Downloaded from orbit.dtu.dk on: May 16, 2024

DTU Library

Assessing the effects of fundamental-frequency dynamics on the intelligibility of competing voices

Mesiano, Paolo Attilio

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Mesiano, P. A. (2022). *Assessing the effects of fundamental-frequency dynamics on the intelligibility of competing voices*. DTU Health Technology. Contributions to Hearing Research Vol. 56

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

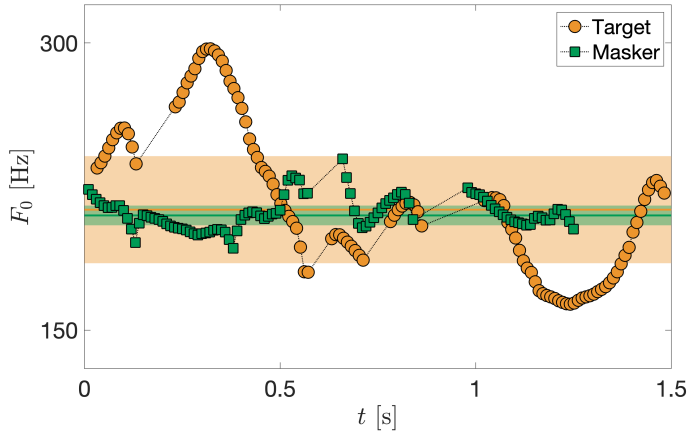
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 56

Paolo Attilio Mesiano

Assessing the effects of fundamental-frequency dynamics on the intelligibility of competing voices



DTU Health Tech
Department of Health Technology

Assessing the effects of fundamental-frequency dynamics on the intelligibility of competing voices

PhD thesis by
Paolo Attilio Mesiano

Preliminary version: July 16, 2022

Technical University of Denmark

2022

© Paolo Attilio Mesiano, 2022
Preprint version for the assessment committee.
Pagination will differ in the final published version.

This PhD dissertation is the result of a research project carried out at the Hearing Systems Section, Department of Health Technology, Technical University of Denmark (DTU).

The project was financed by the Centre for Applied Hearing Research (CAHR) and was carried out in collaboration with Eriksholm Research Centre (Oticon A/S).

Supervisors

Prof. Torsten Dau
Dr. Johannes Zaar
Dr. Helia Relaño-Iborra
Hearing Systems Section
Department of Health Technology
Technical University of Denmark
Kgs. Lyngby, Denmark

Abstract

Young listeners with a healthy auditory system are capable of understanding speech in the presence of one or several interfering voices, even in the most challenging listening scenarios. This ability is crucial for daily social life, but it is compromised in older listeners affected by sensorineural hearing loss, who often experience difficulties understanding speech in complex auditory environments, even when hearing-aid solutions are provided. Investigating auditory scenarios with multiple speakers is thus essential for revealing phenomena that can inspire the development of hearing-loss compensation strategies.

Among many auditory cues, the fundamental frequency (F_0) and its differences between competing voices provide useful information that aids target speech intelligibility, which can be successfully utilized by normal-hearing (NH) listeners, whereas older hearing-impaired (HI) listeners have limited access to it. Evidence for the efficacy of F_0 information has been obtained by means of laboratory simulations of competing-talker scenarios, with highly constrained speech stimuli that are not truly representative of the characteristics and the variety of realistic speech and therefore do not guarantee the reproducibility of the results in the diversity of auditory situations encountered in daily life.

This thesis aims to expand the available knowledge on the role of F_0-related cues in competing-talker scenarios by using naturalistic everyday speech stimuli with a wide variability of F_0 characteristics that is typical of realistic voices. The effects of differences in average F_0 and F_0 dynamic range between competing voices on speech perception were investigated in NH and HI listeners. For NH listeners, the measured effects of these cues on speech intelligibility were small or negligible. The average F_0 difference between competing voices was found to only provide a speech-intelligibility benefit when energetic cues were limited or absent, which occurs especially when the competing voices have unrealistically similar syntactical structure and F_0 trajectories. The effect on speech intelligibility induced by the difference in F_0-dynamic-range between competing voices was found to be negligible. However, it was shown that the presence of a relatively large F_0 dynamic range in at least one of the two competing sentences improved speech intelligibility, regardless of the difference in F_0 dynamic range between sentences. For HI listeners, the inability to utilize these F_0 cues was confirmed: compared to NH listeners, the benefit induced by an average F_0 separation between competing voices was smaller and no significant effect of F_0-dynamic-range of the sentences nor of their difference was observed.

Finally, an analysis of the F_0 properties of speech recordings from naturalistic dialogues was presented, providing a reference for the F_0 properties of realistic speech and describing the changes in F_0 properties that talkers produce in the presence of communication barriers such as background noise and hearing impairment.

This thesis contributes to the body of literature on the role of F_0-related cues in communication scenarios, by proposing new methodologies that focus on the realism of the speech materials and on the numerical control of the experimental method. Overall, the results of this thesis suggest that in realistic competing-talker scenarios, the F_0-related cues contribute to a holistic picture of the auditory scene that involves many auditory cues. In such scenarios, especially the F_0 dynamic range of the individual competing sentences can affect speech intelligibility.

