Exploration of Non-Stationary Speech Protection for Highly Intelligible Time-Scale Compression

Panagiotis Pantalos

Thesis submitted in partial fulfillment of the requirements for the

Masters' of Science degree in Computer Science and Engineering

University of Crete
School of Sciences and Engineering
Computer Science Department
Voutes University Campus, 700 13 Heraklion, Crete, Greece

Thesis Advisor: Prof. *Yannis Stylianou*

Thesis Co-Advisor: Dr. *George Kafentzis*

This work has been performed at the University of Crete, School of Sciences and Engineering, Computer Science Department.

The work has been supported by the Foundation for Research and Technology - Hellas (FORTH), Institute of Applied and Computational Mathematics (IACM).

University of Crete
Computer Science Department

Exploration of Non-Stationary Speech Protection for Highly Intelligible Time-Scale Compression

Thesis submitted by
Panagiotis Pantalos
in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science

THESIS APPROVAL

Author: ______________________
Panagiotis Pantalos

Committee approvals: ______________________
Yannis Stylianou
Professor, Thesis Supervisor

____
Yannis Pantazis
Principal Researcher, Committee Member

____
Grigorios Tsagkatakis
Assistant Professor, Committee Member

Departmental approval: ______________________
Polyvios Pratikakis
Associate Professor, Director of Graduate Studies

Heraklion, November 2023

Exploration of Non-Stationary Speech Protection for Highly Intelligible Time-Scale Compression.

Abstract

Speech recordings are everywhere, from social media, YouTube, and online learning to podcasts and audiobooks. In today's fast-paced world, it is sometimes necessary to speed up speech recordings in order to promote faster information consumption. A population group that benefits the most from such technologies is visually impaired individuals who employ screen reading on their mobile phones. A series of algorithms have been developed for the time-scale expansion or compression of speech recordings. It is well known that fast speech, also known as time-scale compressed speech, is less intelligible due to a loss of speech parts that are important in distinguishing syllables and words. The majority of these parts are non-stationary in nature, such as transient sounds, plosives, and fricatives.

In this work, we investigate algorithms for non-stationary speech protection in order to provide highly intelligible time-scale compression. We base our experiments on the so-called Waveform Similarity Overlap-and-Add (WSOLA) method of time-scale compression. WSOLA is capable of providing both uniform and non-uniform time-scale compression. We propose to characterize speech waveforms according to their non-stationarity using simple time and frequency domain criteria. Utilizing a frame-by-frame analysis, the first criterion (C1) is based on the RMS energy of each frame. Additionally, we implement a Line Spectral Frequency (LSF)-based criterion, named C2, and in combination with C1, we end up with a hybrid non-stationarity detection criterion named C3. C1 and C3 are implemented on dataset of Greek speech recordings named GrHarvard. The latter consists of 720 sentences from both genders that form 72 phonemically balanced lists of 10 sentences each.

Intelligibility and preference experiments were performed on four of the GrHarvard lists involving both sighted and visually impaired individuals. Subsequently, a statistical analysis was carried out to assess the significance of the differences in the results obtained from both experiments' tests. In the first experiment, we conducted a comparative analysis involving uniform WSOLA, non-uniform C1-based WSOLA, and non-uniform C3-based WSOLA. The principal objective was to assess whether the incorporation of protective measures had a positive or negative impact on the intelligibility of speech signals. The findings consistently demonstrated that C1-based WSOLA outperformed the others in both intelligibility and user preference. It was followed by C3-based WSOLA, with uniform WSOLA ranking last. In this experiment, characterized by substantial differences, the majority of observed variations were found to be statistically significant. In the second experiment, our objective was to assess the same three methods under equal words per minute (WPM) conditions. This made it challenging for users to distinguish between different methods and resulted in more uniform outcomes. Differences primarily stemmed from variations within the signals, related to the sizes of their stationary and non-stationary parts. Even though the C1-based method tended to achieve the highest intelligibility (in most cases except at 0.25), it remained challenging to definitively determine which method was superior in both preference and intelligibility tests. Yet, despite our initial expectations of better performance in the results of the visually impaired group compared to the control group, such variations did not materialize, mainly due to the limited number of visually impaired participants willing to participate in our tests. Consequently, all of these challenges led the majority of observed results not to attain statistical significance, even though a discernible pattern was occasionally evident among the methods.

Future work may include further parameter tuning of the stationarity detection algorithm. As an example, different lengths of analysis and hop frames can be used, as well as pitch-synchronous analysis in stationary parts of speech. Furthermore, the base

method used for time-scale compression can be replaced by other more complex models for time-scale compression (such as the Harmonic+Noise model). Finally, further experiments - including a larger sample of visually impaired people - could strengthen statistical conclusions about the performance of each method.

Εξερεύνηση προστασίας μη στάσιμου λόγου για υψηλής καταληπτότητας συμπίεση σε χρονική κλίμακα.

Περίληψη

Ηχογραφήσεις ομιλίας υπάρχουν παντού, από τα μέσα κοινωνικής δικτύωσης, το YouTube και την ηλεκτρονική εκπαίδευση μέχρι τα podcast και τα ηχητικά βιβλία. Στον σημερινό κόσμο με τους γρήγορους ρυθμούς, μερικές φορές είναι απαραίτητο να επιταχυνθούν, προκειμένου να προωθηθεί ταχύτερη κατανάλωση πληροφοριών από τους χρήστες. Μια ομάδα πληθυσμού που επωφελείται περισσότερο από τέτοιες τεχνολογίες είναι τα άτομα με προβλήματα όρασης που χρησιμοποιούν την ανάγνωση οθόνης στα κινητά τους τηλέφωνα. Έχει αναπτυχθεί μια σειρά αλγορίθμων για τη χρονική επέκταση ή συμπίεση των καταγραφών ομιλίας. Είναι γνωστό ότι η γρήγορη ομιλία, γνωστή και ως συμπιεσμένη σε χρονική κλίμακα ομιλία, είναι λιγότερο κατανοητή λόγω της απώλειας τμημάτων ομιλίας που είναι σημαντικά για τη διάκριση συλλαβών και λέξεων. Η πλειονότητα αυτών των τμημάτων είναι μη στάσιμα στη φύση τους, όπως οι μεταβατικοί ήχοι, οι έκκροτοι φθόγγοι και τα τριβόμενα σύμφωνα.

Στην παρούσα εργασία, διερευνούμε αλγορίθμους για την προστασία της μη στάσιμης ομιλίας, προκειμένου να παρέχουμε συμπίεση με υψηλή καταληπτότητα σε χρονική κλίμακα. Βασίζουμε τα πειράματά μας στη λεγόμενη μέθοδο συμπίεσης χρονικής κλίμακας Waveform Similarity Overlap-and-Add (WSOLA). Η WSOLA είναι ικανή να παρέχει τόσο ομοιόμορφη όσο και ανομοιόμορφη συμπίεση χρονικής κλίμακας. Προτείνουμε να χαρακτηρίσουμε τις κυματομορφές ομιλίας ανάλογα με τη μη-στασιμότητά τους χρησιμοποιώντας απλά κριτήρια στο πεδίο του χρόνου και της συχνότητας. Χρησιμοποιώντας μια ανάλυση καρέ-προς-καρέ, το πρώτο κριτήριο (C1) βασίζεται στην ενέργεια RMS κάθε καρέ. Επιπλέον, εφαρμόζουμε ένα κριτήριο που βασίζεται στη φασματική συχνότητα γραμμής (LSF), το οποίο ονομάζεται C2, και σε συνδυασμό με το C1 καταλήγουμε σε ένα υβριδικό κριτήριο ανίχνευσης μη στασιμότητας που ονομάζεται C3. Το C1 και το C3 εφαρμόζονται σε σύνολο δεδομένων από ηχογραφήσεις ελληνικής ομιλίας με την ονομασία GrHarvard. Η τελευταία αποτελείται από 720 προτάσεις και από τα δύο φύλα που σχηματίζουν 72 φωνητικά ισορροπημένες λίστες των 10 προτάσεων η καθεμία.

Πραγματοποιήθηκαν πειράματα καταληπτότητας και προτίμησης σε τέσσερις από τις λίστες του GrHarvard, στα οποία συμμετείχαν και άτομα με υγιή όραση και κάποια άτομα με προβλήματα όρασης. Στη συνέχεια, διενεργήθηκε στατιστική ανάλυση για να εκτιμηθεί η σημασία των διαφορών στα αποτελέσματα που προέκυψαν από τα δύο πειράματα. Στο πρώτο πείραμα, πραγματοποιήσαμε μια συγκριτική ανάλυση που αφορούσε την ομοιόμορφη WSOLA, τη μη ομοιόμορφη WSOLA με βάση το C1 και τη μη ομοιόμορφη WSOLA με βάση το C3. Ο κύριος στόχος ήταν να εκτιμηθεί κατά πόσον η ενσωμάτωση προστατευτικών μέτρων είχε θετικό ή αρνητικό αντίκτυπο στην καταληπτότητα των σημάτων ομιλίας. Τα ευρήματα έδειξαν σταθερά ότι η WSOLA με βάση το C1 υπερείχε των άλλων τόσο στην καταληπτότητα όσο και στην προτίμηση των χρηστών. Μετά, ακολουθούσε η WSOLA με βάση το C3, με την ομοιόμορφη WSOLA να κατατάσσεται τελευταία. Σε αυτό το πείραμα, η πλειονότητα των παρατηρούμενων διακυμάνσεων βρέθηκε να είναι στατιστικά σημαντική.

Στο δεύτερο πείραμα, ο στόχος μας ήταν να αξιολογήσουμε τις ίδιες τρεις μεθόδους υπό ίσες συνθήκες λέξεων ανά λεπτό (WPM). Αυτό έκανε δύσκολο για τους χρήστες να διακρίνουν μεταξύ των διαφορετικών μεθόδων και οδήγησε σε πιο ομοιόμορφα αποτελέσματα. Οι διαφορές προέκυπταν κυρίως από τις διαφοροποιήσεις εντός των σημάτων, που σχετίζονταν με τα μεγέθη των σταθερών και μη σταθερών τμημάτων τους. Παρόλο που η μέθοδος με βάση το C1 έτεινε να επιτυγχάνει την υψηλότερη καταληπτότητα (στις περισσότερες περιπτώσεις εκτός από την περίπτωση των $0,25$), παρέμεινε δύσκολο να προσδιοριστεί οριστικά ποια μέθοδος ήταν ανώτερη ως προς την καταληπτότητα των δειγμάτων και την προτίμηση των χρηστών. Επίσης, παρά τις αρχικές μας προσδοκίες για καλύτερες επιδόσεις στα αποτελέσματα της ομάδας με

προβλήματα όρασης σε σύγκριση με την ομάδα των ατόμων με υγιή όραση, τέτοιες διαφοροποιήσεις δεν υπήρξαν, κυρίως λόγω του περιορισμένου αριθμού συμμετεχόντων με προβλήματα όρασης που ήταν πρόθυμοι να συμμετάσχουν στις δοκιμές μας. Κατά συνέπεια, όλες αυτές οι προκλήσεις οδήγησαν την πλειονότητα των παρατηρούμενων αποτελεσμάτων να μην επιτύχουν στατιστική σημαντικότητα, παρόλο που περιστασιακά ήταν εμφανές ένα διακριτό μοτίβο μεταξύ των μεθόδων.

Σαν μελλοντική εργασία, μπορεί να συμπεριληφθεί περαιτέρω ρύθμιση των παραμέτρων του αλγορίθμου ανίχνευσης στασιμότητας. Για παράδειγμα, μπορούν να χρησιμοποιηθούν διαφορετικά μήκη πλαισίων ανάλυσης και άλματος, καθώς και συγχρονική ανάλυση του τονικού ύψους σε στάσιμα μέρη του λόγου. Επιπλέον, η βασική μέθοδος που χρησιμοποιείται για τη συμπίεση χρονικής κλίμακας μπορεί να αντικατασταθεί από άλλα πιο σύνθετα μοντέλα για τη συμπίεση χρονικής κλίμακας (όπως το μοντέλο Harmonic+Noise). Τέλος, περαιτέρω πειράματα - συμπεριλαμβανομένου ενός μεγαλύτερου δείγματος ατόμων με προβλήματα όρασης - θα μπορούσαν να ενισχύσουν τα στατιστικά συμπεράσματα σχετικά με την απόδοση κάθε μεθόδου.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Yannis Stylianou, whose guidance and unwavering support have been instrumental throughout this journey. His expertise and insightful feedback have truly shaped the trajectory of my work. I am also indebted to my co-supervisor, George Kafentzis, for his continuous contributions and mentorship. His collaborative spirit and constructive criticism have been invaluable in refining the nuances of my research. Also, a big thanks for his help in conducting the listening tests at the Blind People Associations of Kalamata and Ioannina. Big thanks to all the committee members for accepting their position, with a special shoutout to Yannis Pantazis for also helping me secure a scholarship from IACM. Finally, I would also like to thank the blind associations of Heraklion and Kalamata, and the volunteers who offered to perform all the listening tests.

Beyond the academic realm, I extend heartfelt thanks to my family for their unending love, encouragement, and understanding and to my friends who have been a constant source of motivation and joy. Thank you for standing by me during the challenges and celebrations alike, for believing in me, and for being a vital part of this journey.

Ευχαριστίες

Θα ήθελα να εκφράσω τη βαθύτατη ευγνωμοσύνη μου στον επιβλέποντα καθηγητή μου, Γιάννη Στυλιανού, του οποίου η καθοδήγηση και η αμέριστη υποστήριξη υπήρξαν καθοριστικές καθ΄ όλη τη διάρκεια αυτού του ταξιδιού. Η εμπειρογνωμοσύνη του και η διορατική ανατροφοδότηση διαμόρφωσαν πραγματικά την πορεία της εργασίας μου. Είμαι επίσης υπόχρεος στον συνεπιβλέποντα καθηγητή μου, Γιώργο Καφεντζή, για τη συνεχή συμβολή και καθοδήγησή του. Το συνεργατικό του πνεύμα και η εποικοδομητική κριτική του ήταν ανεκτίμητα πολύτιμα για την τελειοποίηση των αποχρώσεων της έρευνάς μου. Επίσης, ένα μεγάλο ευχαριστώ για τη βοήθειά του στη διεξαγωγή των δοκιμών ακρόασης στους Συλλόγους Τυφλών της Καλαμάτας και των Ιωαννίνων. Ένα μεγάλο ευχαριστώ σε όλα τα μέλη της επιτροπής για την αποδοχή της θέσης τους, με μια ιδιαίτερη μνεία στον Γιάννη Πανταζή που με βοήθησε επίσης να εξασφαλίσω μια υποτροφία από το IACM. Θα ήθελα τελος να ευχαριστήσω τους συλλόγους τυφλών Ηρακλείου, και Καλαμάτας, καθώς και τους εθελοντές που προσφέρθηκαν να πραγματοποιήσουν τα τεστ ακρόασης.

Πέρα από το ακαδημαϊκό πεδίο, ευχαριστώ από καρδιάς την οικογένειά μου για την ατελείωτη αγάπη, ενθάρρυνση και κατανόηση και τους φίλους μου που αποτελούν συνεχή πηγή κινήτρων και χαράς. Σας ευχαριστώ που μου συμπαρασταθήκατε τόσο στις προκλήσεις όσο και στις χαρές, που πιστεύατε σε μένα και που υπήρξατε ζωτικό μέρος αυτού του ταξιδιού.

στην οικογένεια μου

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Speech Processing

The primary objective of speech is to simplify communication, which involves sharing messages. According to Shannon's information theory [5], a message, represented as a sequence of discrete symbols, can be quantified in terms of its information content in bits. The rate at which information is transmitted is measured in bits per second (bps). In the domain of speech production, as well as in many human-designed electronic communication systems, the information to be communicated is encoded in the form of a continuously changing (analog) waveform. This waveform can be transmitted, recorded, modified, and ultimately decoded by a human listener.

In the case of speech, the fundamental analog representation of the message is an acoustic waveform, referred to as the speech signal. Speech signals can be transformed into an electrical waveform using a microphone. They can then be further processed through a combination of analog and digital signal processing techniques. Subsequently, the processed signal can be converted back into an acoustic form using a loudspeaker, a telephone handset, or headphones, as desired.

1.2 The Speech Chain

Figure 1.1 shows the comprehensive sequence of generating and comprehending speech, starting with the formation of a message within the speaker's brain, then the creation of the speech signal, and concluding with the interpretation of the message by a listener. In their introduction to the study of speech, Denes and Pinson fittingly coined this progression as the "speech chain" [6]. The procedure initiates in the upper left corner with a message somehow represented in the speaker's brain. Throughout the process of speech production (as indicated in the upper path in Figure 1.1), the message information can be conceived as undergoing various representations.

The complete speech chain contains both a speech production/generation model, as discussed earlier, and a speech perception/recognition model, as depicted in the lower part of Figure 1.1. The speech perception model illustrates a series of stages from capturing speech at the ear to comprehending the message encoded in the speech signal. The initial step involves effectively converting the acoustic waveform into a spectral representation. This transformation occurs within the inner ear, primarily through the basilar membrane, which functions as a non-uniform spectrum analyzer. It spatially separates the spectral components of the incoming speech signal and, in the process, analyzes them by means of a non-uniform filter bank. Subsequently, the speech perception process entails a neural transduction of these spectral features into a set of sound features (often referred to as

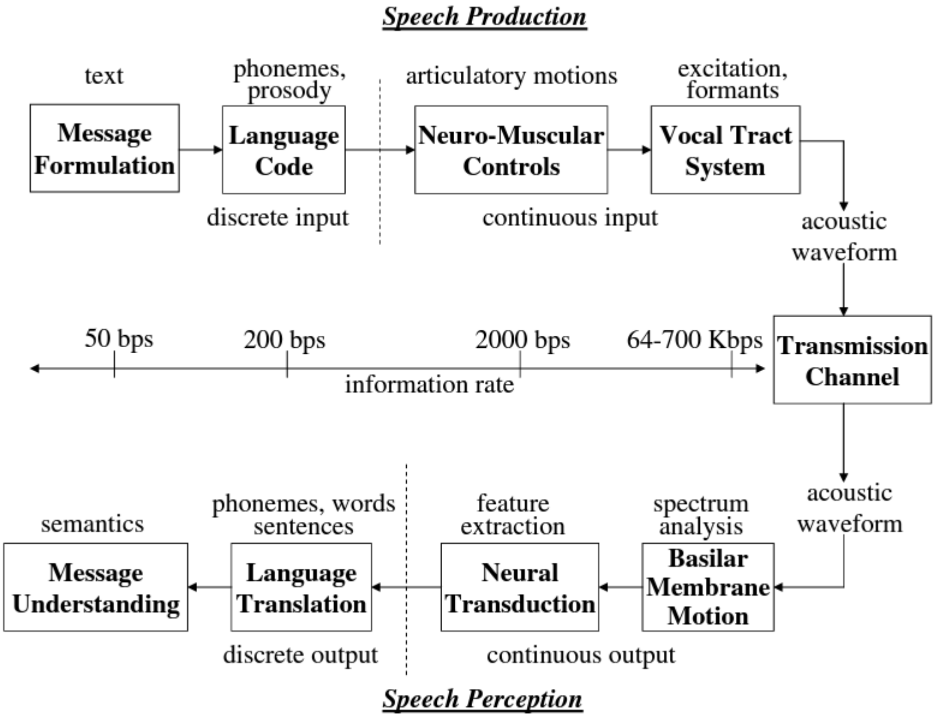


Figure 1.1: *The Speech Chain (from [2]).*

distinctive features in linguistics) that can be decoded and processed by the brain. The following step involves converting these sound features into the set of phonemes, words, and sentences associated with the incoming message through a language translation process within the human brain. Finally, the last phase in the speech perception model involves the conversion of the phonemes, words, and sentences of the message into an understanding of the underlying message's meaning, enabling the listener to respond appropriately or take suitable action. Our fundamental comprehension of the processes within most of the speech perception modules in Figure 1.1 remains rather basic. However, it is generally accepted that some physical correlate of each step in the speech perception model occurs within the human brain. Hence, the entire model serves as a valuable framework for contemplating the processes involved.

In most digital speech processing applications, the initial step involves converting the acoustic waveform into a sequence of numerical values. Modern A-to-D converters typically operate by capturing data at a very high rate, applying a digital low-pass filter with a cutoff set to preserve a predefined bandwidth, and then reducing the sampling rate to the desired level, which can be as low as twice the cutoff frequency of the sharp-cutoff digital filter. This discrete-time representation serves as the foundation for most applications. From this point, various other representations are derived through digital processing. For the most part, these alternative representations are constructed by incorporating knowledge about the inner workings of the speech chain, as depicted in Figure 1.1. As we will observe, it's feasible to include elements from both the speech production and speech perception processes into the digital representation and processing. One can assert that digital speech processing revolves around a set of techniques aimed at reducing the data rate of the speech representation, either along the upper or lower path in Figure 1.1.

1.3 Speech Perception

One of the most important aspects of the Speech Chain is speech perception (SP) [7]. SP is a multifaceted process that can be understood from the perspective of speech signal processing, highlighting the remarkable capabilities of the human auditory system. When we speak, we produce complex acoustic signals that are transmitted through the air as pressure waves. These sound waves are received by the ears, where the external ear captures and funnels them into the ear canal. Within the ear, the middle ear amplifies and transmits these acoustic signals to the cochlea, a spiral-shaped, fluid-filled structure in the inner ear. The cochlea is a critical element in the signal processing chain. It transforms the analog acoustic signal into a digital neural code through hair cells that respond to different frequencies of sound. This transformation is essential for subsequent neural processing [8].

Once the auditory signals are converted into electrical impulses, they travel along the auditory nerve to the brain. This transmission stage corresponds to the initial analog-to-digital conversion in signal processing. In the brain, the signals experience further analysis, starting with spectral and temporal decomposition. Spectral analysis dissects the signal into its constituent frequencies, while temporal analysis tracks how these frequencies change over time. This process is critical for distinguishing between different speech sounds (phonemes) since different phonemes are characterized by specific frequency patterns and temporal sequences. The brain's ability to recognize and discriminate these patterns is vital for speech perception [9].

Another crucial aspect of speech perception is the recognition of phonemes and words. The brain categorizes the speech sounds it processes into familiar linguistic units, facilitating the comprehension of spoken language. This categorization requires pattern recognition, a fundamental concept in signal processing. In addition to identifying individual phonemes, the brain integrates information from preceding and following phonemes, enabling us to understand the coarticulatory cues that occur in connected speech. Speech segmentation, the process of identifying word boundaries in continuous speech, is another signal-processing task that the brain accomplishes efficiently. Moreover, in challenging listening conditions with background noise or distortion, the brain exhibits phonemic restoration capabilities, filling in missing speech sounds based on contextual information, which demonstrates the brain's role as a sophisticated signal processor that can compensate for degraded input.

1.4 Speech Intelligibility and Intelligibility Enhancement

Speech intelligibility is a critical aspect of human communication and language processing [10]. It refers to the degree to which spoken language can be accurately and readily understood by a listener. The assessment of speech intelligibility involves factors such as the clarity of articulation, pronunciation, and the distinctiveness of phonemes, words, and sentences. Various factors influence speech intelligibility, including the acoustic environment, the speaker's articulatory precision, and the listener's familiarity with the language and the specific accent or dialect [11]. Researchers in the field of speech and hearing science employ various methods to quantify and enhance speech intelligibility, from subjective perceptual evaluations to objective acoustic measurements [12, 13]. Understanding and improving speech intelligibility is important for numerous applications, including the design of effective communication systems [14], hearing aid technology [15, 16], speech therapy [17], and the development of robust automatic speech recognition systems [18]. Moreover, it plays a crucial role in ensuring effective interpersonal communication and information dissemination, particularly in challenging acoustic environments [19, 20] or for individuals with hearing impairments [21].

1.5 Time-Scale Compression of Speech

Speeding up speech recordings, or time-scale compression [22, 23], offers a range of essential applications across diverse fields. In the domain of digital content creation, this technique is often used to create engaging and informative video or audio content [24]. Content creators on platforms like YouTube and social media frequently employ time compression to change lengthy explanations, tutorials, or presentations into something more digestible and dynamic. This not only helps maintain viewer engagement but also caters to audiences with shorter attention spans in the digital age.

In the field of transcription services, time-scale compression is a valuable tool for professionals who need to convert spoken content into text efficiently. This approach enables transcriptionists to work more quickly, reducing the time and costs associated with transcribing lengthy interviews, focus groups, or recorded meetings. Moreover, for researchers conducting qualitative analysis, time-scale-compressed transcriptions can expedite data processing and allow for more rapid insights and conclusions.

Education is another domain where time-scale compression has found crucial applications [25]. Instructors often use time-compressed recordings of lectures, webinars, or training sessions to create concise study materials for students. These materials can be instrumental in helping learners review key concepts, save time, and prepare for exams more efficiently. Moreover, in online learning platforms, educators can offer accelerated versions of recorded classes, making it easier for students to navigate content at their own pace and complete courses more rapidly.

In the field of assistive technology, time-scale compression serves as a valuable tool for individuals with various cognitive disabilities, such as auditory processing disorders or slow auditory processing [26]. By changing the speed of audio content, individuals with these conditions can access information more efficiently, reduce the cognitive load associated with slower listening, and enhance comprehension. This application of time compression fosters accessibility and inclusion by ensuring that individuals with cognitive disabilities have equal access to spoken information and educational resources.

Moreover, time compression plays a role in audiobook production, where narrators or publishers may offer audiobooks in both regular and time-scale-compressed versions [27]. This allows readers to choose the listening speed that suits their preferences and time constraints, making audiobooks more versatile and accommodating for a wide range of listeners. Overall, the applications of speeding up speech recordings are diverse, addressing the needs of content creators, transcriptionists, educators, individuals with cognitive disabilities, and audiobook enthusiasts, enhancing the efficiency and accessibility of spoken content in various contexts [28].

Lastly, visually impaired individuals, in contrast to their sighted counterparts, exhibit the capacity to listen to speech recordings at accelerated rates depending on their knowledge of the subject, auditory processing skills, and personal preferences [29]. The heightened reliance on auditory cues for communication and information acquisition among the visually impaired population often results in an enhanced ability to process speech at increased speeds. Assistive technologies and screen reading software further give control over the playback speed of digital content, enabling efficient consumption of spoken material, which is particularly beneficial for visually impaired students and professionals.

1.6 Thesis Motivation

The core aim of this thesis was to identify methods that generate samples at accelerated speeds with a notably high level of intelligibility. To achieve this, three distinct time-compression techniques were implemented and rigorously evaluated. The first one was uniform WSOLA [30] while the other two were two non-uniform extensions of it. These methods were tested within two distinct groups of people, individuals with visual impairments and those with normal vision, through intelligibility and preference tests in high-speed samples. The primary objective was to identify which algorithm performed better in these metrics for most of the tested speedups. For individuals with visual impairments, whose primary source of information acquisition is auditory, ensuring that the recordings they engage with on a daily basis are both intelligible and comprehensible holds great significance. Simultaneously, for individuals with normal vision, the ability to utilize high-speed, intelligible speech when listening to audiobooks or podcasts can be a valuable time-saving feature. Finally, it is crucial to recognize the limitations tied to the perceptual boundaries of auditory processing, which apply to both visually impaired and sighted individuals. Excessively speeding up speech can reduce comprehensibility, presenting a challenge. Hence, it's essential to strike a balance between enhancing speed for efficiency and maintaining content intelligibility.

1.7 Thesis Organization

This thesis is organized as follows: Chapter 2 presents the relevant literature to this work. Chapter 3 describes the dataset used in our experiments. Chapter 4 presents the methodology of our work. Finally, Chapters 5 and 6 discuss our results and propose future research directions, respectively.

Chapter 2

Related Work

In this Chapter we will discuss methods for time-scale modification of speech signals, mostly based on the time-domain. Also, the following methods do not rely on speech modeling, that is, there is no underlying model that is fitted on the speech waveform in order to manipulate its parameters to achieve time-scaling. On the contrary, time-scaling is performed using the raw speech sample data.

2.1 Time-Scale Modification

2.1.1 Definition

Time-scale modification (TSM) is the process of changing the playback speed of an audio signal, making it faster or slower, without altering its pitch.

