potential to completely desynchronize conversations? This is a powerful objection, and led directly to the decision to incorporate simple voice activation detection protocols to the experimental software, so that such pauses would not be expanded. Even so, the possibility remained the desynchronization might happen.

Subtler objections obtained as well, revolving around the idea of feedback effects: If a talker’s speech were expanded by his or her conversational partner, would this lead to a conscious or unconscious change in that talker’s naturally generated speech rate? Would it lead to an unconscious change in the listener’s subsequent speech when speaking roles reversed? Either of these scenarios presented difficulties: If speech rates sped up in the presence of expansion, this would constitute negative, stabilizing feedback which might cancel some or all of the expansion effects. If speech rates slowed down in sympathy with expansion, this could be potentially worse as positive feedback which might cause conversations to spiral out of control or break off entirely as one or both sides slowed their own naturally generated speech more and more.

It was of course not feasible to test all possible combination of audio expansion in the course of a single experiment. These objections, therefore, are still valid in general. However, this study demonstrates that, at least in the circumstances which it tested—multiple combination of a mild audio expansion—none of these ill effects were observed, and no statistically significant changes in speech patterns were detected. Moreover, while some subjects reported an awareness that their audio was being manipulated in some fashion, none was able to discern that the tempo of one or both sides of a conversation was being changed. This bodes well for the idea in general.

The third research question, however, "Is it useful or effective to place the control of speech rate directly into the hands of listeners?" remains not only open, but is now attached to new puzzling questions. All three of the studies which investigated this question (a Speech In Noise test with user-controlled linear expansions, a Speech in Noise test with user-controlled phoneme-based expansions, and a long passage speech comprehension test for ESL subjects) displayed the same general pattern: Subjects were instructed to use the technique in order to increase their ability to understand sentences in noise, or long passages of speech, without indicating how the technique should be used; subjects did generally use the technique rather than rejecting it; subjects reported a belief that the technique helped them accomplish their task; objective tests and scores demonstrated either the opposite (hindrance rather than help) or no statistical effect at all. While it is possible that this is the result of subjects' inferring from the survey questions what the investigators wanted to hear, the uniformity of that result across multiple studies conducted by multiple methods by multiple experimenters makes this less likely. (I conducted the first uniform expansion Speech in Noise experiment in person; my collaborator, Dan Bunn conducted the ESL comprehension study in person; I conducted the second phoneme-based Speech in Noise study entirely on-line through e-mails and websites.) In addition, the first Speech in Noise experiment was able to test subject preferences over a wide range of noise conditions, and there was clear statistical evidence that the subjects' revealed preference is to increase the amount of speech-slowing in direct proportion to the amount of noise present. This revealed preference is in accord with their stated survey preference of audio expansion. Finally, as noted in Chapter Seven, the phrasing of one of the survey question in the phoneme study ("*Compared to normal speech*, how easy or difficult was listening to the changed cadence speech?") shows a strongly bimodal distribution and, further, the subjects' answers to this survey question may predict (although not with statistical significance) how well they perform on the intelligibility tests. The combination of these results suggest to me that the subjects are deriving some as yet unknown benefit from audio expansion, but that this benefit is not what the studies are trying to measure.

8.3 Broader Applications and Implications

Despite the inability (thus far!) to realize intelligibility benefits from the techniques described in this dissertation, I believe the technology is still promising, with several avenues for important applications on the horizon. I also believe that these applications can be broadly divided into two classes: Those that mediate between a listener and real-time but one-way media, and those that mediate between one or more conversational partners. In general, while these applications may benefit from intelligibility benefits if such are demonstrated in the future, they do not rely on it. Rather, they rely on the perceived ease of listening, and/or the demonstrated preferential use of audio expansion in difficult situations.

Conventional radio: The first application in the one-way media application space are simple radio broadcasts. Conceptually, there is no difficulty adding a user-controlled audio expansion function to conventional radio devices, only the technological limitation that most dedicated radio devices are comprised of radio receivers and speakers, lacking the necessary computational hardware to convert received audio broadcasts to sampled streams; store or buffer enough of those streams to be meaningful, and perform the audio expansion algorithms on them. However, modern smartphones contain AM/FM radio receivers which applications can access if they have not been disabled by the vendors; further, there are applications such as SimpleRadio which act to stream live radio stations directly to the smartphone. Any of these applications could be easily modified to add audio expansion as a feature, because all audio manipulations are performed locally—the broadcaster need not know or care about the rate of received speech.

Conventional television: Likewise, there is no *fundamental* barrier to the implementation of audio expansion on a one-way live television platform; similarly to radio, the broadcasters simply do not need to know or care about what is done with their signal as long as it is not re-transmitted. However, there are more serious practical concerns with this application than with radio. Specifically, it is highly likely that the video rate would need to be expanded in the same degree as the audio in order to avoid mismatches both subtle (lips not moving in sync with spoken words) and unsubtle (words not spoken in the correct context of the visual scene). However, it is also not known whether or to what degree a viewer might tolerate slowed video in general. The technique is likely to be far more processor- and memory-intensive as well, due to the greater signal bandwidths involved.

Conventional telephony: Chapters Three and Four demonstrate the feasibility of adding audio expansion features at both a technical level and an interpersonal communication level. Although the study contained in Chapter Four obviously cannot rule in all possible combinations of audio expansion in two-way audio communication, it was able to rule in several combinations in the region of 1.4 expansion factors. However, the applications described and used in those chapters are genuinely *prototype*s, rather than mature, polished applications. However, it is at least in principle possible to include this functionality either in a telephone application itself (which might require the phone manufacturers to do so) or in a more polished, special purpose VOIP system. Regardless, the possibility is there.

Teleconferencing Applications: To the best of my knowledge, no teleconferencing or video chat applications include audio expansion as feature, although again, there is no conceptual bar to it—similarly to television applications, video time expansion is also necessary, here, but laptops and increasingly even smartphones possess powerful GPUs which should make this feasible. However, applying this feature to a multi-party teleconferencing environment opens up significant questions, both of interface design and of basic research into communication dynamics. In a two party conversation (whether audio-only or including video) there are two audio expansion factors to be manipulated and managed; in a fully general N-person conversation there are in principle N(N-1) audio expansion factors to manage. While this presents only a minor software engineering challenge, it presents a more significant challenge to our understanding of how audio expansion might change conversation dynamics.

Although these applications are somewhat narrow in scope, I believe that they can have significant impacts. The study described in [116] argues convincingly that one effect of cognitive load is a perceived "shrinkage of time" and that this includes auditory and linguistic perceptions (indeed, this effect was *measured* using auditory and linguistic perceptions.) Although none of the user studies in this dissertation pertain directly to cognitive load, this linkage between cognitive load and auditory temporal perception suggests that it may be beneficial to slow the speech coming from a radio during activities of high cognitive load. In a general sense, this may benefit drivers who listen talk radio and who wish to better absorb and retain the content they are listening to… preferably without diverting their attention from the task of driving. However, there may be specific situations of higher cognitive burden, where the ability to understand and retain something the first time it is said over a radio is important, such as first responders (while driving) or pilots (while listening to air traffic instructions.) These situations would need to be carefully managed and experimental evidence carefully gathered, but the benefits may be substantial.

A similar study also demonstrates that, for at least some L1/L2 language pairs, foreign languages are perceived as being spoken faster. Although most studies of this effect rely on subject rations, the study described in [117] is more objective and demonstrates this change of perception experimentally. This may partially explain the results of Chapter Six, where subjects employed audio expansion even without benefits of intelligibility; if a second language sounds fast, it is reasonable to slow that speech down given the opportunity. I believe this is important because, while these applications may not (yet!) be suitable for face to face interactions, they may improve the quality of life and quality of learning for second language learners. The radio and television applications may allow for easier consumption of media (especially, in the case of television applications, with closed caption translations) and increased passive practice of the new language. By contrast, the interactive telephone or two-party teleconference applications may allow for easier active practice, and greater confidence during conversations mediated with such applications.

Finally, and in a similar vein, these systems may provide significant quality of life benefits for the elderly. Although much of the literature on speech rate modification for elders involves speech rate *compression* rather than expansion to probe the effects of cognitive slowing in elderly listeners, there is evidence that such listeners' difficulty in tracking rapid speech includes both auditory [118] and cognitive [119] components. That being the case, it is possible that elders may benefit from any and all of the applications described above. Radio and television applications may result in less stressful leisure activities, but perhaps more importantly, real-time and interactive systems such as telephones or two-party video conferencing systems may provide elderly listeners with new avenues to communicate with friends and family. If these new avenues lessen the cognitive load, lessen the stress, or enhance these elders' confidence in their communications—and by doing so, also enhances the frequency or quantity of their engagement with their loved ones—this may profoundly improve the quality of life not only of the elders, but also of their friends and family.

8.4 Future Research Directions

The puzzling results of the three intelligibility studies detailed in the previous section (and the previous three chapters) suggest several promising research directions. These include directed research into the precise nature of the implied benefits that subjects may be deriving from audio expansion, as well as the possibility of sacrificing user choice and listener personalization in order to achieve concrete intelligibility benefits. I will briefly explore these avenues in the remainder of this chapter.

8.4.1 Other Potential Benefits

There are two major conjectures for hypothetical "other benefits" that subjects might be deriving from the speech in noise and ESL experiments, both springing from post-survey interview comments from subjects indicating that the voice of the expanded speech was sometimes easier to extract from the noise without being necessarily more intelligible. One mechanism by which this might be the case, for the particular experimental conditions I used—expanded speech in unexpanded four person babble noise—is that the expansion might be increasing the number and/or sizes of "glimpses" audible during the audio track. A glimpse is a spectrotemporal region where the local SNR is 3 dB or greater, typically during vowels or vocalized phonemes. The act of expanding the desired audio would naturally have the effect of increasing the spectrotemporal glimpse area, with overall effect of making the voice easier to extract from noise. However, somewhat counter-intuitively, expanding the four-person babble noise without expanding the voice signal might also dramatically change the pattern of glimpses by lengthening the naturally-occurring low volume portions of the noise. This might also enhance listeners' ability to extract the voice from the noise. A simple experiment could easily be designed to test the hypothesis that stretching the noise, the signal, or both might lead to an enhanced ability to hear the voice, without necessarily enhancing intelligibility.

The second conjecture is that through some as yet unknown mechanism, audio expansion is reducing the cognitive load without increasing the intelligibility. There are number of results in the literature tying speech rate and/or speech perception in noise to cognitive load, the most salient of which appears to be [120] which highlights the difficulty of measuring cognitive load using pupillometry as a proxy in speech in noise studies. This study argues for a complex interpretation of pupil dilation: that pupil dilation while listening to synthetic voices in noiseless conditions indicates increased or engaged attention; but that in noisy conditions, increased pupil dilation indicates increased listening effort, but varying with the quality of the synthetic voice. This suggest a *conceptually* simple experiment, testing, e.g., a grid of 4 noise conditions (8 dB, 6 dB, 4 dB, and 2 dB) at each of 4 expansion conditions ranging from unity (no dilation) to an expansion factor of 1.5 with modified speech in noise tests. Note that this experiment would be lengthy but, without the element of user choice, somewhat tractable—Speech in noise tests typically take five minutes or less, allowing for 16 tests to be run in slightly over an hour.

8.4.2 Improvement Without Personalization

It may also be the case that audio expansion is simply not the right tool for the emulation of clear speech, i.e. there are no "right choices" because the tool is not sophisticated enough. In this case, machine learning approaches such as convolutional neural networks may provide the sophistication necessary, in at least some circumstances. There is a long tradition of using neural networks to perform voice conversion (i.e., take as input an utterance spoken by Talker A, and provide as output a synthetic utterance in a convincing facsimile of Talker B's voice) using several variations on recurrent neural networks, including LSTMs, GRUs, and bidirectional LSTMs (see [121] for a recent overview). While these approaches have met with varying degrees of success, they all share a fundamental flaw: They do not convert the cadence of Talker A to the cadence of Talker B, which is precisely the element of interest here. Rather, they result in Talker B's voice uttering a sentence at exactly the original cadence of, and exactly the same length as Talker A, when functioning perfectly. However, one recent approach [1] does purport to correctly account for cadence, producing output sequences of unequal length. Neural network voice conversion studies (and Kameoka's work in particular) typically use training sets of approximately one thousand pairs of sentences such as the CMU Arctic database [1].

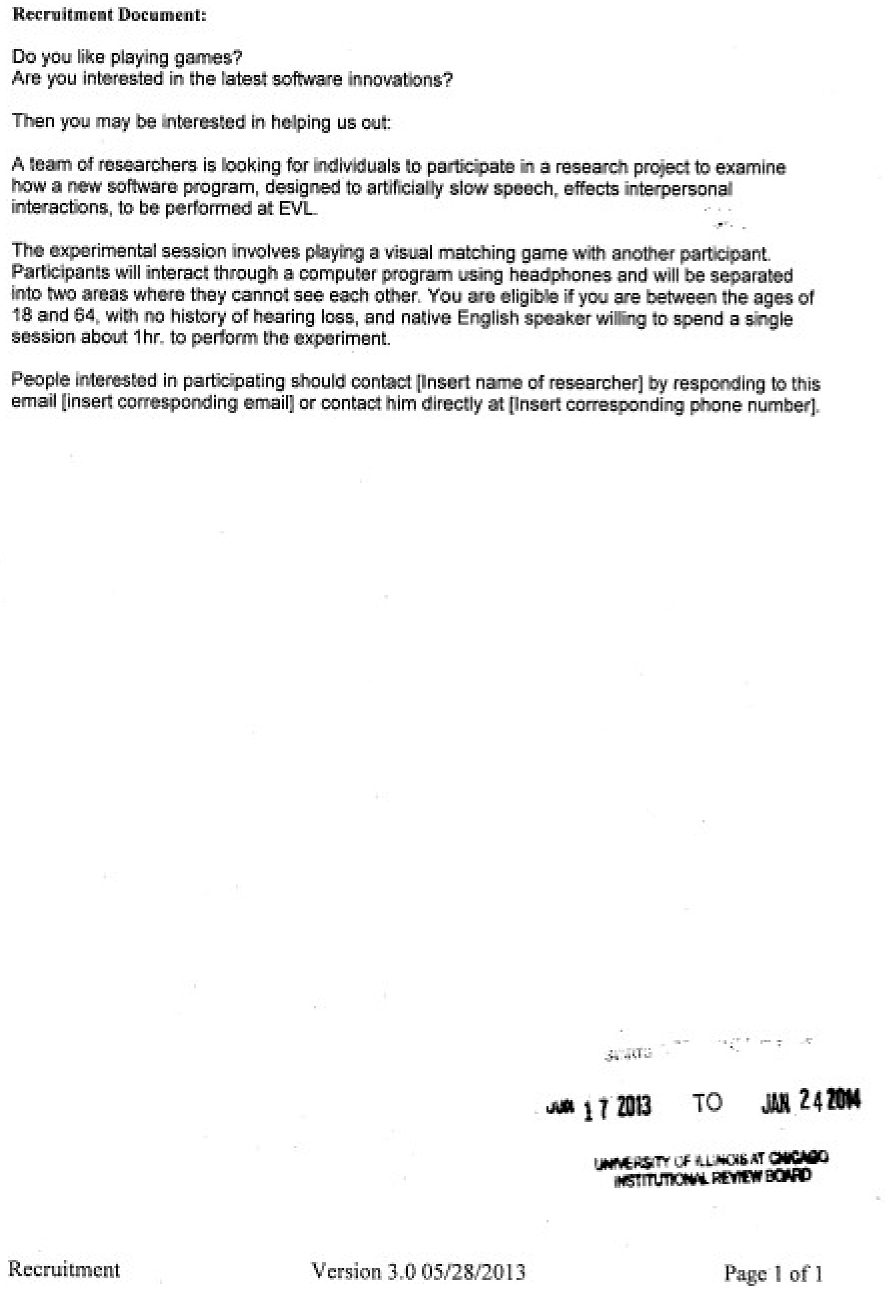
Here, in order to develop a proof of concept, one would need to develop a similarly sized set of paired sentences, spoken by the same person, but one with casual cadence and one with clear speech cadence, to be used as training set. This network would also be personalized to that particular talker, so any study materials (e.g., Harvard lists to enable studies similar to those in Chapters Five, Six, and Seven) must also be recorded in both conditions for future use. Further, the training set itself would need to be evaluated with several listeners to determine if the clear speech portion is, indeed, more intelligible under adverse conditions than its casual counterpart. Once that is done, the neural network could be developed and trained, and a relatively simple three part speech in noise test could be developed, testing unmodified casual speech, unmodified clear speech, and synthetic clear speech converted from casual speech. speech in noise tests, without the element of choice present in previous studies, take approximately five minutes each, meaning that participation in the study can likely be kept to thirty minutes or less per subject.

As noted, however, this experiment would only be a proof of concept, converting the speech of one talker (or at best a few talkers, if resources are sufficient to build multiple parallel corpora.) While it would preserve the vocal qualities of the talker, this approach would require the construction of parallel corpora—and extensive neural network training on those corpora-- for each talker, which is at this time a severe burden. However, if successful, this would be a milestone in the application of machine learning to hearing impairments.

APPENDICES

Appendix A Documentation, Diapix Experiment

Note: I was not the Principal Investigator and am not the custodian of the stamped copies of the recruitment and consent documents.



Recruitment Document:

Do you like playing games?
Are you interested in the latest software innovations?

Then you may be interested in helping us out:

A team of researchers is looking for individuals to participate in a research project to examine how a new software program, designed to artificially slow speech, effects interpersonal interactions, to be performed at EVL.

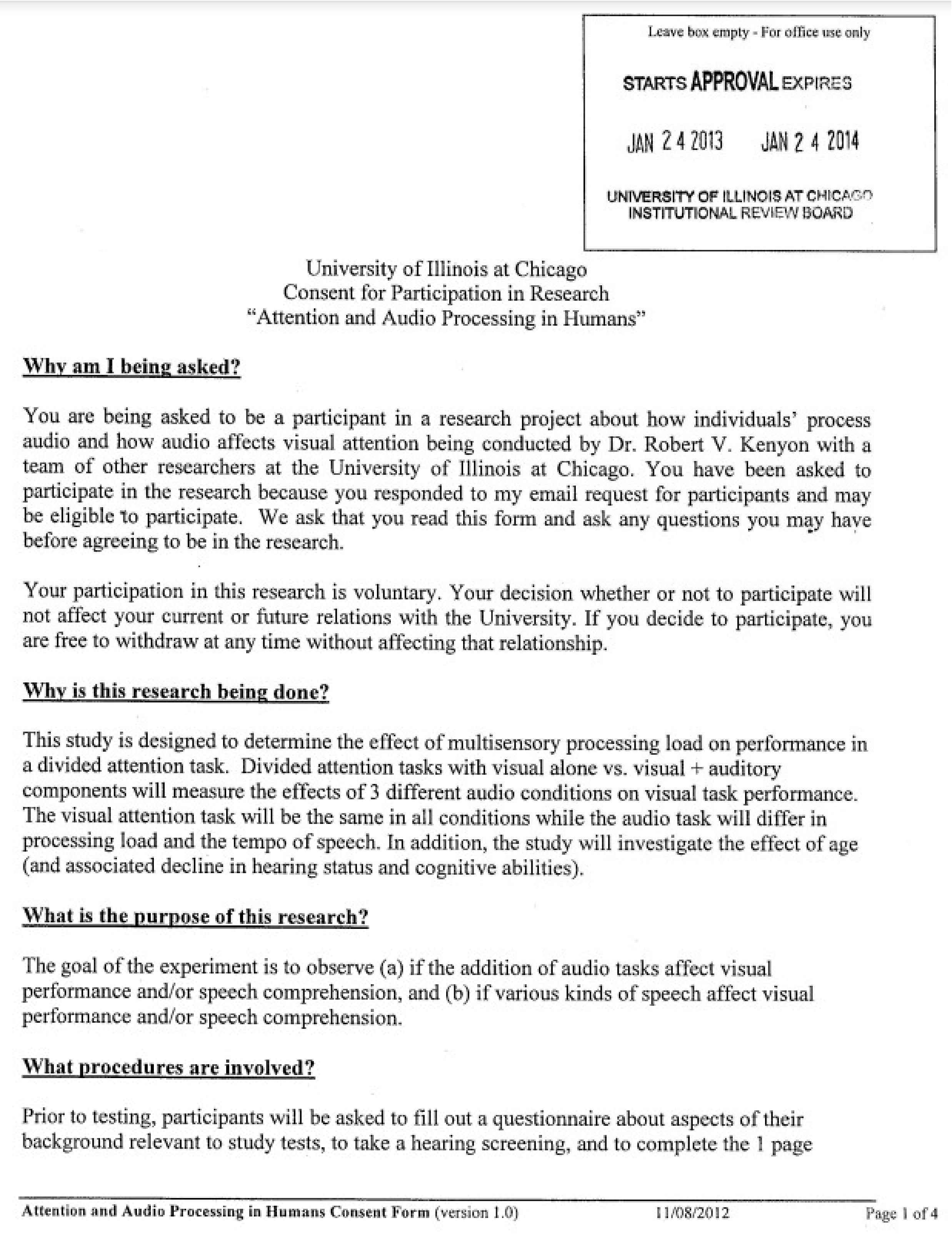
The experimental session involves playing a visual matching game with another participant. Participants will interact through a computer program using headphones and will be separated into two areas where they cannot see each other. You are eligible if you are between the ages of 18 and 64, with no history of hearing loss, and native English speaker willing to spend a single session about 1hr. to perform the experiment.

People interested in participating should contact [Insert name of researcher] by responding to this email [insert corresponding email] or contact him directly at [Insert corresponding phone number].

JAN 17 2013 TO JAN 24 2014

UNIVERSITY OF ILLINOIS AT CHICAGO
INSTITUTIONAL REVIEW BOARD

Recruitment Version 3.0 05/28/2013 Page 1 of 1



Leave box empty - For office use only

STARTS APPROVAL EXPIRES

JAN 2 4 2013 JAN 2 4 2014

UNIVERSITY OF ILLINOIS AT CHICAGO
INSTITUTIONAL REVIEW BOARD

University of Illinois at Chicago
Consent for Participation in Research
"Attention and Audio Processing in Humans"

Why am I being asked?

You are being asked to be a participant in a research project about how individuals' process audio and how audio affects visual attention being conducted by Dr. Robert V. Kenyon with a team of other researchers at the University of Illinois at Chicago. You have been asked to participate in the research because you responded to my email request for participants and may be eligible to participate. We ask that you read this form and ask any questions you may have before agreeing to be in the research.

Your participation in this research is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University. If you decide to participate, you are free to withdraw at any time without affecting that relationship.

Why is this research being done?

