University of Crete
Department of Computer Science

Intelligibility Enhancement of Casual Speech based on Clear Speech Properties

Ph.D. Thesis

Maria C. Koutsogiannaki

Heraklion

March 2016

UNIVERSITY OF CRETE
SCHOOL OF SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE

Intelligibility enhancement of Casual Speech based on Clear Speech Properties

Submitted by

Maria C. Koutsogiannaki

in partial fulfillment of the requirements for the
Doctor of Philosophy degree in Computer Science

Author:

Maria C. Koutsogiannaki
Department of Computer Science

Examination Committee:

Supervisor

Yannis Stylianou, Professor, University of Crete

Member

Athanasios Mouchtaris, Associate Professor, University of Crete

Member

Valerie Hazan, Professor, University College of London

Member

Inma Hernaez, Professor, University of Basque Country

Member

Katerina Nikolaidis, Associate Professor, Aristotle University of Thessaloniki

Member

Martin Cooke, Professor Researcher, Ikerbasque

Member

Daniel Erro, Researcher, Ikerbasque

Departmental Approval:

Chairman
of the Department

Panagiotis Tsakalidis, Professor, University of Crete

Heraklion, March 2016

Αφηερομένο στους γονείς μου
που μου δήδαξαν ότη η αξήα του ανθρώπου
δε μαιτράτε με πτιχεία

Τον γιο του Ευφορίωνα τον Αθηναίο Αισχύλο
κρύβει νεκρόν το μνήμα αυτό της Γέλας με τα στάρια
την άξια νιότη του θα ειπεί του Μαραθώνα το άλσος
κι ο Μήδος ο ακούρευτος οπού καλά την ξέρει.

Acknowledgements

I would like to thank my supervisor Prof. Yannis Stylianou for the opportunity that he gave me to evolve my work and my personality and for his constant motivations to become better. I would also like to thank him for his trust and patience during the years of our collaboration. I would also like to thank Prof. Athanasios Mouchtaris for his support and advice as a member of my committee and Prof. Valerie Hazan of the University College of London for her valuable guidance during the completion of this thesis. At this point, I would like to thank all the members of my committee, Prof. Inma Hernaez (University of Basque Country), Prof. Katerina Nikolaidis (Aristotle University of Thessaloniki), Prof. Martin Cooke (Ikerbasque) and the researcher Dr. Daniel Erro (Ikerbasque).

Many thanks to my colleagues, Elizabeth Godoy, George Kafentzis and Gilles Degottex who advised me and guided me during my PhD thesis. Also, I would like to thank Dr. Anna Sfakiannaki and Prof. Katerina Nikolaidis for their contribution.

This work could not have been accomplished without the assistance of my friends, the people of our laboratory in the University and the people at FORTH who tolerated the numerous intelligibility tests with a smile. I would also like to thank Dr. Yannis Pantazis who inspired me to start my research career.

Finally, I would like to thank Igor Martinez Uriagereka for tolerating me. Last but not least, the greatest "thank you" to my family, my parents and my sister Antigoni for supporting me all these years. Thank you very much.

Ευχαριστίες

'Το διδακτορικό δεν είναι ένα μέρος της ζωής σου αλλά η ζωή σου'. Αυτά ήταν τα λόγια του επιβλέποντα της εργασίας μου καθηγητή Γιάννη Στυλιανού πριν από 4 χρόνια, πριν το ξεκίνημα της διδακτορικής μου διατριβής. Σήμερα, τέσσερα χρόνια μετά αντιλαμβάνομαι την σημασία της φράσης αυτής. Το διδακτορικό δεν είναι ένα ταξίδι που φτάνει στο τέλος του αλλά είναι η αρχή που άλλαξε την πορεία μιας ζωής. Θα ήθελα να ευχαριστήσω τον καθηγητή Γιάννη Στυλιανού για στήριξη που μου παρείχε κατά την διάρκεια του προγράμματος σπουδών μου, για τα κίνητρα και τις ευκαιρίες που μου έδωσε ώστε να γίνω καλύτερη. Θα ήθελα επίσης να τον ευχαριστήσω για το ενδιαφέρον που έδειξε σε προσωπικό επίπεδο αλλά και για την εμπιστοσύνη και την υπομονή του όλα αυτά τα χρόνια συνεργασίας μας.

Επιπροσθέτως, θα ήθελα να ευχαριστήσω θερμά τον καθηγητή Αθανάσιο Μουχτάρη για τις συμβουλές και την συμπαράστασή του ως μέλος της επιτροπής μου και την καθηγήτρια Valerie Hazan του University College of London για την πολύτιμη καθοδήγησή της κατά την διάρκεια εκπόνησης της εργασίας μου. Στο σημείο αυτό θα ήθελα να ευχαριστήσω και όλα τα υπόλοιπα μέλη της εξεταστικής μου επιτροπής, την καθηγήτρια Inma Hernaez (University of Basque Country), την καθηγήτρια Κατερίνα Νικολαΐδη (Αριστοτέλειο Πανεπιστημίο Θεσσαλονίκης), τον καθηγητή Martin Cooke (Ikerbasque) και τον ερευνητή Daniel Erro (Ikerbasque).

Ένα μεγάλο ευχαριστώ στους συνεργάτες μου, Elizabeth Godoy, Γιώργο Καφεντζή και Gilles Degottex που συνεισέφεραν με τις συμβουλές τους και την καθοδήγησή τους στην ολοκλήρωση της εργασίας αυτής. Στο σημείο αυτό θα ήθελα να ευχαριστήσω την διδάκτορα Άννα Σφακιαννάκη και την καθηγήτρια Κατερίνα Νικολαΐδη του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης για την συνεισφορά τους.

Επίσης, ευχαριστώ τους φίλους μου, τα παιδιά του εργαστηρίου μας στο Πανεπιστήμιο και στο ΙΤΕ γιατί άντεξαν τα πολυάριθμα τεστ καταληπτότητας και κυρίως γιατί εξακολουθούσαν να μου χαμογελούν μετά από αυτά. Επίσης, θα ήθελα να ευχαριστήσω τον Δρ. Γιάννη Πανταζή που με ενέπνευσε στο να ξεκινήσω την ερευνητική μου πορεία.

Τέλος, θα ήθελα να ευχαριστήσω τον Ιγόρ Μαρτίνεθ Ουριαγερέκα για την υποστήριξη και την υπομονή του και κυρίως ένα μεγάλο ευχαριστώ στην οικογένειά μου, στους γονείς μου Χαράλαμπο και Ειρήνη και στην αδερφή μου Αντιγόνη για την αμέριστη συμπαράστασή τους όλα αυτά τα χρόνια. Σας ευχαριστώ πολύ.

Abstract

In adverse listening conditions (e.g. presence of noise, hearing-impaired listener etc.) people adjust their speech in order to overcome the communication difficulty and successfully deliver their message. This remarkable adjustment produces different speaking styles compared to unobstructed speech (casual speech) that vary among speakers and conditions, but share a common characteristic; high intelligibility. Developing algorithms that exploit acoustic features of intelligible human speech could be beneficial for speech technology applications that seek methods to enhance the intelligibility of "speaking-devices". Besides the commercial orientation (e.g., mobile telephone, GPS, customer service systems) of these applications, most important is their medical context, providing assistive communication to people with speech or hearing deficits. However, current speech technology is deaf, meaning that it cannot adjust, like humans do, to the dynamically changing real environments or to the listener's specificity.

This work proposes signal modifications based on the acoustic properties of a high intelligible human speaking style, the clear speech, assisting in the development of smart speech technology systems that "mimic" the way people produce intelligible speech. Unlike other speaking styles, clear speech has a high intelligibility impact on various listening populations (native and non-native listeners, hearing impaired, cochlear implant users, elderly people, people with learning disabilities etc.) in many listening conditions (quiet, noise, reverberation).

A significant part of this work is devoted to the comparative analysis between casual and clear speech, which reveals differences on prosody, vowel spaces, spectral energy and modulation depth of the temporal envelopes. Based on these observed and measured differences between the two speaking styles, we propose modifications for enhancing the intelligibility of casual speech. Compared to other state-of-the-art modification systems, our modification techniques (1) do not require excessive computation (2) are speaker and speech independent (3) maintain speech quality (4) are explicit, since they do not require statistical training and the preexistence of clear speech recordings.

Evaluations on intelligibility and quality are performed objectively using recently proposed objective intelligibility scores and subjectively with listening tests conducted by native and non native listeners in noisy environments (speech shaped noise, SSN), reverberation and in quiet. Results show that our modifications enhance speech intelligibility in SSN and reverberation for native and non-native listeners. Specifically, the proposed spectral modification technique, namely Mix-filtering, increases the intelligibility of speech in noise and reverberation while maintains the quality of the original signal, unlike other intelligibility boosters. Moreover, a modulation depth enhancement technique called DMod, increases speech intelligibility more than 30% in SSN. DMod algorithm is inspired by both clear speech properties and by the non-linear phenomena that take place in the basilar membrane. DMod not only achieves to enhance speech intelligibility, but it introduces a novel method for manipulating the modulation spectrum of the signal. Results of this study indicate a connection of the modulations of the temporal envelopes with speech perception and specifically with processes that take place on the basilar membrane of human ear and pave the way for analyzing and comprehending speech

in terms of modulations.

Περίληψη

Όταν ένας άνθρωπος επικοινωνεί με έναν συνάνθρωπό του, προσαρμόζει ανταναικλαστικά την ομιλία του ανάλογα με το περιβάλλον στο οποίο βρίσκεται αυτός (π.χ. παρουσία θορύβου) ή ο συνομιλητής του (π.χ. βαρήκοος), παράγοντας διαφορετικά στυλ ομιλίας (Καθαρή ομιλία, ομιλία Λομβαρδ) σε σχέση με το αν η επικοινωνία του ήταν ανεμπόδιστη (Πρόχειρη ομιλία). Τα στυλ αυτά ομιλίας διαφέρουν ανάλογα με το είδος του εμποδίου στο επικοινωνιακό κανάλι ή/και ανάλογα με τον ομιλητή. Παρουσιάζουν όμως ένα κοινό χαρακτηριστικό: την αυξημένη καταληπτότητα. Η ανάπτυξη αλγορίθμων που εκμεταλλεύονται τα ακουστικά χαρακτηριστικά τέτοιων στυλ ομιλίας θα μπορούσε να είναι επωφελής στην Τεχνολογία Φωνής. Πολλές τεχνολογικές εφαρμογές αναζητούν μεθόδους βελτιστοποίησης της καταληπτότητας των συσκευών που παράγουν συνθετική φωνή. Πέρα από την εμπορική εκμετάλλευση των εφαρμογών αυτών (κινητά τηλέφωνα, συστήματα πλοήγησης, συστήματα τηλεφωνικής υποστήριξης πελατών), πολύ σημαντική είναι η συνεισφορά τους στον ιατρικό τομέα ως βοηθητικά μέσα επικοινωνίας ατόμων με προβλήματα ομιλίας και ακοής. Ωστόσο, η τρέχουσα τεχνολογία φωνής είναι «κωφή», δεν μπορεί δηλαδή να προσαρμοστεί στα δυναμικώς μεταβαλλόμενα περιβάλλοντα ούτε στην ιδιαιτερότητα του ακροατή, όπως ο άνθρωπος.

Η εργασία αυτή προτείνει την ανάπτυξη αλγορίθμων που «μιμούνται» τον τρόπο ανθρώπινης ομιλίας σε δύσκολες συνθήκες επικοινωνίας, συνεισφέροντας στην ανάπτυξη έξυπνων τεχνολογικών συστημάτων φωνής. Συγκεκριμένα, το στυλ ομιλίας του οποίου τα χαρακτηριστικά αναλύονται και χρησιμοποιούνται για την αύξηση της καταληπτότητας της Πρόχειρης ομιλίας είναι η Καθαρή ομιλία. Σε αντίθεση με άλλα στυλ ομιλίας, η Καθαρή ομιλία είναι καταληπτή από διάφορους ακροατές (ομόγλωσσους και μη, με προβλήματα ακοής, με κοχλιακά εμφυτεύματα, ηλικιωμένους, με μαθησιακές δυσκολίες κλπ) σε διάφορες συνθήκες περιβάλλοντος (με ή χωρίς θόρυβο, σε περιβάλλοντα αντήχησης).

Ένα σημαντικό μέρος της εργασίας αυτής αναλύει και συγκρίνει τα χαρακτηριστικά της Πρόχειρης και Καθαρής ομιλίας. Από την σύγκριση αυτή, αναδεικνύονται διαφορές κυρίως στην προσωδία, στον φωνηεντικό χώρο, στην φασματική ενέργεια και στο πλάτος διαμόρφωσης της χρονικής περιβάλλουσας του σήματος φωνής. Βασιζόμενοι στις μετρίσιμες αυτές διαφορές, προτείνουμε μετασχηματισμούς που βελτιώνουν την καταληπτότητα του σήματος Πρόχειρης ομιλίας. Σε σύγκριση με state-of-the-art συστήματα μετασχηματισμού, οι δικές μας τεχνικές (1) είναι χαμηλής υπολογιστικής πολυπλοκότητας (2) μπορούν να εφαρμοστούν ανεξαρτήτως ομιλητή ή σήματος (3) διατηρούν την ποιότητα του αρχικού σήματος (4) εφαρμόζονται άμεσα χωρίς την ανάγκη εκπαίδευσης των δεδομένων και προϋπαρξης του σήματος Καθαρής φωνής. Οι προτεινόμενοι αλγόριθμοι αξιολογηθήκαν ως προς την καταληπτότητα και την ποιότητα με αντικειμενικές μετρικές καταληπτότητας και με υποκειμενικά ακουστικά τέστ από ομόγλωσσους και αλλόγλωσσους ακροατές χωρίς την ύπαρξη θορύβου, μέσα σε θορυβώδη περιβάλλοντα και σε περιβάλλοντα αντήχησης. Η αξιολόγηση δείχνει ότι οι μετασχηματισμοί που προτείνουμε αυξάνουν την καταληπτότητα της πρόχειρης ομιλίας τόσο σε περιβάλλοντα θορύβου όσο και σε περιβάλλοντα αντήχησης για ομόγλωσσους και αλλόγλωσσους ακροατές. Συγκεκριμένα, η τεχνική φασματικού μετασχηματισμού, επονομαζόμενη ως

Mix-filtering, αυξάνει την καταληπτότητα του σήματος ομιλίας σε περιβάλλοντα θορύβου και αντήχησης ενώ διατηρεί την ποιότητα του σήματος, εν αντιθέσει με άλλους αλγορίθμους. Επιπλέον, η προτεινόμενη τεχνική αύξησης του πλάτους των διαμορφώσεων της χρονικής περιβάλλουσας, αναφερθείσα ως DMod, αυξάνει την καταληπτότητα της Πρόχειρης ομιλίας κατά 30% σε περιβάλλοντα θορύβου. Ο αλγόριθμος DMod, εμπνέεται όχι μόνο από χαρακτηριστικά της Καθαρής ομιλίας αλλά και από μη γραμμικές λειτουργίες που λαμβάνουν χώρα στην βασική μεμβράνη του ανθρώπινου κοχλία. Επιτυγχάνει δε, πέρα από την αύξηση της καταληπτότητας, την εισαγωγή μιας νέας μεθόδου χειρισμού των διαμορφώσεων της περιβάλλουσας του σήματος. Τα αποτελέσματα της μελέτης αυτής δείχνουν την ύπαρξη μιας σύνδεσης ανάμεσα στις διαμορφώσεις της χρονικής περιβάλλουσας και στον τρόπο αντίληψης και επεξεργασίας του ήχου από την βασική μεμβράνη του ανθρώπινου κοχλία, ανοίγοντας τον δρόμο για την ανάλυση και κατανόηση της ομιλίας ως κύματα διαμορφώσεων.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Humans use speech to communicate, to express their needs and their emotions, to share their knowledge and their beliefs. Regardless of the communication environment (presence of noise, hearing-impaired listener etc.), people understand how to adjust their speech in order to deliver their message successfully. This is performed usually reflexively, using auditory (people can listen of what they say) and visual feedback (can see the listener's expressions) to estimate whether the listener has understood or not. Therefore, when people communicate in real world environments, they have feedback and this feedback helps them to produce intelligible speech with the least possible effort.

However, there are some cases when human speech may not be intelligible. This is due to the fact that the speaker has no feedback from the listener's side (i.e. airport announcements). In such one-way communication channels where human interaction is absent, recorded speech or synthetic speech by automated systems is often used. The absence of feedback in such applications does not ensure that the message will be understood by the listener. Current speech technology is deaf, meaning that it cannot adjust, like humans do, to the dynamically changing real environments or to the listener's specificity.

With growing numbers of applications using speech technologies in commercial (e.g., mobile telephone, GPS, customer service systems), military (e.g., Air Force,Ground troop relays) and medical (e.g., assisted-speech) contexts, methods that enhance the intelligibility of "speaking-devices" are currently in high demand. Considering the intelligibility gains of the human speaking styles, acoustic features of such styles can be used to inspire speech signal modifications for intelligibility enhancement. This work proposes signal modifications based on the acoustic properties of a highly intelligible human speaking style, the clear speech, assisting in the development of smart speech technology systems that "mimic" the way people produce intelligible speech.

1.1 Recent advances on speech intelligibility enhancement

Generally, speech modifications for intelligibility enhancement can be classified into several groups. First, much work has been done on enhancing the intelligibility of degraded speech, such as noise suppression techniques (Kates, 1994) or methods for dereverberation (Liu et al., 1996). Other techniques focus on prepro-

cessing speech in order to enhance its intelligibility, before transmitted in the communication channel. First, there are techniques to enhancing intelligibility that exploit audio and signal properties, such as the amplitude compression scheme in Niederjohn and Grotelueschen (1976), dynamic range compression in Blesser (1969) and a method for peak-to-rms reduction in Quatieri and McAulay (1991). Second, certain speech intelligibility enhancement methods focus on speech-in-noise and exploit knowledge of the noise masker, such as the optimizations based on a speech intelligibility index in Sauert and Vary (2006) and the glimpse proportion maximization in Tang and Cooke (2011a). Third, in the context of text-to-speech synthesis, adaptation approaches have been explored to increase intelligibility, as in Langner and Black (2005) and Raitio et al. (2011).

Recently, a fourth category of intelligibility enhancement techniques has emerged, aiming on studying the impact on intelligibility of particular acoustic features of naturally produced intelligible speech (Lu and Cooke, 2009; Krause and Braida, 2004a,b). Moreover, several speech intelligibility enhancement approaches exploited these acoustic features, such as the glimpse proportion maximization for HMM-based text-to-speech synthesis in Valentini-Botinhao et al. (2011). Finally, an extensive evaluation of the intelligibility of a variety of methods was recently carried out and described in Cooke et al. (2013). Emerging as the most successful modification from this challenge was the combination of Lombard-like Spectral Shaping (SS) and audio enhancement with dynamic range compression (DRC) proposed in Zorila et al. (2012). This work revealed the advantaged of exploiting human speech characteristics on intelligibility enhancement techniques. At the same time, the techniques developed are speaker- and style-independent in that they can be applied generally to any speech signal (including text-to-speech Erro et al. (2014)), without requiring statistical learning or classification. In this respect, our motivation is to examine modifications based on a specific style of naturally produced speech that has been found to be intelligible in many adverse listening environments and for various listening populations, namely the clear speech.

1.2 Clear and Casual speech

In the absence of communication barriers people produce effortless and unobstructed speech in order to communicate. This type of speech is referred as casual or plain speech to characterize speech produced under casual or typical circumstances when no special speaking effort is made. However, in the presence of a communication difficulty, humans adopt different speaking styles in order to successfully deliver their message. The speaking style they adopt depends mostly on the communication barrier they want to overcome in order to communicate. If the communication barrier that the speaker faces is for example a noisy environment, the speaker will increase the loudness of his/her speech, producing Lombard speech (Bond et al., 1989). Two main interpretations of the Lombard effect have been proposed. The first describes Lombard speech as a physiological audio-phonatory reflex of the person speaking in noise where the speaker adjusts his/her vocal intensity in a way that he/she can hear himself/herself (Lombard, 1911). The second considers Lombard speech as a cognitive process where the speaker is aware of the intelligibility decrease and performs modifications towards

maintaining or improving the intelligibility of speech in noise (Harlan and Tranel, 1971). Other studies support that the Lombard effect phenomenon is a combination of both mechanisms that contribute to the changes the speakers makes in noisy environments (Junqua, 1993; Garnier et al., 2006, 2010).

Alternatively to the Lombard effect, "clear" speech strategies are also adopted by speakers in order to increase the intelligibility of the elicited speech. However, clear speech modifications are mainly cognitive since the speaker does not face any communication barrier. In this case, the speaker is in a quiet environment and deliberately changes his/her speaking style in order to communicate with the listener that faces a communication barrier. For example, the target listener could be either hearing-impaired or a non-native listener (L2). Changes in the articulatory gestures are found both in clear and in Lombard speech, with Lombard speech accompanied more often by increased vocal effort. In both speaking styles, clear and Lombard, the aim of the speaker is to communicate successfully with the least possible speaking effort, described by the Hyper-Hypo-speech model (Lindblom, 1990). In this model, speech modifications and adjustments follow two main principles 1) the communicative efficiency related to speech perception 2) the communicative economy related to speech production. Speakers adjust their speech to achieve the maximum clarity with the minimum speech effort. This explains the high variability of produced speech depending on the speaker, on the communication obstacle etc. Despite many differences in speaker strategies, the most common characteristics of clear speech is hyper-articulation, with increased effort on the part of the speaker to slow down and enunciate. This speaking adjustment is proven to increase the intelligibility of casual speech.

Lombard and clear speech have individually been well-studied in the literature, mainly due to the intelligibility gains of these human speaking styles. However, the intelligibility advantage of clear speech is extended to a wider range of applications than Lombard speech. Clear speech has been found to be highly intelligible on various adverse listening conditions (noisy environments (Payton et al., 1994; Uchanski et al., 1996a), reverberation (Payton et al., 1994)) and for various listening populations (hearing-impaired listeners (Picheny et al., 1985b; Uchanski et al., 1996a; Payton et al., 1994), cochlear implant users (Picheny et al., 1985b; Ferguson, 2004), non-native listeners (Bradlow and Bent, 2002)). This makes clear speech robust in terms of intelligibility under various challenging environments, while its quality is preserved compared to other types of speech (e.g. Lombard speech that sounds unnatural when presented in quiet). Therefore, clear speech is characterized both as natural and robust in terms of intelligibility in various conditions. This renders clear speech the most appropriate speaking style for dynamically changing environments and suggests that its incorporation in telecommunication systems would be beneficial for a majority of users.

Transforming casual speech to clear speech is, therefore, important but also quite challenging. The main difficulty lies in the different linguistic content between the two speaking styles. For example, speech could become clearer by exploiting strategies such as simpler syntax or vocabulary. However, in this work we focus on acoustic differences between the two speaking styles and therefore for comparison purposes we use the same phrases and word sequences for analysis uttered both in clear and casual manner. Even though we choose to analyse pairs of clear and casual sentences with the same linguistic content, there are still linguistic content differences between them. Inside the same sentence, many different phonological phenomena may occur in

casual speech such as merging of phonemes, vowel reduction, co-articulation etc, phenomena which may also affect the intelligibility of casual speech. The degree of these phenomena may also differ across languages (Delattre, 1969; Meunier and Espesser, 2011; Burchfield and Bradlow, 2014; Baltazani, 2007). In this work we focus on acoustic modifications of casual speech without addressing these phonological differences. In fact, our analysis focuses on an English database of read clear and read casual speech corpora (Baker and Hazan, 2010) that contains same utterances read in clear and casual manner from the same speakers. In order to reduce the amount of phonological reductions and assimilations, read speech is chosen rather than spontaneous speech. Furthermore, relative to spontaneous clear speech, read clear speech is considered to be a more exaggerated style of speech and therefore acoustic differences between casual and clear speaking styles are more extreme when clear speech is elicited via explicit instruction (Hazan and Baker, 2010, 2011). Our proposed modifications are evaluated both in the English database and in a Greek database, suggesting that our modifications can be extended to other languages that share similar phonetic-acoustic characteristics with the English language(Bradlow et al., 2010).

Until now, many studies have examined the acoustic and intelligibility differences between casual and clear speech. However, few studies have performed transformations from the one speaking style to the other. Identifying and isolating specific acoustic features that are associated with the speaking adjustments of clear speech and their corresponding contribution to intelligibility is a difficult task. This difficulty lies on (1) speaker variability. Speakers adopt different techniques when speaking clearly. Therefore, analysis on different speakers does not always reveal consistent changes on acoustic features (2) the accuracy of the analysis process algorithms. For example, formant estimation is a difficult task for the analysis algorithms, especially on the transient regions where formant trajectories vary rapidly (3) modifying casual speech to mimic these adjustments. Previous studies that have attempted to modify casual speech towards clear speech, have failed to increase intelligibility. Possible reasons are that the modification techniques introduce disturbing artifacts to the signal or the methods proposed do not have the appropriate intelligibility impact. Last, the different linguistic content between the two speaking styles, clear and casual introduces yet another difficulty, as discussed above. In the sections that follow we will present in detail the acoustic feature differences between the two speaking styles and the modifications performed by related studies for modifying casual speech according to clear speech properties.

1.3 What makes clear speech intelligible?

The two different speaking styles, clear and casual, differ perceptually and acoustically. Overall, clear speech is proven to be more intelligible than casual speech.

On an acoustic and phonetic level of speech, intra-talker differences between the two speaking styles can be found in intensity (Picheny et al., 1986, 1989), speaking rate (Picheny et al., 1986, 1989), number and duration of pauses (Picheny et al., 1986, 1989), pitch (Bradlow et al., 2003; Hazan and Baker, 2010), long term RMS spectra (Picheny et al., 1986; Krause and Braida, 2004a; Hazan and Baker, 2010), modulation spectra (Krause

and Braida, 2004a) and vowel duration and vowel space (Picheny et al., 1986; Ferguson and Kewley-Port, 2002; Bradlow et al., 2003; Hazan and Baker, 2010). Acoustic analysis on clear and casual speech shows that these observed differences are not present for all speakers. For example, speakers can elicit clear speech with and without changing their pitch or increasing their voice volume level.

A comprehensive overview of clear and casual speech differences will be presented on this section. Specifically, we summarize prior work that studied the acoustic feature differences between clear and casual speech and their relationship to intelligibility. The acoustic features of the two speaking styles presented in this section are classified into two levels, segmental (i.e. phonetic) and suprasegmental (i.e. prosody). In linguistics (specifically, phonetics and phonology), a segment is a unit of sound of the size of a consonant or vowel (Ladefoged and Johnson, 2010). Vowels and consonants superimposed on the syllables to create utterances and other acoustic features known as suprasegmentals. Suprasegmental features are considered the stress, length, tone, and intonation, meaning features that do not belong only to a single consonant or vowel. Comparative acoustic analysis between clear and casual speech has revealed differences both on segmental and suprasegmental features. These differences will be described in detail in the section that follows. Also, a brief overview is given of related studies that have aimed to exploit these differences and enhance the intelligibility of casual speech. During this work, our research of clear speech has been guided by the excellent review performed by Uchanski (Uchanski, 2005).

1.3.1 Segmental Features

Duration of phonemes

The first extensive analysis on clear and conversational (casual) speech has been performed by Picheny et al. (1985b, 1986). In this study, the duration of vowels and consonants has been estimated, reporting increased durations of speech segments especially for the tense vowels (Ferguson and Kewley-Port, 2002; Picheny et al., 1985b, 1986). Bradlow et al. (2003) also reports vowel lengthening in clear speech relative to their durations in casual speech both for English and Spanish. A contradictory research performed by Krause and Braida (2004a) has influenced the existing belief of the importance of phoneme duration to intelligibility. Krause and Braida (2004a) trained the speakers to elicit clear speech at normal speaking rates (clear/norm), that is the same speaking rates as casual speech (conv/norm). Comparing clear/norm speech to conv/norm speech, only a small number of segments showed a statistically significant difference in duration.

Hillenbrand and Clark (2000) considered this increased vowel duration in clear speech as an important cue for intelligibility. However, after performing duration modifications on CVC syllables, no intelligibility enhancement was observed between modified and unmodified syllables.

Energy

Analysis on the relative energies between consonants and vowels showed an increased consonant-to-vowel energy ratio (CVR) on clear speech compared to casual speech, particularly for stops and affricates (Bradlow

et al., 2003). Other studies revealed that this CVR difference applies only for affricates (Krause and Braida, 2004a). An increase in the relative power of unvoiced consonants has been also reported by Picheny et al. (1986).

Hazan and Markham (2004) examined the correlation of CVR and intelligibility. In their study, they found no significant correlation between word intelligibility and CVR for nasals, fricatives and stop consonants in naturally-produced speech. However, it has been shown that enhancement of the consonant energy in words, consonant-vowel syllables (CV) and vowel-consonant-vowel (VCV) syllables improves consonant identification for normal hearing listeners (Gordon-Salant, 1986; Hazan and Simpson, 1998) and hearing impaired listeners (Montgomery and Edge, 1988; Gordon-Salant, 1987a). The level of consonant energy enhancement is higher than the energy level met in natural produced speech, but this artificial amplification has been shown to improve intelligibility at both CV, VCV words and nonsense sentences. Skowronski and Harris (2006) performed an energy re-distribution from vowels to consonants. This approach amplified the CV ratio and increased monosyllabic word intelligibility.

Short-term spectra

In the work of Picheny et al. (1985b, 1986), the short-term spectra of unvoiced consonants showed a greater spectral peak value in a higher frequency for clear speech compared to casual speech. Later on, Krause and Braida (2004a) found no consistent differences between the consonants of the two speaking styles. However, for vowel sounds, the short-term spectra revealed higher spectral prominences in clear/norm than in conv/norm speech. Krause (2001) increased the amplitude of the second and third formant of voiced segments as an attempt to simulate the higher spectral prominences that appears in clear/norm speech than in conv/norm. The formant shaping that she performed enhanced the intelligibility of unprocessed conv/norm speech significantly for normal-hearing listeners in noise but not for hearing-impaired listeners.

Formants

In vowel production, the quality of the vowel depends on the position of the articulators: (1) the height of the body of the tongue; (2) the front-back position of the tongue; and (3) the degree of lip rounding. Based on the positions of the articulators (high-low height, front-back position) the vowels are placed in a space as Figure 1.1 depicts. For the cardinal vowels, namely /i/, /u/, /a/, /ɒ/, the position of the tongue takes the more extreme positions in the mouth. Therefore, Figure 1.1 is a graphical method that shows where a vowel is located in the "articulatory" space. This traditional representation for vowels which classifies them according to the position of the articulators has been replaced by acoustic terms, and specifically by the first and second formant. The first formant, F1, is related to the height of the tongue while the second formant to the backness (front-back position) of the tongue. The representation of the articulators with formant values transformed the traditional vowel space into acoustic vowel space, with the formant values F1 and F2 contributing to the quality of the produced vowel.

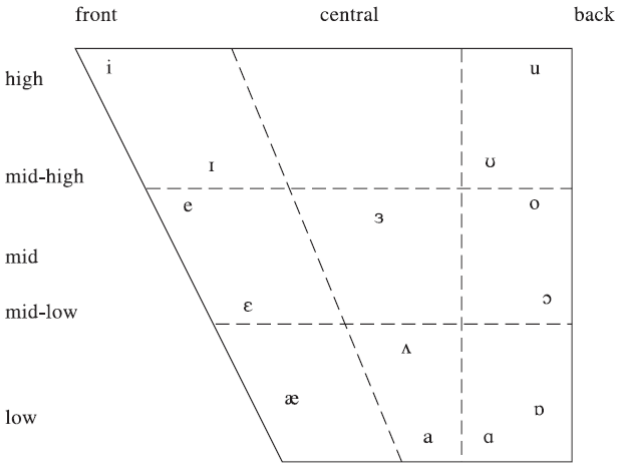


Figure 1.1: Vowel space for English language (Figure 2.2 page 44 Chapter 2 from the book of Ladefoged and Johnson (2010))

Clear speech appears to have a more expanded vowel space compared to casual speech (Chen, 1980; Picheny et al., 1986; Ferguson and Kewley-Port, 2002; Bradlow et al., 2003; Krause, 2001), meaning that the first and second formant frequencies tend to reach their target frequencies and are less variant in clear speech than in casual speech. This expansion is also observed in clear speech at normal speaking rates only for the tense vowels (Krause and Braida, 2004a). There seems to be a correlation between the expansion of the vowel space and intelligibility. Studies on the size of the vowel space indicate that speakers with larger vowel spaces are more intelligible than speakers with reduced spaces (Hazan and Markham, 2004; Bradlow et al., 1996). In particular, speakers who had a wide F1 range appeared to have higher intelligibility scores than speakers with a smaller F1 range. The F2 range was found to be significantly correlated with word intelligibility (Hazan and Markham, 2004), but was not found to be correlated with sentence intelligibility (Bradlow et al., 1996). Correlation between speech intelligibility and vowel space expansion has been also in the study of Sfakianaki et al. (2012) for hearing impaired people that experience difficulties in vowel articulation. Speakers with hearing deficits that produce more intelligible speech have more expanded vowel spaces than those who produce less intelligible speech.

In Krause and Braida (2004a), the authors suggest that vowel space expansion is not a necessary clear speech feature but, rather than that, it may be a concomitant result of slowing down and allowing talkers more time to reach more extreme vowel targets. This assumption is also supported by the fact that the expanded vowel space is observed only on the tense vowels and not on the lax vowels which are shorter. They suggest that the listeners possibly benefit from the entire formant movement rather than the midpoint of the vowel. Indeed, besides the expanded vowel space measured at the midpoint of vowels, clear speech often has increased rates of F2 transitions (Moon and Lindblom, 1994), longer durations of formant transitions (Chen, 1980) and narrower formant bandwidths Krause and Braida (2004a). In Hillenbrand and Nearey (1999) synthesized signals with original formants had higher vowel identification rate than signals with flat formants. In another study (Turner et al., 1997), the identification of stop phonemes increased with the lengthening of the formant transitions.

A first attempt towards vowel space expansion has been made by Mohammadi et al. (2012). A joint-density Gaussian mixture model has been used as the mapping function from casual formant frequencies to clear. Perceptual listening tests involved identification of vowels on CVC syllables in Babble noise of -2 and 3dB SNR by normal hearing listeners. Results showed no significant intelligibility enhancement of transformed speech compared to casual speech. An improvement has been observed for the lower SNR level by combining the vowel space expansion scheme along with duration modifications.

Consonant properties

According to Ladefoged and Johnson (2010), "a consonant can be said to be a particular way of beginning or ending a vowel, and during the consonant articulation itself, there is no distinguishing feature". To better understand what this means lets see Figure 1.2. In this figure, three words are spoken in American English, "bed", "dead" and the nonsense word "gɛg". There is virtually no difference in the sounds during the actual closures of [b, d, g]. What primarily distinguishes these three stops are the onsets and offsets of the second and third formants on the start and end of the vowel, depicted with white lines.

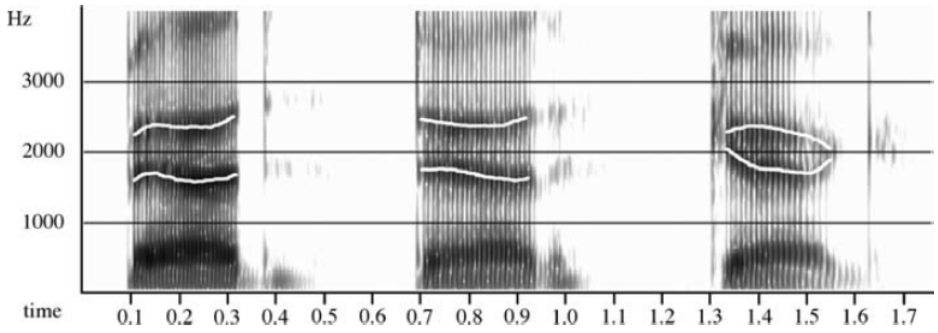


Figure 1.2: Spectrograms of the words bed, dead, and the nonword [gɛg] spoken by an American English speaker

The same applies for the unvoiced consonants [p, t, k], nasals [m, n, ŋ]. and voiceless fricatives [f, θ] and [s, ʃ]. Apart from the acoustic information that is inherent in their own production, consonants convey much of their quality by the effect on the adjacent vowel. This links the properties of consonants with vowel properties and specifically with the formant transitions that are met in clear speech more enhanced than in casual speech (Moon and Lindblom, 1994; Chen, 1980; Krause and Braida, 2004a).

Other properties of consonants (besides energy that is described in the previous corresponding section) is the voice-onset time (VOT). VOT is a feature of stop consonants and is defined as the length of time that passes between the release of a stop consonant and the onset of voicing. Previous research on clear and casual speech (Picheny et al., 1986; Chen, 1980) found longer VOTs in clear speech than in casual speech for unvoiced plosives. Examining ΔVOTs, the difference of VOTs of voiced and unvoiced plosive, no relation was found between ΔVOT and intelligibility (Bond and Moore, 1994). However, in another study of speech produced by deaf talkers (Monsen, 1978), significant correlations were found between ΔVOT and intelligibility.

An increase of VOT degraded the intelligibility of CV syllables (Thomas, 1996) while the combination of

intensity and duration changes of consonants improved the recognition of consonants (Gordon-Salant, 1986; Thomas, 1996). The intensity and duration changes of consonants possibly are correlated with the formant transitions from consonants to vowels. Enhancing formant transitions increases speech intelligibility (Turner et al., 1997; Thomas, 1996). Harris and Skowronski (1988) proposed a spectral energy redistribution algorithm to enhance spectral transitions. Psycho-acoustic tests showed an intelligibility enhancement of speech. Jayan et al. (2008) achieved to detect automatically transient parts of speech and perform on these regions intensity and time-scale modifications. Intelligibility tests reported intelligibility enhancement of non-sense syllables on high noise levels (-9 and -12dB SNR).

Phonological features

Degemination (two phonemes merging into one), vowel reductions (i.e. vowels becoming schwa-like), burst eliminations (i.e. unreleased stops, which are common in final position) appear more frequently in casual speech compared to clear speech. These phonological phenomena in casual and clear speech have been observed in various studies (Picheny et al., 1986; Bradlow et al., 2003). However, in the study of Krause and Braida (2004a), no significant differences in these phenomena have been found between clear speech at normal speaking rates and casual speech, suggesting that the contribution of these phenomena to the intelligibility advantage of clear speech is relatively small.

1.3.2 Suprasegmental Features

Fundamental Frequency (F0)

In their effort to elicit clear speech, many speakers change their pitch both in level and range compared to casual speech (Bradlow et al., 2003; Picheny et al., 1986; Krause and Braida, 2004a), suggesting larger amounts of laryngeal tension. These changes even though apparent for many speakers are not consistent across speakers (Picheny et al., 1985b; Krause and Braida, 2004a), questioning the connection of F0 with intelligibility.

Previous studies on inter-talker variability, found no correlation of the average F0 with intelligibility (Bond and Moore, 1994). More recent studies (Bradlow et al., 1996; Hazan and Markham, 2004) have also reported no correlation between average F0 and range with word intelligibility. However, Bradlow et al. (1996) have found a significant correlation between F0 range and sentence intelligibility in one speaker out of 20. Increase in pitch is also observed in other types of intelligible speech like Lombard speech, advocating a correlation between intelligibility enhancement and F0. There are, nevertheless, other types of speech with extreme pitch height and variation (i.e. infant directed speech) that are proven to be less intelligible than plain speech in noise (Mayo et al., 2012).

Examining further an indirect relation between F0 and intelligibility, Gordon-Salant (1987b) suggested that F1, F2 and F3 formants along with formant transitions can describe English vowels regardless of F0 values. However, other studies (Hoemeke and Diehl, 1994) have reported that the perception of vowel height is influ-

enced by the distance between F0 and F1, revealing that the relationship between F0 and speech intelligibility may be more complex and cannot be measured by metrics like F0 mean and range.

Efforts have been made to increase speech intelligibility by changing the average F0 and F0 range of speech. Such modifications on casual speech did not prove advantageous (Lu and Cooke, 2009; Krause, 2001). Similar results to casual speech have been obtained on synthetic speech. Sommers and Barcroft (2006) decreased and increased the F0 values on synthetic speech with no impact on word recognition. On the other hand, artificial flattening of F0 has been shown to degrade speech intelligibility (Laures and Bunton, 2003; Watson and Schlauch, 2008).

The question of whether F0 is an important cue for intelligibility still remains debatable. It is not evident if these pitch modifications are a mechanism of increasing intelligibility or a result of the increased vocal effort that aims to move the energy on higher frequency regions of the speech spectrum.

Intensity

Clear speech is produced at levels 5 to 8 dB greater than those of conversational speech (Picheny et al., 1985b). However, in all listening experiments that test the intelligibility of the two speaking styles, the overall intensities of clear and casual speech are equalized. Therefore, there are other features contributing to the increased intelligibility of clear speech.

Long Term Average Spectra and spectral tilt

An increase in energy in the 1-3 kHz frequency range of the long-term average spectrum (LTAS) with a simultaneous decrease in the spectral tilt is the prominent characteristic of Lombard speech (Junqua, 1993; Pittman and Wiley, 2001; Summers et al., 1988) revealing changes in the articulation and vocal tension. For clear speech, Picheny et al. (1986) has found slight differences between the LTAS of conversational and clear speech, with all talkers showing a tendency for higher spectrum levels in clear speech at higher frequencies. Krause and Braida (2004a); Hazan and Markham (2004) have also observed an increase in energy in the 1-3 kHz frequency range of the long-term average spectrum. Both studies report that this increase is significantly correlated with intelligibility. Different results, though, have been observed for the spectral tilt on clear speech. Hazan and Markham (2004) has found that the slope of the LTAS is not correlated with intelligibility. The fact that the spectral tilt remains stable while the LTAS in the $1-3$ kHz frequency region increases, indicates that the formant bandwidth, rather than the spectral balance (return phase of the glottal source), contributes to the intelligibility of clear speech.

Speaking rate

The most prominent modification that talkers employ when they elicit clear speech is the decrease of the speaking rate, compared to casual speech. The decrease of speaking rate in clear speech is attributed to an increase in the number of occurrences of pauses and their average duration, and to an increase in the duration

of sound segments (Picheny et al., 1986; Bradlow et al., 2003; Krause and Braida, 2004a) explicitly described in Section 1.3.1 ("Duration of phonemes" and "Consonants").

To quantify the differences on elongation and pause distribution between the two speaking styles, a brief comparative analysis will be provided. Picheny et al. (1986) reported speaking rates of 160 to 205 wpm for conversational speech and 90 to 100 wpm for clear speech. In the work of Bradlow et al. (2003), the overall increase on sentence duration from casual to clear was 51% and 116% for one male and one female talker, respectively. Analysis on pause distribution and duration by Krause and Braida (2004a) revealed 6.36 pauses in average per sentence with an average duration of 42 ms per pause for casual speech (conv/norm) while clear elicited speech at slow speaking rates (clear/slow) had 12.64 pauses per sentence with an average duration of 130 ms.

Findings of related studies have yielded numerous important observations on the clear speech speaking rate. The effect of the duration has been examined on different levels (phoneme, word and sentence level) of intelligibility (Krause and Braida, 2004a; Bradlow et al., 1996; Hazan and Markham, 2004; Uchanski et al., 1996a). Depending on the level, different results have been reported. More analytically, word duration has been found to be positively correlated with word intelligibility (Hazan and Markham, 2004; Bond and Moore, 1994). On the other hand, in Bradlow et al. (1996) no correlation is reported between speaking rate and sentence intelligibility.

There is additional evidence to support that speech intelligibility is independent of speaking rate. In the work of Krause and Braida (2004a), clear speech has also been produced without decreasing the speaking rate (clear/norm), after training the speakers. Clear/norm speech had 6.78 pauses per sentence in average with an average duration of 49 ms, very close to the conv/norm corresponding. Also, other types of intelligible speech in noise are elicited at speaking rates comparable to casual speaking rates. Letowski et al. (1993) reports no difference in speaking rate between Lombard speech and speech produced in quiet. This suggests that acoustic features other than speaking rate contribute to intelligibility.

However, in the study of Krause and Braida (2004a) the intelligibility advantage of clear speech produced at slower speaking rates (clear/slow) has been shown to be greater than that of clear speech produced normal speaking rates (clear/norm). Moreover, deletion of pauses in clear speech has reduced speech intelligibility Uchanski et al. (1996a). Furthermore, the relation between speaking rate and intelligibility is reinforce by psychoacoustic studies. In the work of Ghitza and Greenberg (2009), speech was three times compressed reducing by 50% its intelligibility. Then, periodic pauses of 80 ms were added to the compressed signal. Despite the fact that the resulting signal was disturbed by pauses, with pauses falling between words, its intelligibility increased near 36% proving a strong connection between speaking rate (tempo of speech) and the ability of the brain to process and decode speech. This explanation is similar to the "effortfulness hypothesis" (Rabbitt, 1991, 1968), which assumes an interaction between perceptual and cognitive processes. The "effortfulness hypothesis" suggests that simulated or age-related sensory declines may decrease capacity for semantic integration in language comprehension. By expanding the duration of speech through elongation and pause insertion, listeners can use more time for higher cognitive processes. Possibly clear speech is highly

intelligible to a variety of listeners (normal hearing in noise, listeners with disorders, elderly people) due to this property.