2.1.2 Challenges

TSM encounters significant challenges with speech signals due to their complexity. Maintaining aspects like pitch, timbre, tone, and crispness is essential. To preserve these, various TSM techniques are used. Conventional methods such as WSOLA or PV-TSM excel at handling harmonic sounds but can pose challenges when applied to percussive sounds. However, by combining different TSM methods, we can reduce created artifacts.

In a study [31], the audio was initially divided into harmonic and percussive components. Subsequently, an appropriate TSM method was applied to each component to preserve its distinct characteristics. The final output was generated by merging these two processed signals.

2.1.3 Algorithms

Numerous algorithms are used for this task, with some being standalone solutions, while others build upon or extend existing methods. Methods will be presented in the next sections.

2.1.3.1 Generalized processing pipeline of TSM procedures

The most straightforward algorithm to achieve this goal is illustrated in Figure 2.1. The main idea involves breaking down the input signal into short segments (frames) with a fixed length between 50 and 100 milliseconds. Each frame contains the local pitch characteristics of the signal. These frames are then temporally adjusted to achieve the desired time-scale modification while preserving the original pitch information.

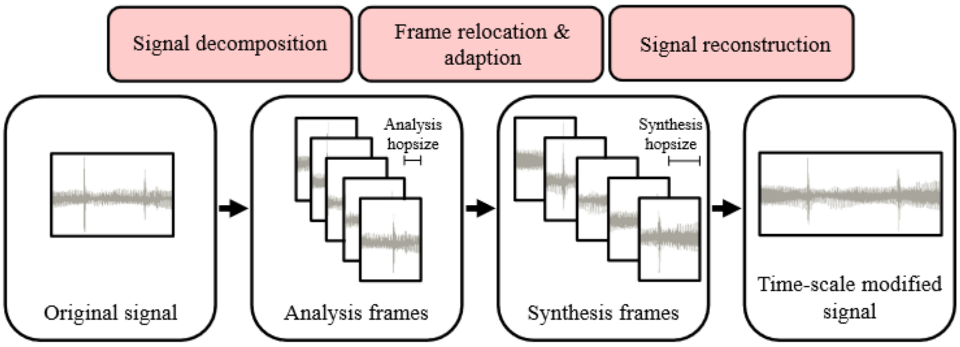


Figure 2.1: *Time-scale modification processing pipeline procedure (from [3]).*

The first step of this method is to split the input signal into fixed-length (equal to N) short frames $x_m, m \in \mathbb{Z}$ that are spaced by an analysis hop size H_a. In the second step, these frames are relocated temporally based on a specified synthesis hop size H_s. The time scale modification of the signal is performed in this step with the relocation of the frames. The stretching factor α is defined as the fraction of the synthesis hop size H_s with the analysis hop size H_a.

$$\alpha = \frac{H_s}{H_a} \tag{2.1}$$

So:

- $H_s > H_a$ and $\alpha > 1$ means signal stretching
- $H_s < H_a$ and $0 \leq \alpha < 1$ means signal shortening

The synthesis hop size Hs can have a fixed value like N/2 or N/4 because it is fine if there is an overlap on the relocated frames. In each case, the analysis hop size H_a can be calculated as:

$$H_a = H_s/\alpha. \tag{2.2}$$

As the final step, we need to define the relation between analysis frames x_m and synthesis frames y_m. Analysis frames could be used directly as synthesis frames. However, directly overlaying the overlapping relocated frames can result in unwanted artifacts like phase discontinuities at frame boundaries and amplitude differences. In the next two paragraphs, we will dive more into the problems that are visualized in Figure 2.2.

Problems At first, during the reconstruction of the output signal, the resulting waveform often has discontinuities on the boundaries of the (unmodified) synthesis frames. These discontinuities can be perceived as clicking sounds.

Secondly, the synthesis hop size (H_s) is typically selected to ensure overlap in the synthesis frames. When overlaying the (unmodified) synthesis frames, each with the same amplitude as the input signal, this often results in an undesirable increase in the output signal's amplitude.

In order to solve these issues, the primary question is how to process the analysis frames (x_m) to create the synthesis frames (y_m). There are multiple approaches to answer this question, resulting in conceptually different TSM procedures.

2.1.3.2 Overlap-Add method (OLA)

One such approach is Overlap-and-Add (OLA). OLA is an iterative method that tries to ensure a smooth transition between frames and to mitigate unintended amplitude fluctuations. The concept of this method involves the application of a window function, usually a

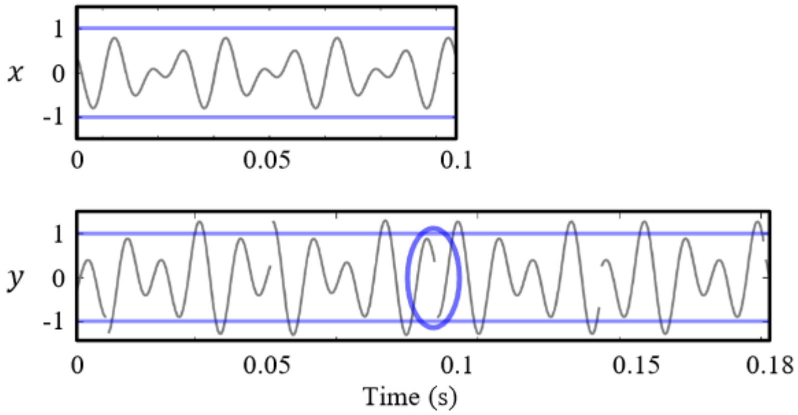


Figure 2.2: *Problems that occur when we select synthesis frames y_m to be the same as analysis frames x_m (with zero processing). You can see discontinuities (oval) and amplitude fluctuations (lines) - extracted from [3].*

Hann window, to the analysis frames before reconstructing the output signal y. This window function tries to eliminate abrupt waveform discontinuities at the boundaries of the analysis frames. Hann window is shown in also in figures 2.3 and 4.1. The Hann window has the nice property that for $Hs = \frac{N}{2}$ the sum of the overlapping windows is equal to 1 (one).

$$\sum_{n\in\mathbb{Z}} w(r - n\frac{N}{2}) = 1 \tag{2.3}$$

for all $r \in \mathbb{Z}$.

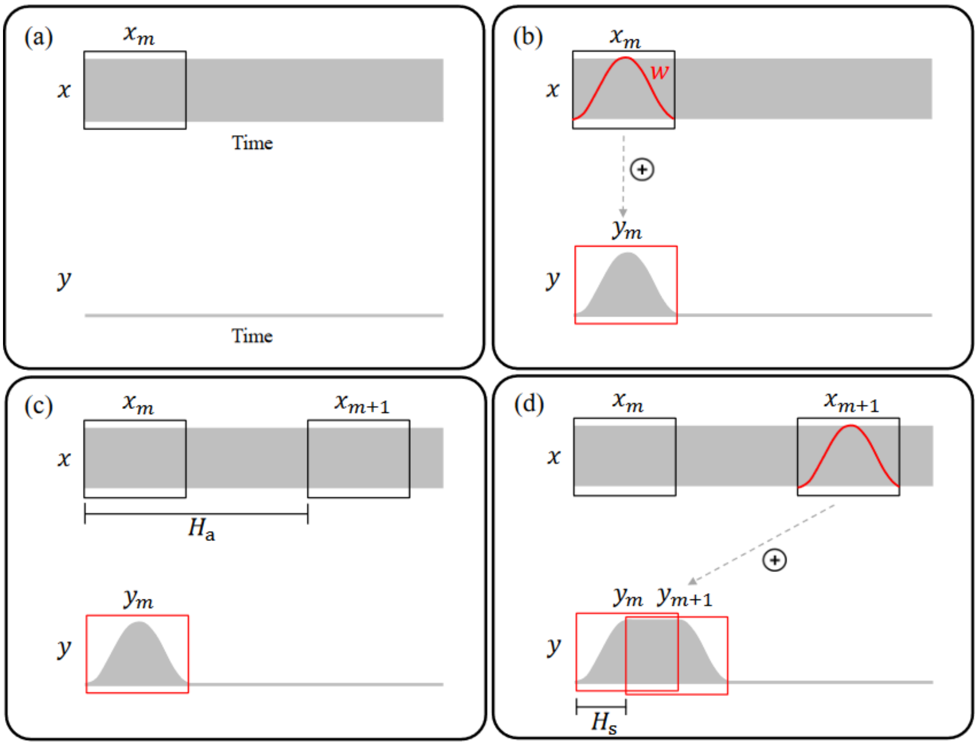


Figure 2.3: *Overlap-Add method process: (a) selection of analysis frame x_m, (b) creations of synthesis frame y_m, (c) selection of analysis frame x_{m+1} (d) Overlap-add y_{m+1} to y_m - (from [3]).*

In each iteration (m), the analysis frame x_m is selected at first (figure 2.3a). Subsequently, it is multiplied with a Hann window, generating the initial synthesis frame, which is then added to the output signal using the Overlap-Add method (figure 2.3b). The next analysis frame is chosen at a distance equal to H_a from the previous frame (figure 2.3c), it is again multiplied with a Hann window (same as before) and is added to the output signal using the Overlap-Add method with a distance equal to H_s from the previous synthesis frame (figure 2.3d).

In order to create the synthesis frame y_m from the analysis frame x_m, we use the following equation:

$$y_m(r) = \frac{w(r)x_m(r)}{\sum_{n\in\mathbb{Z}} w(r - nH_s)} \tag{2.4}$$

In Equation 2.4, the numerator represents the windowing of the analysis frame x_m, achieved by pointwise multiplication with the designated window function w. The denominator helps to normalize the frame by dividing it by the sum of the overlapping window functions, preventing amplitude fluctuations in the output signal. It's worth noting that when the selection for window w is a Hann window and $H_s = N/2$, the denominator always sums to one (1) according to Equation 2.3. This property can be easily seen in figure 2.3b, where the synthesis frame's amplitude is added directly to the output signal y without scaling. Moving on to the next analysis frame x_{m+1} (as shown in Figure 2.3c, this frame experiences windowing, overlaps with the prior synthesis frame, and is integrated into the output signal (Figure 2.3d). Figure 2.3 shows a scenario where the original signal is compressed ($H_a > H_s$). The process remains the same when the signal is stretched ($H_a < H_s$) but in this case, the analysis frames overlap surface is larger than the synthesis frames. Finally, the output signal y is reconstructed as follows:

$$y(r) = \sum_{m\in\mathbb{Z}} y_m(r - mH_s) \tag{2.5}$$

Problems One of OLA's problems is that it lacks signal sensitivity: it involves copying windowed analysis frames from fixed positions in the input signal to fixed positions in the output signal, with the input signal having no impact on the process. When OLA is applied to harmonic signals, the resulting signals are marked by a noticeable warbling effect, defined by periodic frequency modulation [3].

That is because OLA is incapable of preserving the local periodic structures in periodic signals. You can see this problem in Figure 2.4 in which a periodic input signal 'x' is stretched by a factor of $\alpha = 1.8$ using OLA. During the relocation of the analysis frames, the periodic structures in x may no longer align within the superimposed synthesis frames, resulting in distorted periodic patterns in the output signal y. These distortions are commonly referred to as phase jump artifacts.

Advantages OLA is a time-domain TSM procedure. In general, time-domain TSM procedures are not only efficient but also preserve the timbre of the input signal to a high degree. Also, it provides **excellent results** for **purely percussive signals**. This is because audio signals containing percussive elements rarely have local periodic structures. As a result, the phase jump artifacts introduced by OLA in periodic signals, are typically unnoticeable in the percussive signals. However, it's crucial to use a very short frame length 'N' (approximately 10 milliseconds) to minimize the impact of transient doubling, an artifact that can be seen in figure 2.6 and is addressed in the next section.

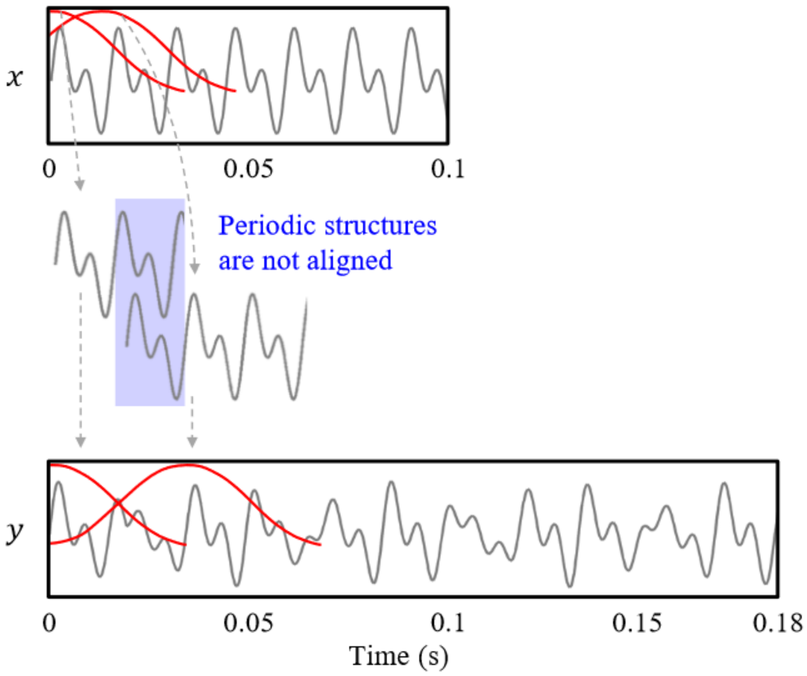


Figure 2.4: *Phase jump artifacts: one of the Overlap-Add problems (from [3]).*

2.1.3.3 Waveform Similarity Overlap-Add (WSOLA)

One approach in the time domain to reduce phase jump artifacts, as induced by OLA, is introducing some flexibility into the TSM process by allowing some degree of tolerance in the positions of analysis or synthesis frames. The main idea is to adjust consecutive synthesis frames in a manner that aligns the periodic structures within the frame waveforms in overlapping regions. This ensures that periodic patterns in the input signal are preserved in the output. In the literature, there are many variations of this concept, including synchronized OLA (SOLA) [32, 33], time-domain pitch-synchronized OLA (TD-PSOLA) [34], and autocorrelation-based methodologies [35]. For now, we will focus on WSOLA, a waveform similarity-based OLA method [30].

The core idea of WSOLA is to permit minor shifts (within a range of $\pm\Delta_{max} \in \mathbb{Z}$ samples) in the positions of the analysis frames. Figure 2.5 illustrates the underlying principle of WSOLA. Much like OLA, WSOLA operates iteratively.

Let' s see more details about WSOLA. Assume that we are in the m^{th} iteration, there are four steps in each iteration. In the first step, suppose the position of the analysis frame x_m is shifted by $\Delta_m \in [-\Delta_{max} : \Delta_{max}]$ samples. This frame is referred to as the adjusted analysis frame $x^{'}_m$ (as depicted in Figure 2.5a) is windowed with a Hann window and added to the output signal as in OLA. $x^{'}_m$ is calculated as follows:

$$x^{'}_m(r) = \begin{cases} x(r + mH_a + \Delta_m), & \text{if } r \in [-N/2 : N/2 - 1] \\ 0, & \text{otherwise} \end{cases} \tag{2.6}$$

In the second-third step, we need to adjust the position of the next analysis frame x_{m+1}. This task can be viewed as a constrained optimization problem. Our objective is to identify the best shift index $\Delta_{m+1} \in [-\Delta_{max} : \Delta_{max}]$ that aligns the periodic structures in the adjusted analysis frame x_{m+1} optimally with the structures of the previously added synthesis frame y_m in the overlapping region when both frames are superimposed at the synthesis hop size H_s. In an unconstrained scenario, the optimal choice for the adjusted

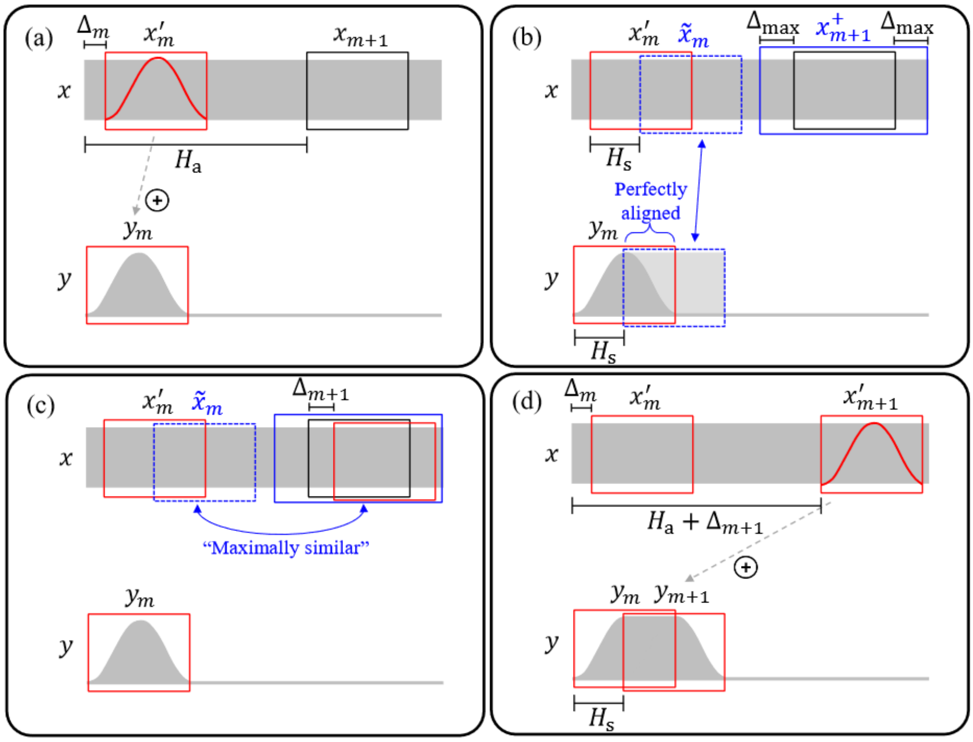


Figure 2.5: *WSOLA method process: (a) adjusted analysis frame $x^{'}_m$ (prime symbol denotes adjustment), windowed and added to the output signal, (b,c) selection of a frame from the extended frame region x^{+}_{m+1} (solid blue box) that closely matches the natural progression $\tilde{x}_m$ (dashed blue box) of the adjusted analysis frame $x^{'}_m$, (d) adjusted analysis frame $x^{'}_{m+1}$ for the subsequent iteration is windowed and incorporated into the output signal y - extracted from [3].*

analysis frame x_{m+1} would be to match the natural progression $\tilde{x}_m$ of the adjusted analysis frame x_m (dashed blue box in Figure 2.5b):

$$\tilde{x}_m(r) = \begin{cases} x(r + mH_a + \Delta_m + H_s), & \text{if } r \in [-N/2 : N/2 - 1] \\ 0, & \text{otherwise} \end{cases} \tag{2.7}$$

This is the case because the adjusted analysis frame $x^{'}_m$ and the synthesis frame y_m are nearly identical (except for the windowing). Consequently, the structures of the natural progression $\tilde{x}_m$ align perfectly with the structures of the synthesis frame y_m when these two frames are superimposed at the synthesis hop size H_s (as shown in Figure 2.5b). However, the constraint $\Delta_{m+1} \in [-\Delta_{max} : \Delta_{max}]$ necessitates that the adjusted frame $x^{'}_{m+1}$ must fall within the bounds of the extended frame region x^{+}_{m+1} (solid blue box in Figure 2.5b):

$$x^{+}_{m+1}(r) = \begin{cases} x(r + (m+1)H_a), & \text{if } r \in [-N/2 - \Delta_{max} : N/2 - 1 + \Delta_{max}] \\ 0, & \text{otherwise} \end{cases} \tag{2.8}$$

Hence, the concept is to select the adjusted frame $x^{'}_{m+1}$ from $'x + m + 1'$ as the frame whose waveform most closely resembles $\tilde{x}_m$.' To accomplish this, we need to establish a metric that assesses the similarity between two frames. One potential option for this metric is cross-correlation of signal s_1 with the signal s_2 that is shifted by $\Delta \in \mathbb{Z}$ samples:

$$\text{crosscorr}(s_1, s_2, \Delta) = \sum_{r \in \mathbb{Z}} s_1(r)s_2(r + \Delta) \tag{2.9}$$

We can subsequently calculate the optimal shift index Δ_{m+1} that maximizes the cross-correlation between $\tilde{x}_m$ and x^+_{m+1} by:

$$\Delta_{m+1} = \arg \max_{\Delta \in [-\Delta_{max}:\Delta_{max}]} (\text{crosscorr}(\tilde{x}_m, x^+_{m+1}, \Delta)) \tag{2.10}$$

The shift index Δ_{m+1} determines the location of the adjusted analysis frame x'_{m+1} within the extended frame region x^+_{m+1} (as shown in Figure 2.5c). Lastly, we calculate the synthesis frame y_{m+1}, following a process similar to OLA, by

$$y_{m+1}(r) = \frac{w(r)x(r + (m+1)H_a + \Delta_{m+1})}{\sum_{n\in\mathbb{Z}} w(r - nH_s)} \tag{2.11}$$

and to reconstruct output signal y, we use the equation:

$$y(r) = \sum_{m\in\mathbb{Z}} y_m(r - mH_s) \tag{2.12}$$

same as OLA's reconstruction equation.

Problems WSOLA-like methods often suffer from a significant issue known as transient doubling or stuttering, as seen in Figure 2.6. In this artifact, a single transient in the input signal falls within the overlapping region of two consecutive adjusted analysis frames x_m and x_{m+1}. As these frames are shifted and added to the output signal, the transient gets duplicated and is heard twice in rapid succession. A related artifact is transient skipping, where transients can be lost during the modification process because they don't align with any of the analysis frames. Transient doubling is more common when stretching a signal ($\alpha > 1$), while transient skipping typically occurs during signal compression ($\alpha < 1$). These artifacts are especially noticeable when working with signals that contain percussive elements, such as instruments with strong onsets like drums or piano.

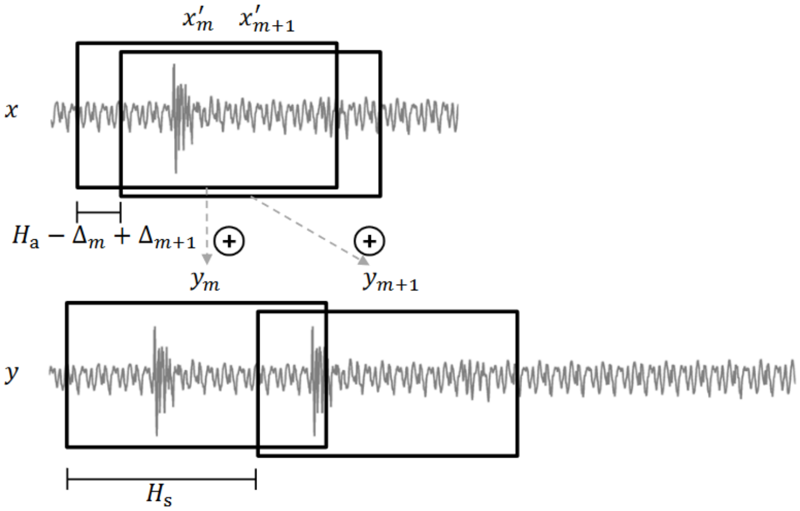


Figure 2.6: *Transient doubling problem in OLA methods (from [3])*

Moreover, WSOLA faces challenges when modifying polyphonic input signals, like orchestral music recordings. In such cases, the output often retains a noticeable warbling effect. This is because WSOLA, by design, can only preserve the most prominent periodic pattern in the input signal's waveform. Consequently, when working with recordings that have multiple harmonic sound sources, only the sound of the dominant source is preserved in the output, while other sources can still introduce phase jump artifacts. While WSOLA is well-suited for modifying monophonic input signals, it may not perform as effectively with more complex audio.

To ensure that WSOLA can effectively adapt to the most dominant periodic pattern in the input signal, it's essential for one frame to capture at least a full period of that pattern. Additionally, the tolerance parameter Δ_{max} needs to be sufficiently large to allow for appropriate adjustment, often set to at least half a period's length. Considering that the lowest frequency audible to humans is approximately 20 Hz, a common choice is to use a frame length N corresponding to 50 ms and a tolerance parameter of 25 ms.

To mitigate transient doubling and skipping artifacts in WSOLA, one approach is to implement a transient preservation scheme. In another method [36], the concept involves initially identifying the temporal positions of transients in the input signal using a transient detector. During the WSOLA process, the analysis hop size is temporarily set to be equal to the synthesis hop size whenever an analysis frame falls within the vicinity of an identified transient. In this area, which encompasses the transient, the frame is copied without modification to the output signal, preventing WSOLA from duplicating or skipping it. Any deviation in the global stretching factor is dynamically compensated for in the regions between transients.

2.1.3.4 Other Models

Sinusoidal models While OLA-based methods are fast and perform well, there are other ways to approach the time-scaling problem. Paremetric speech modeling is one o them. A popular family of models able to time-scale speech waveforms is the so-called *Sinusoidal Models* (SMs). These models are particularly useful for capturing the periodic and harmonic nature of speech, making them valuable for various applications in speech and audio processing, such as speech synthesis, coding, and analysis. In general, SMs usually decompose speech signals into components. This decomposition allows for different handling of each component during time scaling. Several names are given to these components in the literature: periodic and aperiodic, stationary and non-stationary, harmonic and noise, harmonic and stochastic, among others. In sinusoidal modeling, the speech signal $x(t)$ can be written as

$$x(t) = \sum_{k=-N}^{N} a_k(t) e^{j\phi_k(t)} \tag{2.13}$$

where N is the number of sinusoids present in the signal, $a_k(t)$ denote the instantaneous amplitudes, and $\phi_k(t)$ denote the instantaneous phases. Index k denotes the sinusoid or harmonic number. Estimating $a_k(t)$ and $\phi_k(t)$ is the goal of sinusoidal analysis and can be achieved in many ways, such as frequency domain transformations [37], Least Squares estimation [38, 39, 40], or subspace methods [41, 42, 43]. Instantaneous parameter estimation is followed by parameter interpolation in time or frequency domain to obtain time-scaled instantaneous versions of $a_k(t)$ and $\phi_k(t)$. Successful attempts on time scaling modification of these models have been published in literature [44, 45, 46, 47, 48].

Phase Vocoder Another method with very good performance is Phase Vocoder. WSOLA is proficient at preserving the primary periodicity in the input signal. However, to further improve the quality of time-scale modified signals, it's important to retain the periodic characteristics of all signal components. This is where frequency-domain TSM procedures come into play. They interpret each analysis frame as a combination of weighted sinusoidal components with known frequency and phase. By manipulating these components individually, phase jump artifacts across all frequencies in the reconstructed signal can be mitigated. The short-time Fourier transform is a crucial tool for frequency analysis of the input signal. However, the accuracy of frequency estimates may vary depending on the chosen discretization parameters. To address this, the phase vocoder technique [49, 50], which is commonly used for both frequency estimation and time-scale modification, is employed. In time-scale modification procedures based on the phase vocoder (PV-TSM), these refined frequency estimates are used to update the phases of sinusoidal components in the input signal. This process is known as phase propagation.

STRAIGHT More advanced vocoders like Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectra (STRAIGHT) [51, 52] are able to perform modifications and manipulations of speech with high fidelity. STRAIGHT models the periodic part of speech as an AM-FM set of sinusoids:

$$x(t) = \sum_{k \in N} a_k(t) \sin\left(\int_{t_0}^{t} k(\omega(\tau) + \omega_k(\tau))d\tau + \phi_k\right) \tag{2.14}$$

STRAIGHT employs pitch-adaptive spectral analysis in conjunction with a technique for reconstructing the signal in the time-frequency domain. It also utilizes an approach to design the excitation source based on phase manipulation. This combination enables the preservation of the bilinear surface in the time-frequency domain. As a result, STRAIGHT can facilitate extensive adjustments, such as pitch, vocal tract length, and speaking rate, by over 600%, without introducing additional degradation due to these parameter modifications. However, the complexity of these models is relatively high.