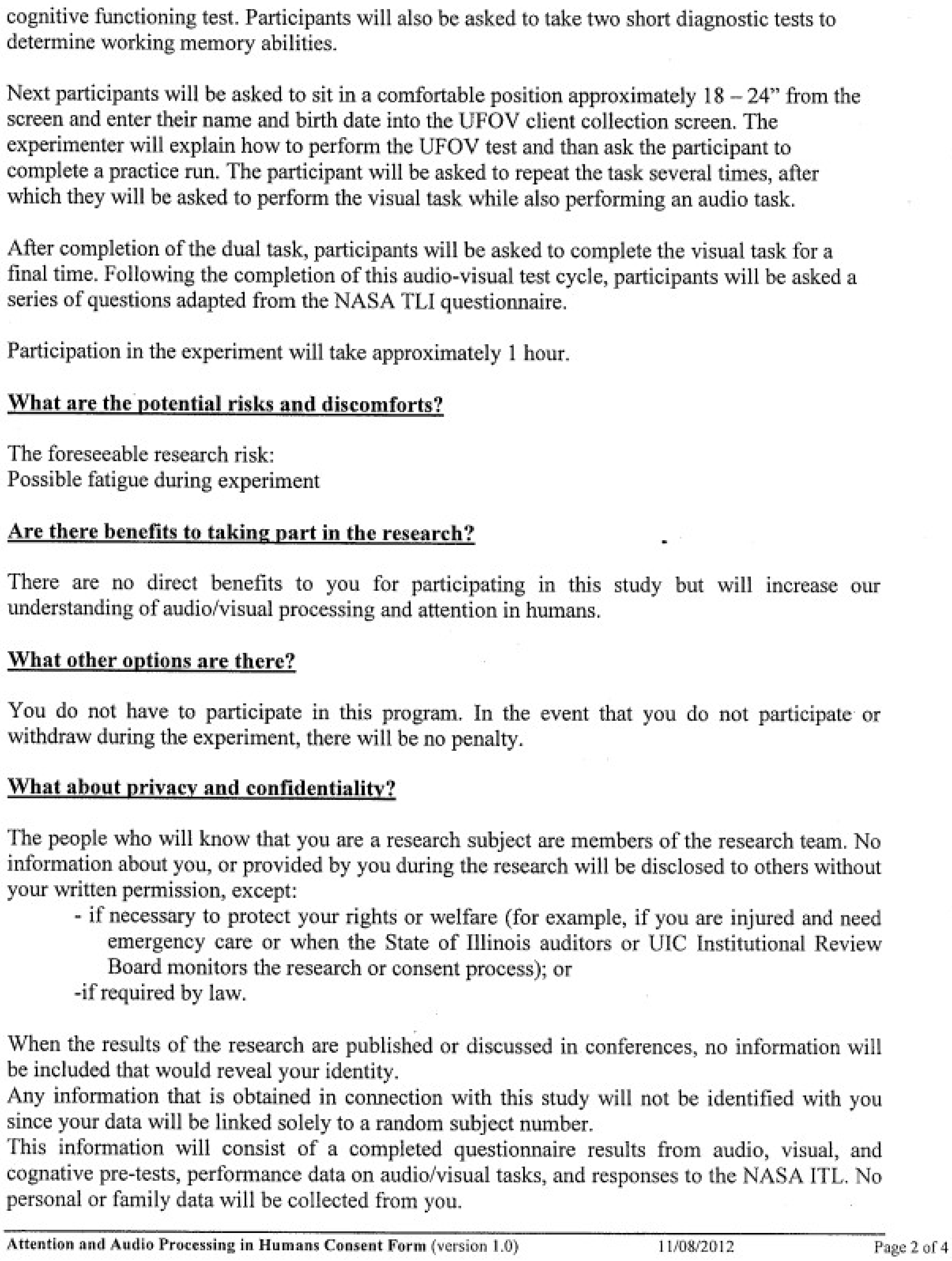
This study is designed to determine the effect of multisensory processing load on performance in a divided attention task. Divided attention tasks with visual alone vs. visual + auditory components will measure the effects of 3 different audio conditions on visual task performance. The visual attention task will be the same in all conditions while the audio task will differ in processing load and the tempo of speech. In addition, the study will investigate the effect of age (and associated decline in hearing status and cognitive abilities).

What is the purpose of this research?

The goal of the experiment is to observe (a) if the addition of audio tasks affect visual performance and/or speech comprehension, and (b) if various kinds of speech affect visual performance and/or speech comprehension.

What procedures are involved?

Prior to testing, participants will be asked to fill out a questionnaire about aspects of their background relevant to study tests, to take a hearing screening, and to complete the 1 page



cognitive functioning test. Participants will also be asked to take two short diagnostic tests to determine working memory abilities.

Next participants will be asked to sit in a comfortable position approximately 18 – 24" from the screen and enter their name and birth date into the UFOV client collection screen. The experimenter will explain how to perform the UFOV test and than ask the participant to complete a practice run. The participant will be asked to repeat the task several times, after which they will be asked to perform the visual task while also performing an audio task.

After completion of the dual task, participants will be asked to complete the visual task for a final time. Following the completion of this audio-visual test cycle, participants will be asked a series of questions adapted from the NASA TLI questionnaire.

Participation in the experiment will take approximately 1 hour.

What are the potential risks and discomforts?

The foreseeable research risk:
Possible fatigue during experiment

Are there benefits to taking part in the research?

There are no direct benefits to you for participating in this study but will increase our understanding of audio/visual processing and attention in humans.

What other options are there?

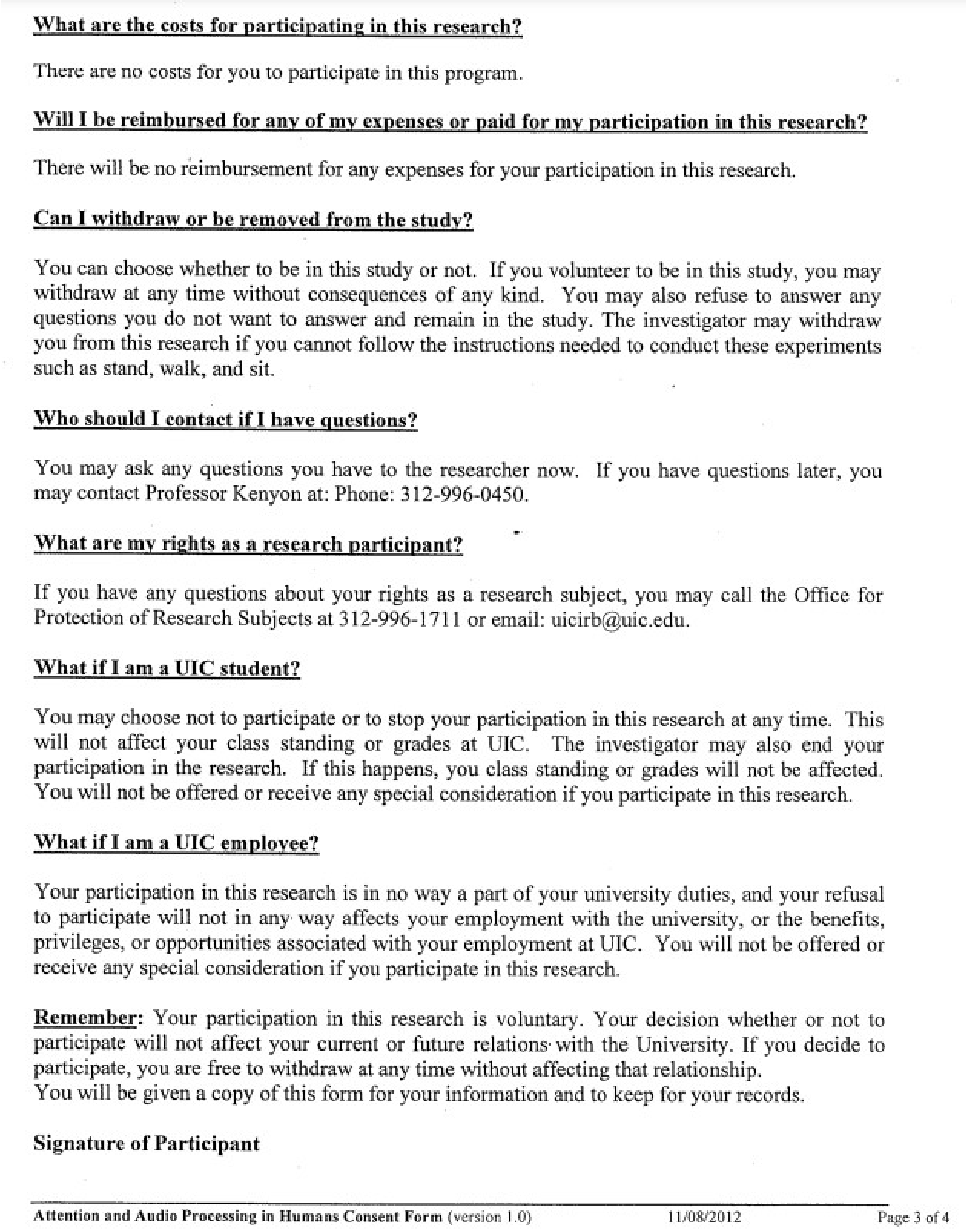
You do not have to participate in this program. In the event that you do not participate or withdraw during the experiment, there will be no penalty.

What about privacy and confidentiality?

The people who will know that you are a research subject are members of the research team. No information about you, or provided by you during the research will be disclosed to others without your written permission, except:

- if necessary to protect your rights or welfare (for example, if you are injured and need emergency care or when the State of Illinois auditors or UIC Institutional Review Board monitors the research or consent process); or
- if required by law.

When the results of the research are published or discussed in conferences, no information will be included that would reveal your identity.
Any information that is obtained in connection with this study will not be identified with you since your data will be linked solely to a random subject number.
This information will consist of a completed questionnaire results from audio, visual, and cognative pre-tests, performance data on audio/visual tasks, and responses to the NASA ITL. No personal or family data will be collected from you.



What are the costs for participating in this research?

There are no costs for you to participate in this program.

Will I be reimbursed for any of my expenses or paid for my participation in this research?

There will be no reimbursement for any expenses for your participation in this research.

Can I withdraw or be removed from the study?

You can choose whether to be in this study or not. If you volunteer to be in this study, you may withdraw at any time without consequences of any kind. You may also refuse to answer any questions you do not want to answer and remain in the study. The investigator may withdraw you from this research if you cannot follow the instructions needed to conduct these experiments such as stand, walk, and sit.

Who should I contact if I have questions?

You may ask any questions you have to the researcher now. If you have questions later, you may contact Professor Kenyon at: Phone: 312-996-0450.

What are my rights as a research participant?

If you have any questions about your rights as a research subject, you may call the Office for Protection of Research Subjects at 312-996-1711 or email: uicirb@uic.edu.

What if I am a UIC student?

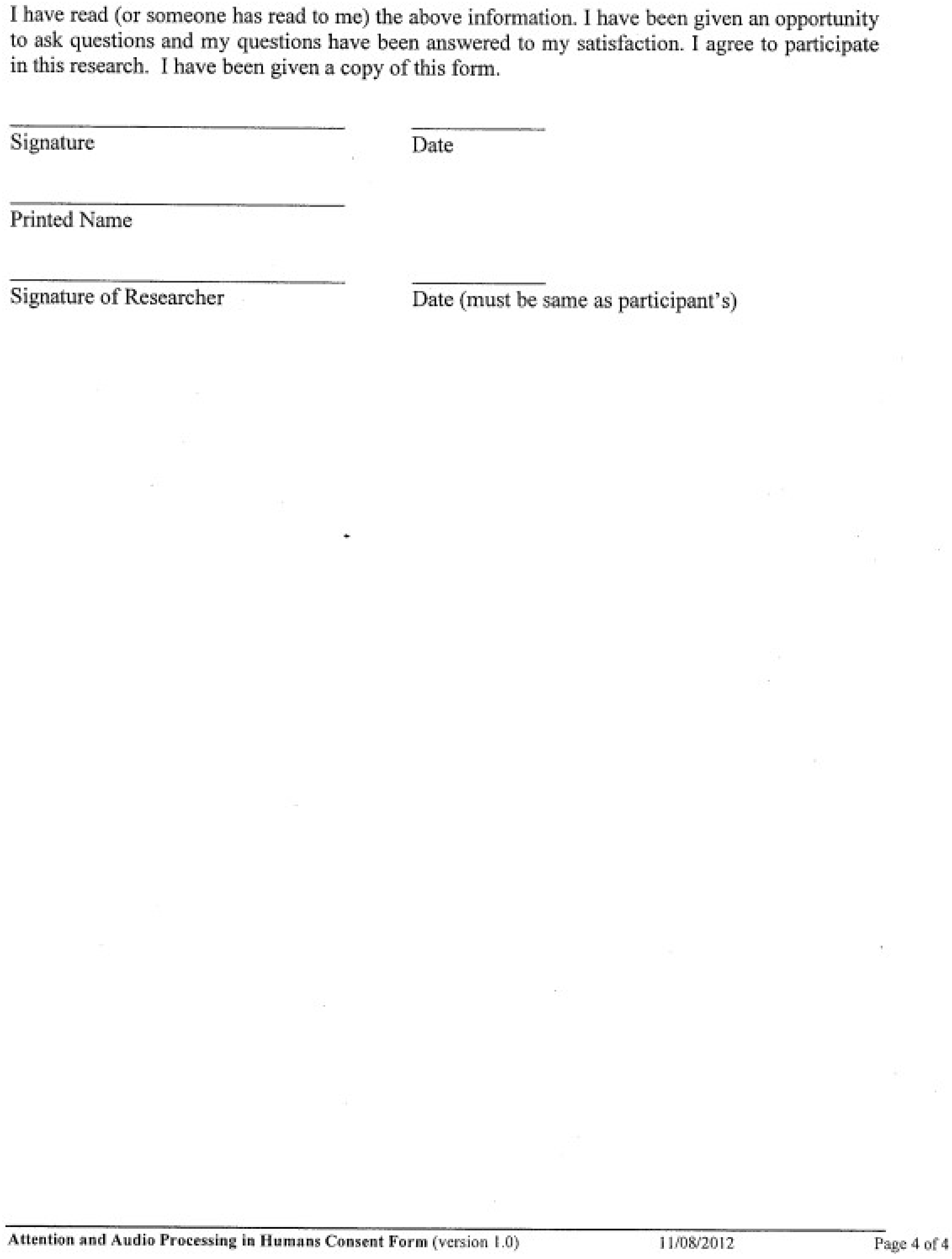
You may choose not to participate or to stop your participation in this research at any time. This will not affect your class standing or grades at UIC. The investigator may also end your participation in the research. If this happens, you class standing or grades will not be affected. You will not be offered or receive any special consideration if you participate in this research.

What if I am a UIC employee?

Your participation in this research is in no way a part of your university duties, and your refusal to participate will not in any way affects your employment with the university, or the benefits, privileges, or opportunities associated with your employment at UIC. You will not be offered or receive any special consideration if you participate in this research.

Remember: Your participation in this research is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University. If you decide to participate, you are free to withdraw at any time without affecting that relationship.
You will be given a copy of this form for your information and to keep for your records.

Signature of Participant



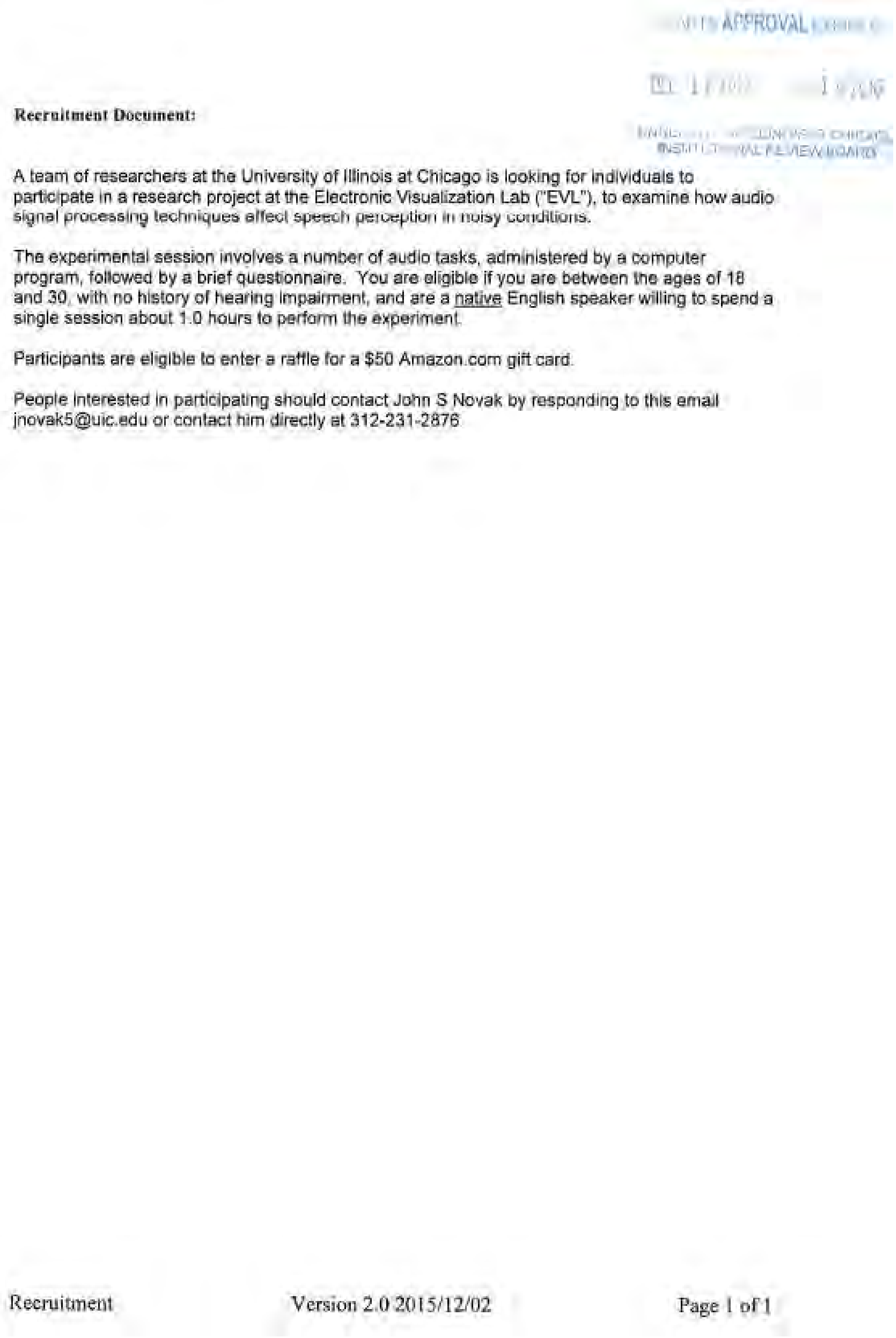
I have read (or someone has read to me) the above information. I have been given an opportunity to ask questions and my questions have been answered to my satisfaction. I agree to participate in this research. I have been given a copy of this form.

______________________________ ____________
Signature Date

____
Printed Name

______________________________ ____________
Signature of Researcher Date (must be same as participant's)

Appendix B Documentation, Speech In Noise Experiment



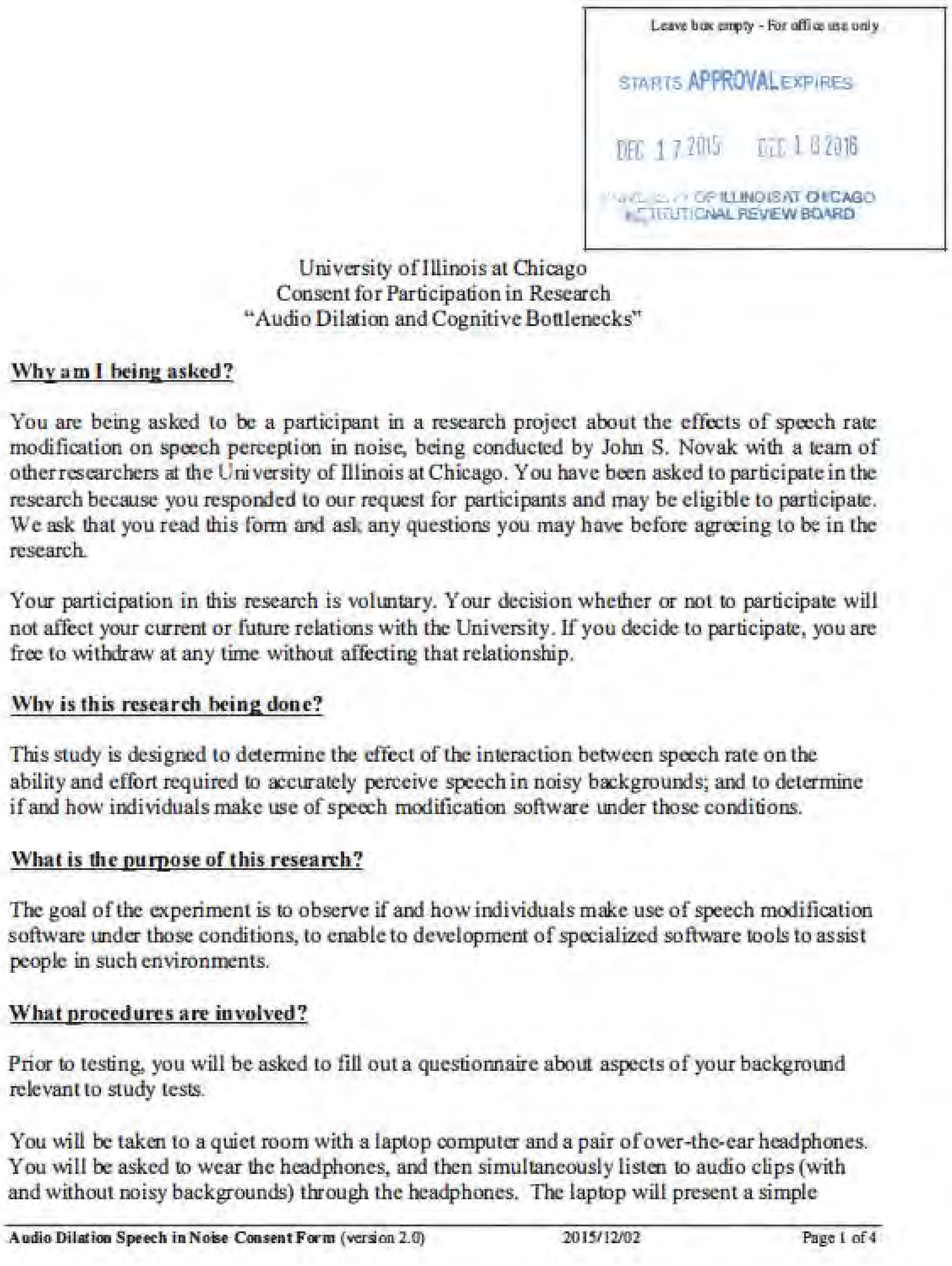
Recruitment Document:

A team of researchers at the University of Illinois at Chicago is looking for individuals to participate in a research project at the Electronic Visualization Lab ("EVL"), to examine how audio signal processing techniques affect speech perception in noisy conditions.