There is a large number of works that examine the influence of the time-expanded speech for normal listeners in noise and on several disorders, like verbal apraxia and aphasia (Coyle et al., 2004; Nejime et al., 1996) and for hearing impaired population (Uchanski et al., 2002; Nejime and Moore, 1998). However, time-scaling transformations had little (Schmitt, 1983) or no beneficial effect on the intelligibility of speech (Small et al., 1997) while in others it degraded the intelligibility of original speech (Nejime and Moore, 1998; Kemper and Harden, 1999). Two studies evaluated artificial manipulations of the speaking rate of casual and clear speech. Picheny et al. (1989) investigated the efficacy of speech-rate reduction for hearing-impaired people using Malah's algorithm. Evaluations on sentence intelligibility showed that the speed-reducing processing led to poorer intelligibility. Later, Uchanski et al. (1996a) evaluated the efficacy of speech-rate reduction using nonuniform processing with elongation of segments and pause insertion. Although the nonuniform time scaling was less deleterious to intelligibility than the uniform time scaling used in earlier studies (Picheny et al., 1985a, 1986, 1989), the intelligibility of the slowed speech was less than that of the original speech for all subjects, possibly attributed to disturbing artifacts. Recently, Liu and Zeng (2006) attempted to identify the importance of temporal information in clear speech perception. In their experiments, they performed uniform time scaling to match the speaking rate between clear and casual speech and decreased the speaking rate in casual speech, without processing artifacts, by increasing silent gaps between phonetic segments. Their results showed that processing artifacts in uniform time scaling reduced speech intelligibility. Inserting gaps in conversational speech improved the intelligibility, but this improvement might be a result of increased short-term signal-to-noise ratios during level normalization.

In summary, even though many studies have been conducted in terms of phoneme, word, and sentence duration, the findings are not all in agreement. However, the decrease in speaking rate allows speakers to enunciate all of the words deliberately, with caution and without omitting phonemes. Moreover, word boundaries are respected by inserting pauses. Consequently, the different articulation adds more acoustic information in clear speech compared to casual speech. With respect to the results presented in Krause and Braida (2004a), clear speech produced at a slower speaking rate potentially achieved higher intelligibility due to the existence of more pauses and to the fact that slowing down the speech gives the listener more time to process the message.

Temporal envelope modulations

The speech signal can be represented as a sum of amplitude-modulated signals in a number of narrow frequency subbands spanning the signal bandwidth (Drullman et al., 1994b). The output waveform of each subband can be considered as an AM modulated signal consisted of a carrier signal (temporal fine structure) and an envelope (temporal envelope). There is a great dichotomy in auditory perception between temporal envelopes and fine structure cues (Smith et al., 2002; Sheft et al., 2008; Liu and Zeng, 2006; Shannon et al., 1995; Zeng et al., 2004; van der Horst et al., 1999). However, in this work we focus on the temporal envelopes.

Previous acoustic analysis between the two speaking styles, clear and casual, revealed higher modulation depth of the temporal envelopes (Krause and Braida, 2004a; Liu and Zeng, 2006) of clear speech. This increased modulation depth is suggested to be independent of speaking rate. Figure 1.3 shows the modulation depth differences between (i) clear speech elicited in slow speaking rate (clr/slow) and conversational (casual) speech (conv/normal) and (ii) clear speech elicited in normal speaking rate (clr/normal) and casual speech (conv/normal). Clear speech elicited in normal speaking rate has been achieved after training the speakers to elicit faster clear speech and has been proven to be more intelligible than casual speech (but a little less effective than clr/slow speech). The majority (five out of six) of the speakers present a higher modulation depth of clr/slow speech than that conv/normal speech. Four out of six speakers, maintain this depth increase when they speak clear but faster. This suggests that modulation depth increase is independent of the speech rhythm. The speaking rate possibly affects the modulation frequency that spans in higher modulation frequency regions in faster clear speech. Therefore, acoustic analysis on casual and clear speech of different speaking rates reveals an increase of the modulation depth of the temporal envelopes of clear speech compared to that of casual speech.

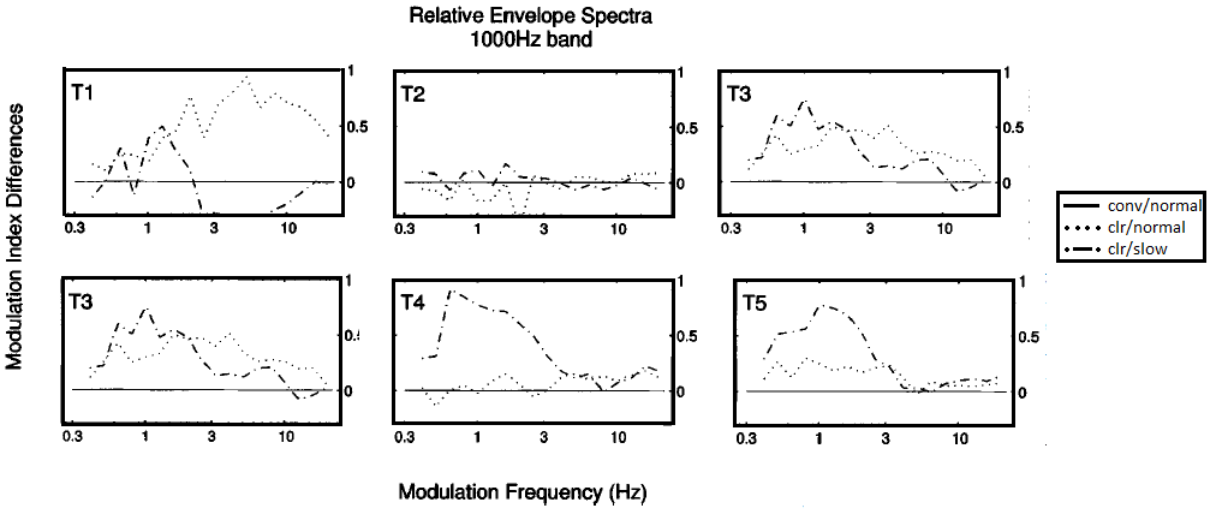


Figure 1.3: Temporal envelope modulation data (Krause and Braida, 2004a). Envelope spectral differences of the 1000-Hz octave band, obtained by subtracting the conv/normal modulation spectrum from the clr/normal and clr/slow modulation spectra for each talker ($T_i, i = 1..5$) and speaking style, depicted by modulation index.

The fact that clear speech appears to have greater modulations than casual speech, does not necessarily link this property with intelligibility. However, there is a growing evidence that there is a significant contribution of the amplitude modulations to the intelligibility advantage of clear speech. First, Houtgast and Steeneken (1985) quantified the modulation depth of the temporal envelopes using the modulation index. The modulation index has been used as an objective measure of intelligibility in room acoustics (Houtgast and Steeneken, 1973). Based on the modulation index the speech transmission index (STI) has been proposed, an objective measure used for prediction of speech intelligibility in noise and reverberation (Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985).

Neurophysiological and psycho-acoustical studies also link the perception of sounds with modulations. Psychophysical experiments by Bacon and Grantham (1989) indicated that there are channels in the auditory system which are tuned to the detection of low frequency modulations. Ewert and Dau (2000) revealed that there is a frequency selectivity in the envelope-frequency domain (i.e., modulation domain), analogous to the frequency selectivity in the acoustic-frequency domain. Other studies have shown that neurons in the auditory cortex are thought to decompose the acoustic spectrum into spectro-temporal modulation content (Mesgarani and Shamma, 2005; Schonwiesner and Zatorre, 2009) and are best driven by sounds that combine both spectral and temporal modulations (Shamma, 1996; Kowalski et al., 1996).

The study of Drullman et al. (1994a,b) came to re-enforce the connection of low frequency modulations with intelligibility. The aforementioned study showed intelligibility degradation of speech after smearing low-frequency modulations. Modulation frequencies between 4 and 16 Hz were found to contribute the most to intelligibility, with the region around 4-5 Hz being the most significant, reflecting the rate at which syllables are produced.

Based on the outcome of the study of Drullman et al. (1994b), modulation processing has been introduced to separate speech from noise and therefore, enhance the intelligibility of speech in noisy environments. The denoising algorithms are based on preserving the low-frequency modulations (4-16Hz) of the spectral envelope which are considered important for intelligibility, while discarding other modulations imposed by the masker (Wójcicki and Loizou, 2012; Paliwal et al., 2010; Won et al., 2008; Mesgarani and Shamma, 2005).

Rather than denoising the speech signal, other studies focus on re-enforcing the amplitude modulations of speech before presented in noise, as naturally happens in clear speech. In Kusumoto et al. (2005) modulation spectral components between 1-16 Hz are enhanced prior to distortion of speech in reverberant environments. The drawback of the designed modulation filters is that their efficiency depends on the reverberation condition. In Krause and Braida (2009) the temporal envelope of casual speech is transformed to have higher modulation depth as in clear speech. The modulation depth enhancement scheme is performed for low modulation frequencies (less than 4Hz) which are considered to be important for phoneme identification. However, this modification technique decreased speech intelligibility due to processing artifacts. Therefore, even though modulations are judged as important for intelligibility and perception, increasing the modulation depth of the temporal envelopes, while enhancing speech intelligibility, has not yet been efficiently addressed.

1.4 Related work on speech modifications from casual to clear speaking style

In the previous section, we presented an overview of the feature differences between the two speaking styles, clear and casual, and attempts that have been made to enhance the intelligibility of casual speech based on the acoustic features of clear speech. In this section, we will summarize these attempts, revealing the necessity for further modifications on casual speech in order to achieve the clear speech benefit.

In terms of prosodic transformations, several studies have examined duration modifications of speech in order to enhance its intelligibility in noise for normal listeners and hearing impaired listeners. Motivated by the

expanded vowel duration that appears in clear speech compared to casual speech, Hillenbrand and Clark (2000) performed modifications on CVC syllables with no intelligibility benefit. However, Gordon-Salant (1986); Montgomery and Edge (1988) performed duration changes of the consonants on CV-syllables, a property that is not met on clear speech. Possibly this choice was perceptually driven since Gordon-Salant (1986) had shown that a combination of amplifying the consonant energy and lengthening consonant durations improved consonant identification for normal-hearing listeners and for elderly hearing-impaired subjects (Gordon-Salant, 1987a). The effect of slowing down speech on intelligibility has been also examined on a sentence level for various populations (cochlear implant users in noise (Nejime and Moore, 1998), on Alzheimer patients (Small et al., 1997) and on elderly listeners (Schmitt, 1983; Kemper and Harden, 1999)). Directly connected with clear and casual speech properties are the works of Picheny et al. (1989) and Uchanski et al. (1996a). These studies performed modifications of casual speech to reach the duration of clear speech using uniform time-scaling (Picheny et al., 1989) and non-uniform time-scaling and pause insertion (Uchanski et al., 1996b; Liu and Zeng, 2006). However, none of the studies found an intelligibility benefit of processed speech. Possible reasons are the introduction of artifacts from the time-scaling approaches (Picheny et al., 1989; Uchanski et al., 1996a) and the fact that the pause insertion scheme copied the location of pauses on clear speech possibly disrupting the casual signal. Last, possible reason of failure of the time-scale modification schemes to improve intelligibility is that the elongation and pause insertion schemes were not combined to perform time-expansion, as in clear speech. On the contrary, each modification scheme was evaluated separately.

Pitch changes occur in many speakers when switching from the casual to the clear speaking style. However, the effect of average F0 and range to the intelligibility enhancement of clear speech is not known. F0 modifications from casual to clear did not prove advantageous, despite the fact that artificial flattening of F0 has been shown to degrade speech intelligibility (Laures and Bunton, 2003; Watson and Schlauch, 2008).

Unlike prosodic transformations which have been proven so far ineffective in increasing the intelligibility of casual speech, spectral transformations have been proven advantageous. Niederjohn and Grotelueschen (1976) increased speech intelligibility in noise by high-pass filtering the signal and applying amplitude compression. Later, Skowronski and Harris (2006) linked the modifications that were performed by Niederjohn and Grotelueschen (1976) to clear speech properties by performing CV ratio boosting. Krause and Braida (2009) enhanced the intelligibility of casual speech for normal hearing listeners by 14% in -1.8 dB speech shaped noise (Nilsson et al., 1994) by increasing the amplitude of the second and third formant of voiced segments. This energy increase in the region of F2 and F3 formants contributed to the amplitude increase in 1-3 kHz frequency region of the LTAS, as observed in clear speech compared to casual speech. The aforementioned studies suggest that consonant emphasis and amplitude boosting in perceptual important frequency regions (1-3kHz) can be beneficial for the intelligibility enhancement of speech. However, these modifications have not been yet combined, meaning that not all clear speech properties have been explored to enhance casual speech intelligibility. Taking into consideration that in the work of Krause and Braida (2009) clear speech has up to 34% intelligibility gain compared to casual speech, there is still much room for improvement.

Spectral transformations inspired by clear speech properties also involve formant modifications. As men-

tioned in the previous section, clear vowels appear to have a more expanded vowel space than casual vowels. Mohammadi et al. (2012) performed vowel space expansion using a Gaussian mixture model to derive the mapping function from casual formant frequencies to clear. Evaluation on normal hearing listeners in noise showed no significant intelligibility enhancement of transformed speech compared to casual speech. Therefore, it is worth exploring different vowel space expansion techniques.

Last, the information contained in the temporal envelopes of narrowband signals is considered to be important for perception and intelligibility. This information is reflected on the modulation depth of the temporal envelopes on low modulation frequencies, between 2 to 16 Hz, which is presented enhanced in clear speech compared to casual speech. Possibly, the intelligibility benefit of clear speech on various listeners and for different environmental conditions is due to its increased modulation depth compared to other speaking styles (Lombard speech, casual speech). Increasing the modulation depth of the temporal envelopes with a positive impact on the intelligibility of the modified speech has been only efficiently addressed by Kusumoto et al. (2005). The designed modulation filter was applied to speech prior to its distortion in reverberation. The drawback of the proposed modulation filter is that its efficiency depends on the reverberation condition and only when consonants are preceded by highly powered segments. These limitations restrict the use of the modulation filters. Therefore, other modulation enhancing techniques can be explored.

1.5 Thesis subject

This work explores novel methods for enhancing the intelligibility of casual speech based on clear speech properties. Casual speech is expected to benefit from such modifications since clear speech is a speaking style with highly intelligibility impact on various listening populations (native and non-native listeners, hearing impaired, cochlear implant users, elderly people, people with learning disabilities etc.) in many listening conditions (quiet, noise, reverberation). Previous research on clear and casual speech has revealed a great number of feature differences between the two speaking styles. However, this work aims on modifying those features that are most frequently met among speakers. These features are possibly connected with intelligibility but corresponding modifications conducted by previous research either degraded intelligibility or gave moderate results. In this thesis, the following modifications on casual speech are performed:

- prosodic modifications (duration and F0 modifications)
- vowel space expansion
- spectral transformations
- modulation depth enhancement

Evaluations are conducted with native and non native listeners in noisy environments (Speech Shaped Noise) and reverberation. However, recently proposed objective intelligibility scores (Rherbergen and Versfeld, 2005; Tang et al., 2013) highly correlated with intelligibility are also used. Listening tests are also

performed by a small number of people with hearing deficits but not for all modifications. For comparison purposes, our modification techniques are compared with clear speech and with the most successful modification algorithm emerged from an extensive evaluation of the intelligibility of a variety of methods (Cooke et al., 2013), namely the Spectral Shaping and Dynamic Range Compression (SSDRC) proposed in Zorila et al. (2012).

Finally, the main objective of this work has been to perform modifications that do not require excessive computation and statistical training. Therefore, the first thought was to create simple and explicit transformations inspired by clear speech properties without requiring the existence of the corresponding clear signal. This designing preference would facilitate the incorporation of the transformation algorithms to a variety of speech systems that demand quick response and preferably no statistical training.

1.6 Contributions

This main contributions of the thesis can be summarized as follows:

- The most prominent modifications that talkers employ when speaking clearly are the decrease of speaking rate and to a less extent the increase of average F0 and range. However, there is a great controversy between research works on whether or not these modifications contribute to intelligibility. We propose a simple method to evaluate the impact of speaking rate and F0 contour on intelligibility. The method is based on performing pitch and duration modifications from casual to clear speech and vice versa. Unlike other studies, transformations did not create artifacts to the signal, allowing us to draw concrete conclusions on what affects intelligibility. This work has been published in **Interspeech 2012** (see Appendix).

- Time-scaling modifications of casual speech have been proven non beneficial for intelligibility. Possible reasons are the introduction of artifacts, the insertion of pauses on places that disrupted casual speech and that no combination of the elongation and pause insertion has been exploited by the time-scaling modification schemes, as it is found on clear speech. This work proposes a novel method that performs nonuniform time-scaling with pause insertion. The method is inspired by the clear speech properties but performs modifications based on the perceptual characteristics of casual speech in order to create the minimum possible disturbances that may degrade intelligibility. The efficiency of the time-scaling technique is compared with state-of-the-art methods (1) segmental time-scaling and (2) uniform time-scaling. Results from this study suggest that time-scaling modifications are beneficial for reverberant environments and reveal a connection of the time-scaling factor with the reverberation time. This work has been published in **eNTERFACE'12** and in **Interspeech 2015** (see Appendix).

- Formant modifications may be proven beneficial for intelligibility since they are strongly connected with articulation and vowel recognition. Previous work on vowel space expansion transformed both formant frequencies and the spectral envelopes of speech. To out knowledge, there has been no previous effort

to address vowel space expansion in isolation and to assess the corresponding intelligibility impact. In this work, we explore vowel space modifications. Vowel space expansion is performed using a clear speech-inspired frequency warping method. Unlike other approaches, this method is explicit and can be applied to any speech signal without the need of training. The results of this study have been published in **Interspeech 2013** and in **Computer Speech and Language** (see Appendix).

- Spectral transformations inspired by clear speech properties have been proposed by previous studies. Specifically consonant emphasis and energy boosting of $1-3$kHz frequency region of the LTAS have been proposed to enhance the intelligibility of clear speech. However, these modifications have not been combined. Studies have not explored all clear speech properties since their analysis was limited on specific speakers or on specific segments (voiced parts). We propose a different approach. Our analysis is performed on a large number of sentences of various speakers for the two speaking styles. This averaging shows a general trend between the two speaking styles, revealing more than one frequency regions with enhanced spectral content in clear speech. We propose a method that boosts the spectral content of these frequency regions. The proposed technique has multiple benefits on (1) intelligibility (2) quality (3) computation. This work has been published in **Interspeech 2014**. A combination of this method with time-scaling modifications has been proposed in **Interspeech 2015** for enhancing the intelligibility of speech in reverberation. This work has been **patented** (see Appendix) for revealing a relation between the time-scaling factor of duration expansion techniques with the reverberation time in order to have an intelligibility benefit.

- Enhancing the modulation depth of low frequency modulations has been addressed only for reverberation with moderate results (less than 10% of intelligibility enhancement) and limitations. The contribution of this work in terms of modulations is twofold. First, a novel method is proposed to enhance the intelligibility advantage of speech in noise. The proposed method is based on the idea of coherent demodulation (Atlas and Janssen, 2005), where the signal has to be analyzed in very short narrowbands to reveal the fine structure and the temporal envelope of the signal. The idea of coherent demodulation is applied in this work by decomposing the speech signal into quasi-harmonics using the extended Adaptive quasi-harmonic model (Kafentzis et al., 2014). eaQHM is an accurate analysis and synthesis model, highly adaptive to the signals characteristics. After decomposing speech into a carrier (fine structure) and an amplitude (temporal envelope), a transforming function is applied to the temporal envelopes to increase their modulation depth. The transforming function is based on clear speech properties and on perceptual characteristics related to the non-linear process of sounds by the ear. The re-synthesized signal is highly intelligible in noise compared to the original signal (more than 30% intelligibility enhancement). Second, to our knowledge none of the previous works has examined the impact of the spectral enhancing techniques to the modulations of the temporal envelopes. This work shows that the spectral modification algorithms enhance the modulation depth of the temporal envelopes on specific frequency regions. Therefore, the intelligibility effectiveness of these techniques may be attributed to a

degree to this modulation depth enhancement.

Other less significant contributions are:

- Comparison of clear speech and Lombard speech terms of intelligibility and comparative acoustic analysis on specific features.
- Highlighting and dealing with the high variability in scores among listeners. This variability has been greater for the non-native listeners showing differences on language processing. Moreover, problems with variability have been observed with sentence or word difficulty. This led us to adopt more controlled "random" scenarios on intelligibility tests in order to ensure that all algorithms will be evaluated under the same circumstances (i.e. same difficult sentences).
- An extensive analysis is performed on the clear and casual speech corpus of the LUCID database (Baker and Hazan, 2010). The database has been annotated. Pause distributions have been estimated per speaking style, vowel spaces have been extracted per speaker and per speaking style etc.
- A great number of modification algorithms has been implemented. Modification algorithms included application of A-weighting filters, transition detection and enhancement, formant shaping using line spectral frequency pairs. However, these techniques were not advantageous and are not reported.

1.7 Limitations

Prior to presenting this work, the main limitations are pointed out:

- Listening tests have been performed by different listener populations and in different conditions. Indeed, our modification algorithms have been evaluated in different conditions (reverberation, speech-shaped noise of various SNR levels) and for various populations (native, non-native) which is desirable. However, in some cases a more consistent methodology could be followed. For example, some modifications have been evaluated both by native and non-native listeners while other modifications have been evaluated only by native or non-native listeners. This mandatory option (due to the difficulty in finding subjects for performing the intelligibility tests) should not affect the major findings of our research, since comparisons have been made with original casual speech in all listening experiments.
- The majority of the listening tests has been designed to test intelligibility. In some cases (e.g for the time-scaling and pitch modifications) the listening effort was measured rather than intelligibility. However, it is made clear to the reader when the listening experiment measures intelligibility or listening effort.
- Two different languages are used for analysis and evaluation. Specifically, transformations are inspired by speech analysis on a English database while evaluation is conducted both in the English and in a Greek database. The choice for this "inconsistency" in the speech corpus is deliberate in order to show

that some of the proposed modification techniques may be applied on other languages without the need of additional training.

- A random subset of speakers is selected from the total number of speakers in the English database. A more refined selection of speakers could have been used based on the maximum intelligibility distance between their speaking styles, clear and casual.

- The modification techniques proposed by this work are speaker independent. The extracted features used on our modification algorithms are derived by a large number of sentences uttered by male and female speakers. Modifications adapted to speaker (or to gender) could be more beneficial. However, our aim was to perform explicit and speaker independent modifications.

- There is a lack of statistical evaluation of the effect of speaking style on some acoustic measures (e.g F0 variation). However, the database used for analysis has been analysed in depth by other studies and the statistical significance has been shown to apply on the features selected for modification. The primary goal of this study was not to reveal the statistical significance of the acoustic features with the speaking style but to quantify the differences between clear and casual speech and incorporate these measured differences in the proposed modifications. However, for completeness, statistical evaluation could have been done for the subset of the English database used for training.

1.8 Structure of thesis

The rest of this thesis is organized in 7 chapters, as follows:

- Chapter 2 presents our acoustic analysis of the two speaking styles, clear and casual, focusing on the main features that inspire our proposed modifications.

- Chapter 3 examines the impact of speaking rate and pitch on intelligibility and explores prosodic modifications for enhancing the intelligibility of casual speech. Objective and subjective results are presented in speech shaped noise (SSN).

- Chapter 4 presents the vowel space expansion technique. Objective and subjective intelligibility scores are presented on clear, casual and modified casual speech in SSN.

- Chapter 5 presents our spectral modification technique for intelligibility enhancement with quality restrictions. First, a comparative analysis on the intelligibility between clear, casual and Lombard speech is presented in order to show the intelligibility differences between natural speaking styles. Then, a Lombard-like modification technique, namely the SSDRC is presented. In this work, SSDRC serves as an upper limit of intelligibility since it has been proven the most powerful modification technique in terms of intelligibility in SSN. Next, our clear-inspired proposed spectral modification is described and

evaluated in two conditions, SSN and reverberation, combined also with time-scaling transformations. Quality tests (preference tests) and objective and subjective intelligibility scores are presented.

- Chapter 6 describes the proposed modulation enhancement technique, DMod. Subjective intelligibility tests in SSN are used to evaluate the efficiency of our method.
- Chapter 7 concludes the thesis and suggests future research directions.
- Appendix contains links to supportive sound material for each modification algorithm proposed. Last, it summarizes the publications that describe this work.

Chapter 2

Acoustic analysis on Read Clear and Read Casual speech

This chapter describes the clear and casual speech corpora used for algorithm development and evaluation. The corpora used for our analysis is the read clear and read casual speech from the LUCID database (Baker and Hazan, 2010). Relative to spontaneous speech (Hazan and Baker, 2010, 2011), read speech is an exaggerated form of speech. Specifically, read clear speech compared to spontaneous clear speech shows greater change in F0 range and decrease in speaking rate (Laan, 1997; Hazan and Baker, 2011).

Read clear speech appears to be more intelligible than spontaneous clear speech (Smith, 1982). One probable reason for this enhancement, is that when the talkers are instructed to speak clearly in the read task, since they do not have any communication feedback, they follow consistently the instructions throughout the task. On the other hand, in spontaneous dialogue tasks, the talkers use the feedback from the listener's side in order to communicate effectively and simultaneously minimize their speaking effort. With the passage of time, the speaker reduces his/her speaking effort until a misunderstanding from the listener's side adjusts his/her speaking to higher intelligibility levels. Therefore, spontaneous speech is characterized by parts of high and low clarity (Hazan and Baker, 2011).

In LUCID corpora, read casual speech was produced after instructing speakers to read the sentences "casually as if talking to a friend" whereas for read clear speech the instructions were to speak "clearly as if talking to someone who is hearing impaired". Forty speakers participated in the recordings (20 Male and 20 Female). Speakers in this database are Southern British English between 19 and 29 years old with no speech or language disorders. 144 sentences were recorded for each speaker and for each speaking style, clear and casual. The sentences are meaningful and simple in syntax.

Acoustic analysis performed on the database (Hazan and Baker, 2010, 2011; Baker and Hazan, 2010) showed differences between read clear and read casual speech in the spectral and time domain. Specifically, when speakers read clearly, they produced speech with higher fundamental frequency, higher energy in the frequency band 1-3kHz and higher range in first and second formant than when reading casually. However, these adaptations showed variability among speakers, depending on the strategy that each speaker adopted to

produce the specific style of speech (Lindblom, 1990). The feature that was prevalent for all speakers, when eliciting read clear speech, was the decrease of the speaking rate. Speakers slowed down their speech to a greater extent when reading clearly, by extending the mean word duration and inserting pauses. As reported by Hazan and Baker (2011), the mean word duration was approximately 1.7 times longer in read clear speech than in read casual speech.

In the work of Hazan and Baker (2010, 2011), explicit statistical analysis is performed on acoustic differences between read clear and read casual speech, examining for each acoustic difference the effect of speaking style, the effect of talker gender etc. In the sections that follow, we will present a comparative analysis between the two speaking styles, read clear and read casual speech aiming on quantifying differences between the two speaking styles and use the observed and measured differences on our modifications. Therefore, in order to lower the computational cost imposed by the number of data need to be processed and analyzed, we reduce the number of speakers in the LUCID database and we focus our analysis on a subset of speakers consisted by 4 female speakers (F12, F13, F14, F15) and 4 male speakers (M11, M13, M15, M33). The most important result from this acoustic analysis is that the observed and measured differences between the two speaking styles will be incorporated in our proposed modifications for enhancing the intelligibility of casual speech. For simplicity, the term clear and casual will be used to refer to read clear and read casual speech respectively, throughout the document.

2.1 Pause frequency, pause duration and speech duration

Reduction in speaking rate includes an increase in pause frequency and duration, as well as an elongation of the speech segments. The acoustic analysis performed by Hazan and Baker (2010) reveals a significant increase in the mean word duration when switching from the casual to the clear speaking style with 64% increase for female speakers (20 speakers) and 73% for male speakers (20 speakers), with the effect of talker gender not being statistically significant. In this work, an acoustic analysis is performed on our subset. For the 8 speakers that form our subset, the duration difference is estimated between each clear sentence and its corresponding casual sentence. Then, the average difference duration of 144 sentence pairs is depicted on Figure 2.1. The standard deviation of the sentence duration per speaker is also illustrated. As we can see, speakers decrease their speaking rate to a greater extent when speaking clearly.

Examining further the decrease of speaking rate in clear speech, the contribution of pauses is estimated. For measuring the number of pauses and their duration, we have implemented an automatic pause and speech detection algorithm based on loudness criteria. Specifically, the pause detector relies on a low-loudness detection function based on the Perceptual Speech Quality measure (PSQ) described in *ITU Standard REC-BS.1387-1-2001*. First, the total loudness of the speech signal is computed by PSQ and then the normalized loudness is estimated, dividing by the maximum loudness of the signal. Then, a frame of the signal is considered not-speech, if its normalized loudness is less than 15%. After cross-validation (Stylianou et al., 2012) using a subset of files with manually-detected pauses (50 files from spontaneous speech of the LUCID database), it

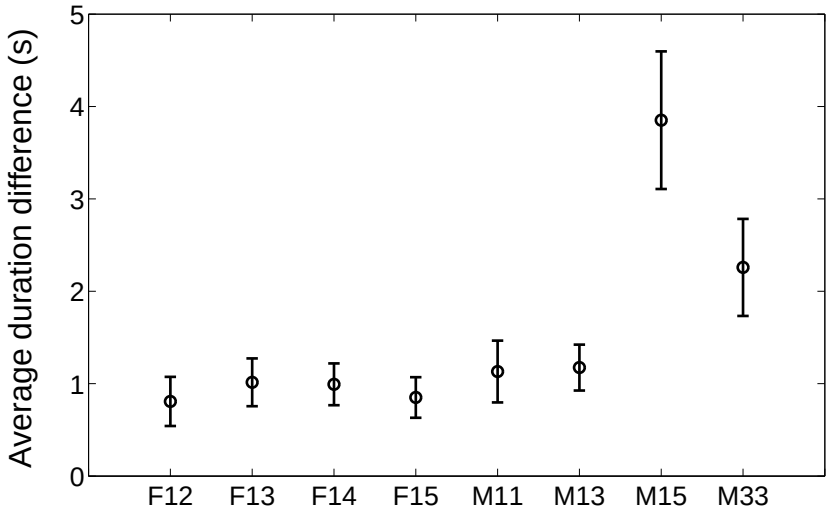


Figure 2.1: Average duration difference between the two speaking styles, clear and casual

was found consistent.

Table 2.1 presents differences on the number of pauses and on the pause and speech duration between clear and casual speech. The average number of pauses per sentence is computed as the ratio of the number of pauses detected on all sentences to the number of sentences. Mean pause duration is the ratio of the sum of all pause durations on the whole dataset to the number of pauses. Mean speech duration per sentence is the mean of the speech duration for all sentences. As Table 2.1 depicts, the number and duration of pauses as well as the mean speech duration is greater in clear than in casual speech. Moreover, the pause duration varies substantially in clear speech.

	Number of pauses per sentence		Mean Pause duration in msec		Mean Speech duration per sentence in sec	
	Female	Male	Female	Male	Female	Male
Clear	4.7	5.2	75.0 (59.2)	101.7 (83.1)	2.3 (0.3)	2.5 (0.5)
Casual	2.6	1.9	43.0 (28.8)	40.4 (28.3)	1.6 (0.3)	1.6 (0.2)

Table 2.1: Average number of pauses per sentence, mean pause duration (in msec), mean speech duration (in sec) per sentence in clear and casual style produced by 4 female and 4 male speakers. Standard deviations are given on parenthesis.

2.2 F0 distribution

Focusing on F0 differences between the two speaking styles, clear and casual, acoustic analysis is performed on our speech corpora. Specifically, 50 sentences out of 144 per speaking style are randomly selected for each speaker on our subset. The F0 values are extracted using SWIPEP (Camacho and Harris, 2008). SWIPEP estimates the pitch of the signal every 10ms within the value range 75-500 Hz. The spectrum is sampled uniformly in the ERB scale every 1/20 of ERB, using an overlap factor of 50%. The pitch is fine-tuned

using parabolic interpolation with a resolution of 1 cent samples. Pitch estimates with values lower than 0.2 are discarded. The histogram of the F0 values for all sentences of a male speaker is depicted in Figure 2.2(b) for clear speech and Figure 2.2(c) for casual speech. The density estimate is also calculated and depicted in Figure 2.2(a) to provide a smoother version of the histograms and a better illustration for comparing the two speaking styles. As we can see, clear speech has higher F0 average and range compared to casual speech for this male speaker (M33).

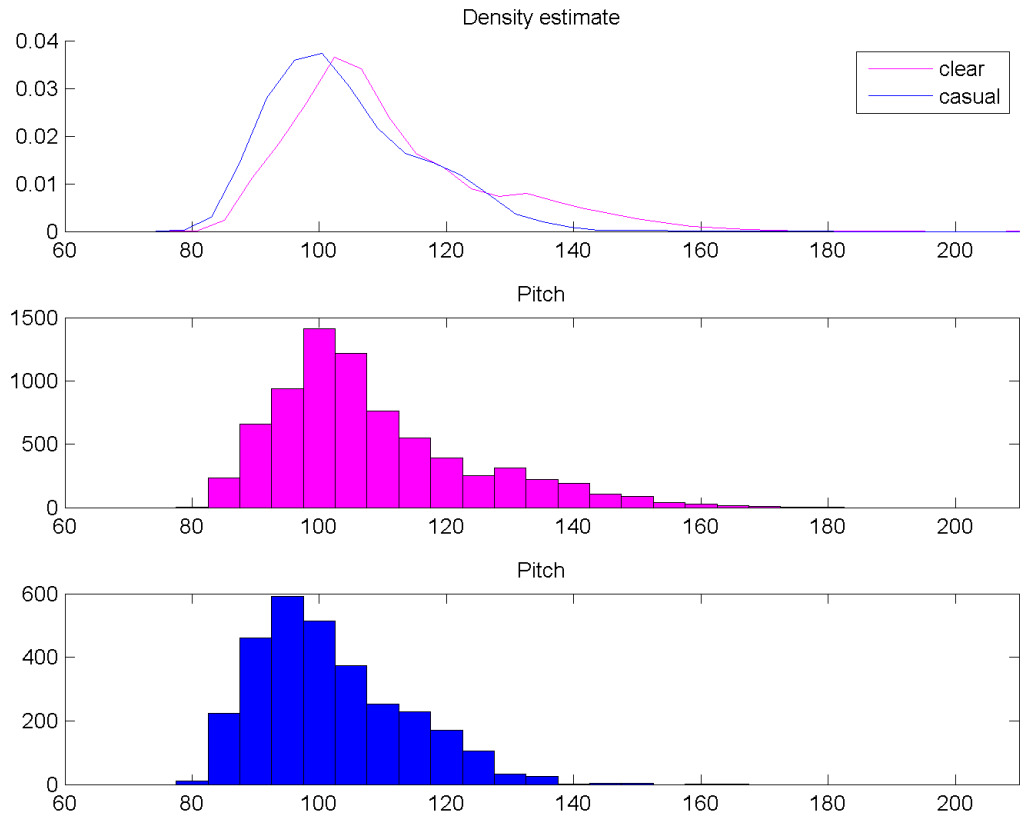


Figure 2.2: The distribution of F0 among the two speaking styles, clear and casual for M33 (a) Density estimates (b) Histogram of clear speech (c) Histogram of casual speech

Figure 2.3 shows the density estimates for each speaker on our subset. As we can see, the higher F0 mean and range observed in clear speech compared to casual speech is not consistent across speakers. For example, the similar F0 distributions of clear and casual speech for speakers M13 and M15 (Figure 2.3) show that these speakers do not modify their pitch when speaking clearly.

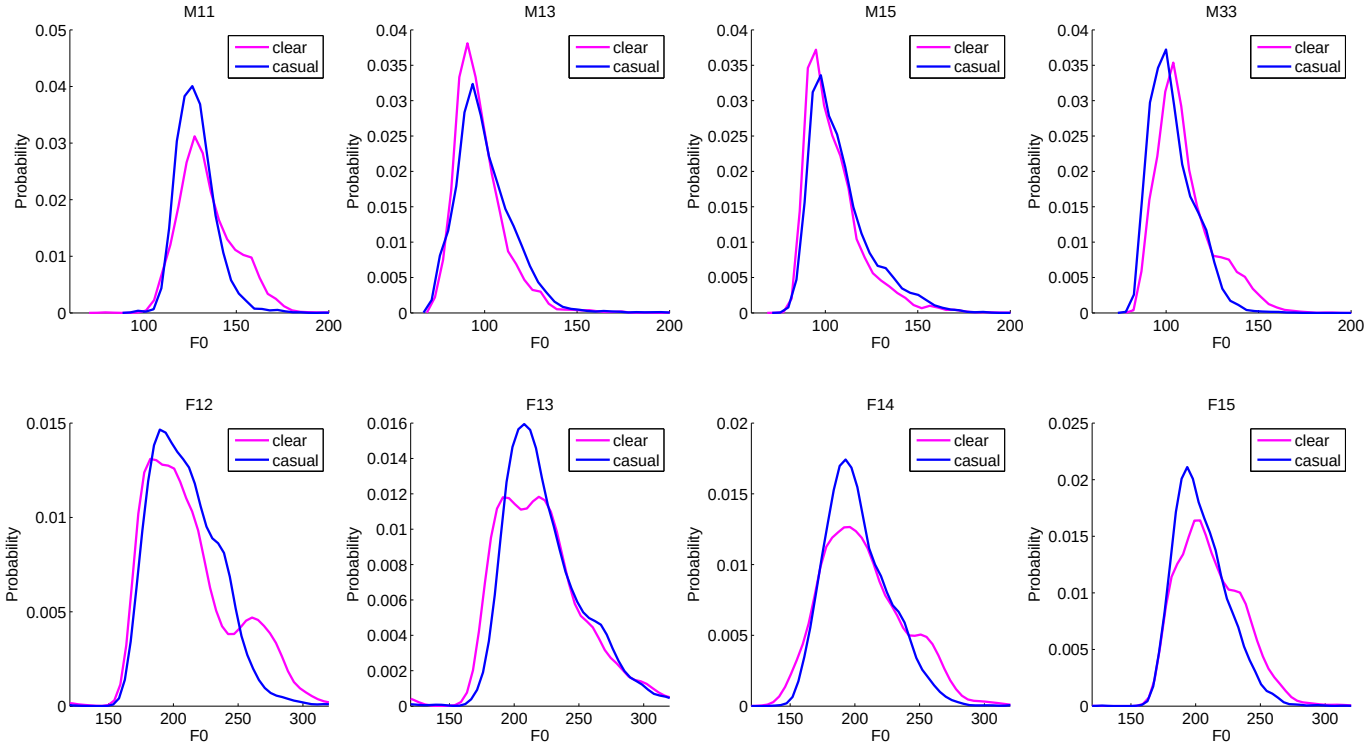


Figure 2.3: Pitch differences between the two speaking styles, clear and casual

2.3 Vowel spaces

Acoustic analysis of the vowel spaces is also performed on clear and casual speech. 50 sentences per speaker (40 speakers) and per speaking style are automatically segmented using an HTK-based audio-to-text aligner[1], without manual corrections. After segmentation, the vowel spaces are generated as follows. First, formant analysis is performed using Praat, which exploits the Burg algorithm (Boersma, 2001) to estimate the formant values of speech segments. The values at the center of the speech segment are the representative pair of F1 and F2 values for each vowel instance. Figure 2.4 depicts the F1-F2 values of all vowel instances of a female speaker (F12) classified into two vowel categories, lax and tense. Note that we used the IPA symbol system. However, there is some inconsistency of the symbols visualized in the graphs with the IPA system. Therefore, the ə, eə and ɔː IPA symbols are represented by @, e@ and o: respectively in the graphs. It can be observed that there is a high dispersion of the formant values in the vowel space for each vowel, especially for the lax category. This dispersion is more prominent in casual speech, whereas for clear speech there is a tendency even for the lax vowels to create clusters, even though these clusters overlap significantly. For the tense vowels, cluster formation is observed both for clear and casual speech with the overlap of clusters being less prominent in clear speech.

Expanding our analysis to all speakers of LUCID, the mean of each vowel instance is estimated per speaker

[1] many thanks to Paul Iverson, Mark Wibrow, Jos Joaqun Atria and Valerie Hazan for providing the authors with the HTK aligner.

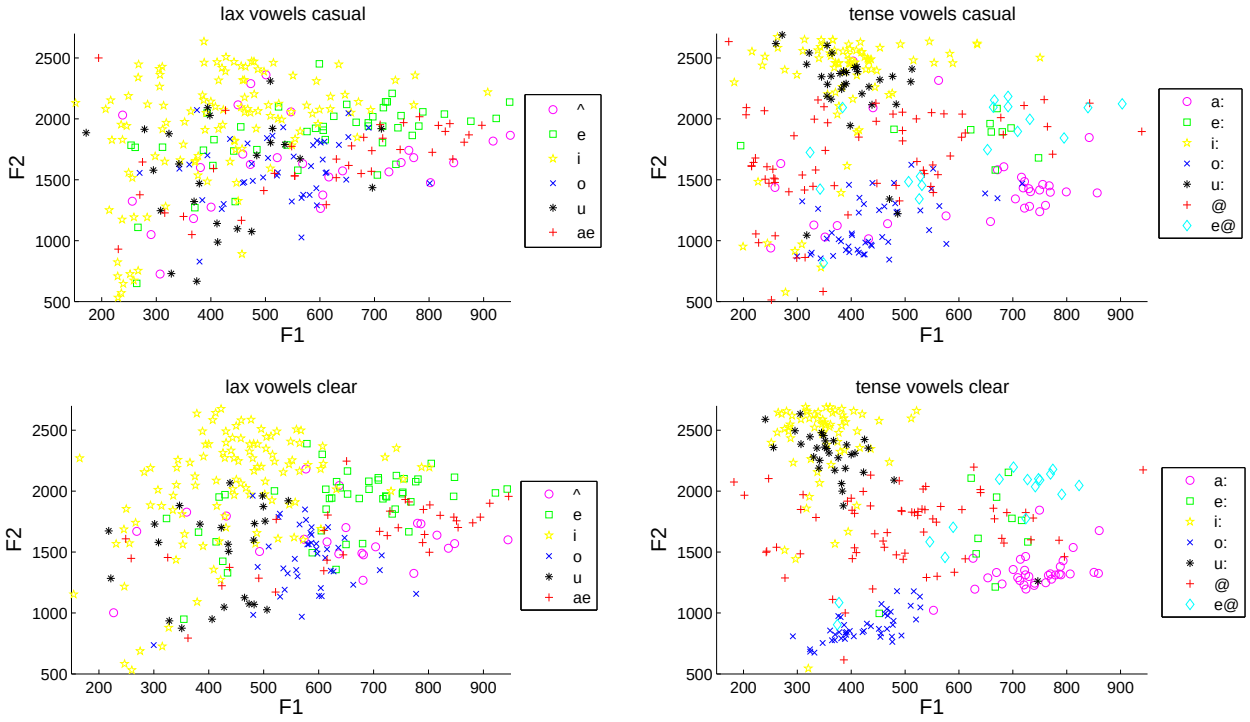


Figure 2.4: Casual and clear vowel space of lax and tense vowels for F12

and the vowel spaces are depicted per speaker category, namely female (Figure 2.5) and male (Figure 2.6) and for all speakers (Figure 2.7). Based on visual inspection of the graphs, two observations can be made. First, the vowel space seems to be more expanded in clear speech than in casual speech both for lax and tense vowels, for the latter though to a greater extent. Second, the male speakers appear to have more compact clusters than female speakers. However, this is possibly attributed to differences in the vocal tract length between male and female speakers.

In order to visualize and measure the vowel space differences between the two speaking styles, for each vowel (except for e: and e@ which were the least frequent in our dataset), the mean is calculated to create a single value in the graph. Specifically, for each vowel, the mean over all of the vowel instances is trimmed, with 95% of the data kept, in order to limit the influence of potential outliers. Then, the convex hull (i.e., a polygon fit that encompasses all of the data points) is used to represent the vowel space area, as it effectively captures the maximal area that the points in the vowel space span. Figure 2.9 shows the vowel spaces calculated using the aforementioned vowel instances for all speakers in our dataset, while Figure 2.8 shows the vowel spaces for each individual speaker. Additionally, Table 2.2 indicates the convex hull vowel space areas for the overall average plot in Figure 2.9 as well as the average of the male (M) and female (F) speakers for each style. It is apparent from Figure 2.9 and Table 2.2 that the clear speech vowel space area is expanded with respect to that of the casual. Moreover, in Figure 2.8, the vowel space expansion for clear speech is consistent across speakers despite the speaker variability in the degree of expansion, with the convex hull shape and orientation also remaining almost intact.