Resumé

Unge personer med rask hørelse er i stand til at forstå tale trods indblanding af en eller flere samtidige talere og selv i meget vanskelige lyttesituationer. Denne evne, som er essentiel for hverdagens sociale liv, er forringet hos ældre personer med sensorineuralt høretab, og de møder ofte udfordringer i komplekse lyttesituationer trods brug af høreapparat. Derfor er en undersøgelse af lyttesituationer med flere samtidige talere essentiel for at afdække de underliggende mekanismer og dermed inspirere til udvikling af bedre signalbehandling i høreapparater.

Grundtonen i tale (F_0) og forskellen fra taler til taler er vigtige informationer for god taleforståelighed, til hjælp for personer med normal hørelse (NH) personer, hvorimod personer med høretab (HT) har mindre gavn af dem. Nytteværdien af F_0 er tidligere blevet undersøgt med forsimplede simuleringer i laboratoriet af lyttesituationer med samtidige talere, for udvalgte talesignaler. Disse afspejler ikke nødvendigvis hverdagstale og dækker derfor ikke nødvendigvis variationen i hverdags lyttesituationer.

Målet med denne afhandling er at tilføje viden om betydningen af F_0 informationen i situationer med samtidige talere ved brug af talesignaler med stor variation i F_0 forskelle, svarende til typiske talere. Effekten af såvel stationære og dynamiske F_0 forskelle mellem samtidige talere blev undersøgt hos både NH og HT-personer. For NH-gruppen var effekten af disse manipulationer forsvindende. En stationær F_0 forskel mellem samtidige talere var kun gavnligt for taleforståeligheden, når der var små eller ingen energimæssige forskelle til stede, og i den urealistiske situation, hvor de samtidige talere følger sammenlignelig syntaks og F_0 forløb over tid. Dynamiske F_0 forskelle var kun betydende ved specifikke kombinationer af talere som repræsenterer et mindre udsnit af realistiske scenarier med samtidige talere. Derimod blev det vist at et stort F_0 dynamikområde hos mindst en af de to samtidige talere forbedrer taleforståeligheden, uanset forskelle i F_0 dynamikområde mellem de to sætninger. For HT-personer blev den manglende evne til at udnytte F_0 forskelle bekræftet: sammenlignet med NH personer, var gevinsten ved statiske F_0 forskelle mindre og ikke til stede for ændret F_0 dynamikområde eller dynamiske forskelle i F_0 mellem de to sætninger. Endeligt blev F_0 egenskaberne for forskellige talematerialer analyseret og sammenlignet med realistiske taleoptagelser. hvilket giver et godt overblik til fremtidigt valg af talemateriale til undersøgelser af taleforståelighed, med fokus på grundtone. Sammenlagt antyder resultaterne fra denne afhandling at i almindelige lyttesituationer med samtidige talere, bidrager F_0 til et holistisk

billede af lyttesituationen som også involverer mange andre taleegenskaber. Især viste F_0 dynamikområdet for den individuelle samtidige sætning en påvirkning på taleforståelighed, og der foreslås derfor yderligere undersøgelser af det individuelle F_0 forløb i sætninger frem for statiske forskelle.

Acknowledgments

This PhD project was carried out at the Hearing Systems Section at DTU, with the financial support of the Centre for Applied Hearing Research (CAHR) and in collaboration with Eriksholm Research Centre (Oticon A/S). It has been an extremely stimulating and challenging experience, thanks to which I now carry with me valuable lessons that enriched me professionally and personally.

I conducted this project under the main supervision of Prof. Torsten Dau. To him, I would like to express my gratitude for offering me the opportunity of conducting this research, for allowing the freedom in managing and shaping my project and for providing the highest flexibility I could desire in the work environment. I would also like to thank my co-supervisors Johannes Zaar and Helia Relaño-Iborra for their infinite patience, for constantly encouraging me and for providing an impressive motivational propeller. My gratitude goes also to Lars Bramsløw, from Eriksholm Research Centre, for his technical contribution to the project and for being always available to clear up my doubts. I had an excellent team of supervisors.

I would also like to thank my colleagues at the Hearing Systems Section and in particular Caroline van Oosterhout for her efforts in keeping the Section always up and running. A special thanks goes to Borys Kowalewski, a former colleague and now a good friend, with whom I shared the office and a very pleasant part of my time at DTU.

I finished this thesis while staying in the beautiful city of Pisa, whose tower constantly reminded me that a mistake by an engineer not always leads to a failure. I will keep that in mind.

This journey lasted four years, during which I met a lot of challenges (interesting fact: the scientific ones were not the most difficult). Luckily, I always had someone on my side, showing me how to adapt and respond to difficulties. With very few exceptions, it has always been a woman, probably because resilience is just another ability where they outperform me. So, here is my gratitude to them: Mamma, Roberta, Moneypenny, Ninni, Francine, Emma, Angelo. Dad.