The experimental session involves a number of audio tasks, administered by a computer program, followed by a brief questionnaire. You are eligible if you are between the ages of 18 and 30, with no history of hearing impairment, and are a native English speaker willing to spend a single session about 1.0 hours to perform the experiment.

Participants are eligible to enter a raffle for a $50 Amazon.com gift card.

People interested in participating should contact John S Novak by responding to this email jnovak5@uic.edu or contact him directly at 312-231-2876



Leave box empty - For office use only

STARTS APPROVAL EXPIRES

DEC 17 2015 DEC 10 2016

UNIVERSITY OF ILLINOIS AT CHICAGO
INSTITUTIONAL REVIEW BOARD

University of Illinois at Chicago
Consent for Participation in Research
"Audio Dilation and Cognitive Bottlenecks"

Why am I being asked?

You are being asked to be a participant in a research project about the effects of speech rate modification on speech perception in noise, being conducted by John S. Novak with a team of other researchers at the University of Illinois at Chicago. You have been asked to participate in the research because you responded to our request for participants and may be eligible to participate. We ask that you read this form and ask any questions you may have before agreeing to be in the research.

Your participation in this research is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University. If you decide to participate, you are free to withdraw at any time without affecting that relationship.

Why is this research being done?

This study is designed to determine the effect of the interaction between speech rate on the ability and effort required to accurately perceive speech in noisy backgrounds; and to determine if and how individuals make use of speech modification software under those conditions.

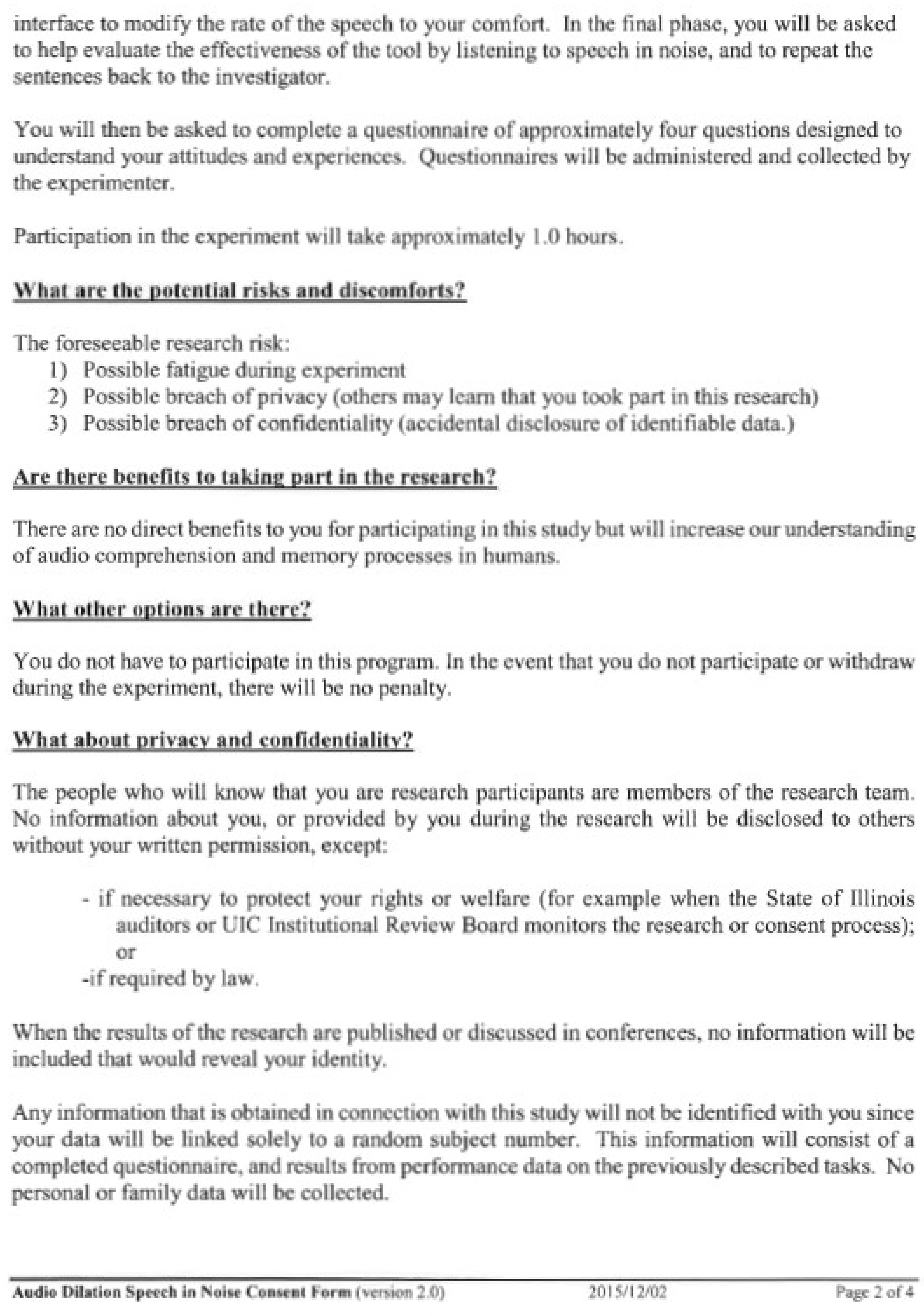
What is the purpose of this research?

The goal of the experiment is to observe if and how individuals make use of speech modification software under those conditions, to enable to development of specialized software tools to assist people in such environments.

What procedures are involved?

Prior to testing, you will be asked to fill out a questionnaire about aspects of your background relevant to study tests.

You will be taken to a quiet room with a laptop computer and a pair of over-the-ear headphones. You will be asked to wear the headphones, and then simultaneously listen to audio clips (with and without noisy backgrounds) through the headphones. The laptop will present a simple



interface to modify the rate of the speech to your comfort. In the final phase, you will be asked to help evaluate the effectiveness of the tool by listening to speech in noise, and to repeat the sentences back to the investigator.

You will then be asked to complete a questionnaire of approximately four questions designed to understand your attitudes and experiences. Questionnaires will be administered and collected by the experimenter.

Participation in the experiment will take approximately 1.0 hours.

What are the potential risks and discomforts?

The foreseeable research risk:

1) Possible fatigue during experiment
2) Possible breach of privacy (others may learn that you took part in this research)
3) Possible breach of confidentiality (accidental disclosure of identifiable data.)

Are there benefits to taking part in the research?

There are no direct benefits to you for participating in this study but will increase our understanding of audio comprehension and memory processes in humans.

What other options are there?

You do not have to participate in this program. In the event that you do not participate or withdraw during the experiment, there will be no penalty.

What about privacy and confidentiality?

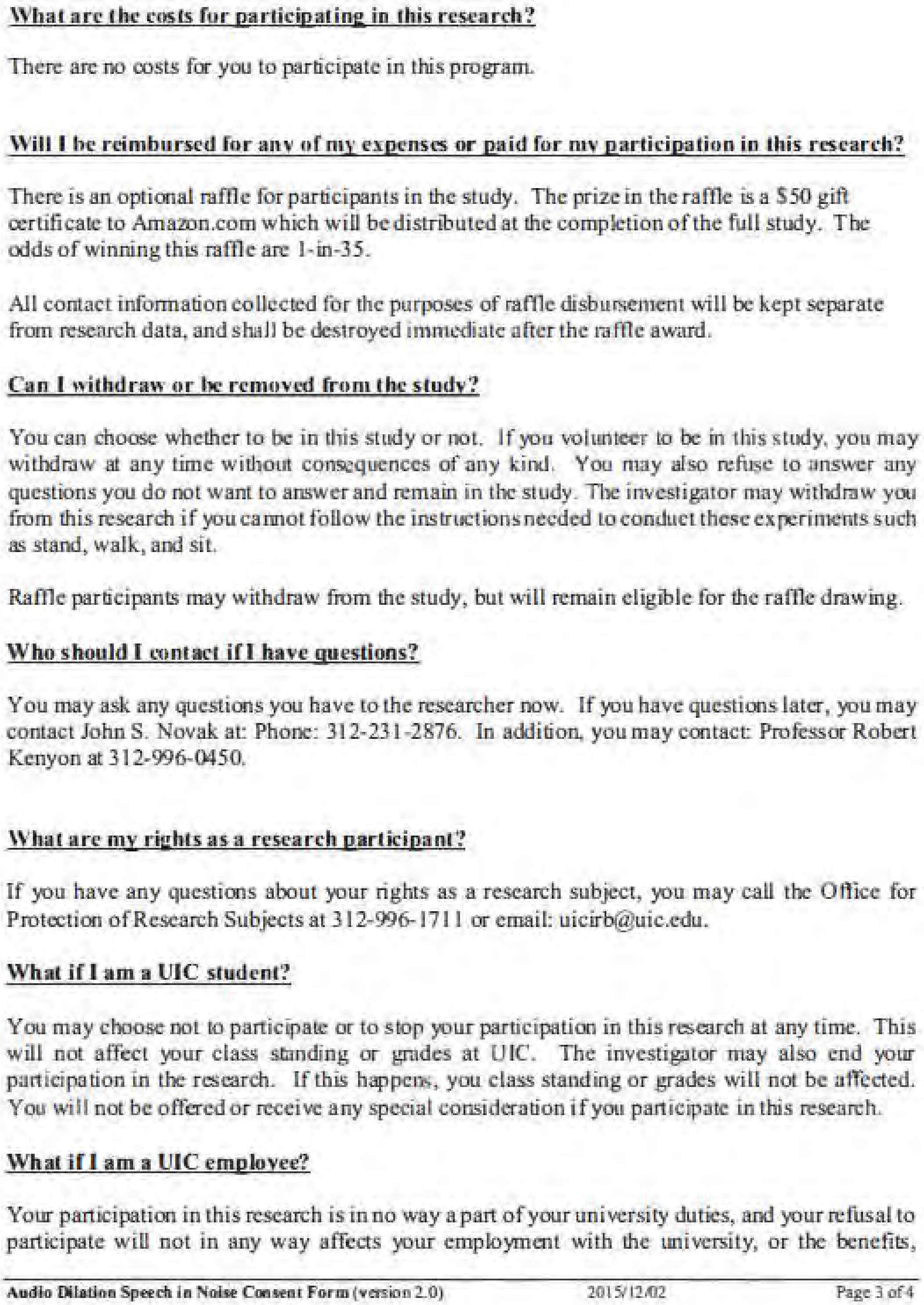
The people who will know that you are research participants are members of the research team. No information about you, or provided by you during the research will be disclosed to others without your written permission, except:

- if necessary to protect your rights or welfare (for example when the State of Illinois auditors or UIC Institutional Review Board monitors the research or consent process); or

-if required by law.

When the results of the research are published or discussed in conferences, no information will be included that would reveal your identity.

Any information that is obtained in connection with this study will not be identified with you since your data will be linked solely to a random subject number. This information will consist of a completed questionnaire, and results from performance data on the previously described tasks. No personal or family data will be collected.



What are the costs for participating in this research?

There are no costs for you to participate in this program.

Will I be reimbursed for any of my expenses or paid for my participation in this research?

There is an optional raffle for participants in the study. The prize in the raffle is a $50 gift certificate to Amazon.com which will be distributed at the completion of the full study. The odds of winning this raffle are 1-in-35.

All contact information collected for the purposes of raffle disbursement will be kept separate from research data, and shall be destroyed immediate after the raffle award.

Can I withdraw or be removed from the study?

You can choose whether to be in this study or not. If you volunteer to be in this study, you may withdraw at any time without consequences of any kind. You may also refuse to answer any questions you do not want to answer and remain in the study. The investigator may withdraw you from this research if you cannot follow the instructions needed to conduct these experiments such as stand, walk, and sit.

Raffle participants may withdraw from the study, but will remain eligible for the raffle drawing.

Who should I contact if I have questions?

You may ask any questions you have to the researcher now. If you have questions later, you may contact John S. Novak at: Phone: 312-231-2876. In addition, you may contact: Professor Robert Kenyon at 312-996-0450.

What are my rights as a research participant?

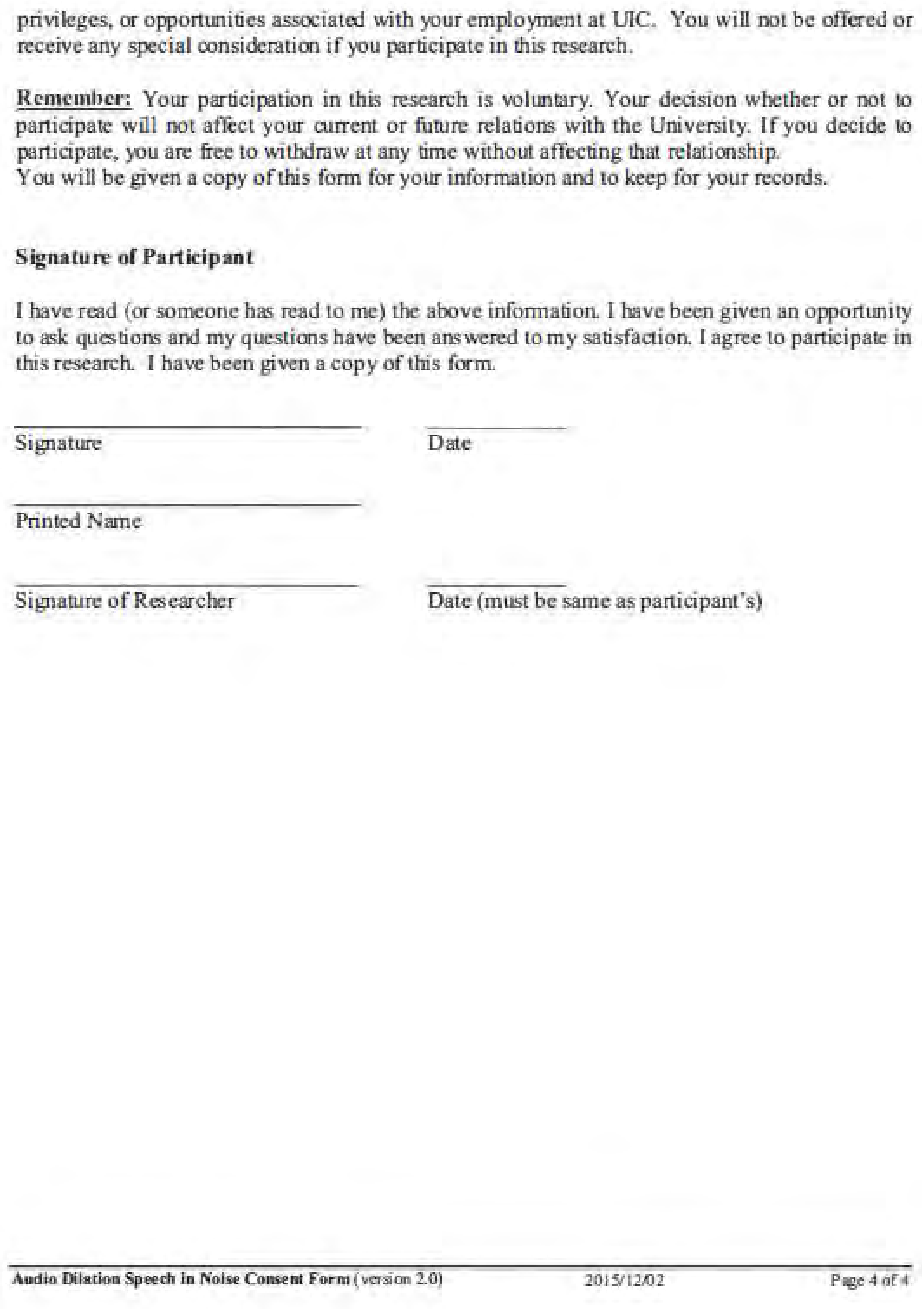
If you have any questions about your rights as a research subject, you may call the Office for Protection of Research Subjects at 312-996-1711 or email: uicirb@uic.edu.

What if I am a UIC student?

You may choose not to participate or to stop your participation in this research at any time. This will not affect your class standing or grades at UIC. The investigator may also end your participation in the research. If this happens, you class standing or grades will not be affected. You will not be offered or receive any special consideration if you participate in this research.

What if I am a UIC employee?

Your participation in this research is in no way a part of your university duties, and your refusal to participate will not in any way affects your employment with the university, or the benefits,



privileges, or opportunities associated with your employment at UIC. You will not be offered or receive any special consideration if you participate in this research.

Remember: Your participation in this research is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University. If you decide to participate, you are free to withdraw at any time without affecting that relationship.
You will be given a copy of this form for your information and to keep for your records.

Signature of Participant

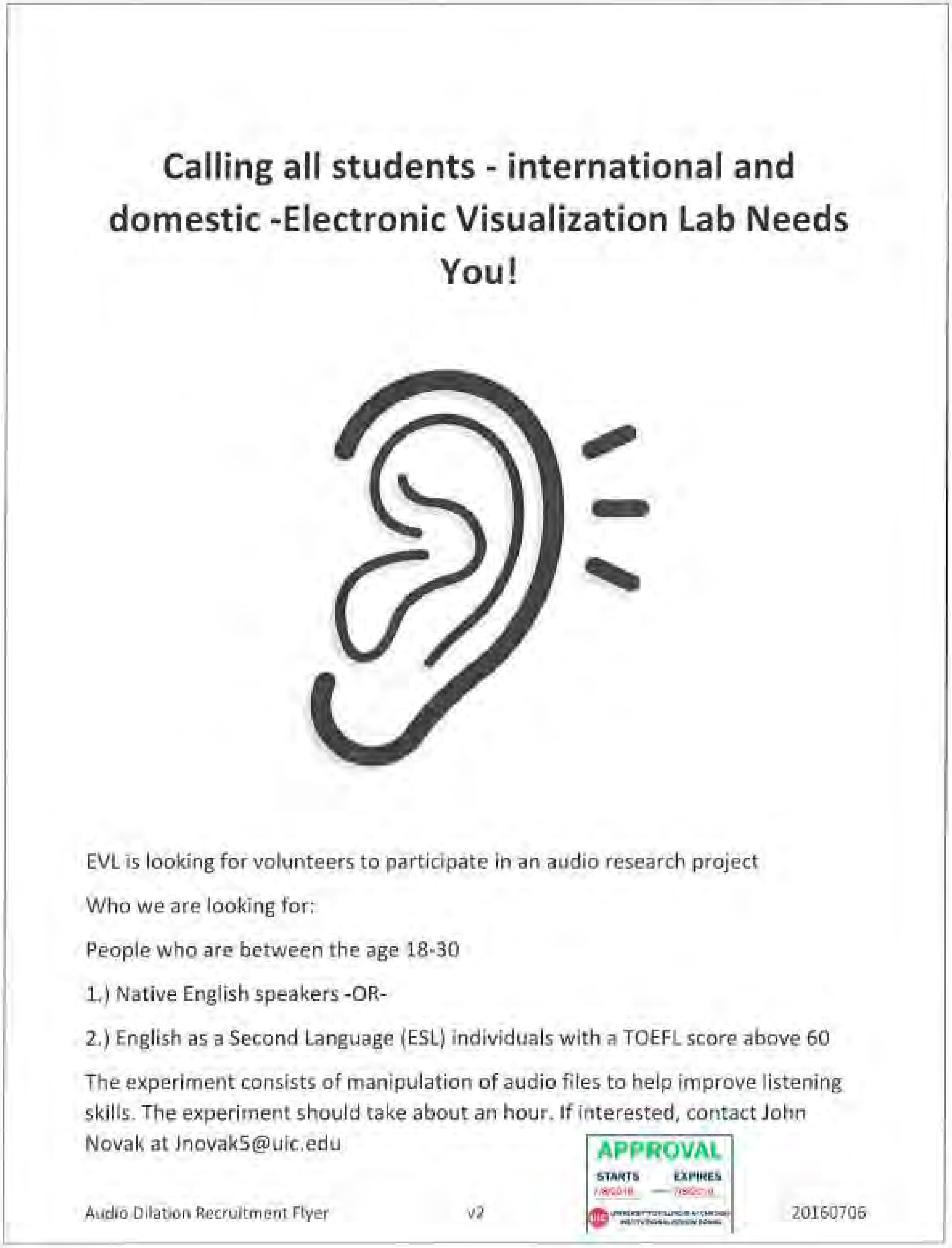
I have read (or someone has read to me) the above information. I have been given an opportunity to ask questions and my questions have been answered to my satisfaction. I agree to participate in this research. I have been given a copy of this form.

______________________________ ____________
Signature Date

____
Printed Name

______________________________ ____________
Signature of Researcher Date (must be same as participant's)

Appendix C Documentation, English As A Second Language Experiment



Calling all students - international and domestic -Electronic Visualization Lab Needs You!