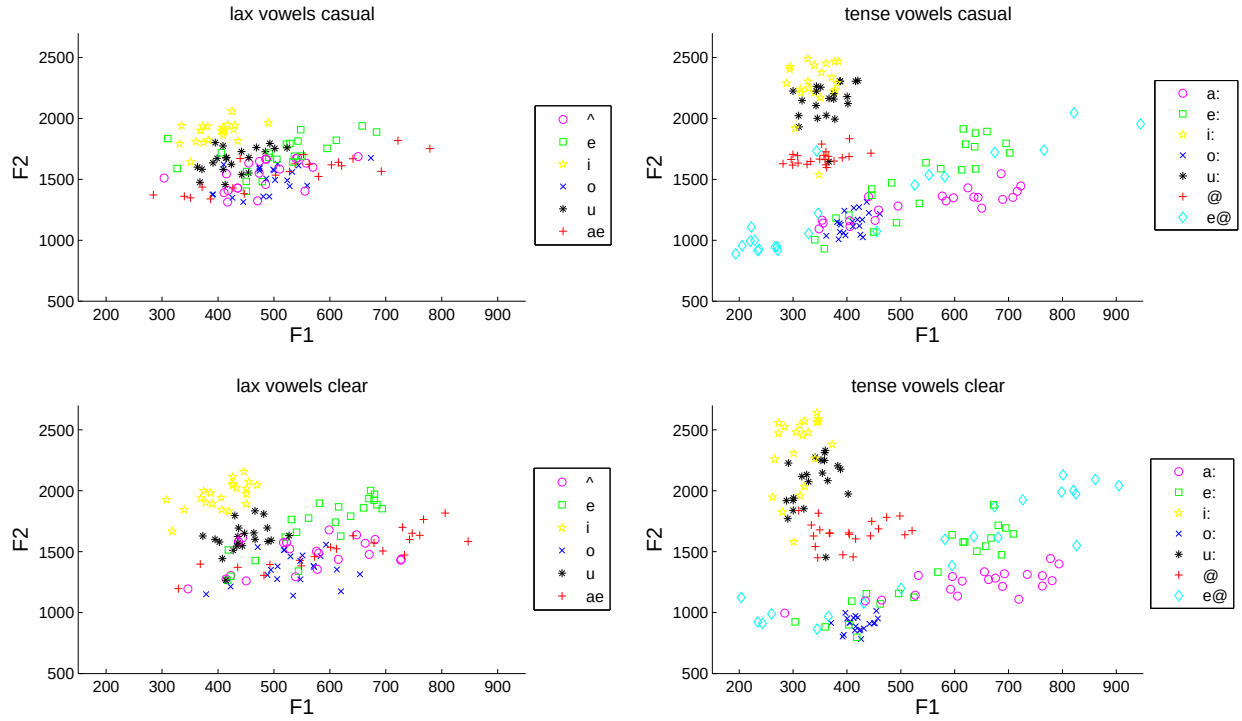


Figure 2.5: Casual and clear vowel space of lax and tense vowels for all female speakers in LUCID

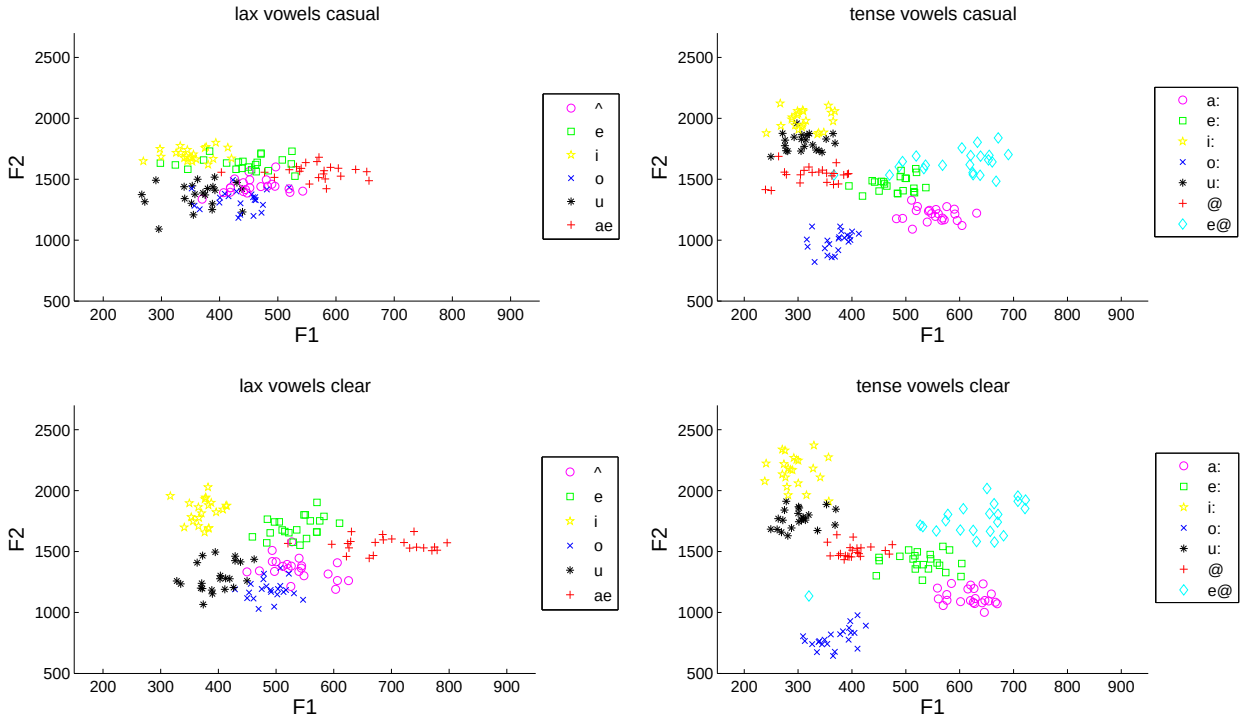


Figure 2.6: Casual and clear vowel space of lax and tense vowels for all male speakers in LUCID

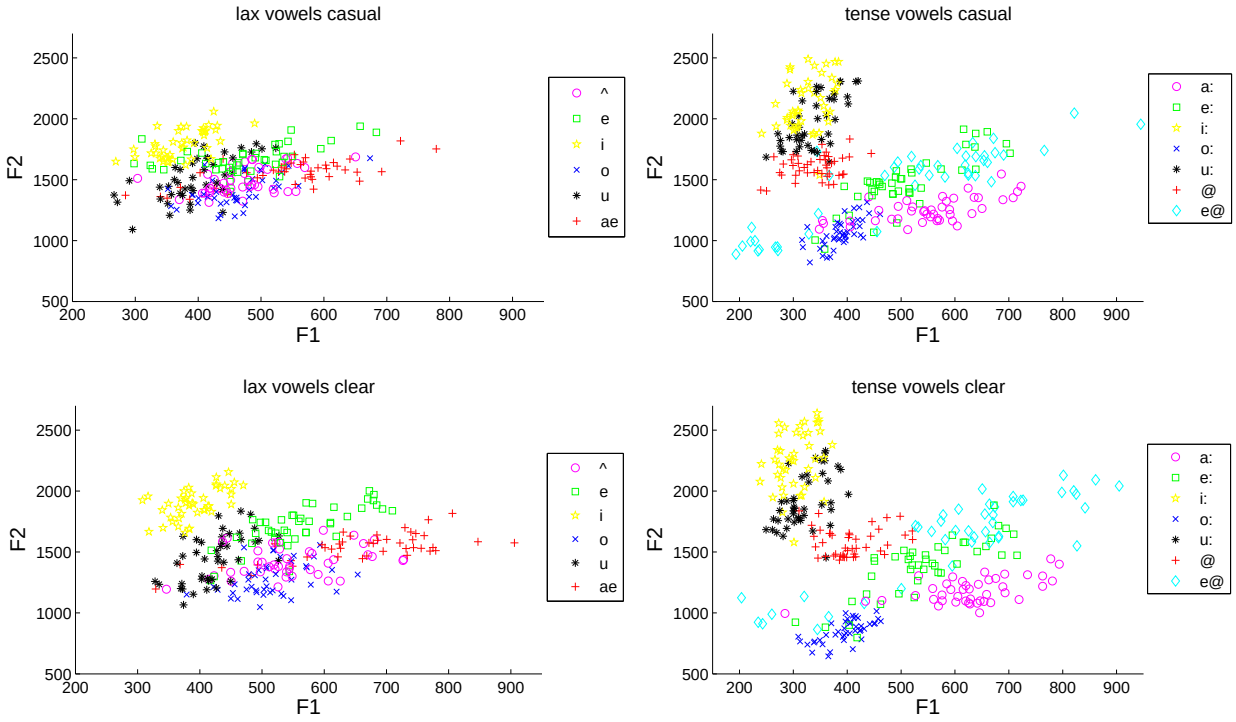


Figure 2.7: Casual and clear vowel space of lax and tense vowels for all speakers in LUCID

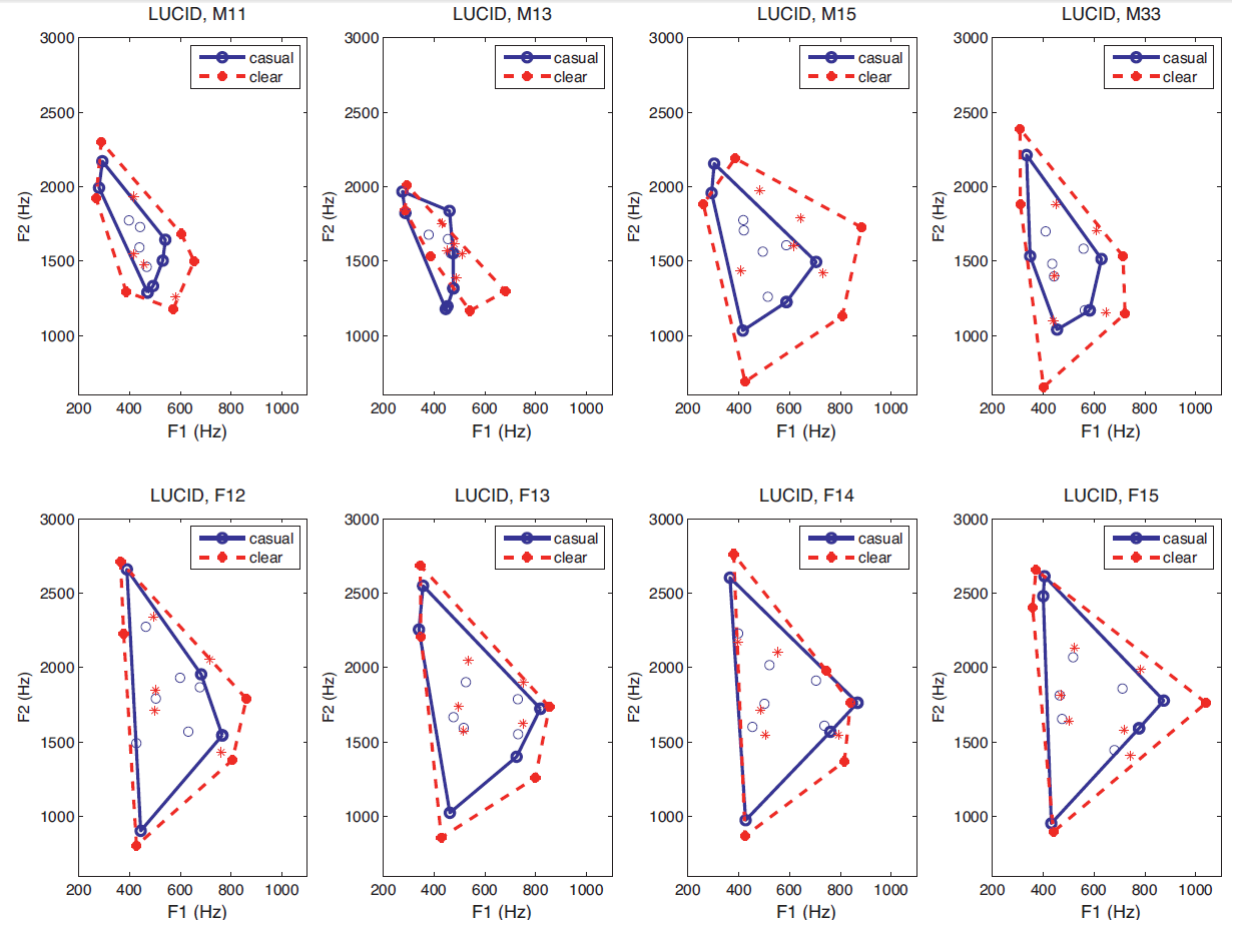


Figure 2.8: Convex hull of clear and casual vowel spaces for lax and tense vowels per speaker

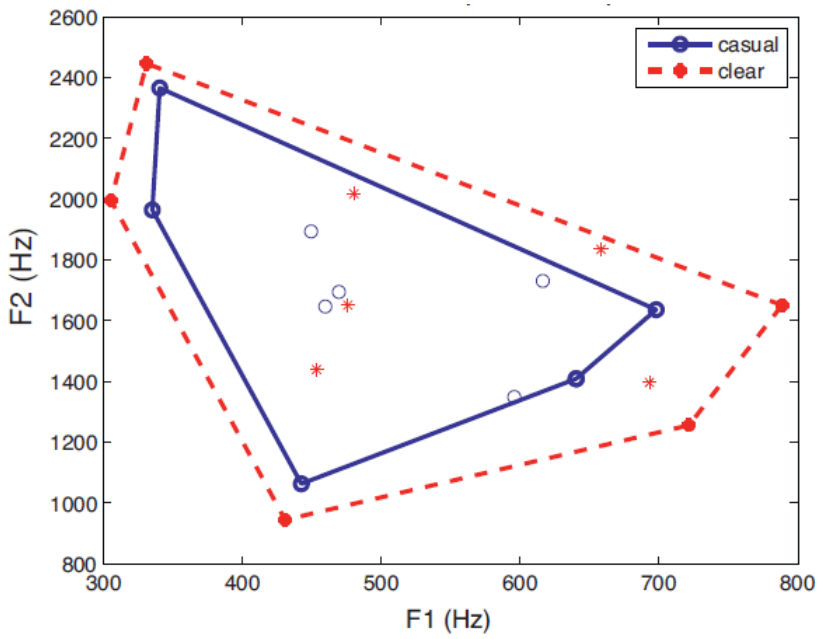


Figure 2.9: Convex hull of clear and casual vowel spaces for lax and tense vowels for all speakers of our dataset

	Clear	Casual	Difference (%)
ALL	3.93	2.32	69%
M	2.44	1.16	110%
F	5.15	3.58	44%

Table 2.2: Average vowel space area ($x10^5$ Hz^2) determined by the convex hull, given for all speakers as well as for male (M) and female (F) speakers separately. The percentage of the average expansion is also reported.

2.4 Spectral envelopes

Spectral energy enhancement in the 1-3kHz frequency region has been observed in clear compared to casual speech and was found statistical significant for the clear and casual speech corpora (Hazan and Baker, 2010). The following analysis explicitly examines the differences in spectral energy distributions of clear speech with respect to its casual counterpart, specifically via the average relative amplitude spectra (Krause and Braida, 2004b; Godoy and Stylianou, 2012) aiming on incorporating these measured differences on modifications of casual speech. First, all sentence pairs (clear-casual) are normalized to have the same Root Mean Square (RMS) and downsampled to 16kHz. Frame-by-frame estimation of the true envelope as proposed by Imai (1983); Roebel et al. (2007) is performed for the voiced segments. For unvoiced segments, spectral envelope estimation directly from the LPC analysis is employed. The true envelope estimation is based on cepstral smoothing of the amplitude spectrum. The cepstrum order is set to 15 in order to estimate an overall energy of the frequency bands and avoid energy canceling due to different formant positions of each speaker. Again, for each spectral envelope, the DC component is set to zero. Then, the spectral envelope is normalized by its RMS to eliminate intensity differences between clear and casual speech. The "relative spectrum" for each speaker is then defined as the log-difference between the average clear and average casual spectral envelopes, calculated using all frames. As in Godoy and Stylianou (2012), the DC component of this difference is removed in order to avoid a constant bias related to frame energy. This DC component would simply shift the relative spectrum up or down by a constant amount, without altering the overall curve shape. The average relative spectra for the clear-casual corpora are shown in Figure 2.10 for individual speakers.

Clear-casual relative spectra show a significant variation across speakers indicating that each speaker employs a different strategy to produce clear speech. For example, speaker M15 shows an increased energy near 1000Hz while for M11 an energy enhancement above 3000Hz is observed. Possibly, M11 employs a consonant emphasis technique to produce clear speech.

Figure 2.11 shows the average relative spectra of all individuals depicted in Figure 2.10. The averaged spectral envelopes of clear and casual speech are computed as the mean of all frames for each speaking style separately. Figure 2.11 shows the outcome that derives from the subtraction of the log average spectral envelopes of casual speech from that of clear speech. A limited number of sentences (20 sentences per speaking style) is used to derive the average relative spectra since it will serve as a training dataset (see Chapter 5). Positive difference suggests that the energy of clear speech is higher than that of the casual. As we can see, clear speech appears to have higher energy in two frequency bands, $B1 = [2000, 4800]$ and $B2 = [5600, 8000]$.

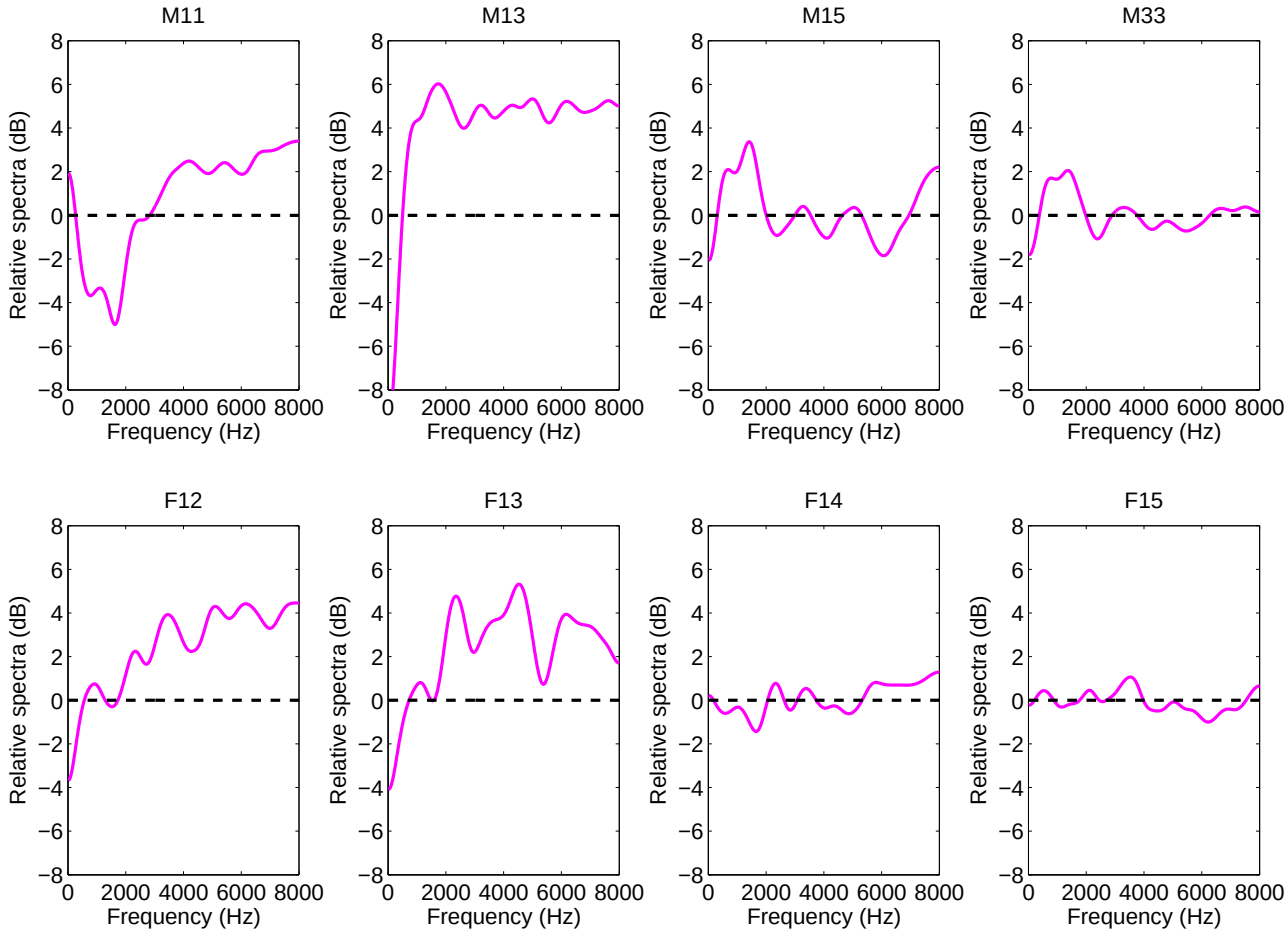


Figure 2.10: Relative spectra for each of the 8 speakers on LUCID dataset

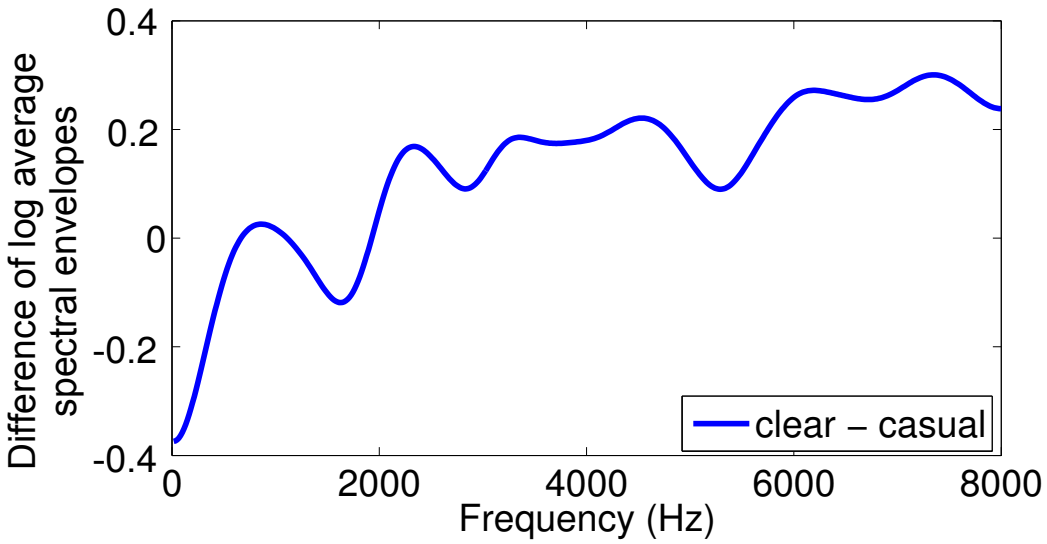


Figure 2.11: Difference of the log average spectral envelopes of clear speech minus casual speech for the 8 speakers on the LUCID dataset

2.5 Connection of observed differences with speaker comprehensibility

In the previous sections we have seen the acoustic differences between read clear and read casual speech, focusing on a subset of speakers from the LUCID database, F12, F13, F14, F15 (female) and M11, M13, M15, M33 (male). A ranking of speakers according to their comprehensibility is reported in this section in order to have a measure of how effective is their speaking adjustment. However, we have not conducted intelligibility listening tests on read clear and read casual speech to rate the speakers in terms of intelligibility. The comprehensibility ranking of speakers is obtained through subjective evaluations performed in a different task, described in Baker and Hazan (2010). This task includes spontaneous speech recordings rather than read sentences uttered by the specific speakers. Specifically, the speakers of LUCID database have participated in a dialogue condition, namely the diapixUK task described in Baker and Hazan (2010). It was an interactive "spot the difference" game for two people that allows for recordings of natural spontaneous speech. The task was evaluated in two conditions. In the first condition there was no noise (no barrier, NB) on the listener's side. In the second condition, the speech elicited by the talker was spectrally degraded before presented to the listener (VOC condition). The speakers in VOC condition had to increase their clarity so that the listener could accomplish successfully the diapix task. Figure 2.12 depicts the subjective ranking of speakers according to the listeners' comprehension when switching from the NB to the VOC condition [2]. Even those these measurements have not been estimated on our database, the ratings among speakers should give an indication of their relative clarity and of the degree to which they modify their speech in the presence of a communication barrier.

Comparing the comprehensibility among speakers (Figure 2.12), M15 and F15 seem to produce clearer speech compared to other speakers. These speakers modify their speech to a great extent from NB to VOC condition. Examining which of the above features, namely {speaking rate, pitch, vowel space, spectral envelopes}, change the most for these two speakers, vowel spaces show the greatest difference between the two speaking styles for these speakers (Figure 2.8). M15 produces also the slowest speech compared to other speakers, whereas F15 exhibits small modifications on the speaking rate (Figure 2.1). Moreover, in comparing Figure 2.10 and Figure 2.3 with the subjective ratings, differences in pitch and spectral energy do not appear to be correlated with speaker comprehensibility.

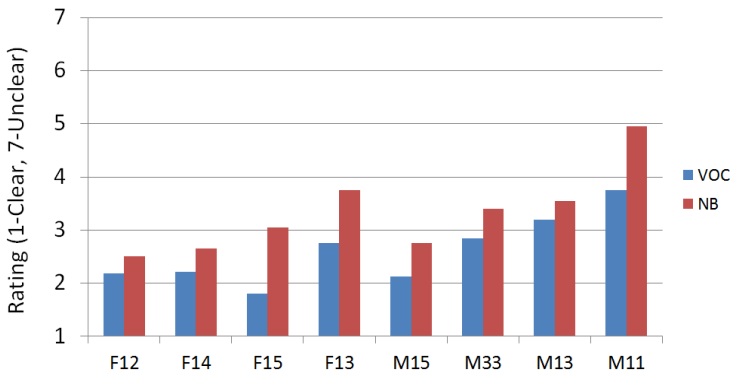


Figure 2.12: Intelligibility scores from NB to VOC condition per speaker

[2] The data of the graph has been provided to us by the authors in (Hazan and Baker, 2010)

2.6 Discussion

In this Chapter, we presented a the comparative analysis performed between clear and casual speech which focuses on the most prominent features that differ between the two speaking styles. First, by examining differences on the speaking rate between clear and casual speech, we found that clear speech exhibits a 42-60% increase of the mean pause duration compared to casual speech and roughly the same percentage of increase in the average number of pauses, with female speakers being more conservative than male speakers to changes between styles. The mean speech duration also increases, although to a less extent (nearly 30%), when switching from casual to clear speech. Second, exploring changes in F0 distribution revealed that clear speech has higher F0 range than casual speech. However, this is not consistent in all speakers. On the other hand, vowel space expansion in clear speech is a prominent feature difference to all speakers. Indeed, all speakers that produce clear speech have more expanded vowel spaces compared to their casual counterpart. Moreover, speakers with the highest intelligibility benefit also have the more expanded vowel spaces compared to other speakers. This suggests a possible correlation of the vowel space expansion with intelligibility. Finally, differences in the relative spectra is also presented between clear and casual speech. However, unlike other studies, this work reveals two important frequency regions where clear speech is more enhanced than casual speech. The analyses performed in this Chapter are incorporated in the sections that follow to our proposed modifications for intelligibility enhancement.

Chapter 3

Prosody Transformations

In the previous chapter we observed differences between the two speaking styles in pauses and elongation of speech segments and in the distribution of F0. However, these two features may not be connected with the intelligibility advantage of clear speech. Previous studies suggest that clear speech can also be produced without decreasing the speech rate, after training the speakers (Krause and Braida, 2004a). Moreover, our analysis on the LUCID database has revealed that natural increase of F0 average and range in clear speech is not displayed for all speakers. Therefore, in this chapter we examine whether or not the speaking rate and pitch contribute to the intelligibility benefit of clear speech. To achieve this, we exclude these acoustic features from clear speech by modifying clear signals to match the duration and pitch of the casual signals. Then, changes in the intelligibility are measured with objective and subjective listening tests.

The above method suggests that the speaking rate is indeed important for intelligibility. Therefore, less intrusive time-scaling approaches than segmental time-scaling are explored in Section 3.2 for enhancing casual speech intelligibility.

3.1 Studying the effect of speaking rate and pitch to intelligibility

A simple method is implemented in order to explore whether or not duration and pitch contribute to the intelligibility advantage of clear speech. The general concept of our method is to modify these acoustic attributes from the one speaking style towards the other and evaluate the changes on intelligibility subjectively and objectively. This is implemented in two sequential experiments. In the first experiment, we attempt to explore whether and to what degree pitch and duration contribute to the intelligibility benefit of clear speech. Therefore, we modify clear speech to match the corresponding acoustic properties of casual speech and we measure the intelligibility degradation from these modifications. In the second experiment, we incorporate the acoustic properties of clear speech to casual speech aiming on enhancing the intelligibility of casual speech. The second experiment evaluates only duration transformations from casual to clear as dictated by the outcome of the first experiment.

Both experiments share the same speech corpus. The speech corpus selected for evaluation is a random

subset from the LUCID database. Specifically, 69 distinct pairs of clear and casual sentences are selected, uttered by the same speakers. Before modified and presented for evaluation, the dataset is preprocessed. The preprocessing involves downsampling to 16 kHz and high-pass filtering the signals in order to remove low-pass noise introduced by breath and lip effects. The filter is a 5-order high pass digital elliptic filter with 80Hz cut-off frequency. For the duration modifications, segmental time-scaling is performed. Therefore, time alignment values are estimated by hand at a segmental level. Then, the time-alignment information feeds the Waveform Similarity based Overlap-Add algorithm (WSOLA-Demol et al. (2004)) that modifies the duration of the one speaking style to match the duration of the other.

3.1.1 Experiment I

The methodology of the first experiment is depicted in Figure 3.1. In this experiment, WSOLA time-scales clear sentences to match the duration of casual sentences. Then, the time-compressed clear signal is modified in pitch to match the corresponding F0 values of casual, using the following pitch equalization function:

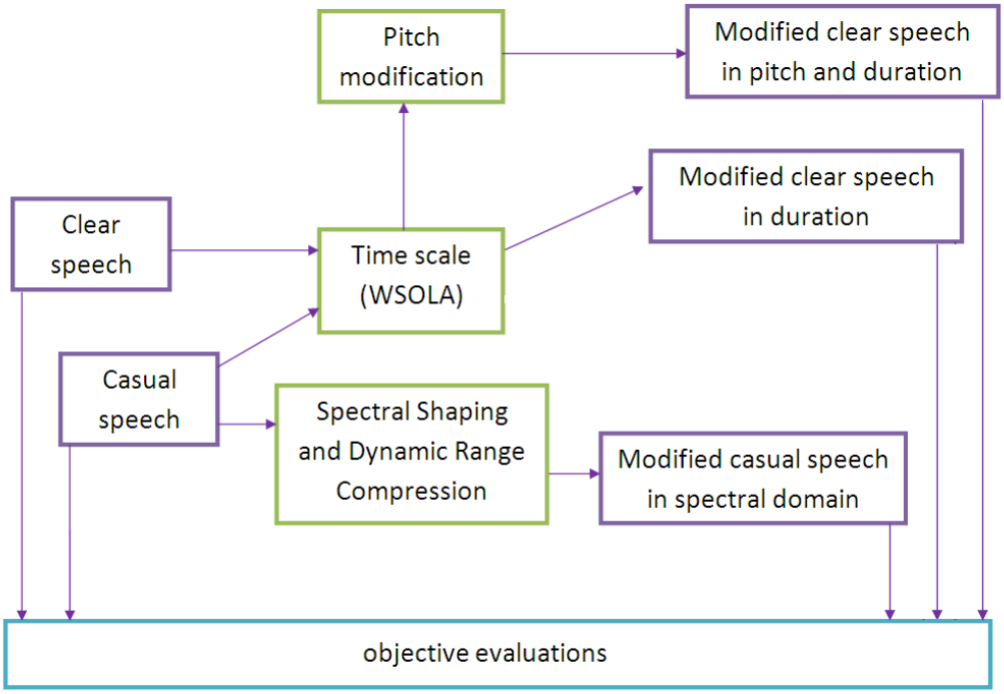


Figure 3.1: Defining the impact of duration and pitch to intelligibility

$$F0_{clear}^{new} = \frac{F0_{clear} - \overline{F0}_{clear}}{\sigma_{F0_{clear}}} \sigma_{F0_{casual}} + \overline{F0}_{casual} \tag{3.1}$$

where $\overline{F0}_{\{\cdot\}}$ and $\sigma_{\{\cdot\}}$ are the mean and standard deviation of the F0, respectively. The pitch modification is performed by PSOLA (Charpentier and Stella, 1986). Figure 3.2 shows the F0 track (Drugman and Alwan, 2011) estimated on clear and casual speech for the M35 speaker, whose F0 distributions differ significantly between the two speaking styles (see Chapter 2). The F0 of the pitch-scaled clear signal is also depicted,

showing a reduced F0 range compared to unprocessed clear speech and a similar average value and range to that of casual speech.

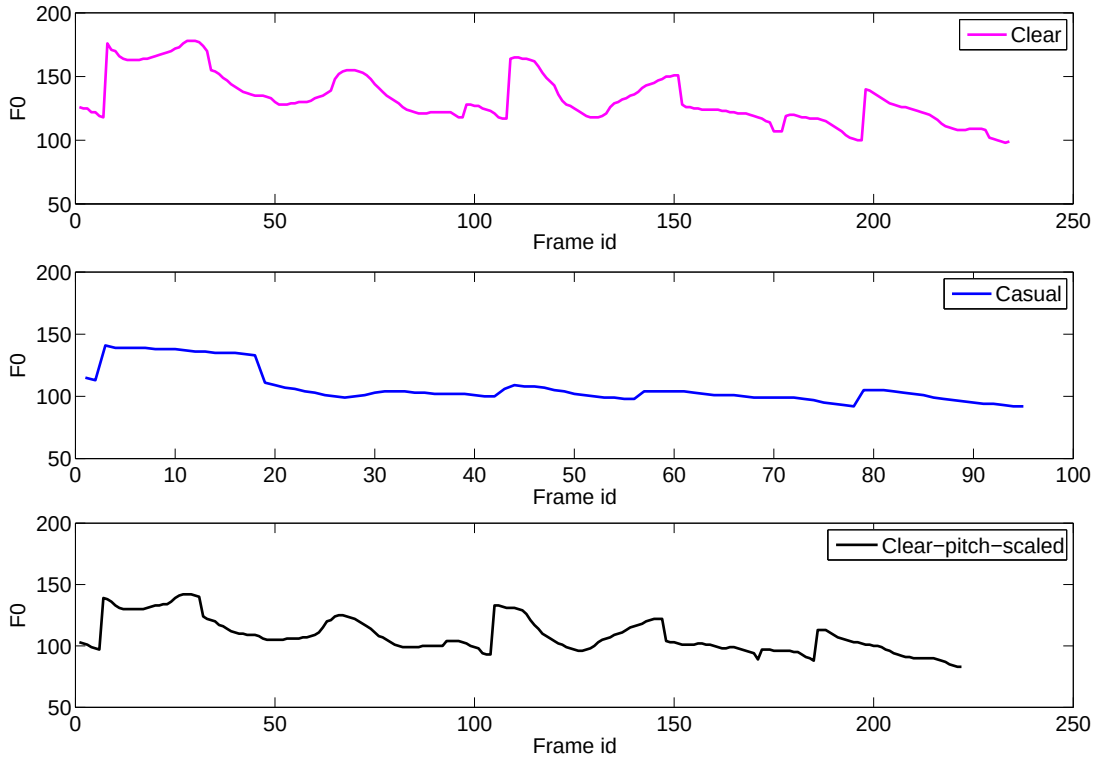
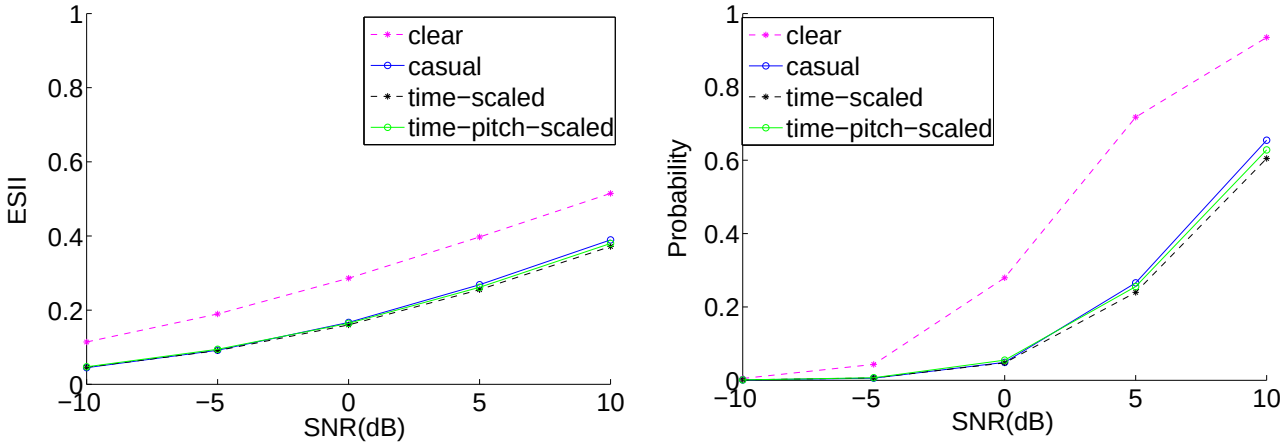


Figure 3.2: Fundamental frequency (F0) of a clear sentence and its corresponding casual sentence uttered by M35 speaker. The clear sentence is modified in pitch using equation (3.1) in order to approach the F0 average value and range of the casual sentence. Non-zero values of F0 in the sentence are not depicted.

The modifications are objectively evaluated in the presence of Speech Shaped Noise (SSN) using a modified version of the extended Speech Intelligibility Index (extSII), the ESII. For the computation of extSII we followed the steps described in Rherbergen and Versfeld (2005). First, an FIR Filter Bank is used to filter speech and noise signals into 21 critical bands (Fraser, 1999). Each filter in the filter bank is a linear FIR filter of type I and order 200. Next, the time varying intensity of the signal is computed for each output of the filter bank. For this, non overlapped rectangular windows are used with window lengths ranging from 35 ms at the lowest band (center frequency 50 Hz) to 10 ms at the highest band (center frequency 7000 Hz). The windows are aligned such that they ended simultaneously (Rherbergen and Versfeld, 2005). The intensity level is normalized to dBSPL. More specific, the Root-Mean-Square energy for each windowed frame is divided by the absolute threshold of hearing (10^{-12} Watts) and then this intensity value is estimated in dB. At a given instant, the instantaneous extSII is computed following a standard procedure (ANSI-S3.5-1997, 1997) using the so-called speech perception in noise (SPIN) weighting function and the estimated speech and noise normalized intensities. Finally, the extSII for a speech-in-noise condition is determined by simple averaging across all the instantaneous extSII values. ESII differs from extSII in that ESII considers also the duration of the signal as an intelligibility factor. The objective intelligibility score ESII was successfully validated subjectively in the

work of Valentini-Botinhao et al. (2011).



(a) (b)

Figure 3.3: Objective Measure Score for the four sets of signals for different levels of SNR: (a) Speech Intelligibility Index (b) Probability of correctly identifying a sentence

Figure 3.3(a) shows the objective intelligibility score ESII of the casual, the clear, the modified clear speech in duration, and the modified clear speech in duration and pitch towards that of casual speech in SSN of different SNR levels. Figure 3.3(b) depicts the probability of correctly identifying a sentence for various SNR levels. Therefore, the probability of identifying correctly a sentence uttered in clear style on 5dB SNR is 60% with an intelligibility score near 40%. Figure 3.3 reports that time-compressed clear speech has lower intelligibility than unprocessed clear speech. Moreover, it suggests that pitch modifications do not affect the intelligibility of clear speech, since the degradation of intelligibility derives from the time-compression scheme. This is re-enforced by pilot listening tests (with a small number of non-native listeners on 0dB SNR in SSN) that verify that pitch modifications do not seem to contribute to speech intelligibility. These results suggest that the intelligibility benefit of clear speech is possibly attributed to its lower speaking rate. Therefore, the second experiment focuses only on duration modifications between the two speaking styles and performs both objective and subjective evaluations.

3.1.2 Experiment II

The methodology followed by the second experiment is depicted in Figure 3.4. In this experiment, casual speech is time-expanded using segmental time-scaling and is compared with unprocessed clear and casual speech and with time-compressed clear speech. SSDRC modified casual speech is also presented since it serves as an upper bound for intelligibility in most of our evaluations. These modifications are evaluated subjectively by native and non-native listeners and objectively in presence of SSN noise. Evaluations are analytically described below.

Subjective evaluations

Subjective evaluations include perceptual tests performed by native and non-native listeners on SSN noise.

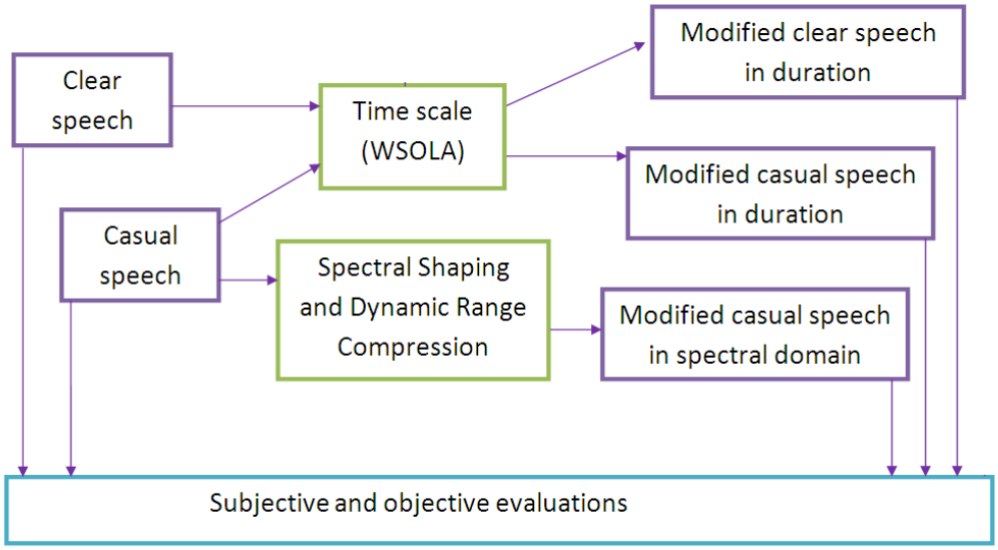


Figure 3.4: Defining the intelligibility impact of segmental time-scaling modifications on casual speech

Specifically, SSN is added to the five set of signals, {clear, casual, SSDRC, casual-time-expanded, clear-time-compressed}, to create the test signals. Different levels of SNR are explored, namely $\{-3, 0, 5\}$ dB. Therefore, for the five set of signals and for the 3 different SNRs, a dataset of 5x3 test sentences is created. From this dataset, each listener randomly hears signals with the limitation of hearing each sentence only once. Then, the listener evaluates a sentence based on the description provided in Table 3.1. The listening test is performed by 24 native English speakers and 15 non-native speakers. For each set, the average score value across participants is calculated. Then, the average score values are normalized by the maximum theoretical score (5) and are depicted in Figure 3.5(a) and Figure 3.5(b) for the native and non-native population, respectively. The subjective scores are normalized.

Score	Description
5	if you understood the whole sentence
4	if you understood the sentence except one or two words
3	if you could barely understand the sentence
2	if you could understand some words but not the message
1	if you could not understand anything at all

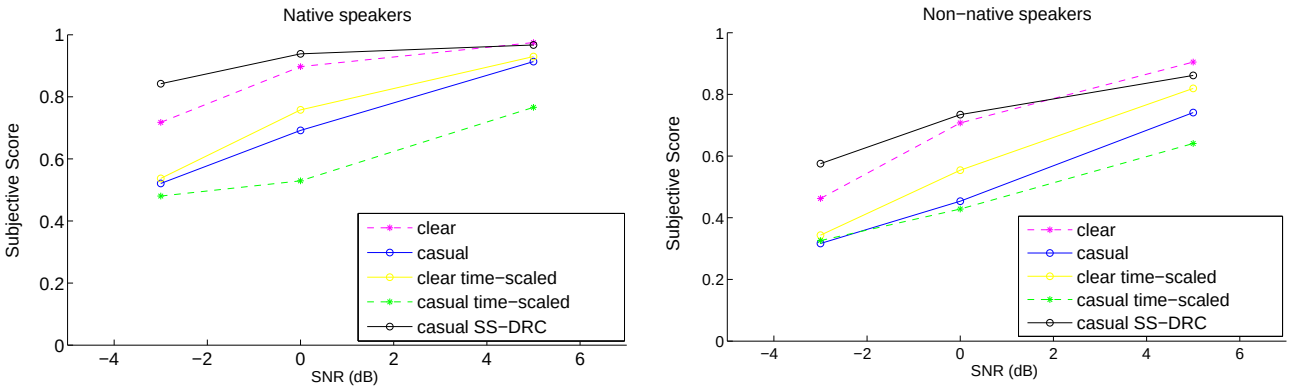
Table 3.1: Subjective Scores Description

Evaluation results show that in the presence of relatively low noise ($SNR = 5$dB), clear speech is more intelligible than casual speech both for native (7%) and for non-native speakers (17%). Comparing native and non-native speakers, the intelligibility advantage of clear speech is much higher for the non-native population. Time-compression of clear speech reduces its intelligibility around 5% for native speakers and 8% for non-native speakers in high SNR condition ($SNR = 5$dB). However, in more adverse listening conditions ($SNR = -3$dB), making clear speech faster reduces its intelligibility to 18% for native speakers and 11% to non-native speakers. This suggests that speaking rate plays a significant role in the intelligibility of clear

speech.