Related publications

Journal papers

- Mesiano, P.A., Zaar, J., Relaño-Iborra, H., Bramsløw, L., and Dau, T. (**2022**). "The role of average fundamental frequency difference on the intelligibility of real-life competing sentences," Journal of Speech, Language, and Hearing Research, *under revision*.
- Mesiano, P.A., Zaar, J., Relaño-Iborra, H., Bramsløw, L., and Dau, T. (**2022**). "Effects of the fundamental-frequency dynamics on sentence intelligibility in competing-talker scenarios," *in preparation*.
- Mesiano, P.A., Zaar, J., Relaño-Iborra, H., Bramsløw, L., and Dau, T. (**2022**). "Effects of fundamental-frequency differences between competing sentences on speech intelligibility for listeners with hearing impairment," *in preparation*.
- Mesiano, P.A., Zaar, J., Relaño-Iborra, H., Bramsløw, L., and Dau, T. (**2022**). "On the influence of the auditory environment on fundamental-frequency production," *in preparation*.

Conference papers

- Mesiano, P.A., Zaar, J., Bramsløw, L., Pontoppidan, N.H., and Dau, T. (**2019**). "Assessing the impact of fundamental frequency on speech intelligibility in competing-talker scenarios," Proceedings of the International Symposium on Auditory and Audiological Research **7**, 77-84.

Published abstracts

- Mesiano, P.A., Zaar, J., Kowalewski, B., Fan, L., Carney, L.H., and Dau, T. (**2019**). "Characterizing the Role of Hearing Loss in Comodulation

Masking Release: Behavioral Measurements and Computational Model Predictions," Association for Research in Otolaryngology (ARO), 42nd MidWinter Meeting, Baltimore, MA, February 2019.

- Mesiano, P.A., Zaar, J., Bramsløw, L., Pontoppidan, N.H., and Dau, T. (**2020**). "The Role of Fundamental Frequency in Competing-Talker Scenarios," Association for Research in Otolaryngology (ARO), 43rd MidWinter Meeting, San Jose, CA, January 2020.

Contents

1

General introduction

The ability to hear and understand speech in the presence of one or many interfering voices (i.e., competing-talker scenarios; Bronkhorst, 2000; Cherry, 1953) is essential for daily social life and is a major topic in auditory research. Young normal-hearing listeners typically perform this task successfully, even in adverse listening conditions, thanks to a variety of auditory cues that are available to them. Listeners with sensorineural hearing loss, instead, appear to have limited access to the relevant auditory cues and their ability to understand speech in these challenging scenarios is impaired, even when hearing-loss compensation solutions are provided, compromising the participation in conversation (Bramsløw et al., 2018; Kochkin, 2002; Neher et al., 2007). It is therefore important to understand which acoustic features of the competing speech signals are relevant for speech perception and how they are processed by the healthy auditory system to provide insights for the development of hearing-aid signal-processing strategies that can support the speech understanding for people affected by hearing loss.

The auditory cues that help segregate the target speech from the interfering voices in competing-talker environments can depend on the properties of the auditory scene, such as the spatial configuration of the talkers, or on the vocal attributes of the talkers and the prosodic features of the speech signals, such as fundamental frequency (F_0), formant structure and amplitude modulations. The spatial configuration of the talkers with respect to the listener is encoded in the interaural level differences (ILDs) and interaural time differences (ITDs) between the speech signals arriving at the listener's ears and both ILDs and ITDs have been shown to provide powerful cues for multi-talker speech segregation (e.g., Culling et al., 1994; Freyman et al., 1999; Hawley et al., 2004; Lőcsei et al., 2016; Plomp, 1976). The F_0 and the formant structure, as well as their differences across competing talkers, can also provide prominent cues aiding the perception of the target speech (Assmann, 1999; Binns and Culling, 2007; Calandruccio et al., 2019; Darwin and Hukin, 2000a,b; Darwin et al., 2003; Flaherty et al., 2021).

In particular, the time-average F_0 of the individual talker can help distinguish the talker's sex (Honorof and Whalen, 2010), while the F_0 trajectory (i.e., the time-evolution of the F_0 along the speech signal) carries prosodic information that reflects the syntactical and grammatical content of speech (Ladd, 2008). Spatial cues (i.e., ILDs and ITDs) can be limited in reverberant environments (Culling et al., 1994; Darwin and Hukin, 2000b; Viveros Muñoz et al., 2019) or when the competing talkers are close in space, especially when the listener is affected by hearing loss (Best et al., 2011; Marrone et al., 2008; Viveros Muñoz et al., 2019). In these situations, vocal and prosodic cues (such as the F_0), which are more resilient to reverberation than spatial cues (Culling et al., 1994; Darwin and Hukin, 2000b), may become crucial for speech intelligibility.

The investigation of the effects of these auditory cues on speech intelligibility is commonly carried out in controlled laboratory settings that represent merely a simulation of real life. Such settings typically replicate only certain aspects of realistic auditory environments and may be highly selective with respect to the availability and variability of the acoustic features described above. The choice of the experimental speech material is largely influential in this regard, as the use of speech materials that are constrained in their linguistic or acoustic properties (often aimed at isolating specific attributes of speech) might not faithfully represent real-life speech. It is therefore essential to verify how scientific evidence obtained in constrained laboratory settings transfers to less constrained experimental conditions that better represent the realism of the acoustic features under study.