2.2 Conclusions

From all of these methods that were presented, we selected WSOLA as the base method because it is simple and fast, as it is a time domain method, and with very good quality results despite the transient skipping or transient doubling effect that can be reduced. We are trying to solve its intelligibility issues with the protection of non-stationary parts of the signal that can be lost (in signal shrinking) or stretched badly (in signal stretching), Finally, we compared and tested these methods on the dataset [1] that is also presented in chapter. 3. The results of these tests as well as discussion can be found in chapter 5

Among the various methods discussed, we chose WSOLA as the foundational technique due to its simplicity, speed (as it operates in the time domain), and the commendable quality of results it offers, notwithstanding potential transient skipping or doubling effects, which can be mitigated. To address its intelligibility challenges, we focused on protecting the non-stationary segments of the signal from potential loss during shrinking or excessive stretching. Next, we proceeded to evaluate and compare these methods using the dataset [1] introduced in Chapter 3. For a comprehensive examination of the results and an in-depth discussion, please refer to Chapter 5.

Chapter 3

Dataset

3.1 Description

The dataset [1] is structured following the Harvard/IEEE sentence material format in American English as outlined in [53]. This material contains 72 phonetically balanced lists each containing ten meaningful sentences (e.g., ”Rice is often served in round bowls”) but in Greek. In our listening tests, we only used two lists for the intelligibility test (20 samples) and two lists for the preference test (20 different samples). Notably, each of these sentences contains five keywords, each consisting of one or two syllables.

In the subsequent section, we will dive into the initial stages of developing the GrHarvard Corpus, a sentence corpus designed in the Harvard/IEEE tradition, tailored for Modern Greek.

3.2 Design

3.2.1 Criteria

The sentences were crafted based on specific objective criteria:

1. Each sentence consists of precisely five keywords,
2. The sentence length ranges from five to nine words,
3. All words within the sentences have a maximum of three syllables,
4. Sentences are exclusively either statements or commands.

3.2.2 Keywords

In most cases, keywords are content words, but depending on sentence structure and meaning, function words can also serve as keywords. Notable function words marked as keywords include δεν/δε (not), μην/μη (do not), σαν/σα (like/as), πιο (more), πια (any longer), προς (to/towards), ενώ (although), μπρος (ahead), τόσος-η-ο (such/so much), όσος-η-ο (as much as), αλλά (but), αντί (instead). Proverbs and most of the proper names were excluded.

Considerable effort was invested to minimize the repetition of keywords, including their various conjugated forms, throughout the corpus. Additionally, few foreign loanwords, such as μπουφάν (jacket), κολιέ (necklace), τρακτέρ (tractor), and others, were incorporated as keywords.

3.2.3 Translation challenge and change

Translating the original Harvard/IEEE sentences into Greek proved to be a formidable task, mainly due to the syntax and grammar differences between English and Greek. A significant number of English words, consisting of just one or two syllables, corresponded to Greek words with more than three syllables, making them unsuitable candidates for translation. Moreover, the original sentences were written decades ago, during a wartime context, many of which would appear illogical or out of context today.

For instance, the sentence from the original Harvard Corpus: ”These days a chicken leg is a rare dish.” has been translated in the Greek Harvard Corpus as ¨Ο κόσμος τρώει τακτικά ψητό κρέας' (People have roast meat regularly). Thus, while the material drew inspiration from the original American English Harvard/IEEE sentences, most of the sentences have been modified or have low to no resemblance to the original content.

3.2.4 Greek words selection

The majority of Greek words were manually selected from Greeklex 2, a lexical database with comprehensive information on part-of-speech, syllabication, phonological attributes, and stress patterns [54]. The selection process was guided by the principle that the combination of precisely five words should yield a coherent and non-repetitive sentence, one that easily can be a part of an everyday conversation.

3.2.5 Phonetic Transcriptions

The sentences went through phonetic transcription using the online tool provided by IPLR [55], a comprehensive online resource for Greek word-level and sublexical information, followed by a manual review.

To simplify cases involving nasals followed by homorganic stops, a consistent approach was applied, in line with [55] and [54]. For instance, the word ¨λάμπει' was transcribed as [”labi], rather than [”lambi].

The transcriptions for the GrHarvard sentences follow the SAMPA format, as demonstrated in [56]. For example, the sentence ¨Το σχέδιο δράσης είναι ασαφές προς το παρόν' is transcribed as [to.”sCeDio.”Drasis.”ine.asa”fes.pros.to.pa”ron] (The plan of action is unclear at the moment).

The GrHarvard sentence material, presented in Greek orthography along with its SAMPA transcriptions, is accessible to the research community. Additionally, it includes essential meta-data such as the word count, syllable count, and phoneme count per sentence, as well as information about keywords and their syllable counts. For more information check [1].

3.3 Statistical Infomation

3.3.1 Sentence lengths

As previously indicated, the GrHarvard Corpus contains 720 sentences. And given that each sentence contains five keywords, the number of keywords is $3,600$ keywords. The sentence length ranges from a minimum of five to a maximum of nine words. Sentence lengths can be described as a distribution centered around seven, eight, or nine words, with a smaller portion consisting of sentences with five or six words, as illustrated in Table 3.1 taken from [1].

Table		
Number of words per sentence	Percentage (and absolute number) of sentences in corpus	Sentence example (%)
5	1.5% (11)	Ψύχοντας νερό φτιάχνεις καθαρό πάγο. (By freezing water one makes clear ice.)
6	9.6% (69)	Βρέθηκε νέο φάρμακο κατά του διαβήτη. (A new drug against diabetes was found.)
7	28.5% (205)	Εννιά εργάτες σκάβουν τον τόπο για αρχαία. (Nine workers are digging the site for ancient artifacts.)
8	38.9% (280)	Χώμα και σκόνη έτσουξαν τα μάτια του κοριτσιού. (Soil and dust stung the girl’s eyes.)
9	21.5% (155)	Ο τολμηρός λοχίας σύρθηκε στο πεδίο με τις νάρκες. (The bold sergeant dragged himself on the minefield.)

Table 3.1: *Number of words per sentence in the corpus (from [1]).*

.

3.3.2 Syllables frequency distribution

Concerning the syllable count within keywords, it’s worth noting that the majority of keywords consist of either two syllables (42.2%) or three syllables (54.2%). The percentage of one-syllable keywords is very small (3.58%), mainly because there are relatively few monosyllabic content words in Greek. In sentences, the total number of syllables spans a range of 10 to 22 syllables. Most sentences fall within the 15 to 18 syllable range, as illustrated in Figure 3.1.

3.3.3 Phonemes frequency distribution

The GrHarvard Corpus holds a total count of 20, 230 phonemes and allophones scattered across its 3, 600 keywords. Detailed information on the frequency of vowel and consonant phonemes and allophones can be found in Tables 3.2 and 3.3, respectively.

The collective number of phonemes found within keywords per sentence varies from 16 to 38. However, in the majority of sentences, keywords contain between 24 to 34 phonemes, as depicted in Figure 3.2.

3.3.4 Comparison with other corpora

To offer insights into the phoneme frequency distribution within the GrHarvard Corpus, they performed a comparative analysis with two other Modern Greek corpora.

The first, known as the ”C Corpus,” contains an extensive collection of 34 million tokens derived from journalistic, legal, and literary texts in the Hellenic National Corpus (written

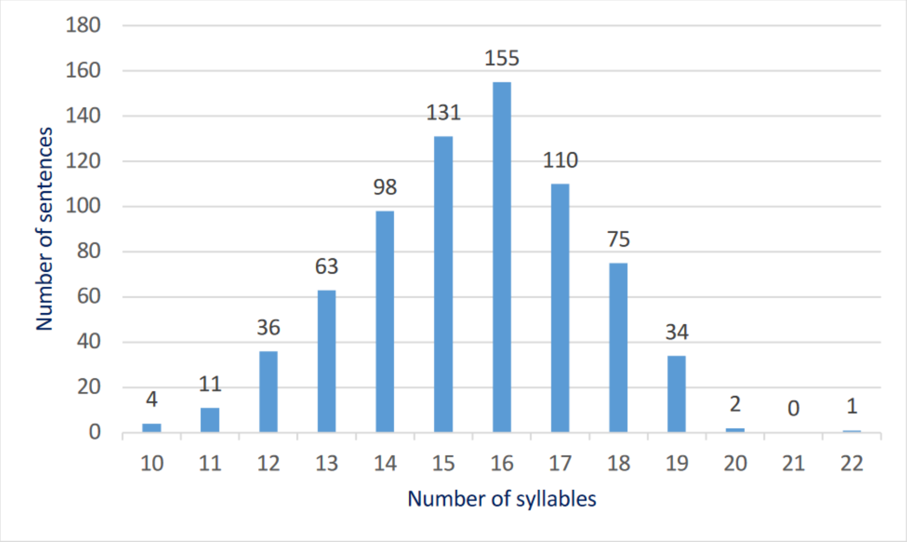


Figure 3.1: *Histogram of the number of sentences given the number of syllables in the corpus (from [1]).*

Table			
APA/SAMPA Vowels	Allophones in APA (and SAMPA) 1 (%)	Allophones in APA (and SAMPA) 2 (%)	Total (%)
i	8.14	4.74	12.88
e	4.84	3.30	8.14
a	7.92	4.69	12.61
o	5.22	3.26	8.48
u	1.42	1.08	2.50
j			0.44

Table 3.2: *Vowels Frequency Distribution (from [1]).*

corpus), carefully curated and cross-verified with an online Greek dictionary [54]. While there are more extensive corpora available, they selected the ”C Corpus” for comparison due to its comprehensive verification.

The second corpus (spoken corpus) is a singular corpus that had been analyzed for phonemic frequency. This corpus includes 102,934 words extracted from 100 television and radio shows of the Hellenic Broadcasting Corporation [57] (spoken corpus). The frequency of phonemes in this spoken corpus was analyzed and recorded.

Figure 3.3 provides an illustrative comparison of phoneme frequency distribution between the GrHarvard Corpus and the two corpora mentioned earlier. In general, the phoneme frequency distribution within the GrHarvard Corpus aligns well with both the written and spoken corpora. As expected, some variations exist, primarily in phonemes that appear in frequently used function words, because they were excluded from their keyword-based analysis. Consequently, the deletion of definite articles like ό/η/το,′ indefinite articles like ′ένας/μία/ένα,′ and the conjunction ′και′ (and) contributes to a reduced representation of /o/, /i/, /e/, /t/, [c], and [n] within the GrHarvard Corpus.

Table			
APA/SAMPA Consonants	Unstressed (%)	Stressed (%)	Total (%)
p (p)			4.01
t (t)			4.80
k (k)	k (k): 3.75	c (c): 1.25	5.00
b (b)			0.50
d (d)			0.63
g (g)	g (g): 0.20	ï (gj): 0.07	0.27
f (f)			1.93
v (v)			1.62
T (T)			1.22
D (D)			2.08
s (s)			9.23
z (z)			0.98
x (x)	x (x): 1.15	X (C): 1.12	2.27
⊗ (G)	⊗ (G): 1.17	⊗ (jj): 0.88	2.05
ts (ts)			0.26

Table 3.3: *Consonants Frequency Distribution (from [1]).*

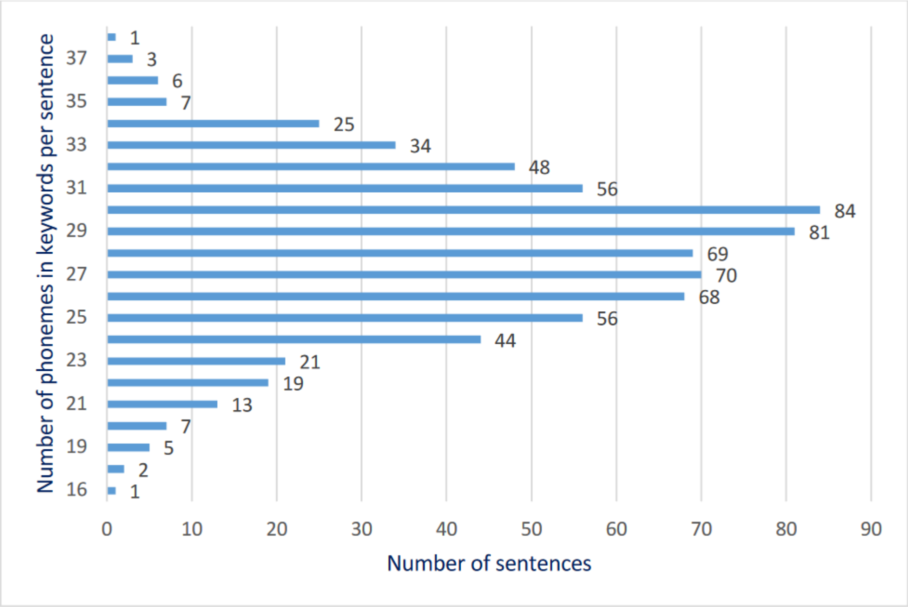


Figure 3.2: *Histogram of the number of sentences given the number of phonemes in keywords per sentence (from [1])*

.

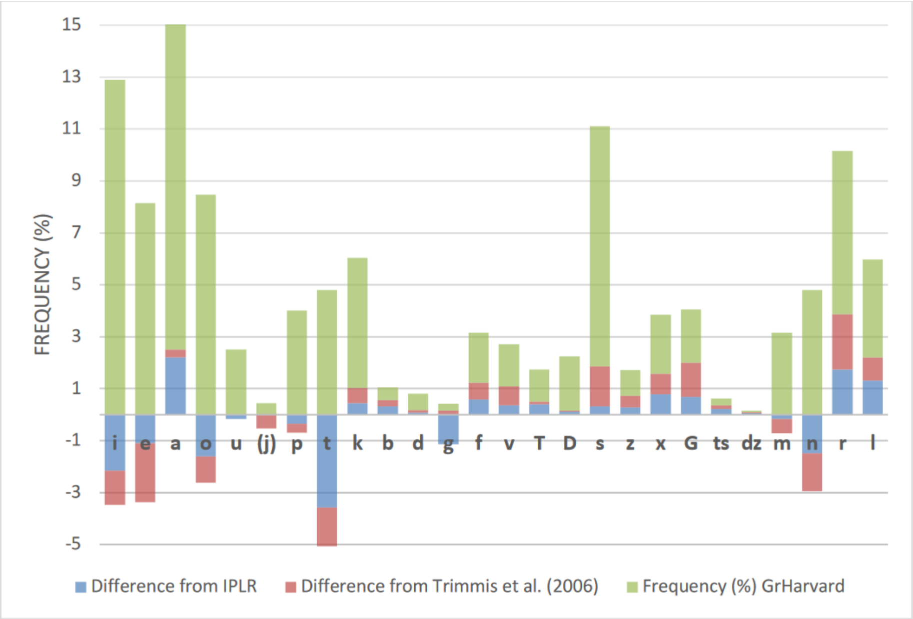


Figure 3.3: *The graph illustrates how phoneme frequencies in the GrHarvard Corpus compare to those in other corpora. Negative values indicate that a particular phoneme appears less frequently in the GrHarvard Corpus compared to the corresponding phoneme in the reference corpus. SAMPA notation is used for phoneme representation (from [1]).*

.

Chapter 4

Non-Stationarity Detection Method

This paper [58] proposes a powerful tool for automatically controlling time-scale factors in speech signals by analyzing their characteristics and also protecting the input signal's sensitive parts. This tool consists of two parts, the front-end and the back-end.

4.1 Back-end - Time Scale Modification

Let's start with the back-end method. To produce fast speech from normal speech, the time-domain algorithm for time-scale modification of speech that used is called WSOLA [30]. WSOLA was selected because of its advantages in contrast to other signal processing methods as we explained in the chapter section 2.2. WSOLA takes as input the signal we want to change its scale and the desired scale factor that we want this signal to have and as an output gives the final signal. The desired scale factor can be one value (i.e. 0.5) or a vector of values with the same length as the input signal. This allows us to control differently every region of the input signal with different scales. The problems were two. At first, at large scale factors, WSOLA suffered from unnatural artifacts and tonalities and the amount of these perceived artifacts depended on the time scale factor and the speaker. On the other hand, at small scale factors, some parts of the signal vanished. So, we used a front-end method to diminish those issues.

4.2 Front-end - Detection of non-stationarity

4.2.1 Overview

The front-end method's main purpose was to protect signal regions sensitive to scale factor changes. So, we needed a method that protects them by changing their scale factor to be closer to 1 (neutral) so, they will not be affected by the audio speed changes. It was found that the areas of the speech signal that were most affected were those that could be described as non-stationary. By controlling the time scale factors in these non-stationary areas, it is possible to reduce the tonalities and disappearance of sounds in the time-expanded signal. To find these areas, this paper uses three features including Line Spectrum Frequencies LSFs (C2), the derivative of Root Mean Square RMS values (C1), and a combination of both features (C3) that measure non-stationarity over the signal. Only one of these features is needed to change the scale factor across the signal. After the protection of the scale factor, We pass the signal and the final scale factor vector in the back-end method to produce the final signal.

4.2.2 Method

To identify non-stationarity in speech signals, we use a frame-by-frame analysis method that calculates three criteria in each frame (C1, C2, C3) of the signal. As a frame, we used a Hanning window with size length(C_{win}) which was being shifted $\frac{\text{length}(C_{win})}{2}$ each time (50% window coverage). Hanning is a window that helps you deal with the windows'coverage as you can see from the figure 4.1 on page 24. After that, we can use one of the three criteria to finally protect the signal. We will see how in the next paragraphs. First, let's analyze the criteria.

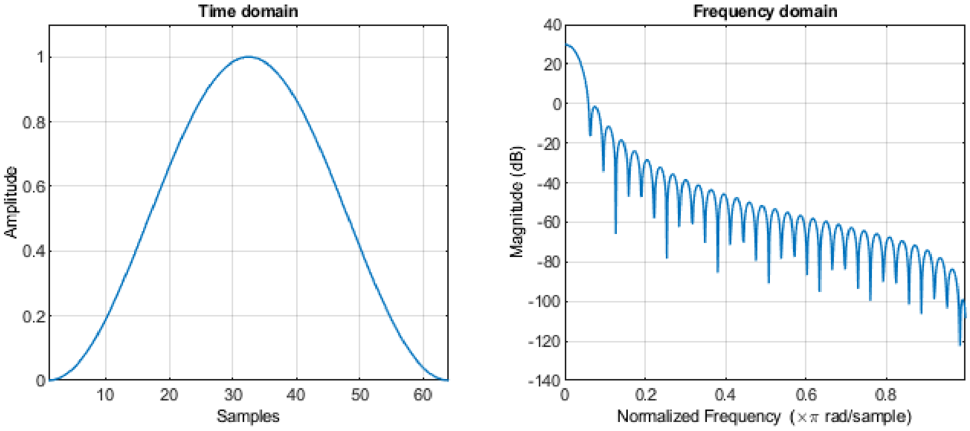


Figure 4.1: *Example of Hann(Hanning) window (from [4])*

4.2.3 Criteria

Implementation details Following an exploration of window sizes, we aimed to select a window size that introduces variability into the three criteria while minimizing noise. That's why we selected a window size equal to 20 milliseconds and a shift size equal to 10 milliseconds. You can see some plots in figure 4.2 (in page 25) in which we used values of criterion C2 to compare. Our purpose was for the result to be smooth because we did not want to create any weird tonalities in the output signal but not close to linear. The optimization is not yet complete, making it a potential avenue for future research to investigate how varying sizes may impact the quality of the output audio samples.

4.2.3.1 C1

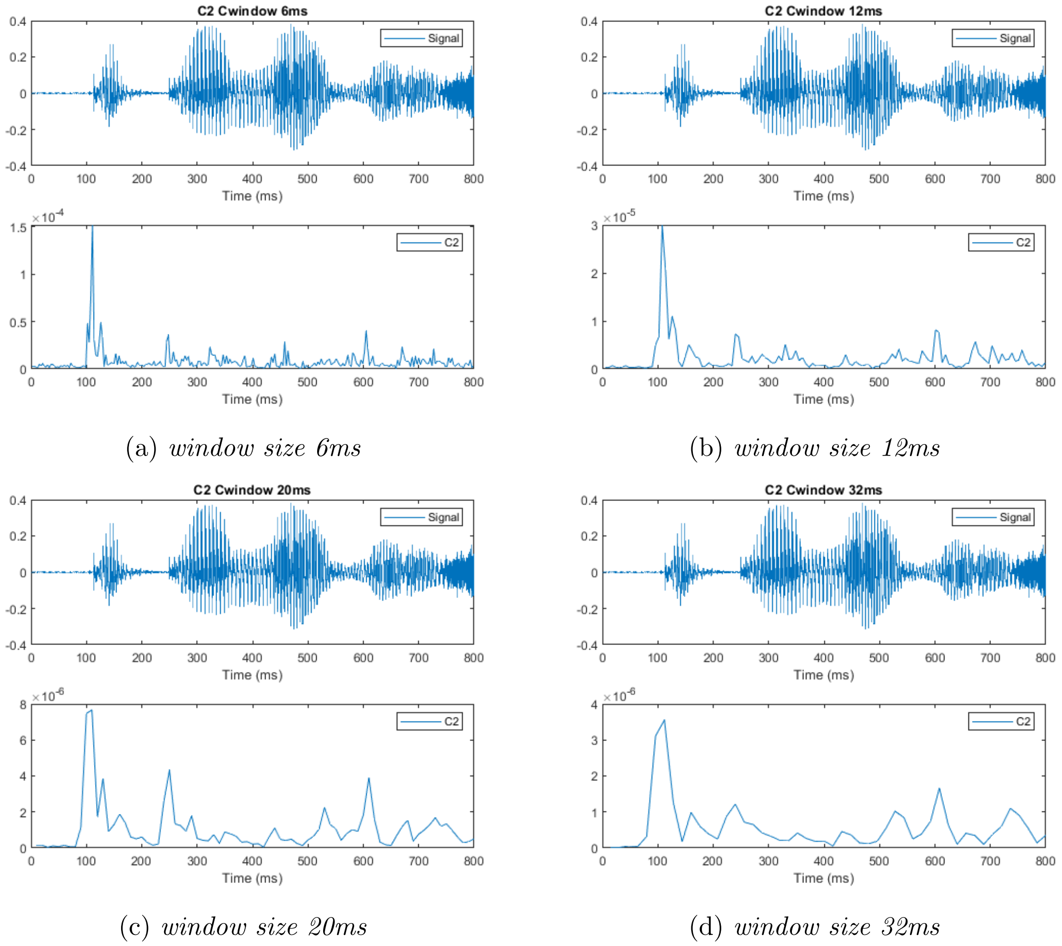
Description The first criterion C1 is based on the transition rate of the Root Mean Square (RMS) values E_n and It is a time-domain criterion. As we can find in [59], RMS means the root mean square value of a signal. You can calculate it by squaring each value, calculating their mean, and at the end, taking the square root of the result. They represent the average "power" of a signal and they are only related to the signal amplitudes.

A normalized transition rate of the RMS values is given by:

$$C_n^1 = \frac{|E_n - E_{n-1}|}{E_n + E_{n-1}} \tag{4.1}$$

where the transition rate of the RMS value E_n is given by:

$$E_n = \sqrt{\frac{1}{N+1} \sum_{m=-N/2}^{N/2} x^2[n+m]} \tag{4.2}$$



(a) *window size 6ms*

(b) *window size 12ms*

(c) *window size 20ms*

(d) *window size 32ms*

Figure 4.2: *C2 plots over different analysis window sizes.*

This criterion can take values:

$$C_n^1 = \begin{cases} \sim 1, & \text{if } |E_n - E_{n-1}| \text{ is large} \\ \sim 0, & \text{if } |E_n - E_{n-1}| \text{ is small} \end{cases} \tag{4.3}$$

Figure 4.3 shows four segments of speech with the measurements of the C1 criterion below them. It is easy to see from Figures 4.3b and 4.3d that C1 is good for detecting abrupt changes in speech signals (such as the stop burst at time ≈100ms). However, it is easy to see that C1 is too noisy in voiced speech (check at time after 300ms).

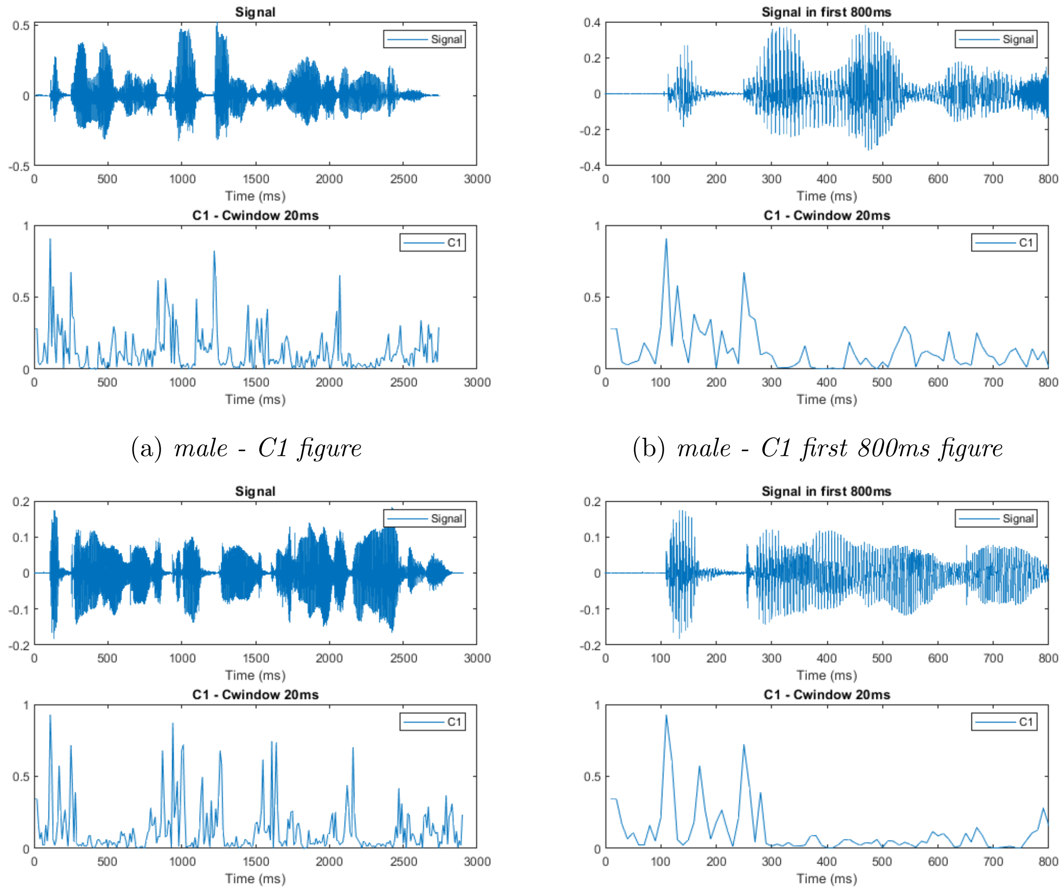
Implementation detail We could not calculate the first value of C_n^1 because we could not calculate E_0 (there is not a valid frame before the first frame given that frames started from 1).

$$C_n^1(1) = \frac{|E_1 - E_0|}{E_1 + E_0} \tag{4.4}$$

This means that in the first iteration, we only calculated E_1. After, we continued to the next frame and we calculated the second value:

$$C_n^1(2) = \frac{|E_2 - E_1|}{E_2 + E_1} \tag{4.5}$$

.