EVL is looking for volunteers to participate in an audio research project

Who we are looking for:

People who are between the age 18-30

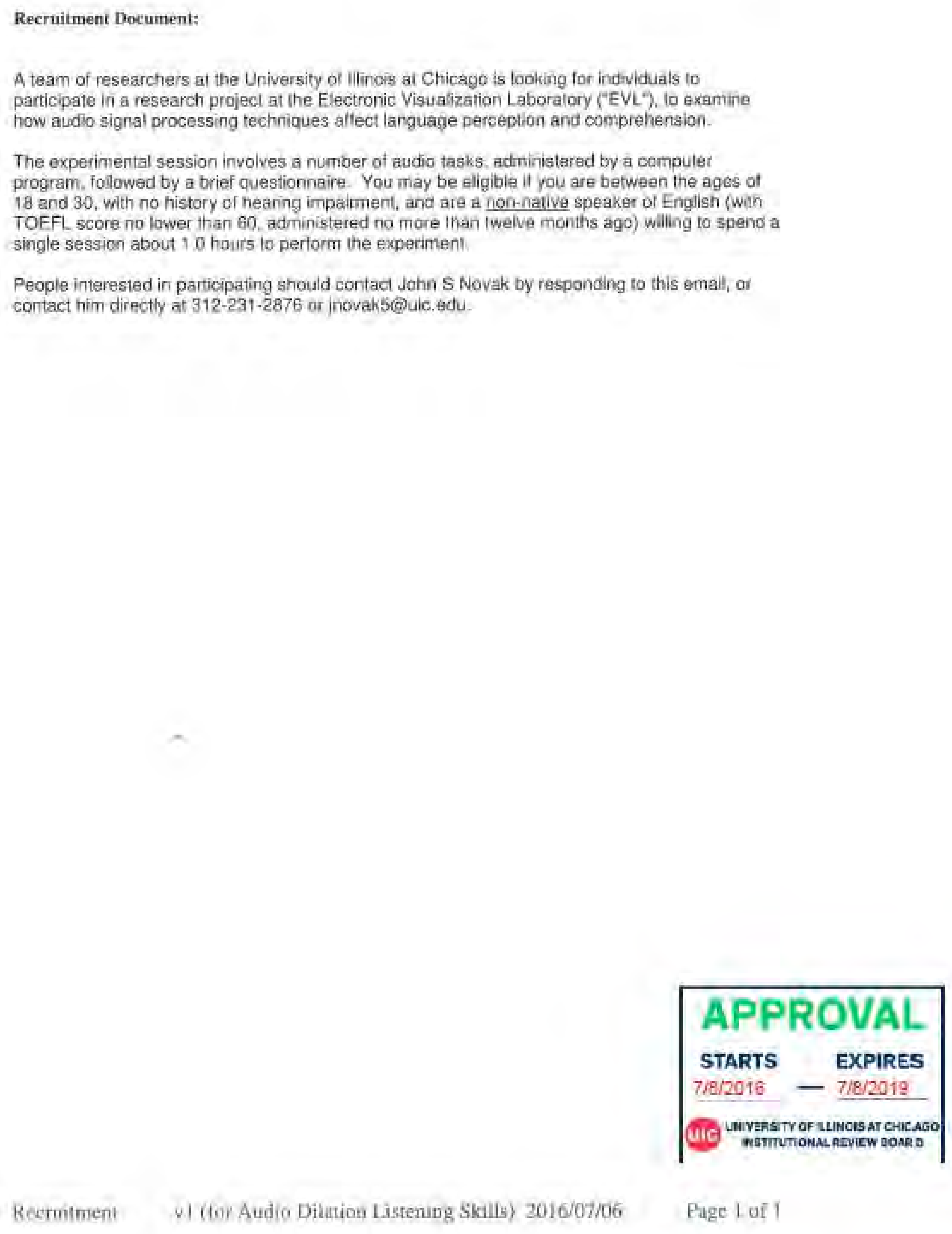
1.) Native English speakers -OR-

2.) English as a Second Language (ESL) individuals with a TOEFL score above 60

The experiment consists of manipulation of audio files to help improve listening skills. The experiment should take about an hour. If interested, contact John Novak at Jnovak5@uic.edu

APPROVAL
STARTS [illegible] — EXPIRES [illegible]
UNIVERSITY OF ILLINOIS AT CHICAGO INSTITUTIONAL REVIEW BOARD

Audio Dilation Recruitment Flyer v2 20160706



Recruitment Document:

A team of researchers at the University of Illinois at Chicago is looking for individuals to participate in a research project at the Electronic Visualization Laboratory ("EVL"), to examine how audio signal processing techniques affect language perception and comprehension.

The experimental session involves a number of audio tasks, administered by a computer program, followed by a brief questionnaire. You may be eligible if you are between the ages of 18 and 30, with no history of hearing impairment, and are a non-native speaker of English (with TOEFL score no lower than 60, administered no more than twelve months ago) willing to spend a single session about 1.0 hours to perform the experiment.

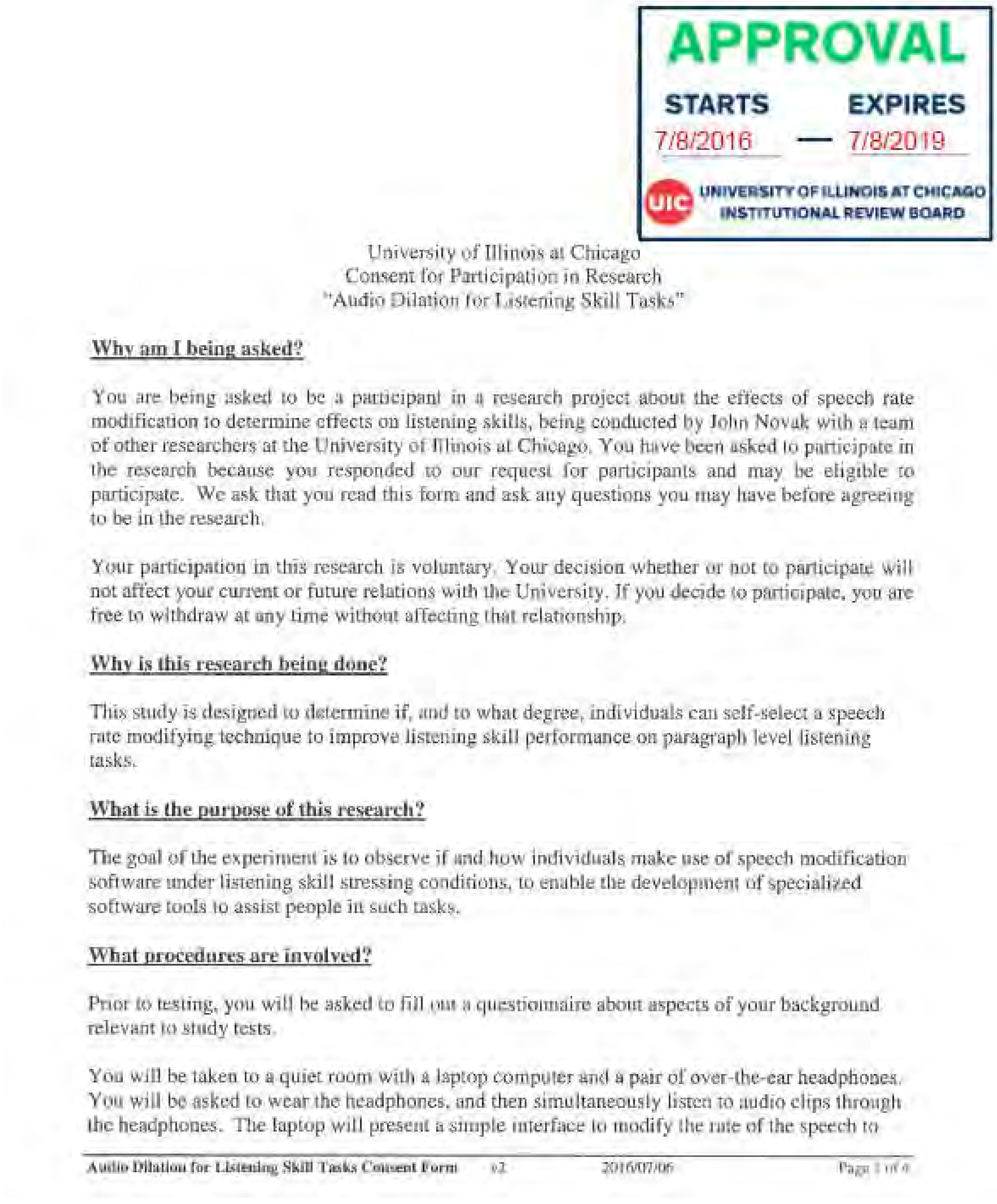
People interested in participating should contact John S Novak by responding to this email, or contact him directly at 312-231-2876 or jnovak5@uic.edu.

APPROVAL

STARTS 7/8/2016 — EXPIRES 7/8/2019

UIC UNIVERSITY OF ILLINOIS AT CHICAGO INSTITUTIONAL REVIEW BOARD

Recruitment v1 (for Audio Dilation Listening Skills) 2016/07/06 Page 1 of 1



APPROVAL

STARTS 7/8/2016 — EXPIRES 7/8/2019

UIC UNIVERSITY OF ILLINOIS AT CHICAGO INSTITUTIONAL REVIEW BOARD

University of Illinois at Chicago
Consent for Participation in Research
"Audio Dilation for Listening Skill Tasks"

Why am I being asked?

You are being asked to be a participant in a research project about the effects of speech rate modification to determine effects on listening skills, being conducted by John Novak with a team of other researchers at the University of Illinois at Chicago. You have been asked to participate in the research because you responded to our request for participants and may be eligible to participate. We ask that you read this form and ask any questions you may have before agreeing to be in the research.

Your participation in this research is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University. If you decide to participate, you are free to withdraw at any time without affecting that relationship.

Why is this research being done?

This study is designed to determine if, and to what degree, individuals can self-select a speech rate modifying technique to improve listening skill performance on paragraph level listening tasks.

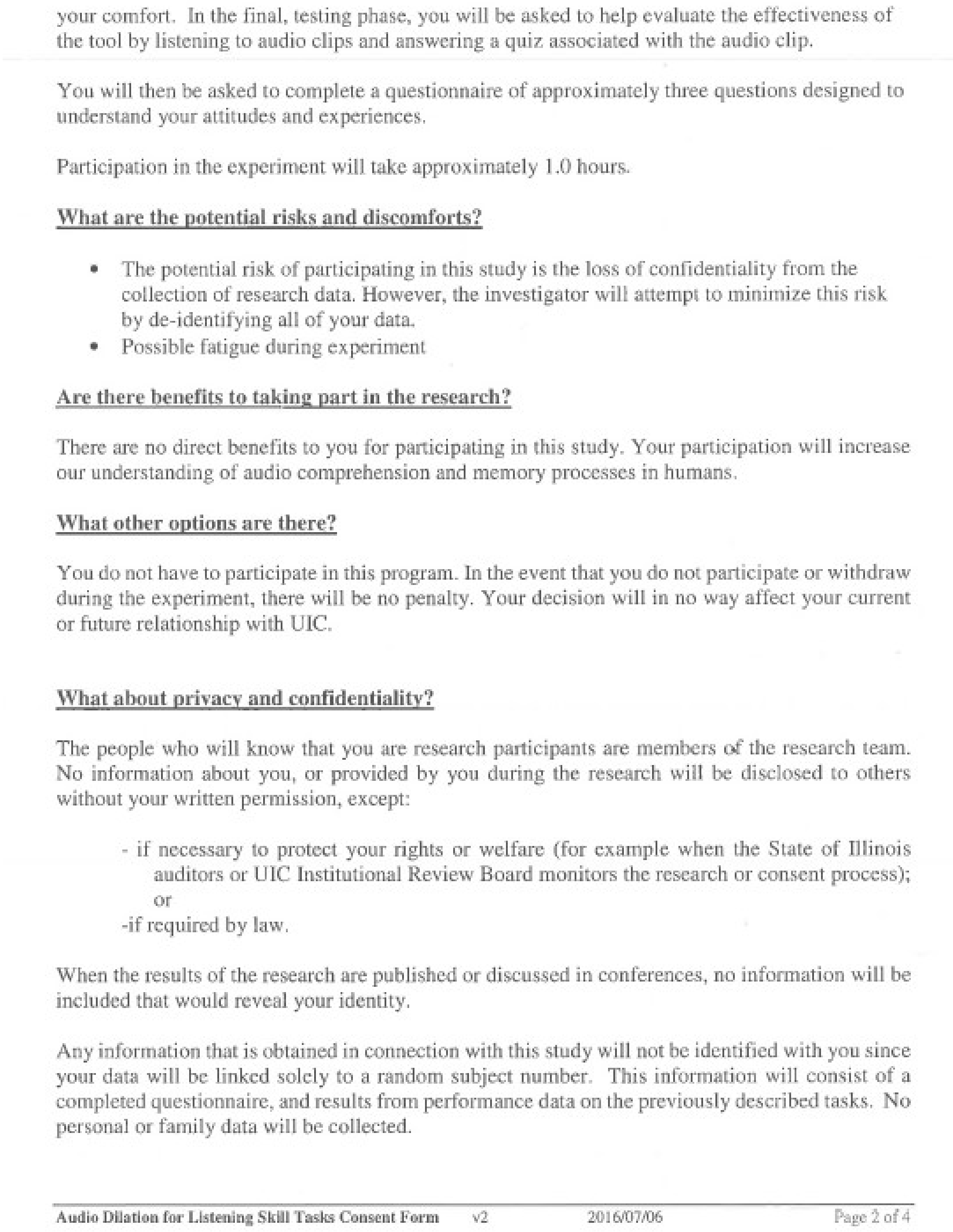
What is the purpose of this research?

The goal of the experiment is to observe if and how individuals make use of speech modification software under listening skill stressing conditions, to enable the development of specialized software tools to assist people in such tasks.

What procedures are involved?

Prior to testing, you will be asked to fill out a questionnaire about aspects of your background relevant to study tests.

You will be taken to a quiet room with a laptop computer and a pair of over-the-ear headphones. You will be asked to wear the headphones, and then simultaneously listen to audio clips through the headphones. The laptop will present a simple interface to modify the rate of the speech to



your comfort. In the final, testing phase, you will be asked to help evaluate the effectiveness of the tool by listening to audio clips and answering a quiz associated with the audio clip.

You will then be asked to complete a questionnaire of approximately three questions designed to understand your attitudes and experiences.

Participation in the experiment will take approximately 1.0 hours.

What are the potential risks and discomforts?

- The potential risk of participating in this study is the loss of confidentiality from the collection of research data. However, the investigator will attempt to minimize this risk by de-identifying all of your data.
- Possible fatigue during experiment

Are there benefits to taking part in the research?

There are no direct benefits to you for participating in this study. Your participation will increase our understanding of audio comprehension and memory processes in humans.

What other options are there?

You do not have to participate in this program. In the event that you do not participate or withdraw during the experiment, there will be no penalty. Your decision will in no way affect your current or future relationship with UIC.

What about privacy and confidentiality?

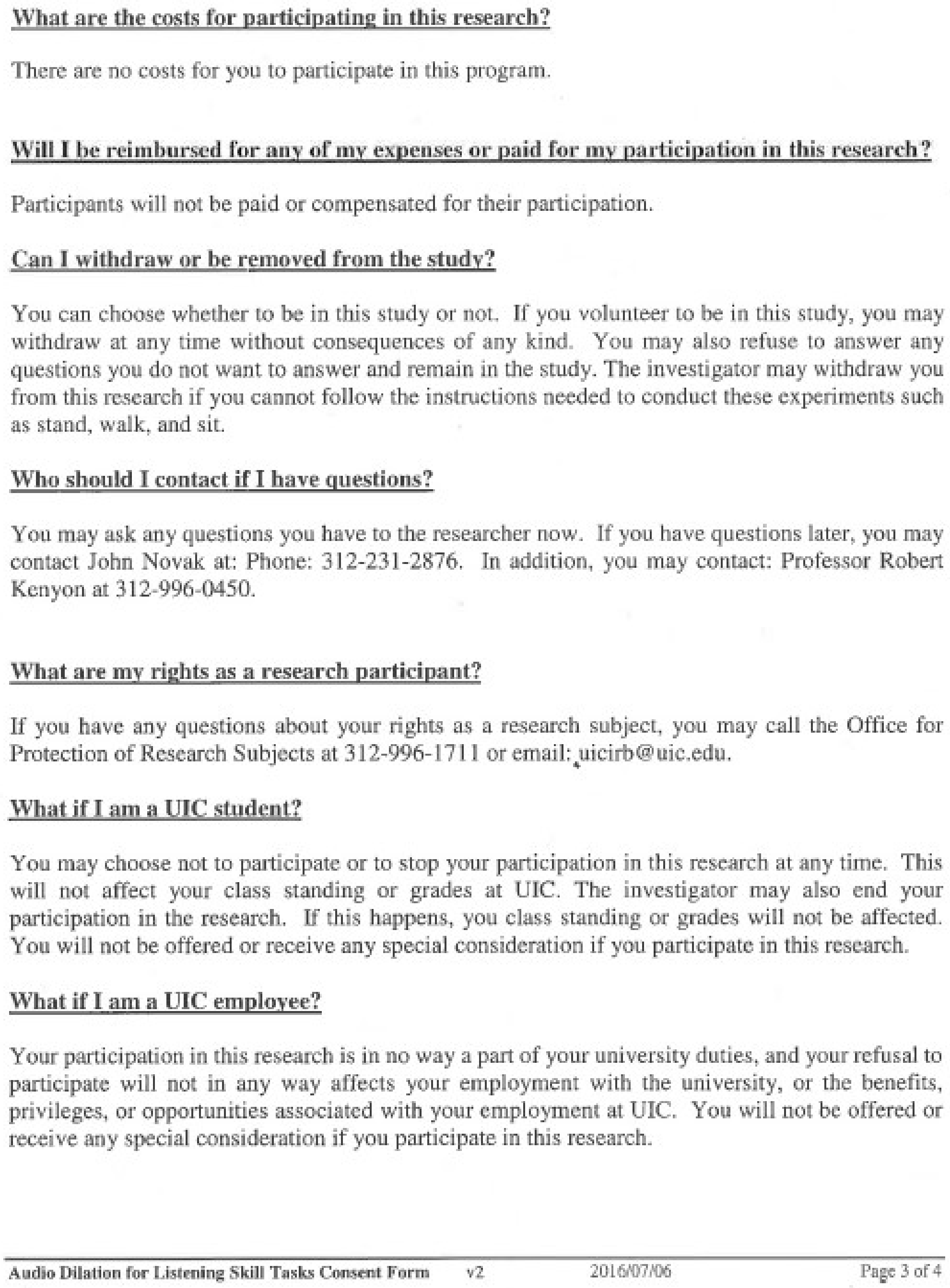
The people who will know that you are research participants are members of the research team. No information about you, or provided by you during the research will be disclosed to others without your written permission, except:

- if necessary to protect your rights or welfare (for example when the State of Illinois auditors or UIC Institutional Review Board monitors the research or consent process); or

-if required by law.

When the results of the research are published or discussed in conferences, no information will be included that would reveal your identity.

Any information that is obtained in connection with this study will not be identified with you since your data will be linked solely to a random subject number. This information will consist of a completed questionnaire, and results from performance data on the previously described tasks. No personal or family data will be collected.



What are the costs for participating in this research?

There are no costs for you to participate in this program.

Will I be reimbursed for any of my expenses or paid for my participation in this research?

Participants will not be paid or compensated for their participation.

Can I withdraw or be removed from the study?

You can choose whether to be in this study or not. If you volunteer to be in this study, you may withdraw at any time without consequences of any kind. You may also refuse to answer any questions you do not want to answer and remain in the study. The investigator may withdraw you from this research if you cannot follow the instructions needed to conduct these experiments such as stand, walk, and sit.

Who should I contact if I have questions?

You may ask any questions you have to the researcher now. If you have questions later, you may contact John Novak at: Phone: 312-231-2876. In addition, you may contact: Professor Robert Kenyon at 312-996-0450.

What are my rights as a research participant?

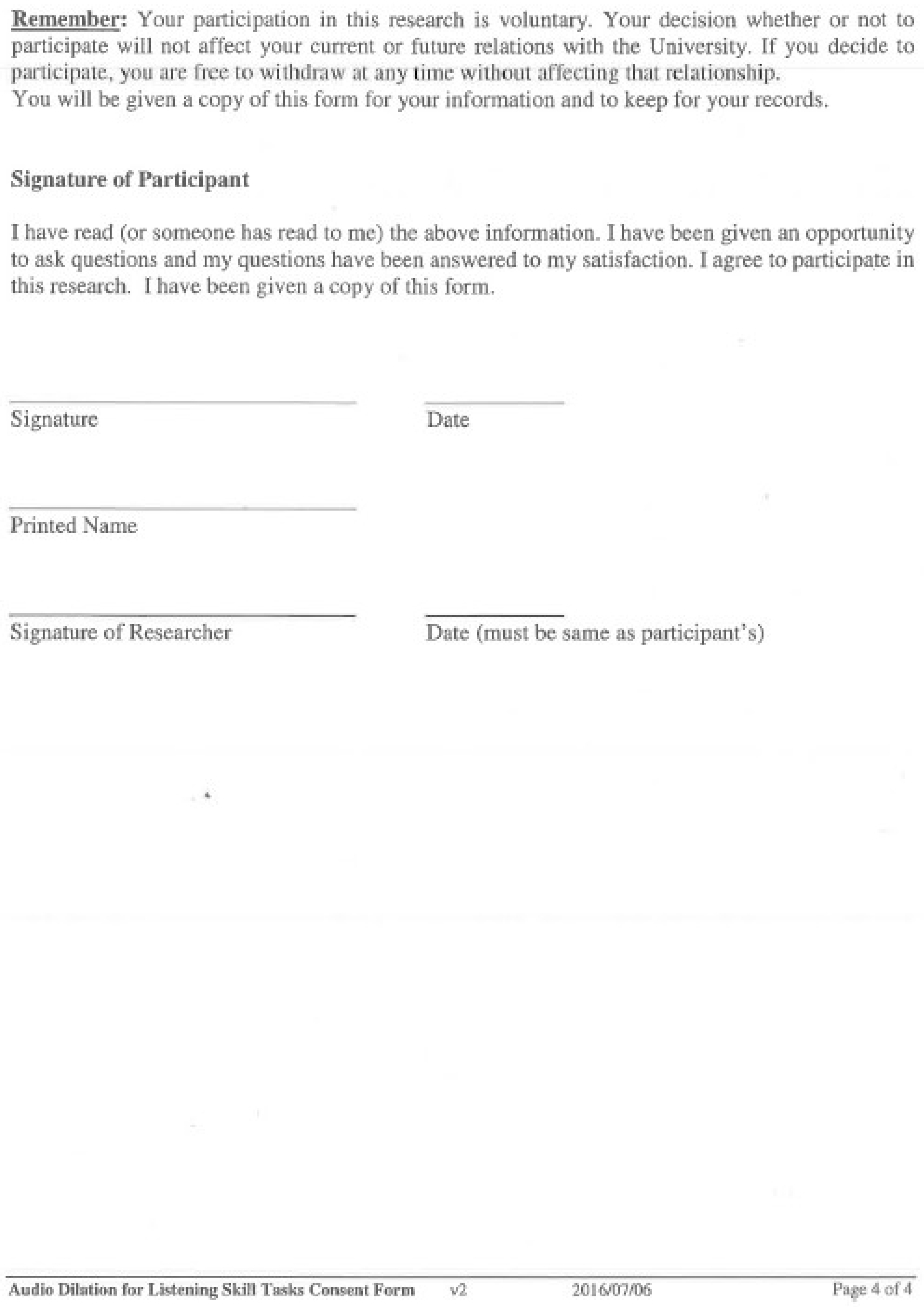
If you have any questions about your rights as a research subject, you may call the Office for Protection of Research Subjects at 312-996-1711 or email: uicirb@uic.edu.

What if I am a UIC student?

You may choose not to participate or to stop your participation in this research at any time. This will not affect your class standing or grades at UIC. The investigator may also end your participation in the research. If this happens, you class standing or grades will not be affected. You will not be offered or receive any special consideration if you participate in this research.

What if I am a UIC employee?

Your participation in this research is in no way a part of your university duties, and your refusal to participate will not in any way affects your employment with the university, or the benefits, privileges, or opportunities associated with your employment at UIC. You will not be offered or receive any special consideration if you participate in this research.



<u>Remember:</u> Your participation in this research is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University. If you decide to participate, you are free to withdraw at any time without affecting that relationship.
You will be given a copy of this form for your information and to keep for your records.