The work described in this thesis focused on the effects of F_0 and its dynamics on the intelligibility of competing voices. In previous studies that dedicated attention to this topic (Binns and Culling, 2007; Calandruccio et al., 2019; Darwin et al., 2003; Flaherty et al., 2021; Summers and Leek, 1998), the employed experimental design and speech material might have largely differed from real-life speech in its linguistic and acoustic properties and may have enhanced the effects of F_0-related cues under study beyond their real-life importance. In fact, some of the used experimental speech materials lacked linguistic cues, such as syntax, grammar and meaning, or were limited in the variability of F_0 and its dynamics to specific values that are not thoroughly representative of real-life speech. It was the aim of this thesis to extend the knowledge of how average F_0 differences and differences in the dynamic properties of the F_0 (as reflected by a contrast in F_0 dynamic range) between competing voices affect

the perception of target speech in competing-talker scenarios, using a speech material that well replicates the F_0 properties and variety of real-life speech. This was done by conducting three experiments that explored the effects on speech intelligibility induced by differences in F_0 between two competing sentences, for young normal-hearing (NH) listeners and older hearing-impaired (HI) listeners. The three experiments shared the same experimental paradigm, where two sentences (one cued as the target), manipulated in their F_0 properties to obtain a desired average F_0 difference and/or F_0-dynamic-range contrast, were presented simultaneously to the listener who was asked to repeat the words of the target sentence. All three experiments employed the Danish Hearing in Noise Test (HINT; Nielsen and Dau, 2011) in an extended version with several male and female talkers.

The thesis is structured as follows. *Chapter 2* describes the speech materials used in the work presented in the thesis, the methods used for the analysis and manipulation of their F_0 information, as well as the metrics used for quantifying the statistical properties of the F_0. The same chapter provides a comparative analysis of the F_0 statistics measured in HINT and in other experimental speech materials used in previous investigations. The F_0 statistics of the different speech materials were also compared to those measured on speech from laboratory-recordings of naturalistic dialogues, providing an argument as to why HINT can offer a more faithful representation of the F_0 information of real-life speech. *Chapter 3* and *Chapter 4* describe the first two experiments, where NH listeners were tested to measure the effects of an average F_0 separation and of an F_0-dynamic-range contrast between two competing voices on speech intelligibility. These measures were extended to a group of older HI listeners in the third experiment, described in *Chapter 5*. *Chapter 6* presents an analysis of the F_0 information measured in speech produced by pairs of NH or HI talkers who were conducting naturalistic dialogues in laboratory settings under conditions with various degrees of acoustic or communicative barriers, such as the presence of background noise or of hearing-impaired interlocutors. This analysis offers an assessment of how the F_0 of the voice is adapted by the talker to overcome potential auditory barriers in complex acoustic scenarios and to improve speech intelligibility for their interlocutors. Finally, in *Chapter 7*, the results of the experiments are discussed in relation to current literature, dedicating particular attention to the differences in the employed speech materials and experimental paradigms. This last chapter provides an outlook of potential directions for

future research on the effects of F_0-related cues in competing-talker scenarios.

2

Speech materials: Description, analysis and processing

This chapter describes the speech materials employed in the work presented in this thesis, the methods used for their analysis and the methods used for manipulating them to generate the acoustic stimuli used in the experiments described in the next chapters. First, the speech materials are described. Then, an overview is provided of the signal analysis and processing methods used for extracting and modifying the F_0 information of the speech recordings as well as the statistical metrics used for describing it. Attention is dedicated to assessing the precision of the F_0 extraction method and to the possible steps for limiting the impact of potential F_0 extraction errors. An analysis of the F_0 information of the speech material utilized in the experiments is provided, together with a comparison with the F_0 information measured in other speech materials used in auditory research. The differences between speech materials and their accuracy in representing the F_0 information of real-life speech are discussed by comparing them with recordings of naturalistic dialogues.

2.1 Speech materials

2.1.1 Experimental speech materials

The three experimental studies presented in Chapters 3, 4 and 5 employed the Danish Hearing In Noise Test (HINT; Nielsen and Dau, 2011) as speech material. The Danish HINT consists of 200 open-set, five-word, daily-life sentences, divided into ten phonetically balanced lists. Recordings of the HINT speech material were available from twelve different talkers (six males and six females, all native Danish speakers): the original recordings of a male talker (later labelled as 'M1') from Nielsen and Dau (2011) and eleven additional recordings provided by Eriksholm Research Centre (Bramsløw et al., 2019).

Other experimental speech materials, which were used in previous investi-

gations on the role of F_0-related cues on speech intelligibility, were analyzed. These additional speech materials are the Coordinate Response Measure (CRM; Bolia et al., 2000) and the Bamford-Kowal-Bench (BKB; Bench et al., 1979). The CRM speech corpus contains 256 closed-set sentences with the fixed structure "ready *call-sign* go to *color number* now", where call-sign, color and number are words selected from a closed set of alternatives. Recordings of the CRM sentences from eight talkers (four males and four females) were analyzed. The BKB speech corpus consists of 336 meaningful English sentences, characterized by simple syntax and grammar and with word number varying between three and seven (for example, "The clown had a funny face"). The recordings of the BKB corpus analyzed here were the speech materials used by Calandruccio et al. (2019) and Wasiuk et al. (2020) and consisted of all 336 sentences spoken by a female talker (later labelled as 'Talker A') and mixtures of two streams of 50 concatenated BKB sentences spoken by two other female talkers (later labelled as 'Talker B' and 'Talker C'). In these two-talker streams (later labelled as spoken by 'Talker BC'), the same 50 BKB sentences were spoken by talkers B and C but they were concatenated in different order to ensure that the same sentence was never spoken by both talkers simultaneously. All three talkers were trained actors and all sentences were recorded with three different speaking styles, obtained by instructing the talkers to speak with a flat, normal, and exaggerated intonation, to produce three different degrees of F_0 dynamic range. To produce the flat and the exaggerated speaking styles, the actors were instructed to speak "as if they were sad" and "as if they were happy and excited", respectively.