(a) *male - C1 figure*

(b) *male - C1 first 800ms figure*

(c) *female - C1 figure*

(d) *female - C1 first 800ms figure*

Figure 4.3: *Signal (top plots) and C1 criterion (bottom plots) of two recording versions (Male-Female) of file 'P1' extracted from [1]. The left figures contain the whole signal while the right figures contain only the first 800ms of the signal.*

Finally, to fix the problem of missing value in this feature, we copied the second value of C_n^1 to the first one in the second frame.

$$C_n^1(1) = C_n^1(2) \tag{4.6}$$

4.2.3.2 C2

Description The second criterion C2 is based on the slope of the regression line that represents the change in Line Spectrum Frequencies (LSFs) over time. It is a frequency domain criterion. The number of LSFs is represented with P. So, for the LSF l, for $l = 1, \cdots, P$, at instant n, within the time interval $[n - M, n + M]$, the slope for the LSF l, y_l, is calculated as:

$$g_n^l = \frac{\sum_{m=-M}^{M} m y_l(n+m)}{\sum_{m=-M}^{M} m^2} \tag{4.7}$$

where the transition rate m_n is given by:

$$m_n = \sum_{l=1}^{P} \left(g_n^l\right)^2 \tag{4.8}$$

Normalization of m_n between 0 and 1 is achieved through the function:

$$C_n^2 = \frac{2}{1 + e^{-\beta_1 m_n}} - 1 \tag{4.9}$$

In the normalization looks like they are using a sigmoid function [60]

$$f(x) = \frac{1}{1 + e^{-x}} \tag{4.10}$$

As parameters, $P = 10$ and $\beta_1 = 20$ are used. The weight β_1 was defined based on measurements they made on known stable and voiced signals and on signals with fast transitions. These values were found inside the paper [58]

Fig.4.4 shows the same segments of speech as Fig.4.3 on the top plots but now, on the bottom plots exists the C2 criterion. In this figure, stop signals have not-so-strong values compared to this criterion's max values, but you can see in fig 4.7 (in page 31) that in the voiced sections it is not as noisy as the first criterion C1. Thus, while this criterion appears to be more effective in distinguishing voiced sections, when compared to the initial criterion it isn't suitable for sound types like stops, or, in a broader sense, speech events with brief durations. That's because the regression line gradient in these instances is nearly zero. As a result, they determined that combining criteria C1 and C2 is a more suitable approach for detecting transitions in all types of speech events.

Implementation details To compute criterion C2, it was essential to have already calculated the Line Spectral Frequencies (LSFs) for all frames. Therefore, we computed LSFs as part of the preprocessing step prior to criteria computation. This was accomplished by using a 30 ms window with a 5 ms frame rate, using a standard Linear Prediction procedure (a pre-emphasis [61] of the signal before the computation of the AR filter has been used).

For the calculation of the LSF frequencies, in each frame, we calculated the Linear Prediction Coefficients. Then, using MATLAB's poly2lsf function, we converted the LPC coefficients to LSFs. Up to that point, we had LSF values calculated for each frame's midpoint, but we aimed to obtain values for the entire signal. To achieve this, we performed linear interpolation with extrapolation (to create also the values before the first and after the last midpoint). All LSFs are saved into a matrix with size P x N, where N is the number of samples.

Finally, for consistency, we copied the second value of C_n^2 to the first one in the second frame. That is because we could not calculate C_n^1 and C_n^3 on the first frame (because we did not know the value of E_0 as it is explained in implementation details of C_n^1). So,

$$C_n^2(1) = C_n^2(2) \tag{4.11}$$

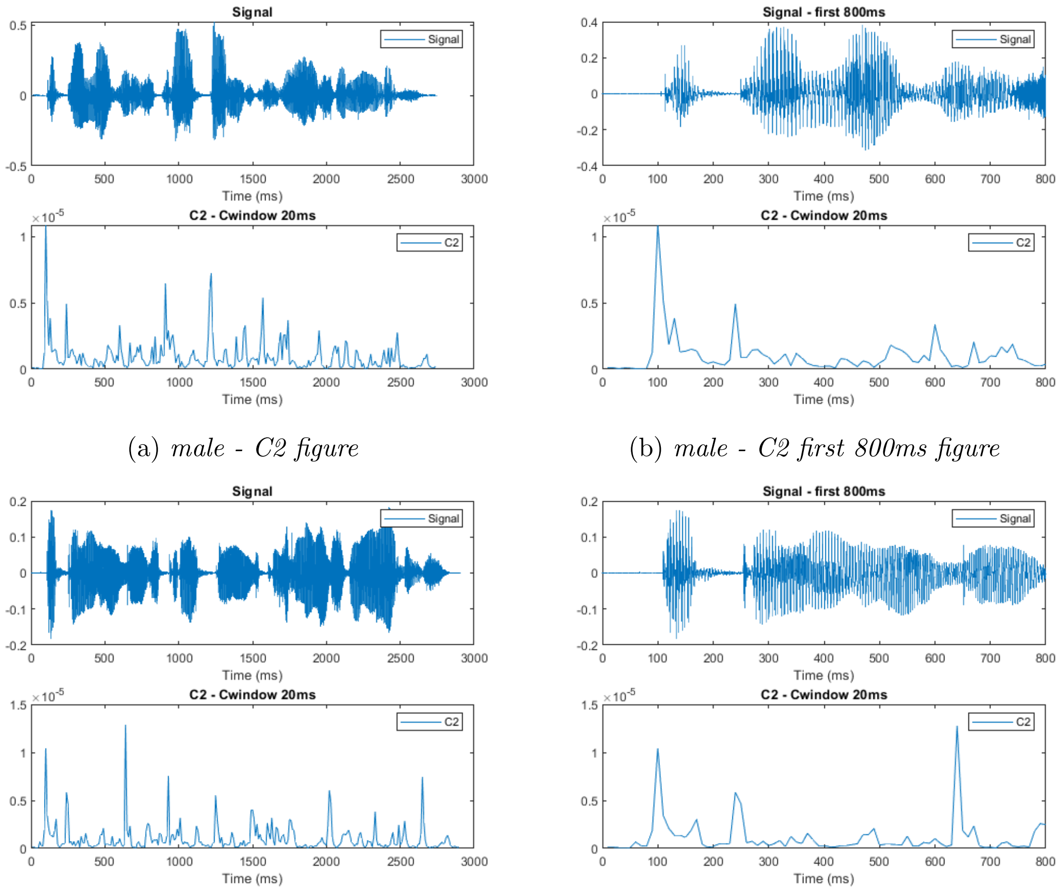
4.2.3.3 C3

Description Trying to improve the performance of the two metrics a third criterion was proposed that tried to join the benefits of the above two criteria. It is calculated as follows:

$$C_n^3 = \frac{2}{1 + e^{-\beta_2 m_n - \alpha C_n^1}} - 1 \tag{4.12}$$

As parameters, they used $\beta_2 = 17$ while α is given from the following equation:

$$\alpha = \begin{cases} 18.43 \left(1.001 - 1.0049 e^{C_n^1} + C_n^1 e^{C_n^1}\right), & \text{if } C_n^1 \leq 0.5 \\ 0.5, & \text{if } C_n^1 > 0.5 \end{cases} \tag{4.13}$$



(a) *male - C2 figure*

(b) *male - C2 first 800ms figure*

(c) *female - C2 figure*

(d) *female - C2 first 800ms figure*

Figure 4.4: *Signal (top plots) and C2 criterion (bottom plots) of two recording versions (Male-Female) of file 'P1' extracted from [1]. The left figures contain the whole signal while the right figures contain only the first 800ms of the signal.*

The values of parameter α have been determined using a least squares approach (using an exponential model) to normalize the criterion C_n^3 between 0 and 1.

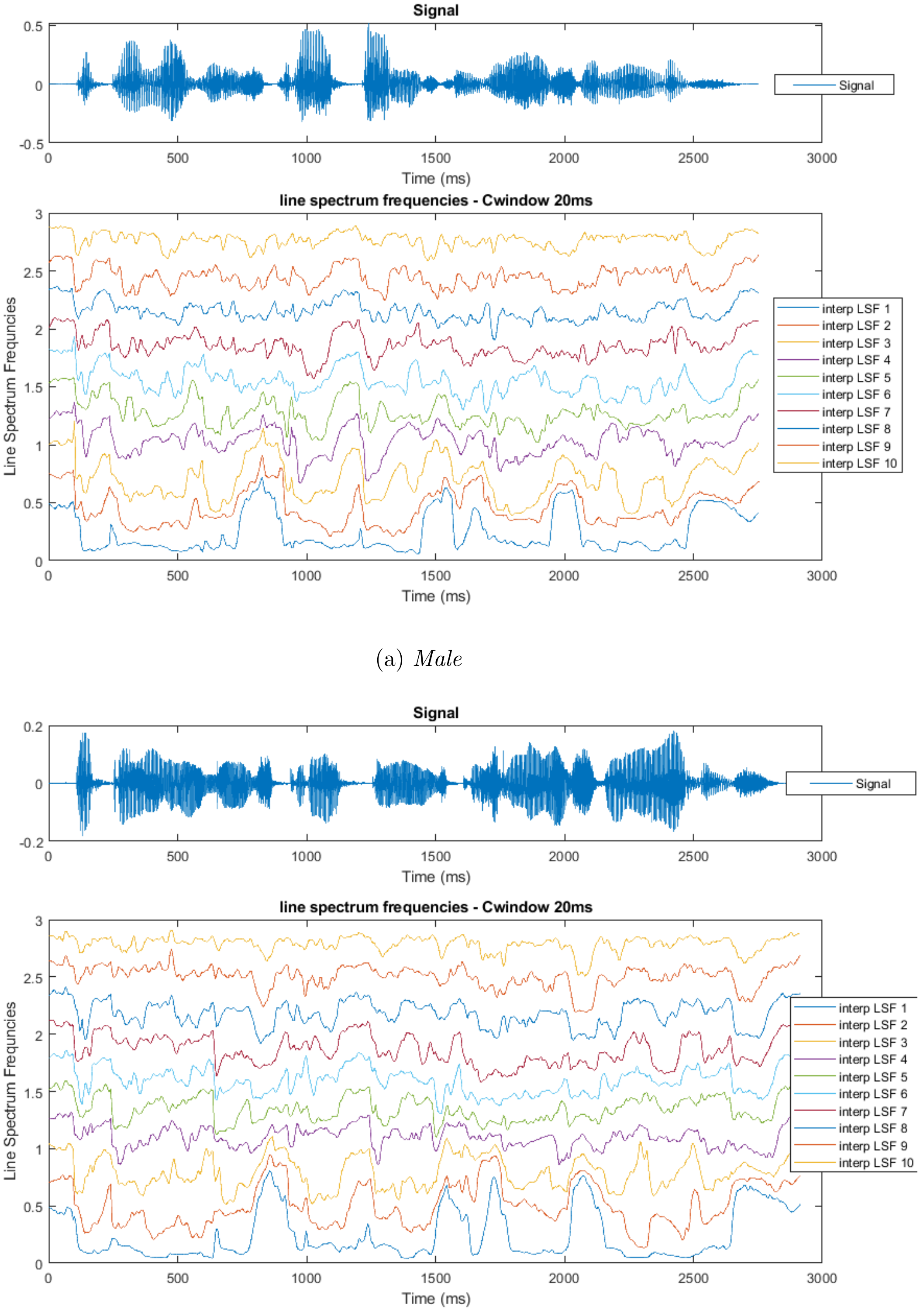
Fig.4.6 shows also the same speech segments as Fig.4.3 and Fig.4.4 along with the third criterion C_n^3. It is easily seen that criterion C3 combines C1 and C2.

Implementation detail C_n^3 uses β_2 and m_n from C_n^2 and C_n^1 itself for its calculation. That's why we also could not calculate the first value of C_n^3. For that reason, we again use the second frame's value as we did in C_n^1 and C_n^2.

$$C_n^3(1) = C_n^3(2) \tag{4.14}$$

4.2.3.4 Comparison

You can the differences between all the criteria (we explained above) in figure 4.7. C1 exhibits more significant discrimination but introduces some noise in voiced regions. On the other hand, C2 performs better in voiced regions but struggles in areas with sudden transitions. Lastly, C3 attempts to harness the advantages of both C1 and C2.



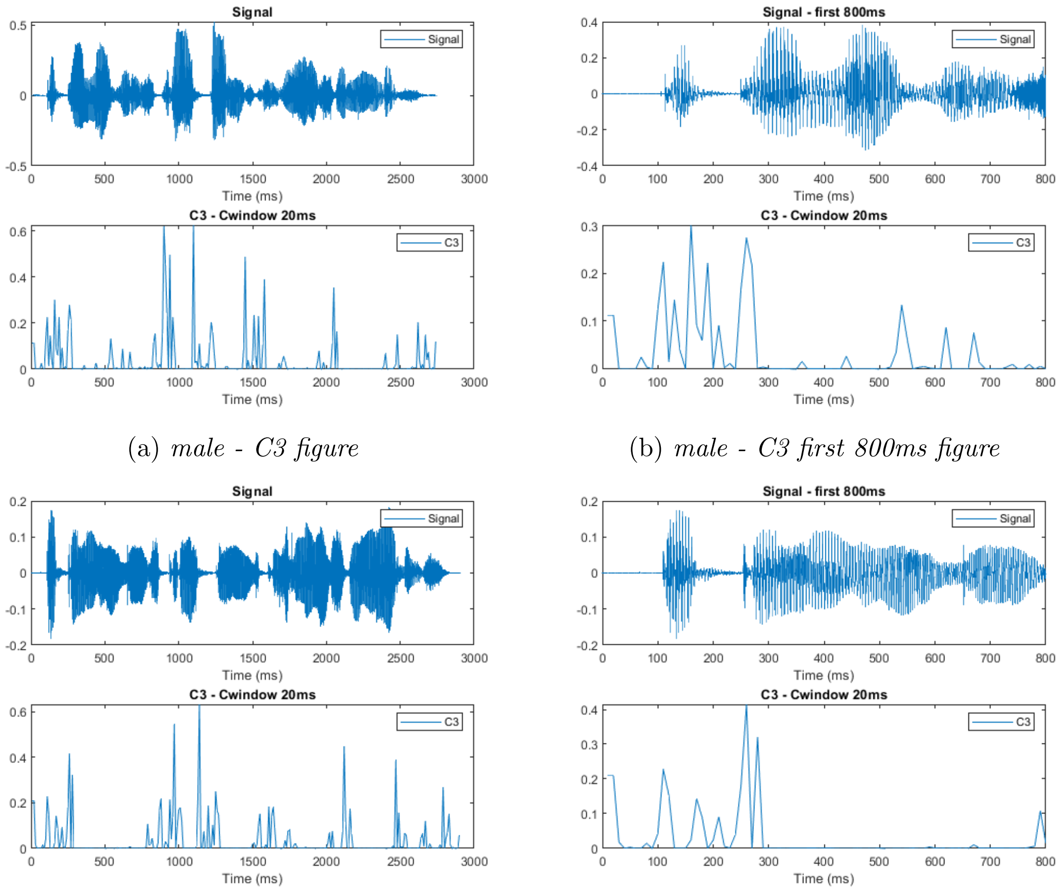
(a) *Male*

(b) *Female*

Figure 4.5: *Male voice and Female voice LSFs figures over the signal P1.*

4.2.3.5 Challenge

Even though it is said that the criteria are ready to be used in scale factor protection, we found that their values are not always inside 0 and 1 as you can see in Figures 4.3, 4.4, 4.6.



(a) *male - C3 figure*

(b) *male - C3 first 800ms figure*

(c) *female - C3 figure*

(d) *female - C3 first 800ms figure*

Figure 4.6: *Signal (top plots) and C3 criterion (bottom plots) of two recording versions (Male-Female) of file 'P1' extracted from [1]. The left figures contain the whole signal while the right figures contain only the first 800ms of the signal.*

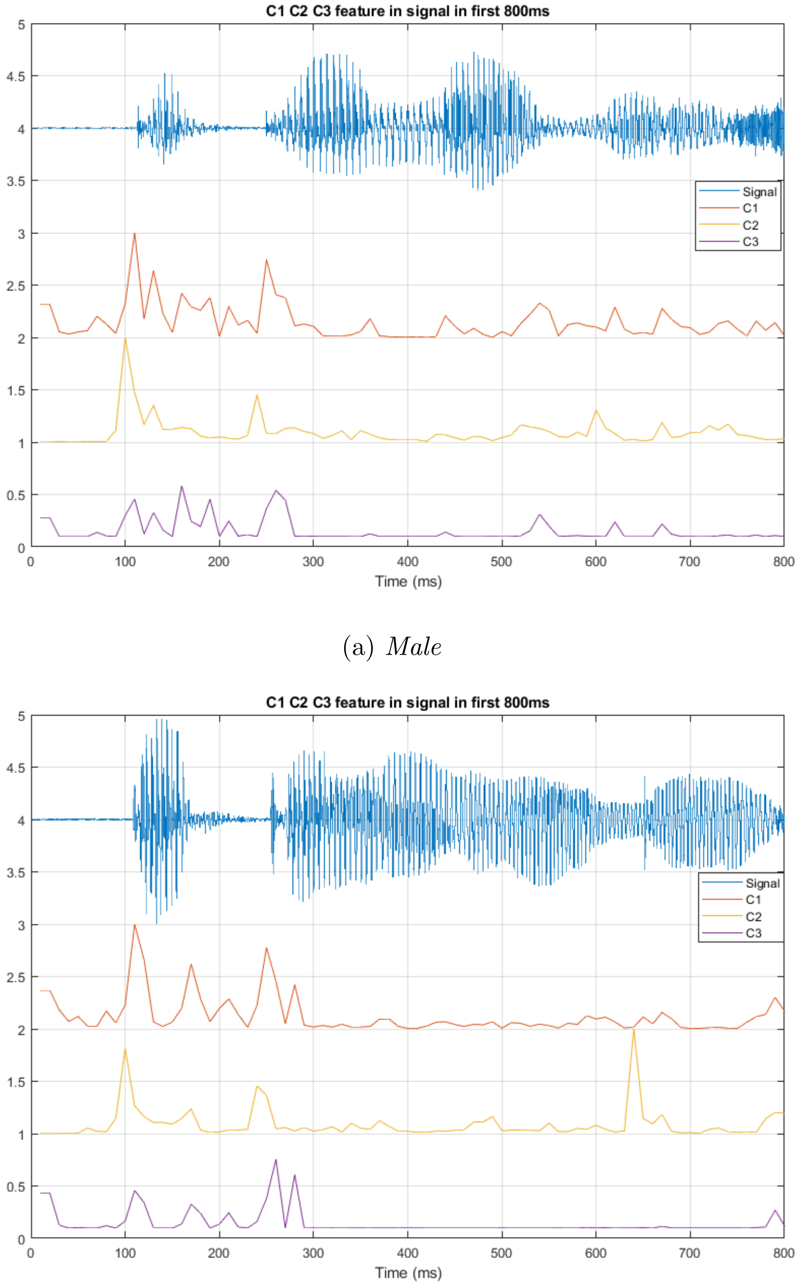
Especially C2's values are very small (10^{-5}). To solve this issue, after we calculate all the features, we normalize all features C_n^* so that:

$$0 \leq C_n^* \leq 1 \tag{4.15}$$

We interpolated them as well with extrapolation (in order to have one value per signal sample) to be used in the scale factor protection we will see in the next section.

4.2.4 Scale Factor Protection

To understand how this protection works, let us remember the problem at hand. Time domain techniques, such as TD-PSOLA and WSOLA, have been suggested for their simplicity and flexibility in time-scaling speech signals. However, when the time-scaling factor differs significantly from 1 (neutral), the quality of the time-scaled signal worsens, resulting in tonalities and artifacts in some parts of the output signal. These issues do not occur in the whole signal but can be found in transitional segments and thus in segments where the signal is non-stationary. So, to find regions where the signal is non-stationary, one of the 3 criteria can be used because all of them can be defined as the following function:



(a) *Male*

(b) *Female*

Figure 4.7: *Male (a) and Female (b) speech waveform along with C1, C2, C3 respectively.*

$$f(t) = \begin{cases} \sim 0, & \text{when a speech segment is stationary} \\ \sim 1, & \text{when a speech segment is non-stationary} \end{cases} \tag{4.16}$$

It can be seen in Fig. 4.8 at page 32 an ideal function $f(t)$ with the above characteristics.

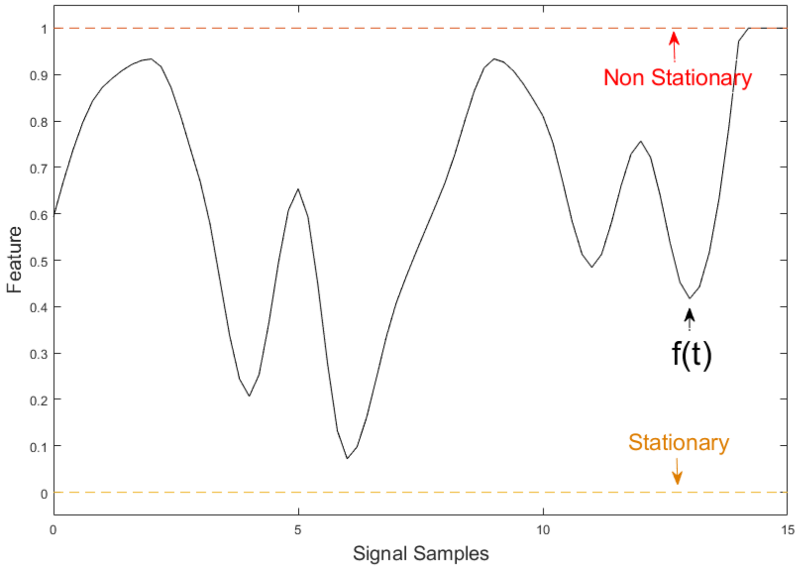


Figure 4.8: *An ideal function $f(t)$ for detection of non-stationarity.*

The protection function is the following:

$$\beta = 1 + d(t)b \tag{4.17}$$

where

- $d(t) = 1 - f(t)$ and has the opposite behavior than $f(t)$. $d(t) = 1$ when the speech segment is stationary while $d(t) = 0$ when the speech segment is non-stationary (and needs protection)

$$d(t) = \begin{cases} \sim 1, & \text{when a speech segment is stationary} \\ \sim 0, & \text{when a speech segment is non-stationary} \end{cases} \tag{4.18}$$

- β: The final scale factor of the output signal after protection with function $f(t)$ that is declared above.

- b: The desired relative modification of the original duration of the output signal. For stretching, For example, without protection ($d(t) = 1$), $b = 0.25$ means 25% stretching and thus $\beta = 1.25$. On the other hand, $b = -0.25$ means 25% shrinking and thus $\beta = 0.75$

Let's see a complete example of protection: If a speech segment is going to shrink by 25%, there are two cases as we examined before. The first case is to be stationary. This means that the shrinking is done without any protection:

$$b = -0.25 \tag{4.19}$$
$$f(t) \simeq 0 \tag{4.20}$$
$$d(t) = 1 - f(t) \simeq 1 \tag{4.21}$$
$$\beta = 1 + d(t)b \simeq 1 + 1 \cdot (-0.25) = 0.75 \tag{4.22}$$

The second case is to be non-stationary. This means that the more non-stationary is this part, the more this change will fade:

$$b = -0.25 \tag{4.23}$$

$$f(t) \simeq 1 \tag{4.24}$$

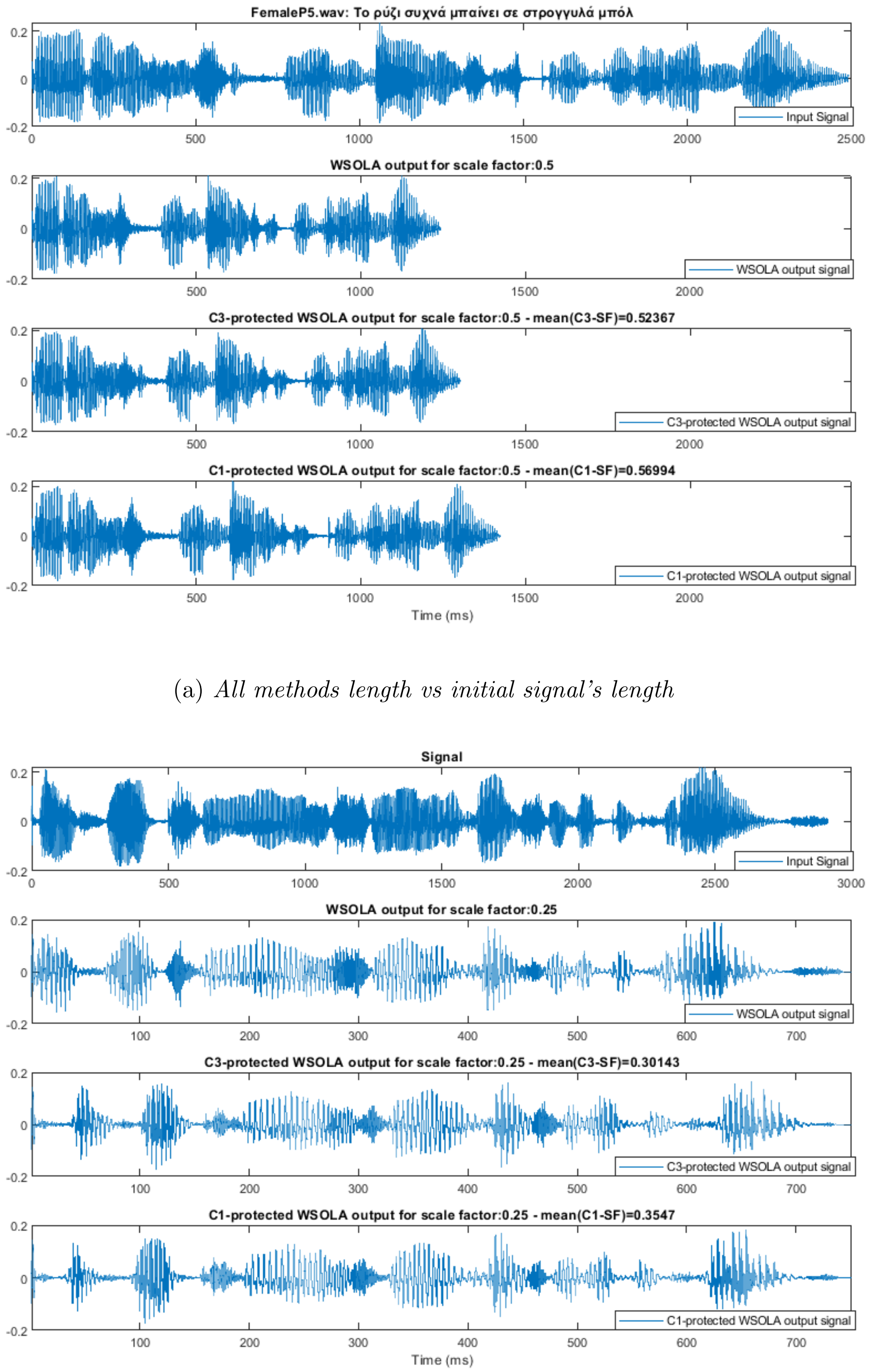
$$d(t) = 1 - f(t) \simeq 0 \tag{4.25}$$

$$\beta = 1 + d(t)b \simeq 1 + 0 \cdot (-0.25) = 1 \tag{4.26}$$

After the protection, we observed that C1 protects more than C3 as you can see in the figure 4.10. As a result, the output signal of C1 is larger than the output signal of C3 which is also larger than the output signal of uniform WSOLA. You can easily see it if you check in figure 4.9 in which comparing the three methods, WSOLA's average scale factor is 0.5 (uniform), C3-WSOLA's average scale factor is 0.52367 (non-uniform) and C1-WSOLA's average scale factor is 0.56994 (non-uniform). So,

$$\text{ScaleFactor}_{WS} \leq \text{ScaleFactor}_{C3} \leq \text{ScaleFactor}_{C1} \tag{4.27}$$

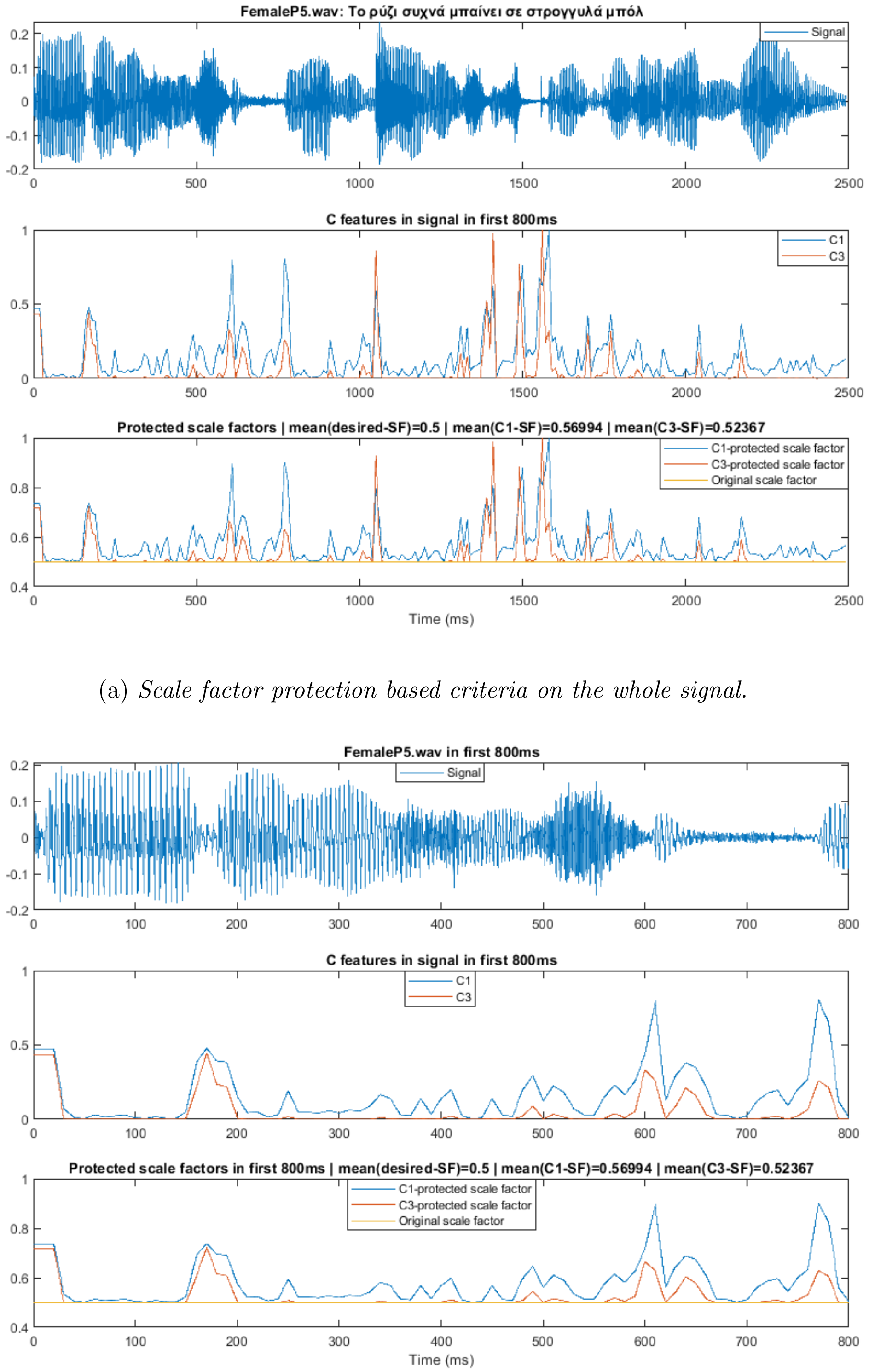
Additionally, in the second subfigure of Figure 4.9, it is evident how the non-uniform scale factor protects the input signal in non-stationary regions by slowing them down. As observed in the first 50 milliseconds of each plot, uniform WSOLA’s output is similar to the input signal, while in the next two outputs, there is a noticeable delay in the signals’ start.