However, the intelligibility benefit of clear speech is possibly attributed to other acoustic properties besides speaking rate. This is suggested by the fact that clear speech maintains a part of its intelligibility after time-compression compared to casual speech especially for non-native listeners (Figure 3.5(b)). This assumption is also enforced by the fact that time-expansion does not help casual speech to increase its intelligibility. On the contrary, time-expanded casual speech gives lower intelligibility scores than the unprocessed casual speech, with a more negative impact to the native speakers. Native speakers actually reported that the expanded in duration casual signals sounded irritating.

Transforming casual speech in spectral domain significantly raises intelligibility for all SNR levels, both for native and non-native listeners (Figure 3.5). Native speakers reported a 32% increase of intelligibility after SSDRC modifications on casual speech, while non-native speakers reported a 27% of raise for low SNR (-3dB). For the same condition the transformed SSDRC casual speech is 11% more intelligible than clear speech, as Figure 3.5 reports, while for higher SNR values (5dB), SSDRC-modified casual speech and clear speech share the same intelligibility score.



(a) (b)

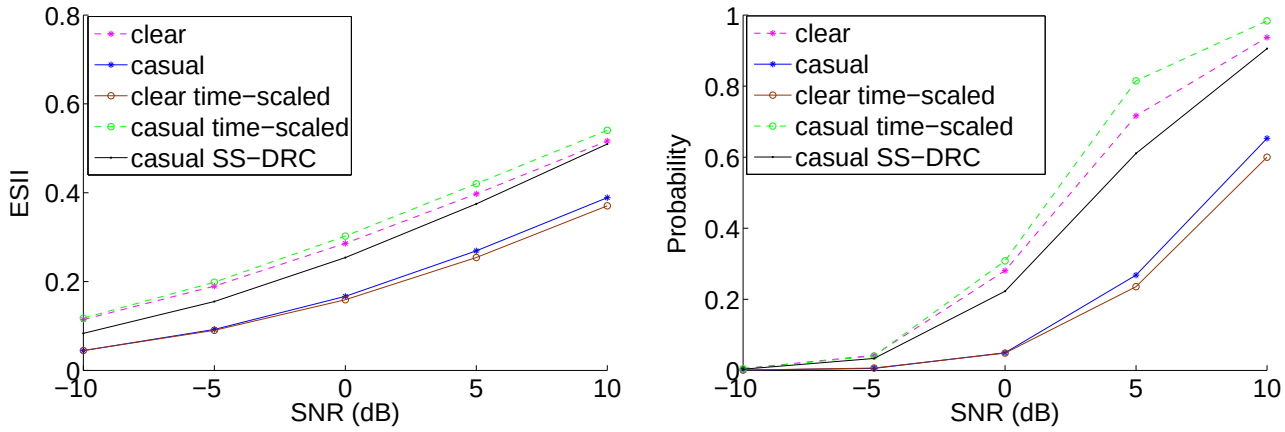
Figure 3.5: Subjective Measure Score for the 5 set of signals for different levels of SNR. a) Native Speakers b) Non-native speakers

Objective evaluations

Objective measure tests based on the ESII were also performed in this experiment. Figure 3.6 depicts the ESII intelligibility scores of the five set of signals, namely {casual, clear, clear time-compressed, casual time-expanded, SSDRC modified casual} in SSN for various SNR levels {-10, -5, 0, 5, 10}. According to the ESII measure, clear speech and SSDRC modified speech have higher intelligibility scores than casual speech (Figure 3.6(a)) with higher probability (Figure 3.6(b)) of correctly identifying a sentence for SNR levels above -5dB. On the other hand, casual speech and time-compressed clear speech that have the same duration, give the same score of ESII independent of the SNR level (Figure 3.6(a)). This is consistent with subjective evaluations for native listeners, despite differences in the score values. In general, ESII scores are much lower than subjective scores reported by the listeners. A possible reason is that the linguistic content helps the listeners to guess the meaning of the sentence. Furthermore, the subjective scores provided by the listeners reflect more

than intelligibility, the listening effort of the speaker to comprehend the message, since speakers scored what they think they heard rather than writing down the sentences. An intelligibility test would probably lower the overall scores but it is not expected to change the relative scores between modifications and speaking styles.

For the time-expanded speech, objective measures give contradictory results with the subjective scores. This is due to the fact that the ESII takes into consideration the duration of the signal that is examined compared to the reference signal. The reference signal is the casual signal whose duration is much smaller. Increasing the duration of casual speech contributes positively to the intelligibility score of ESII.



(a) (b)

Figure 3.6: Objective Measure Score for the five sets of signals for different levels of SNR: (a) Extended Speech Intelligibility Index (b) Probability of correctly identifying a sentence

3.1.3 Discussion

Differences obtained from the acoustical analysis between the two speaking styles show a decrease of the speaking rate and a higher intelligibility advantage of clear speech compared to casual speech. Since clear speech exhibits a sizable intelligibility advantage over casual speech, it can be assumed that at least some of the essential acoustical characteristics of clear speech are preserved after its time-compression. Indeed, excluding the duration factor from clear speech, its intelligibility decreases but does not reach the intelligibility levels of casual speech. This suggests that duration is an important intelligibility feature but not the only contributing factor to the intelligibility of clear speech. Clear speech is enriched with other acoustic-level information that casual speech does not have (i.e pauses). Possibly this is the reason why the time-expansion of casual speech failed to increase intelligibility since it cannot fill the gap of this missing phonetic-level and acoustic-level information.

This study motivates us to explore two different approaches for enhancing speech intelligibility. On the one hand, duration modifications seem to be important for intelligibility. On the other hand, segmental time-scaling has been proven harmful for intelligibility. Possibly, segmental time-scaling is a rather intrusive way to time-scale casual speech, since it is based on the acoustic properties of clear speech, disregarding the acoustics of casual speech. Moreover, the absence of pauses leads to extreme elongation of casual speech segments,

degrading speech quality and intelligibility. Therefore, different time-scaling modifications are explored in the section that follows, considering the acoustic properties of casual speech and differences in the pause distribution between the two speaking styles. Furthermore, SSDRC transformations of casual speech increase its intelligibility near the levels of clear speech, suggesting that spectral transformations are advantageous for enhancing speech intelligibility. It is therefore, worth exploring clear-speech inspired modifications and also examine whether or not SSDRC shares similar acoustic features with clear speech.

Last but not least, one could comment on the subjective and objective evaluations. First, the contradictory results between subjective and objective scores, obtained from the above experiments, reveal the necessity of performing subjective intelligibility tests in order to evaluate the efficiency of our modifications. Last, in this section the subjective listening tests cannot be characterized as intelligibility tests since this would require asking the listeners to write down what they understood rather than asking them if they understood. The subjective listening tests conducted measure the listening effort rather than the intelligibility. However, the listening effort and intelligibility are correlated as suggested by studies (Rennies et al., 2014; Brons et al., 2014). Moreover, the classification of modifications proposed by this test is in accordance with the intelligibility classification of speaking styles and modifications provided in Chapter 5 (SSDRC and clear speech have higher subjective scores than casual speech both in this subjective test and in the intelligibility test presented in Chapter 5)and therefore is considered trustworthy.

3.2 Exploring Clear-inspired time-scaling modification techniques

The intelligibility advantage of clear speech has proved itself to be linked to a reduction in the speaking rate. However, this reduction is associated with the hyper-articulation of the style, involving numerous cues that carry multi-level (e.g. linguistic, perceptual, phonetic, acoustic) significance. For example, to illustrate the multi-level significance of the pronunciation of one word, consider the word "insert". The word insert can be either pronounced as [ˈɪn.sɜːt] or [ɪnˈsɜː t] depending if the word is a noun or a verb. Therefore, these words differ both linguistically (the two words have different meaning and stress) and acoustically. Since the placement of the intonation differs, there must be also a pitch change on the two segments of speech in different placements for the first and the second word. However, when stressing a syllable a durational change often accompanies a pitch change (Ladefoged and Johnson, 2010). Therefore, one expects that an elongation might happen where the pitch is higher. This simple example shows the interaction of the different levels of speech for the pronunciation of one word.

When comparing casual and clear speech, this analysis becomes even more complex. For example, a "sloppy" enunciation of the word "insert" differs from a hyper-articulation of the same word: the "r" could be pronounced distinctly or only present in r-coloring of the "e" vowel, the length of the "i" can vary, the absence of the burst "t" is more likely to happen on the sloppy enunciation, etc. These differences can be translated, on a general level, into variations in burst frequencies, vowel spaces, vowel durations, pitch variations, etc. Therefore, there are several "events" on multiple levels that take place when speech is slowing down. On

the linguistic level of analysis, deletion of phonemes and co-articulation are more likely to happen for casual speech, making the speech faster (Krause and Braida, 2004a). On an acoustic-phonetic level of analysis, differences between clear and casual speech can be seen in the formant movements (Rossing, 2007), consonant aspiration etc. On purely acoustic level, the reduced speaking rate can be described in terms of the existence of more pauses and elongations compared to normal speaking rates.

The problem, therefore, of increasing the duration of casual speech is rather complex, considering that the reduced speaking rate is accompanied by modifications on three levels of analysis. One way of approaching this complex problem is to focus only on one level (e.g., acoustic) and a few of its corresponding isolated features (e.g., phoneme duration, pauses). Indeed, this approach has been adopted by previous works and is also introduced in this study.

Focusing only on the acoustic level, clear speech exhibits an increase in pause frequency, pause duration and word duration compared to its casual counterpart, as reported from previous studies and as quantified in this work as well (Chapter 2). However, our previous effort to time-scale casual speech in order to match the duration of clear speech using segmental time-scaling has failed to increase intelligibility. Other studies have reported similar results (Uchanski et al., 1996a; Picheny et al., 1986) as we have described in Chapter 1. Possible reason for degrading the intelligibility of casual speech includes the introduction of artifacts in the modified speech due to the segmental time-scaling techniques (Uchanski et al., 1996a; Picheny et al., 1986). Also, the absence of pauses and other acoustic or linguistic information in some parts of the casual signal compared to the clear signal have resulted in an inappropriate (e.g. overly-exaggerated) elongation of these parts. For example, Figure 3.7 depicts the waveform of the phrase "full of". In the clear signal, each phoneme is elicited carefully and a pause exists between the words "full" and "of" as shown in Figure 3.7. The casual signal, however neither contains pauses nor all of the acoustic-phonetic information that the clear signal has. Segmental time-scaling consequently results in a exaggerated elongation of casual speech, as the lower graph of Figure 3.7 depicts. Other studies (Uchanski et al., 1996a) attempted to add pauses to casual speech to reduce its speaking rate. However, the location of the pauses inserted on casual speech was indicated by the corresponding locations of clear speech. This method is not necessarily correct, since the pauses may be inserted at implausible locations (e.g places with high energy).

The objective of this work is to explore different acoustically-driven time-scaling techniques for decreasing the speaking rate of casual speech and evaluate the impact of these techniques on intelligibility. Inspired by the above properties of clear speech on a purely acoustic level, the proposed modifications insert pauses and elongate parts of the casual signal. Unlike other time-scaling techniques that try to copy the duration characteristics of clear speech to casual speech, the presented modifications consider only the acoustic properties of casual speech. Specifically, two general classes of techniques are explored. First, uniform time-scaling is employed as a baseline modification so that the casual speech sentence duration matches that of the clear. The uniform time-scaling slows down prosody uniformly without changing the relative durations between segments of speech. Second, a refined approach is proposed considering loudness and stationarity, in determining where and how to elongate certain parts of the signal and insert pauses. The goal of this technique is to elon-

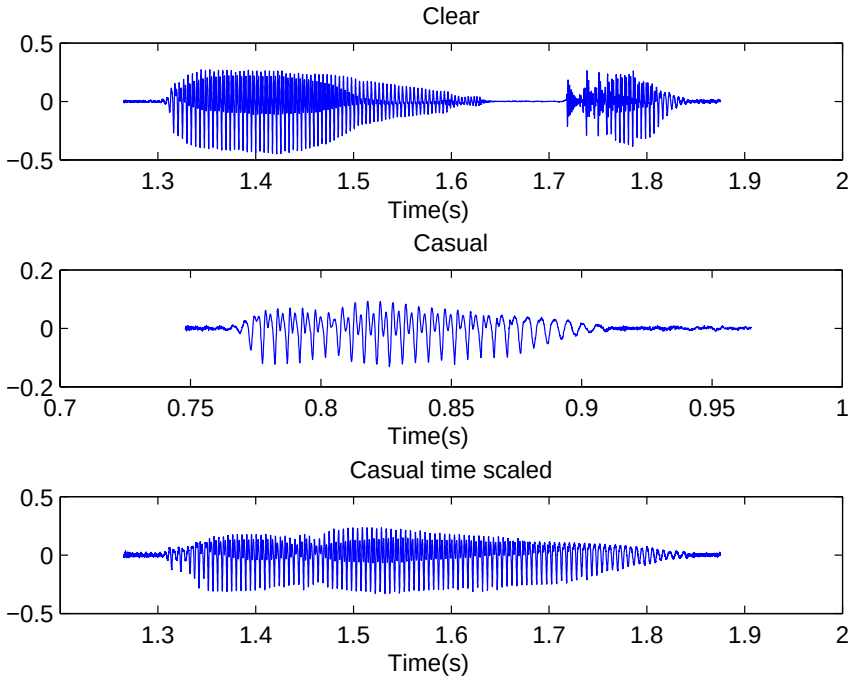


Figure 3.7: Segmental time-scaling of casual speech to match the duration of clear phrase "Full of" a) clear speech (top) b) casual speech (middle) c) modified casual speech using segmental time-scaling (bottom).

gate speech, while avoiding disturbing artifacts from the time-scaling of non-stationary parts of speech and to insert pauses on low energy parts of casual speech. Compared to segmental time-scaling this technique is considered less intrusive, since it takes into consideration the acoustics of casual speech (stationarity, loudness etc.) inspired, however, by clear speech properties (increased number of pauses, increased vowel duration).

3.2.1 Proposed time-scaling modifications

Unlike previous techniques, this work uses a combination of pause insertion and elongations that seek to respect the acoustics of the signal, in order to limit excessive elongations and artifacts. Specifically, the pause insertion scheme does not copy the locations of pauses on clear speech but is an unsupervised method that respects the acoustic properties of casual speech (pauses are not inserted in high loudness positions). Moreover, the proposed pause insertion scheme can be seen as also being inspired by the duration manipulations of clear speech to make word boundaries clearer (Cutler and Butterfield, 1990). That is, the hope in inserting pauses is to help distinguish word boundaries. Additionally, the elongations are carried out either uniformly or based on loudness and stationarity criteria in order to limit artifacts. These elongations primarily serve to offer the listener more time to understand a given sound. Explicitly, it should be noted that, even if time-scaling is employed to generate casual speech with lower speaking rate, the modified casual speech will still not carry the enriched acoustic-phonetic information of clear speech. However, the motivation for these modifications is that the elongation of speech segments and the discrimination between words using pauses may prove to

be beneficial for the listener, providing him/her with the appropriate time to process the message (Ghitza and Greenberg, 2009).

Two general classes of time-scaling techniques are explored; blind time-scaling and refined time-scaling modifications. First, blind time-scaling is based on uniform time-scaling, taking into account only sentence durations in order to match those of casual speech to clear speech. This technique aims to slow down prosody uniformly, without the insertion of pauses, keeping the ratio of the durations of the speech segments unaffected. Second, refined time-scaling firstly inserts pauses in casual speech and then elongates parts of speech, combining two acoustic features that are observed in clear speech. The refined time-scaling is based on the Perceptual Quality Measure model (PSQ). It uses a loudness criterion to insert pauses in casual speech and to elongate mainly the stationary parts of speech. The goal of this technique is to introduce acoustic properties that are observed in clear speech, namely the elongation along with the existence of pauses, while limiting speech degradations that may be introduced by attempting to time-scale non-stationary parts of speech. Moreover, the pause insertion scheme is acoustically meaningful in that it is based only on the acoustic properties of casual speech, inspired nevertheless by the properties of clear speech.

In this work, three time-scaling methods are evaluated subjectively: 1) Uniform time-scaling 2) our novel time-scaling approach based on the Perceptual Quality Measure (PSQ) model, and 3) Time-scaling using a Rhythmogram - based approach [1]. The Rhythmogram (RM) described in Stylianou et al. (2012) is used here for comparison reasons. RM uses similar to PSQ criteria for elongation and pause insertion. The Rhythmogram time-scaling method uses the Rhythmogram (Todd and Brown, 1994, 1996) to capture and visualize the overall rhythm of the speech. The Rhythmogram level curve is used to elongate louder parts of speech and to detect where to insert pauses. While PSQ adds a steady pause, the Rhythmogram-inspired algorithm adds pauses proportional to the rhythm of speech. The Uniform time-scaling expands the casual signal uniformly to reach clear speech duration without adding pauses. On the other hand, Perceptual-Speech-Quality-based (PSQ) and Rhythmogram-based (RM) approaches are trying to avoid speech artefacts created by scaling non-stationary parts of speech. Therefore, PSQ and RM perform time-expansion of the casual speech by elongating the stationary speech portions of the signal and by adding pauses to the speech signal at specific instances.

Uniform time-scaling

Uniform time-scaling provides a non-invasive way to time-scale a sentence without disturbing its intra-segmental characteristics. The uniform time-scaling respects the relative durations between segments of speech. In contrast to segmental alignment, the uniform time-scaling does not force specific segments of casual speech to be perfectly aligned to the corresponding clear speech parts but takes into account only the duration of the clear signal. Uniform time-scaling is performed by feeding the Waveform Similarity Based Overlap-Add algorithm - WSOLA (Demol et al., 2004) a constant scale factor, that is the ratio of casual signal duration to that of the clear signal. Then, WSOLA time scales the casual signal to match the duration of the

[1] The author would like to thank Elizabeth Godoy postdoctoral research fellow at ICS-FORTH and Vincent Aubanel, postdoctoral research fellow at Laslab for providing the Rhythmogram algorithm.

clear signal. This scheme does not account for pause insertion.

Perceptual Speech Quality Measure based Time-Scale Modifications

In this work, the Perceptual-Speech-Quality measure (PSQ) is used to elongate the stationary parts of casual speech and to define where to insert pauses to the signal. The Perceptual Speech Quality measure is based on the basic version of *ITU Standard REC-BS.1387-1-2001*, a method for objective measurements of perceived speech quality. It estimates features such as loudness and modulations in specific frequency bands, in order to describe the input signal with perceptual attributes.

Elongation of voiced parts of speech

Two metrics of the PSQ model are used to detect the stationary parts of speech, where time-scaling can be applied: the perceived loudness of the signal in low frequency bands and the loudness modulations in high frequency bands. Analytically, PSQ estimates the perceived loudness on the low frequency bands (0-300Hz) of the signal, where unvoiced speech is less likely to be present. However, some voiced stop consonants have high energy in low frequency bands. A characteristic example is shown on Figure 3.8. The top graph of this Figure depicts the speech signal that corresponds to the phrase "made a sign" and its average perceived loudness in low frequency bands, as calculated by PSQ. The loudness is depicted with a green curve. Using only the loudness metric to distinguish stationary from non-stationary parts of speech is not sufficient, since consonants like /d/ (as depicted in Figure 3.8, at time instant 2.1s) have high energy on low frequency bands. Time-scaling voiced stop consonants would cause distortion. Therefore, the loudness is not the appropriate metric to decide which parts of speech should be elongated.

However, by combining the loudness with a another metric, namely the loudness modulations of high frequency bands (around 4000Hz), non-stationary parts of high loudness values on low frequency bands can be detected as non-stationary. The loudness modulations in high frequency bands are strongly correlated with the non-stationarity of the signal and are able to detect voiced stop consonants. On the top graph of Figure 3.8, the red curve corresponds to the loudness modulations in high frequency bands for the speech segment "made a sign". These modulations have high values for plosives and very low values for vowels. Subtracting the modulation values from the loudness values, we detect the stationary parts more efficiently. Analytically, let us denote with L the average perceived loudness in low frequency bands and with M the loudness modulations in high frequency bands calculated by PSQ for a speech frame and S their difference, $S = L - M$. The idea behind this subtraction is that for a speech frame corresponding to a vowel, L is high and M is almost zero. Therefore, S has approximately the same values as L (this applies more to the center of the vowels rather than to the edges, where the loudness is rather low). For a fricative, L is almost zero and M is high leading to a negative value of S. Finally, in the case of plosives both M and L are high and, after subtraction, S values are expected to concentrate around zero or around a threshold. Frames that correspond to negative values of S or values of S close to the threshold are treated as non-stationary.

The threshold, below which the frames are categorized as non-stationary, is defined experimentally. Specif-

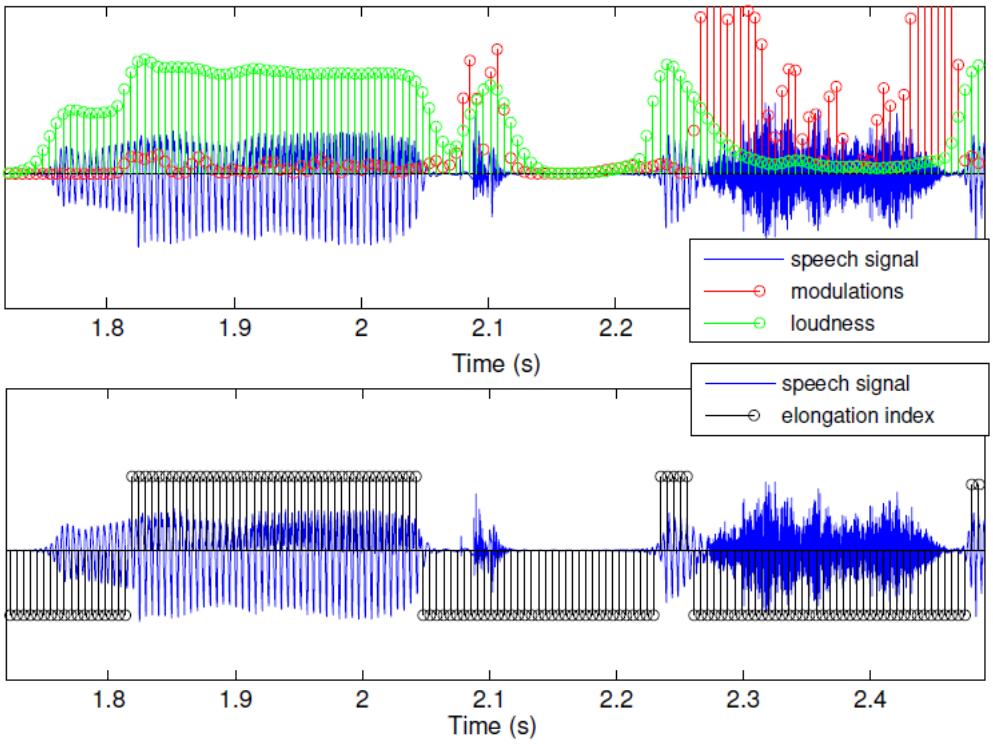


Figure 3.8: Detection of non-stationary parts using PSQ model on the sentence "made a s(ign)" a) Loudness in low frequency bands and modulations in high frequency bands (top) b) Elongation index (bottom)

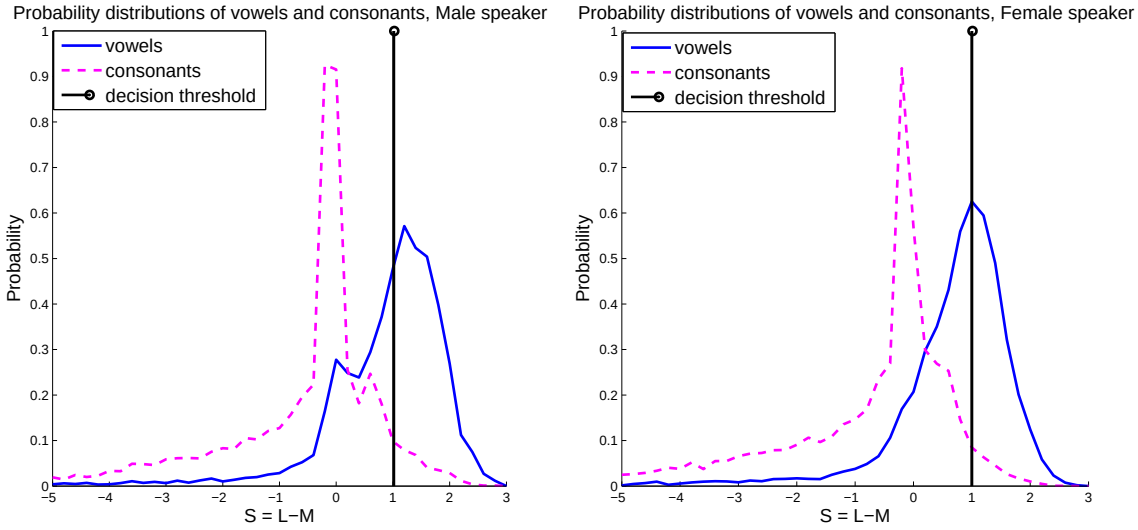


Figure 3.9: Defining the decision threshold for detecting stationary parts of speech. For each speech frame of the 100 sentences uttered by a Male (left) and a Female (right) speaker, the loudness L and the loudness modulation M are estimated. The histogram of the difference $S = L - M$ for all vowel-frames and all consonant-frames is computed and the normalized histogram (probability distribution) is depicted for each category $\{consonants, vowels\}$. The horizontal line at the value 1 is the decision threshold that classifies stationary from non-stationary parts of speech, taking into account a high cost in case of consonant misclassification.

ically, for two speakers on the database one male and one female and for 100 sentences for each speaker, automatic aligner is used to distinguish consonant-frames from vowel-frames. Then, the average perceived loudness in low frequency bands L and loudness modulations in high frequency bands M are calculated per frame. The difference S between these values for each frame is calculated and the normalized histograms of S for vowel and consonant frames are depicted on Figure 3.9. The left and right plot of Figure 3.9 depicts the probability distribution of the value S for two frame categories {consonant-frame, vowel frame} for 100 sentences uttered by a male and a female speaker, respectively. Indeed, consonant-frames concentrate near and below zero, while vowel-frame values fluctuate between 0 and 2. Based on Figure 3.9, the optimal threshold for classification would be the point where the two probability curves cross. However, while the cost for a vowel misclassification is zero, a consonant misclassification introduces a cost, since in this case consonants are treated as vowels and their elongation degrades the signal. Here, "cost" is a general term. No attempt has been made to introduce a cost function and minimize it. For avoiding introducing artifacts on the modified signal, a threshold of value 1 is selected. This threshold value rejects many voiced speech frames. However, it ensures that the majority of non-stationary parts of speech will not be elongated. The number of mistakingly classified consonant-frames is much lower using the proposed S metric comparatively to the loudness criterion. This applies especially for the voiced consonant-frames. Figure 3.10 depicts the probability distributions of the loudness L and of the proposed metric S, computed for all the vowel frames and the voiced consonants {b, g, w, l} for the male speaker. Assuming that the decision threshold is again at 1, if the loudness metric L is selected for classification, the number of voiced frames allowed to be elongated is indeed greater than that of the proposed metric, but so is the number of consonant frames and therefore the misclassification error. In Figure 3.10 the misclassification error is the area below the consonant curve on the interval [1, 3].

Therefore, the combination of the two metrics decides the elongation or not of a speech frame. This binary decision for each frame is incorporated into an index, called Elongation Index (EI):

$$EI = \begin{cases} 1 & \text{if } S \geq threshold \\ -1 & \text{if } S < threshold \end{cases} \tag{3.2}$$

where S is the metric described above, namely the difference of the loudness modulations in high frequency bands from the average perceived loudness in low frequency bands. The value of the threshold is set to 1, estimated as described. A speech frame will be elongated if the value of EI for that frame equals to 1, otherwise if $EI = -1$ the corresponding frame will not be elongated. The lower graph of Figure 3.8 shows the elongation index EI calculated for each frame of speech. Indeed, frames belonging to phonemes /ey/ and /e/ will be elongated ($EI = 1$), whereas frames belonging to consonants /n/, /s/ and /d/ are indexed with $EI = -1$ and therefore will not be elongated. Each frame with $EI = 1$ is time-scaled 20% of its original duration. The time-scaling is performed by WSOLA.

Pause insertion

Pause insertion is also implemented using the PSQ model. The proposed pause insertion scheme is a purely

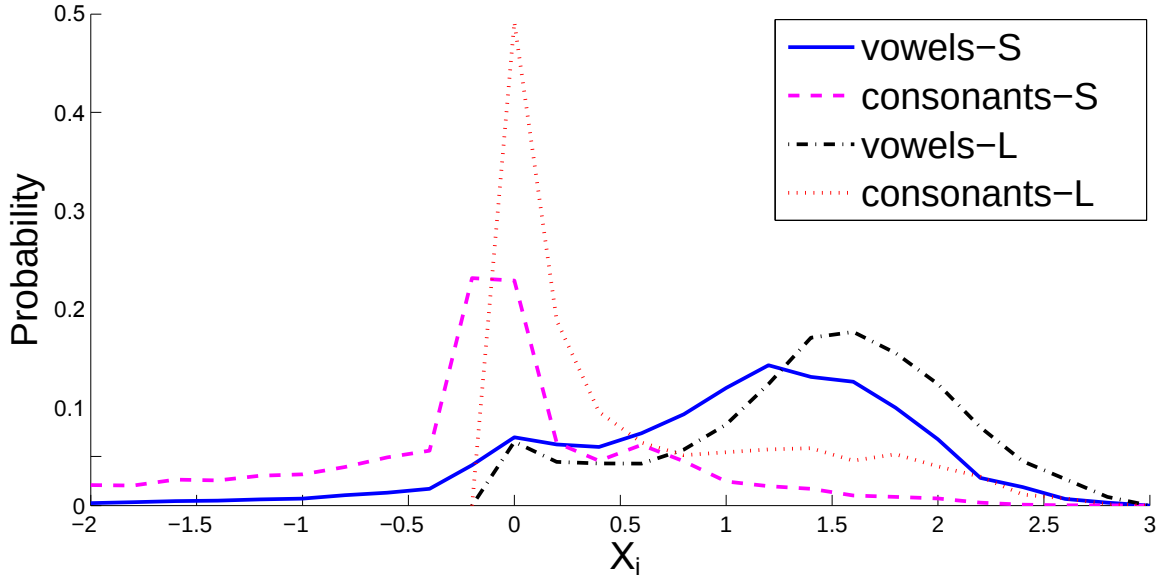


Figure 3.10: Comparing the classification error of the two metrics, the loudness metric L and the proposed metric S on vowel and voiced consonants frames. For each speech frame of the 100 sentences uttered by a Male speaker the loudness L and the metric S are estimated. The histograms of the values S and L for all vowel-frames and for the voiced consonant-frames {b,g,d,l} are computed and the corresponding normalized histograms (probability distributions) are depicted. Selecting a decision threshold $T > 0.5$ for consonant and vowel classification, the misclassification error of the proposed metric S for the consonants (the area below the consonant curve on the interval [0.5, 3]) is lower than that of the L metric.

unsupervised and takes into consideration the acoustic properties of casual speech. A pause is inserted if the loudness of the signal lowers in that part of speech and after a pre-processing is performed on the estimated word boundary. This pre-processing involves elongation of the part of speech before the word boundary, inspired by the the fact that speakers increase the duration of the last part of a word before the next word is elicited in order to distinguish two words (Cutler and Butterfield, 1990).

Specifically, the perceived loudness of the speech signal in the whole frequency band is estimated (in dB SPL). Then, loudness is normalized by the maximum loudness of the signal. After normalization, the values of the normalized curve lie in the interval [0,1]. Then, all valleys are detected on the normalized loudness curve. The valleys with very low values, less than 10% of the normalized loudness of the signal, can be considered silences. Therefore, pauses can be inserted on the places of these valleys. These pauses are named non-aggressive pauses because they are inserted on very low energy parts of speech. On the other hand, it is observed that the valleys that fall within the interval (10%, 20%] of the normalized loudness are usually in the middle of word boundaries and are appropriate for inserting pauses without distorting the signal. The pauses that result from these valleys are called aggressive pauses to distinguish them from the pauses derived from the valleys with very low values of loudness (non-aggressive). The PSQ algorithm adds both non-aggressive and aggressive pauses to the signal.

The reason for the distinction between aggressive and non-aggressive pauses is that the algorithm uses different techniques to do the insertion. First, the non-aggressive pauses are inserted on the signal. Then, in order to insert the non-aggressive pauses on the location where the signal has higher loudness, a pre-processing of the signal before and after the location of the valley must be made. The pre-processing involves a time-scaling of the signal around the location where the gap will be inserted, if this is allowed by the stationarity restriction. Then, after scaling, a hamming window is applied on the center of the valley so that the transition from speech to silence will be more smooth. Both aggressive and non-aggressive pauses have a fixed pause length of 90 ms, based on the average pause duration on clear speech.

3.2.2 Evaluations

In this section the proposed time-scaling techniques are evaluated. Subjective evaluations were performed by native and non-native speakers for the five set of signals, namely the clear (Clear), the casual (Casual), the casual time-scaled to match the duration of the clear speech using the uniform time - scaling (Ucasual), the PSQ (Pcasual) and the Rhythmogram (Rcasual). For the evaluation purposes, the PSQ modified sentences and the Rhythmogram modified sentences were also time-scaled uniformly to match exactly the duration of the clear signal. This modification, even though it had a minor effect on the duration of the already modified signals, was done in order to ensure that all the sentences, apart from the sentences corresponding to casual speech, will have the same duration in the evaluation set.

From the corpus of LUCID database, 124 distinct sentences were selected uttered by 4 female and 4 male speakers in a clear and casual style. A pre-processing was performed on the dataset to remove low-pass noise from breath and lip effects, using a 5-order highpass digital elliptic filter with 80Hz cut-off frequency. In the perceptual tests, SSN was added to the signals to create the test signals, with SNR of 0 dB. This noise level was considered the optimum for revealing intelligibility differences between the speaking styles, since for this SNR it was observed a maximum distance between the intelligibility scores of clear and casual (see Figure 3.5). From the dataset consisting the five set of signals {Clear, Casual, Ucasual, Pcasual, Rcasual}, 12 sentences were randomly selected from each set. Therefore, the total set contained 60 unique sentences for evaluation. Then, the listener heard each sentence once and was instructed to write down whatever she/he perceived to have heard. The listening test was performed in a quiet room with headphones of good quality. Each listener was allowed to select the listening level of his/her comfort using a test sentence that was presented to the listener prior to the listening test.

The listening test was performed by 8 native (British English) listeners[2] and 10 non-native listeners with a good comprehension level of English. For each sentence of a set, the percentage of the correct perceived words in the sentence (function words were not taken into consideration) was computed and then the average of these percentages for all sentences of the set was calculated to define the percentage of correct perceived words in the set. Figure 3.11 presents the intelligibility scores (% of correct perceived words) on the 5 sets for

[2]Many thanks to Dr. Sonia Granlund, from University College of London for organizing the listening tests.

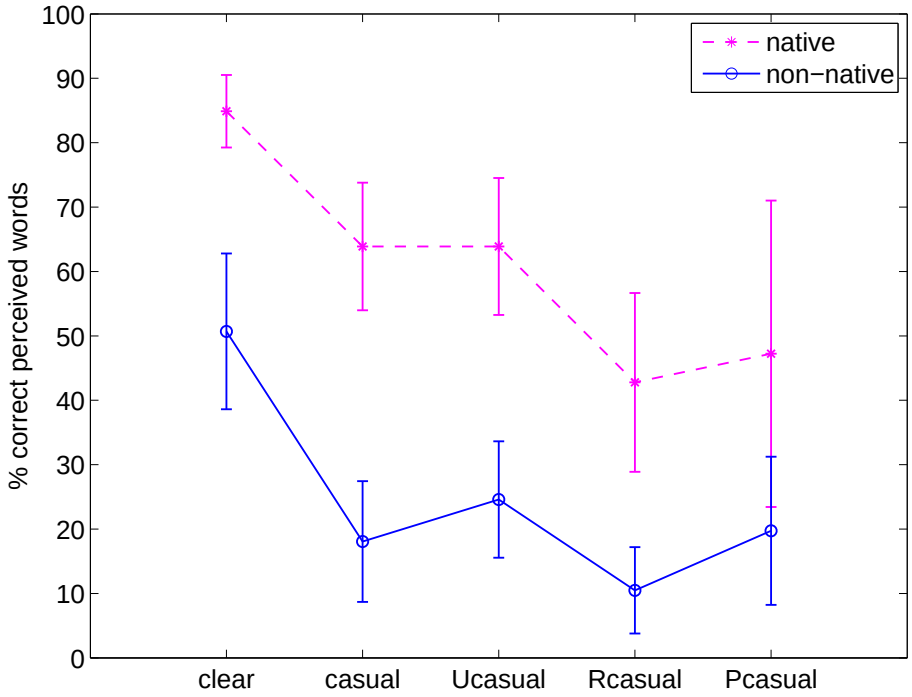


Figure 3.11: Subjective Intelligibility Score for the 5 set of signals for 0dB SNR. The percentage of correctly perceived words for each set for native and non-native listeners and the corresponding standard deviations. The Ucasual, Rcasual and Pcasual refer to the casual speech modified by the corresponding time-scaling techniques Uniform, Rhythmogram-based and PSQ-based.

native and non-native speakers. Figure 3.11 shows that clear speech has a profound intelligibility advantage over casual speech and that time-scaled casual speech is almost the same or less intelligible than unmodified casual speech, regardless of the modification technique. There seems to be a slight advantage of the uniformly time-scaled casual speech over original casual speech. However, this advantage is not statistically significant.

In order to look for statistical significant differences between the mean scores of intelligibility for each category {Clear, Casual, Ucasual, Pcasual, Rcasual}, tests of ANOVA were performed. Firstly, the ANOVA null hypothesis that all mean values of the intelligibility scores for every category were equal was rejected using the F-test ($F(4, 85) > 8.4, p < 0.01$). Then, pairwise comparisons of the means were performed using Fisher's Least Significant Difference (LSD) test in order to derive which of the groups differ significantly. The test was performed for the whole subjective test corpus and for the two groups of native and non-native listeners separately. The results (Table 3.2) show that there is a significant categorical difference between Clear and Casual. However, no significant categorical difference was found between Casual, Ucasual, Pcasual and Rcasual. This reveals that the two groups {Clear} and {Casual, Ucasual, Pcasual, Rcasual} are statistically different but no significant differences are found within a group. Examining separately the two populations, native and non-native, results resemble those described in Table 3.2.

Figure 3.12 shows the percentage difference of correctly perceived words between each set and the casual speech, for non-native (top) and native listeners (bottom). Specifically, using the casual set as a reference set,

	Clear	Casual	Ucasual	Pcasual	Rcasual
Clear		**27.61**	**23.83**	**35.72**	**38.24**
Casual	**-27.61**		-3.79	8.10	10.62
Ucasual	**-23.83**	3.79		11.89	14.41
Pcasual	**-35.72**	-8.10	-11.89		2.52
Rcasual	**-38.24**	-10.62	-14.41	-2.52	

Table 3.2: Significant Categorical Difference between the five sets of speech for all the listeners, using Fisher's Least Significant Difference (LSD) test. The standardized difference is given for each pair of sets. Significant differences are in bold.

we subtract the percentage of correctly identified words on the casual set from the correctly identified words on each set and we depict every difference in Figure 3.12. For example, in Figure 3.12 for the non-native listener 1, positive difference for clear speech means that the amount of correctly perceived words in clear speech was greater than that of casual speech, whereas negative difference for Rcasual shows that the particular listener identified correctly more words for unmodified casual speech than for this set that was modified with the Rhythmogram-based approach. Eleven out of eighteen listeners, indeed, achieved better score when casual speech was uniformly time-scaled than when it was presented unprocessed. This percentage is much higher if we consider only the non-the native speakers. Seven out of ten non-native listeners seem to prefer the uniform time-scaled sentences than unprocessed casual speech. Even though there seems to be no statistical significance between the Casual and the Ucasual set for the non-native listeners, it is important that the majority of the non-native listeners finds the uniform time-scaling technique beneficial.

3.2.3 Discussion

Modifying casual speech in order to achieve the intelligibility gain of clear speech is quite a challenging task. The difficulty in the modification not only resides in signal processing aspects (formant estimation, stationarity limitations etc.) but also on the fact that no prominent features have been found to uniquely and unequivocally increase intelligibility, despite the extensive research on the topic. Moreover, since speech production and perception function on numerous linguistic and acoustic levels, exploring features (or even better the combination of features) that contribute to the intelligibility gain of clear speech becomes even trickier.

Consequently, this work tried to approach this complex problem focusing only on one level of speech, the acoustic level, and examining the most prominent feature that differs between the two speaking styles, the speaking rate. Focusing only on the acoustic level, compared to its casual counterpart, clear speech exhibits an increase in pause frequency, pause duration and word duration. The proposed time-scaling techniques tried to incorporate these acoustic features in the casual speech with intent of increasing its intelligibility. Unlike previous works, though, the suggested pause insertion scheme respects the acoustic properties of casual speech. The locations of the pauses were not copied from the clear signal and then incorporated on the casual signal. Intervening a pause on a high loudness/energy location would disrupt the speech flow and would create

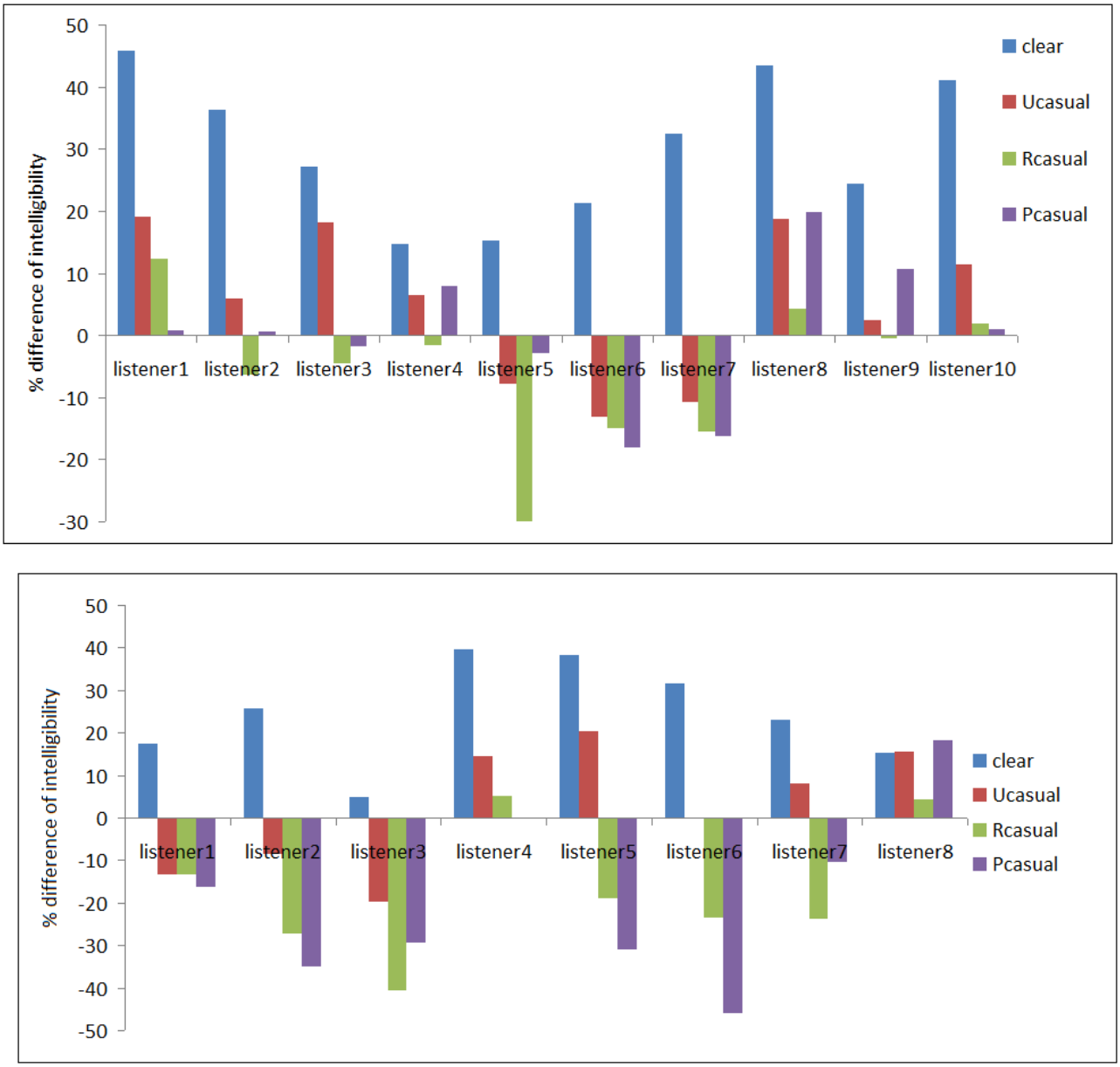


Figure 3.12: Difference of the percentages of correctly perceived words between each set and the casual speech, for non-native (top) and native listeners (bottom). The Ucasual, Rcasual and Pcasual refer to the casual speech modified by the corresponding time-scaling techniques Uniform, Rhythmogram-based and PSQ-based.

undesirable energy discontinuities. Both the Rhythmogram and the PSQ took into consideration this energy restriction. Specifically, the Rhythmogram-based approach added pauses to casual speech proportional to its rhythm whereas the PSQ based approach elongated the last parts of the syllables before inserting a pause resulting in a smoother pause transition.