Signature of Participant

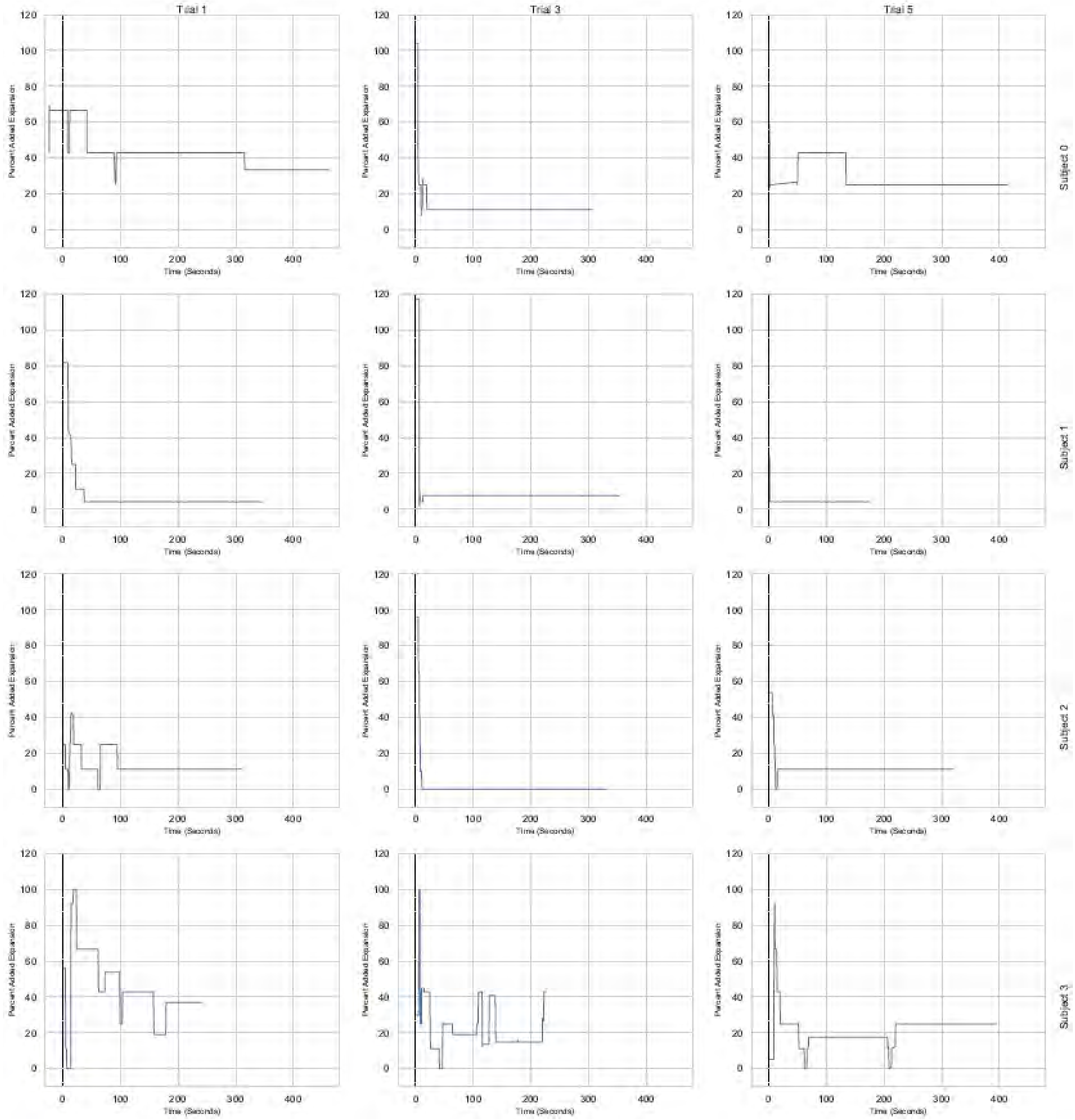
I have read (or someone has read to me) the above information. I have been given an opportunity to ask questions and my questions have been answered to my satisfaction. I agree to participate in this research. I have been given a copy of this form.

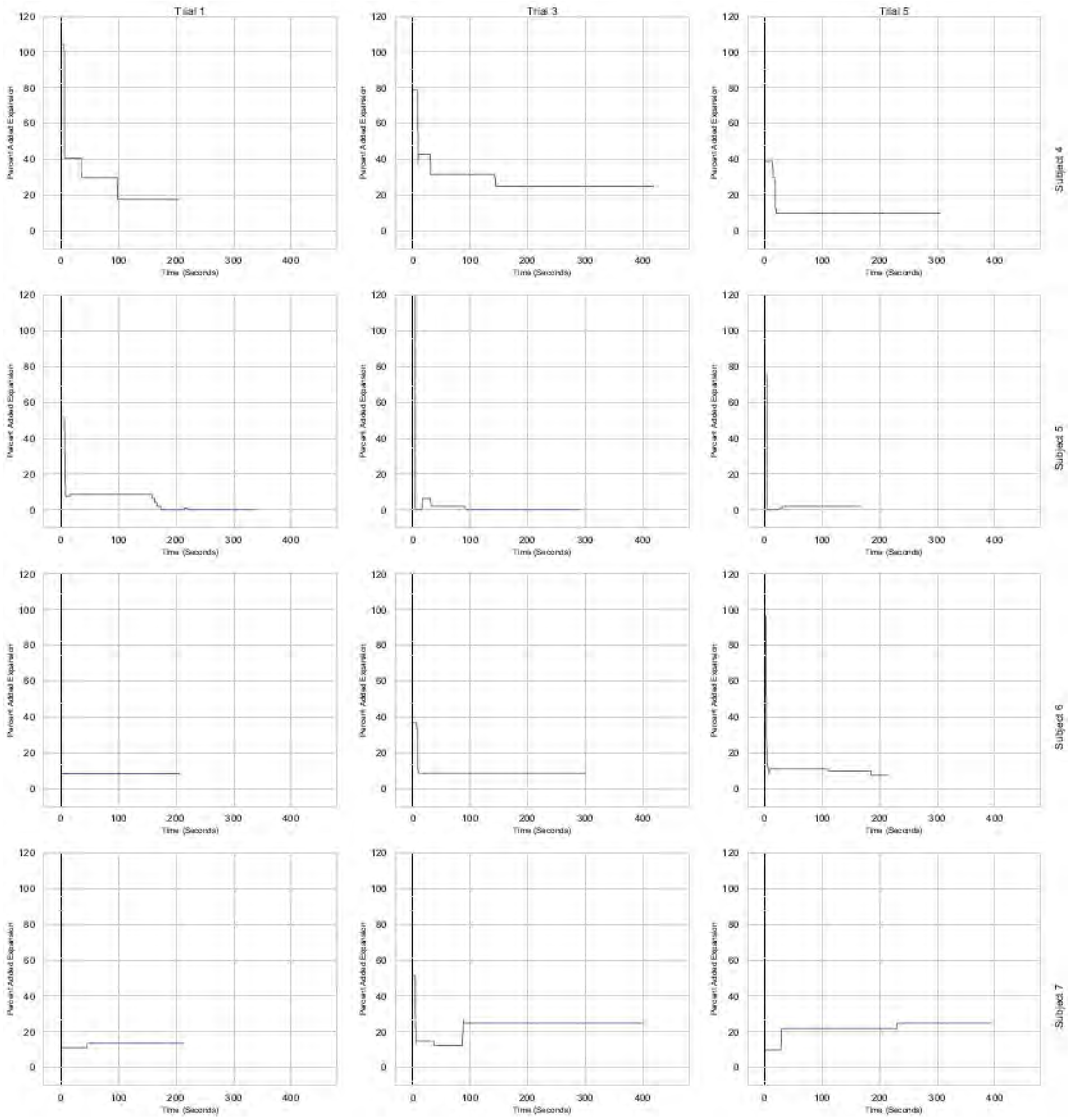
Signature Date

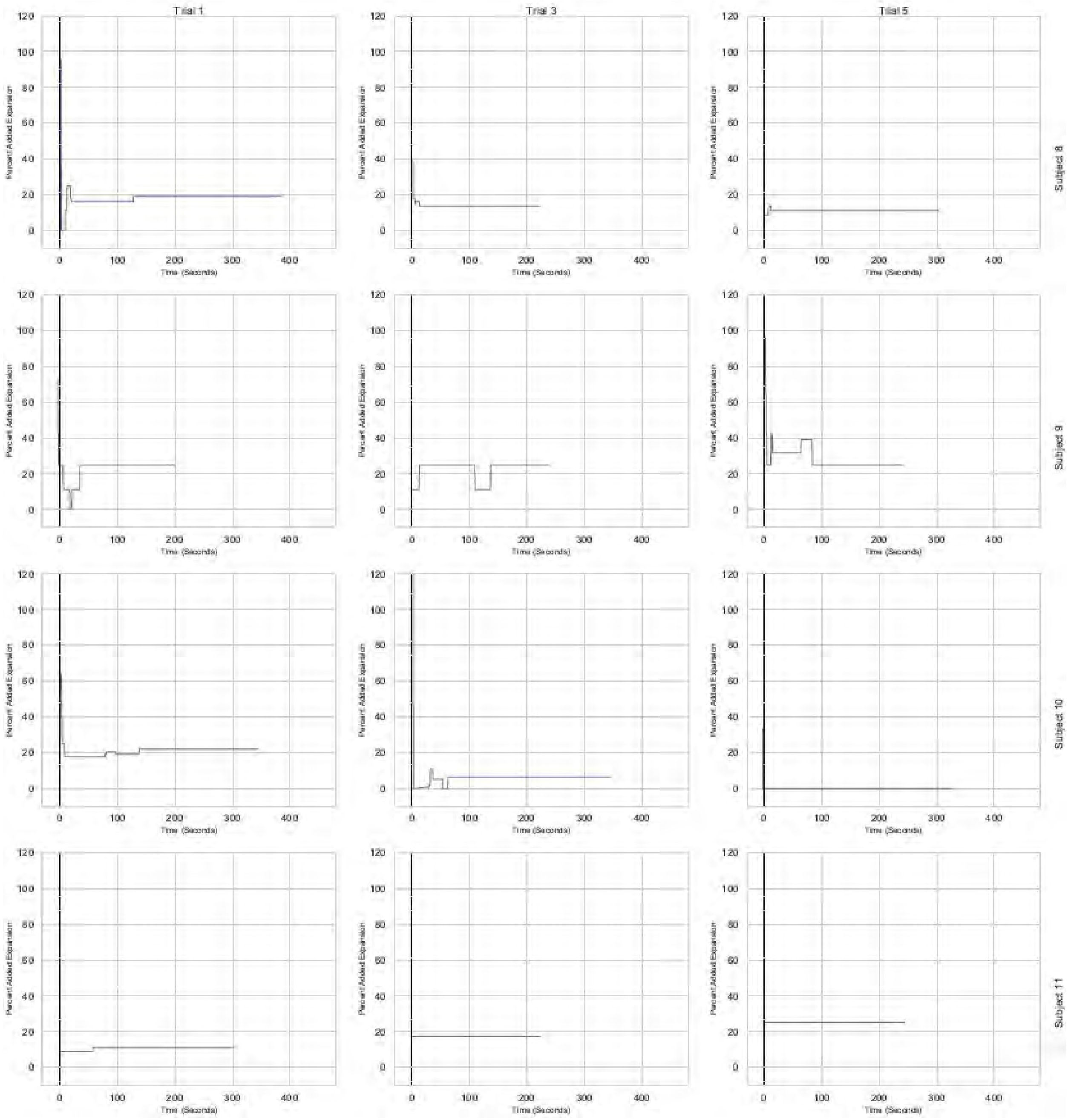
Printed Name

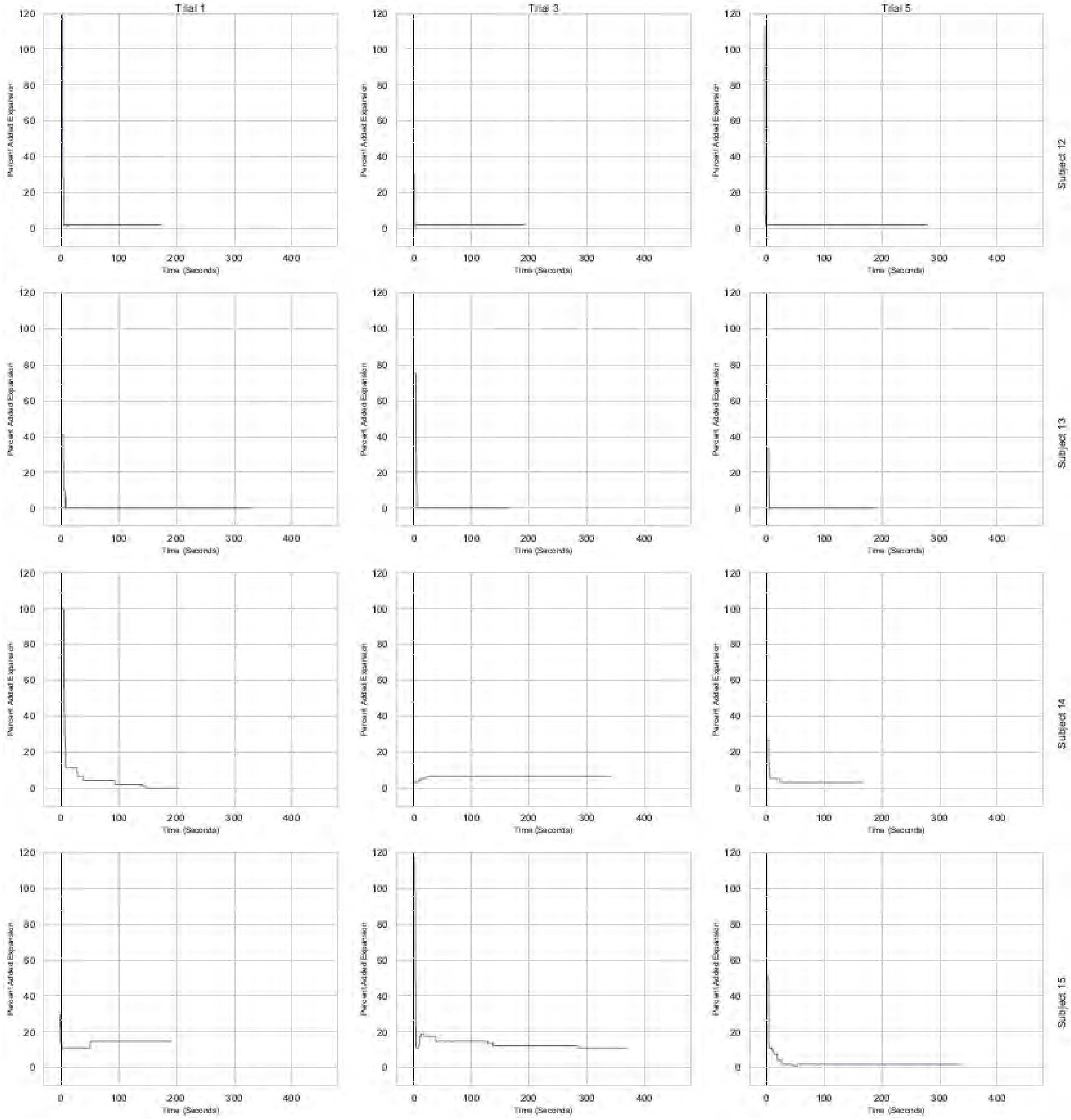
Signature of Researcher Date (must be same as participant's)

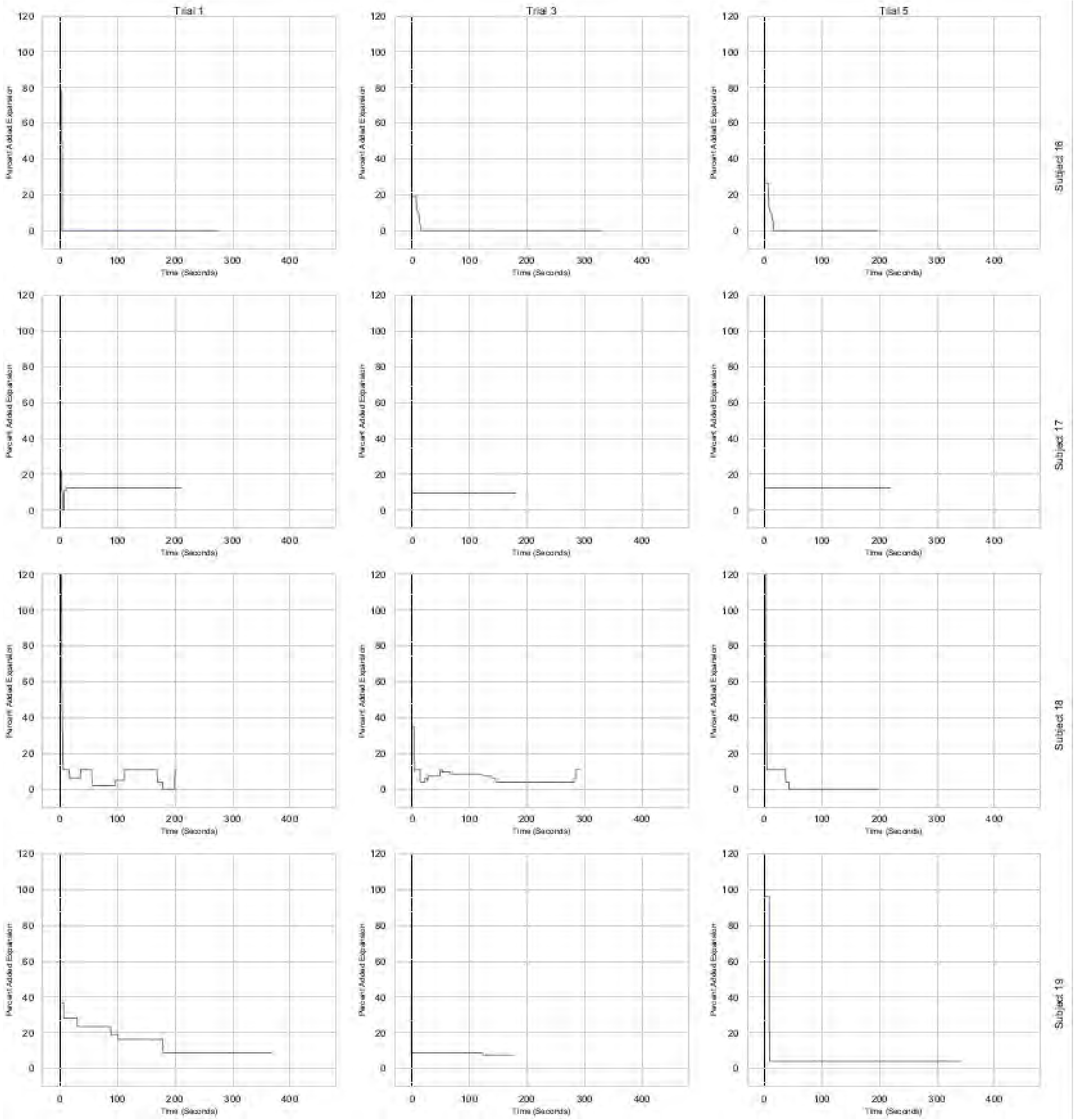
Appendix D Time Activity of All Subjects

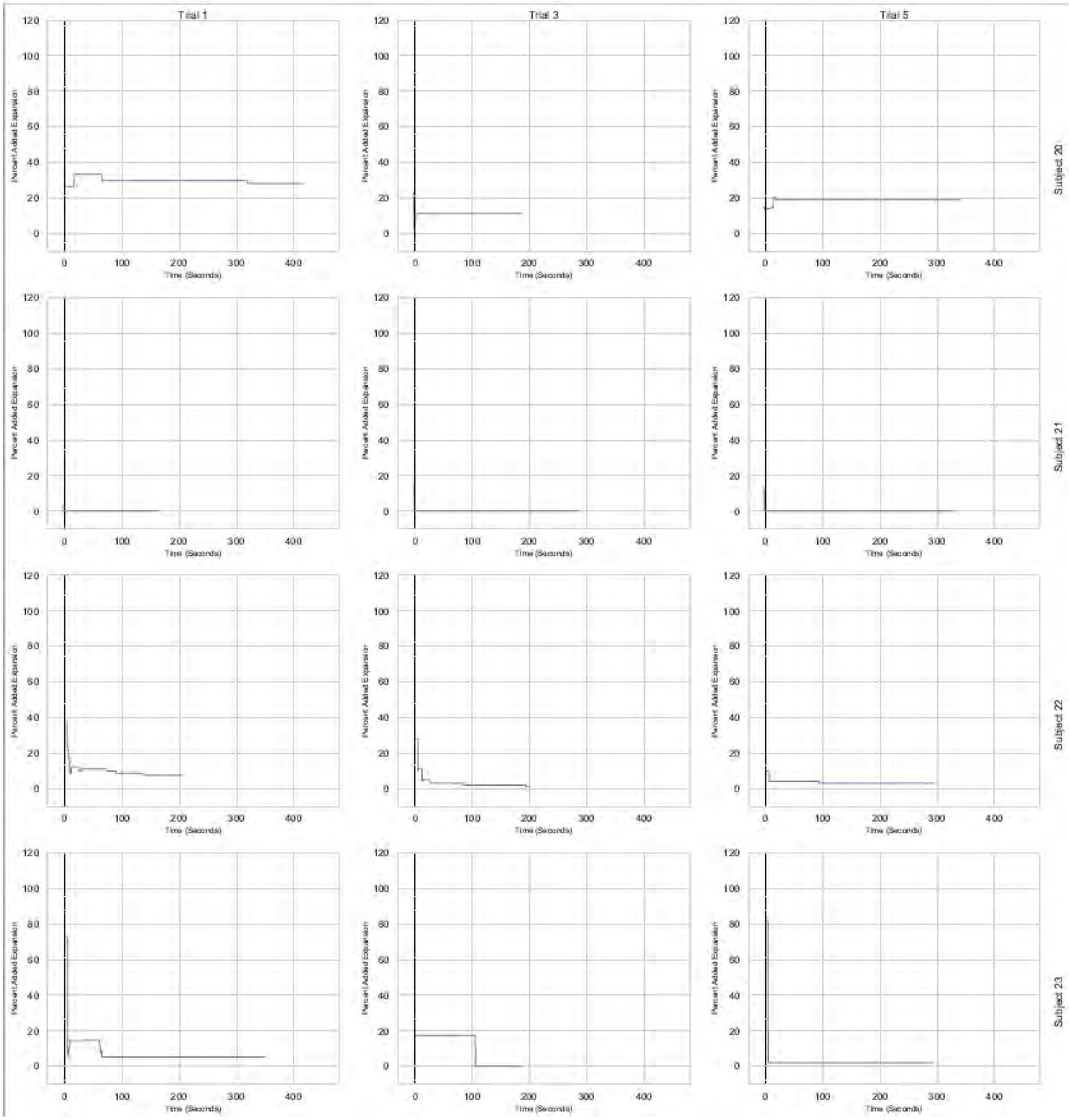


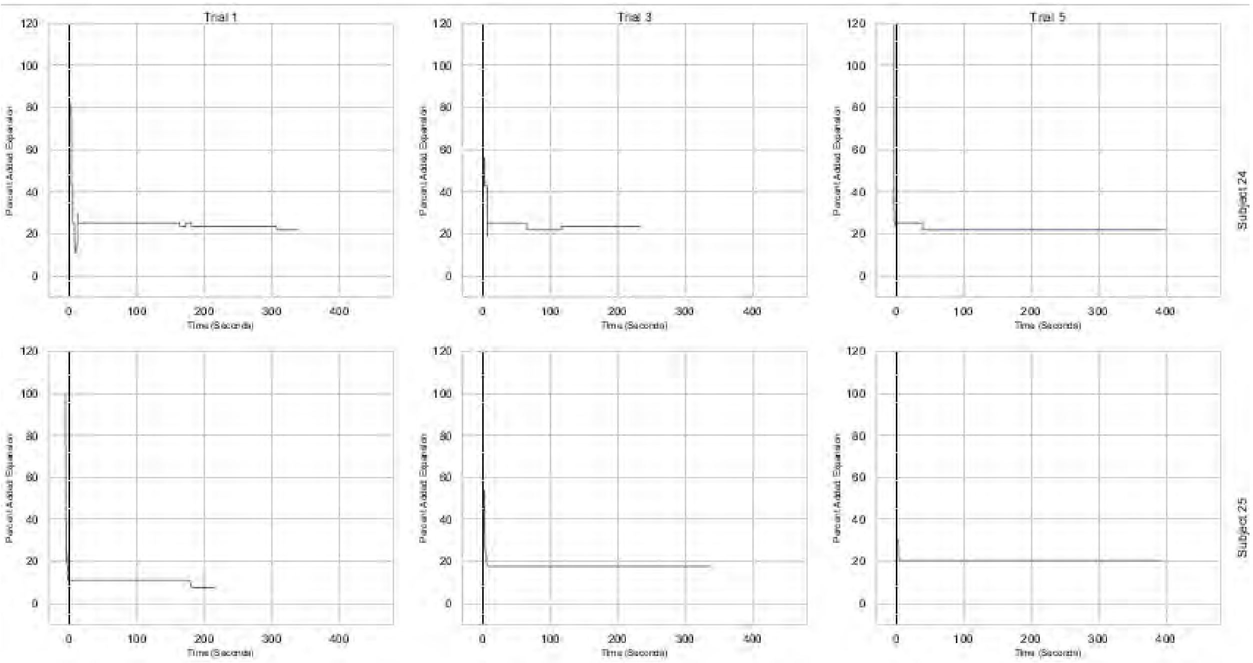

Trial 1
Trial 3
Trial 5
Subject 4
Subject 5
Subject 6
Subject 7
Percent Added Expansion
Time (Seconds)
120
100
80
60
40
20
0
0
100
200
300
400