(a) *All methods length vs initial signal's length*

(b) *All methods first 800 ms comparison*

Figure 4.9: *Comparing output signals generated by their respective methods.*



(a) *Scale factor protection based criteria on the whole signal.*

(b) *Scale factor protection based criteria on the first 800ms of the signal.*

Figure 4.10: *Scale factor protection: scale factors change as a function of the criteria values.*

Chapter 5

Results & Discussion

We performed two experiments of listening tests. The first experiment, conducted in June 2023, aimed to determine whether protecting non-stationarities enhanced or diminished the intelligibility of the signals. The second experiment was carried out in August 2023 and focused on comparing all methods under the same words-per-minute condition. We compared the methods: (1) uniform WSOLA, (2) protected WSOLA using C3, and (3) protected WSOLA using C1, differently in each experiment.

Listening tests were conducted on two groups of people. The first group consisted of visually impaired people (being in that condition for a lot of years) who used a screen reading app while the second one was the control group. For each individual, the conducted test consisted of two components: an intelligibility test and a preference test. All tests were created by randomly sampling a subset of corpus [1] (different subsets were used for the different types of tests). It is important to mention that we manually removed the initial and final silences from each sample in this dataset before using them.

All tests were conducted using a set of high-quality over-ear headphones (Seinheiser HD 650) connected to a dedicated laptop. Most tests involving subjects from the control group were performed within a soundproof room. In contrast, tests involving subjects from the visually impaired group were conducted at their homes or the corresponding Association. All subjects had no reported speech, language, or learning difficulties.

In the subsequent sections, we will delve into each experiment in greater detail, presenting and analyzing the results.

5.1 Experiment 1: Methods Comparison

In this experiment, our purpose was to assess whether the incorporation of protective measures discussed in chapter 4 had a positive or negative impact on the intelligibility of speech signals.

5.1.1 Participants

We conducted these tests with a total of 36 participants (24 males and 12 females, 20.8±0.8 and 23.1 ± 1.3 years old, respectively) from the control group, along with three visually impaired individuals (2 males and 1 female). Finding additional visually impaired participants in Heraklion proved to be very challenging, as we were informed that many were unavailable during our testing period or had permanently relocated to another place in Greece. Consequently, our analysis will primarily emphasize individuals from the control group. All participants were informed about the experiment, its goals, its limitations, and all signed an informed consent form. Consent forms are collected and safely stored by thesis supervisor, Prof. Yannis Stylianou. All consent forms will be destroyed two years after the completion of this thesis.

5.1.2 Description

For this experiment, we created a listening test for each user, with samples that were randomly sampled from a subset of the main corpus [1]. This test was used to analyze both the intelligibility and preference of the users. We first performed the **intelligibility test** because users had to **listen** to every sample **only once** before giving their answers about what they understood. After, the same test was used as a preference test and users selected their preferred method in each scale factor. This was valid because, in the **preference test**, they could **listen** to every sample as **many times** as they wanted. In Figure 7.1 (Appendix), the reader can find an example of a test used as an intelligibility test at first and then as a preference test.

5.1.3 Scale factors

In both tests (intelligibility and preference) we tested the signal's scale factors

$$\text{scale factors} = [0.1, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50] \tag{5.1}$$

In other words, we tested signals with speedups

$$\text{speedup factors} = [10.0, 6.6, 5.0, 4.0, 3.3, 2.8, 2.5, 2.2, 2.0] \tag{5.2}$$

respectively (1.0 is zero speedup). We tested a lot of speedups to find the first speedup which made it difficult for users to understand the content of the samples. Every speedup was calculated from the scale factor as follows:

$$\text{speedup} = \frac{1}{\text{scale factor}} \tag{5.3}$$

Speedups were presented to users in ascending order from the slower one (x2) to the faster one (x10). For every one of them, we randomly selected a sample from a specific subset (12 male and 12 female voices) from GrHarvard speech corpus [1] without using any sample twice in the test, in the same or any other scale factor.

5.1.4 Samples per scale factor

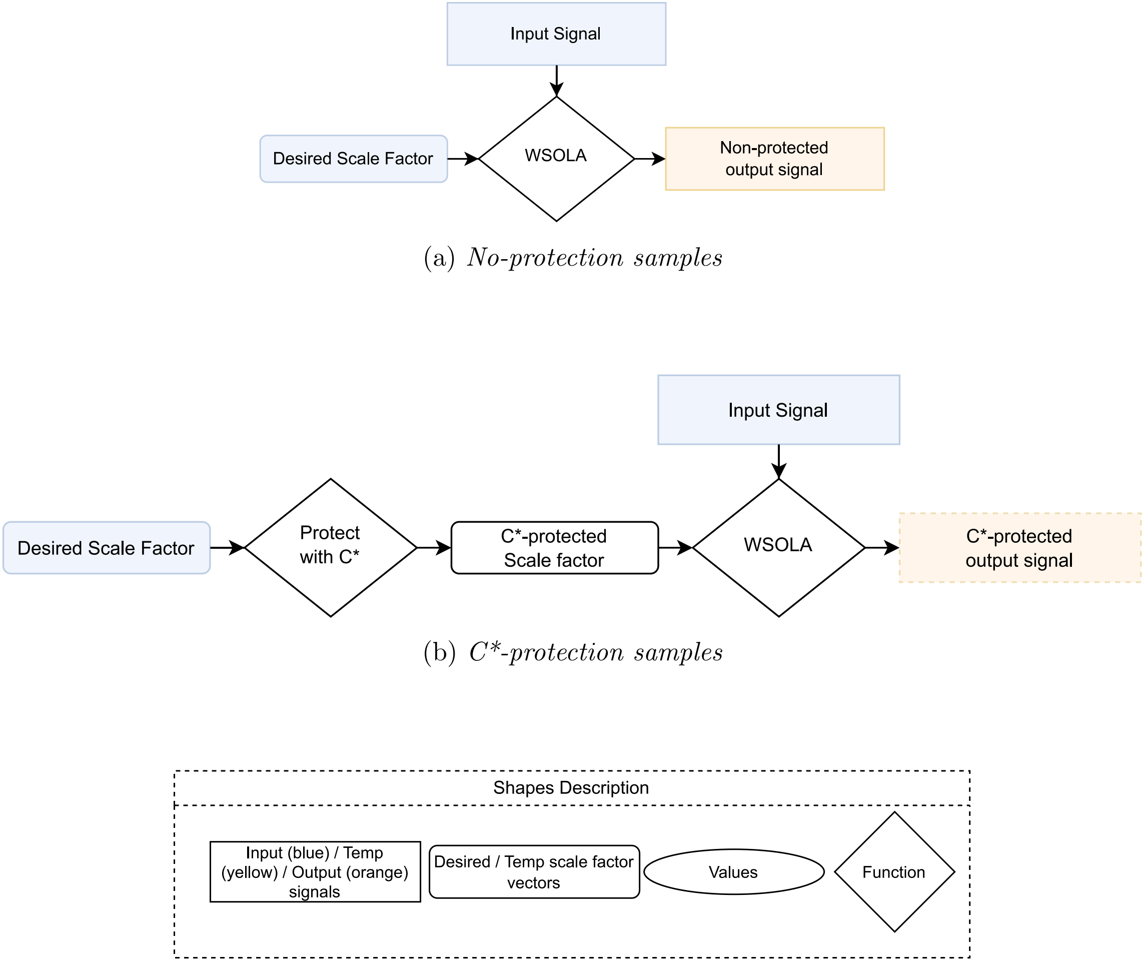
We use each selected sample to produce the following samples:

1. Non-protected WSOLA output sample: this is simple WSOLA output, using a constant time-scale factor resulting in a uniform time-scale compression.

2. C1-protected WSOLA output sample: this is simple WSOLA output, using a non-constant time-scale factor obtained from criterion C1, resulting in a non-uniform time-scale compression.

3. C3-protected WSOLA output sample: this is simple WSOLA output, using a non-constant time-scale factor obtained from criterion C3, resulting in a non-uniform time-scale compression.

Figure 5.1 illustrates a series of flowcharts that produce (a) samples with uniform time-scale compression and (b) samples with non-uniform, C_i-based time-scale compression. Figure 5.1(c) provides an explanation of each shape used in the flowcharts.


Input Signal
Desired Scale Factor
WSOLA
Non-protected output signal
Input Signal
Desired Scale Factor
Protect with C*
C*-protected Scale factor
WSOLA
C*-protected output signal
Shapes Description
Input (blue) / Temp (yellow) / Output (orange) signals
Desired / Temp scale factor vectors
Values
Function

(a) *No-protection samples*

(b) *C*-protection samples*

(c) *Flowchart labels*

Figure 5.1: *Experiment 1: System that produces time-scale compressed samples.*

We presented the three methods in the following order: the first method presented was a non-protected method WSOLA, the second method presented was a C3-protected method WSOLA and the last one presented a C1-protected method WSOLA. That is because more protection means slower (and larger) output and we know that C1 protects more than C3 (as it is shown in Figure 4.9).

5.1.5 Intelligibility test

In every intelligibility test, after listening to every method sample once, the users had to say exactly what they heard from it in order to measure the percentage of correct keywords found in each case. Before every sample, we used a warning tone that the sentence was about to begin. The final samples had the following form: [silence, warning tone, silence, signal].

5.1.5.1 Results & Discussion

As we can see in the Figures 5.3 and 5.2, we can draw the following conclusions given that all users listened to the samples of each scale factor starting with the 'fastest' sample (method of least protection) and ending with the 'slowest' sample (method of most protection):

- For scale factors > 0.40, we observe that all methods are almost balanced (high intelligibility in all methods $> 90\%$). This indicates that users found it challenging to distinguish differences in larger scale factors (slower speedups) among the samples. This slight difference may also be attributed to the fact that users listened to the same sample three times. The balance in the preference tests is further illustrated in Figure 5.6.

- For scale factors ≤ 0.40, we observe that C1-WSOLA was the most intelligible method with a big difference from the other methods (C3-WSOLA, WSOLA). That is because C1-WSOLA results were less time-compressed (more protected) than the other methods and users were able to understand the speech content better. The difference was high enough to bypass the bias from listening the same sample three times. This is because the results for C1-WSOLA were less time-compressed, indicating better preservation of the speech content compared to the other methods and thus better understanding from the users. This is why there is a noticeable difference in intelligibility, and it is significant enough to overcome any potential bias that might arise from listening to the same sample three times.

- The variance in correctly understood words was higher in the scale factors ranging from 0.2 to 0.4. However, it was not so high for very large and very small scale factors. Nearly everyone could comprehend the samples under the large scale factors (0.45 and 0.5), while only a few users were able to understand those at the smaller scale factors (0.1 and 0.15).

- The results might have been more consistent if we could have included more users from this group, but as previously mentioned, this was quite challenging. This is why we can observe some unusual results, such as the percentage of correct words in the 0.35 scale factor shown in Figure 5.3c, where the percentage was lower than in the 0.30 scale factor.

Statistical Significance - Intelligibility For each scale factor, a separate one-way analysis of variance (ANOVA) test was conducted to assess the statistical significance of differences among the various methods. ANOVA techniques are designed to determine whether a set of group means (representing treatment effects) are equal or not. Rejecting the null hypothesis indicates that not all group means are equal. In our case, the null hypothesis, denoted as H_0, assumed that the results of all methods were drawn from the same distribution, while the alternative hypothesis, H_1, posited that there was at least one pair of methods that originated from different distributions. Detailed ANOVA results for each scale factor are provided inside Table 5.1.

ScaleFactor	Source	SS	df	MS	F	Prob > F
0.1	Columns	0.06838	2	0.03419	14.05	3.51877e-06
	Error	0.27744	114	0.00243		
	Total	0.34581	116			
0.15	Columns	0.31607	2	0.15803	15.32	1.2788e-06
	Error	1.1759	114	0.01031		
	Total	1.49197	116			
0.20	Columns	1.94581	2	0.97291	38.17	2.0422e-13
	Error	2.90564	114	0.02549		
	Total	4.85145	116			
0.25	Columns	6.1827	2	3.09137	64.15	2.16753e-19
	Error	5.4938	114	0.04819		
	Total	11.6766	116			
0.30	Columns	4.7646	2	2.38231	20.16	3.1912e-08
	Error	13.4723	114	0.11818		
	Total	18.2369	116			
0.35	Columns	4.875	2	2.43752	28.1	1.20233e-10
	Error	9.8903	114	0.08676		
	Total	14.7653	116			
0.40	Columns	1.01744	2	0.50872	7.13	0.0012
	Error	8.13333	114	0.07135		
	Total	9.15077	116			
0.45	Columns	0.36376	2	0.18188	4.42	0.0142
	Error	4.6959	114	0.04119		
	Total	5.05966	116			
0.50	Columns	0.00838	2	0.00419	3.03	0.0521
	Error	0.15744	114	0.00138		
	Total	0.16581	116			

Table 5.1: *Experiment 1: Intelligibility test: ANOVA Results per Scale Factor. Red values indicate that the null hypothesis H_0 cannot be rejected while purple values are close to the threshold, suggesting that they are close to being significant enough to reject the null hypothesis.*

We observe that scale factor 0.50 has a P-value equal to 0.05 and [0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45] have a P-value smaller than 0.05. This means that we can reject the null hypothesis H_0 for almost all the scale factors. Thus, there is a statistically significant difference in results between at least one pair of methods within each of these specific scale factors.

It's important to note that this outcome, by itself, doesn't specify which group means differ from one another. So, to compare methods inside the scale factors, for each scale factor, we performed a multiple-comparisons post-hoc test using the MATLAB's function multcompare(stats). This function takes as argument stats that is one of the outputs of the MATLAB's one-way ANOVA function anova1(test_results). [62] Detailed multi-comparison [63] test results for each scale factor are provided inside Table 5.2.

Based on the observations in Tables 5.1 and 5.2 and in Figures 5.4 and 5.5, we can

Method A	Method B	p-value
0.1-WSOLA	0.1-C3	1
0.1-WSOLA	0.1-C1	3.3936e-05
0.1-C3	0.1-C1	3.3936e-05
0.15-WSOLA	0.15-C3	1
0.15-WSOLA	0.15-C1	1.4789e-05
0.15-C3	0.15-C1	1.4789e-05
0.20-WSOLA	0.20-C3	0.58283
0.20-WSOLA	0.20-C1	9.5915e-10
0.20-C3	0.20-C1	1.4596e-09
0.25-WSOLA	0.25-C3	0.10219
0.25-WSOLA	0.25-C1	9.5604e-10
0.25-C3	0.25-C1	9.5617e-10
0.30-WSOLA	0.30-C3	0.023588
0.30-WSOLA	0.30-C1	1.634e-08
0.30-C3	0.30-C1	0.0011251
0.35-WSOLA	0.35-C3	3.8539e-06
0.35-WSOLA	0.35-C1	1.0772e-09
0.35-C3	0.35-C1	0.076998
0.40-WSOLA	0.40-C3	0.050102
0.40-WSOLA	0.40-C1	0.0008691
0.40-C3	0.40-C1	0.36722
0.45-WSOLA	0.45-C3	0.11479
0.45-WSOLA	0.45-C1	0.012324
0.45-C3	0.45-C1	0.64606
0.50-WSOLA	0.50-C3	0.087766
0.50-WSOLA	0.50-C1	0.087766
0.50-C3	0.50-C1	1

Table 5.2: *Experiment 1: Intelligibility test: Multi-comparison results for all scale factor pairs. Red values indicate that the null hypothesis H_0 cannot be rejected while purple values are close to the threshold, suggesting that they are close to being significant enough to reject the null hypothesis H_0.*

deduce that in scale factors greater than 0.35, it becomes more challenging to reject the null hypothesis. This phenomenon is primarily attributed to the decreasing magnitude of differences in results among the various methods. Conversely, for scale factors below 0.30, the only null hypothesis H_0 that we are unable to reject is the one concerning the comparison between WSOLA and C3. This suggests that these two methods consistently exhibited similar performance.

5.1.6 Preference test

In every preference test, we presented the three methods as we did in the Intelligibility test but when we played the different methods to users we used a random order each time. We used this approach to prevent users from consistently choosing the same number in the method sequence. After users listened to all the samples of a specific speedup, they had to select the faster method that was intelligible as well (first, second, or third method). Thus, the main purpose here was to examine the method preference of users in each speedup. After the selection of the best methods from a user (in all scale factors), we downloaded them with the button "Show results".

5.1.6.1 Results & Discussion

As we can see in the figure 5.6, we can draw the following conclusions given that all users listened to the samples of each scale factor in random order:

- For scale factors > 0.40, the balance between methods can be easily seen here. It is because of the same reason we analyzed in Intelligibility test's results.
- For scale factors ≤ 0.40, we observe that C1-WSOLA was the most preferable method with again a big difference from the other methods (C3-WSOLA, WSOLA). That's also due to that users selected the most intelligible method, and thus the slowest method. If we combine this with the intelligibility test's results, in these scale factors they could barely understand 1-2.5 keywords (0-50% correct words in fig 5.3) with WSOLA and 1-3.5 words with C3-WSOLA (0-70% correct words in fig 5.3) while for C1-WSOLA it was 1-5 words (0-100% correct words in fig 5.3).
- Once again, the results from the control group were similar to those of the visually impaired group, although they could potentially be improved from a broader user base. Consequently, we cannot draw any conclusions for the visually impaired group.

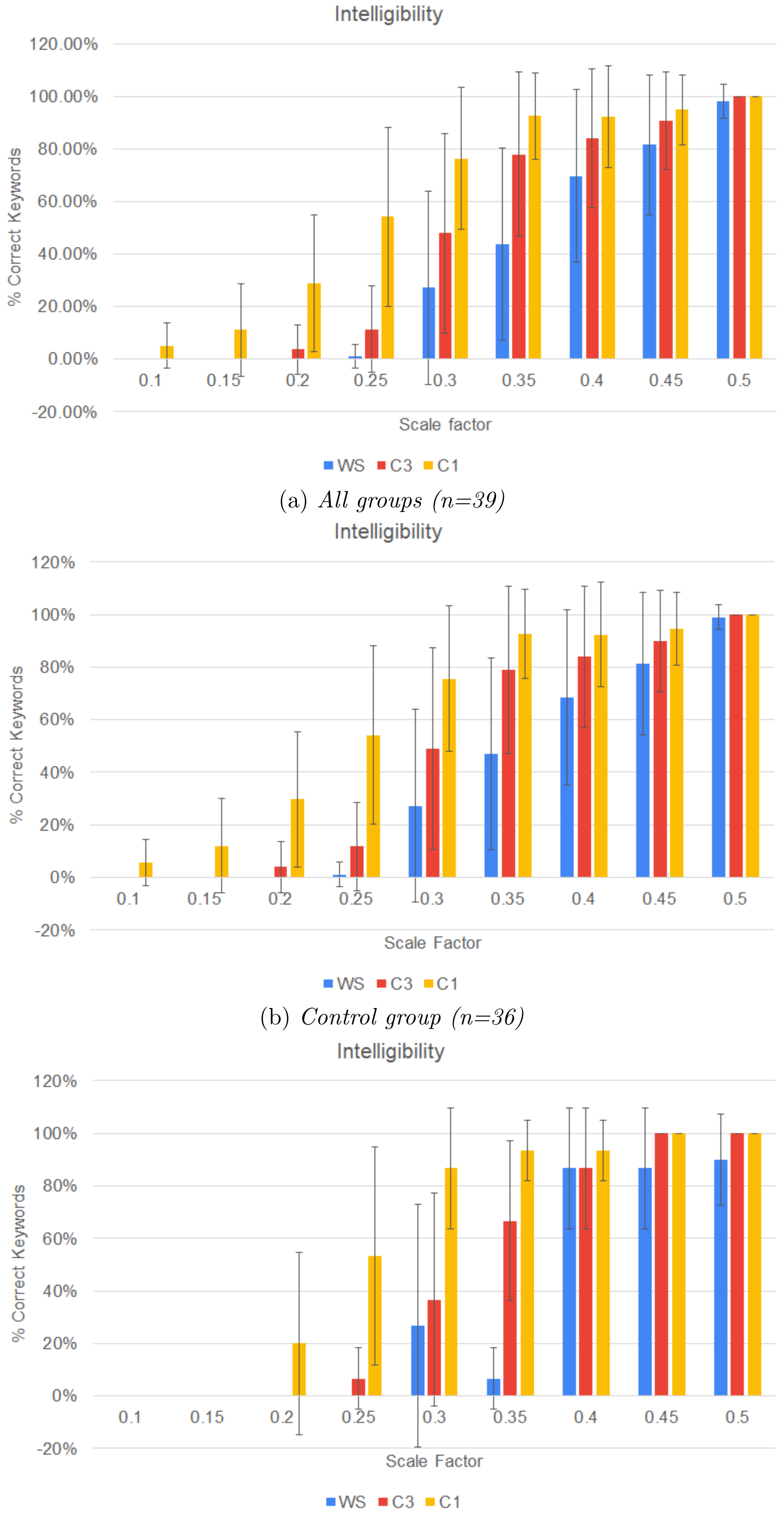
Statistical Significance - Preference For the preference tests, based on [64] [65], we performed a chi-squared test for each scale factor measuring the difference of our results (WS selections, C3 selections, C1 selections, None selections) from a uniform distribution. MATLAB's method 'chi2gof' [66] was used. The null hypothesis H_0 was that the distribution of the results followed a uniform while the alternative hypothesis H_1 stated that the results do not follow the uniform distribution. Results for each scale factor are provided inside the first tabular of Table 5.3. We observe that all p-values are below the significance threshold of 0.05, allowing us to reject the null hypothesis H_0.

Subsequently, in search of a method for conducting a post-hoc test on the preference test results, one approach similar to it was to calculate standardized residuals [67] [68]. They measure the difference of each value from the mean value of a uniform distribution among the results. In our case, residual values results can be found inside the second tabular of Table 5.3 in which they indicate that in the majority of the scale factors, C1 is the preferred method.