Evaluations of the time-scaling approaches were conducted via listening tests considering speech-in-noise intelligibility. While the uniform time-scaling achieved essentially the same intelligibility as the unmodified casual speech, with some improvement observed for non-native listeners (7 out of 10 non-native listen-

ers reported an increase in intelligibility compared to unmodified casual speech as Figure 3.12 shows), the more refined approaches with elongations and pause insertions did not prove advantageous. PSQ-based and Rhythmogram-based techniques thus implement a largely successful pause insertion method in terms of acoustics but not in terms of intelligibility.

One possible reason that the time-scaling approaches could not improve speech intelligibility is the change in naturalness of speech. However, naturalness has a multi-level significance. A modified/synthetic speech signal may sound unnatural if its rhythm differs from natural elicited speech (text to speech synthesizers for example sound unnatural in terms of prosody) or because artifacts are introduced on the signal by the modification algorithms etc. Moreover, even natural elicited speech may sound unnatural, e.g Lombard speech in quiet. Unnatural speech, however, may be intelligible. In the previous section, spectral transformed casual speech (SSDRC modified) had higher intelligibility than clear speech in noisy conditions although it sounded unnatural in noiseless conditions (harsh). One could, therefore, suggest that naturalness and intelligibility are not related. However, there seems to be a strong connection between naturalness in rhythm and intelligibility. The degradation of intelligibility after the removal of pauses (Uchanski et al., 1996a) does not only show that the existence of pauses is important for speech intelligibility. It reveals that the rhythm of speech is important in intelligibility. Many text-to-speech systems that lack of naturalness in rhythm also appear to have lower scores on intelligibility than natural speech in noisy environments (Valentini-Botinhao et al., 2012). The refined time-scaling methods implemented on this work change the relative durations between segments of speech impacting the naturalness of the speech rhythm and possibly intelligibility.

Finally, another difficult issue that we faced in this study and was also reported by other related studies of speech intelligibility is the listener variability. Speech modifications may be beneficial to some target groups of listeners but not beneficial to others (Narne and Vanaja, 2008). Listener variability in judgements of intelligibility also appears across speakers or tasks (McHenry, 2011). In addition, native listeners process the speech different than non-native listeners since the linguistic experience of the native listeners provide them with a greater intelligibility advantage. Furthermore, non-native listeners seem also to be more affected in noisy-conditions than native listeners (Cooke and Lecumberri, 2012a; Smiljanic and Bradlow, 2009). Potentially, the slowing down of speech using the uniform time-scaling approach is beneficial to the majority of the non-native listeners, as it is reported by our subjective evaluations (Figure 3.12). Possibly, a more extensive listening test on the population of the non-native listeners could reduce this variability but for sure it cannot eliminate it. Only clear speech is judged by all listeners as more intelligible than casual speech. Therefore, in order to increase the intelligibility gain of casual speech significantly, future directions towards casual speech modifications should account not only for one level of exploration but for parallel modifications on spectral-time domain (e.g formant movements), also incorporating knowledge from the linguistic level of clear speech. That is, the key for defining which of the numerous features prominent in clear speech are responsible for positively contributing to its intelligibility may be a deeper understanding of how the various levels of linguistic structure interact.

Chapter 4

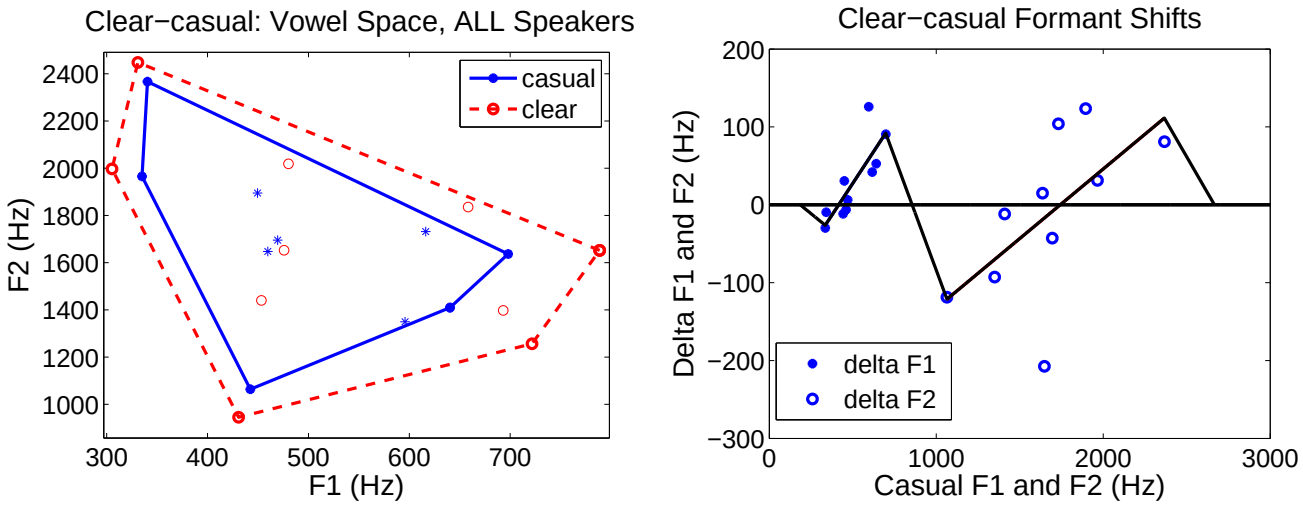
Vowel Space Expansion

Among the key acoustic features attributed with the intelligibility gain of clear speech is the observed expansion of vowel space, representing greater articulation and vowel discrimination. The gain, however, resulting from this expansion remains obscure. Possible reason is the difficulty of manipulating vowel spaces due to limitations imposed by accurate formant estimation and modification. The recent work in Mohammadi et al. (2012) statistically transform both formant frequencies and the spectral envelope of casual vowels to resemble those of clear. However, the vowel space expansion is not addressed in isolation and therefore the intelligibility impact due to vowel space modifications is not assesed. The present work focuses on revealing possible intellligibility gains from expanding the vowel space of casual speech. Motivated by the acoustic analysis performed in Chapter 2, where the most efficient speakers in terms of intelligibility had the more expanded vowel spaces, we attempt to assess the intelligibility impact of expanding the vowel space of casual speech to mimic that of clear speech. Specifically, a clear speech-inspired frequency warping method is described for expanding the vowels space of casual speech. The frequency-warping scheme has been successfully used in the past for voice conversion proposed by Godoy et al. (2012). The method successfully achieves vowel space expansion when applied to casual speech. The intelligibility impact resulting from this expansion is then evaluated objectively and subjectively through formal listening tests.

4.1 Observed Vowel Space Expansion and Formant Shifts

In Chapter 2, the vowel spaces of clear and casual speech have been estimated using all of the vowel instances for the speakers in the specified LUCID corpora (Figure 2.9). For simplicity reasons, Figure 2.9 is again depicted in Figure 4.1(a). It is evident from Figure 4.1(a) that the clear speech vowel space is expanded compared to the casual speech. Additionally, Figure 4.1(b) shows the F1 and F2 differences between the clear and casual vowel spaces, showing the amount that each formant in the casual vowel space needs to be shifted in order to match the corresponding formant in the clear vowel space. Rather than a uniform increase or decrease, it can be seen that both F1 and F2 are shifted either up or down, depending on the formant frequency. Specifically, low F1 or F2 of the casual speech are decreased, while high F1 or F2 are increased

going from casual to clear, ultimately expanding the vowel space. This trend is shown more explicitly by fitting the formant shifts with a linear regression, as shown by the solid lines in Figure 4.1(b). Then, in order to visualize frequency shifts of a piecewise linear function that would emulate an expanding vowel space, a linear interpolation between the F1 and F2 boundaries and a return to zero-shift at the endpoints is also indicated. Ultimately, the overall form of this function inspires the frequency warping approach considered in this work.



(a) (b)

Figure 4.1: (a) Casual and clear vowel spaces. (b) Casual-to-clear formant shifts with piecewise linear fitting.

4.2 Frequency Warping for V.S. Expansion

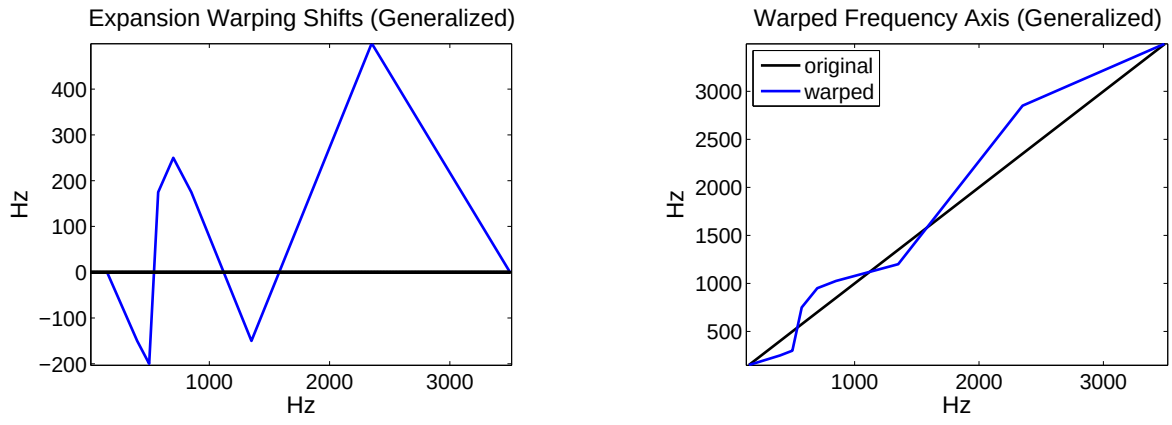
Typically used in voice conversion (Valbret et al., 1992; Godoy et al., 2012; Erro et al., 2010), frequency warping is employed here in a novel manner as a means for vowel space expansion. The appeal of frequency warping for this expansion is that it offers a way of shifting speaker formants, while both avoiding notable speech degradations and limiting dependence on accurate formant detection. In particular, this work proposes a frame-based piecewise linear frequency warping function based on the related Dynamic Frequency Warping (DFW) algorithms used for voice conversion (Godoy et al., 2012). However, the intervals of the warping function are defined in this work by sampling a curve (based on spectral peak locations) of exaggerated formant shifts that is drawn from clear-casual vowel space analyses. In examining the vowel space of warped casual speech, it is confirmed that the proposed approach successfully yields expansion. Then, the corresponding intelligibility impact is assessed using objective and subjective evaluations. In the end, results ultimately motivate more careful consideration and qualification of the clear speech intelligibility advantage in relation to vowel space expansion.

4.2.1 Method Description

The proposed frequency warping approach for vowel space expansion can be described in two stages. The first defines a curve of generalized warping shifts, inspired by the formant shifts observed in the clear speech vowel expansion. The second stage then outlines the frequency warping algorithm based on sampling the aforementioned curve on a frame-by-frame basis. The following respectively describes these stages in more detail.

Clear Speech-Inspired V.S. Expansion Shifts

Working from the trends shown in Figure 4.1(b), the curve $\Delta(f)$ of generalized warping shifts (Figure 4.2(a)) used in the proposed approach was determined after several trials observing warped vowel spaces and by taking into account certain practical considerations. First, it is noted from Figure 4.1 that the magnitude of the formant shifts, on average, is quite small, especially compared to the separation of harmonic peaks (e.g., about 150 Hz on average) in the amplitude spectrum. Consequently, to define the curve of generalized warping shifts, the magnitude of the shifts must be significantly larger in order to overcome the harmonic structure in the amplitude spectrum and effectively shift a formant. Second, since the span of F1 is less than F2, care should be taken such that the F1 warping will achieve the desired effect without approaching DC or overlapping with F2. Consequently, the slope from negative to positive F1 shifts is exaggerated and the maximal shift is rounded out. Finally, the defined shifts should ensure that any warped frequency axis is always monotonically non-decreasing. Figure 4.2(b) displays the warped frequency axis generated from $\Delta(f)$, confirming that the slope is always non-negative. Figure 4.2(b) similarly indicates bounds on all possible warped axes. Specifically, the warped frequency axis of any frame will be defined as a set of lines connecting points (i.e., detected spectral peak frequencies) on the generalized warped axis.



(a) (b)

Figure 4.2: (a) $\Delta(f)$ - Generalized curve of exaggerated warping shifts. (b) Corresponding warped frequency axis.

Frequency Warping Algorithm

Before outlining the proposed frequency warping approach, some details of the speech analysis and synthesis are presented. Specifically, the analysis and synthesis is pitch-asynchronous, using a 30ms Hamming window and 10ms step. Each frame is analyzed using a 2048-pt DFT. In the case of speech modification, frequency warping filters are applied to the amplitude spectrum of a frame before synthesis using the inverse DFT (with the original phases) and overlap-add.

Now, let the spectral envelope from frame n of unmodified speech, $S_n^X(f)$, be represented here by the True Envelope using a cepstral order of 48 (Roebel and Rodet, 2005). The frequency warping filter for frame n, $H_n^W(f)$, is then defined as

$$H_n^W(f) = S_n^W(f)/S_n^X(f) \tag{4.1}$$

where $S_n^W(f)$ is the warped spectral envelope

$$S_n^W(f) = S_n^X(W_n^{-1}(f)) \tag{4.2}$$

and $W_n(f)$ is the warping function for frame n, defined as follows. First, the spectral tilt, $S_n^{Xtilt}(f)$, is generated from the first two cepstral coefficients (zeroth and first order) of the True Envelope analysis. The spectral envelope peaks in the warping frequency range $f \in [150Hz, 3500Hz]$ are then detected from the tilt-normalized spectral envelope as

$$f_{n,i}^X = peak\_detect(S_n^X(f)/S_n^{Xtilt}(f)) \tag{4.3}$$

where $f_{n,i}^X$ indicates the frequency of the i^{th} spectral peak detected in frame n, $i = 1, ...M_n$. The peak detection algorithm defines peaks as local maxima preceded by local minima that are more than 10% lower than the maximum value in the frequency range (so as to avoid ripples or slight fluctuations and inflection points in the envelope). Next, the detected peaks for frame n sample $\Delta(f)$ to provide the intervals defining the frequency warping for the frame. Note that using the detected spectral peaks in this way tailors the frequency warping to the acoustic characteristics of the frame, while avoiding explicit formant estimation (e.g., limiting estimated peaks to F1 and F2), which can be quite error-prone. Specifically, the warped spectral peak frequencies are given by

$$f_{n,i}^W = f_{n,i}^X + \Delta(f_{n,i}^X) \tag{4.4}$$

and these frequencies, together with $f_{n,i}^X$, define a piecewise linear warping function with the form given in Godoy et al. (2012), Erro et al. (2010). Specifically, for $f \in [f_{n,i}^X, f_{n,i+1}^X]$, the warping function for frame n is

$$W_n(f) = A_{n,i}f + B_{n,i} \tag{4.5}$$

where $f^X_{n,0} = f^W_{n,0} = 150Hz$, $f^X_{n,M_n+1} = f^W_{n,M_n+1} = 3500Hz$, and

$$A_{n,i} = \frac{f^W_{n,i+1} - f^W_{n,i}}{f^X_{n,i+1} - f^X_{n,i}} = \Delta(f^X_{n,i+1}) - \Delta(f^X_{n,i}) \tag{4.6}$$

$$B_{n,i} = f^W_{n,i} - A_{n,i} f^X_{n,i} \tag{4.7}$$

With $W_n(f)$ defined from above, the warping filter $H_n(f)$ is calculated and applied to the amplitude spectrum of each frame. Application of the warping to all frames ensures that vowels spaces are expanded, without need for speech segmentation and labeling, while the influence on voiced non-vowels and unvoiced parts of speech is perceptually negligible (as confirmed upon listening to numerous warped speech samples). Furthermore, it should be emphasized that the focus on overall average trends for the vowel space expansion makes the proposed algorithm speaker-independent and thus generalized.

4.2.2 Results

The vowel space for the frequency-warped casual speech is shown in Figure 4.3. The warped vowel space is generated in the same way as the casual and clear vowel spaces in Chapter 2, but with the data being the warped casual sentences. Figure 4.3 shows that the casual speech vowel space is successfully expanded. The warped vowel space area (3.58 compared to 2.32 x 10^5 Hz^2), also confirms this expansion emulating that of clear speech. Moreover, the structure of the vowel space is largely maintained, ensuring that the perceptual distinctions between vowels is respected, with only the distance or discriminability between them being increased. It should be noted that the proposed frequency warping approach is a generalized approximation, rather than deterministic replication, of the vowel space expansion observed in clear speech. Overall, however, the vowel space expansion is achieved and largely respects observations from clear speech via the proposed frequency warping, without explicit vowel or formant identification.

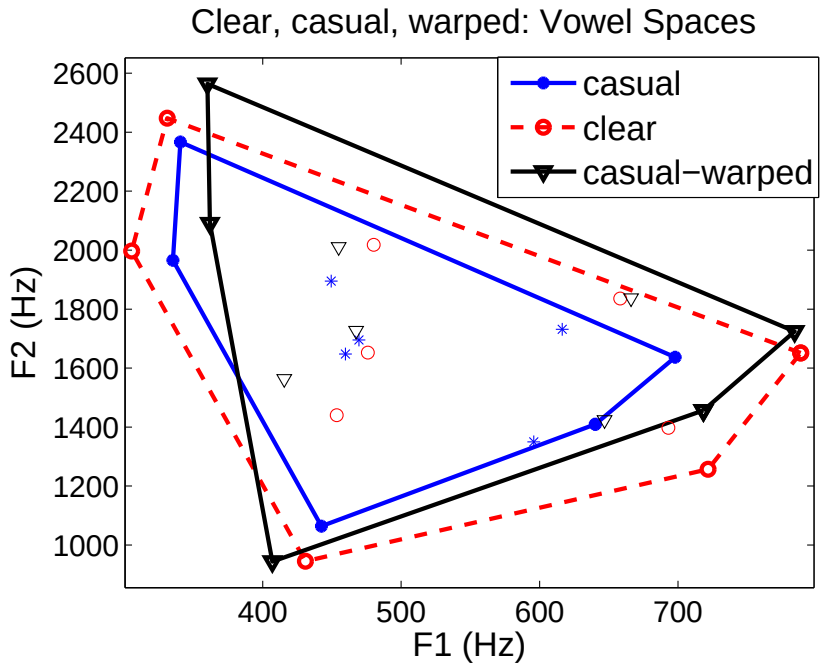


Figure 4.3: Clear, casual and casual-warped vowel spaces, with respective areas: 3.93, 2.32 and 3.58 ($x10^5$ Hz^2).

4.3 Evaluations

4.3.1 Objective evaluations

Preliminary evaluations of intelligibility are conducted using the extSII (Rherbergen and Versfeld, 2005). The extSII was calculated using Speech Shaped Noise (SSN) added to yield a 0dB Signal to Noise ratio (SNR). Table 4.1 gives the average (median) of the extSII distributions for each of the conditions examined in evaluations.

There are a few points to be gleaned from the extSII results. In terms of the frequency warping, there is a very slight gain observed in extSII over the unmodified casual speech, though this factor is probably too small to be meaningful. However, it should also be noted that the extSII fails to capture the intelligibility gain of clear speech, highlighting some potential limitations of objective intelligibility metrics. Consequently, listening tests are required in an effort to capture more subtle acoustic modifications, such as vowel space expansion, in the clear and warped speech.

	Casual	Casual-Warped	Clear
extSII	.311	.316	.312

Table 4.1: Average extSII for Casual, Casual-Warped and Clear speech. The noise masker was SSN added to yield 0dB SNR.

4.3.2 Subjective evaluations

In formal listening tests[1], 20 native English-speakers evaluated LUCID sentences from each of the conditions (casual, casual-warped, clear) described above, with SSN added at two levels to yield 0 and -4 dB SNR, respectively. It should be mentioned that, perceptually, no significant artifacts resulted from the frequency warping, though voice quality was noticeably altered. In the test, listeners were asked to type what they think they heard after hearing each sentence once. Of the listeners, 6 were removed due to inconsistencies in their scores using established conditions (clear speech as a reference). Specifically, those listeners showed very poor scores of clear speech even for 0 SNR. Figure 4.4 shows the intelligibility scores (i.e., percent of words correctly identified) from the remaining listeners for each condition, at the high and low SNR levels. It should be noted that the inter-speaker variability was quite high and, in some cases, frequency warping did improve scores over unmodified casual speech. However, the general trend is displayed in Figure 4.4, indicating no significant improvement and even slight degradation overall.

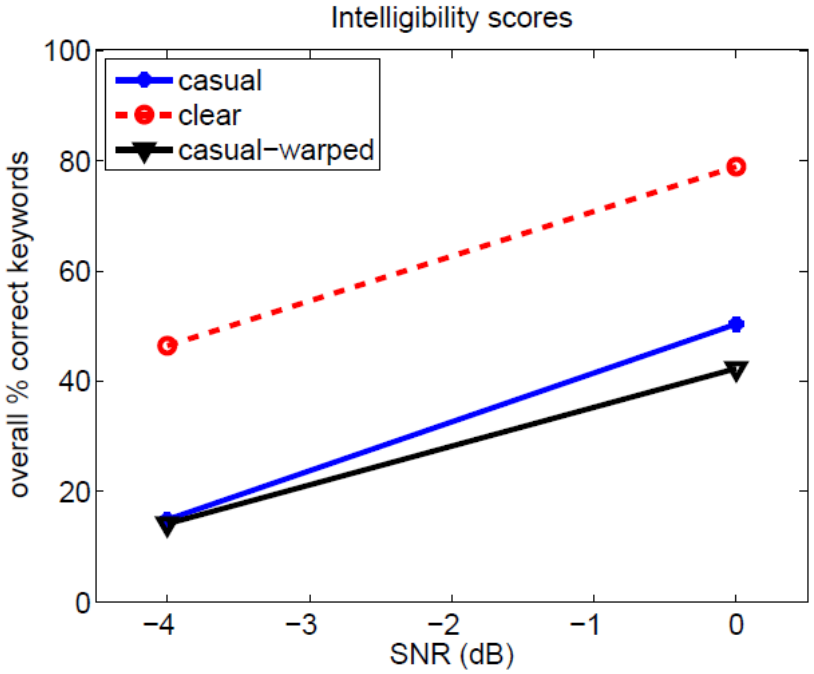


Figure 4.4: Intelligibility test scores for casual, casual-warped and clear speech. The percent of overall correct keyword identification is given for the low (-4dB) and high (0dB) SNR values with a SSN masker.

In order to evaluate the statistical significance of these results, ANOVA tests were performed. Specifically, the ANOVA null hypothesis (i.e., the average values of the intelligibility scores for every condition are equal) was rejected using the F-test (SNR-4dB: $F(4,65)>28.719$, $p<0.05$ SNR0dB: $F(4,65)>25307$, $p<0.05$). Then, pairwise comparisons of the averages were performed using Fisher's Least Significant Difference (LSD) test in order to derive which of the groups differ significantly. The standardized difference between condition pairs is provided in Table 4.2. Analysis of the differences between the conditions confirms a significant categorical

[1]Thank you to Catherine Mayo and CSTR at the University of Edinburgh for their help administering the listening tests.

SNR:	0dB (high)	-4dB (low)
Casual & Casual-Warped	1.41	0.041
Casual & Clear	**5.14**	**5.66**
Casual-Warped & Clear	**6.55**	**5.70**

Table 4.2: Results of Significant Difference Analysis between the clear, casual and casual-warped intelligibility scores. The standardized differences are given and significant differences are indicated in bold.

difference between Clear and casual speech, in agreement with observations from previous works. Similarly, the difference between the Clear and casual-warped speech is also significant. However, no significant categorical difference is found between the casual and casual-warped speech for either SNR. Thus, overall, there is essentially no intelligibility gain observed from the frequency warping for vowel space expansion.

4.3.3 Discussion

This work presented and evaluated an approach to expand vowel space via frequency warping inspired by clear speech analyses. Results indicate that, while vowel space is successfully expanded, there is no significant intelligibility gain from the frequency warping. The evaluation results can be explained from two perspectives. First, the lack of effectiveness of the frequency warping at increasing intelligibility could be due to the specific algorithm itself. That is, while successfully achieving vowel space expansion and altering voice characteristics without noticeable artifacts, the warping (by design) does not exactly replicate the expansion observed in clear speech. Perhaps there exists a more effective approach to expanding vowel space. For example, vowel hyperarticulation should be applied on tense rather than lax vowels. However no discrimination is performed in the proposed approach possibly leading to a perceptual mismatch if all vowels are enhanced. Second, considering clear speech, the observed vowel space expansion might be reflecting more important acoustic-phonetic modifications that occur on an increasingly detailed, spectro-temporally localized level. In other words, the observed vowel space expansion represents a static, average view of the articulation in the style that might not be highlighting the most pertinent cues. For example, the formant dynamics could be playing a significant role in enhancing the speech intelligibility and examination of these features in future work could prove fruitful. Therefore, a more localized level of analyses and consequent modifications (e.g. within phones and considering formant transitions) might expose more perceptually relevant differences that positively impact intelligibility.

Chapter 5

Spectral Transformations

In this chapter, the problem of modifying casual speech to reach the intelligibility level of clear speech is addressed via spectral transformations. In Chapter 2, we have performed acoustic analysis between clear and casual speech and we have observed different energy distributions in the spectral envelopes between the two speaking styles, with varying patterns across speakers. Examining the average relative spectra for all speakers of our subset, we have discovered an energy boosting in clear speech compared to casual speech on the upper midrange frequency region (2-4 kHz) and on the brilliance range (above 6kHz). We exploit these observations by introducing a simple method that boosts these frequency regions on casual speech. The proposed method, called Mix-filtering, uses a multi-band filtering scheme to isolate the information of these frequency bands and then, add this information to the original signal. In terms of intelligibility and quality, our method is compared to unmodified casual speech and to a highly intelligible spectral modification technique, namely the Spectral Shaping and Dynamic Range Compression (SSDRC). While Mix-filtering is "clear-inspired", SSDRC is "Lombard-inspired". Therefore, the first section of this chapter presents the SSDRC. This will give the reader an insight of what are the spectral differences between the two speaking styles, clear and Lombard, and what is the impact of performing clear-based and Lombard-based spectral transformations on the intelligibility and quality of casual speech.

Two different objective measures that are highly correlated with subjective intelligibility scores are used for estimating the intelligibility, whereas for evaluating the quality, preference listening tests are performed. Results show that the Mix-filtering technique increases the intelligibility of casual speech in SSN noise, while maintains its quality. On the other hand, while SSDRC outperforms on intelligibility, it degrades significantly the quality of casual speech.

The intelligibility benefit of the Mix-filtering method for SSN is also explored for reverberant environments. The Mix-filtering scheme is combined with the time-scaling techniques proposed in Chapter 3. Subjective evaluations by non-native, native and hearing impaired listeners are performed. Results reveal a significant benefit of our proposed modification for non-native listeners and suggest a connection between the amount of the time-scaling and the reverberation time for beneficial intelligibility enhancement in reverberation.

5.1 Related work: Lombard-like modifications - the SSDRC

Considering the intelligibility gains of the human speaking styles, acoustic phenomena observed in Lombard speech have be used to inspire speech signal modifications for intelligibility enhancement. Unlike clear speech, Lombard speech is produced when the speaker communicates inside a noisy environment. The adjustments that the speaker makes in order to be heard (by the listener and by himself) share some similarities with clear speech; Lombard speech shows decreased speaking rate, increased pitch, higher energy, spectral and vowel-to-consonant energy re-distribution, compared to its "normal" counterpart (Summers et al., 1988; Junqua, 1993; Garnier et al., 2006; Lu and Cooke, 2009). However, while both in clear speech and in Lombard speech there is an active reorganisation of articulation gestures, an increased oral pressure due to greater vocal effort is met in Lombard speech (and to a less extent in clear speech), characterizing Lombard as "tense" and "loud" speech compared to clear speech. Among these observations, the Lombard increase in intelligibility has been shown to be largely attributed to spectral modifications (Lu and Cooke, 2009), particularly increased spectral energy in an inclusive formant band or, otherwise stated, a decreased spectral "tilt". These spectral modifications have been exploited by the work of Zorila et al. (2012), yielding their algorithm, namely the SSDRC, the most successful modification from a challenge task, containing extensive evaluation of various intelligibility enhancement techniques (Cooke et al., 2013). Therefore, before performing spectral transformations based on clear speech, it is important to demonstrate how spectral energy is distributed in Lombard speech and how it differs from clear speech. Second, and equally important is to examine Lombard and clear speech in terms of intelligibility in order to compare the effectiveness of the two speaking styles. Finally, the SSDRC algorithm is presented thoroughly in this section, since it incorporates Lombard-inspired transformations and is used as comparative standard for the modifications examined in this work.

5.1.1 Lombard vs. clear speech

The most significant spectral trait attributed with the intelligibility gain of Lombard speech is a boosting of spectral energy in a frequency region spanning the range of formants, sometimes referred to as a flattening of the amplitude spectrum (Summers et al., 1988; Lu and Cooke, 2009; Godoy and Stylianou, 2012). Spectral energy boosting in given frequency bands or tilt changes have also been attributed with intelligibility gains of clear speech (Krause and Braida, 2004b; Amano-Kusumoto and Hosom, 2011). Differences in spectral energy distributions of clear speech with respect to its casual counterpart have been examined in Chapter 2 via the average relative amplitude spectra (Krause and Braida, 2004b; Godoy and Stylianou, 2012). The following analyses explicitly shows the differences in spectral energy distributions of Lombard speech compared to normal speech.

The Lombard (and normal) speech data is from the Grid corpora presented in Cooke et al. (2006); Lu and Cooke (2008, 2009). The sentences have a simple 6-word structure (e.g., "place red in G 9 soon"), as defined in the Grid multi-talker speech corpus (Cooke et al., 2006). Each sentence was read and recorded both in quiet conditions (normal) and while the speaker listened through headphones to SSN at a 96dB level

(Lombard). The corpus recording and processing is detailed in Lu and Cooke (2009). The Lombard speech corresponding to the highest noise level (i.e., Ninf96) in Lu and Cooke (2008, 2009) was selected so that the Lombard acoustic-phonetic characteristics would be most apparent. For the analyses in this work, 50 sentences per speaker, from 8 British English speakers (4 male, 4 female) are examined. The preprocessing of the speech corpora and the computation of the average relative amplitude spectra of Lombard speech vs. normal speech is performed following the methodology described in Chapter 2 for clear and casual speech.

The average relative spectra for the Lombard-normal case are shown in Figure 5.1 and Figure 5.2, for individual speakers as well as the overall average for the respective corpora. For the Lombard case, a variability among speakers is also observed as in clear speech (Figure 2.10). However, there is evident a more consistent trend of boosting the energy near 500-4500 Hz region, comprising the broad frequency range in which formants are located. This is observable on average in Figure 5.2. Nonetheless, as it is evident in Figure 5.2 and Figure 2.11, different frequency regions are emphasized for the clear and Lombard speaking style. Clear speech appears to give emphasis in upper midrange (2000-4800 Hz) and brilliance range (5600-8000 Hz) frequencies while Lombard speech boosts midrange (500-2000Hz) and upper midrange frequencies (2000-4000Hz).

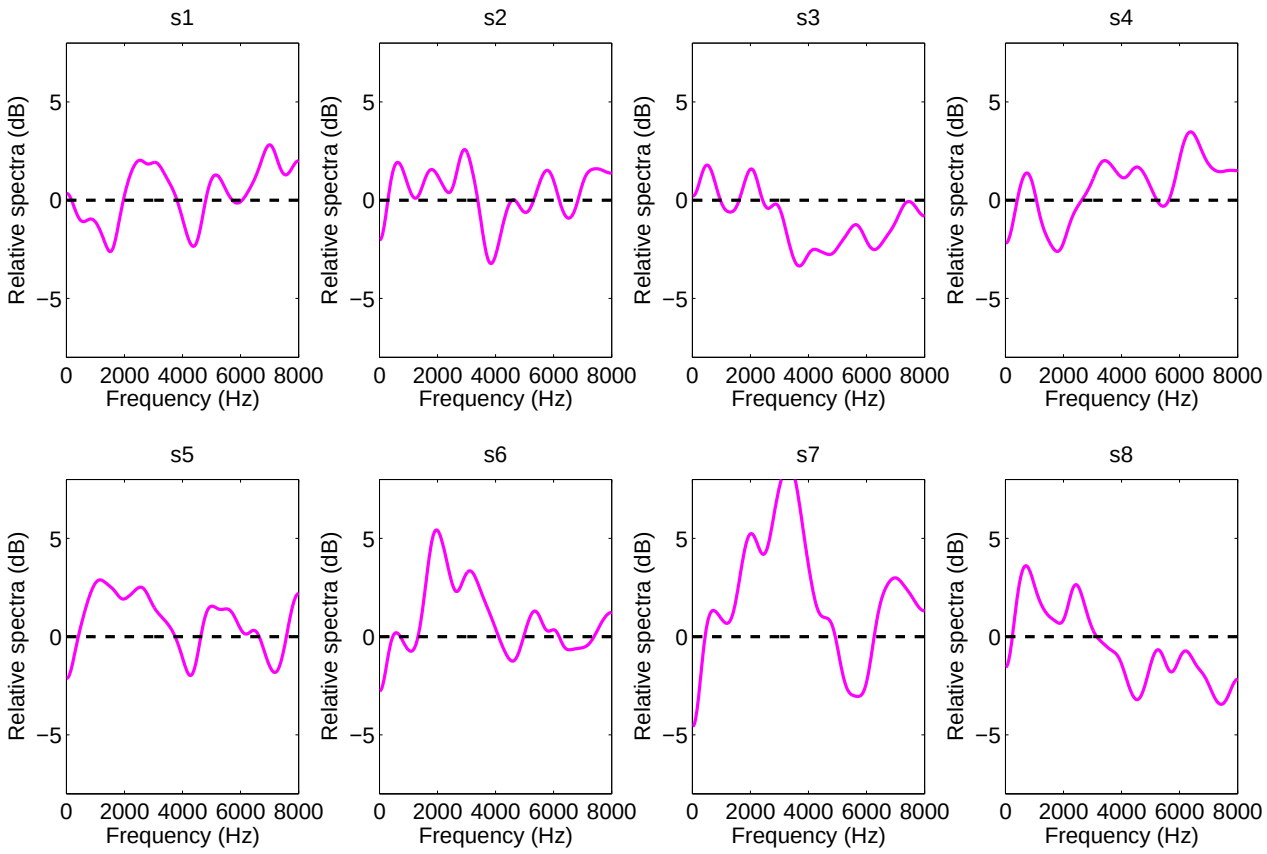


Figure 5.1: Relative spectra for 8 speakers on Grid database

Examining the efficiency of the speaking styles, clear and Lombard, in terms of intelligibility, formal

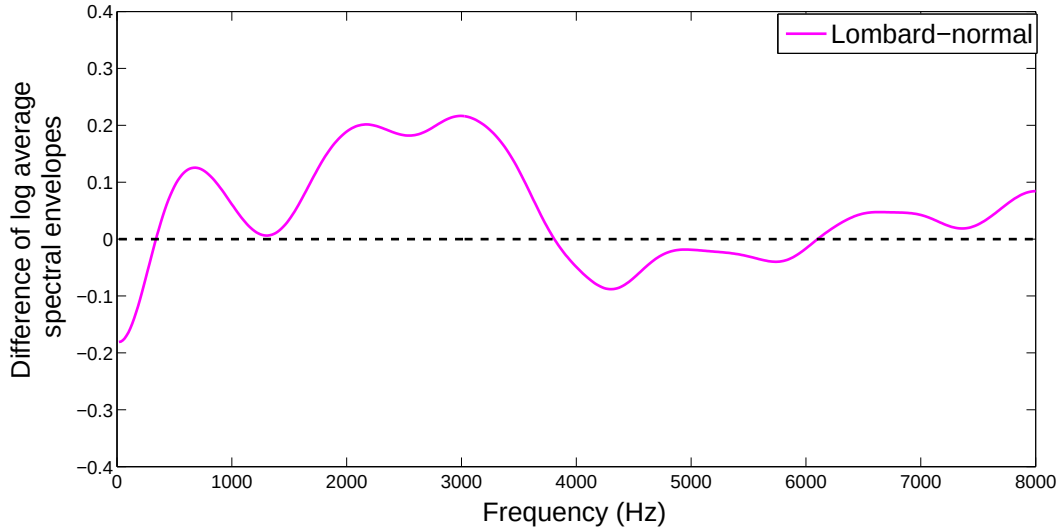


Figure 5.2: Relative spectra: average spectral envelope of Lombard minus normal in log scale

listening tests[1] were conducted by native listeners. Specifically, 20 native English-speakers heard Grid and LUCID sentences, with SSN added at two levels to yield 0 ("high") and -4 ("low") dB SNR, respectively. The normal, casual, clear and Lombard sentences were included in the test. For each listener, the test was split into two parts, involving the Grid and LUCID sentences, respectively. The order of these parts was randomized. Additionally, the sentences within each part were randomly ordered and the test was designed with the goal that speakers and conditions be equally represented (i.e., appear approximately the same number of times across the tests). When taking the test, listeners were asked to type what they think they heard after hearing each sentence once.

It should be noted that the variability among listeners was quite high. Nonetheless, the average trends in intelligibility scores are summarized in Table 5.1 and are depicted in Figure 5.3. In order to evaluate the statistical significance of these results, ANOVA tests were performed. Specifically, the ANOVA null hypothesis (i.e.,the average values of the intelligibility scores for every condition are equal) was rejected using the F-test. Then, pairwise comparisons of the averages were performed using Fisher's least significant difference (LSD) test, with a confidence interval of 95%, in order to derive which of the groups differ significantly. The standardized difference between condition pairs is provided in Table 5.2.

Comparing the Lombard and clear speech corpora, Figure 5.3 shows the intelligibility scores (i.e., percent of keywords correctly identified) from the listeners for each condition, at the two SNR levels. The first and last columns in Table 5.1 similarly indicate the percent of overall correct keywords identified for these conditions. It should be noted that all conditions shown in Figure 5.3, except Lombard-normal at 0dB SNR, were found to be statistically significant, as shown in Table 5.2. In examining Figure 5.3 and Table 5.1, several trends are observed for the Lombard and clear speech corpora. First, comparing the clear and casual speech scores, there is a +31% and +29% gain in intelligibility of clear speech over casual for low and high SNR, respectively.

[1]Thank you to Catherine Mayo and CSTR at the University of Edinburgh for their help administering the listening tests

SNR	-4				0			
Style	clear	Lombard	casual	normal	clear	Lombard	casual	normal
Score	46	50	15	33	79	67	50	62

Table 5.1: Overall percent of correct keyword identification for clear/Lombard, casual/normal speech.

Intelligibility Difference			
-4 SNR		0 SNR	
clear-casual	**5.66**	clear-casual	**5.14**
Lombard-normal	**2.03**	Lombard-normal	0.709

Table 5.2: Results of significant different analysis between conditions for the Grid and LUCID corpora. The standardized difference is given for pairing between clear-casual, Lombard-normal. Significant differences are in bold.

Thus, the intelligibility gain of clear speech over casual is consistent across both SNR. On the other hand, in comparing the gain of Lombard over normal speech, +17% and +5% are observed for low and high SNR, respectively. Thus, the intelligibility advantage of Lombard speech over normal speech is noticeably reduced for high SNR, e.g., situations in which the noise level is low. This observation is in line with related work in Lecumberri (2012) suggesting that Lombard speech does not necessarily provide an intelligibility advantage at high SNR. The intelligibility scores would thus suggest that clear speech is always helping to make speech more intelligible, however, Lombard speech is only advantageous when hearing speech in noisy conditions.

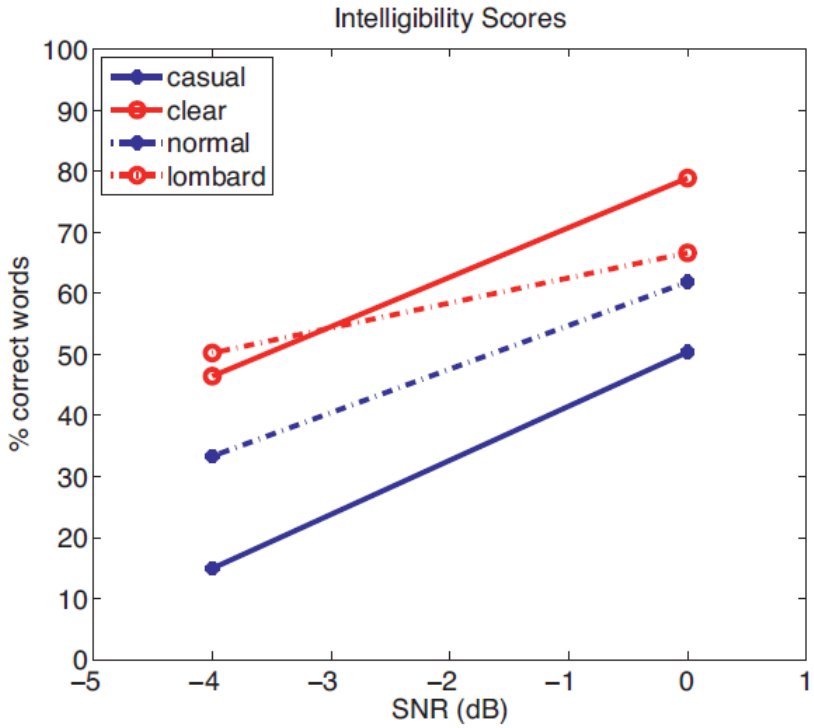


Figure 5.3: Intelligibility scores: clear/Lombard, casual/normal

In order to draw concrete conclusions regarding the clear, casual and Lombard speech benefit, the corpus of the above natural speech modifications should be recorded with the same utterances and listeners in different

simulating conditions or instructions in order to produce the corresponding speaking style. Therefore, the above observations from the listening test should be treated with consideration, since the Lombard speech corpora differ from the clear speech corpora in the amount of contextual information, on the recorded speakers etc. For example, considering the "standard" conditions, the normal speech from Grid was judged to be more intelligible than the casual speech from LUCID. One likely explanation for this intelligibility difference is the sentence structure used in the corpora. Specifically, the Grid corpus has a constrained pattern and identifiable structure. Consequently, it is easier to guess the keywords once some highly intelligible speech examples have been heard. For example, the structure of the Grid sentences "Bin blue at G 6 again" and "Place red by Z six now" are more similar than the LUCID sentence examples "Wasps and bees are part of summer" and "Jonathan gave his wife a bush". Moreover, the British accents of the speakers in the corpora are different and this could also play a role in the intelligibility scores. Given these observations, the results comparing the different corpora are accordingly tempered.

Nonetheless, the results of the subjective tests cannot be considered misleading. And this is due to the fact that even though the comparison between Lombard and clear speech cannot be made directly, the intelligibility distances between normal and Lombard compared to the intelligibility differences between clear and casual speech differ from -4 to 0dB SNR. While the intelligibility scores of Lombard and normal speech approach in 0dB SNR, clear and casual speech maintain their intelligibility distance compared to -4dB SNR. The intelligibility gap difference between Lombard-normal and clear-casual from -4 to 0dB SNR is such that it cannot be convincingly supported by differences in the speech corpus. Moreover, these results are consistent with observations made in related works, in particular concerning the intelligibility gains of Lombard and clear speech at low and high SNR. Therefore, these results support our initial motivation to enhance speech intelligibility based on clear speech properties in order to create a desirable signal both in terms of quality and intelligibility.