2.1.2 Recordings of naturalistic speech

For comparison, recordings of spontaneous, naturalistic speech were also analyzed. These recordings consisted of lab-recorded dialogues conducted in the Danish language between young normal-hearing (NH) talkers or between a young NH talker and an older hearing-impaired (HI) talker (all native Danish speakers), who were performing the Diapix task (Baker and Hazan, 2011), speaking either in quiet or in presence of background noise. The dialogues between NH talkers were recorded by Sørensen et al. (2021) from 19 pairs of young NH talkers (indicated as 'NH1'), who were speaking in quiet and in presence of a 6-talker speech-shaped noise (ICRA 7; Dreschler et al., 2001) at a 70-dBA sound pressure level. Speech recordings from dialogues between NH1 talkers (38 in total) are labelled here as 'NH1-NH1'. The dialogues between the NH (a group

different from NH1, therefore indicated as 'NH2') and HI listeners were recorded by Sørensen et al. (2019) for 12 pairs of talkers who were talking in quiet and in three different noise conditions (20-talker babble noise at 60-, 65- and 70-dBA sound pressure level). Speech recordings from the 12 NH2 talkers participating in dialogues with the 12 HI talkers are labelled here as 'NH2-HI', while speech recordings from HI talkers conversating with NH2 talkers are labelled as 'HI-NH2'. For each pair and each condition, three recording sessions were available for both the NH-versus-NH and the NH-versus-HI dialogues.

For each speech material, the number of talkers for which the recordings were available and the time durations of the recordings are reported in Table A.1 in Appendix A.

2.2 F_0 extraction and manipulation

For the purposes of the studies presented in the following chapters, all speech materials described in the previous section were analyzed in terms of their F_0 information. For each audio file, an estimate of the F_0 trajectory (i.e., the time evolution of the F_0) was obtained using the autocorrelation method implemented in the software PRAAT (Boersma et al., 1993) and stored as a Matlab array. PRAAT is widely used for analyzing and manipulating different properties of speech and its use is largely documented. The obtained F_0 trajectories were used for estimating long-term F_0 statistics (time average and dynamic range) of the different speech materials. For the experimental speech materials (HINT, CRM and BKB), talker-specific F_0 statistics were also computed by concatenating the F_0 trajectories of all sentences spoken by each talker. Similarly, the talker-specific F_0 statistics of the real-life recordings of young NH and older HI talkers were computed across the concatenated F_0 trajectories of the recording sessions of each talker.

For the experiments presented in Chapters 3, 4 and 5, the F_0 trajectories of the HINT speech material were modified in Matlab (e.g., shifted in frequency, compressed or expanded in their frequency dynamic range) in different ways to create F_0 trajectories with desired statistical properties that were used as controlled variables in the different experimental conditions. After the required modifications in Matlab, the F_0 trajectories were applied to the experimental speech stimuli that were resynthesized using PRAAT, with the Pitch-Synchronous Overlap-Add (PSOLA) algorithm (Moulines and Charpentier, 1990).

The specific settings utilized in the F_0-estimation algorithm and the F_0 manipulations applied to the F_0 trajectories differed between the experiments and are reported in dedicated methods sections in the following chapters.

2.2.1 F_0 extraction artifacts in PRAAT

One limitation of the autocorrelation algorithm implemented in PRAAT is the estimation of potentially erroneous F_0 values and F_0 variations, such as (sub)-multiple of the actual F_0 produced by the voice (i.e., 'octave jumps'). Figure 2.1 illustrates two examples of F_0 trajectories extracted with PRAAT from HINT sentences (labelled as sentence 'A' and sentence 'B') showing potentially erroneous F_0 values that can be categorized as octave jumps. In the figure, the F_0 trajectories originally extracted with PRAAT are shown as black squares, together with their medians (indicated by solid black lines) and dynamic ranges (indicated by grey areas). In these examples, the F_0 values that are unrealistically-far from the median (occurring about one octave below the F_0 values in the neighboring time frames) were identified, shifted to what was estimated as their proper octave (multiplying their values by a factor of two) and re-plotted as red circles.