Chi-Squared Test Results									
SF	WS	C3	C1	Nn	Exp.WS	Exp.C3	Exp.C1	Exp.N	**Pvalue**
0.50	6	41	39	14	25	25	25	25	1.3773e-07
0.45	1	36	36	17	25	25	25	25	2.5092e-04
0.40	8	25	64	3	25	25	25	25	4.7973e-18
0.35	3	22	75	0	25	25	25	25	2.6098e-28
0.30	0	39	58	3	25	25	25	25	8.0802e-19
0.25	0	17	83	0	25	25	25	25	1.5019e-36
0.20	0	8	92	0	25	25	25	25	4.6440e-47
0.15	0	0	91	9	25	25	25	25	9.1176e-46
0.10	0	0	91	9	25	25	25	25	9.1176e-46

Chi-Squared Test Results with Residuals									
SF	WS	C3	C1	Nn	**Res.WS**	**Res.C3**	**Res.C1**	**Res.N**	Pvalue
0.50	6	41	39	14	-4.1288	2.5153	2.1356	-2.6102	1.3773e-07
0.45	1	36	36	17	-3.1797	1.5661	1.5661	-2.0407	2.5092e-04
0.40	8	25	64	3	-3.7492	-0.5220	**6.8814**	-4.6983	4.7973e-18
0.35	3	22	75	0	-4.6983	-1.0915	**8.9695**	-5.2678	2.6098e-28
0.30	0	39	58	3	-5.2678	2.1356	**5.7424**	-4.6983	8.0802e-19
0.25	0	17	83	0	-5.2678	-2.0407	**10.4882**	-5.2678	1.5019e-36
0.20	0	8	92	0	-5.2678	-3.7492	**12.1967**	-5.2678	4.6440e-47
0.15	0	0	91	9	-5.2678	-5.2678	**12.0068**	-3.5593	9.1176e-46
0.10	0	0	91	9	-5.2678	-5.2678	**12.0068**	-3.5593	9.1176e-46

Table 5.3: *Experiment 1: Preference test: Chi-squared test results with Residuals (WS: Uniform WSOLA, C3: C3-Protected WSOLA, C1:C1-Protected WSOLA, Nn: None, Exp: Expected results if we had a uniform distribution, Res: Standardized Residual value). All p-values are small enough to reject the null hypothesis.*

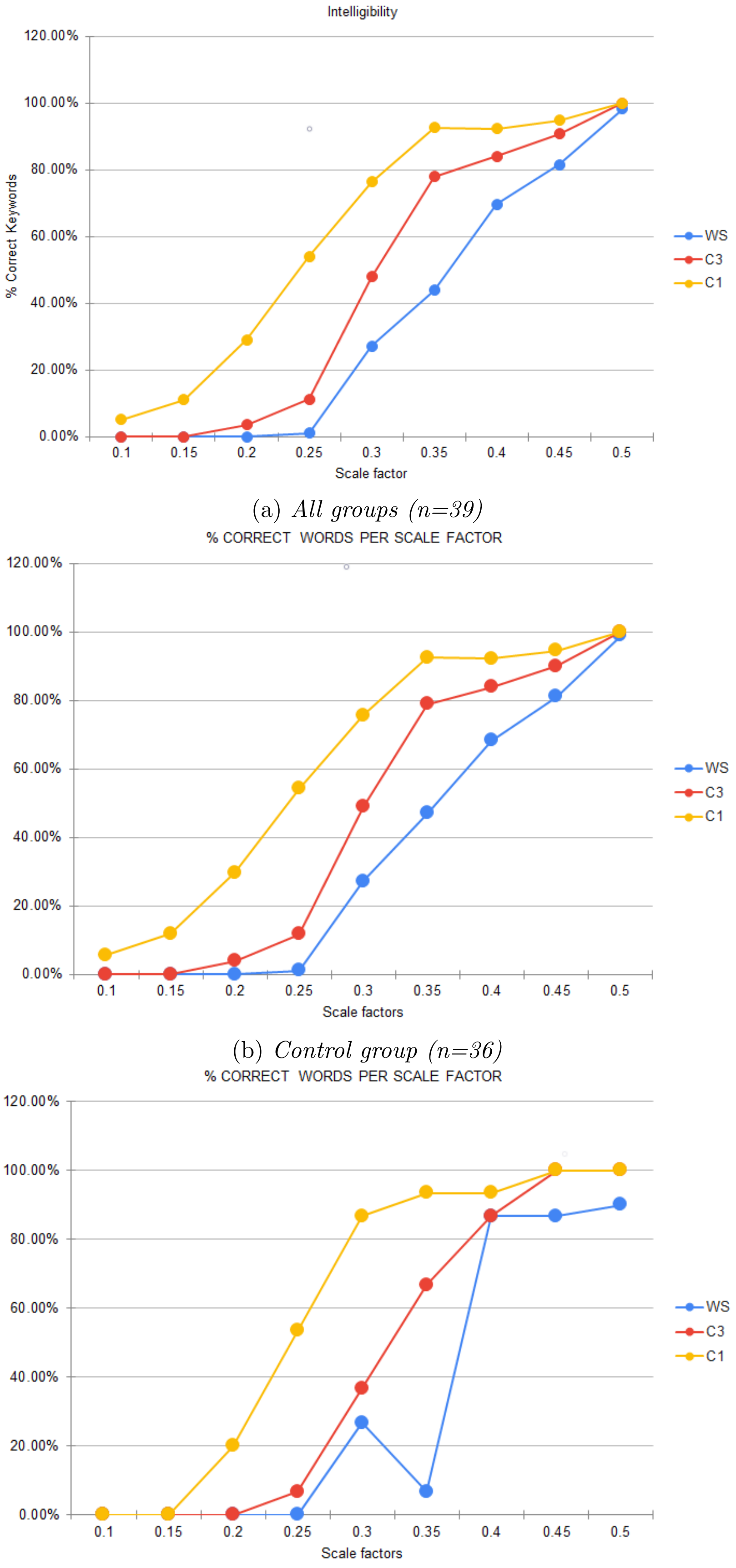


(a) *All groups (n=39)*

(b) *Control group (n=36)*

(c) *Visually impaired group (n=3)*

Figure 5.2: *Experiment 1: Intelligibility Listening Tests Results with bar plots.*

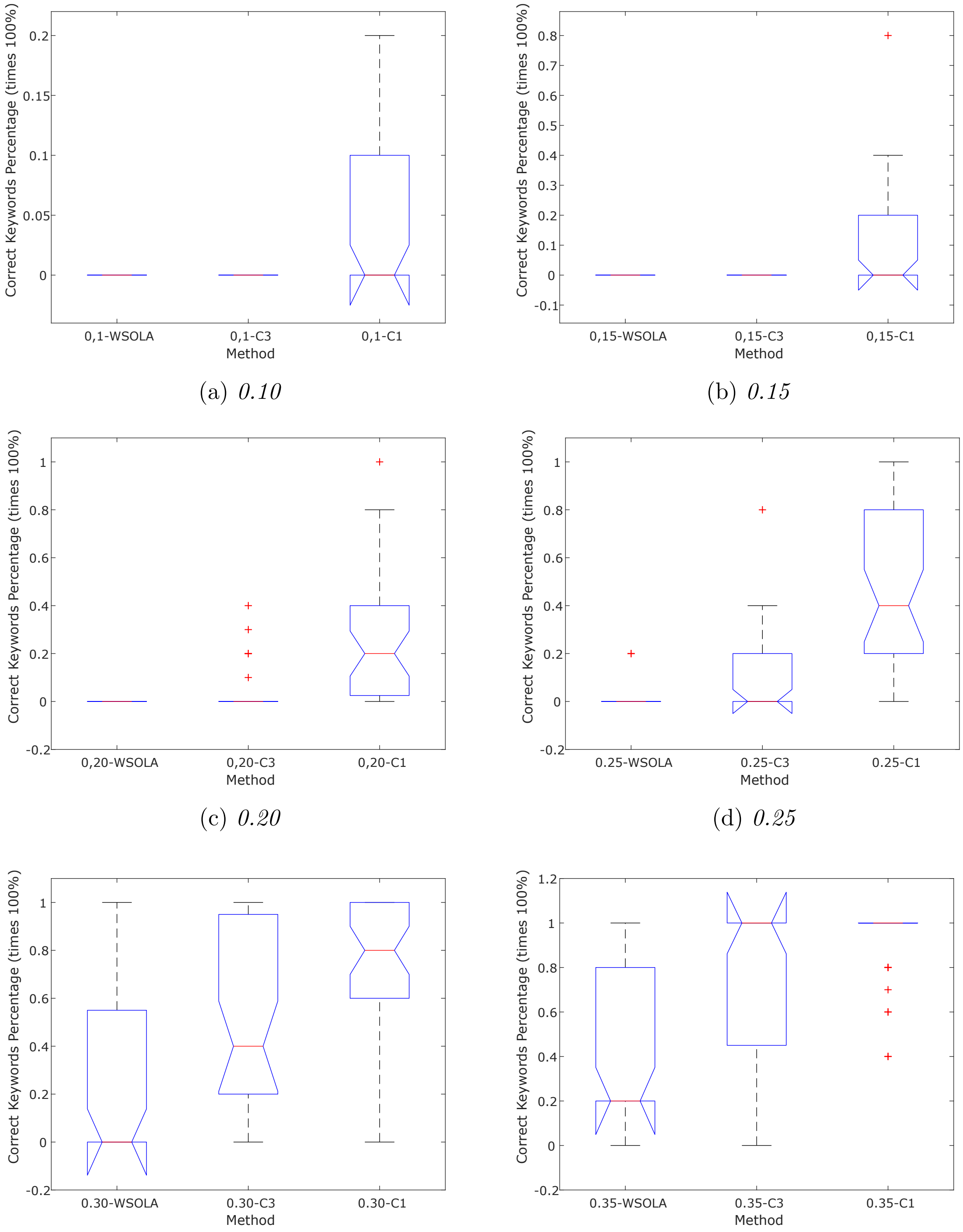


(a) *All groups (n=39)*

(b) *Control group (n=36)*

(c) *Visually impaired group (n=3)*

Figure 5.3: *Experiment 1: Intelligibility Listening Tests Results with line plots.*



(a) *0.10*

(b) *0.15*

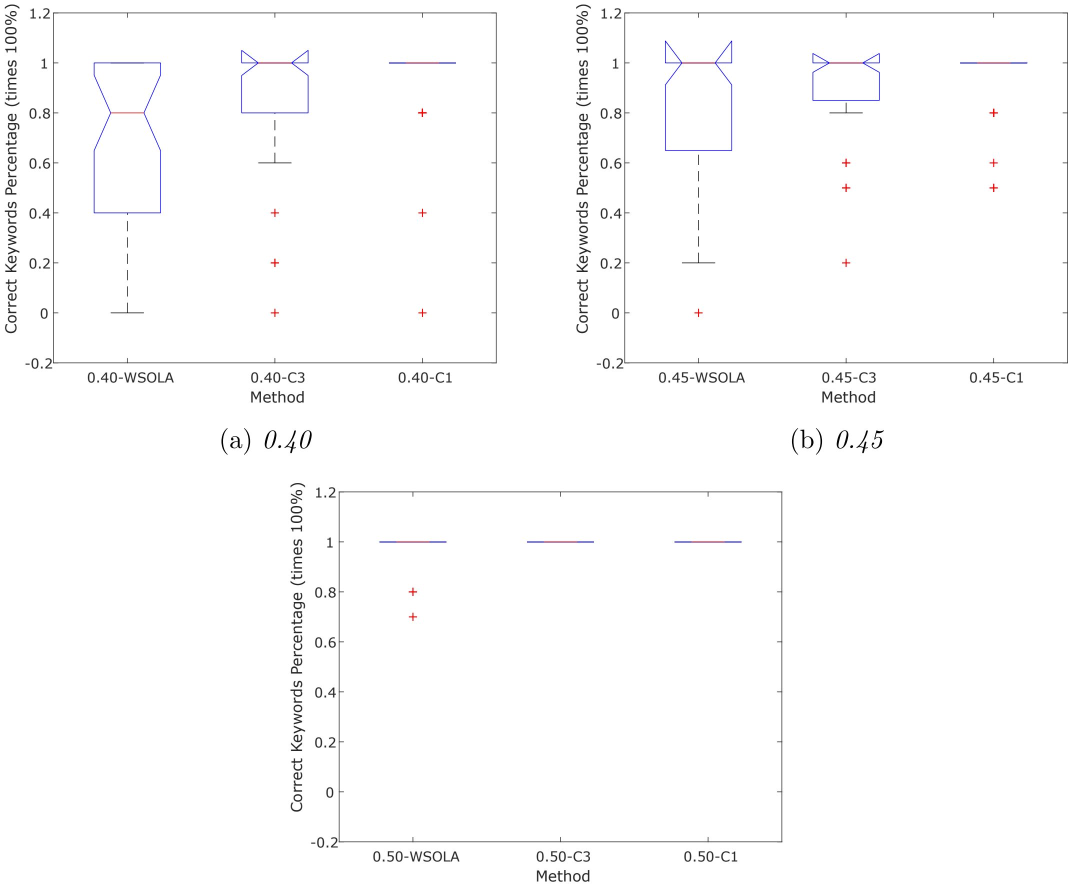
(c) *0.20*

(d) *0.25*

(e) *0.30*

(f) *0.35*

Figure 5.4: *Experiment 1: Listening test: Boxplots containing the three methods results in the following order WS-C3-C1. (PART A)*

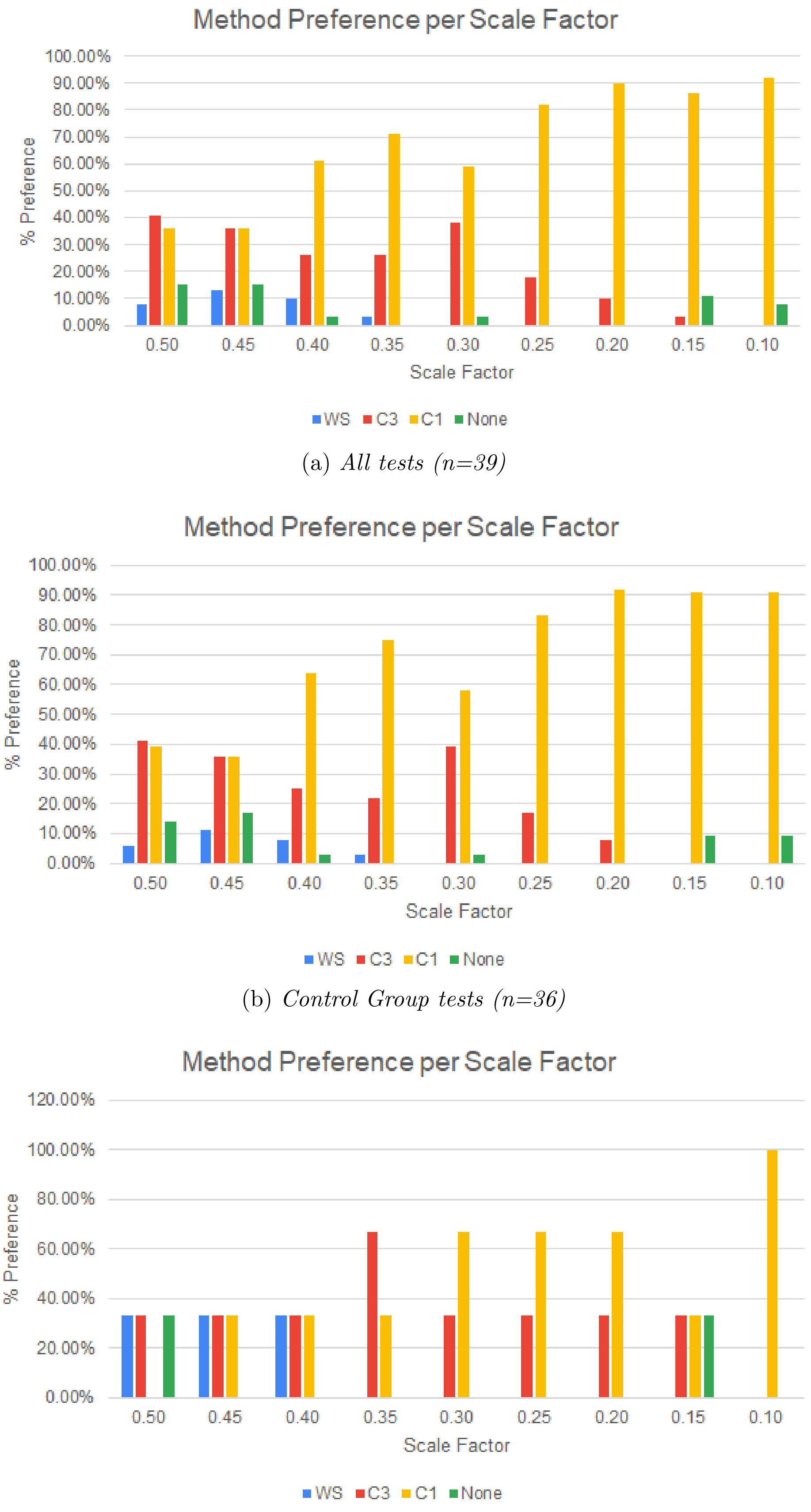


(a) *0.40*

(b) *0.45*

(c) *0.50*

Figure 5.5: *Experiment 1: Listening test: Boxplots containing the three methods results in the following order WS-C3-C1. (PART B)*



(a) *All tests (n=39)*

(b) *Control Group tests (n=36)*

(c) *Visually impaired Group tests (n=3)*

Figure 5.6: *Experiment 1: Preference Tests Results with bar plots.*

5.2 Experiment 2: Same words per minute

In this experiment, our purpose was to compare all methods given that they produce samples with the same words-per-minute (WPM) ratio that is equal to uniform WSOLA's WPM. To do that, we pass all other methods outputs from a uniform WSOLA again with a scale factor that corrected their length as it can be seen in figure 5.7. Also, we solve some of the previous listening test issues discussed in the next section.

5.2.1 Experiment 1: Challenges and new approaches

After the first experiment, we identified numerous issues:

1. Both Intelligibility and preference tests used the same subset of samples. Solution: We made them more independent by choosing disjoint subsets.

2. We used the same sample in all methods of a scale factor. This made it easier for the user to focus only on the words that did not listen the first time (especially in the intelligibility test). Solution: In this test, we used different samples in the different methods per scale factor.

3. We found only 3 visually impaired individuals in Crete in the first experiment. Solution: This time, we also went to the local association of visually impaired people in Kalamata and Ioannina, Greece, where we found 6 more individuals willing to participate.

4. In the previous experiment, C3-WSOLA and (especially) C1-WSOLA had an advantage over the other methods because they made the output slower. The more protected the output signal was, the slower it was. Solution: This time, we forced all methods output to have the same size. This means that output samples had the same words per minute number.

5. In preference tests, it was hard for users to remember the differences between three samples for each scale factor. Solution: we now compare methods in couples (2-by-2).

This is why we performed a second set of tests aimed at addressing these concerns.

5.2.2 Participants

We conducted these tests with 21 participants (14 males and 7 females) from the control group and 9 individuals who were visually impaired. In order to find more visually impaired individuals, we visited more related associations including the ones in Heraklion, Kalamata, and Ioannina, Greece. Specifically, we found 5 participants in Heraklion and 4 in Kalamata. Similarly to the first experiment, informed consent forms are signed by all participants.

5.2.3 Description

For this experiment, we also created two listening tests for each user (an intelligibility and a preference test). Each listening test had samples that were randomly sampled from a different subset of the main corpus [1]. These subsets contained two (out of 72) phonetically balanced lists from the corpus.

We first performed the **intelligibility test** in which again users had to **listen** to every sample **only once** before giving their answers about what they understood. After, the second test was used as a preference test and users selected their preferred method in each scale factor.

5.2.4 Scale factors

In the previous experiment, we tested 9 scale factors with some of them having zero variance among the users' results. This is why we only tested the 3 most important scale factors in this experiment including [0.25, 0.33, 0.50] and thus speedups [4.0, 3.0, 2.0]. Speedups are also calculated with the same method:

$$\text{speedup} = \frac{1}{\text{scale factor}} \tag{5.4}$$

and presented to users in ascending order.

For every scale factor, we randomly selected a sample from a specific subset (listening and preference tests have different and disjoint subsets).

5.2.5 Samples per scale factor

We again used samples in subsets to produce the following three speech samples:

1. Non-protected WSOLA output sample: this is simple WSOLA output, using a constant time-scale factor resulting in a uniform time-scale compression.

2. C1-protected WSOLA output sample: this is simple WSOLA output, using a non-constant time-scale factor obtained from criterion C1, resulting in a non-uniform time-scale compression.

3. C3-protected WSOLA output sample: this is simple WSOLA output, using a non-constant time-scale factor obtained from criterion C3, resulting in a non-uniform time-scale compression.

but the difference now is that every one of them has the same size (none of them will have the advantage of the slower output anymore). The reader can see how in Algorithm 1 in page 54 and the system in Figure 5.7.

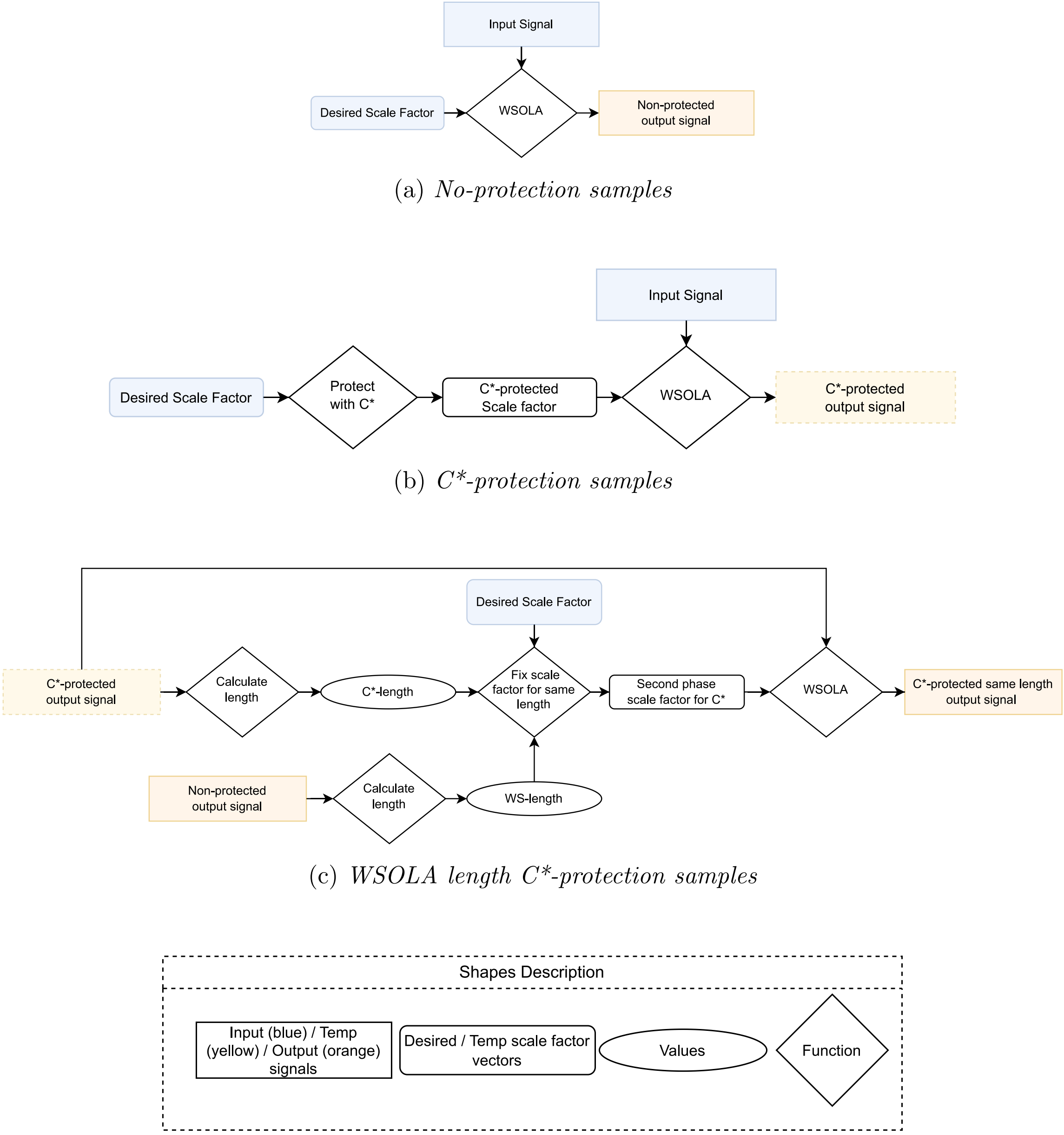
This means that since all of them have the same length, the differences will be inside the signals, in the stationary and non-stationary parts, More specifically, ”protection” with ”slow down” non-stationary parts while stationary parts will be ”sped up” in the output signal. You can see the differences between all methods in figure 5.8. On the top subfigure, it is evident that all methods produce output signals of the same length. However, in the bottom subfigure, you can observe the differences in the lengths of the stationary and non-stationary parts of the output signals.

5.2.6 Intelligibility test

To eliminate the bias introduced by listening to the same sample across all methods for a particular scale factor, this time, we used distinct samples. Methods were again presented the same way as before: (1) non-protected method WSOLA, (2) C3-protected method WSOLA, and (3) C1-protected method WSOLA as you can see in the following example. The reader can also see the basic idea in Algorithm 2. Again, we used a warning tone that the sentence was about to begin. So, the final samples had the following structure: [silence, warning tone, silence, signal]. An example of an intelligibility test can be found in Figure 7.2 (Appendix).

5.2.6.1 Results & Discussion

Drawing insights from the subfigures in Figure 5.9, we can conclude the following:


Input Signal
Desired Scale Factor
WSOLA
Non-protected output signal
Input Signal
Desired Scale Factor
Protect with C*
C*-protected Scale factor
WSOLA
C*-protected output signal
Desired Scale Factor
C*-protected output signal
Calculate length
C*-length
Fix scale factor for same length
Second phase scale factor for C*
WSOLA
C*-protected same length output signal
Non-protected output signal
Calculate length
WS-length
Shapes Description
Input (blue) / Temp (yellow) / Output (orange) signals
Desired / Temp scale factor vectors
Values
Function

(a) *No-protection samples*

(b) *C*-protection samples*

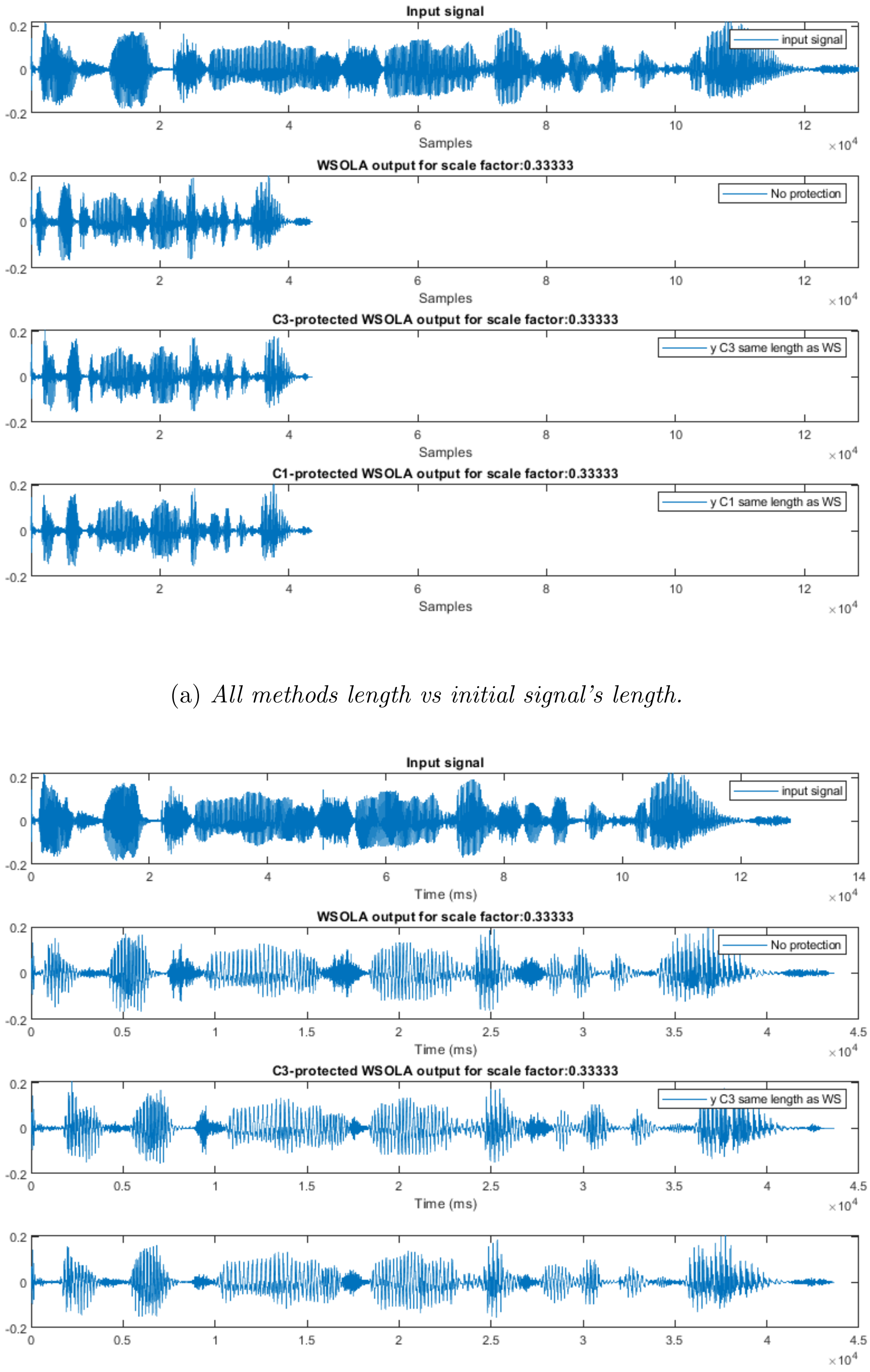
(c) *WSOLA length C*-protection samples*

(d) *Flowchart labels*

Figure 5.7: *Experiment 2: System that produces samples with the same words per minute.*

- The results exhibited a trend wherein, in the majority of cases, the C1 method yields the highest intelligibility value (with the exception of 0.25) but it's difficult to definitively determine which method is superior. The next two positions are typically occupied by the other two methods, C3 and WS. However, due to the same number of words per minute, it was quite challenging for the user to understand the distinctions between the compared methods, as further evidenced in the preference test results discussed below.

- The results from the control group and the visually impaired group are similar. We anticipated that visually impaired individuals might perform better, but the limited sample size made it challenging to discern significant differences.