5.1.2 Spectral Shaping and Dynamic Range Compression, SSDRC

Previous work on speech intelligibility enhancement has incorporated spectral modifications based on the Lombard speaking style. The Spectral Shaping and Dynamic Range Compression (SSDRC) proposed by Zorila et al. (2012) uses a Lombard-inspired fixed spectral gain filter, in order to increase loudness and mimic Lombard speech. The corrective filter is shown in Figure 5.4. Other corrective filters have been proposed to modify normal speech (Godoy and Stylianou, 2012; Lu and Cooke, 2009), following the main shape of the overall average Lombard-normal relative spectra shown in Figure 5.2. However, the fixed filter from SS ($Hr(f)$), described in Zorila et al. (2012) and depicted in Figure 5.4 has been proven the most effective in an extensive evaluation of speech intelligibility enhancement modifications (Cooke et al., 2013).

Figure 5.5 depicts the main steps performed by SSDRC. In addition to the fixed filter $H_r(f)$ described above, the SS described in Zorila et al. (2012) and evaluated in Cooke et al. (2013) also incorporates adaptive components. The adaptive spectral shaping takes into account the probability of voicing given a speech frame, while the fixed spectral shaping is independent of the probability of voicing. The adaptive spectral shaping

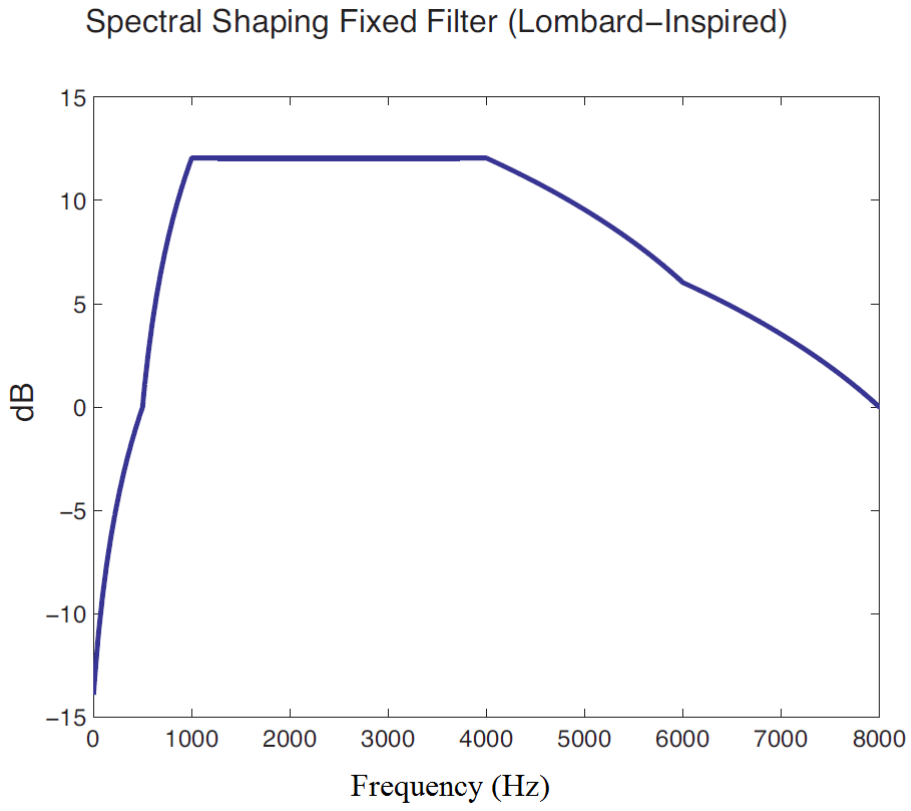


Figure 5.4: The SS fixed filter $H_r(f)$ from Zorila et al. (2012)

consists of (i) adaptive sharpening where the formant information is enhanced ($H_s(f)$), and (ii) an adaptive pre-emphasis filter ($H_p(f)$).

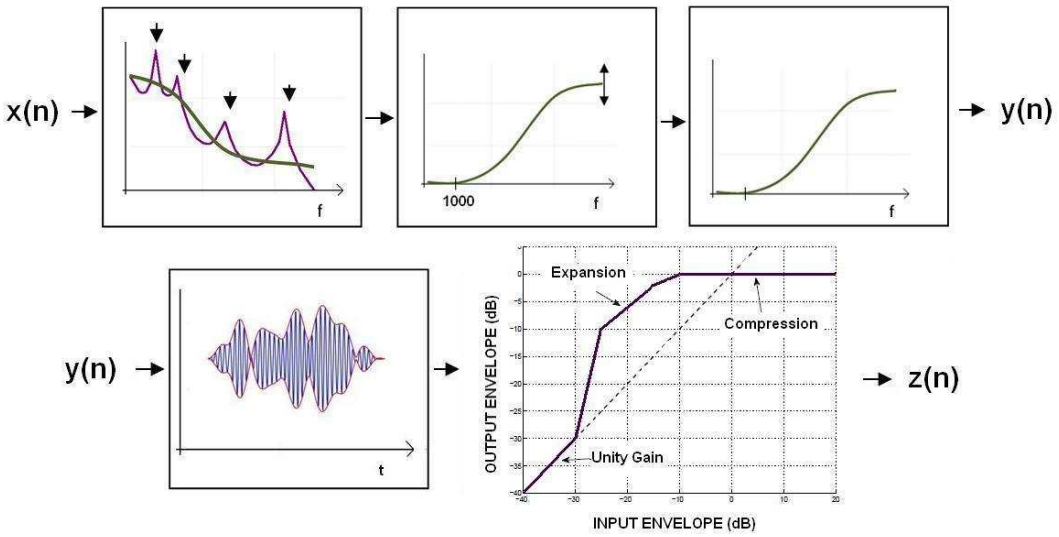


Figure 5.5: The SSDRC from Zorila et al. (2012)

The output of the Spectral Shaping system is the input to the Dynamic Range Compressor, namely the

DRC (Blesser, 1969; Quatieri and McAulay, 1991). DRC has a dynamic and a static stage. During the dynamic stage, the envelope of the signal is dynamically compressed with $2ms$ release time constant and almost instantaneous attack time constant. The signal envelope is based on the Hilbert transform and a moving average operator with order determined by the average pitch of speakers gender. After the dynamic compression of the signal envelope, a static amplitude compression is applied. During the static amplitude compression, the 0 dB reference level is a key element in forming the Input/Output Envelope Characteristics (IOEC). For the current system this was set to 0.3 of the peak of the signal. The whole system is based on a frame-by-frame analysis and synthesis. In each frame the magnitude spectrum is computed using FFT and then manipulated in the way mentioned above. Overlap and add is then used to reconstruct the modified signal. The whole process is very fast and can run in real time.

5.2 Clear-inspired spectral modifications: the Mix-filtering

In this section, the idea of applying a corrective filter similar to the SS fixed filter is explored in order to enhance the intelligibility of casual speech. However, the corrective filter proposed follows the shape of the overall average clear-casual relative spectra depicted in Figure 2.11. The authors expect that transforming the spectral content of casual speech similar to that of clear speech, the intelligibility gain of casual speech will increase, while its quality will remain intact.

The method proposed for the intelligibility enhancement of casual speech is simple and is based on the analysis described in Chapter 2. Specifically, from our dataset (4 female speakers and 4 male speakers from the LUCID database), 60 sentences were randomly selected per speaker and per speaking style. Then, this dataset was split in two parts. The first part contained 20 sentences out of 60 and was used as an analysis dataset (dataset A). Then, the second part (dataset B) was used as an evaluation dataset and contained 40 sentences per speaker but only for the casual speaking style. The intersection of the two datasets A and B was null.

The analysis performed and described on Chapter 2 on the dataset A reveals spectral differences on two frequency bands between clear and casual. Figure 2.11 shows the difference of the log average spectral envelopes of clear speech minus casual speech. Positive difference suggests that the energy of clear speech is higher than that of the casual speech. As we can see, clear speech appears to have higher energy in two frequency bands, $B_1 = [2000, 4800]$ and $B_2 = [5600, 8000]$. The method proposed in this work for enhancing the intelligibility of casual speech involves the isolation of these important frequency bands B_1 and B_2 and then the addition of their energy to the original signal with different weighting factors for each frequency band. Hopefully, this addition will boost the important frequency regions on casual speech, as it naturally happens in clear speech.

For the isolation of the frequency bands a simple method is used. Casual speech s is filtered with a 5-order bandpass digital elliptic filter with $0.1dB$ of ripple in the passband, and $60dB$ ripple in the stopband and bandpass edge frequencies $[2000, 4800]$. An IIR filter is selected for the frequency band isolation in order to have an abrupt transition from passband to stopband. No phase adjustment is performed. However,

distortions due to the non-uniform group-delays are not perceptually noticeable as will be reported later by the quality evaluations. The output of the filter is signal s_1 which contains information on the B_1 frequency band. Moreover, casual speech s is filtered with a 5-order highpass digital elliptic filter with normalized passband edge frequency $f_c = 5600Hz$. The output of this filter is the signal s_2 which contains information on the frequency band B_2. Then, the original signal s and the filtered signals s_1 and s_2 are combined with different weighting factors to form the modified signal y, which is normalized to have the same RMS energy as original speech:

$$y[i] = w_0 s[i] + w_1 s_1[i] + w_2 s_2[i] \tag{5.1}$$

$$y_{mixF}[i] = y[i] \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} s^2[i]}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} y^2[i]}} \tag{5.2}$$

where, y_{mixF} is the proposed modified signal, N is the number of samples of the casual signal s and y, and w_0, w_1, w_2 are the weighting factors of the signals s, s_1 and s_2, respectively.

The selection of the proper combination of the weights is important both for intelligibility and quality. In Niederjohn and Grotelueschen (1976) it has been shown that high pass filtering speech above 1.5kHz increases its intelligibility in noise. However, the absence of information on lower frequency bands can degrade the quality of speech. Therefore, this information is contained on the modified speech y_{mixF} by choosing to keep the original speech signal weighted by w_0. Then, the selection of the other two weights is inspired by clear speech properties. Specifically, focusing on the energy differences between clear and casual speech, it can be observed from Figure 2.11 that the energy in B_2 frequency band is greater than that of B_1. Possibly, this energy difference is attributed to a consonant emphasis that possibly happens in clear speech. Therefore, we choose $w_2 > w_1$ to account for the slight higher energy difference of B_2 frequency band compared to B_1 between the two speaking styles.

Summarizing the above, the set of the possible weighting combinations can be described by the following equations:

$$w_0 = 1 - \sum_{i=1}^{2} w_i \tag{5.3}$$

$$w_2 > w_1 \tag{5.4}$$

$$w_i \neq 0, i = 0, 1, 2 \tag{5.5}$$

In order to select the proper weight combination $\{w_0, w_1, w_2\}$ we consider w_0 as a dependent variable. Then, the two variables w_1, w_2 can vary between (0, 1) respecting the restrictions described by equations (5.3), (5.4) and (5.5). As we are interested on enhancing the intelligibility of casual speech, the proper values $\{w_0, w_1, w_2\}$ are those that maximize the intelligibility score of modified speech compared to unmodified speech. To define these values, the casual speech of dataset A is used as a training dataset. Specifically, the

casual signals of dataset A are modified using different weight combinations that satisfy the above equations. The intelligibility of the modified sentences using the Mix-filtering approach (mixF) and the unmodified casual sentences is evaluated objectively in the presence of SSN at -10dB SNR. The best combination of weights is the one that maximizes the objective intelligibility difference of the modified speech minus the unmodified speech.

The objective metric used to predict intelligibility is the Glimpse Proportion (GP) proposed by M.Cooke (2006); Tang and Cooke (2011b). The Glimpse measure comes from the Glimpse model for auditory processing. As an intelligibility predictor, the model is based on the assumption that in a noisy environment humans listen to the glimpses of speech that are less masked. Therefore, the GP measure is the proportion of spectral-temporal regions where speech is more energetic than the noise.

Figure 5.6 shows for various weight combinations the difference on the GP intelligibility scores of unmodified casual speech from mixF modified casual speech. Note, that w_0 is not present as it is assumed from equation (5.3) to be the dependent variable. The optimal weight combination that maximizes this difference is $\{0.1, 0.4, 0.5\}$. Figure 5.7 depicts the relative spectra of mixF modified speech and casual speech. Specifically, the average spectral envelope of casual speech is subtracted from the corresponding spectral envelope of mixF modified casual speech with the optimal weight combination. As we can see from Figure 5.7, the important frequency bands are boosted "stealing" from the lower frequency bands.

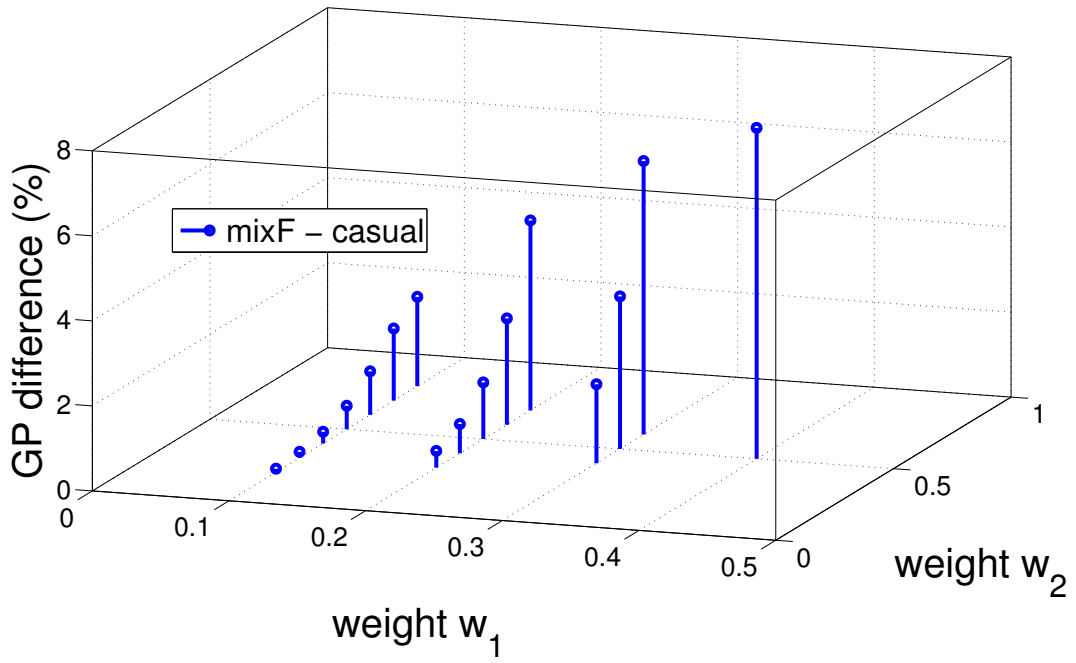


Figure 5.6: % difference of GP scores between modified mix-filtered speech (mixF) minus unmodified casual speech. MixF is derived using various weights combinations that verify equations (5.3), (5.4), (5.5). The maximum difference is 7.78% *and corresponds to* $\{w_0, w_1, w_2\} = \{0.1, 0.4, 0.5\}$.

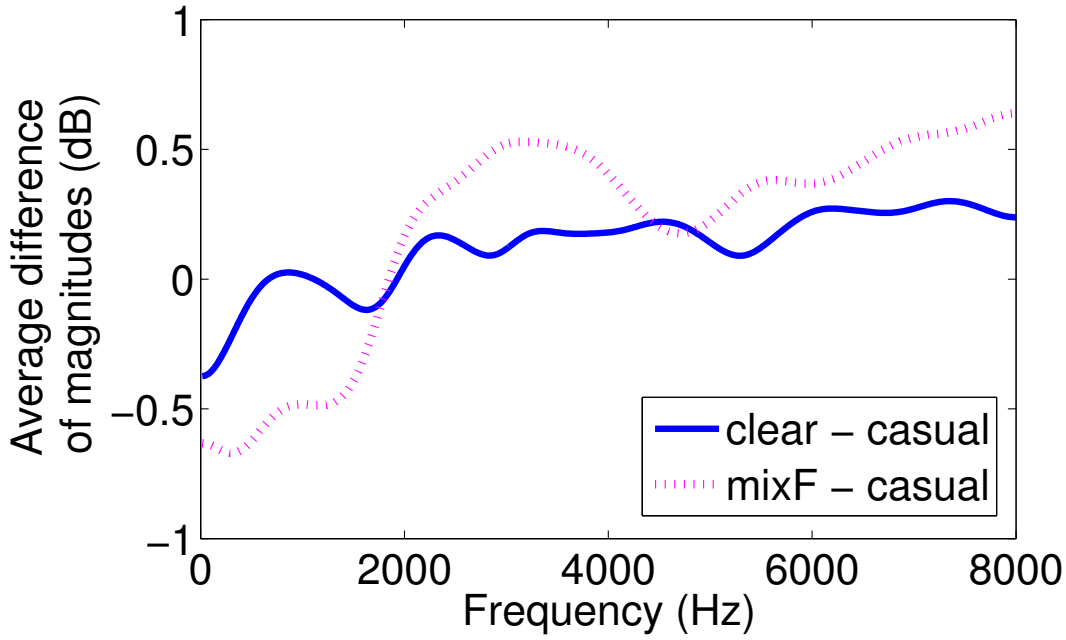


Figure 5.7: Difference of the log average spectral envelopes.

5.2.1 Evaluations on intelligibility and quality

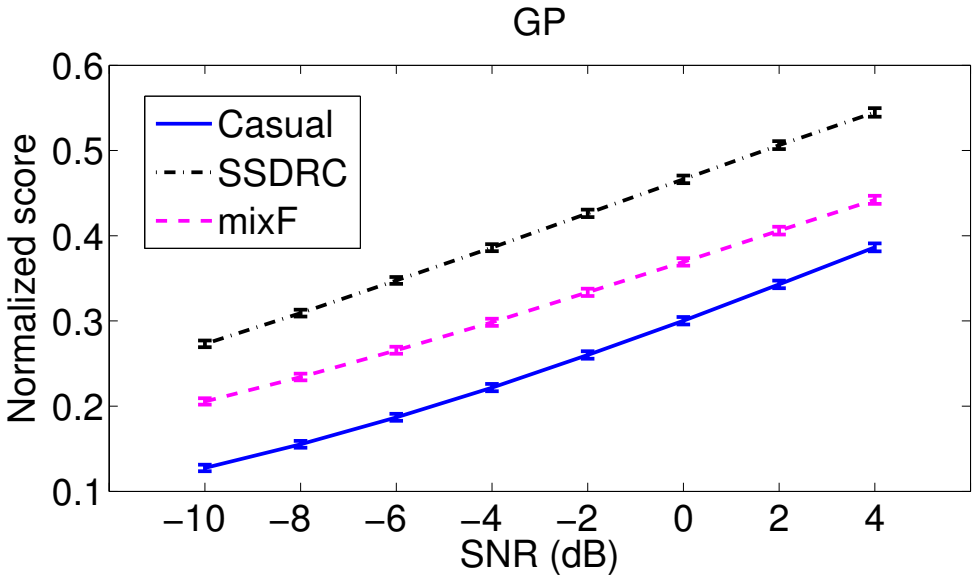
The modified casual speech derived from the Mix-filtering approach is compared in terms of intelligibility and quality with unmodified casual speech and with SSDRC modified casual speech. The evaluation of Mix-filtering approach in terms of intelligibility is done using two different objective measures, the GP measure described above (M.Cooke, 2006; Tang and Cooke, 2011b) and the Distortion-Weighted Glimpse Proportion (DWGP) (Tang et al., 2013). DWGP [2] has been shown to have a better correlation with subjective intelligibility evaluations than GP (Tang et al., 2013). The DWGP measure computes the correlation between frequency bands of clean speech and speech in noise, weighting these correlations according to the importance of each frequency band. The prediction of intelligibility is estimated by the correlation which gives a measure of how much noise affects the signal. Then, for the evaluation of the quality of speech a preference test is made between three different speech signals.

Objective evaluations of intelligibility

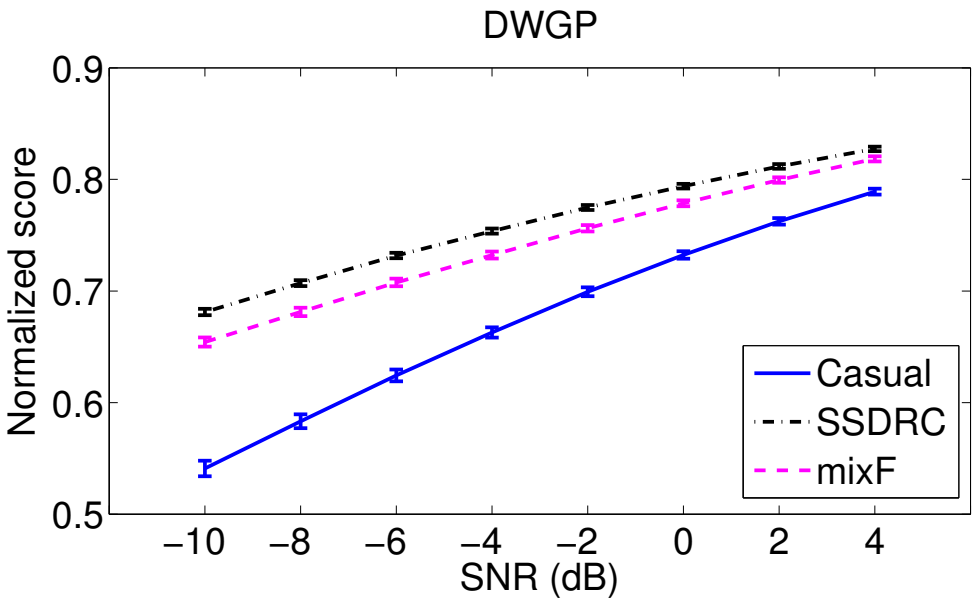
For assessing the intelligibility impact of the Mix-filtering approach objective evaluations of intelligibility were performed on the testing dataset B. The sentences of this dataset are modified using SSDRC and Mix-filtering with the optimal weight combination. Then, GP and DWGP scores are extracted for the three categories of speech, casual speech, mixF and SSDRC modified speech. SSN of various SNR levels is used for evaluating objectively the intelligibility of each category in noise. Figures 5.8(a) and 5.8(b) depict the objective scores predicted by GP and DWGP respectively for SNR levels varying from -10 to 4 dB. GP reports that the SSDRC outperforms in terms of intelligibility while our proposed scheme increases the intelligibility

[2]The authors would like to thank Dr. Yan Tang and Prof. Martin Cooke from the Ikerbasque research center for providing the objective intelligibility score DWGP

of casual speech by 8% on low SNR. On the other hand, DWGP predicts that the intelligibility advantage of our proposed method is more than 10% on casual speech on low SNR (-10 dB) , approaching the intelligibility scores of SSDRC. Overall both objective scores predict an intelligibility increase of our proposed scheme for every SNR, with DWGP reporting intelligibility levels of mixF close to those of SSDRC.



(a)



(b)

Figure 5.8: Objective scores for predicting intelligibility of each speech category in speech-shaped noise: mean values and 95% confidence intervals.

Subjective evaluations on quality

For evaluating the quality of our method, mixF, casual and SSDRC are compared in terms of quality using preference listening tests that have been conducted without the presence of noise. 10 random distinct sentences from dataset B were presented to 18 listeners. Each sentence was modified by SSDRC and mixF and was heard 6 times, two times for each pair {casual-mixF, mixF-SSDRC, SSDRC-casual}. Listeners had to select from -3 to 3 the degree of preference between those pairs in terms of quality with 0 corresponding to the same quality and 3 (-3) to the much better (worse) quality of the one signal compared to the other. Despite the fact that the energy of the signal was the same for the three categories, the loudnesses was higher for SSDRC and mixF. Therefore, in order to avoid loudness differences that could influence the perceptual judgement of the listeners, all signals were normalized in loudness using ACTIVLEV (ITU-T P.56).

Figure 5.9 summarizes the scores of preference of each category against the two others. Confidence intervals are also provided. As we can see, casual and mixF appear to have similar scores of preference whereas SSDRC gives negative quality scores against the other two categories, casual and mixF. The proposed Mix-filtering approach preserves the quality of casual speech.

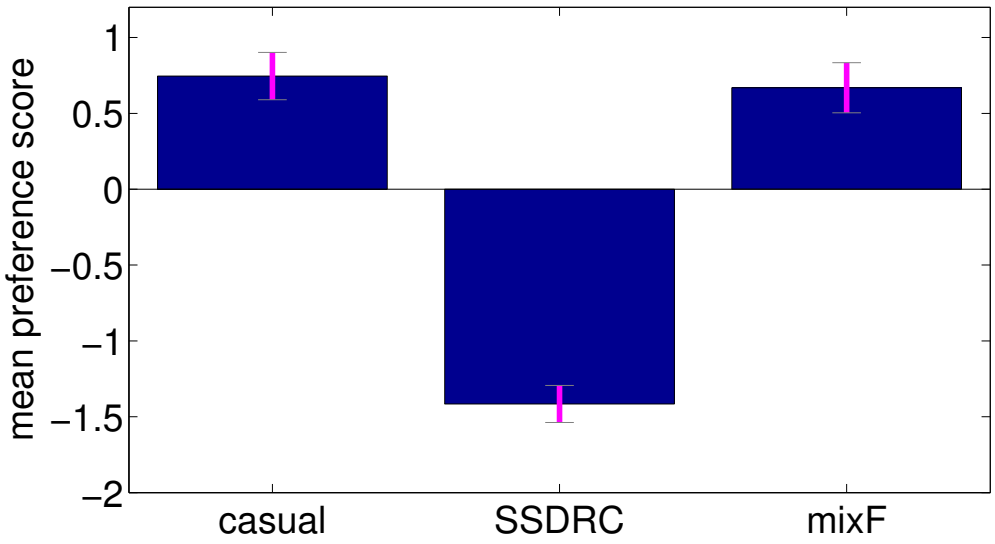


Figure 5.9: Subjective quality evaluation: mean values and 95% confidence intervals of the preference scores of each category against the two others.

5.2.2 Discussion

Focusing on the spectral domain, one feature that is possibly associated with the intelligibility of clear speech is an energy increase above 1000Hz (Krause and Braida, 2004a; Hazan and Baker, 2010). This increase of energy compared to casual speech occurs also in other speaking styles, like on Lombard speech, in similar frequency regions. It has been shown that performing Lombard-like modifications on plain speech by boosting the frequency region $1 - 4$kHz while maintaining the overall RMS energy of the signal, has an

intelligibility increase (Godoy and Stylianou, 2012). In Krause and Braida (2009) a similar approach has been used for clear speech, amplifying the energy around F2 and F3 formants on casual speech on voiced segments; intelligibility tests for normal hearing listeners in noise (SNR=-1.8dB) showed that modified speech was more intelligible than unmodified casual speech and less intelligible than clear speech. In addition, other simpler spectral modifications can increase speech intelligibility. Performing high-pass filtering on speech with cut-off frequency 1.5 kHz increases its intelligibility in noise (Niederjohn and Grotelueschen, 1976).

The aforementioned studies report that spectral modifications of casual speech may be proven beneficial for its intelligibility. However, none of the studies is concerned with the quality degradations imposed to original speech. In Chapter 3 we have seen that SSDRC can increase the intelligibility of casual speech to levels higher than that of clear speech on low SNR. However, the quality of modified speech is quite degraded, as reported in this Chapter. The majority of the studies that examine speech intelligibility, test the speech signals in noise, masking all the artifacts that may be introduced on processed speech. Even if it is preferable to test the intelligibility of speech in noise, as normal-hearing subjects can be used for evaluations, speech is not always intended in noise. On some applications it is important to preserve the quality of speech (e.g applications for hearing impaired listeners).

This work tries to address the problem of increasing the intelligibility of casual speech while maintaining its quality. Motivated by previous studies that achieve to increase intelligibility using spectral modifications, this study also modifies the spectral characteristics of casual signals imposing however, quality restrictions. The proposed method, inspired by the properties of clear speech, amplifies the energy of specific frequency bands of original casual speech. The advantage of the method is its simplicity and efficiency. Firstly, unlike other techniques (Krause and Braida, 2009; Zorila et al., 2012) it does not require frame-based analysis and modifications (detection of voiced/unvoiced regions, formant shaping, maximum voice frequency estimation etc). On the contrary, it isolates frequency bands on casual speech by simply performing multi-band filtering and then adds back to the initial signal the filter outputs. Secondly, results show that the proposed modified scheme increases the intelligibility of casual speech while maintains its quality. As the proposed method is less intrusive, the intelligibility benefit is less compared to SSDRC. However, unlike SSDRC, the Mix-filtering approach does not degrade the speech quality, as reported by subjective quality tests. Last, the Mix-filtering approach is speaker and sentence independent and can be applied to any speech signal.

5.3 Combining the Mix-filtering approach with time-scaling: application to reverberation

In the previous section the Mix-filtering approach has been shown to have an intelligibility increase over unmodified speech in Speech-shaped noise. Speech-shaped noise is used extensively as a masker in studies of speech perception (Nelson et al., 2003; Qin and Oxenham, 2003) since it is simulates the "multi-speaker" babble effect which is considered to be the most common masker in real life scenarios. However, there are other difficult environments where casual speech has significantly less intelligibility than clear speech (Payton

et al., 1994). Therefore, possible benefits of clear-speech-inspired modifications in such environments should be also explored. Reverberation is one such environment where the masker is actually a delayed version of the signal due to its reflection inside a room.

The intelligibility benefit of the Mix-filtering method for SSN is explored here for reverberant environments. The motivation for proposing this technique for enhancing speech intelligibility in reverberation is that the mix-filtered modified speech simulates clear speech in terms of spectral energy distribution, which is resistant to reverberant environments (Payton et al., 1994). In such environments, the intelligibility decrease of speech is due to (1) overlap masking effect where the energy of a phoneme is masked by the preceding one (Nabelek et al., 1989) (2) self-masking where the information is smeared inside a phoneme possibly as a result of flattened formant transitions (Nabelek et al., 1989). As in clear speech, the Mix-filtering approach boosts higher spectral regions, where transient parts are more likely to be found, and "steals" spectral energy from low-frequency energy which usually causes the overlap masking effect on the energy of a preceding phoneme. Other studies that successfully address the problem of intelligibility degradation on reverberant environments use steady-state suppression techniques to reduce steady-state portions of speech like vowel nuclei and to increase transient information (Nabelek et al., 1989; Arai et al., 2002; Hodoshima et al., 2006). Mix-filtering achieves, with less complexity, a similar acoustic result as steady state suppression and consonant emphasis since it does not require classification of speech portions.

The combination of the Mix-filtering spectral technique along with time-scaling is explored, since spectral and time-scaling transformations, either natural (clear speech) or synthetic (Arai, 2005; Arai et al., 2007), have been proven advantageous for speech intelligibility in reverberation. Time-scaling schemes may enhance the intelligibility of unmodified speech through repetition of the information in time, reducing the overlap-masking and self-masking effect. The performance of two time-scaling techniques is evaluated for reverberant environments: 1) Uniform time-scaling 2) Time-scaling based on the Perceptual Quality Measure (PSQ) model. Both time-scaling schemes have been already described in Chapter 3. Uniform time-scaling changes the overall duration while respects the "local" speech rhythm. PSQ proposes both an elongation and a pause insertion scheme that could be beneficial inside reverberant environments, as the energy of a speech segment falls into pauses and does not mask following segments. Unlike other proposed pause insertion schemes that are used for reverberation (Arai, 2005), this work explores a pause insertion scheme that inserts pauses in acoustically meaningful places.

Subjective evaluation of modified and unmodified speech is carried out via intelligibility tests performed by non-native, native and hearing-impaired listeners on two reverberation times. Unlike other studies that use a carrier sentence and non-sense syllables or rhyming words (Nakata et al., 2006; Arai, 2005; Arai et al., 2007) to test word intelligibility, a more realistic scenario is used by testing sentence intelligibility, using the LUCID corpora.

5.3.1 Evaluations

In this section the proposed modifications are evaluated in reverberant conditions. Reverberation is simulated using a room impulse response (RIR) model obtained with the source-image method (Allen and Berkley, 1979). The hall dimensions are fixed to $20\ \mathrm{m} \times 30\ \mathrm{m} \times 8\ \mathrm{m}$. The speaker and listener locations used for RIR generation are $\{10\ \mathrm{m},\ 5\ \mathrm{m},\ 3\ \mathrm{m}\}$ and $\{10\ \mathrm{m},\ 25\ \mathrm{m},\ 1.8\ \mathrm{m}\}$ respectively. The propagation delay and attenuation are normalized to the direct sound. Effectively, the direct sound is equivalent to the sound output from the speaker. Convolving the modified speech signals with RIR produces the signals for evaluation.

Seven sets of signals are evaluated: (1) the clear speech (CL), (2) the casual speech (CV) (3) the Mix-filtering spectrally modified casual speech signal (M) (4) the uniformly time-scaled casual speech signal (U) (5) the PSQ-based time-scaled casual speech signal (P) and the combinations of the above modifications (6) uniform time-scaling and Mix-filtering of casual speech (UM) (7) PSQ-based time-scaling and Mix-filtering of casual speech (PM). The term Categories will be used to refer to the seven sets of signals. 56 randomly selected distinct sentences from the LUCID corpus are presented to the listeners, uttered from 2 Male and 2 Female speakers (14 sentences per speaker, 8 sentences per set of signals, 4 sentences per Category per reverberant condition). The reverberation times are $RT_1 = 0.8s$ and $RT_2 = 2s$ to simulate low and high reverberant environments, respectively. The listener heard each sentence once and was instructed to write down whatever he/she perceives to have heard.

As sentence difficulty may affect the intelligibility scores (especially for the non-native population), 7 different listening scenarios have been created to ensure that each sentence will be presented in a $\{$CL, CV, M, U, P, UM, PM$\}$ manner to different listeners (as each listener cannot hear the same sentence twice). For example, if a specific sentence is presented to the listener in CL manner on RT_1 condition on the listening Scenario 1, then the same sentence will be presented to another listener in CV manner on the same reverberant condition on listening Scenario 2 etc. This allows us to "denoise" the performance evaluation from the sentence dependency.

32 listeners participated in the intelligibility test, 7 native speakers, 4 hearing-impaired listeners, and 21 non-native speakers. The intelligibility test was performed online and was split in three parts. The first part contained information of what is an intelligibility test and what is the listener's contribution to the test and its oriented applications. In the second part, the listeners had to answer to some questions regarding their origin (whether they were native British English speakers or not) and their hearing ability (whether they had a hearing-impaired problem or not according to their knowledge). It should be noted here that for the hearing-impaired population we have no specification about the degree and type of hearing loss. Moreover, in this part detailed instructions were given to the listeners of how to perform the listening test. Specifically, the listeners were asked to evaluate the sentences only with the use of headphones in a quiet environment and to write down what they understood. The online listening test was designed in such a way that did not allow the listeners to hear each sentence more than once. In order to control the concentration level and the good perception of English especially for the non-native listeners, 5 difficult sentences were presented in the second

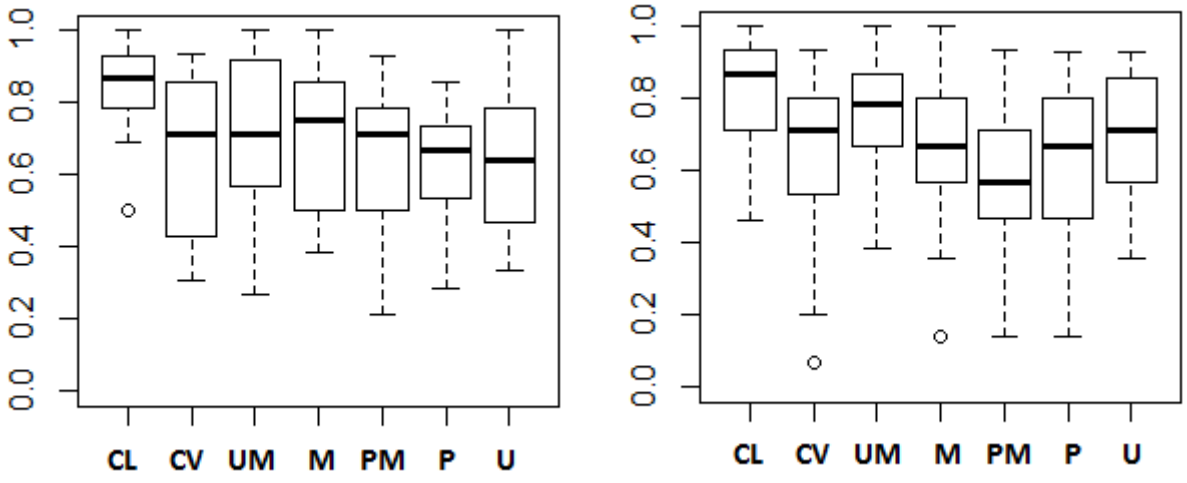
part without reverberation conditions. Then, the listeners could proceed to the third part which contained the main intelligibility test. The instructions of how to perform the test were also contained in the third part. A "header" of 4 sentences was added to the listening test to serve as a preparation set for the listeners to the reverberant environment (these sentences were not evaluated).

Statistical analysis of the intelligibility scores derived from the listening test is performed for each Category across listeners for all populations (native, non-native and hearing impaired). As the majority of the listeners are non-native, explicit statistic analysis is performed for this population, presenting statistical significance of the intelligibility scores across sentences and different reverberant conditions.

Non-native speakers

Performance evaluation for the non-native speakers contains three parts of statistical analysis. The first part presents the intelligibility scores of each Category across listeners, in order to reveal possible intelligibility benefits of the proposed modifications for the non-native population. The second part of analysis computes the intelligibility scores of each Category across sentences, to parcel out the possible variability due to sentence difficulty and reveal the Category main effect. Lastly, the third part of analysis presents the intelligibility scores of each Category across the two different reverberant conditions.

For each reverberant condition, the ratio of the correct key-words to the number of total keywords per sentence is estimated per listener and per Category. Then, the mean of the ratios for all sentences is estimated per listener and per Category. Figure 5.10 shows the $\{$min, 1^{st} quartile, median, 3^{rd} quartile, max$\}$ of intelligibility scores per Category across all listeners. CL appear to have a higher intelligibility advantage over all Categories for both reverberant conditions while the UM has a benefit over CV on RT_2 (Figure 5.10).



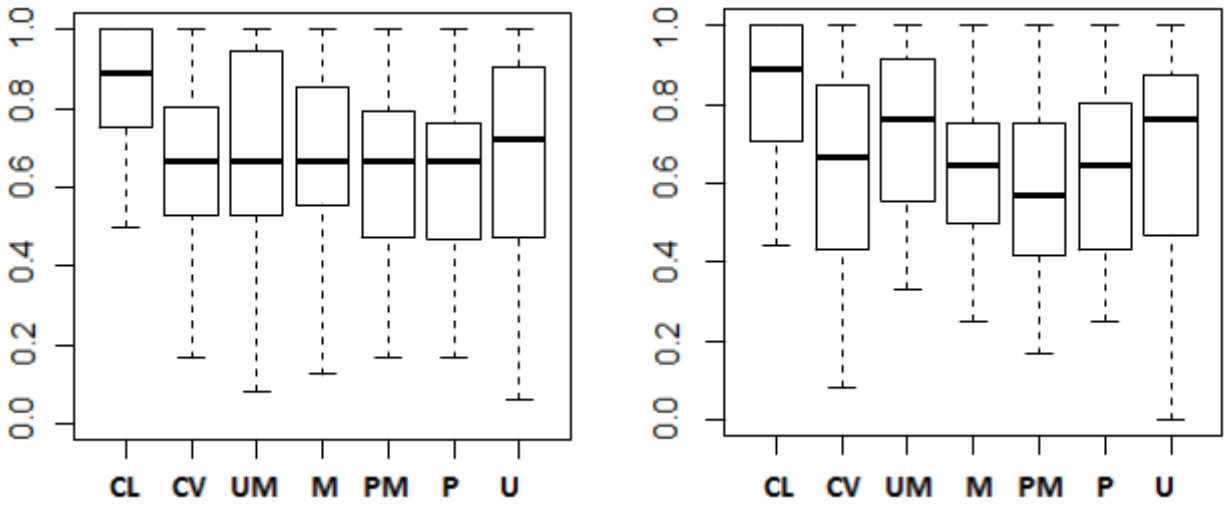
(a) (b)

Figure 5.10: Intelligibility scores per Category across listeners on (a) $RT_1 = 0.8s$ (b) $RT_2 = 2s$

In order to evaluate the statistical significance of these results, a repeated-measures ANOVA is performed on intelligibility with Category nested within each listener. Results reveal significant intelligibility differences

among Categories, for both reverberant conditions RT_1 ($F(6, 20) = 5.601, p < 0.001$) and RT_2 ($F(6, 20) = 7.167, p < 0.001$). Post-hoc comparisons using pairwise paired t-tests reveal that the mean intelligibility score of CL ($M = 0.86, SD = 0.13$, M stands for mean and SD for standard deviation) is significantly different ($p < 0.001$) from CV ($M = 0.67, SD = 0.22$) in RT_1 while in RT_2 both CL ($M = 0.83, SD = 0.16$) and UM ($M = 0.77, SD = 0.17$) have significantly different means ($p < 0.01$) from CV ($M = 0.64, SD = 0.23$). No significant difference between means of CL and UM are reported ($p = 0.07$).

For each reverberant condition, the ratio of the correct key-words to the number of total keywords per listener is estimated per sentence and per Category. Then, the mean of the ratios for all listeners is estimated per sentence and per Category. Figure 5.11 shows the {min, 1^{st} quartile, median, 3^{rd} quartile, max} of intelligibility scores per Category across all sentences. CL appear to have a higher intelligibility advantage over all Categories for both reverberant conditions while the UM seems to have a benefit over CV on RT_2 (Figure 5.11).



(a) (b)

Figure 5.11: Intelligibility scores per Category across sentences on (a) $RT_1 = 0.8s$ (b) $RT_2 = 2s$

A repeated-measures ANOVA on intelligibility with Category nested within each sentence is performed to remove possible dependencies of the intelligibility scores on sentence difficulty. ANOVA null hypothesis of equal means of the intelligibility scores for every Category, is rejected using the F-test for RT_1 ($F(6, 27) = 6.634, p < 0.001$) and RT_2 ($F(6, 27) = 7.268, p < 0.001$). Post-hoc comparisons using pairwise paired t-tests reveal that the mean intelligibility score of CL ($M = 0.87, SD = 0.15$) is significantly different ($p < 0.01$) from CV ($M = 0.66, SD = 0.20$) in RT_1 while in RT_2 both CL ($M = 0.83, SD = 0.18$) and UM ($M = 0.75, SD = 0.21$) have means different from CV ($M = 0.63, SD = 0.26$) and this result is statistical significant ($p < 0.001$ for CL, $p < 0.01$ for UM). No significant differences are reported between the means of CL and UM ($p = 0.07$). The mean of UM is significantly different from the means of all other modifications ($p < 0.01$). Last, pairwise paired t-tests showed no significant difference between means per Category in RT_1

with their corresponding in RT_2.

Native listeners and Hearing-Impaired

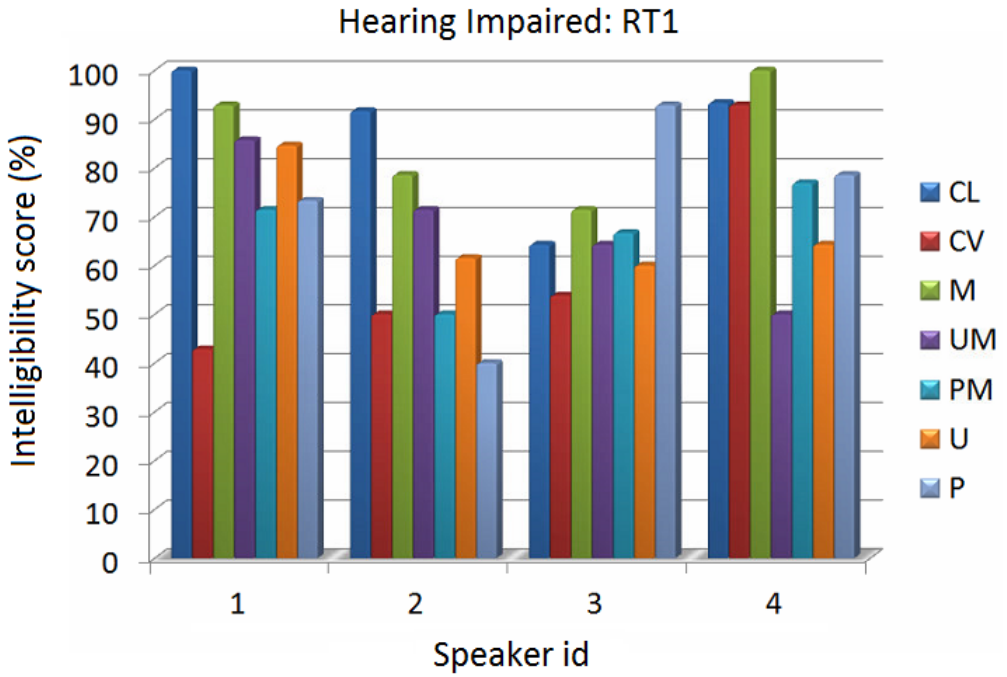
Subjective evaluations are also performed by 7 native listeners. As the sentences were meaningful, the content helped the native listeners to understand both CL and CV speech almost 100%. One listener appeared to have intelligibility score below 70% for both speaking styles. That listener benefits from all modification techniques in RT_2, and in RT_1 from all modifications except uniform-time scaling. Repeated measures ANOVA showed no statistical significant differences between Categories both for RT_1 ($F(6,6) = 1.544, p = 0.192$) and RT_2 ($F(6,6) = 1.781, p = 0.131$).