The presence of these potentially erroneous F_0 estimates (including, but not limited to octave jumps) in the F_0 trajectories of the employed speech material was assessed by means of a relatively simple method based on the detection of F_0 variations that were considered unrealistically fast. The reason why this method was based on the speed of F_0 variation rather than simply the F_0 variation itself is that identifying large F_0 variations as erroneous without considering the time duration within which they occur might lead to inaccurate error detections. It was assumed that the human voice can produce large changes in F_0, such as octave jumps, within relatively long time durations, for example in correspondence of silences or unvoiced portions of the trajectory that encompass several F_0 sampling periods (i.e., between voiced segments of the F_0 trajectory). However, large F_0 variations of one octave or more produced within short time durations of few milliseconds, such as the 10-ms sampling period used for the extraction of the F_0 trajectories, were assumed to be unnatural. Therefore, it was possible to identify potentially erroneous F_0 estimates by considering the speed of change in F_0: large F_0 variations that occur between consecutive time frames of the F_0 trajectories were more likely to be erroneous estimates than large F_0 variations occurring over longer time durations.

Figure 2.2 illustrates the probability density histogram of the speed of F_0

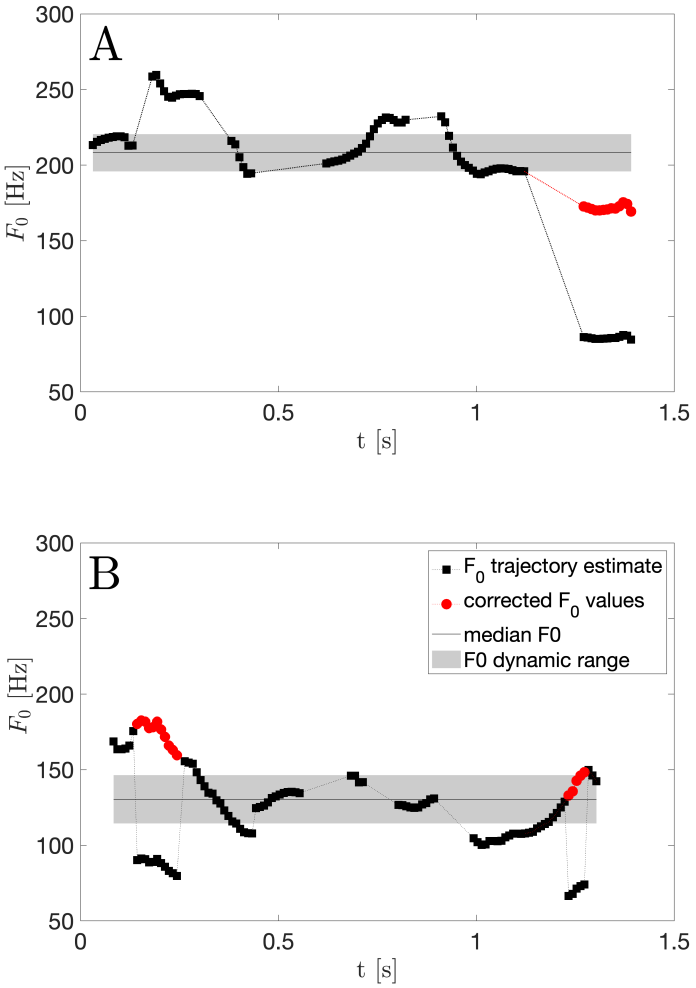


Figure 2.1: Examples of potentially erroneous F_0 estimates from two HINT sentences. Each panel shows: the extracted F_0 trajectory from PRAAT (black squares connected by a dotted line), its median (solid black line) and dynamic range (measured as plus/minus one MAD from the median value and indicated with a grey area), a possible correction of the potentially erroneous F_0 estimate (red circles). Top panel (A) shows a HINT sentence from talker F2 ("Bussen kan ikke komme frem", "The bus cannot come forward"). Bottom panel (B) shows a HINT sentence from talker M3 ("Filmen er rigtig godt lavet", "The film is really well done").

variation between consecutive F_0 values (indicate as 'F_0 speed', measured in semitones per centisecond, ST/cs), computed over the entire HINT corpus of 12 talkers. In the histogram, the F_0-speed values are binned into 0.1-ST/cs wide intervals and the y-axis is shown with a log scale. The highest probability is found at the lowest values of F_0 speeds and decays for higher values. However,

in proximity of F_0 speeds of about 12 ST/cs, a higher probability was measured, indicating unrealistically-fast F_0 variations of about one octave occurring in time frames of the order of the 10-ms sampling period used for extracting the F_0 trajectory from the speech signal. To illustrate this, the histogram bars corresponding to F_0 speeds larger than 11 ST/cs are shown in red in the figure.