(a) *All methods length vs initial signal's length.*

(b) *All methods full length comparison.*

Figure 5.8: *Same Words Per Minute Methods Output Comparison.*

Statistical Significance - Intelligibility For each scale factor, a separate one-way analysis of variance (ANOVA) test was again conducted to assess the statistical significance of differences among the various methods. The null hypothesis H_0 assumed again

Algorithm 1 *Create same size samples (WS, C1, C3) for tests*

```
Input: $X$: the input sample
Input: $SF$: the desired scale factor vector with the same size as the input sample
Input: $C$: the criterion that will be used for the protection of the input signal. It takes
  values [NP-WS, $C_1$-WS, $C_3$-WS] where 'NP-WS' means zero-protection pass to WSOLA,
  '$C_1$-WS' means $C_1$ protection but length equal to 'NP-WS' and '$C_3$-WS' means $C_3$
  protection but length equal to 'NP-WS'
Output: $Y$: the output sample
  //All variables starting with 'T' and 'L' are used for temporary calculations (signals and lengths)

  if $C$ = 'NP' then
      $Y$ = wsolaScale($X$, ..., $SF$);
  else if $C$ = '$C_1$' then
      //first calculate NP-WS
      $T_{WS}$ = wsolaScale($X$, ..., $SF$);
      $L_{WS}$ = length($T_{WS}$);
      //then calculate $C_1$ as we explained in chapter 4
      $SF_{C_1}$ = protect($SF$, '$C_1$');
      $T_{C_1}$ = wsolaScale($X$, ..., $SF_{C_1}$);
      $L_{C_1}$ = length($T_{C_1}$)
      //make $T_{C_1}$ with same duration as $T_{WS}$
      $SF_{input2wsola}$ = $SF$; //WSOLA Second pass - not always close to the desired scale after protection
      $SF_{C_1-WS}$ = ($L_{WS}$ / $L_{C_1}$) * ($SF$ / $SF_{input2wsola}$);
      $Y$ = wsolaScale($T_{C_1}$, ..., $SF_{C_1-WS}$);
  else if $C$ = '$C_3$' then
      //first calculate NP-WS
      $T_{WS}$ = wsolaScale($X$, ..., $SF$);
      $L_{WS}$ = length($T_{WS}$);
      //then calculate $C_3$ as we explained in chapter 4
      $SF_{C_3}$ = protect($SF$, '$C_3$');
      $T_{C_3}$ = wsolaScale($X$, ..., $SF_{C_3}$);
      $L_{C_3}$ = length($T_{C_3}$)
      //make $T_{C_3}$ with same duration as $T_{WS}$
      $SF_{input2wsola}$ = $SF$; //WSOLA Second pass - not always close to the desired scale after protection
      $SF_{C_3-WS}$ = ($L_{WS}$ / $L_{C_3}$) * ($SF$ / $SF_{input2wsola}$);
      $Y$ = wsolaScale($T_{C_3}$, ..., $SF_{C_3-WS}$);
  end if
```

that the results of all methods were drawn from the same distribution, while the alternative hypothesis, H_1, posited that there was at least one pair of methods that originated from different distributions. Detailed ANOVA results for each scale factor are provided in Table 5.1.

We observe that scale factors $[0.25, 0.50]$ have a small P-value, smaller or equal to 0.05. This means that we can (or are close to) reject the null hypothesis H_0. Thus, there is a statistically significant difference in results between at least one pair of methods within each of these specific scale factors.

This doesn't specify which group means differ from one another. So we again performed a post-hoc multiple-comparison test using MATLAB's function 'multcompare(stats)' [63]. where 'stats' was one of the outputs of MATLAB's one-way ANOVA function 'anova1' [62]. Detailed multi-comparison post-hoc results are presented in Table 5.5.

Algorithm 2 *Create an intelligibility test*

```
Input: $S$: the subset of corpus for intelligibility tests
Output: $I$: the intelligibility listening test (one per user)
  for scalefactor in $[0.5, 0.33, 0.25]$ do
      Randomly pick 3 samples $\alpha, \beta, \gamma$ (no replacement) from $S$
      Generate $WS$ sample from sample $\alpha$
      Generate $C1$ sample from sample $\beta$
      Generate $C3$ sample from sample $\gamma$
      Add warning tone in $\alpha$ as $\alpha_{new} = [silence, warning, silence, \alpha]$
      Add warning tone in $\beta$ as $\beta_{new} = [silence, warning, silence, \beta]$
      Add warning tone in $\gamma$ as $\gamma_{new} = [silence, warning, silence, \gamma]$
      Place them inside $I$
  end for
```

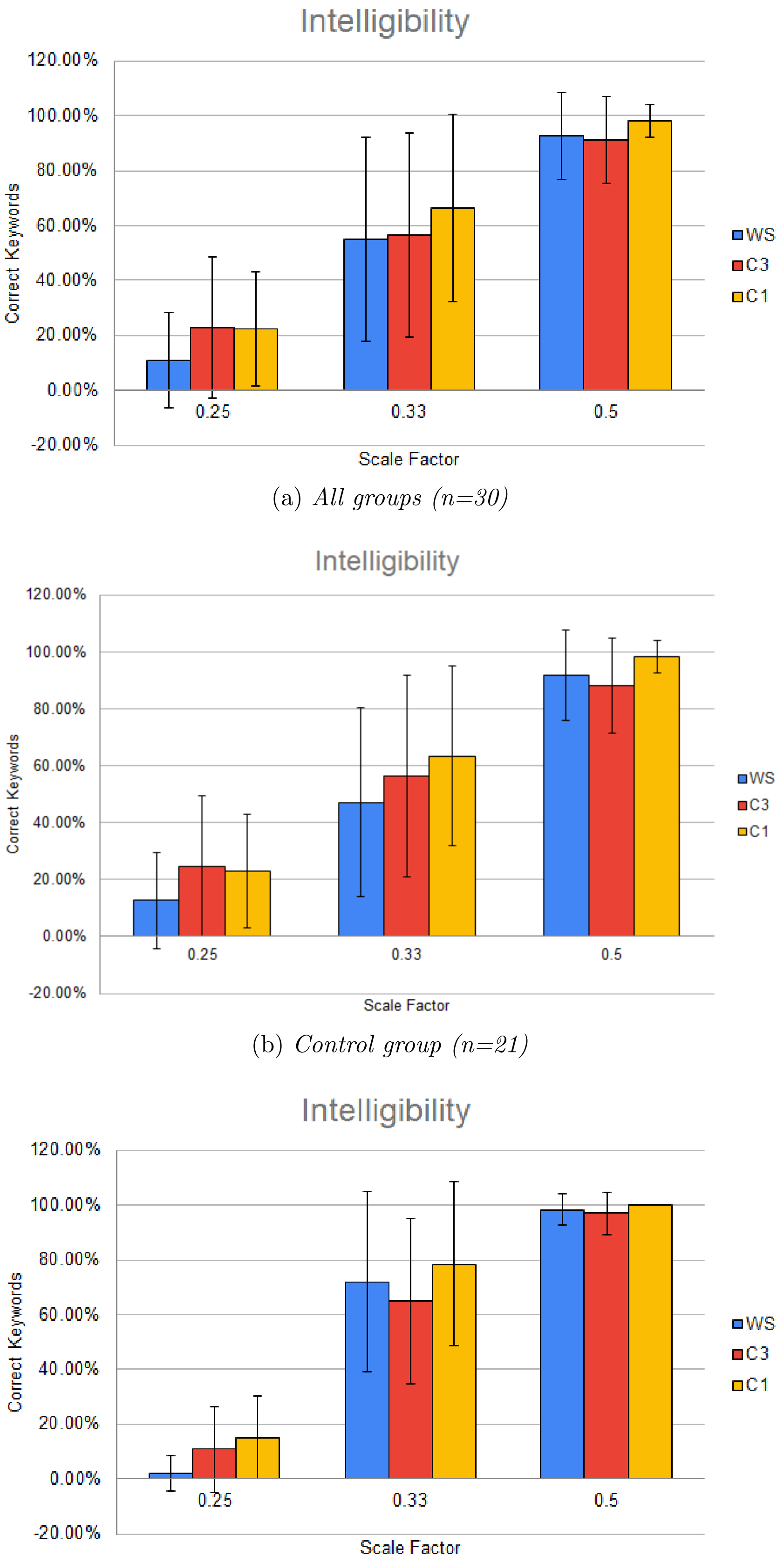
ScaleFactor	Source	SS	df	MS	F	p-value
0.25	Columns	0.242	2	0.121	3.07	0.0517
	Error	3.434	87	0.03947		
	Total	3.676	89			
0.33	Columns	0.3282	2	0.16411	1.47	0.2348
	Error	9.69	87	0.11138		
	Total	10.0182	89			
0.50	Columns	0.10689	2	0.05344	3.44	0.0364
	Error	1.351	87	0.01553		
	Total	1.45789	89			

Table 5.4: *Experiment 2: Intelligibility test: All groups ANOVA results per Scale Factor. Red values indicate that the null hypothesis H_0 cannot be rejected* while *purple values are close to the threshold, suggesting that they are close to being significant enough to reject the null hypothesis.*

Method A	Method B	p-value
0.25-WSOLA	0.25-C3	0.086915
0.25-WSOLA	0.25-C1	0.086915
0.25-C3	0.25-C1	1
0.3-WSOLA	0.3-C3	0.88818
0.3-WSOLA	0.3-C1	0.22513
0.3-C3	0.3-C1	0.45687
0.5-WSOLA	0.5-C3	0.62138
0.5-WSOLA	0.5-C1	0.22742
0.5-C3	0.5-C1	0.029968

Table 5.5: *Experiment 2: Intelligibility test: All groups results multi-comparison for all scale factor pairs. Red values indicate that the null hypothesis H_0 cannot be rejected* while *purple values are close to the threshold, suggesting that they are close to being significant enough to reject the null hypothesis H_0.*

Based on the observations in Tables 5.4, 5.5 and in Figure 5.10, we have arrived at the conclusion that rejecting the null hypothesis across all scale factors was a challenging task. This observation can be further substantiated by the results depicted in the intelligibility and preference test figures, where it becomes evident that the two methods consistently



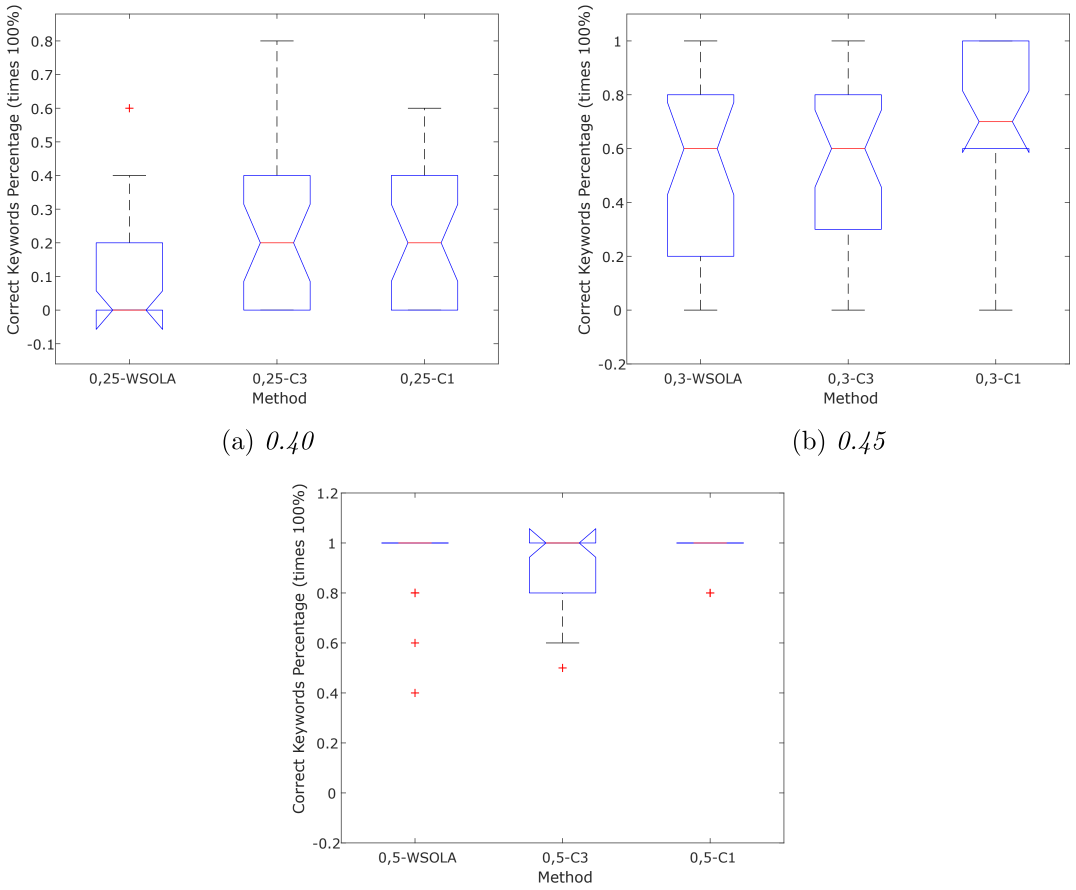
(a) *All groups (n=30)*

(b) *Control group (n=21)*

(c) *Visually impaired group (n=9)*

Figure 5.9: *Experiment 2: Intelligibility Test Results with bar plots.*

demonstrated nearly identical performance as we discussed above.



(a) *0.40*

(b) *0.45*

(c) *0.50*

Figure 5.10: *Experiment 2: Intelligibility test: All groups boxplots containing the three methods results in the following order WS-C3-C1.*

5.2.7 Preference test

In the preference test, our objective was to help users concentrate on comparing only two samples at a time. Consequently, for each scale factor, we offered three separate comparisons, each involving a different combination of methods from the set [WS-C1, WS-C3, C1-C3]. In every comparison, the order of methods was randomized to ensure that users made selections exclusively based on what they heard, rather than relying on memory from prior choices. The basic idea is described in Algorithm 3 on page 58 and an example of a preference test exists in Figure 7.3 (Appendix).

5.2.7.1 Results & Discussion

Preference test results can be found in Figures 5.11 and 5.12. Based on them we conclude the following:

- In Figure 5.11, a comparison of all methods is presented regarding how frequently they were selected. Notably, in scale factors 0.33 and 0.25, WS appears to be more preferable compared to the other methods. However, upon closer examination of the results, it becomes apparent that there is no dominant method. This suggests

Algorithm 3 *Create a preference test*

```
Input: $S$: the subset of corpus for preference tests
Output: $P$: the preference listening test (one per user)
  for scalefactor in $[0.5, 0.33, 0.25]$ do
      Randomly pick 3 samples $\alpha, \beta, \gamma$ (no replacement) from $S$
      Generate $WS - C3$ samples from sample $\alpha$
      Change their position with 50% probability
      Generate $WS - C1$ samples from sample $\beta$
      Change their position with 50% probability
      Generate $C1 - C3$ samples from sample $\gamma$
      Change their position with 50% probability
      Place them inside $P$
  end for
```

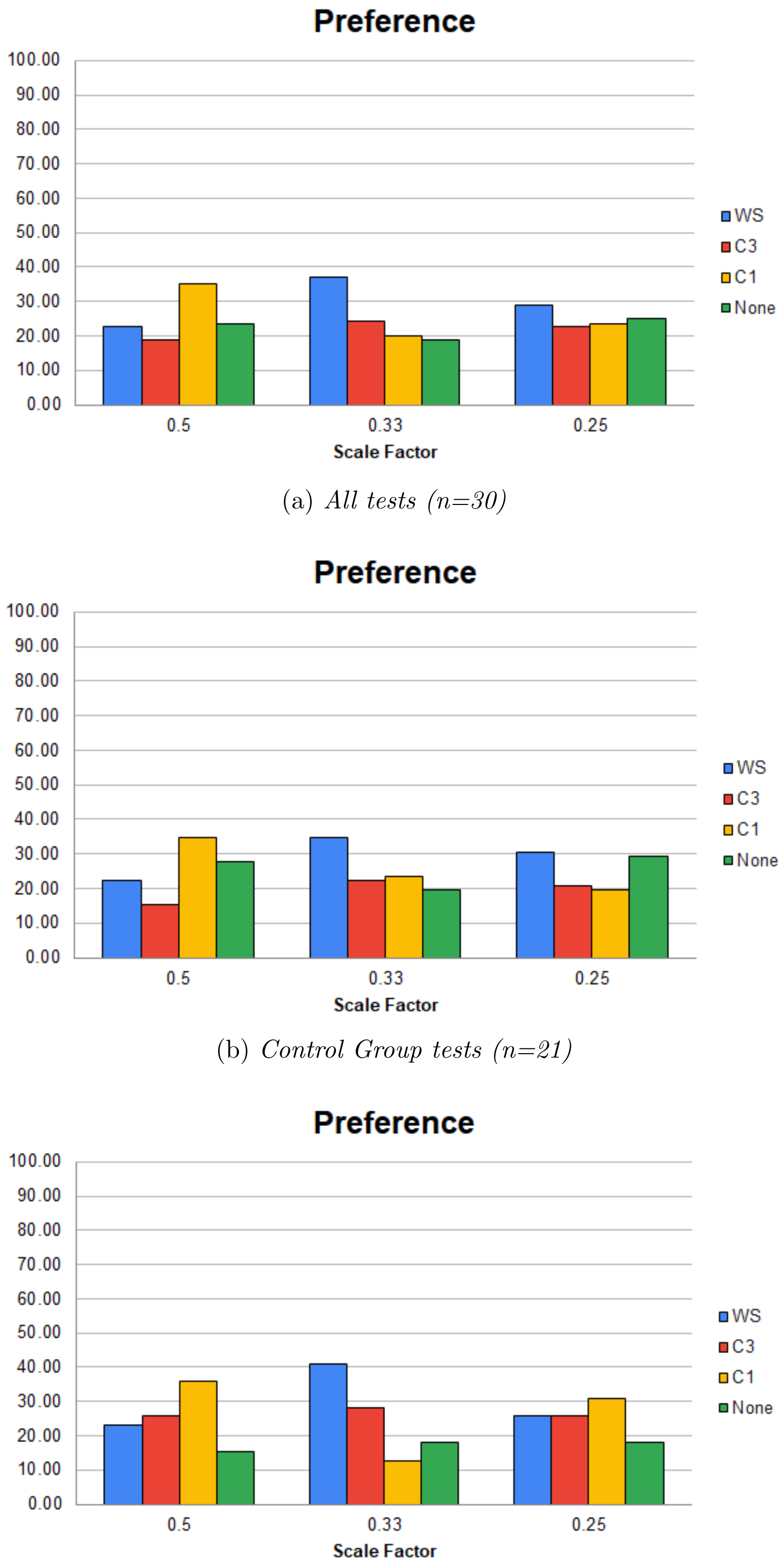
that all methods exhibit similar performance, and the differences among the samples are relatively minor and challenging for users to distinguish. Consequently, drawing further conclusions from these figures is a complex task.

- However, things are getting better when we compare methods in pairs as we do inside Figure 5.12. Comparing WS and C1, it is not immediately evident which one is dominant. Users preferred WS more at the scale factor of 0.33, while they favored C1 more at scale factors 0.25 and 0.5. When comparing WS and C3, WS was consistently more preferable in each scale factor but in the comparison between C1 and C3, a certain level of confusion emerged again, especially at a scale factor of 0.33. At the scale factor of 0.50, C1 was more preferable, whereas at 0.25, C3 received higher user preference. So, $WS \simeq C1$, $WS > C3$ and $C1 \simeq C3$

- Yet, both the control and visually impaired groups exhibited similar preferences, with the control group's preferences aligning more closely with the overall preferences due to its larger sample size of 21 individuals compared to 9 in the visually impaired group.

Statistical Significance - Preference For the preference tests, based on [64] [65], we performed a chi-squared test for each scale factor measuring the difference of our results (WS selections, C3 selections, C1 selections, None selections) from a uniform distribution. MATLAB's method 'chi2gof' [66] was used. The null hypothesis H_0 was again that the distribution of the results followed a uniform while the alternative hypothesis H_1 stated that the results do not follow the uniform distribution. Results for each scale factor are provided inside the first tabular of Table 5.6. We observe that the p-values of scale factors 0.33 and 0.50 are close rejecting the null hypothesis H_0 while the p-value of 0.25 is not.

Subsequently, in search of a method for conducting a post-hoc test on the preference test results, we again calculated standardized residuals [67] [68]. As we said earlier, they measure the difference of each value from the mean value of a uniform distribution among the results. Residual values results can be found inside the second tabular of Table 5.3 in which they indicate that there is no dominant method, especially in scale factor 0.25 where results were more uniform.

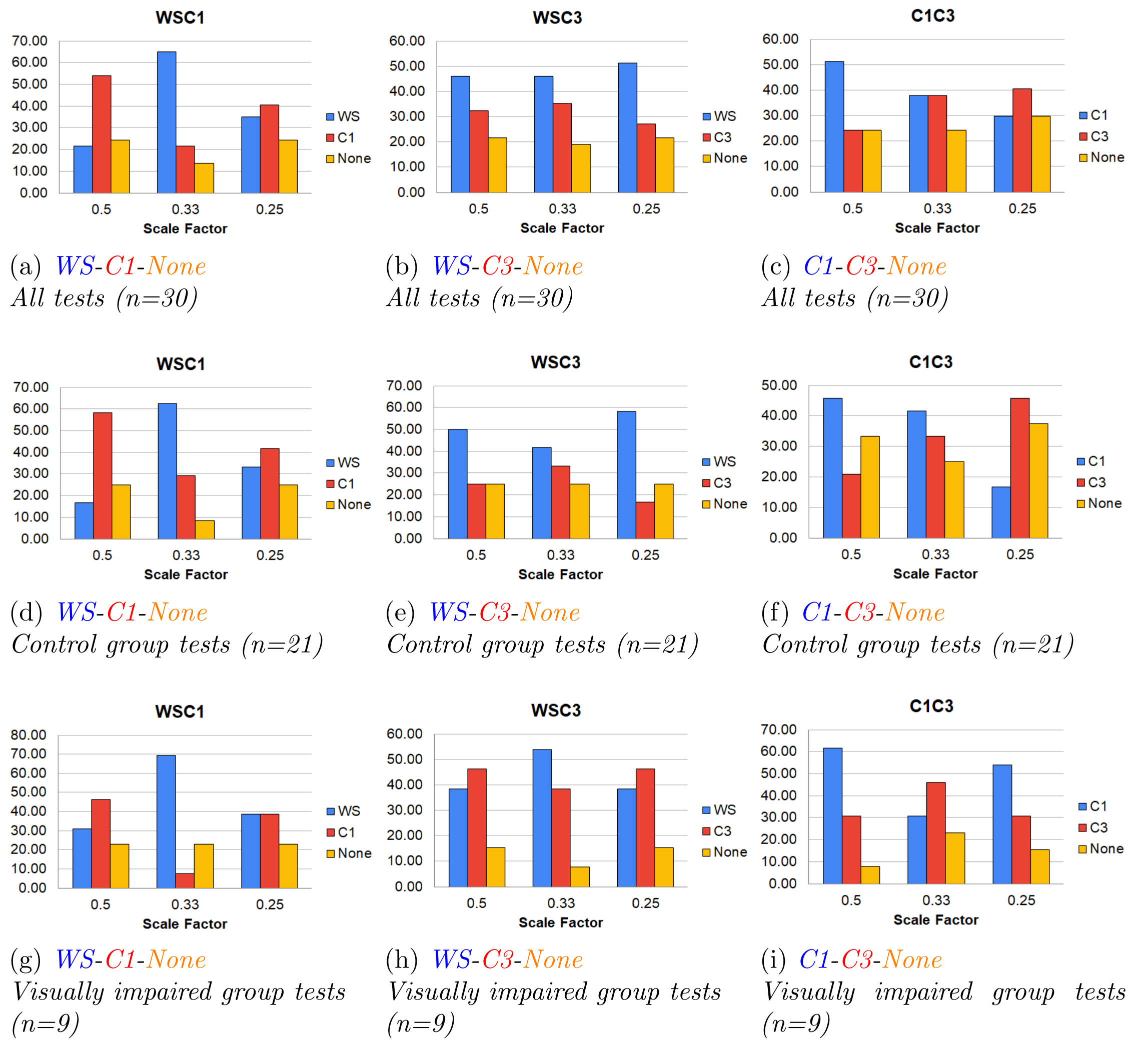


(a) *All tests (n=30)*

(b) *Control Group tests (n=21)*

(c) *Visually impaired Group tests (n=9)*

Figure 5.11: *Experiment 2: Preference Tests Results with bar plots.*



(a) *WS-C1-None*
All tests (n=30)

(b) *WS-C3-None*
All tests (n=30)

(c) *C1-C3-None*
All tests (n=30)

(d) *WS-C1-None*
Control group tests (n=21)

(e) *WS-C3-None*
Control group tests (n=21)

(f) *C1-C3-None*
Control group tests (n=21)

(g) *WS-C1-None*
Visually impaired group tests (n=9)

(h) *WS-C3-None*
Visually impaired group tests (n=9)

(i) *C1-C3-None*
Visually impaired group tests (n=9)

Figure 5.12: *Experiment 2: Preference Test Results with bar plots - 2-by-2 Methods. In the x-axis you can see the scale factors (0.50, 0.33 and 0.25 respectively)*

Chi-Squared Test Results									
SF	WS	C3	C1	Nn	Exp.WS	Exp.C3	Exp.C1	Exp.N	**Pvalue**
0.50	25	21	39	26	27.75	27.75	27.75	27.75	0.0863
0.33	41	27	22	21	27.75	27.75	27.75	27.75	0.0270
0.25	32	25	26	28	27.75	27.75	27.75	27.75	0.7925

Chi-Squared Test Results with Residuals									
SF	WS	C3	C1	Nn	**Res.WS**	**Res.C3**	**Res.C1**	**Res.N**	Pvalue
0.50	25	21	39	26	-0.5220	-1.2814	**+2.1356**	-0.3322	0.0863
0.33	41	27	22	21	**+2.5153**	-0.1424	-1.0915	-1.2814	0.0270
0.25	32	25	26	28	**+0.8068**	-0.5220	-0.3322	+0.0475	0.7925

Table 5.6: *Experiment 2: Preference test: Chi-squared test results (WS: Uniform WSOLA, C3: C3-Protected WSOLA, C1:C1-Protected WSOLA, Nn: None, Exp: Expected results if we had a uniform distribution, Res: Standardized Residual value).). In red values, null hypothesis H_0 cannot be rejected* while *purple values are close to the threshold.*

Chapter 6

Conclusion & Future Work

In this work, we demonstrated the effectiveness of signal processing methods based on the Waveform Similarity Overlap-Add (WSOLA) for highly intelligible time-scale compression of speech signals. Examples of time-scale compression in everyday life include listening to audiobooks and podcasts, as well as in online learning and speech transcription. Moreover, in the visually impaired population, high speech rates are very useful during screen-reading from a mobile phone, enabling them to maintain a connection to societal events.

WSOLA is a well-known algorithm for time scaling, able to handle both uniform and non-uniform time-scaling factors while minimizing the perceived artifacts that can occur when altering the tempo of a speech recording. WSOLA is known for its ability to provide high-quality time-stretching results with minimal audible artifacts. However, in very high time-scale compression rates, speech intelligibility decreases rapidly. We investigated the potential of protecting sensitive to time-compression parts of speech via signal processing techniques able to detect transient and non-stationary speech segments. The two methods, called C1 and C3, are based on signal RMS energy and Line Spectral Frequencies (LSFs): C1 detects non-stationarity by computing the RMS energy on a frame-by-frame basis while C3 combines the latter with the inclusion of the gradient of the regression line of LSFs over time as a measurement of transition rate. C1 and C3 can be used to build a time-varying function of stationarity of a speech signal. This function, when fed to WSOLA, was able to guide time-compression in a way that non-stationary and transient parts of speech are protected from time-scale compression.