Subjective evaluations were also performed by 4 non-native hearing impaired listeners. The type and severity of hearing loss was not reported by the subjects, since the intelligibility test was conducted online. Figure 5.12 depict the intelligibility scores for each Category in two conditions RT_1 and RT_2. CL speech is more intelligible than CV speech in RT_1 ($M_{CL} = 0.87, SD_{CL} = 0.16$, $M_{CV} = 0.60, SD_{CV} = 0,22$) and RT_2 ($M_{CL} = 0.80, SD_{CL} = 0.23$, $M_{CV} = 0.57, SD_{CV} = 0.17$). In RT_2 condition, modification schemes failed to increase the intelligibility of casual speech. However, for RT_1, all listeners showed an intelligibility increase of modified casual speech with the Mix-filtering modification ($M = 0.86, SD = 0.26$). Repeated measures ANOVA showed no statistical significant differences between Categories for RT_1 ($F(6,3) = 1.754, p = 0.166$) and RT_2 ($F(6,3) = 3.228, p = 0.0248$).

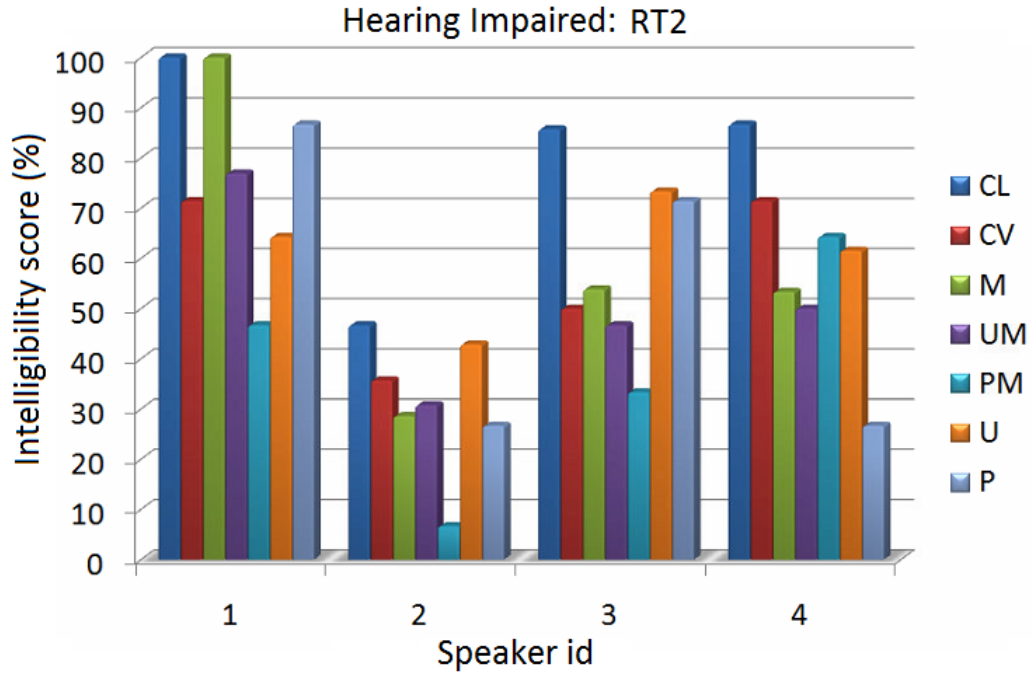
5.3.2 Discussion

Subjective evaluations presented in this experiment confirm that clear speech is more intelligible than casual speech in reverberant conditions for the non-native listeners. Indeed, CL outperforms CV by 19% in 0.8s and 2s reverberant time. Non-native listeners also report that the combination of uniform time scaling and Mix-filtering technique is advantageous for RT_2 since the intelligibility benefit is 13%, that is 6% lower from the upper bound (CL). However, in less reverberation, the benefit of this modification drops. This inefficiency is possibly due to the selection of the uniform-time scaling factor. Figure 5.10(a) shows that the Mix-filtering technique has a slight advantage over casual speech. Then, when uniform-time scaling is combined with the spectral boosting, the median intelligibility score drops and the variance increases. Therefore, this result indicates that the duration of the speech signal is linked to the reverberation condition and the selection of the time-scaling factor should be proportional to the reverberation time. Also, the PSQ-based modification fails to increase intelligibility of casual speech. One possible reason for this is the change of rhythm between speech segments and the extreme elongation in some cases. A more conservative time-scaling factor could be proven more advantageous for the time-scaling techniques and is to be explored in the future.

The hearing-impaired people reported that clear speech has 23% and 27% higher intelligibility than casual speech for RT_1 and RT_2, respectively and that the Mix-filtering in RT_1 increases the intelligibility of casual speech by 26%. However, the hearing-impaired population is rather small to draw any concrete conclusions.



(a)



(b)

Figure 5.12: Intelligibility scores per Category for each hearing-impaired listener on (a) $RT_1 = 0.8s$ *(b)* $RT_2 = 2s$

Finally, native listeners do not benefit from the transformations since the intelligibility of CV is as high as that of CL, highlighting the importance of the semantic content and/or the amount of reverberation, above

which their perception is degraded (possibly on higher reverberation times).

Results indicate that modifications based on clear speech properties can be beneficial for the intelligibility enhancement of casual speech in reverberant environments. The proposed modification uses a combination of spectral boosting and uniform time-scaling. Our spectral transformation applies a multi-band filtering on casual speech, enhancing information of important frequency bands indicated by clear speech and it has low computational complexity, since it does not require detection of steady-state portions. The Mix-filtering and uniform time-scaling combination increases the intelligibility of casual speech in high reverberant environments ($RT = 2s$) for the non-native population. Future work could focus on validating the linear connection of the uniform-time scaling factor to the reverberation time.

Chapter 6

Modulation Enhancement

In Chapter 2, the importance of modulations for speech perception has been underlined, reporting also differences on the temporal envelope modulations of clear and casual speech. Previous efforts to enhance the modulation depth of the temporal envelopes have also been reported, revealing the difficulty of related studies to enhance speech intelligibility by changing the temporal envelopes of speech. In this work, we have achieved to enhance the intelligibility of speech by manipulating the modulation spectrum of the signal. Our method is based on the concept of coherent speech demodulation. First, the signal is decomposed into Amplitude Modulation (AM) and Frequency Modulation (FM) components using a high resolution adaptive quasi-harmonic model of speech. Then, the AM part of frequencies above the midrange region of the speech spectrum is modified by applying a perceptually motivated compression rule, mimicking the non-linear compression activity on the basilar membrane, as well following characteristics of the clear style of speaking. This results in increasing the modulation depth of the temporal envelopes of weak - in terms of loudness - components. The modified AM components of speech are then combined with the original FM parts to synthesize the final processed signal. Subjective listening tests evaluating the intelligibility of speech in noise showed that the suggested approach increases the intelligibility of plain speech by 40% on average, while it is comparable with recently suggested state-of-the-art algorithms of intelligibility boosters, the SSDRC and Mix-filtering technique.

This Chapter is organized as follows. First, the AM-FM coherent demodulation algorithm will be presented, namely the extended adaptive quasi-harmonic model (eaQHM). This model is used for analysis and synthesis of speech. Then, the transforming function that boosts the low-frequency amplitude modulations of the temporal envelope is defined. Next, the evaluation section is presented, describing the intelligibility benefit of our modulation enhancement technique via subjective intelligibility tests performed by native listeners on SSN noise. SSDRC and Mix-filtering are also evaluated for comparison reasons. Last, we attempt to find a relation between beneficial spectral modifications and modulation enhancement. Therefore, our proposed modulation enhancement technique is compared in terms of modulations with the SSDRC and Mix-filtering.

6.1 Coherent demodulation of temporal envelopes

Our novel method for increasing the modulation depth of the temporal envelopes of speech and simultaneously its intelligibility is based on the concept of coherent demodulation. Coherent demodulation approaches (Atlas and Janssen, 2005; Schimmel and Atlas, 2005) suggest that for decomposing properly the signal into amplitudes and carrier frequencies, the initial signal should be filtered into bandpass analytic signals of relative narrow bandwidth. This concept motivated us to separate the temporal envelopes (amplitudes) from the temporal fine structure (carriers) using a Sinusoidal model, rather than the state-of-the-art technique, namely the Hilbert transform. Specifically, a powerful quasi-harmonic model is introduced to decompose speech into time-varying amplitudes, frequencies and phases. Each time-varying amplitude is considered to be the "temporal envelope" of the signal in the corresponding frequency. Then, a transforming function is applied to the time-varying amplitudes to increase their modulation depth. The speech signal is then synthesized using the modified amplitudes and the initial frequencies and phases. The analysis and synthesis of the speech signal is performed by the extended adaptive Quasi-Harmonic Model (eaQHM) proposed by Kafentzis et al. (2014, 2012) which can decompose and reconstruct the signal with high accuracy.

The proposed modification algorithm changes the temporal envelope modulations of speech similar to clear speech. The transforming function is inspired by clear speech properties and by the input-output functions on the basilar membrane of the cochlea. When presented in noise, the modified signal has higher intelligibility than the unprocessed signal. The advantage of our method is that it does not require filterbank analysis and synthesis and temporal envelope estimation using the Hilbert transform. Such techniques reduce the effectiveness of the modulation filters (Atlas and Janssen, 2005) or introduce artifacts to the signal detrimental for its intelligibility. Furthermore, the proposed method does not require the design of modulation filters which their efficiency may depend to the type of noise (Kusumoto et al., 2005). Instead, a simple transforming function is used to change the modulation depth of the time-varying amplitudes. In the subsection that follows, the eaQHM algorithm is described.

6.1.1 Decomposition and reconstruction of speech: the extended adaptive quasi-harmonic model (eaQHM)

Generally, the speech signal, containing both voiced and unvoiced segments, can be described as an AM-FM decomposition:

$$x(t) = \sum_{k=-K}^{K} \alpha_k(t) e^{j\Phi_k(t)} \tag{6.1}$$

where $\alpha_k(t)$, $\Phi_k(t)$ are the instantaneous amplitude and the instantaneous phase of the k^{th} component, respectively. A means to accurately compute these parameters is the full-band extended adaptive Quasi-Harmonic

Model (Kafentzis et al., 2014, 2012) which has been successfully used for analysis and synthesis of speech:

$$\hat{x}(t) = \sum_{k=-K}^{K} (\hat{\alpha}_k + t\hat{b}_k)\hat{A}_k(t)e^{j\hat{\phi}_k(t)} \tag{6.2}$$

In this model, $\hat{\alpha}_k$ is the complex amplitude and $\hat{b}_k$ is the complex slope of the k^{th} component, and $\hat{A}_k(t), \hat{\phi}_k(t)$ are functions of the instantaneous amplitude and phase of the k^{th} component, respectively (Kafentzis et al., 2012). These estimates are iteratively updated via Least Squares until a convergence criterion is met, which is related to the overall Signal-to-Reconstruction-Error Ratio (SRER) (Kafentzis et al., 2014). Then, the overall signal is synthesized using equation (6.1) where $\hat{\Phi}_k(t)$ is formed by a frequency integration scheme using the estimated frequencies and phases (Pantazis et al., 2011) and $\alpha_k(t)$ is simply $|\hat{\alpha}_k(t)|$ via linear interpolation.

6.2 Modulation enhancement based on the non-linear compression function of the basilar-membrane and on clear speech properties

After AM-FM decomposition of speech, the time-varying amplitudes need to be properly modified in order to increase their modulation depth. Two major concepts are introduced to propose the appropriate transforming function. First, following the characteristics of clear speech, not all time-varying amplitudes will be modified. Comparative acoustic analysis between clear speech and casual speech has shown increased modulation depth of the temporal envelopes of clear speech in midrange frequencies (Krause and Braida, 2004a). Therefore, the temporal envelopes of low midrange frequencies will not be modified. Second, the degradation of the temporal envelope sensitivity is connected to hearing loss. This degradation may be attributed to the inability of the basilar-membrane to perform non-linear compression (Moore and Oxenham, 1998). This suggests that the low loudness parts of the temporal envelope are equally important. Boosting this information can be beneficial both for hearing impaired and normal-hearing people. Therefore, we propose a transforming function which approaches the compressive nonlinearity that takes place in the basilar-membrane (Moore and Oxenham, 1998):

$$m_k(t) = \hat{\alpha}_k(t)^{\gamma}, \quad |k| = 4, .., K \tag{6.3}$$

where $\frac{1}{3} < \gamma < 1$. After modifying the time-varying amplitudes using equation (6.3), the signal is synthesized using equation (6.1) where $\alpha_k(t) = m_k(t)$. Then, the synthesized signal is normalized to have the same Root Mean Square energy (RMS) as the original unmodified signal.

The proposed transforming function, called DMod, has the ability to significantly increase the very low values of the time-varying amplitude component, while maintaining or conservatively increasing the higher ones. The motivation of proposing this transforming function is based both on the non-linear compressive function of the basilar membrane and on the comparative analysis between clear and casual speech. Figure 6.1(a) shows the temporal envelopes (time-varying amplitudes) near 3000 Hz (15^{th} harmonic) of a clear sentence and its

casual counterpart, while Figure 6.1(b) depicts the corresponding normalized amplitudes. Comparing clear and casual temporal envelopes two observations can be made. First, the amplitude harmonic of clear speech is in average greater than that of casual speech, showing an increased spectral energy of clear speech compared to casual speech for this harmonic (Figure 6.1(a)). Second, removing this effect by normalizing, we can observe that casual speech has weaker - in terms of amplitude - components (near 0.25s and 2s) which may be masked when presented in noise. Applying the compression function of equation (6.3), we can see that these weak components are enhanced by DMod compared to the original casual signal (Figure 6.1(b)) while the average amplitude of the temporal envelope is boosted (Figure 6.1(a)), as in clear speech. Furthermore, it should be noted that not all the amplitude components are modified by the transforming function described in equation (6.3). This is due to the fact that comparing the temporal envelopes of clear and casual speech on low frequency regions, we have observed that amplitude components of casual speech seem equally enhanced and in sometimes more boosted than that of clear speech.

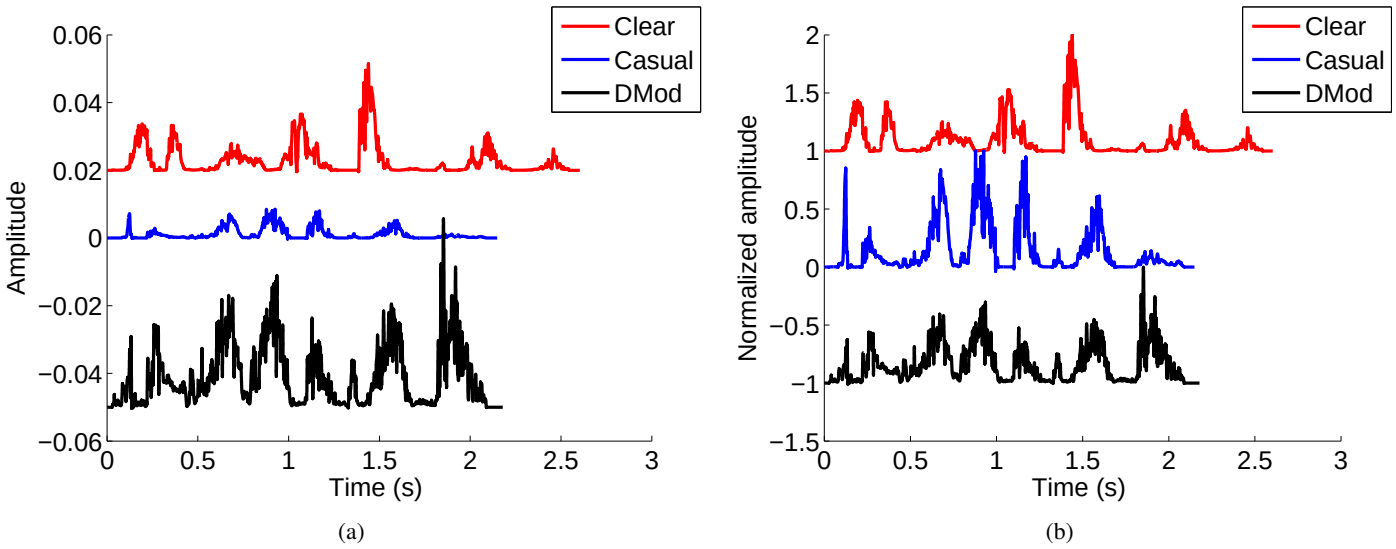


Figure 6.1: Time-varying amplitude of 15 quasi-harmonic (around 3000Hz) estimated by eaQHM for the same sentence uttered in clear and casual style. The modified amplitude harmonic by the proposed modulation enhancement technique, DMod is also depicted. (a) Amplitudes (b) Normalized amplitudes. Note that only for visualization purposes different mean values are added to the amplitudes, therefore only the scale of the vertical axis is informative.

To quantify the amount of modulation enhancement, we estimate the modulation depth for a casual sentence before and after applying DMod. Figure 6.2 depicts the mean modulation depth, $\overline{D(t)}$, of the temporal envelopes on three acoustic frequency regions for clear, casual and modified casual signal using the transforming function (DMod) with $\gamma = 0.5$. The mean modulation depth, $\overline{D(t)}$, is estimated as follows: eaQHM decomposes speech into acoustic frequencies and amplitudes. The time-varying amplitudes that correspond to the acoustic frequency regions depicted in Figure 6.2 are summed to derive the temporal envelope for each frequency region. Now, let us denote as $p(t)$, the temporal envelope whose modulation depth needs to be estimated. In equation (6.1), $\hat{x}(t) = p(t)$ and $K = 8$ in order to capture modulation frequencies from 2 to 8

Hz. Using equation (6.1), eaQHM decomposes the temporal envelope into modulation amplitudes $\alpha_k(t)$ and modulation frequencies from 2-8 Hz. The modulation depth $D(t)$ is computed using equation (6.4), which sums the modulation amplitudes of the above modulation frequencies. Note that $D(t)$ is also time-varying. The average of $D(t)$ in time, namely $\overline{D(t)}$ is then depicted in Figure 6.2.

$$D(t) = \sum_{k=-8, k\neq 0,1}^{8} |\hat{\alpha}_k(t)| \tag{6.4}$$

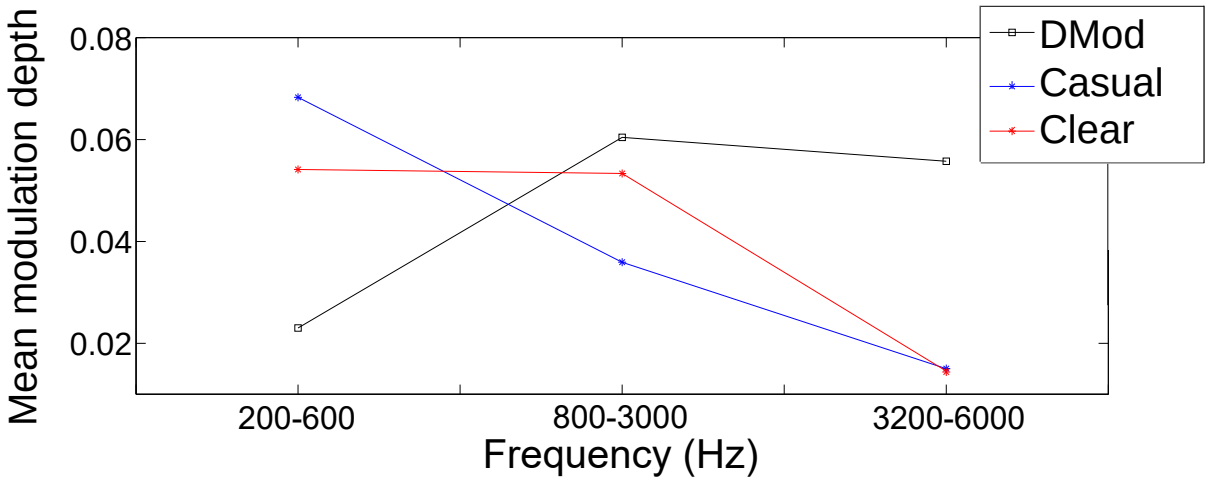


Figure 6.2: Mean modulation depth for modulation frequencies $2-8$ Hz of the temporal envelopes of $\{$Clear, Casual, DMod ($\gamma = 0.5$)$\}$ on different frequency regions for the same sentence.

Figure 6.2 illustrates that clear speech (Clear) has higher mean modulation depth than casual speech (Casual) on midrange and upper midrange frequencies (800-3000Hz, 4^{th}-15^{th} component) while on low midrange frequencies (200-600Hz, 1^{st}-3^{rd} component) clear speech has lower mean modulation depth than casual speech, supporting our initial observations. The transforming function with $\gamma = 0.5$ (DMod) follows this clear speech characteristic; it increases the modulation depth of casual speech significantly for frequencies above the midrange while decreases its modulation depth on low midrange frequencies. Trials and errors along with informal listening tests on values of γ showed that the intelligibility of casual speech significantly increases around the area of $\gamma = 1/2$. For higher values of γ the modulation depth of the modified time-varying amplitudes increases with less intensity and the modified signal is acoustically closer to the original signal. For higher SNR levels a lower value of 0.5 for γ can be selected. Therefore, the parameter γ can be proportional to the noise level. However, very low values of γ should be avoided, since they create signal distortions that may affect speech intelligibility. Figure 6.3 presents the spectrogram of the original signal and the modified signal using the transforming function with $\gamma = 0.5$. It is worth noticing how the harmonic structure is emphasized. The same sentences in SSN are depicted in Figure 6.4 where SSDRC in SSN is also illustrated. With visual inspection it is observed that both SSDRC and DMod emphasize speech segments "popping them out" of noise.

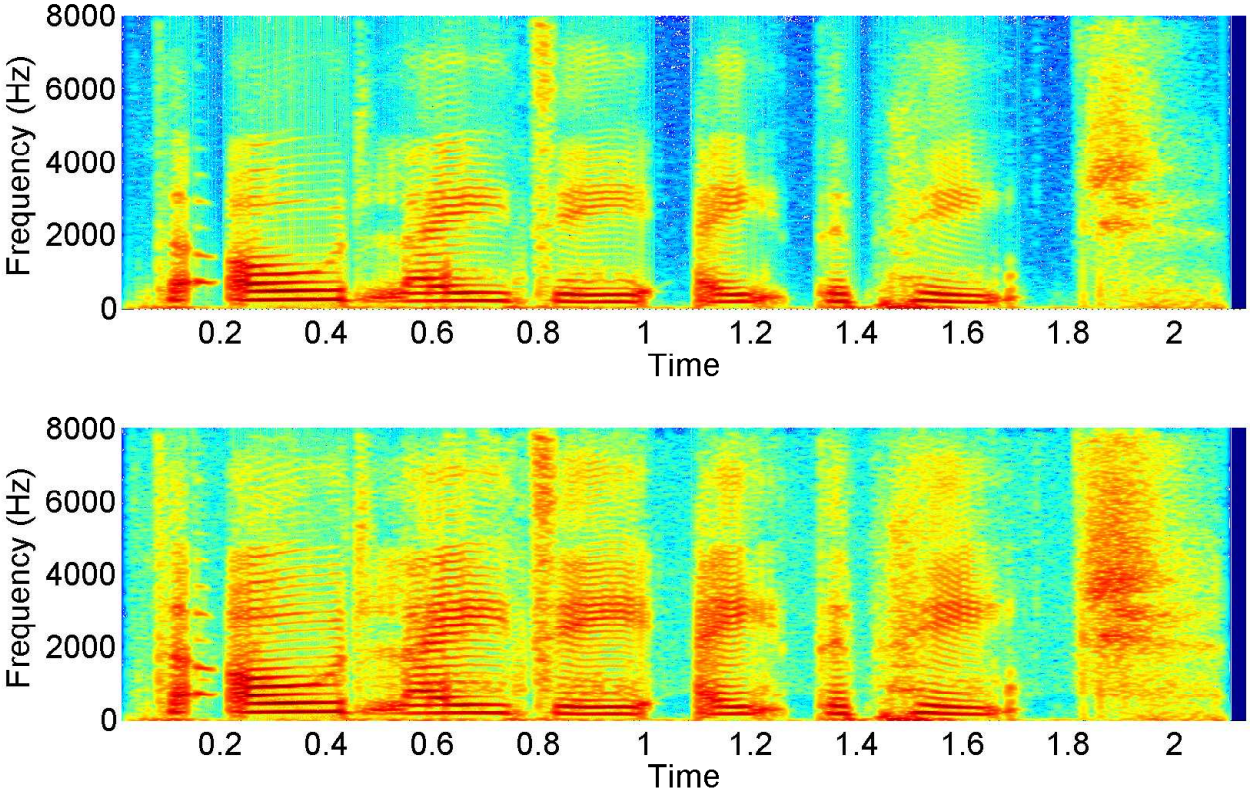


Figure 6.3: Spectrogram of the casual signal (upper panel) and the modified casual signal (lower panel) using the transforming function with $\gamma = 0.5$.

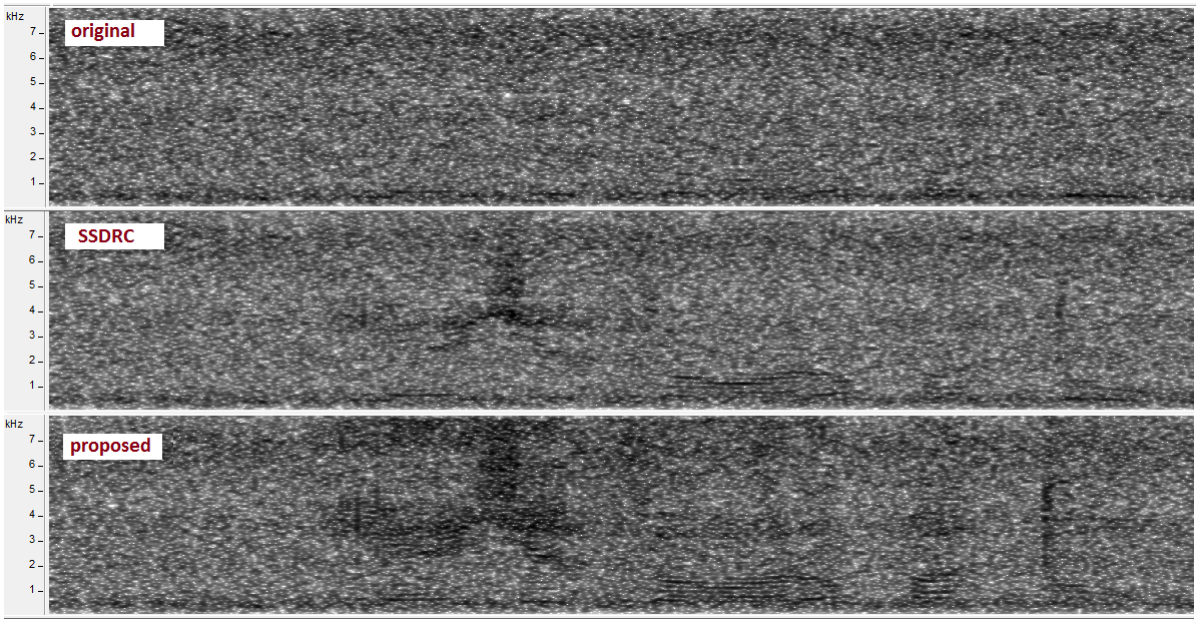


Figure 6.4: Spectrogram of the casual signal (upper panel), the SSDRC modified casual signal (middle panel) and the DMod modified casual signal (lower panel) in noise using the transforming function with $\gamma = 0.5$.

6.3 Evaluations

In the previous section it was shown that the proposed compression function leads to increased modulation depth of the temporal envelopes for low modulation frequencies (2-8 Hz) (Figure 6.2). In this section we will evaluate our algorithm in terms of intelligibility via subjective intelligibility tests.

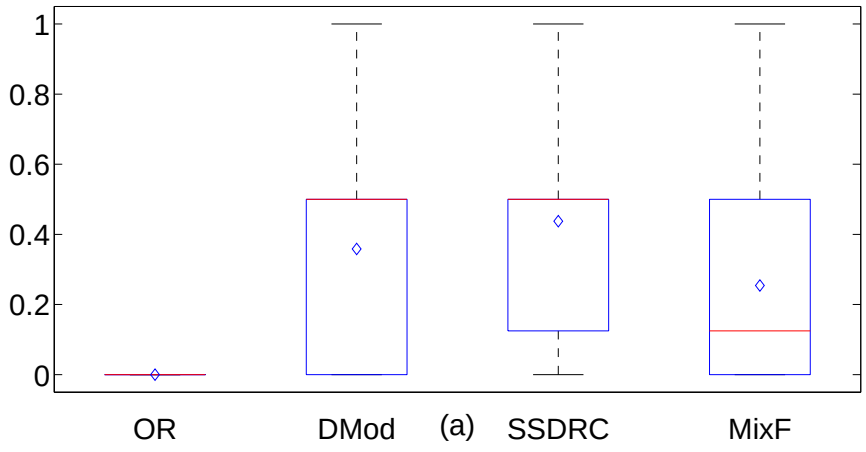
For comparison purposes, our proposed algorithm is compared to two other intelligibility enhancing techniques, the SSDRC and the Mix-filtering described analytically in Chapter 5. Four sets of signals are evaluated: (1) the original speech (OR), (2) the proposed modified speech using modulation-depth enhancement (DMod) (3) the SSDRC modified speech (SSDRC) (4) the mix-filtering modified speech (MixF). The term Categories will be used to refer to the 4 sets of signals, {OR, DMod, SSDRC, MixF}.

The database for evaluating the proposed modification is different from the LUCID database. The database, called SpeakGreek[1] contains sentences in Greek uttered by normophonic female and male speakers. The database is used to test word intelligibility by native (Greek) listeners. The corpus contains sentences with one keyword inside the carrier sentence uttered in Greek "Lége [léksi klidí] padú" ("Say [keyword] everywhere") (http://speakgreek.web.auth.gr/, 2013; Nicolaidis et al., 2014, 2015a,b). The keyword is a CVCV word. Each sentence is recorded five times. The speakers were instructed to speak at a comfortable speaking rate. Therefore, among the five repetitions the speaking style presents a variability. There are differences in the carefulness and style of production with some sentences being more hyper-articulated and carefully elicited than others. The more casually pronounced sentences are selected from this corpus (casual sentences). The sentences are presented in SSN of low ($SNR_1 = -8dB$) and mid ($SNR_2 = -2dB$) levels of SNR. The listeners are asked to write down the keyword that they hear. Each keyword is presented only once to the listeners. 16 distinct sentences are presented to the listeners for evaluation uttered by a female and a male speaker (8 sentences per speaker (8x2 = 16 sentences), 4 sentences per speaker per noise level (4x2x2 = 16 sentences), 2 sentences per Category per noise level (2x4x2 = 16 sentences)). All sentences are normalized to have the same RMS energy and then noise is added to the sentences. First, the low SNR sentences are presented to the listeners and next the sentences on higher SNR. 4 sentences are presented as a header to the listeners to adjust their hearing to the noise level (20 sentences in total). The "header sentences" are not evaluated. The scoring system is based on previous research on English intelligibility tests (Monsen, 1978, 1982; Picheny et al., 1985a), supported also by researchers for Greek language (Sfakianaki, 2012). Each word is considered incorrect even if there is a mismatch in one phoneme e.g "fíki" instead of "thíki" ("seaweed" instead of "case"). However, incorrect person of verb and number of noun is considered half-correct e.g. "dóra" instead of "dóro" ("gifts" instead of "gift").

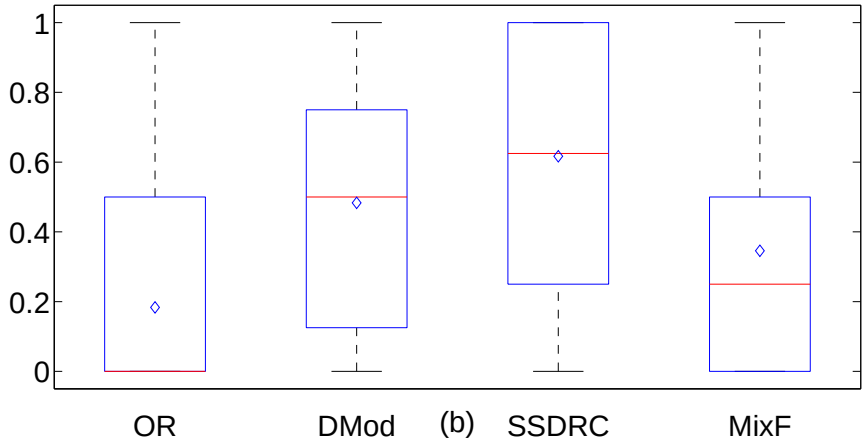
As word difficulty may affect the intelligibility scores, 4 different listening scenarios were created to ensure that each word would be presented in a {OR, DMod, SSDRC, MixF} manner to different listeners (as each listener cannot hear the same sentence twice). For example, if a specific word was presented to the listener in

[1]The authors would like to thank Prof. Katerina Nicolaidis and Dr. Anna Sfakianaki for providing the database: "SpeakGreek: Developing a biofeedback speech training tool for Greek segmental and suprasegmental features: Application in L2 learning/teaching and clinical intervention", co-financed by the European Union (ESF) and Greek national funds (ARISTEIA II).

OR manner on SNR_1 on the listening Scenario 1, then on listening Scenario 2 the same word was presented in SSDRC manner to another listener on the same SNR condition etc. This allowed us to "denoise" the performance evaluation from the word dependency.



(a)



(b)

Figure 6.5: Intelligibility score across listeners per Category (a) SNR=-8dB (b) SNR=-2dB

60 listeners (15 listeners per scenario), all native Greek speakers, participated in the intelligibility test. Performance evaluation contains two parts of analysis. The first part presents the intelligibility scores of each Category across listeners, in order to reveal possible intelligibility benefits of the proposed modifications for the native population. The second part of analysis computes the intelligibility scores of each Category across sentences, to parcel out the possible variability due to word difficulty.

For each SSN condition, the score of the correct and the half correct keywords is estimated and divided by the score of the total keywords. This normalized score is estimated per listener and per Category. Figure 6.5 shows the {min, 1st quartile, median, 3rd quartile, max} of intelligibility scores per Category across all listeners. As it is expected, there is a high variability across listeners attributed to the differences in the

degree of influence of noise on listener's perception (Cooke and Lecumberri, 2012b). Mean values are also depicted (rhombus symbol). SSDRC appears to have a higher intelligibility advantage over all Categories for both SSN conditions. Our proposed modification DMod has higher intelligibility score, that is 36% and 48% (mean values) compared to unmodified speech (0% and 18%) for SSN_1 and SSN_2 respectively, approaching the intelligibility benefit of SSDRC (44% and 62% respectively). MixF achieves lower intelligibility scores than DMod (25% for SSN_1 and 35% for SSN_2).

In order to evaluate the statistical significance of these results, a repeated-measures ANOVA was performed on intelligibility with Category nested within each listener. Results reveal significant intelligibility differences among Categories, for both SSN conditions SSN_1 ($F(3;59) = 35.91$; $p < 10^{-15}$) and SSN_2 ($F(3;59) = 15.46$; $p < 10^{-8}$). Post-hoc comparisons using pairwise paired t-tests with Holm adjustment reveal that the mean intelligibility scores of DMod ($M = 0.36$ (mean); $SD = 0.26$ (standard deviation)), SSDRC ($M = 0.44$; $SD = 0.32$) and MixF ($M = 0.25$; $SD = 0.28$) are significantly different ($p_{DMod} < 10^{-14}$, $p_{SSDRC} < 10^{-13}$, $p_{MixF} < 10^{-8}$) from OR ($M = 0$; $SD = 0$) in SSN_1. For SSN_2, both DMod ($M = 0.48$; $SD = 0.37$) and SSDRC ($M = 0.62$; $SD = 0.38$) have significantly different means ($p < 10^{-4}$) from OR ($M = 0.18$; $SD = 0.25$) while no statistical significance ($p = 0.028$) was found between the mean of MixF ($M = 0.35, SD = 0.39$) and OR ($M = 0.18$; $SD = 0.25$). No statistical significant difference was found between MixF, DMod and SSDRC on SSN_1. On SSN_2 condition there was no statistical significant difference between SSDRC and DMod.

In order to investigate possible dependencies of the intelligibility scores on word difficulty, intelligibility scores for each word was computed for all Categories. Figure 6.6 shows the normalized scores for each word for the two SSN conditions, SSN_1 (Figure 6.6(a)) and SSN_2 (Figure 6.6(b)). As we can see there are words that influence the efficiency of the modification algorithms possibly due to their difficulty. This variability on the intelligibility scores due to word selection justifies the high variance on the intelligibility scores for the modification techniques reported in Figure 6.5. In words "zoúla, laliá" the intelligibility scores of {SSDRC, DMod, MixF} are higher than that of like "theté, goní". Possibly in the latter words, the use of different grammatical cases than the nominative made it difficult for the listeners to guess the word meaning. It is worth noticing that original speech, OR, has also higher intelligibility score in some words for the SSN_2 condition. Finally, 9 out of 16 words have the highest intelligibility scores when modified by SSDRC, while 4 out of 16 words have the highest intelligibility score when modified by DMod.

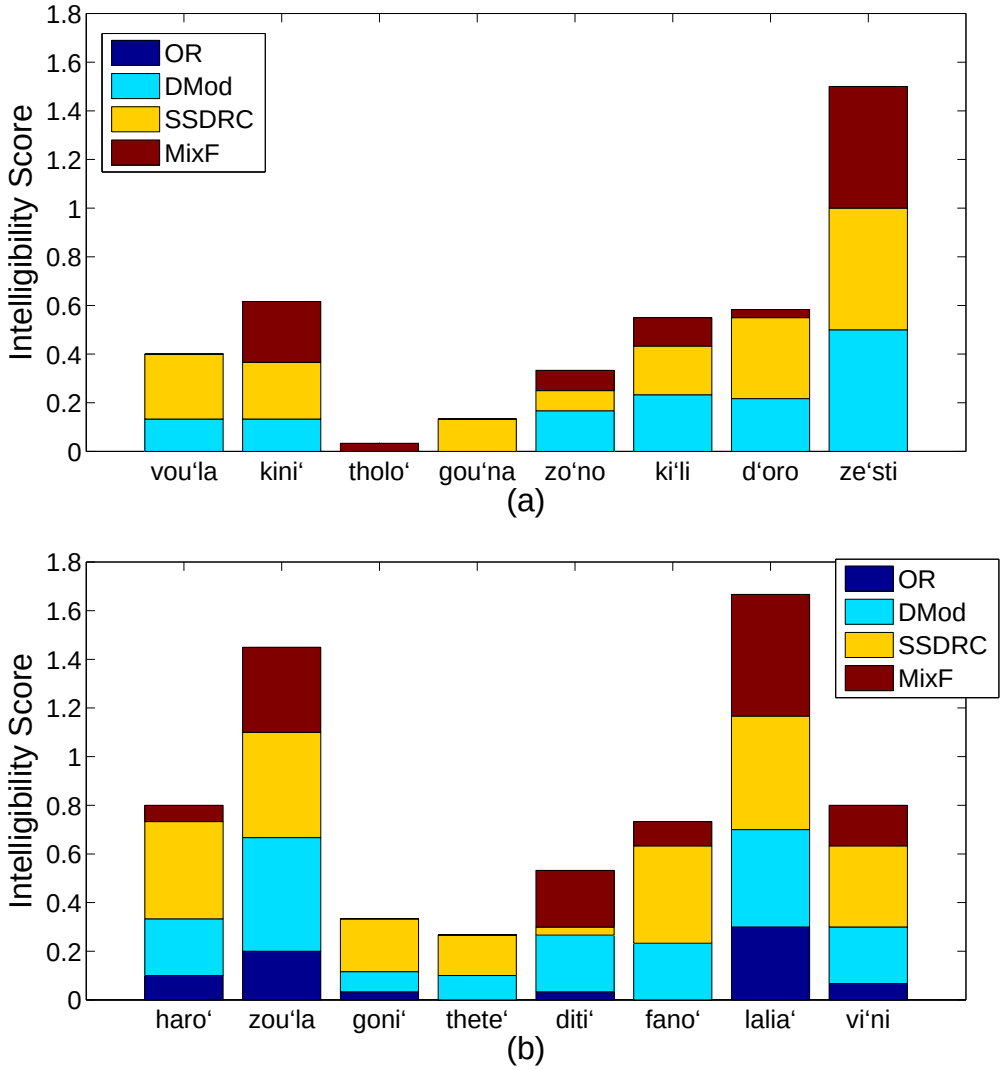


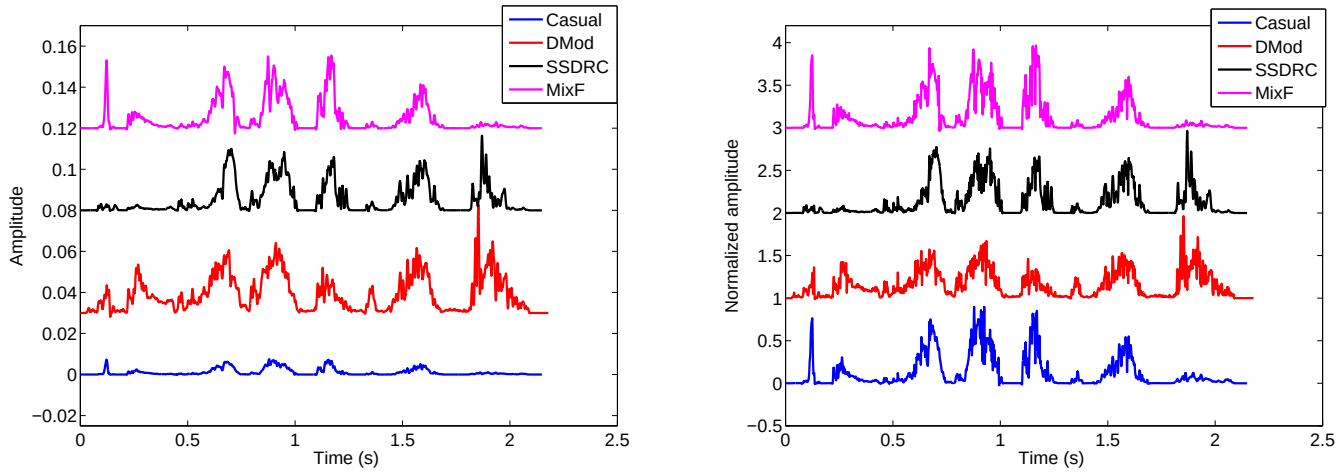
Figure 6.6: Intelligibility score of each word per Category (a) SNR=-8dB (b) SNR=-2dB

6.4 Examining the relation of spectral transformations and modulation enhancement

Motivated by the above intelligibility scores, a future work would be to combine the gains of DMod and SSDRC in order to create a high intelligible signal in noise. However, first we should examine if SSDRC changes the modulation depth of the temporal envelopes, through energy reallocation in the spectral and time domain. In this case, a combination of DMod with spectral modification schemes would probably be less promising than expected. The main question that we need to answer is *why when boosting specific frequency content speech intelligibility increases in SSN noise? Is this attributed to the fact that the spectral energy of the signal exceeds the noise level or this spectral energy re-distribution enhances the temporal envelope modulations of speech, which are linked to speech perception in noise?* If spectral transformation techniques indeed increase the modulation depth of temporal envelopes, then the intelligibility increase that casual speech experiences is possibly due to modulation enhancement and not due to energy reallocation. This result would be very

important since it would suggest to shift our research interest from indirect (through spectral transformations) to direct modulation enhancement, focusing on specific frequency bands.

In this work, we have achieved to increase the intelligibility of speech in noise by 30-40%, reaching the intelligibility scores of SSDRC. However, this result has been accomplished only by manipulating the temporal envelope modulations of speech. We assume that SSDRC and probably Mix-filtering also enhance the modulation depth of the temporal envelopes. Therefore, we are motivated to investigate how the temporal envelopes change in time and to which extent their modulations are affected for the spectral modification techniques. Following the analysis described in Section 6.2 for SSDRC and Mix-filtering, Figure 6.7 depicts the amplitudes (Figure 6.7(a)) and normalized amplitudes (Figure 6.7(b)) of the 15 harmonic estimated by eaQHM for the same sentence in Casual, SSDRC, MixF (Mix-filtering) and DMod. As we can see, the spectral modifications also enhance the average amplitude value of the specific harmonic (Figure 6.7(a)). MixF seems to preserve the overall shape of the casual temporal envelope waveform (Figure 6.7(b)), while SSDRC and DMod change the shape of the waveform by highlighting different time regions. It is obvious from Figure 6.7 that only by enhancing the amplitude of the temporal envelope using spectral modifications, the modulation depth of the temporal envelope is also enhanced.