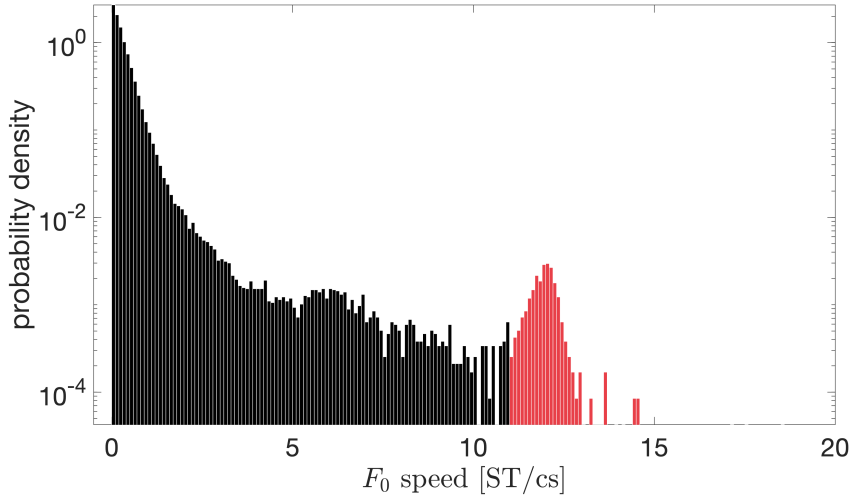


Figure 2.2: Probability density histogram of the F_0 speed of variation between consecutive F_0 values in the F_0 trajectories of HINT, measured in semitones per centisecond (ST/cs). The histogram includes the F_0 trajectories of all sentences from all the 12 talkers in the HINT speech material. The histogram bins are 0.1-ST/cs wide. The histogram bars corresponding to speed values larger than 11 ST/cs (corresponding to F_0 variations that occur unrealistically fast) are shown in red.

Overall, in the entire HINT speech material, the occurrence of F_0 variations faster than 11 ST/cs was found to be negligible (0.23% of all estimated F_0 values). In general, regardless of their speed, F_0 variations larger than 11 semitones occurred only in 0.54% of the estimated F_0 values.

Despite the negligible probability of occurrence, an octave jump can affect an entire portion of the F_0 trajectory (as shown by the examples in Figure 2.1), which can be assigned to the wrong octave, producing in some cases highly inaccurate F_0 trajectories and corresponding statistics. The only way to obtain a 'clean' F_0 trajectory would consist in identifying these F_0 artifacts and correcting them by assigning the F_0 values to the proper octave (as done with the trajectories in Figure 2.1). However, without a reliable identification of such F_0 artifacts (i.e., in absence of ground-truth F_0 information), this method would be rather error prone and could introduce an additional source of error in the

F_0 estimates. Therefore, since the experiments presented in this thesis focused on the long-term statistics of the F_0, it was decided to limit the impact of octave jumps by using median moments, i.e., median for the time average (later indicated as $(\overline{F_0})$ and median absolute deviation (MAD) for the dynamic range (later indicated as $\sigma(F_0)$). Table 2.1 shows how the mean, median, standard deviation and median absolute deviation change between the original and the modified F_0 trajectories (i.e., before and after the 'octave-jump correction' shown in Figure 2.1). These changes are measured in Hz and in percentage with respect to the values computed on the modified trajectory (considered as the one free from erroneous F_0 values). This analysis provides an estimate of how the potentially erroneous F_0 estimates can affect the different statistical measures. As shown by these examples, the median moments are less affected by the presence of octave jumps in the F_0 trajectories.

Table 2.1: Statistical measures of the original F_0 trajectory extracted with PRAAT from sentence A and B (shown in Figure 2.1, top and bottom panels, respectively), of the F_0 trajectory after the correction of potentially erroneous F_0 values and the difference between the two.

	mean [Hz]	median [Hz]	STD [Hz]	MAD [Hz]
Sentence A				
Extracted from PRAAT	197	208	50.0	12.3
Corrected	210	208	23.6	12.3
Difference	-13 (-6%)	0 (0%)	26.4 (112%)	0.0 (0%)
Sentence B				
Extracted from PRAAT	120	125	24.3	16.8
Corrected	134	130	22.8	15.9
Difference	-14 (-10%)	-5 (-4%)	1.5 (7%)	0.9 (6%)

A method for reducing the probability of unrealistic F_0 estimates produced by PRAAT has been proposed by De Looze and Hirst (2008) and consists in adapting the frequency range for searching F_0 values (defined by the 'pitch floor' and 'pitch ceiling' parameters of the autocorrelation algorithm in PRAAT) to the specific audio signal. De Looze and Hirst (2008) suggested to extract the F_0 trajectory from the signal using a default frequency range for the search of F_0 estimates (i.e., from 50 to 750 Hz), and then to re-extract the trajectory using a frequency range whose boundaries are estimated from the first and third quartiles of the distribution of F_0 values from the first estimate of the F_0 trajectory. However, the author of this thesis became aware of this method only recently and its use and efficacy could not be explored in the work presented here. This method is not mentioned in any of the studies referenced in this

thesis that employed PRAAT. Nevertheless, future works involving the use of PRAAT for F_0 estimate may consider assessing the quality of this method in reducing the occurrence of F_0 estimate errors, for example by measuring how the probability distribution of F_0 speed changes when this method is employed.

2.3 Comparison of the F_0 statistics of different speech materials

The F_0 information of the different experimental speech materials (HINT, CRM, BKB) were analyzed and compared to the F_0 information of naturalistic speech from the laboratory-recordings of dialogues between NH talkers (NH1-NH1) conducted in quiet. The purpose of this comparative analysis was to assess how faithfully the experimental speech materials represent the F_0 information found in real-life speech. The specific settings of the autocorrelation method used in PRAAT for extracting the F_0 information are shown in Table 2.2.

Table 2.2: Parameters of the autocorrelation algorithm implemented in PRAAT, used for extracting the F_0 trajectories from the speech signals.