Trial 1
Trial 3
Trial 5
Subject 8
Subject 9
Subject 10
Subject 11
Percent Added Expansion
0
20
40
60
80
100
120
Time (Seconds)
0
100
200
300
400

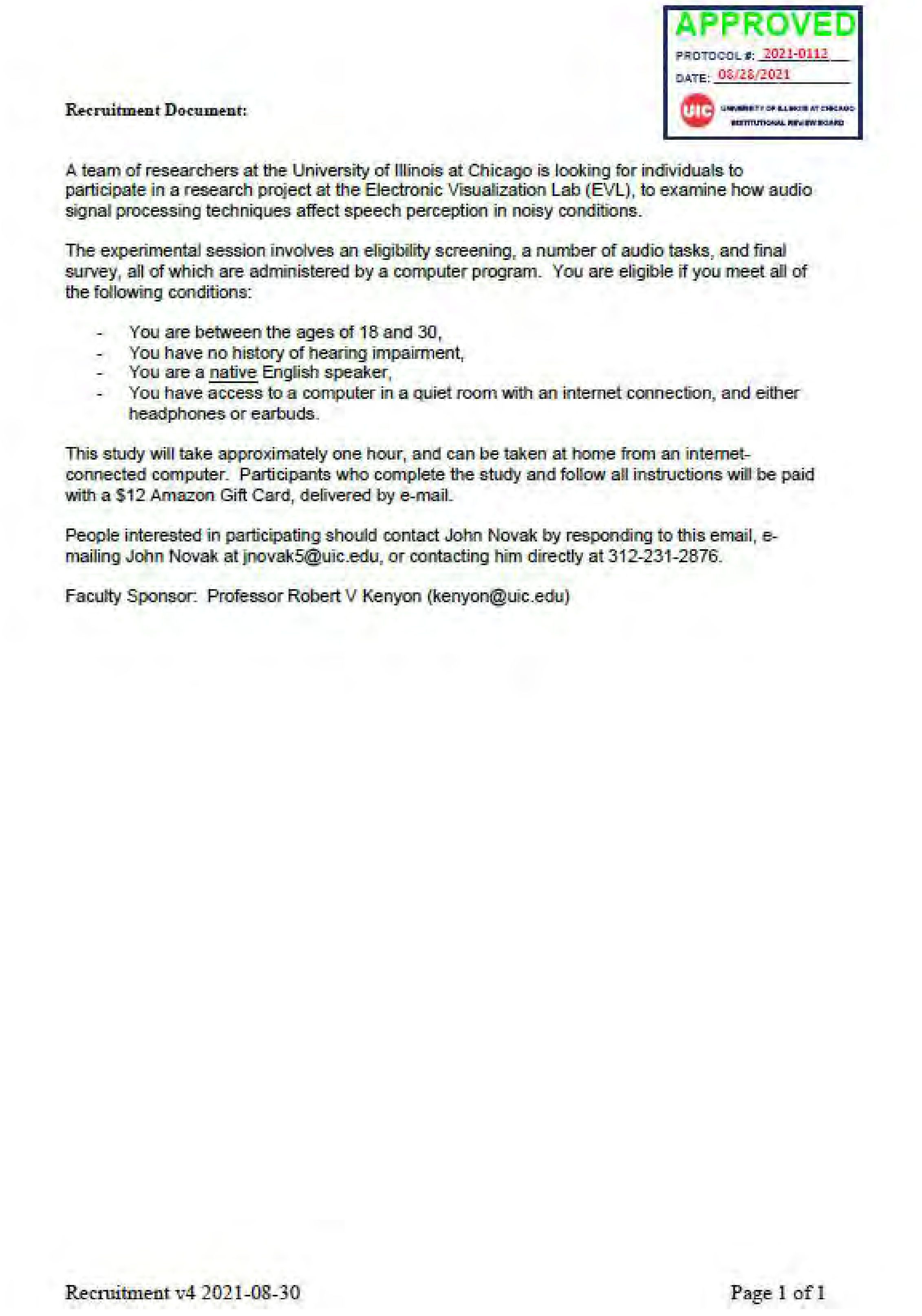

Trial 1
Trial 3
Trial 5
Subject 12
Subject 13
Subject 14
Subject 15
Percent Added Expansion
Time (Seconds)
120
100
80
60
40
20
0
0
100
200
300
400


Trial 1
Trial 3
Trial 5
Subject 16
Subject 17
Subject 18
Subject 19
Percent Added Expansion
Time (Seconds)
0
20
40
60
80
100
120
0
100
200
300
400


Trial 1
Trial 3
Trial 5
Subject 20
Subject 21
Subject 22
Subject 23
Percent Added Expansion
Time (Seconds)
120
100
80
60
40
20
0
0
100
200
300
400


Trial 1
Trial 3
Trial 5
Subject 24
Subject 25
Percent Added Expansion
Time (Seconds)
120
100
80
60
40
20
0
0
100
200
300
400

Appendix E Documentation, Phoneme Specific Speech In Noise



Recruitment Document:

A team of researchers at the University of Illinois at Chicago is looking for individuals to participate in a research project at the Electronic Visualization Lab (EVL), to examine how audio signal processing techniques affect speech perception in noisy conditions.

The experimental session involves an eligibility screening, a number of audio tasks, and final survey, all of which are administered by a computer program. You are eligible if you meet all of the following conditions:

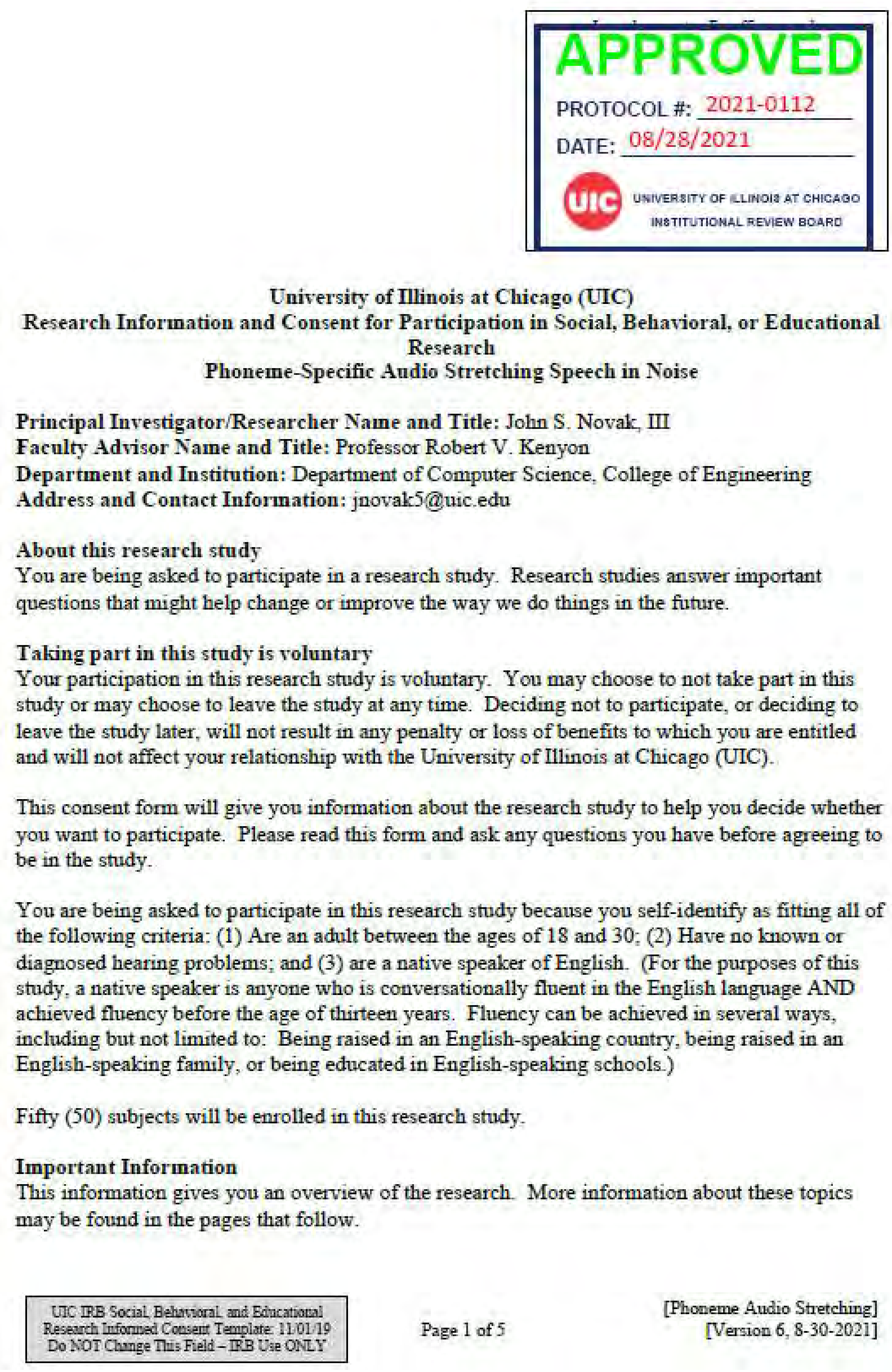
- You are between the ages of 18 and 30,
- You have no history of hearing impairment,
- You are a native English speaker,
- You have access to a computer in a quiet room with an internet connection, and either headphones or earbuds.

This study will take approximately one hour, and can be taken at home from an internet-connected computer. Participants who complete the study and follow all instructions will be paid with a $12 Amazon Gift Card, delivered by e-mail.

People interested in participating should contact John Novak by responding to this email, e-mailing John Novak at jnovak5@uic.edu, or contacting him directly at 312-231-2876.

Faculty Sponsor: Professor Robert V Kenyon (kenyon@uic.edu)

Recruitment v4 2021-08-30 Page 1 of 1



APPROVED

PROTOCOL #: 2021-0112

DATE: 08/28/2021

UIC UNIVERSITY OF ILLINOIS AT CHICAGO
INSTITUTIONAL REVIEW BOARD

University of Illinois at Chicago (UIC)
Research Information and Consent for Participation in Social, Behavioral, or Educational Research
Phoneme-Specific Audio Stretching Speech in Noise

Principal Investigator/Researcher Name and Title: John S. Novak, III
Faculty Advisor Name and Title: Professor Robert V. Kenyon
Department and Institution: Department of Computer Science, College of Engineering
Address and Contact Information: jnovak5@uic.edu

About this research study
You are being asked to participate in a research study. Research studies answer important questions that might help change or improve the way we do things in the future.

Taking part in this study is voluntary
Your participation in this research study is voluntary. You may choose to not take part in this study or may choose to leave the study at any time. Deciding not to participate, or deciding to leave the study later, will not result in any penalty or loss of benefits to which you are entitled and will not affect your relationship with the University of Illinois at Chicago (UIC).

This consent form will give you information about the research study to help you decide whether you want to participate. Please read this form and ask any questions you have before agreeing to be in the study.

You are being asked to participate in this research study because you self-identify as fitting all of the following criteria: (1) Are an adult between the ages of 18 and 30; (2) Have no known or diagnosed hearing problems; and (3) are a native speaker of English. (For the purposes of this study, a native speaker is anyone who is conversationally fluent in the English language AND achieved fluency before the age of thirteen years. Fluency can be achieved in several ways, including but not limited to: Being raised in an English-speaking country, being raised in an English-speaking family, or being educated in English-speaking schools.)

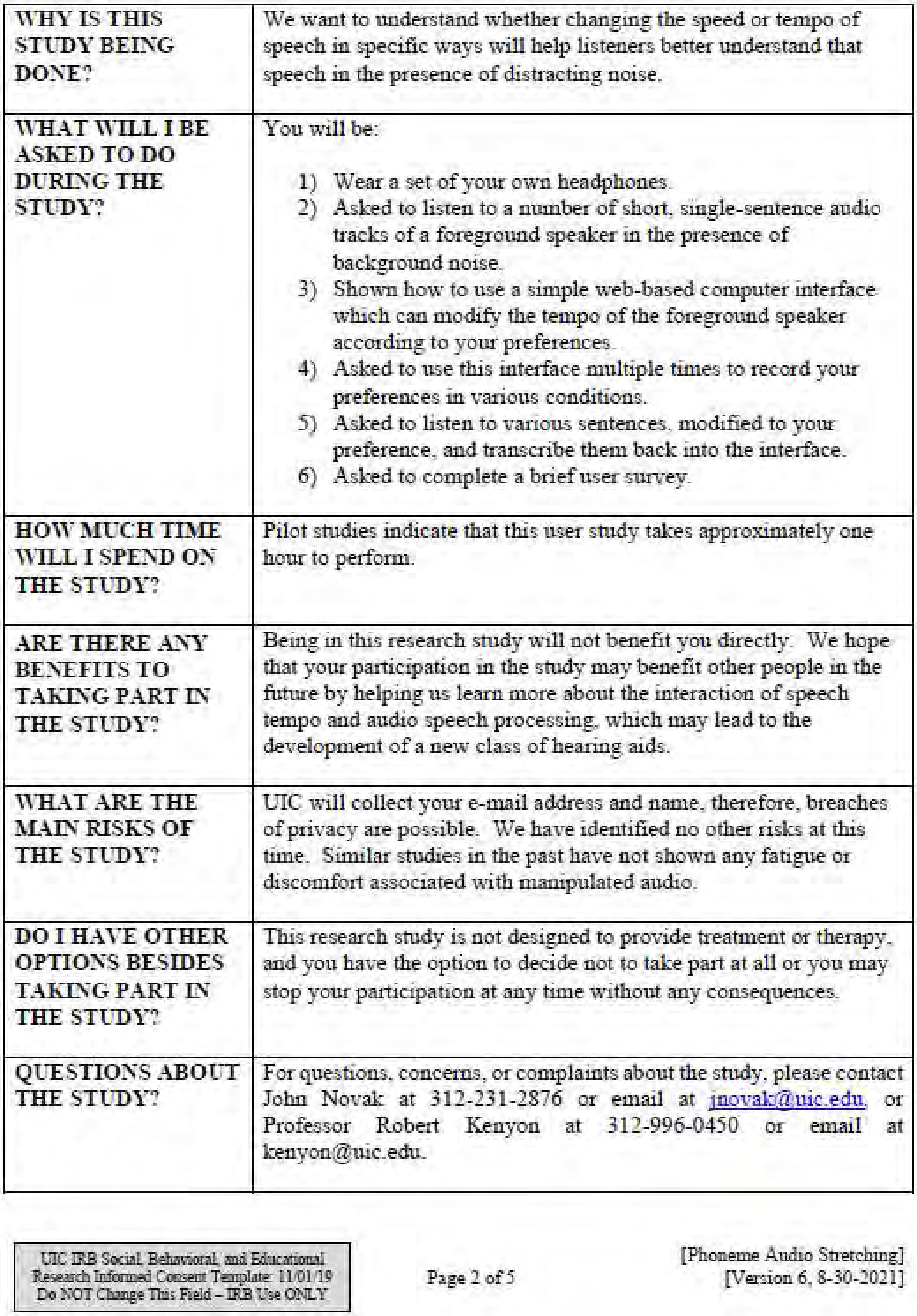
Fifty (50) subjects will be enrolled in this research study.

Important Information
This information gives you an overview of the research. More information about these topics may be found in the pages that follow.

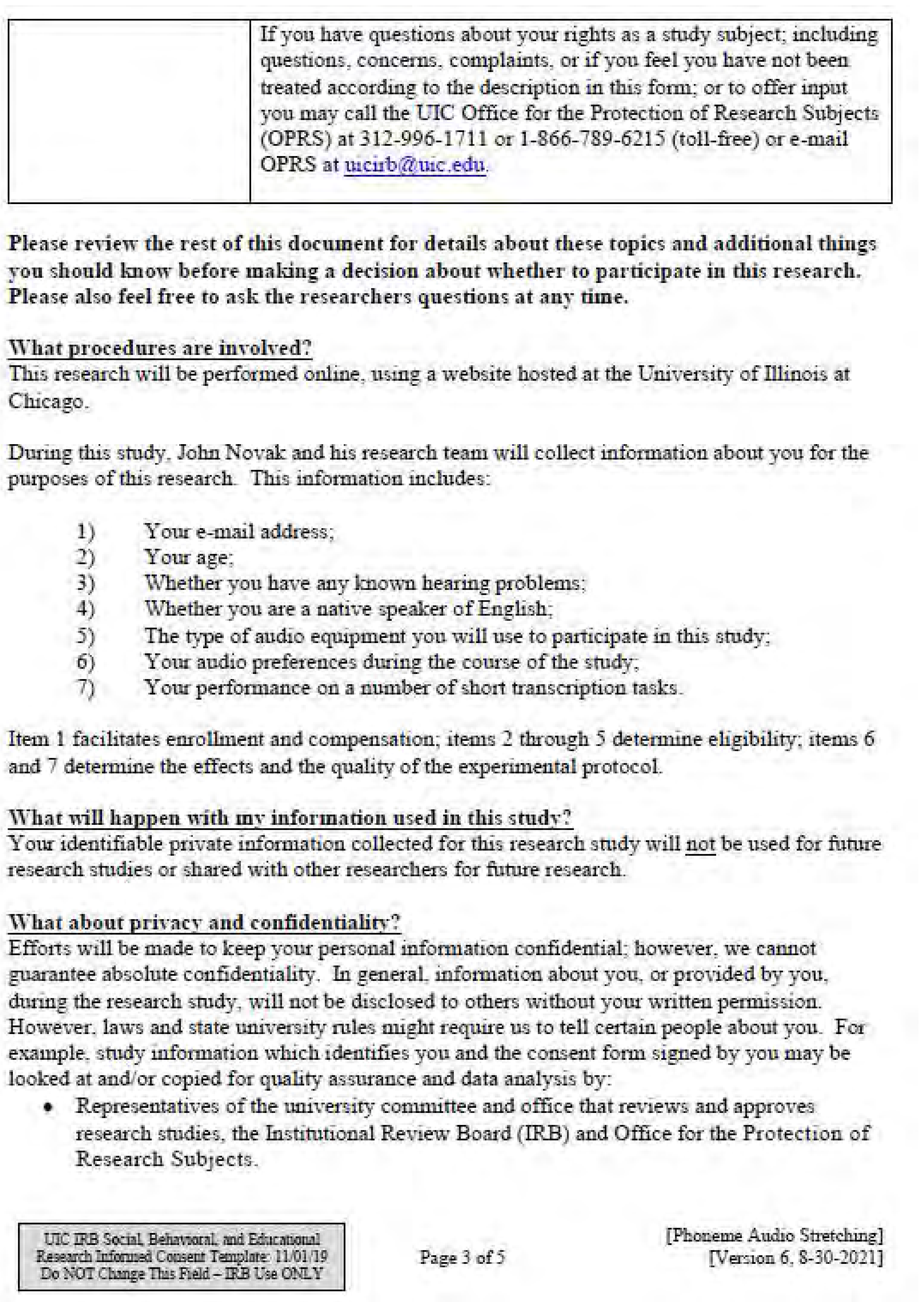
UIC IRB Social, Behavioral, and Educational Research Informed Consent Template: 11/01/19
Do NOT Change This Field – IRB Use ONLY

Page 1 of 5

[Phoneme Audio Stretching]
[Version 6, 8-30-2021]



WHY IS THIS STUDY BEING DONE?	We want to understand whether changing the speed or tempo of speech in specific ways will help listeners better understand that speech in the presence of distracting noise.
WHAT WILL I BE ASKED TO DO DURING THE STUDY?	You will be: 1) Wear a set of your own headphones. 2) Asked to listen to a number of short, single-sentence audio tracks of a foreground speaker in the presence of background noise. 3) Shown how to use a simple web-based computer interface which can modify the tempo of the foreground speaker according to your preferences. 4) Asked to use this interface multiple times to record your preferences in various conditions. 5) Asked to listen to various sentences, modified to your preference, and transcribe them back into the interface. 6) Asked to complete a brief user survey.
HOW MUCH TIME WILL I SPEND ON THE STUDY?	Pilot studies indicate that this user study takes approximately one hour to perform.
ARE THERE ANY BENEFITS TO TAKING PART IN THE STUDY?	Being in this research study will not benefit you directly. We hope that your participation in the study may benefit other people in the future by helping us learn more about the interaction of speech tempo and audio speech processing, which may lead to the development of a new class of hearing aids.
WHAT ARE THE MAIN RISKS OF THE STUDY?	UIC will collect your e-mail address and name, therefore, breaches of privacy are possible. We have identified no other risks at this time. Similar studies in the past have not shown any fatigue or discomfort associated with manipulated audio.
DO I HAVE OTHER OPTIONS BESIDES TAKING PART IN THE STUDY?	This research study is not designed to provide treatment or therapy, and you have the option to decide not to take part at all or you may stop your participation at any time without any consequences.
QUESTIONS ABOUT THE STUDY?	For questions, concerns, or complaints about the study, please contact John Novak at 312-231-2876 or email at jnovak@uic.edu, or Professor Robert Kenyon at 312-996-0450 or email at kenyon@uic.edu.