(a) (b)

Figure 6.7: Time-varying amplitude of 15 quasi-harmonic (around 3000Hz) estimated by eaQHM for the same sentence in Casual, SSDRC, MixF and DMod (a) Amplitudes (b) Normalized amplitudes. Note that only for visualization purposes different mean values are added to the amplitudes, therefore only the scale of the vertical axis is informative.

Examining more than one harmonics, the modulation spectra (Atlas and Janssen, 2005) of unmodified (Clear, Casual) and modified speech (DMod, SSDRC, MixF) are illustrated in Figure 6.8. Clear speech appears to have more intense modulations all over the spectrum compared to casual speech, whose modulations are limited in the midrange. Both SSDRC and MixF seem to enhance the modulations of casual speech, approaching the modulation spectrum of clear speech. Indeed, comparing the modulation spectra of clear speech and MixF there is a great similarity. Therefore, our initial suggestion that spectral modification techniques

probably enhance the modulations of the temporal envelopes is verified. DMod that directly focuses on modulation enhancement, boosts significantly modulations all over the spectrum.

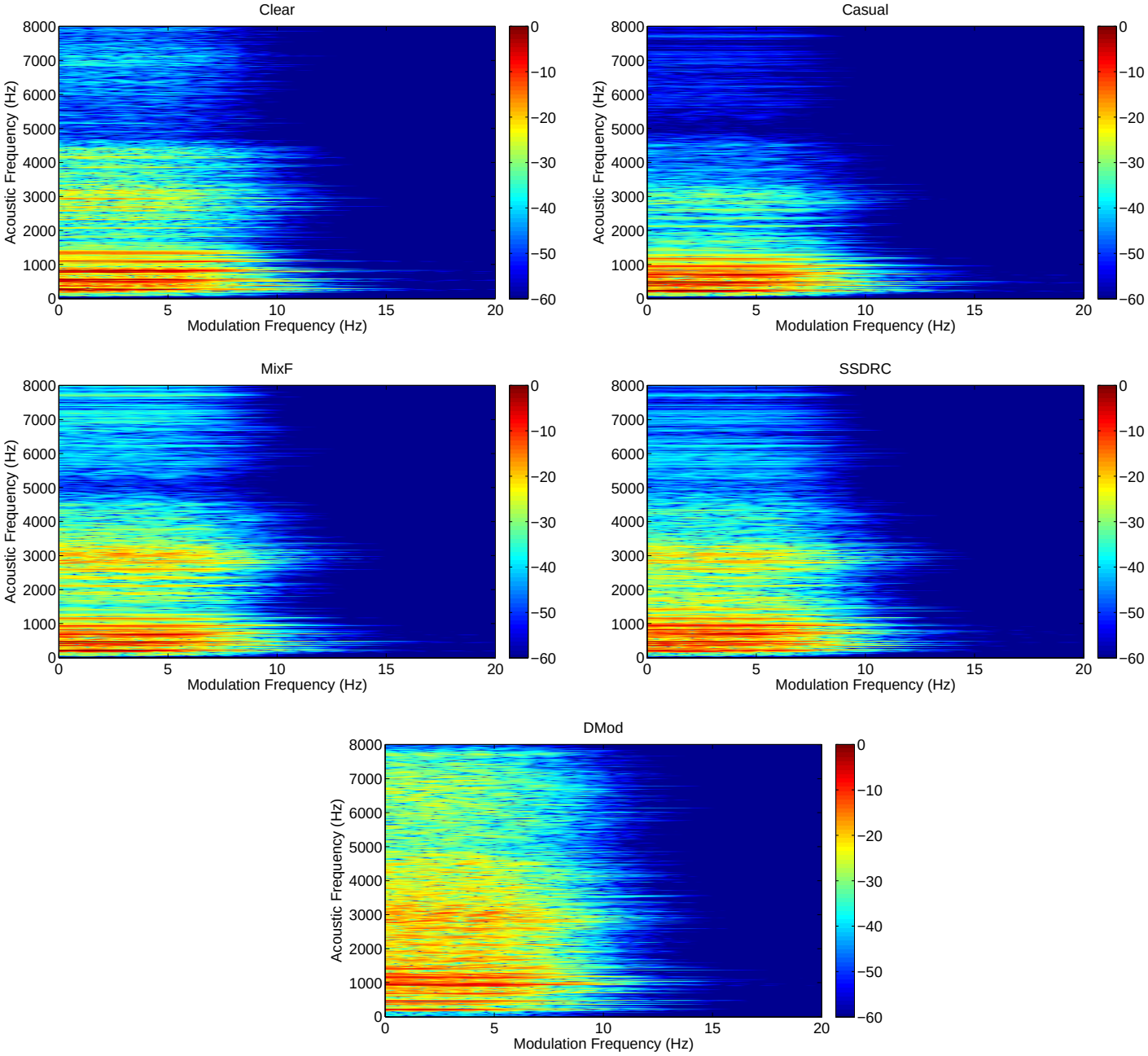


Figure 6.8: Mean modulation depth for modulation frequencies $2-8$ Hz of the temporal envelopes of $\{$Clear, Casual, DMod ($\gamma = 0.5$), SSDRC, MixF$\}$ on different quasi-harmonic regions for the same sentence.

6.5 Discussion

Subjective evaluations report that the proposed modification method, DMod, increases speech intelligibility in SSN. This intelligibility improvement is inspired by acoustic differences between clear and casual speech and psychoacoustic studies on the peripheral auditory system. The transforming function increases the

modulation depth of low modulation frequencies (2-8Hz) of the temporal envelopes, as it naturally happens on clear speech. However, unlike other studies the intensity envelope is not extracted using filterbank analysis on frequency bands. On the other hand, the time-varying amplitudes of the quasi-harmonics are extracted using an AM-FM decomposition algorithm. This alleviates possible distortions of the envelope during the carrier and envelope extraction process (Atlas and Janssen, 2005; Won et al., 2008) since the eaQHM model, used for analysis-synthesis, is highly adaptive to the signal parameters (amplitudes, frequencies and phases). Statistical analysis has also shown that DMod and SSDRC are not statistically different groups. However, SSDRC seems to have a slight advantage over DMod. Analysis on the SSDRC signals reveals that SSDRC increases the modulation depth of the intensity envelope of the unmodified speech, to a less extent though, than DMod. However, this lower modulation depth possibly is not the reason for the slight intelligibility benefit of SSDRC over DMod. This is supported by the fact that lower modulation depths deriving from the DMod modification scheme with e.g $\gamma = 0.8$, lead to lower intelligibility of speech in noise (tested via informal listening tests).

The work presented in this Chapter addresses directly and efficiently the problem of increasing the modulation depth of the temporal envelopes and simultaneously enhancing speech intelligibility. This problem has been examined by previous studies (Krause and Braida, 2009; Kusumoto et al., 2005) with moderate results on intelligibility and limitations imposed by the designed modulation filters. Equally important to enhancing the modulations of the temporal envelopes is the finding that spectral modifications change the low frequency modulations of the temporal envelopes. This suggests that the intelligibility benefit of such modifications is due to the modulation enhancement rather than the spectral energy redistribution itself. Having an algorithm that directly manipulates the low-frequency modulations on specific harmonics is very important. Using DMod, further investigation can be performed to explore which frequency regions are the most important for higher intelligibility benefits. This can be beneficial especially for hearing impaired listeners whose hearing loss affects specific frequency bands.

Last but equally important is to explore modifications that enhance speech intelligibility while preserving its quality. DMod degrades the quality of the speech signal as SSDRC does. This motivates us to explore in the future a different transforming function of DMod or a transforming function with γ values proportional to noise level. Also, the transforming function could be applied to specific harmonics. It is possible that for higher frequency regions (brilliance range) the modulation enhancement technique is rather invasive. As clear speech dictates, the modulation boosting in this region is rather small. DMod paves the way for further research on speech intelligibility and modulation enhancement for a variety of applications. Detailed steps for future directions are commented on the last chapter.

Chapter 7

Conclusions and Future Directions

In this thesis, we studied the problem of enhancing the intelligibility of speech based on natural modifications that humans do when the listener with whom they communicate faces a communication barrier. The intelligibility benefit of clear speech in various conditions and for many populations was the motivation to propose clear-based modifications on casual speech. However, transforming casual speech to clear speech is not an easy task. Indeed, comparative acoustic analysis performed on the two speaking styles, clear and casual, in English language revealed acoustic feature differences. However, not all features are connected to intelligibility. Detecting the features that make clear speech more intelligible than casual is not trivial. Moreover, while many explored modifications from the one speaking style to the other were advantageous in terms of implementation they did not prove advantageous in terms of intelligibility. This thesis,

- has attempted to quantify the contribution of human prosodic modifications to the intelligibility of clear speech. Results suggest that the duration is important for intelligibility, impacting to a greater extent the non-native population rather than the native listeners. Focusing on speech intelligibility enhancement, this work explored different acoustically-driven time-scaling algorithms to slow down casual speech. The three techniques explored are segmental time-scaling, uniform time-scaling and a novel method for time-scaling and pause insertion, the Perceptual Speech Quality Measure. This new introduced technique tries to mimic the acoustic properties of clear speech by inserting pauses and elongating the stationary parts of speech to avoid artifacts. However, although this approach is inspired by clear speech properties, it does not require any knowledge of the clear speech counterpart to insert pauses to the casual signal. On the contrary, it considers the acoustics of casual speech (loudness and stationarity) to insert pauses and to elongate. Subjective evaluations in SSN and reverberant environments show great variability among listeners in the preference of the techniques. However, the general trend is that none of the techniques was able to significantly increase the intelligibility of casual speech in SSN. The uniformly time-scaled casual speech, overall, had almost the same intelligibility with unmodified casual speech, with a slight improvement on the non-native population, whereas segmental time-scaling and PSQ decreased the intelligibility of the unmodified casual speech. Thus, this work suggests that acoustically-driven time-scaling modifications based on loudness criteria may not be appropriate for

increasing the intelligibility of casual speech in SSN. On reverberant environments, the uniform time-scaling technique was proven advantageous for the non-native population when combined with spectral modifications. Results indicate the connection of the uniform-time scaling factor to the degree of reverberation for successful intelligibility enhancement.

- presented and evaluated a novel approach to expand vowel space via frequency warping inspired by clear speech analyses. Unlike other studies, we perform vowel space expansion in isolation in order to evaluate its impact on intelligibility. Our modification scheme is explicit and speaker independent. Results indicate that, while vowel space is successfully expanded, there is no significant intelligibility gain from the frequency warping. These observations suggest that further detailed evaluation of the clear speech intelligibility gain related to vowel space expansion is merited. In particular, a more localized level of analyses and consequent modifications (e.g. within phones and considering formant transitions) might expose more perceptually relevant differences that positively impact intelligibility.

- performed acoustic analyses on Lombard and clear speaking styles in order to isolate and compare pertinent spectral phenomena that later inspire speech modifications to increase intelligibility. While Lombard speech consistently exhibits spectral energy boosting in an inclusive formant region, effectively increasing loudness, clear speech shows spectral energy redistribution on higher frequency regions than Lombard, possibly attributed to consonant emphasis and hyper-articulation. Examined in terms of intelligibility, both styles exhibit an intelligibility benefit compared to casual speech, with Lombard being more effective in lower SNR and clear speech in higher SNR SSN conditions. However, the intelligibility gain of clear speech on different populations and on a variety of difficult conditions motivated us to propose modifications based on the spectral characteristics of clear speech, namely the Mix-filtering method. The Mix-filtering approach proposed in this work has multiple benefits on (1) intelligibility tested in SSN and reverberant environments combined with time-scaling modifications for native and non-native listeners (as well as for hearing-impaired but this is a suggestion and cannot be proven statistically) (2) quality, keeping the quality of original speech unaffected compared to SSDRC (3) complexity, since it is speaker and speech independent and does not require frame-by-frame analysis and voice-unvoiced discrimination.

- proposes a novel method for enhancing speech intelligibility by increasing the modulation depth of the time-varying amplitude components of speech. To the best of our knowledge, there has been no similar study that addresses directly and effectively the modulation boosting of the temporal envelopes with an intelligibility impact of more than 30% for native listeners. The proposed modification scheme is based on a powerful analysis and synthesis Sinusoidal model, the eaQHM. The instantaneous amplitudes are extracted from eaQHM and are modified using a transforming function that approaches clear speech characteristics and perceptual attributes. Then, the signal is reconstructed by the eaQHM model using the modified amplitudes and the unmodified instantaneous frequencies and phases. Intelligibility tests

from native listeners in SSN of low (-8dB) and mid (-2dB) SNR have shown significant intelligibility improvement of speech using the proposed method.

This thesis suggests that there is still much room for the intelligibility enhancement of speech, while the modifications proposed open up several avenues for future work. Specifically,

- The most beneficial modification technique proposed for intelligibility enhancement in noise is based on modulation boosting of specific frequency regions (quasi-harmonics). The intelligibility impact of this technique is evaluated on native listeners in SSN. However, this technique may be also beneficial:
  - for hearing impaired listeners. The transforming function may compensate for the loss of the non-linear compression of the basilar membrane (Moore and Glasberg, 2004) on subjects with hearing loss. Furthermore, using DMod, modulation boosting can be selectively performed on specific frequency regions (quasi-harmonics) where the listener is not deaf, "saving" energy from spectral regions non-beneficial to the subject's hearing.
  - for reverberant environments. It has been shown that when speech is presented in reverberation, the peak of the modulation spectrum shifts to a lower modulation frequency (Houtgast and Steeneken, 1985). Nonetheless, modulation features are more robust against room reverberation than conventional cepstral and dynamic features, motivating previous studies on using modulation filtering in order to enhance speech intelligibility (Kusumoto et al., 2005). Our proposed modification performs modulation enhancement and therefore, can be proven advantageous for speech intelligibility in reverberation. The efficiency of DMod can be also evaluated on other types of noise (competitive speaker, babble noise etc).
- DMod like SSDRC reduces the quality of original speech. It is important, therefore, to focus on the quality of our modulation enhancement scheme. This can be performed by manipulating the transforming function. The transforming function increases the "noise" on the temporal envelopes. One way to confront the noisy components is to isolate the low frequency modulations and add them back to the temporal envelope. Moreover, different transforming functions or combinations of the γ value with the harmonic components or/and the noise type/level are to be explored in the future for a better intelligibility and quality outcome.
- The merit of the modification techniques proposed in this work in SSN and reverberation should be also explored for other types of noise and speech. Indeed, SSN noise simulates babble noise which is the most common masker in real environments. However, SSN is a rather strong masker which may mask quality degradations of the modifications. Furthermore, our proposed modifications have been evaluated in recorded speech in two different languages English and Greek. It would be very interesting to extend the evaluation of this work to other non-tonal languages and verify the amount of intelligibility enhancement in spontaneous speech. It is expected that with small adaptations (i.e using a voice activity

detection algorithm to avoid e.g respiration areas in spontaneous speech and therefore degradations) the modification techniques proposed will also be beneficial.

- Time-scaling approaches failed to increase the intelligibility of speech in SSN, while in reverberant environments changing the duration of the signal has been proven advantageous. Considering that time-scaling transformations change the modulation frequencies of each temporal envelope of the speech signal, it is worth exploring modulation-based time-scaling techniques for intelligibility enhancement. This is also suggested by the work of Kusumoto et al. (2005), where the efficiency of the modulation filtering approach was dependent on the reverberation condition and it is also supported by our study, by revealing a connection of the time-scaling factor with the reverberation time (also to be proved in the future with intelligibility evaluations). Therefore, we suggest as a future work, for enhancing the intelligibility of speech in reverberation, a simultaneous change of the modulation frequency of the signal dependent on the reverberation time with a parallel boosting of modulation depth. For SSN noise, the time-scaling techniques proposed, should be examined in terms of modulation frequencies and amplitudes and should be compared with that of clear speech to deeply comprehend why these techniques failed to increase intelligibility.

- Our frequency warping method has been proven successful in expanding the vowel space but not in enhancing intelligibility. Possibly this is attributed to the fact that the generalized curve of warping shifts derives from the formant analysis on the center of vowels. However, equally important are the beginning and ending of the vowel, where the co-articulation of consonant and vowel takes place. Exploring relative differences on the vowel spaces between clear and casual in these transition regions may result to a beneficial modification scheme for intelligibility. Moreover, a combination of vowel space expansion with other modifications could be proven advantageous. Treating each modification in isolation indeed could help distinguish the intelligibility impact of the method, however this is not necessarily correct since the combination of modifications may be beneficial. Last, it should be noted that the vowel space expansion method proposed in this work is rather generalized, averaging many speakers. It is possible that performing vowel space expansion using a frequency warping function for each individual speaker or modifying only content words rather than function words could be advantageous. However, the motivation of this work was to perform speaker independent and speech independent modifications (even though this is partially true for vowel space expansion since adaptation would be required for different languages).

- The transforming function of DMod is motivated by the non-linear compression function in the basilar membrane. This suggests that for the intelligibility enhancement of speech, modification algorithms could benefit from recent advances in the field of psychoacoustics.

- The results of our study can be expanded to other applications besides intelligibility enhancement. Using eaQHM we are able to manipulate the temporal envelopes of speech, perform modifications and

re-synthesize speech without imposing artifacts due to the analysis and re-synthesis process. However, temporal envelopes carry both prosodic and spectral information, important for the perception of speech in terms of intelligibility and emotions. Exploring the time-varying characteristics of the temporal envelopes and the modulation amplitudes and frequencies for each emotional state can introduce a novel method for emotional-based applications (emotion recognition and modification).

Appendix

.1 Publications

Conference Papers, Journals and Workshops

- Maria Koutsogiannaki and Yannis Stylianou. Intelligibility enhancement of casual speech based on clear speech properties. On preparation.
- Maria Koutsogiannakiand Yannis Stylianou: Quasi-Harmonic Amplitude Modulation Enhancement for Increasing Speech Intelligibility. Under submission.
- Maria Koutsogiannaki, Petko Petkov and Yannis Stylianou: Intelligibility Enhancement of Casual Speech for Reverberant Environments inspired by Clear Speech Properties. **Interspeech 2015** Dresden, Germany.
- Maria Koutsogiannaki and Yannis Stylianou: Simple and Artefact-free Spectral Modifications for Enhancing the Intelligibility of Casual Speech. In: **ICASSP 2014**, Frorence, May 2014.
- Elizabeth Godoy, Maria Koutsogiannaki, Yannis Stylianou: Approaching speech intelligibility enhancement with inspiration from Lombard and Clear speaking styles. **Journal: Computer Speech and Language** 28(2): 629-647 (2014)
- Elizabeth Godoy, Maria Koutsogiannaki, Yannis Stylianou: Assessing the intelligibility impact of vowel space expansion via clear speech-inspired frequency warping. In: **Interspeech 2013**, pp. 1169-1173, Vancouver Kanada 2013.
- Koutsogiannaki, M., Pettinato, M, Mayo, C, Kandia, V. and Stylianou, Y. (2012). Can modified casual speech reach the intelligibility of clear speech? In: **Interspeech 2012**, Portland Oregon, USA, 9-13 September 2012
- Stylianou, Y., Hazan, V., Aubanel, V., Godoy, E., Granlund, S., Huckvale, M., Jokinen, E., Koutsogiannaki, M., Mowlaee, P., Nicolao, M., Raitio, T., Sfakianaki, A. and Tang, Y. P8-Active Speech Modifications. Final Project Report. In Proceedings of the 8th International Summer Workshop on Multimodal Interfaces 2013, pp. 61-82, **eNTERFACE '12**

- Koutsogiannaki, M., Mayo, C, Kandia, V., and Stylianou, Y. (2012). On the detection of the intelligibility advantage of clear speech vs. casual speech. The listening Talker - An interdisciplinary workshop on natural and synthetic modification of speech, **LISTA Workshop**, 2-3 May 2012, Edinburgh.

.2 Acoustic material

In the following url the reader can find acoustic samples of the beneficial modifications proposed in this work: http://www.csd.uoc.gr/ mkoutsog/sound.php

Bibliography

Allen, J. and Berkley, D. (1979). Image method for efficiently simulating small–room acoustics. *J. Acoust. Soc. Am.*, 65(4):943–950.

Amano-Kusumoto, A. and Hosom, J. (2011). 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)*.

ANSI-S3.5-1997 (1997). American national standard methods for calculation of the speech intelligibility index. Technical report, American National Standards Institute, New York.

Arai, T. (2005). Padding zero into steady-state portions of speech as a preprocess from improving intelligibility in reverberant environments. *Acoust. Sci. Tech.*, 25(5):459–461.

Arai, T., Kinoshita, K., Hodoshima, N., Kusumoto, A., and Kitamura, T. (2002). Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments. *Acoust. Sci. Tech.*, 23(4):229–232.

Arai, T., Nakata, Y., Hodoshima, N., and Kurisu, K. (2007). Decreasing speaking rate with steady-state suppression to improve speech intelligibility in reverberant environments. *Acoust. Sci. Tech.*, 28(4):282–285.

Atlas, L. and Janssen, C. (2005). Coherent modulation spectral filtering for single-channel music source separation. *ICASSP*, 4:461–464.

Bacon, S. and Grantham, D. (1989). Modulation masking: Effects of modulation frequency, depth and phase. *J. Acoust. Soc. Am.*, 85:2575–2580.

Baker, R. and Hazan, V. (2010). Lucid: a corpus of spontaneous and read clear speech in british english. *DiSS-LPSS*, pages 3–6.

Baltazani, M. (2007). Prosodic rhythm and the status of vowel reduction in greek. *17th International Symposium on Theoretical and Applied Linguistics, Department of Theoretical and Applied Linguistics, Salonica, Greece*, pages 31–43.

Blesser, B. (1969). Audio dynamic range compression for minimum perceived distortion. *IEEE Transactions on Audio and Electroacoustics*, 17(1):22–32.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Bond, Z. and Moore, T. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication*, 14:325–337.

Bond, Z., Moore, T., and Gable, B. (1989). Acoustic phonetic characteristics of speech produced in noise and while wearing an oxygen mask. *J. Acoust. Soc. Am.*, 85:907–912.

Bradlow, A. and Bent, T. (2002). The clear speech effect for non-native listeners. *J. Acoust. Soc. Am.*, 112(1):272–284.

Bradlow, A., Clopper, C., Smiljanic, R., and Walter, M. (2010). A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech communication*, 52(11-12):930–942.

Bradlow, A., Torretta, G., and Pisoni, D. (1996). Intelligibility of normal speech i: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20:255–272.

Bradlow, A. R., Kraus, N., and Hayes., E. (2003). Speaking clearly for learning-impaired children: sentence perception in noise. *Journal of Speech, Language, and Hearing Research*, 46:80–97.

Brons, I., Houben, R., and Dreschler, W. (2014). Effects of noise reduction on speech intelligibility, perceived listening effort, and personal preference in hearing-impaired listeners. *Trends in Hearing*, 18.

Burchfield, L. and Bradlow, A. (2014). Syllabic reduction in mandarin and english speech. *Journal of the Acoustical Society of America*, 135(6):270–276.

Camacho, A. and Harris, J. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *J. Acoust. Soc. Am.*, 124:1638–1652.

Charpentier, F. and Stella, M. (1986). Diphone synthesis using an overlap-add technique for speech waveforms concatenation. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86*, 11:2015–2018.

Chen, F. (1980). Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level. *Master Thesis, MIT, Cambridge*.

Cooke, M., Barker, J., Cunningham, S., and Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.*, 120:2421–2424.

Cooke, M. and Lecumberri, M. (2012a). The intelligibility of Lombard speech for non-native listeners. *J. Acoust. Soc. Amer., Letters to the Editor*, 132:1120–1129.

Cooke, M. and Lecumberri, M. (2012b). The intelligibility of Lombard speech for non-native listeners. *J. Acoust. Soc. Amer., Letters to the Editor*, 132:1120–1129.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., and Tang, Y. (2013). Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55(4):572–585.

Coyle, E., Donnellan, O., Jung, E., Meinardi, M., Campbell, D., MacDonailli, C., and Leung, P. (2004). Intelligibility of modified speech for young listeners with normal and impaired hearing. *In Proceedings of the 5th Intl. Conf. Disability, Virtual Reality and Assoc. Tech. Oxford*.

Cutler, A. and Butterfield, S. (1990). Durational cues to word boundaries in clear speech. *Speech Communication*, 9:485–495.

Delattre, P. (1969). An acoustic and articulatory study of vowel reduction in four languages. *International Review of Applied Linguistics in Language Teaching*, 7:295–325.

Demol, M., Struyve, K., Verhelst, W., Paulussen, H., Desmet, P., and Author, P. V. (2004). Efficient non-uniform time-scaling of speech with wsola for call applications. *Proceedings of InSTIL/ICALL2004 • NLP and Speech Technologies in Advanced Language Learning Systems*.

Drugman, T. and Alwan, A. (2011). Joint robust voicing detection and pitch estimation based on residual harmonics. *Interspeech*.

Drullman, R., Festen, J., and Plomp, R. (1994a). Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.*, 95(5):2070–2680.

Drullman, R., Festen, J., and Plomp, R. (1994b). Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, 95(2):1053–1064.

Erro, D., Moreno, A., and Bonafonte, A. (2010). Voice conversion based on weighted frequency warping. *IEEE Trans Audio, Speech, Lang Processing*, 18(5):922–931.

Erro, D., Zorila, T., and Stylianou, Y. (2014). Enhancing the intelligibility of statistically generated synthetic speech by means of noise-independent modifications. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 22(12):2101–2111.

Ewert, S. D. and Dau, T. (2000). Characterizing frequency selectivity for envelope fluctuations. *J. Acoust. Soc. Am.*, 108:1181–1196.

Ferguson, S. (2004). Talker differences in clear and conversatoinal speech: Vowel intelligibility for normal-hearing listeners. *J. Acoust. Soc. Am.*, 116(4):2365–2373.

Ferguson, S. and Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 112:259–271.

Fraser, S. (1999). *Handbook for Acoustic Ecology*. University and ARC Publications, 2 edition.

Garnier, M., Bailly, L., Dohen, M., Welby, P., and Loevenbruck, H. (2006). An acoustic and articulatory study of Lombard speech: global effects on the utterance. *Interspeech*.

Garnier, M., Henrich, N., and Dubois, D. (2010). Influence of sound immersion and communicative interaction on the Lombard effect. *Journal of Speech, Language and Hearing Research*, 53:588–608.

Ghitza, O. and Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, 66(1-2):113–126.

Godoy, E., Rosec, O., and Chonavel, T. (2012). Ieee trans audio, speech, lang processing. *Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora*, 20(4):1313–1323.

Godoy, E. and Stylianou, Y. (2012). Unsupervised acoustic analyses of normal and Lombard speech, with spectral envelope transformation to improve intelligibility. *Interspeech*.

Gordon-Salant, S. (1986). Recognition of natural and time/intensity altered cvs by young and elderly subjects with normal hearing. *J. Acoust. Soc. Am.*, 82(6):1599–1607.

Gordon-Salant, S. (1987a). Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects. *J. Acoust. Soc. Am.*, 81(4):1199–1202.

Gordon-Salant, S. (1987b). Review of text-to-speech conversion for english. *J. Acoust. Soc. Am.*, 82(1):737–793.

Harlan, L. and Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, 14:677–709.

Harris, J. and Skowronski, M. (1988). Energy redistribution speech intelligibility enhancement, vocalic and transitional cues. *jasa*, pages 2305–2305.

Hazan, V. and Baker, R. (2010). Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style? *DiSS-LPSS*, pages 7–10.

Hazan, V. and Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *J. Acoust. Soc. Am.*, 130:2139–2152.

Hazan, V. and Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *Journal of the American Academy of Audiology*, 116(5):3108–3118.

Hazan, V. and Simpson, A. (1998). The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24:211–226.

Hillenbrand, J. M. and Clark, M. J. (2000). Some effects of duration on vowel recognition. *J. Acoust. Soc. Am.*, 108(6):3013–3022.

Hillenbrand, T. and Nearey, T. (1999). Identification of resynthesized /hvd/ utterances: Effects of formant contour. *J. Acoust. Soc. Am.*, 406(6):3509–3523.

Hodoshima, N., Behne, D., and Arai, T. (2006). Steady-state suppression in reverberation: a comparison of native and non-native speech perception. *Interspeech 2006*, pages 873–876.

Hoemeke, K. A. and Diehl, R. L. (1994). Perception of vowel height: The role of f1-f0 distance. *J. Acoust. Soc. Am.*, 96:661–674.

Houtgast, T. and Steeneken, H. (1973). The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acustica*, 28:66–73.

Houtgast, T. and Steeneken, H. (1985). A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.*, 77:1069–1077.

http://speakgreek.web.auth.gr/ (2007-2013). http://speakgreek.web.auth.gr/wp/research-goals/.

Imai, S. (1983). Cepstral analysis synthesis on the mel frequency scale. *ICASSP*, 8:93–96.

Jayan, A., Pandey, P., and Lehana, P. (2008). Automated detection of transition segments for intensity and time-scale modification for speech intelligibility enhancement. *IEEE International Conference on Signal Processing, Communications and Networking*, pages 63–68.

Junqua, J. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *J. Acoust. Soc. Am.*, 93:510–524.

Kafentzis, G., Pantazis, Y., Rosec, O., and Stylianou, Y. (2012). An extension of the adaptive quasi-harmonic model. *ICASSP*, pages 4605–4608.

Kafentzis, G., Rosec, O., and Stylianou, Y. (2014). Robust full-band adaptive sinusoidal analysis and synthesis of speech. *ICASSP*, pages 6260–6264.

Kates, J. (1994). Speech enhancement based on a sinusoidal model. *Journal of Speech and Hearing Research*, 37:449–464.

Kemper, S. and Harden, T. (1999). Experimentally disentangling what's beneficial about elder-speak from what's not. *Psychology and Aging*, 14:656–670.

Kowalski, N., Depireux, D., and Shamma, S. (1996). Analysis of dynamic spectra in ferret primary auditory cortex: I. characteristics of single unit responses to moving ripple spectra. *Journal of Neurophysiology*, 76:3503–3523.

Krause, J. (2001). Properties of naturally produced clear speech at normal rates and implications for intelligibility enhancement. *Doctoral Thesis, MIT, Cambridge*.

Krause, J. and Braida, L. (2004a). Acoustic properties of naturally produced clear speech at normal speaking rates. *J. Acoust. Soc. Am.*, 115:362–378.

Krause, J. and Braida, L. (2004b). Investigating alternative forms of clear speech: the effects of speaking rate and speaking mode on intelligibility. *J. Acoust. Soc. Am.*, 112(5):2165–2172.

Krause, J. and Braida, L. (2009). Evaluating the role of spectral and envelope characteristics in the intelligibility advantage of clear speech. *J. Acoust. Soc. Am.*, 125(5):3346–3353.

Kusumoto, A., Kinoshita, T., Hodoshima, K., and Vaughan, N. (2005). Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments. *Speech Communication*, 45:101–113.

Laan, G. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous ans a read speaking style. *Speech Communication*, 22(1):43–965.

Ladefoged, P. and Johnson, K. (2010). *A course in phonetics*. Wadsworth Cengage Learning, 6 edition.

Langner, B. and Black, A. (2005). Improving the understandability of speech synthesis by modeling speech in noise. *ICASSP*, pages 265–268.

Laures, J. and Bunton, K. (2003). Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions. *Journal of Communication Disorders*, 36:449–464.

Lecumberri, M. (2012). The intelligibility of Lombard speech for non-native listeners. *J. Acoust. Soc. Am.*, 132:1120–1129.

Letowski, T., Frank, T., and Caravella, J. (1993). Acoustical properties of speech produced in noise presented through supraaural earphones. *Ear and Hearing*, 14:332–338.

Lindblom, B. (1990). *Speech Production and Speech Modelling*, chapter Explaining Phonetic Variation: A Sketch of the HH Theory, pages 403–439. Springer Netherlands, Dordrecht.

Liu, Q., Champagne, B., and Kabal, P. (1996). A microphone array processing technique for speech enhancement in a reverberant space. *Speech Communication*, 18:317–334.

Liu, S. and Zeng, F. (2006). Temporal properties in clear speech perception. *J. Acoust. Soc. Am.*, 120(1):424–432.

Lombard, E. (1911). Le signe de l'elevation de la voix, annals maladiers oreille. *Larynx Nez Pharynx*, 37:101–119.

Lu, Y. and Cooke, M. (2008). Speech production modifications produced by competing talkers, babble and stationary noise. *J. Acoust. Soc. Am.*, pages 3261–3275.

Lu, Y. and Cooke, M. (2009). The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, pages 1253–1262.

Mayo, C., Aubanel, V., and Cooke, M. (2012). Effect of prosodic changes on speech intelligibility. *Interspeech 2012, 9-13 September, Portland, Oregon*.

McHenry, M. (2011). An exploration of listener variability in intelligibility judgments. *American Speech Language and Hearing Association*, 20:119–123.

M.Cooke (2006). A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*, 119:1562–1573.

Mesgarani, N. and Shamma, S. (2005). Speech enhancement based on filtering the spectrotemporal modulations. *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 1:1105–1108.

Meunier, C. and Espesser, R. (2011). Vowel reduction in conversational speech in french: The role of lexical factors. *Journal of Phonetics*, 39(3):271–278.

Mohammadi, S., Kain, A., and Santen, J. (2012). Making conversational vowels more clear. *Interspeech 2012*.

Monsen, R. (1978). Toward measuring how well hearing-impaired children speak. *Journal of Speech, Language and Hearing Research*, 21(2):197–219.

Monsen, R. (1982). The oral speech intelligibility of hearing-impaired talkers. *Journal of Speech, Language and Hearing Research*, 48(3):286–296.

Montgomery, A. and Edge, R. (1988). Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults. *Journal of Speech and Hearing Research*, 31:386–393.

Moon, S. and Lindblom, B. (1994). Interaction between duration, context, and speaking style in english stressed vowels. *jasa*, 96:40–55.

Moore, B. and Oxenham, A. (1998). Psychoacoustic consequences of compression in the peripheral auditory-system. *Psychological review*, 105(1):108–124.

Moore, B. C. J. and Glasberg, B. R. (2004). A revised model of loudness perception applied to cochlear hearing loss. *Hearing Research*, 132:70–88.

Nabelek, A. K., Letowski, T. R., and Tucker, F. M. (1989). Reverberant overlap- and self-masking in consonant identification. *J. Acoust. Soc. Am.*, 86(4):1259–1265.

Nakata, Y., Murakami, Y., Hodoshima, N., Hayashi, N., Miyauchi, Y., Arai, T., and Kurisu, K. (2006). The effects of speech-rate slowing for improving speech intelligibility in reverberant environments. *International Workshop on Frontiers in Speech and Hearing Research, Technical Report of IEICE Japan*.

Narne, V. and Vanaja, C. (2008). Effect of envelope enhancement on speech perception in individuals with auditory neuropathy. *Ear and Hearing*, 29:45–53.

Nejime, Y., Aritsuka, T., Imamura, T., Ifukubei, T., and Matsushima, J. (1996). A portable digital speech-rate converter for hearing impairment. *IEEE Transactions Rehabilitation. Engineering*, 4:73–83.

Nejime, Y. and Moore, B. (1998). Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss. *J. Acoust. Soc. Am.*, 103:572–576.

Nelson, P. B., Jin, S., Carney, A. E., and Nelson, D. A. (2003). Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. *Journal of the Acoustical Society of America*, 113:961–968.

Nicolaidis, K., Papanikolaou, G., Avdelidis, K., Kainada, E., Sfakianaki, A., Vrisis, L., Konstantoudakis, K., Starchenko, I., and Kelmali, E. (2014). Speakgreek: Development of an online speech training system. *Proceedings of the 7th Panhellenic Conference "Acoustics 2014"*.

Nicolaidis, K., Papanikolaou, G., Kainada, E., and Avdelidis, K. (2015a). Speakgreek: An online speech training tool for l2 pedagogy and clinical intervention. *Accepted at the 18th International Congress of Phonetic Sciences*.

Nicolaidis, K., Papanikolaou, G., Sfakianaki, A., Kainada, E., Avdelidis, K., and Konstantoudakis, K. (2015b). Computer assisted teaching of vowel production to learners of greek as an l2 and individuals with speech disorders. *22nd International Symposium on Theoretical and Applied Linguistics, Aristotle University of Thessaloniki*.

Niederjohn, R. and Grotelueschen, J. (1976). The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. *IEEE Trans. Audio Speech Lang. Process*, 24:277–282.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.*, 95:1085–1099.

Paliwal, K., Wójcicki, K., and Schwerin, B. (2010). Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech Communication*, 52 (5):450–475.

Pantazis, Y., Rosec, O., and Stylianou, Y. (2011). Adaptive AM–FM signal decomposition with application to speech analysis. *IEEE Trans. on Audio, Speech and Language Processing*, 19:290–300.

Payton, K., Uchanski, R., and Braida, L. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *J. Acoust. Soc. Am.*, 95(3):1581–1592.

Picheny, M., Durlach, N., and Braida, L. (1985a). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech, Language and Hearing Research*, 28(1):96–103.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1985b). Speaking clearly for the hard of hearing i: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, 28:96–103.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1986). Speaking clearly for the hard of hearing ii: acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, 29:434–446.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1989). Speaking clearly for the hard of hearing iii: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *Journal of Speech and Hearing Research*, 32:600–603.

Pittman, A. L. and Wiley, T. L. (2001). Recognition of speech produced in noise. *Journal of Speech Language and Hearing Research*, 44:487–496.

Qin, M. K. and Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *Journal of the Acoustical Society of America*, 114:446–454.

Quatieri, T. and McAulay, R. (1991). Peak-to-rms reduction of speech based on a sinusoidal model. *IEEE Transactions on Audio and Electroacoustics*, 39:273–288.

Rabbitt, P. (1968). Channel capacity, intelligibility, and immediate memory. *Quarterly Journal of Psychology*, 20:241–248.

Rabbitt, P. (1991). Mild hearing loss can cause apparent memory failures which increase with age and reduce with iq. *ActaOtolaryngologia*, 476:167–176.

Raitio, T., Suni, A., Vainio, M., and Alku, P. (2011). Analysis of HMM-based Lombard speech synthesis. *Interspeech*, pages 2781–2784.

Rennies, J., Schepker, H., Holube, I., and Kollmeier, B. (2014). Listening effort and speech intelligibility in listening situations affected by noise and reverberation. *Journal of the Acoustical Society of America*, 136(5):2642–53.

Rherbergen, K. S. and Versfeld, N. J. (2005). Speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *JASA*, pages 2181–2192.

Roebel, A. and Rodet, X. (2005). Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. *Digital AudioEffects*, pages 30–35.

Roebel, A., Villavicencio, F., and Rodext, X. (2007). On cepstral and all-pole based spectral envelope modeling with unknown model order. *Pattern Recognition Letters*, 28(11).

Rossing, T. (2007). *Springer handbook of Acoustics*. Springer.

Sauert, B. and Vary, P. (2006). Near end listening enhancement: speech intelligibility improvement in noisy environments. *ICASSP*, pages 493–496.

Schimmel, S. and Atlas, L. (2005). Coherent envelope detection for modulation filtering of speech. *ICASSP*, pages 221–224.

Schmitt, J. F. (1983). The effects of time compression and time expansion on passage comprehension by elderly listeners. *Journal of Speech and Hearing Research*, 26:373–377.

Schonwiesner, M. and Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fmri. *Proceedings of National Academy of Sciences*, 106:14611–14616.

Sfakianaki, A., Nicolaidis, K., and Okalidou, A. (2012). Intelligibility and production in greek hearing impaired speech. *The Listening-talker workshop, Edinburgh*.

Sfakianaki, A. M. (2012). An acoustic study of coarticulation in the speech of greek adults with normal hearing and hearing impairment. *PhD thesis in Linguistics, Aristotle University of Thessaloniki*.

Shamma, S. (1996). Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method. *Network: Computation in Neural Systems*, 7:439–476.

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

Sheft, S., Ardoint, M., and Lorenzi, C. (2008). Speech identification based on temporal fine structure cues. *J. Acoust. Soc. Am.*, 124(1):562–575.

Skowronski, M. and Harris, J. (2006). Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communication*, pages 549–558.

Small, J. A., S., K., and K., L. (1997). Sentence comprehension in alzheimer's disease: Effects of grammatical complexity, speech rate and repetition. *Psychology and Aging*, 12:3–11.

Smiljanic, R. and Bradlow, A. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistic Compass*, 3(1):236–264.

Smith, C. (1982). Differences between read and spontaneous speech of deaf children. *J. Acoust. Soc. Am.*, 72(4):1304–06.

Smith, Z., Delgutte, B., and Oxenham, A. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Letters to Nature*, 416:87–90.

Sommers, M. and Barcroft, J. (2006). Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *J. Acoust. Soc. Am.*, 119(4):2406–2416.

Steeneken, H. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *J. Acoust. Soc. Am.*, 67:318–326.

Stylianou, Y., Hazan, V., Aubanel, V., Godoy, E., Granlund, S., Huckvale, M., Jokinen, E., Koutsogiannaki, M., Mowlaee, P., Nicolao, M., Raitio, T., Sfakianaki, A., and Tang, T. (2012). P8-active speech modifications. *eNTERFACE '12 Metz, France*.

Summers, W., Pisoni, D., Bernacki, R., Pedlow, R., and Stokes, M. (1988). Effects of noise on speech production: Acoustic and perceptual analysis. *J. Acoust. Soc. Am.*, 84:917–928.

Tang, Y. and Cooke, M. (2011a). Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. *Interspeech*, pages 345–348.

Tang, Y. and Cooke, M. (2011b). Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. *Florence, Italy*, pages 345–348.

Tang, Y., Cooke, M., and Valentini-Botinhao, C. (2013). P16: A distortion-weighted glimpse-based intelligibility metric for modified and synthetic speech. *Speech in Noise Workshop*.

Thomas, T. (1996). Experimental evaluation of improvement in speech perception with consonantal intensity and duration modification. *Ph. D. Thesis, Indian Institute of Technology*.

Todd, N. and Brown, G. (1994). A computational model of prosody perception. *Proc. IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, 10:127–130.

Todd, N. and Brown, G. (1996). Visualization of rhythm time and meter. *Artificial Intelligence Review*, 10:91–113.

Turner, C., Smith, S., Aldridge, P., and S.L. (1997). Formant transition duration and speech recognition in normal and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 101(5):2822–2825.

Uchanski, R. (2005). *Clear Speech*. Blackwell Publishing Ltd, UK.

Uchanski, R., Choi, S., Braida, L., Reed, C., and Durlach, N. (1996a). Speaking clearly for the hard of hearing IV: further studies of the role of speaking rate. *J. of Speech and Hearing*, 39:494–509.

Uchanski, R., Choi, S., Braida, L. D., Reed, C. M., and Durlach, N. I. (1996b). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech and Hearing Research*, 39(3):494–509.

Uchanski, R., Geersi, A., and Protopapas, A. (2002). Intelligibility of modified speech for young listeners with normal and impaired hearing. *Journal of Speech, Language, and Hearing Research*, 45:1027–1038.

Valbret, H., E., M., and Tubach, J. (1992). Voice transformation using psola technique. *Speech Communication*, 11(2-3):175–187.

Valentini-Botinhao, C., Yamagishi, J., and King, S. (2011). Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise? *Interspeech*.

Valentini-Botinhao, C., Yamagishi, J., and King, S. (2012). Evaluating speech intelligibility enhancement for HMM - based synthetic speech in noise. *Proc. SAPA Workshop, Portland, USA*.

van der Horst, R., Leeuw, A., and Dreschler, W. (1999). Importance of temporal-envelope cues in consonant recognition. *J. Acoust. Soc. Am.*, 105(3):1801–1809.

Watson, P. and Schlauch, R. (2008). The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours. *American Journal of Speech-Language Pathology*, 17(4):348–355.

Wójcicki, K. and Loizou, P. (2012). Channel selection in the modulation domain for improved intelligibility in noise. *Journal of the Acoustical Society of America*, 131 (4):2904–2913.

Won, J. H., Schimmel, S. M., Drennan, W. R., Souza, P. E., Atlas, L., and Rubinstein, J. T. (2008). Improving performance in noise for hearing aids and cochlear implants using coherent modulation filtering. *Hearing Research*, 239:1–11.

Zeng, F., Nie, K., Liu, S., Stickney, G., Del Rio, E., Kong, Y., and Chen, H. (2004). On the dichotomy in auditory perception between temporal envelope and fine structure cues. *J. Acoust. Soc. Am.*, 116:1351–1354.

Zorila, T., Kandia, V., and Stylianou, Y. (2012). Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. *Interspeech 2012, Portland Oregon, USA*, pages 635–638.