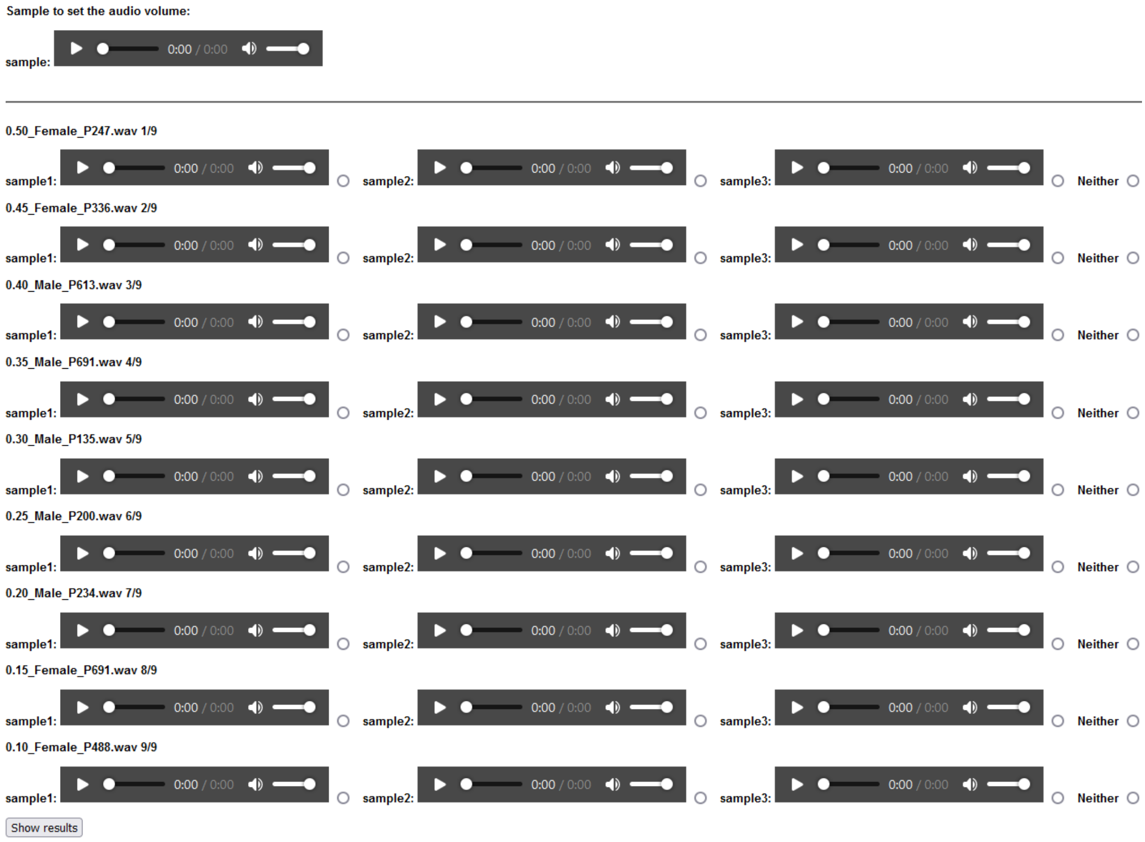
Experiments were conducted on four of the GrHarvard lists, encompassing both sighted and visually impaired participants. Following the experiments, a comprehensive statistical analysis was executed, employing ANOVA and post-hoc tests, to determine the significance of the differences in the results derived from the intelligibility tests of both experiments. In the first experiment, a comparison between uniform WSOLA, non-uniform C1-protected WSOLA, and non-uniform C3-protected WSOLA was conducted. The primary aim of this test was to evaluate whether the protective measures enhanced or diminished the intelligibility of the speech signals. The results revealed that the most protective method consistently outperformed the others in terms of both intelligibility and user preference. In our study, this top-performing method was C1-Protected WSOLA, the second position was occupied by the other protected method, C3-protected WSOLA, and the uniform WSOLA method took the last position. In this experiment, where substantial differences existed, most of the observed variations were indeed statistically significant. In the second experiment, we aimed to evaluate the same three methods under conditions where they produced speech at an equal number of words per minute. This approach led to more similar outcomes, making it challenging for users to distinguish significant differences between the methods. The variations in results were primarily attributed to differences within the signals, which related to the sizes of their stationary and non-stationary components. Despite the tendency for the C1 method to exhibit higher intelligibility (in most cases, except at 0.25), it remained challenging to conclusively determine the superior method due to the overall similarities in performance, as observed in both preference and intelligibility test results. Additionally, despite the initial expectations of enhanced performance in the visually impaired group compared to the control group, such differences did not materialize, primarily owing to the limited number of visually impaired participants. Consequently, the majority of observed differences failed to achieve statistical significance, despite the occasional discernible patterns among the methods.

Future work can include further tuning of parameters that affect C1 and C3 computation. As an example, different lengths of analysis and hop frames can be used, as well as pitch-synchronous analysis in stationary parts of speech. Moreover, further experiments - including a larger sample from visually impaired people - could strengthen statistical conclusions about the performance of each method. Furthermore, research can be done towards including these methods in popular audio processing tools and screen-reading software. Another noteworthy research direction is the training of a neural vocoder to generate controllable output signals, enabling adjustments to speed up and ensuring non-stationarity protection. Finally, further investigation into the analysis and preservation of the phonemes targeted by this method is a promising avenue of study.

Chapter 7

Appendix

7.1 More Figures



Sample to set the audio volume:

sample: ▶ 0:00 / 0:00

0.50_Female_P247.wav 1/9

sample1: ▶ 0:00 / 0:00 ○ sample2: ▶ 0:00 / 0:00 ○ sample3: ▶ 0:00 / 0:00 ○ Neither ○

0.45_Female_P336.wav 2/9

sample1: ▶ 0:00 / 0:00 ○ sample2: ▶ 0:00 / 0:00 ○ sample3: ▶ 0:00 / 0:00 ○ Neither ○

0.40_Male_P613.wav 3/9

sample1: ▶ 0:00 / 0:00 ○ sample2: ▶ 0:00 / 0:00 ○ sample3: ▶ 0:00 / 0:00 ○ Neither ○

0.35_Male_P691.wav 4/9

sample1: ▶ 0:00 / 0:00 ○ sample2: ▶ 0:00 / 0:00 ○ sample3: ▶ 0:00 / 0:00 ○ Neither ○

0.30_Male_P135.wav 5/9

sample1: ▶ 0:00 / 0:00 ○ sample2: ▶ 0:00 / 0:00 ○ sample3: ▶ 0:00 / 0:00 ○ Neither ○

0.25_Male_P200.wav 6/9

sample1: ▶ 0:00 / 0:00 ○ sample2: ▶ 0:00 / 0:00 ○ sample3: ▶ 0:00 / 0:00 ○ Neither ○

0.20_Male_P234.wav 7/9

sample1: ▶ 0:00 / 0:00 ○ sample2: ▶ 0:00 / 0:00 ○ sample3: ▶ 0:00 / 0:00 ○ Neither ○

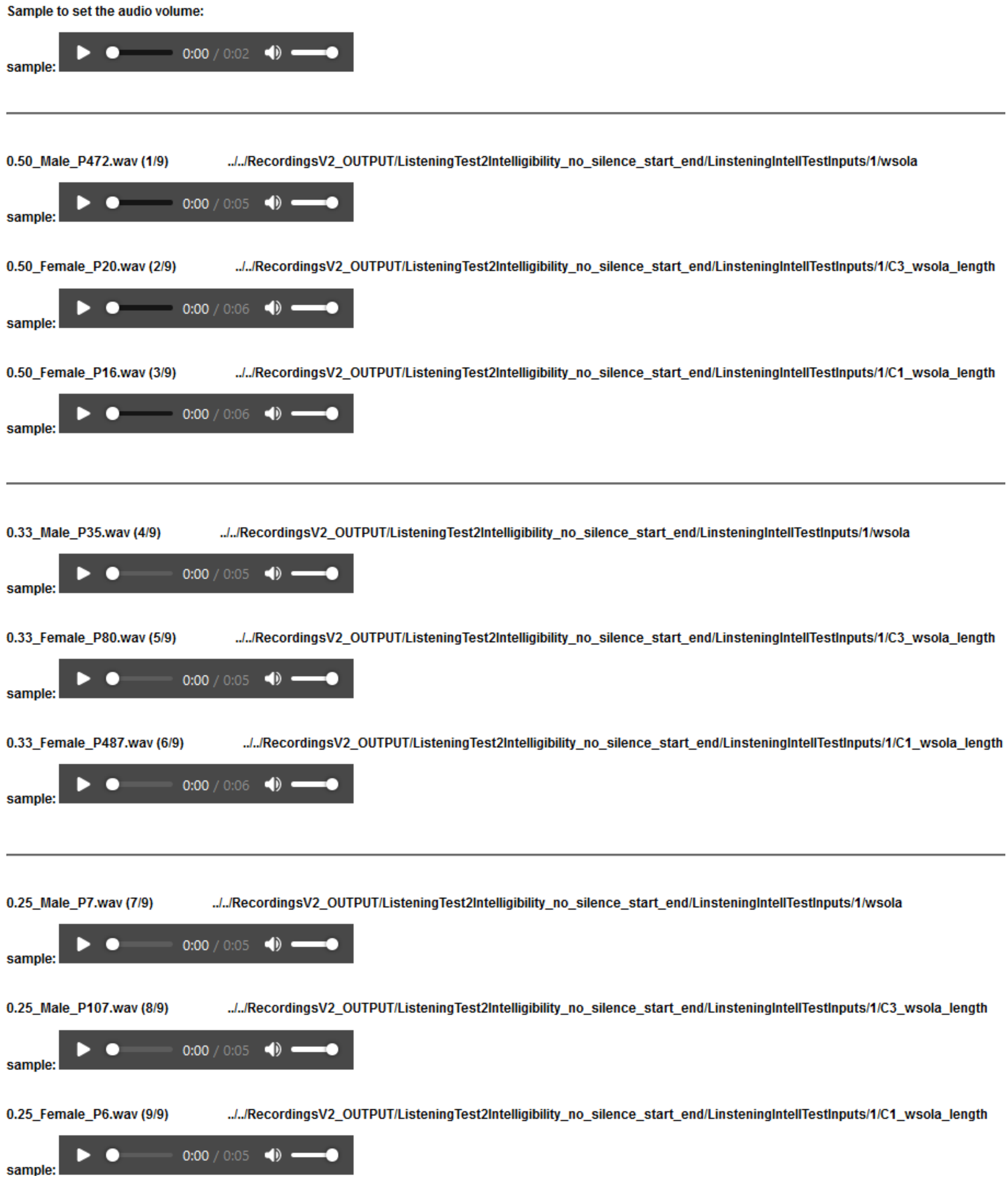
0.15_Female_P691.wav 8/9

sample1: ▶ 0:00 / 0:00 ○ sample2: ▶ 0:00 / 0:00 ○ sample3: ▶ 0:00 / 0:00 ○ Neither ○

0.10_Female_P488.wav 9/9

sample1: ▶ 0:00 / 0:00 ○ sample2: ▶ 0:00 / 0:00 ○ sample3: ▶ 0:00 / 0:00 ○ Neither ○

Figure 7.1: *Intelligibility & Preference test for group 1 of listening tests.*



Sample to set the audio volume:

sample:

0.50_Male_P472.wav (1/9) ../../RecordingsV2_OUTPUT/ListeningTest2Intelligibility_no_silence_start_end/LinsteningIntellTestInputs/1/wsola

sample:

0.50_Female_P20.wav (2/9) ../../RecordingsV2_OUTPUT/ListeningTest2Intelligibility_no_silence_start_end/LinsteningIntellTestInputs/1/C3_wsola_length

sample:

0.50_Female_P16.wav (3/9) ../../RecordingsV2_OUTPUT/ListeningTest2Intelligibility_no_silence_start_end/LinsteningIntellTestInputs/1/C1_wsola_length

sample:

0.33_Male_P35.wav (4/9) ../../RecordingsV2_OUTPUT/ListeningTest2Intelligibility_no_silence_start_end/LinsteningIntellTestInputs/1/wsola

sample:

0.33_Female_P80.wav (5/9) ../../RecordingsV2_OUTPUT/ListeningTest2Intelligibility_no_silence_start_end/LinsteningIntellTestInputs/1/C3_wsola_length

sample:

0.33_Female_P487.wav (6/9) ../../RecordingsV2_OUTPUT/ListeningTest2Intelligibility_no_silence_start_end/LinsteningIntellTestInputs/1/C1_wsola_length

sample:

0.25_Male_P7.wav (7/9) ../../RecordingsV2_OUTPUT/ListeningTest2Intelligibility_no_silence_start_end/LinsteningIntellTestInputs/1/wsola

sample:

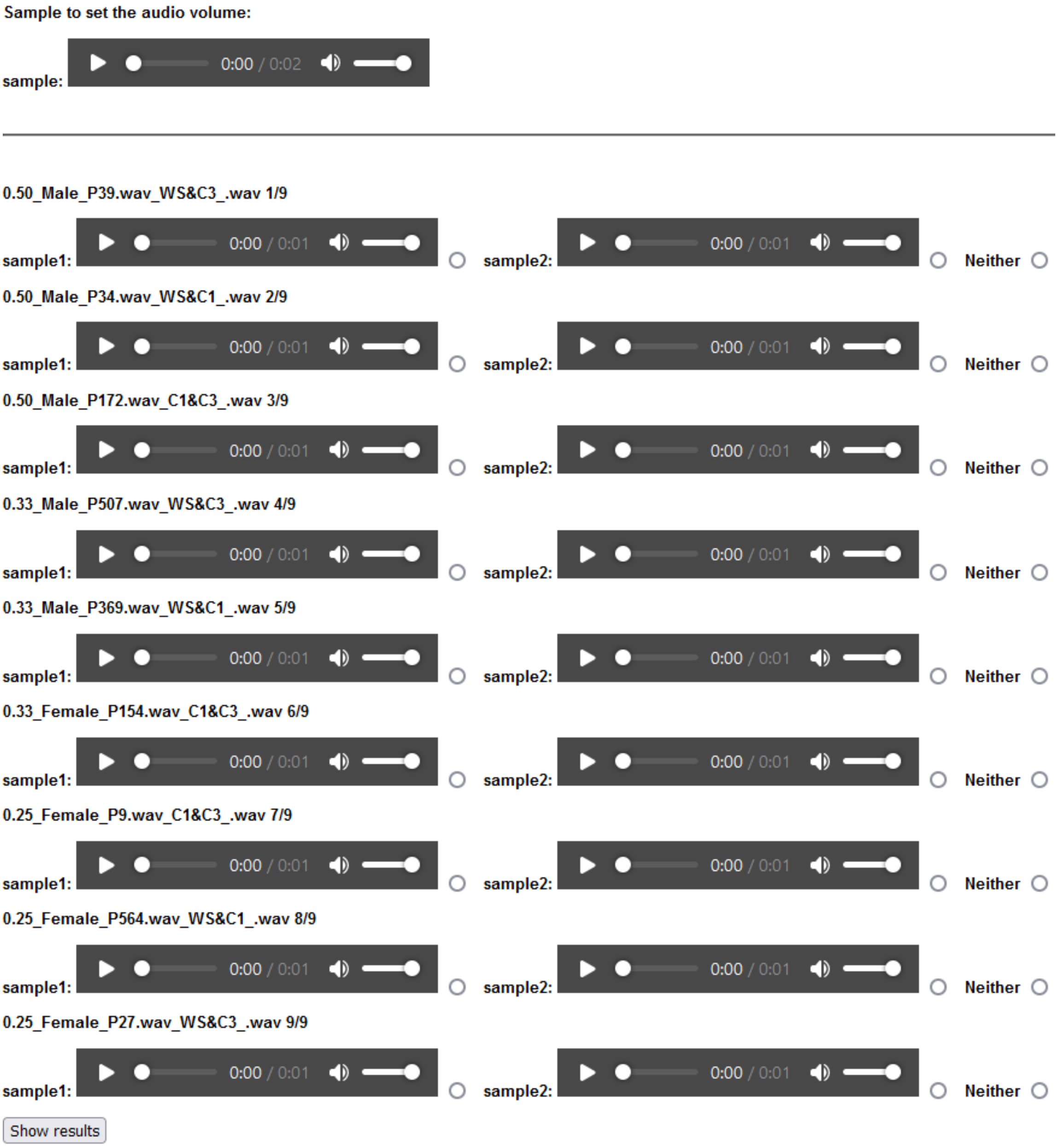
0.25_Male_P107.wav (8/9) ../../RecordingsV2_OUTPUT/ListeningTest2Intelligibility_no_silence_start_end/LinsteningIntellTestInputs/1/C3_wsola_length

sample:

0.25_Female_P6.wav (9/9) ../../RecordingsV2_OUTPUT/ListeningTest2Intelligibility_no_silence_start_end/LinsteningIntellTestInputs/1/C1_wsola_length

sample:

Figure 7.2: *Intelligibility test for group 2 of listening tests.*



Sample to set the audio volume:

sample:

0.50_Male_P39.wav_WS&C3_.wav 1/9

sample1: ○ **sample2:** ○ **Neither** ○

0.50_Male_P34.wav_WS&C1_.wav 2/9

sample1: ○ **sample2:** ○ **Neither** ○

0.50_Male_P172.wav_C1&C3_.wav 3/9

sample1: ○ **sample2:** ○ **Neither** ○

0.33_Male_P507.wav_WS&C3_.wav 4/9

sample1: ○ **sample2:** ○ **Neither** ○

0.33_Male_P369.wav_WS&C1_.wav 5/9

sample1: ○ **sample2:** ○ **Neither** ○

0.33_Female_P154.wav_C1&C3_.wav 6/9

sample1: ○ **sample2:** ○ **Neither** ○

0.25_Female_P9.wav_C1&C3_.wav 7/9

sample1: ○ **sample2:** ○ **Neither** ○

0.25_Female_P564.wav_WS&C1_.wav 8/9

sample1: ○ **sample2:** ○ **Neither** ○

0.25_Female_P27.wav_WS&C3_.wav 9/9

sample1: ○ **sample2:** ○ **Neither** ○

Show results

Figure 7.3: *Preference test for group 2 of listening tests.*

Bibliography

[1] A. Sfakianaki, G. Kafentzis, and Y. Stylianou, “The GrHarvard Corpus: A Greek sentence corpus for speech technology research and applications,” in *14th International Conference On Greek Linguistics*, 07 2021.

[2] L. R. Rabiner and R. W. Schafer, “Introduction to digital speech processing,” *Foundations and Trends in Signal Processing*, vol. 1, jan 2007.

[3] J. Driedger and M. Müller, “A review of time-scale modification of music signals,” *Applied Sciences*, vol. 6, no. 2, 2016.

[4] MATLAB, “Hanning Window function.” `https://www.mathworks.com/help/signal/ref/hann.html`. [Online].

[5] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication.* Urbana: University of Illinois Press, 1949.

[6] S. P. Whiteside, “The speech chain: The physics and biology of spoken language,” *Journal of the International Phonetic Association*, vol. 23, no. 2, p. 98–101, 1993.

[7] R. L. Diehl, A. J. Lotto, and L. L. Holt, “Speech perception,” *Annu. Rev. Psychol.*, vol. 55, pp. 149–179, 2004.

[8] J. B. Allen, “Nonlinear cochlear signal processing and masking in speech perception,” *Springer handbook of speech processing*, pp. 27–60, 2008.

[9] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound.* MIT press, 1994.

[10] A. Bronkhorst, “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,” *Acta Acustica united with Acustica*, vol. 86, pp. 117–128, 01 2000.

[11] A. Amano-Kusumoto and J.-P. Hosom, “A review of research on speech intelligibility and correlations with acoustic features,” *Center for Spoken Language Understanding, Oregon Health and Science University (Technical Report CSLU-011-001)*, 2011.

[12] N. Miller, “Measuring up to speech intelligibility,” *International Journal of Language & Communication Disorders*, vol. 48, no. 6, pp. 601–612, 2013.

[13] C. V. Pavlovic, “Review of speech intelligibility measures,” *The Journal of the Acoustical Society of America*, vol. 88, pp. S31–S31, 08 2005.

[14] C. E. Williams, M. Hecker, K. N. Stevens, and B. Woods, “Intelligibility test methods and procedures for evaluation of speech communication systems,” *National Technical Information Service AD 646e781*, 1966.

[15] J. M. Alexander, “Hearing aid technology to improve speech intelligibility in noise,” *Seminars in hearing*, vol. 42, p. 175—185, August 2021.

[16] B. Edwards, “Beyond amplification: Signal processing techniques for improving speech intelligibility in noise with hearing aids,” *Seminars in Hearing*, vol. Volume 21, pp. 0137–0156, 01 2000.

[17] A. K. Namasivayam, M. Pukonen, D. Goshulak, V. Y. Yu, D. S. Kadis, R. Kroll, E. W. Pang, and L. F. De Nil, “Relationship between speech motor control and speech intelligibility in children with speech sound disorders,” *Journal of Communication Disorders*, vol. 46, no. 3, pp. 264–280, 2013.

[18] M. Karbasi and D. Kolossa, “Asr-based speech intelligibility prediction: A review,” *Hearing Research*, vol. 426, p. 108606, 2022.

[19] J. S. Bradley, R. D. Reich, and S. G. Norcross, “On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility,” *The Journal of the Acoustical Society of America*, vol. 106, pp. 1820–1828, 10 1999.

[20] G. Kim and P. C. Loizou, “Improving speech intelligibility in noise using environment-optimized algorithms,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2080–2090, 2010.

[21] S. D. Soli and L. L. Wong, “Assessment of speech intelligibility in noise with the hearing in noise test,” *International Journal of Audiology*, vol. 47, no. 6, pp. 356–361, 2008.

[22] J. Driedger and M. Müller, “A review of time-scale modification of music signals,” *Applied Sciences*, vol. 6, no. 2, 2016.

[23] B. Ninness and S. J. Henriksen, “Time-scale modification of speech signals,” *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1479–1488, 2008.

[24] A. Amir, D. Ponceleon, B. Blanchard, D. Petkovic, S. Srinivasan, and G. Cohen, “Using audio time scale modification for video browsing,” in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, pp. 9–pp, IEEE, 2000.

[25] O. Donnellan, E. Jung, and E. Coyle, “Speech-adaptive time-scale modification for computer assisted language-learning,” in *Proceedings 3rd IEEE International Conference on Advanced Technologies*, pp. 165–169, IEEE, 2003.

[26] M. H. Stollman and T. S. Kapteyn, “Effect of time scale modification of speech on the speech recognition threshold in noise for elderly listeners,” *Audiology*, vol. 33, no. 5, pp. 280–290, 1994.

[27] M. Wlodarczyk and P. Sekalski, “Evaluation of time-scale modification methods for audio signals on mobile devices with android os,” in *2014 Proceedings of the 21st International Conference Mixed Design of Integrated Circuits and Systems (MIXDES)*, pp. 451–454, IEEE, 2014.

[28] M. Slaney, “Normalizing non-linear speech speed for maintaining listener comprehension at increased playback speeds,” *Technical Disclosure Commons Technical Disclosure Commons*, 2021.

[29] D. Choi, D. Kwak, M. Cho, and S. Lee, “” nobody speaks that fast!” an empirical study of speech rate in conversational agents for people with vision impairments,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020.

[30] W. Verhelst and M. Roelands, “An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 554–557 vol.2, 1993.

[31] J. Driedger, M. Müller, and S. Ewert, “Improving time-scale modification of music signals using harmonic-percussive separation,” *Signal Processing Letters, IEEE*, vol. 21, pp. 105–109, 01 2014.

[32] S. Roucos and A. Wilgus, “High quality time-scale modification for speech,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 10, pp. 493–496, 1985.

[33] E. Hardam, “High quality time scale modification of speech signals using fast synchronized-overlap-add algorithms,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 409–412 vol.1, 1990.

[34] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[35] J. Suzuki, “Speech processing by splicing of autocorrelation function,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 713–716, IEEE, 1976.

[36] S. Grofit and Y. Lavner, “Time-scale modification of audio signals using enhanced wsola with management of transients,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 106–115, 2008.

[37] R. J. McAulay and T. F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, pp. 744–754, 1986.

[38] G. Degottex and Y. Stylianou, “Analysis and synthesis of speech using an adaptive full-band harmonic model,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 2085–2095, 2013.

[39] J. Laroche, Y. Stylianou, and E. Moulines, “Hnm: a simple, efficient harmonic+noise model for speech,” *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 169–172, 1993.

[40] G. P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou, “An extension of the adaptive quasi-harmonic model,” *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4605–4608, 2012.

[41] R. Badeau, R. Boyer, and B. David, “EDS Parametric Modeling and Tracking of Audio Signals,” in *DAFx*, 2002.

[42] J. Jensen, R. Heusdens, and S. Jensen, “A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids,” *Speech and Audio Processing, IEEE Transactions on*, pp. 121 – 132, 04 2004.

[43] T. K. Nguyen, S. M. Vlasov, A. A. Pyrkin, A. S. Kirsanova, and M. Korotina, "Estimation of the parameters of exponentially damped sinusoidal signals," *2022 European Control Conference (ECC)*, pp. 309–314, 2022.

[44] M. R. Portnoff, "Time-scale modification of speech based on short-time fourier analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, pp. 374–390, 1981.

[45] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. Signal Process.*, vol. 40, pp. 497–510, 1992.

[46] J. Laroche, Y. Stylianou, and É. Moulines, "Hns: Speech modification based on a harmonic+noise model," *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 550–553 vol.2, 1993.

[47] G. P. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou, "Time-scale modifications based on a full-band adaptive harmonic model," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8193–8197, 2013.

[48] N. K. Sharma, S. Potadar, S. R. Chetupalli, and T. V. Sreenivas, "Mel-scale subband modelling for perceptually improved time-scale modification of speech and audio signals," *2017 Twenty-third National Conference on Communications (NCC)*, pp. 1–5, 2017.

[49] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio processing*, vol. 7, no. 3, pp. 323–332, 1999.

[50] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966.

[51] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3933–3936, 2008.

[52] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2001.

[53] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[54] A. Kyparissiadis, W. J. B. van Heuven, N. J. Pitchford, and T. Ledgeway, "Greeklex 2: A comprehensive lexical database with part-of-speech, syllabic, phonological, and stress information," *PLoS ONE*, vol. 12, no. 2, 2017.

[55] A. Protopapas, M. Tzakosta, A. Chalamandaris, and P. Tsiakoulis, "Iplr: An online resource for greek word-level and sublexical information," *Language Resources and Evaluation*, vol. 46, p. 449–459, 09 2012.

[56] V. Aubanel, M. Lecumberri, and M. Cooke, "The sharvard corpus: a phonemically-balanced spanish sentence resource for audiology.," *Epub 2014 May 26. PMID: 24863133*, vol. 46, 09 2014.

[57] N. Trimmis, E. Papadeas, T. Papadas, S. Naxakis, P. Papathanasopoulos, and P. Goumas, “Speech audiometry: The development of modern greek word lists for suprathreshold word recognition testing.,” *The Mediterranean Journal of Otology*, vol. 3, p. 117–126, 2006.

[58] D. Kapilow, Y. Stylianou, and J. Schroeter, “Detection of non-stationarity in speech signals and its application to time-scaling,” in *EUROSpeech*, 09 1999.

[59] StackExchange, “What is the RMS value of a signal.” `https://dsp.stackexchange.com/questions/19337/what-is-the-rms-value-of-a-signal-and-how-to-use-it`. [Online].

[60] “Sigmoid function Wiki.” `https://en.wikipedia.org/wiki/Sigmoid_function`. [Online].

[61] “Pre-emphasis signal calculation MATLAB.” `https://www.mathworks.com/matlabcentral/answers/414488-pre-emphasis-signal-processing`. [Online].

[62] MATLAB, “Anaysis of Variance function.” `https://www.mathworks.com/help/stats/anova1.html`. [Online].

[63] MATLAB, “Post-hoc multi-compare test.” `https://www.mathworks.com/help/stats/multcompare.html`. [Online].

[64] “Statistical test for preference data.” `https://measuringu.com/preference-data/`. [Online].

[65] “Anova-and-chi-square tests on medium.” `https://medium.com/@chandradip93/anova-and-chi-square-aea693c4eb96`. [Online].

[66] MATLAB, “Chi-square goodness-of-fit test.” `https://www.mathworks.com/help/stats/chi2gof.html`. [Online].

[67] “Standardized residuals on statology.” `https://www.statology.org/standardized-residuals/`. [Online].

[68] “Learn what are residuals on displayr.” `https://www.displayr.com/learn-what-are-residuals/`. [Online].