	If you have questions about your rights as a study subject; including questions, concerns, complaints, or if you feel you have not been treated according to the description in this form; or to offer input you may call the UIC Office for the Protection of Research Subjects (OPRS) at 312-996-1711 or 1-866-789-6215 (toll-free) or e-mail OPRS at uicirb@uic.edu.

Please review the rest of this document for details about these topics and additional things you should know before making a decision about whether to participate in this research. Please also feel free to ask the researchers questions at any time.

What procedures are involved?

This research will be performed online, using a website hosted at the University of Illinois at Chicago.

During this study, John Novak and his research team will collect information about you for the purposes of this research. This information includes:

1) Your e-mail address;
2) Your age;
3) Whether you have any known hearing problems;
4) Whether you are a native speaker of English;
5) The type of audio equipment you will use to participate in this study;
6) Your audio preferences during the course of the study;
7) Your performance on a number of short transcription tasks.

Item 1 facilitates enrollment and compensation; items 2 through 5 determine eligibility; items 6 and 7 determine the effects and the quality of the experimental protocol.

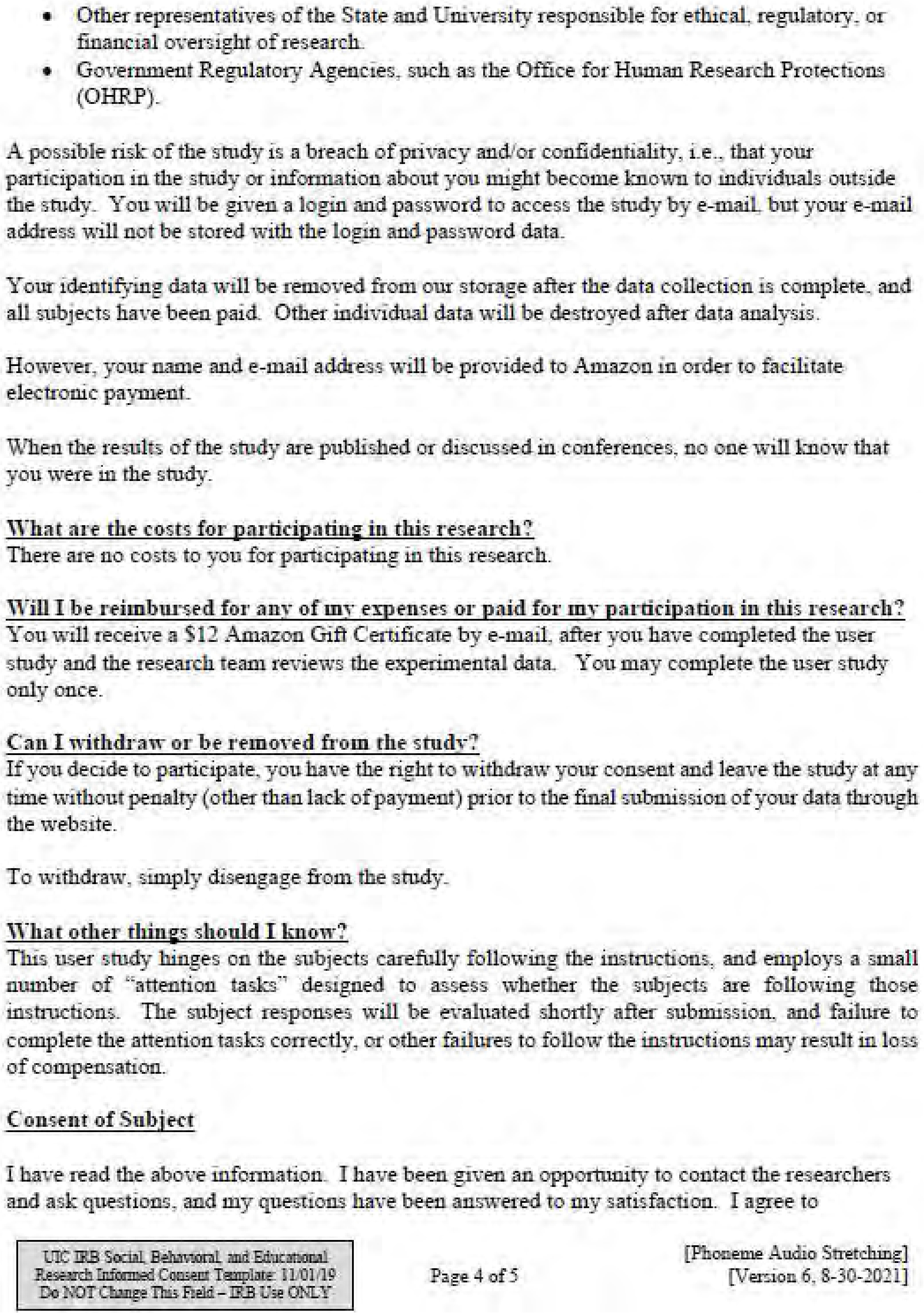
What will happen with my information used in this study?

Your identifiable private information collected for this research study will not be used for future research studies or shared with other researchers for future research.

What about privacy and confidentiality?

Efforts will be made to keep your personal information confidential; however, we cannot guarantee absolute confidentiality. In general, information about you, or provided by you, during the research study, will not be disclosed to others without your written permission. However, laws and state university rules might require us to tell certain people about you. For example, study information which identifies you and the consent form signed by you may be looked at and/or copied for quality assurance and data analysis by:

- Representatives of the university committee and office that reviews and approves research studies, the Institutional Review Board (IRB) and Office for the Protection of Research Subjects.



- Other representatives of the State and University responsible for ethical, regulatory, or financial oversight of research.
- Government Regulatory Agencies, such as the Office for Human Research Protections (OHRP).

A possible risk of the study is a breach of privacy and/or confidentiality, i.e., that your participation in the study or information about you might become known to individuals outside the study. You will be given a login and password to access the study by e-mail, but your e-mail address will not be stored with the login and password data.

Your identifying data will be removed from our storage after the data collection is complete, and all subjects have been paid. Other individual data will be destroyed after data analysis.

However, your name and e-mail address will be provided to Amazon in order to facilitate electronic payment.

When the results of the study are published or discussed in conferences, no one will know that you were in the study.

What are the costs for participating in this research?

There are no costs to you for participating in this research.

Will I be reimbursed for any of my expenses or paid for my participation in this research?

You will receive a $12 Amazon Gift Certificate by e-mail, after you have completed the user study and the research team reviews the experimental data. You may complete the user study only once.

Can I withdraw or be removed from the study?

If you decide to participate, you have the right to withdraw your consent and leave the study at any time without penalty (other than lack of payment) prior to the final submission of your data through the website.

To withdraw, simply disengage from the study.

What other things should I know?

This user study hinges on the subjects carefully following the instructions, and employs a small number of "attention tasks" designed to assess whether the subjects are following those instructions. The subject responses will be evaluated shortly after submission, and failure to complete the attention tasks correctly, or other failures to follow the instructions may result in loss of compensation.

Consent of Subject

I have read the above information. I have been given an opportunity to contact the researchers and ask questions, and my questions have been answered to my satisfaction. I agree to

participate in this research. **PLEASE PRINT OUT A COPY OF THIS DOCUMENT FOR YOUR RECORDS.**

Appendix F Phoneme Confusion Matrices

Table XII Phoneme Confusion Matrix, Training Set

Predicted	Silence	Stop	Fricative	Nasal	Approx.	Vowel	Actual
Silence	264653	387	141	187	132	217	265717
Stop	175	313451	76	264	297	1825	316088
Fricative	443	3908	282780	156	386	2112	289785
Nasal	63	263	56	113421	140	1145	115088
Approx.	83	104	20	66	144769	2055	147097
Vowel	26	541	166	163	1484	677239	679619
	265443	318654	283239	114257	147208	684593	1813394

Table XIII Phoneme Confusion Matrix, Test Set

Predicted	Silence	Stop	Fricative	Nasal	Approx.	Vowel	Actual
Silence	77257	2168	853	716	591	745	82330
Stop	1615	84653	1818	1739	1244	4395	95464
Fricative	1520	6135	76609	730	883	3234	89111
Nasal	534	1260	329	30008	610	2908	35649
Approx.	563	1124	560	702	37119	11288	51356
Vowel	521	3277	1809	1874	8217	199498	215196
	82010	98617	81978	35769	48664	222068	569106

Appendix G Pairwise Statistics

Table XIV Pairwise Intelligibility, All Data (Wilcoxon Test, Benjamini-Hochberg Corrected)

Condition 1	Condition 2	p-value	Critical	Significant
SNR Comparisons, Modified vs Unmodified				
2 dB, Mod	2 dB, Unmod	1.26E-02	0.04	Yes
5 dB, Mod	5 dB, Unmod	1.42E-01	0.05	No
8 dB, Mod	8 dB, Unmod	1.26E-06	0.03	Yes
SNR Comparisons within Modified Condition				
2 dB, Mod	5 dB, Mod	1.75E-07	0.02	Yes
5 dB, Mod	8 dB, Mod	1.58E-03	0.04	Yes
2 dB, Mod	8 dB, Mod	2.09E-08	0.01	Yes
SNR Comparisons within Unmodified Condition				
2 dB, Unmod	5 dB, Unmod	3.90E-05	0.03	Yes
5 dB, Unmod	8 dB, Unmod	2.51E-08	0.00	Yes
2 dB, Unmod	8 dB, Unmod	7.45E-09	0.01	Yes
Other SNR Comparisons				
2 dB, Mod	5 dB, Unmod	1.63E-07	0.02	Yes
5 dB, Mod	8 dB, Unmod	6.2E-07	0.02	Yes
2 dB, Mod	8 dB, Unmod	8.2E-09	0.01	Yes
2 dB, Unmod	5 dB, Mod	3.0E-03	0.04	Yes
2 dB, Unmod	8 dB, Mod	0.18E-01	0.05	No
5 dB, Unmod	8 dB, Mod	3.21E-05	0.03	Yes

This table represents all 15 possible pairwise tests for statistical significance of the intelligibility data of Chapter Seven (Phoneme Aware Speech In Noise), for all data collected (N=44). This data is not normally distributed, therefore the p-values are determined with Wilcoxon Signed Rank tests, which are further corrected by a Benjamini-Hochberg procedure controlling for a false detection rate of 0.05.

Table XV Pairwise Intelligibility, 'Hard' (Wilcoxon Test, Benjamini-Hochberg Corrected)

Condition 1	**Condition 2**	**p-value**	**Critical**	**Significant**
SNR Comparisons, Modified vs Unmodified				
2 dB, Mod	2 dB, Unmod	4.56E-03	0.03	Yes
5 dB, Mod	5 dB, Unmod	5.16E-02	0.04	No
8 dB, Mod	8 dB, Unmod	2.69E-03	0.03	Yes
SNR Comparisons within Modified Condition				
2 dB, Mod	5 dB, Mod	1.17E-03	0.02	Yes
5 dB, Mod	8 dB, Mod	5.19E-02	0.04	No
2 dB, Mod	8 dB, Mod	8.29E-04	0.01	Yes
SNR Comparisons within Unmodified Condition				
2 dB, Unmod	5 dB, Unmod	1.12E-02	0.03	Yes
5 dB, Unmod	8 dB, Unmod	8.32E-04	0.02	Yes
2 dB, Unmod	8 dB, Unmod	2.90E-04	0.01	Yes
Other SNR Comparisons				
2 dB, Mod	5 dB, Unmod	7.65E-04	0.01	Yes
5 dB, Mod	8 dB, Unmod	9.92E-04	0.02	Yes
2 dB, Mod	8 dB, Unmod	2.88E-04	0.00	Yes
2 dB, Unmod	5 dB, Mod	5.17E-01	0.05	No
2 dB, Unmod	8 dB, Mod	1.33E-01	0.04	No
5 dB, Unmod	8 dB, Mod	4.69E-01	0.05	No

This table represents all 15 possible pairwise tests for statistical significance of the intelligibility data of Chapter Seven (Phoneme Aware Speech In Noise), for all responses answering that listening to the speech with changed cadence was in ‘Very difficult’, ‘Somewhat difficult’, or ‘A little difficult’ (N=17). This data is not normally distributed, therefore the p-values are determined with Wilcoxon Signed Rank tests, which are further corrected by a Benjamini-Hochberg procedure controlling for a false detection rate of 0.05.

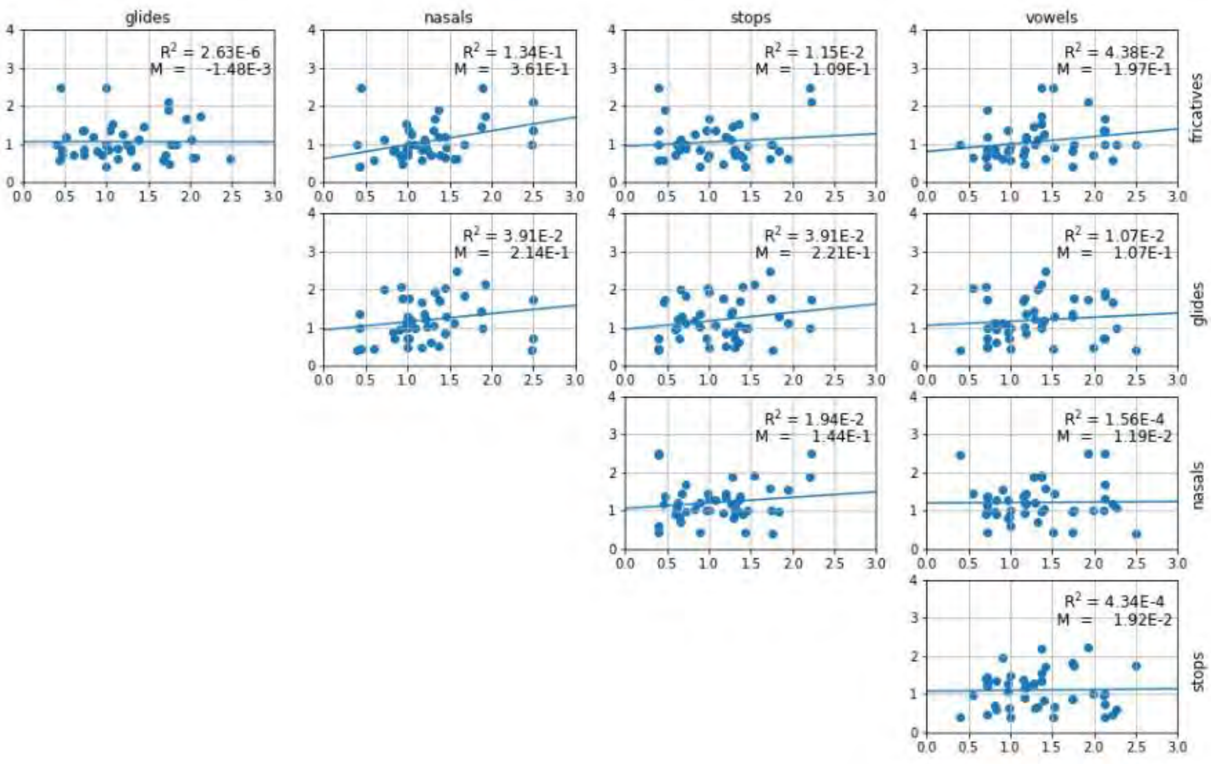
Table XVI Pairwise Intelligibility, 'Easy' (Wilcoxon Test, Benjamini-Hochberg Corrected)

Condition 1	Condition 2	p-value	Critical	Significant
SNR Comparisons, Modified vs Unmodified				
2 dB, Mod	2 dB, Unmod	3.46E-01	0.05	No
5 dB, Mod	5 dB, Unmod	8.45E-01	0.05	No
8 dB, Mod	8 dB, Unmod	5.98E-04	0.03	Yes
SNR Comparisons within Modified Condition				
2 dB, Mod	5 dB, Mod	1.35E-04	0.02	Yes
5 dB, Mod	8 dB, Mod	4.01E-02	0.04	No
2 dB, Mod	8 dB, Mod	5.60E-05	0.01	Yes
SNR Comparisons within Unmodified Condition				
2 dB, Unmod	5 dB, Unmod	8.67E-03	0.04	Yes
5 dB, Unmod	8 dB, Unmod	4.90E-05	0.01	Yes
2 dB, Unmod	8 dB, Unmod	3.90E-05	0.00	Yes
Other SNR Comparisons				
2 dB, Mod	5 dB, Unmod	4.45E-04	0.02	Yes
5 dB, Mod	8 dB, Unmod	6.52E-04	0.03	Yes
2 dB, Mod	8 dB, Unmod	4.50E-05	0.01	Yes
2 dB, Unmod	5 dB, Mod	5.02E-03	0.03	Yes
2 dB, Unmod	8 dB, Mod	2.31E-04	0.02	Yes
5 dB, Unmod	8 dB, Mod	5.95E-02	0.04	No

This table represents all 15 possible pairwise tests for statistical significance of the intelligibility data of Chapter Seven (Phoneme Aware Speech In Noise), for all responses answering that listening to the speech with changed cadence was in ‘Very easy, ‘Somewhat easy’, or ‘A little easy (N=22). This data is not normally distributed, therefore the p-values are determined with Wilcoxon Signed Rank tests, which are further corrected by a Benjamini-Hochberg procedure controlling for a false detection rate of 0.05.

Appendix H Correlations Of Phoneme Expansion Factors

Figure 28 Correlations of Phoneme Expansion Factors



Appendix I Republication Permissions (Text)

The text of Chapter 4 is based on published in Novak, John S., Jason Archer, Robert V. Kenyon, and Valeriy Shafiro. "Audio dilation in real time speech communication." In *Proceedings of Meetings on Acoustics 169ASA*, vol. 23, no. 1, p. 050008. Acoustical Society of America, 2015.

The Proceedings of Meetings on Acoustics "Permission to Reuse Content" policy can be found at the following URL, and the policy is reproduced here for convenience:

https://publishing.aip.org/resources/researchers/rights-and-permissions/permissions/

> *Authors do* ***not*** *need permission from AIP Publishing to:*
>
> - *quote from a publication (please include the material in quotation marks and provide the customary acknowledgment of the source)*
> - *reuse any materials that are licensed under a Creative Commons CC BY license (please format your credit line: "Author names, Journal Titles, Vol.#, Article ID#, Year of Publication; licensed under a Creative Commons Attribution (CC BY) license.")*
> - *reuse your own AIP Publishing article in your thesis or dissertation (please format your credit line: "Reproduced from [FULL CITATION], with the permission of AIP Publishing")*
> - *reuse content that appears in an AIP Publishing journal for republication in another AIP Publishing journal (please format your credit line: "Reproduced from [FULL CITATION], with the permission of AIP Publishing")*
> - *make multiple copies of articles–although you must contact the Copyright Clearance Center (CCC) at www.copyright.com to do this*

The text of Chapter 5 is based on material presented at the “Interspeech 2018” conference, and published in Novak III, John S., Daniel Bunn, and Robert V. Kenyon. "The Effects of Time Expansion on English as a Second Language Individuals." In *INTERSPEECH*, pp. 2643-2647. 2019.

The text of Chapter 6 is based on material presented at the “Interspeech 2019” conference, and published in Bunn, Daniel. "Effects of Audio Dilation and Listening Skill Ability for English as a Second Language." Master’s thesis, University of Illinois at Chicago, 2018.

The Interspeech 2018 and Interspeech 2019 republication policies can both be found at the following URL, and the policy is reproduced here for convenience:

https://www.isca-speech.org/iscaweb/index.php/component/content/article?id=33&Itemid=144

> *“Q Can I place a copy of my paper in my institional [sic] or other repository?”*
>
> *“A Yes. For any paper published in the proceedings of INTERSPEECH or other ISCA sponsored events whose copyright is transferred to ISCA, ISCA grants each author permission to use the paper in that author's dissertation or in institutional and public (such as arXiv) repositories (paper and/or electronic versions), provided that the paper is correctly referenced with the proceeding information including page numbers and DOI (if available). All authors of the article have the same permission for reprinting, under the same conditions.”*
>
> *Moreover, all authors are required to refer to the paper in the proceedings of INTERSPEECH or ISCA sponsored event, and not use the reference for the institutional or public repository copy of the paper.”*

Appendix J Republication Permissions (Figures)

Figure 1 has been adapted from [9] with modest additions and changes to labels in accord with CC BY 2.5

Figure 4 has been adapted from [12] with modest additions and changes to labels in accord with CC BY 2.5

Figure 5 has been adapted from [13] with modest additions and changes to labels in accord with CC BY 2.5

A link to the CC BY 2.5 license can be found at the following URL: https://creativecommons.org/licenses/by/2.5/

Figure 19 has been adapted from [73] unaltered, in accord with CC 4.0

A link to the CC4.0 license can be found at the following URL: https://creativecommons.org/licenses/by/4.0/legalcode

Figure 23 has been re-printed from Fig 1a of [6], which was presented at the "Interspeech 2019" conference.

Figure 24 has been reprinted from Fig 1c and Fig 1c of [6] , which was presented at the "Interspeech 2019" conference.

The Interspeech 2019 republication guidelines can be found at the following URL, and the policy is reproduced here for convenience: https://www.isca-speech.org/iscaweb/index.php/component/content/article?id=33&Itemid=144

> *"Q Can I place a copy of my paper in my institional [sic] or other repository?"*
>
> *"A Yes. For any paper published in the proceedings of INTERSPEECH or other ISCA sponsored events whose copyright is transferred to ISCA, ISCA grants each author permission to use the paper in that author's dissertation or in institutional and public (such as arXiv) repositories (paper and/or electronic versions), provided that the paper is correctly referenced with the proceeding information including page numbers and DOI (if available). All authors of the article have the same permission for reprinting, under the same conditions."*

Moreover, all authors are required to refer to the paper in the proceedings of INTERSPEECH or ISCA sponsored event, and not use the reference for the institutional or public repository copy of the paper."

Appendix K Uploaded Material

This dissertation contains the following supplemental uploaded materials:

1. A zipped archive of an Eclipse java workspace (“Audio Dilation Interspeech.zip”) containing the Single-Talker Half-Prototype presented in Chapter 3.1.

2. A zipped archive of an Eclipse android workspace (“voicereceiver.zip”) containing the Two-Talker Wireless Prototype presented in Chapter 3.2.

3. A zipped archive of an Eclipse java workspace (“Dilation_Server.zip”) containing the server portions of the Full Transmit-Receive Server Prototype presented in Chapter 3.3.

4. A zipped archive of an Eclipse android workspace (“VOIP_TxClient.zip”) containing the mobile portions of the Full Transmit-Receive Server Prototype presented in Chapter 3.3.

For simplicity, these are all contained in the single archive file, “Expansion.zip”.

Cited Literature

1. Novak III, John S., Jason Archer, Valeriy Shafiro, Robert V. Kenyon, and Jason Leigh. "On-line audio dilation for human interaction." In *INTERSPEECH*, pp. 1869-1871. 2013.

2. Novak III, John S., Aashish Tandon, Jason Leigh, and Robert V. Kenyon. "Networked on-line audio dilation." In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pp. 255-258. 2014.

3. Novak, John S., Jason Archer, Robert V. Kenyon, and Valeriy Shafiro. "Audio dilation in real time speech communication." In *Proceedings of Meetings on Acoustics 169ASA*, vol. 23, no. 1, p. 050008. Acoustical Society of America, 2015.

4. Novak III, John S., and Robert V. Kenyon. "Effects of User Controlled Speech Rate on Intelligibility in Noisy Environments." In *INTERSPEECH*, pp. 1853-1857. 2018.

5. Novak III, John S., Daniel Bunn, and Robert V. Kenyon. "The Effects of Time Expansion on English as a Second Language Individuals." In *INTERSPEECH*, pp. 2643-2647. 2019.

6. Bunn, Daniel. "Effects of Audio Dilation and Listening Skill Ability for English as a Second Language." Master's thesis, University of Illinois at Chicago, 2018.

7. Bradlow, Ann R., Rachel E. Baker, Arim Choi, Midam Kim, and Kristin J. Van Engen. "The Wildcat corpus of native-and foreign-accented English." *Journal of the Acoustical Society of America* 121, no. 5 (2007): 3072.

8. Bentley, Lauren E., and Hua Ou. "Using QuickSIN speech material to measure acceptable noise level for adults with hearing loss." (2017).

9. Lynch, Patrick J, "Mouth Anatomy", 2006, *https://commons.wikimedia.org/wiki/File:Mouth_anatomy-de.svg*

10. Van den Berg, Janwillem. "Myoelastic-aerodynamic theory of voice production." *Journal of speech and hearing research* 1, no. 3 (1958): 227-244.

11. Coupé, Christophe, Yoon Mi Oh, Dan Dediu, and François Pellegrino. "Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche." Science advances 5, no. 9 (2019): eaaw2594.

12. Chitttka, Lars, and Axel Brockmann, "Anatomy of the Human Ear", 2009, https://commons.wikimedia.org/wiki/File:Anatomy_of_the_Human_Ear.svg

13. Kern A, Heid C, Steeb W-H, Stoop N, Stoop R, "Uncoiled Cochlea with Basilar Membrane", 2006, https://commons.wikimedia.org/wiki/File:Uncoiled_cochlea_with_basilar_membrane.png

14. Zwicker, Eberhard. "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)." *The Journal of the Acoustical Society of America* 33, no. 2 (1961): 248-248.

15. Olson, Harry Ferdinand. *Music, physics and engineering.* Vol. 1769. Courier Corporation, 1967.

16. Picheny, Michael Alan, and N. I. Durlach. "Speaking clearly for the hard of hearing." *The Journal of the Acoustical Society of America* 65, no. S1 (1979): S135-S136.

17. Chen, Francine Robina. "Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level." PhD diss., Massachusetts Institute of Technology, 1980.

18. Krause, Jean C., and Louis D. Braida. "Acoustic properties of naturally produced clear speech at normal speaking rates." *The Journal of the Acoustical Society of America* 115, no. 1 (2004): 362-378.

19. Picheny, Michael A., Nathaniel I. Durlach, and Louis D. Braida. "Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech." *Journal of Speech, Language, and Hearing Research* 32, no. 3 (1989): 600-603.

20. Ferguson, Sarah Hargus, and Diane Kewley-Port. "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners." *The Journal of the Acoustical Society of America* 112, no. 1 (2002): 259-271.

21. Papoušek, Mechthild, and Shu-Fen C. Hwang. "Tone and intonation in Mandarin babytalk to presyllabic infants: Comparison with registers of adult conversation and foreign language instruction." *Applied Psycholinguistics* 12, no. 4 (1991): 481-504.

22. Uther, Maria, Monja A. Knoll, and Denis Burnham. "Do you speak E-NG-LI-SH? A comparison of foreigner- and infant-directed speech." *Speech communication* 49, no. 1 (2007): 2-7.

23. Moon, Seung-Jae, and Björn Lindblom. "Formant undershoot in clear and citation-form speech: A second progress report." *STL-QPSR 30* (1989): 121-123.

24. Moon, Seung-Jae, and Björn Lindblom. "Interaction between duration, context, and speaking style in English stressed vowels." *The Journal of the Acoustical society of America* 96, no. 1 (1994): 40-55.

25. Bradlow, Ann R., Nina Kraus, and Erin Hayes. "Speaking Clearly for Children With Learning Disabilities: Sentence Perception in Noise." *Journal of Speech, Language, and Hearing Research* 46: 80-97.

26. Payne, Elinor, Brechtje Postb, Lluïsa Astrucc, Pilar Prietod, Maria del Mar, and Pompeu Fabra. "Rhythmic Modification in Child Directed Speech." *Oxford University Working Papers in Linguistics, Philology & Phonetics*: 123.

27. Bradlow, Ann R. "Confluent talker-and listener-oriented forces in clear speech production." *Laboratory phonology* 7 (2002): 241-273.

28. Kondaurova, Maria V., and Tonya R. Bergeson. "The effects of age and infant hearing status on maternal use of prosodic cues for clause boundaries in speech." (2011).

29. Grieser, DiAnne L., and Patricia K. Kuhl. "Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese." *Developmental psychology* 24, no. 1 (1988): 14.

30. Mayo, Catherine, Vincent Aubanel, and Martin Cooke. "Effect of prosodic changes on speech intelligibility." In *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

31. Burnham, Denis, Sebastian Joeffry, and Lauren Rice. "Computer-and human-directed speech before and after correction." space 6 (2010): 7.

32. Burnham, Denis, Sebastian Joeffry, and Lauren Rice. "" Does-Not-Compute": Vowel Hyperarticulation in Speech to an Auditory-Visual Avatar." *Auditory-Visual Speech Processing* 2010. 2010.

33. Castellanos, Antonio, José-Miguel Benedí, and Francisco Casacuberta. "An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect." *Speech* Communication 20, no. 1-2 (1996): 23-35.

34. Garnier, Maëva, and Nathalie Henrich. "Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?" *Computer Speech & Language* 28, no. 2 (2014): 580-597.

35. Brunskog, Jonas, Anders Christian Gade, Gaspar Payá Bellester, and Lilian Reig Calbo. "Increase in voice level and speaker comfort in lecture rooms." *The Journal of the Acoustical Society of America* 125, no. 4 (2009): 2072-2082.

36. Michael, Deirdre D., Gerald M. Siegel, and Herbert L. Pick Jr. "Effects of distance on vocal intensity." *Journal of Speech, Language, and Hearing Research* 38, no. 5 (1995): 1176-1183.

37. Fux, Thibaut, Véronique Aubergé, Gang Feng, and Véronique Zimpfer. "Speaker's prosodic strategy for a large physical distance communication task." *Acoust. Soc. Am* 45, no. 1 (2012): 47-53.

38. Picheny, M. A., N. I. Durlach, and L. D. Braida. "Speaking clearly for the hard of hearing. II: Acoustic characteristics of clear and conversational speech." *Journal of speech and hearing research* 29, no. 4 (1986): 434-446.

39. Picheny, Michael A., Nathaniel I. Durlach, and Louis D. Braida. "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech." *Journal of Speech, Language, and Hearing Research* 28, no. 1 (1985): 96-103.

40. Payton, Karen L., Rosalie M. Uchanski, and Louis D. Braida. "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing." *The Journal of the Acoustical Society of America* 95, no. 3 (1994): 1581-1592.

41. Tuomainen, Outi, Valerie Hazan, Chris Davis, and Jeesun Kim. "Intelligibility of conversational and clear speech in young and older talkers as perceived by young and older listeners." *The Journal of the Acoustical Society of America* 146, no. 1 (2019): EL28-EL33.

42. Lu, Youyi, and Martin Cooke. "Speech production modifications produced by competing talkers, babble, and stationary noise." *The Journal of the Acoustical Society of America* 124, no. 5 (2008): 3261-3275.

43. Bradlow, Ann R., and Tessa Bent. "The clear speech effect for non-native listeners." *The Journal of the Acoustical Society of America* 112, no. 1 (2002): 272-284.

44. Cooke, Martin, Simon King, Maëva Garnier, and Vincent Aubanel. "*The listening talker: A review of human and algorithmic context-induced modifications of speech.*" Computer Speech & Language 28, no. 2 (2014): 543-571.

45. Smiljanic, Rajka. "Clear speech perception: Linguistic and cognitive benefits." *The handbook of speech perception* (2021): 177-205.

46. Ferguson, S. H. "Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners." *The Journal of the Acoustical Society of America* 116, no. 4 (2004): 2365-2373.

47. Studebaker, Gerald A. "A 'rationalized' arcsine transform." *Journal of Speech, Language, and Hearing Research* 28, no. 3 (1985): 455-462.

48. Sherbecoe, Robert L., and Gerald A. Studebaker. "Supplementary formulas and tables for calculating and interconverting speech recognition scores in transformed arcsine units." *International Journal of Audiology* 43, no. 8 (2004): 442-448.

49. Limb, Charles J., and Lawrence R. Lustig. "Hearing loss and the invention of the phonograph: The story of Thomas Alva Edison." *Otology & Neurotology* 23, no. 1 (2002): 96-101.

50. Kimizuka, Masanori. "Historical development of magnetic recording and tape recorder." *Survey reports on the systemization of technologies* 17 (2012): 185-273.

51. Gold, Bernard, and C. Rader. "The channel vocoder." *IEEE Transactions on Audio and Electroacoustics* 15, no. 4 (1967): 148-161.

52. Dudley, Homer. "Remaking speech." *The Journal of the Acoustical Society of America* 11, no. 2 (1939): 169-177.

53. Flanagan, James L., and Roger M. Golden. "Phase vocoder." *Bell System Technical Journal* 45, no. 9 (1966): 1493-1509.

54. Portnoff, Michael. "Implementation of the digital phase vocoder using the fast Fourier transform." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24, no. 3 (1976): 243-248.

55. Portnoff, Michael. "Time-scale modification of speech based on short-time Fourier analysis." *IEEE Transactions on Acoustics, Speech, and Signal Processing 29*, no. 3 (1981): 374-390.

56. Portnoff, M. R. "Mathematical framework for time-scale modification of speech signals." *The Journal of the Acoustical Society of America* 61, no. S1 (1977): S68-S68.

57. Goodwin, Michael M. "The STFT, sinusoidal models, and speech modification." *In Springer handbook of speech processing*, pp. 229-258. Springer, Berlin, Heidelberg, 2008.

58. Ellis, Dan. "A Phase Vocoder in Matlab." n.d. www.ee.columbia.edu. Accessed September 5, 2022. https://www.ee.columbia.edu/~dpwe/resources/matlab/pvoc/.

59. Baken, Ronald J., and Robert F. Orlikoff. Clinical measurement of speech and voice. Cengage Learning, 2000.

60. Pichora-Fuller, M. Kathleen. "Cognitive aging and auditory information processing." International journal of audiology 42, no. sup2 (2003): 26-32.

61. Humes, Larry E., and Judy R. Dubno. "Factors affecting speech understanding in older adults." The aging auditory system (2010): 211-257.

62. Rönnberg, Jerker, Mary Rudner, Catharina Foo, and Thomas Lunner. "Cognition counts: A working memory system for ease of language understanding (ELU)." *International journal of audiology* 47, no. sup2 (2008): S99-S105.

63. Lunner, Thomas, and Elisabet Sundewall-Thorén. "Interactions between cognition, compression, and listening conditions: Effects on speech-in-noise performance in a two-channel hearing aid." *Journal of the American Academy of Audiology* 18, no. 07 (2007): 604-617.

64. Griffiths, Roger. "Speech rate and NNS comprehension: A preliminary study in time-benefit analysis." Language Learning 40, no. 3 (1990): 311-336.

65. Higginbotham, D. Jeffery, Anne Drazek, Kim Kowarsky, Chris Scally, and Erwin Segal. "Discourse comprehension of synthetic speech delivered at normal and slow presentation rates." Augmentative and Alternative Communication 10, no. 3 (1994): 191-202.

66. Freud, Debora, Ruth Ezrati-Vinacour, and Ofer Amir. "Speech rate adjustment of adults during conversation." Journal of fluency disorders 57 (2018): 1-10.

67. Foo, Edwin W., and Gregory F. Hughes. "Hearing assistance system for providing consistent human speech." U.S. Patent 8,781,836, issued July 15, 2014.

68. Agnew, Jeremy, and Jeffrey M. Thornton. "Just noticeable and objectionable group delays in digital hearing aids." Journal of the American Academy of Audiology 11, no. 06 (2000): 330-336.

69. Stone, Michael A., and Brian CJ Moore. "Tolerable hearing aid delays. II. Estimation of limits imposed during speech production." Ear and Hearing 23, no. 4 (2002): 325-338.

70. Grant, Ken W., Virginie van Wassenhove, and David Poeppel. "Discrimination of auditory-visual synchrony." In AVSP 2003-International Conference on Audio-Visual Speech Processing. 2003.

71. Pardo, Jennifer S. "On phonetic convergence during conversational interaction." The Journal of the Acoustical Society of America 119, no. 4 (2006): 2382-2393.

72. Van Engen, Kristin J., Melissa Baese-Berk, Rachel E. Baker, Arim Choi, Midam Kim, and Ann R. Bradlow. "The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles." Language and speech 53, no. 4 (2010): 510-540.

73. Hazan, Valerie, and Baker, Rachel. 2020. “Photoshop Files of Diapixuk Picture Materials - Original English Version”. Zenodo. doi:10.5281/zenodo.3739053.

74. De Jong, Nivja H., and Ton Wempe. "Praat script to detect syllable nuclei and measure speech rate automatically." Behavior research methods 41, no. 2 (2009): 385-390.

75. Simantiraki, Olympia, and Martin Cooke. "Exploring Listeners' Speech Rate Preferences." INTERSPEECH. 2020.

76. Zhao, Yong. "The effects of listeners' control of speech rate on second language comprehension." Applied linguistics 18, no. 1 (1997): 49-68.

77. Fernald, Anne, and Thomas Simon. "Expanded intonation contours in mothers' speech to newborns." *Developmental psychology* 20, no. 1 (1984): 104.

78. Van de Weijer, Joost. "Language input to a prelingual infant." In *the GALA'97 Conference on Language Acquisition*, pp. 290-293. Edinburgh University Press, 1997.

79. Nejime, Yoshito, and Brian CJ Moore. "Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss." The Journal of the Acoustical Society of America 103, no. 1 (1998): 572-576.

80. Piquado, Tepring, Jonathan I. Benichov, Hiram Brownell, and Arthur Wingfield. "The hidden effect of hearing acuity on speech recall, and compensatory effects of self-paced listening." International Journal of Audiology 51, no. 8 (2012): 576-583.

81. McArdle, Rachel A., and Richard H. Wilson. "Homogeneity of the 18 QuickSIN™ lists." Journal of the American Academy of Audiology 17, no. 03 (2006): 157-167.

82. Wilson, Richard H., Rachel A. McArdle, and Sherri L. Smith. "An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss." (2007).

83. Rothauser, E. H. "IEEE recommended practice for speech quality measurements." IEEE Trans. on Audio and Electroacoustics 17 (1969): 225-246.

84. Tillman, T. W., and W. O. Olsen. "Speech audiometry." *Modern developments in audiology* 2 (1973): 37-74.

85. A Reading of the Gettysburg Address [Internet]. NPR.org. 2016 [cited 14 December 2016]. Available from: http://www.npr.org/templates/story/story.php?storyId=1512410

86. Killion, Mead C. "New thinking on hearing in noise: A generalized articulation index." In Seminars in Hearing, vol. 23, no. 01, pp. 057-076. Copyright© 2002 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA. Tel.:+ 1 (212) 584-4662, 2002.

87. Niquette, P., G. Gudmundsen, and M. Killion. "QuickSIN Speech-in-Noise Test Version 1.3." Elk Grove Village, IL: Etymotic Research (2001).

88. Cooke, Martin. "A glimpsing model of speech perception in noise." *The Journal of the Acoustical Society of America* 119, no. 3 (2006): 1562-1573.

89. Brown, Guy J., and Martin Cooke. "Computational auditory scene analysis." Computer Speech & Language 8, no. 4 (1994): 297-336.

90. Laroche, Jean, and Mark Dolson. "Phase-vocoder: About this phasiness business." In Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 4-pp. IEEE, 1997.

91. Bigi, Brigitte, and Daniel Hirst. "SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody." In Speech prosody, pp. 19-22. 2012.

92. Gygi, Brian, and Valeriy Shafiro. "Spatial and temporal modifications of multitalker speech can improve speech perception in older adults." Hearing research 310 (2014): 76-86.

93. Laroche, Jean, and Mark Dolson. "Phase-vocoder: About this phasiness business." In Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 4-pp. IEEE, 1997.

94. Uchanski, Rosalie M., Sunkyung S. Choi, Louis D. Braida, Charlotte M. Reed, and Nathaniel I. Durlach. "Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate." Journal of Speech, Language, and Hearing Research 39, no. 3 (1996): 494-509.

95. Davis, Steven, and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE transactions on acoustics, speech, and signal processing 28, no. 4 (1980): 357-366.

96. Meskill, Carla, and Jonathan Mossop. "Technologies use with ESL learners in New York State: Preliminary report." Journal of Educational Computing Research 22, no. 3 (2000): 265-284.

97. Feyten, Carine M. "The power of listening ability: An overlooked dimension in language acquisition." The modern language journal 75, no. 2 (1991): 173-180.

98. Vandergrift, Larry. "1. Listening to learn or learning to listen?." Annual review of applied linguistics 24 (2004): 3-25.

99. Bradlow, Ann R., and Jennifer A. Alexander. "Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners." The Journal of the Acoustical Society of America 121, no. 4 (2007): 2339-2349.

100. Foulke, Emerson. "Listening comprehension as a function of word rate." Journal of Communication 18, no. 3 (1968): 198-206.

101. Foulke, Emerson, and Thomas G. Sticht. "Review of research on the intelligibility and comprehension of accelerated speech." Psychological bulletin 72, no. 1 (1969): 50.

102. Carver, Ronald P. "Effects of increasing the rate of speech presentation upon comprehension." Journal of Educational Psychology 65, no. 1 (1973): 118.

103. Blau, Eileen K. "The effect of syntax, speed, and pauses on listening comprehension." TESOL quarterly 24, no. 4 (1990): 746-753.

104. Graves, Alex, Santiago Fernández, and Jürgen Schmidhuber. "Bidirectional LSTM networks for improved phoneme classification and recognition." In International conference on artificial neural networks, pp. 799-804. Springer, Berlin, Heidelberg, 2005.

105. Disc, NIST Speech, John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, and Nancy L. Dahlgren. "Acoustic-Phonetic Continuous Speech Corpus."

106. Davis, Steven, and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE transactions on acoustics, speech, and signal processing 28, no. 4 (1980): 357-366.

107. Oh, Donghoon, Jeong-Sik Park, Ji-Hwan Kim, and Gil-Jin Jang. "Hierarchical phoneme classification for improved speech recognition." Applied Sciences 11, no. 1 (2021): 428.

108. Panfili, L. M., J. Haywood, D. R. McCloy, P. E. Souza, and R. A. Wright. "The UW/NU corpus, version 2.0." (2017).

109. C. J. Steinmetz and J. Reiss, “pyloudnorm: A simple yet flexible loudness meter in Python”, in *Audio Engineering Society Convention 150*, 2021.

110. A. Davies, “17 the native speaker in applied linguistics”, *The handbook of applied linguistics*, bl 431, 2004.

111. K. B. Sheehan and M. Pittman, Amazon’s Mechanical Turk for academics: The HIT handbook for social science research. Melvin & Leigh, Publishers, 2016.

112. Bapineedu, G. "Analysis of Lombard effect speech and its application in speaker verification for imposter detection." PhD diss., MS thesis, Language Technologies Research Centre, International Institute of Information Technology, Hyderabad, India, 2010.

113. Stanton, Bill J., George D. Allen, and Leah H. Jamieson. "Acoustic-phonetic analysis of normal, loud, and Lombard speech in simulated cockpit conditions." The Journal of the Acoustical Society of America 84, no. S1 (1988): S115-S115.

114. J. H. L. Hansen, “Analysis and compensation of stressed and noisy speech with application to robust automatic recognition”, Georgia Institute of Technology, 1988.

115. Smiljanic, Rajka, and Ann R. Bradlow. "Stability of temporal contrasts across speaking styles in English and Croatian." Journal of Phonetics 36, no. 1 (2008): 91-113.

116. Bosker, Hans Rutger, Eva Reinisch, and Matthias J. Sjerps. "Cognitive load makes speech sound fast, but does not modulate acoustic context effects." Journal of Memory and Language 94 (2017): 166-176.

117. Bosker, Hans Rutger, and Eva Reinisch. "Foreign languages sound fast: Evidence from implicit rate normalization." Frontiers in Psychology 8 (2017): 1063.

118. Pichora-Fuller, M. Kathleen, and Gurjit Singh. "Effects of age on auditory and cognitive processing: implications for hearing aid fitting and audiologic rehabilitation." Trends in amplification 10, no. 1 (2006): 29-59.

119. Janse, Esther. "Processing of fast speech by elderly listeners." The Journal of the Acoustical Society of America 125, no. 4 (2009): 2361-2373.

120. Govender, Avashna, Anita E. Wagner, and Simon King. "Using Pupil Dilation to Measure Cognitive Load When Listening to Text-to-Speech in Quiet and in Noise." In INTERSPEECH, pp. 1551-1555. 2019.

121. Huang, Tzu-hsien, Jheng-hao Lin, and Hung-yi Lee. "How far are we from robust voice conversion: A survey." In 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 514-521. IEEE, 2021.

122. Kameoka, Hirokazu, Kou Tanaka, Damian Kwaśny, Takuhiro Kaneko, and Nobukatsu Hojo. "ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 1849-1863.

123. Kominek, John, and Alan W. Black. "The CMU Arctic speech databases." In Fifth ISCA workshop on speech synthesis. 2004.