Setting	**Value**
Time step	0.01 s
Pitch floor	50 Hz
Pitch ceiling	500 Hz
Max. number of candidates	5
veryAccurate	False
Silence threshold	0.03
Voicing threshold	0.45
Octave cost	0.01
Octave-jump cost	0.5
Voiced/unvoiced cost	0.14

The statistical properties of the F_0 ($\overline{F_0}$ and $\sigma(F_0)$) of the different speech materials (HINT, CRM, BKB and NH1-NH1 in quiet) were analyzed and compared. For the recording of NH1-NH1 naturalistic dialogues, one talker had a high $\sigma(F_0)$ value (48 Hz) that was identified as an outlier (more than three standard deviations away from the mean computed over all talkers in this speech material) and was therefore excluded from the analysis. Figure 2.3 shows the $\sigma(F_0)$ as a function of $\overline{F_0}$ for each talker in the HINT, CRM and BKB speech corpora, as well as for the NH1-NH1 talkers in the recordings of naturalistic dialogues. The speech from naturalistic dialogues (indicated by open red diamonds) revealed a trend of increasing $\sigma(F_0)$ with increasing $\overline{F_0}$, indicating that voices with

higher registers also have stronger fluctuations in intonation. This trend was well replicated in the HINT and CRM speech corpora (green circles and blue squares, respectively).

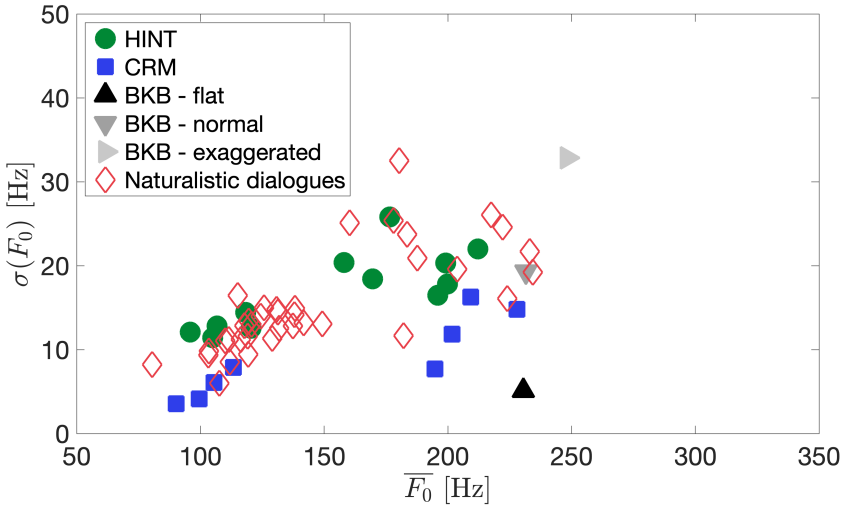


Figure 2.3: Overview of the F_0 statistics (F_0 dynamic range as a function of F_0 median) of the talkers in the different speech materials analyzed: the 12 talkers from the HINT corpus (green circles), the eight talkers from the CRM corpus (blue squares), the single-talker sentences (talker A) from the BKB corpus spoken with flat, normal and exaggerated speaking styles (black upward triangle, dark-grey downward triangle, light-grey rightward triangle, respectively) and the recordings from naturalistic dialogues between NH interlocutors (open red diamonds).

Figure 2.4 offers a separate analysis along the $\overline{F_0}$ and $\sigma(F_0)$ dimensions (shown in top and bottom panels, respectively), for each talker and speech material. In the figure, the data of each talker is shown with black circles, and the average values across talkers are shown in read. For the BKB speech material, the red circles indicate the data for the single-talker recordings (available only for Talker A) while the blue circles indicate the data for the two-talker recordings (Talker BC). The $\overline{F_0}$ of each talker in the HINT and CRM materials was found within the variability of real-life speech. However, the CRM talkers appeared separated in two compact groups along the $\overline{F_0}$ dimension, corresponding to the male (with overall lower $\overline{F_0}$s) and female talkers (with overall higher $\overline{F_0}$), with the voices in each group having very similar $\overline{F_0}$ values. The 12 talkers from the HINT corpus also showed lower $\overline{F_0}$s for the males and higher $\overline{F_0}$s for the females, but their values were more continuously distributed along the $\overline{F_0}$ axis and better replicated the distribution of the talker's $\overline{F_0}$ from naturalistic voices. For the single-talker BKB speech material, similar $\overline{F_0}$ values were measured for

the different speaking styles, indicating that the speaking style did not affect the overall height of the F_0. These values were high (as often observed for female talkers) and at the upper limit of the $\overline{F_0}$ measured in the naturalistic speech available here. For the two-talker mixture of BKB speech material, the $\overline{F_0}$ of flat and normal speaking styles were comparable to those of the single talker in the same style, but the two-talker $\overline{F_0}$s in the exaggerated speaking style were found higher than those for the single talker. Regarding the F_0 dynamic range, the $\sigma(F_0)$ of the 12 talkers in the HINT corpus well replicated the values and variability of naturalistic speech. The CRM corpus showed overall lower $\sigma(F_0)$ values, often below the values found in naturalistic voices with the same $\overline{F_0}$ (see Figure 2.3), indicating a rather monotonous speech intonation for all talkers. The $\sigma(F_0)$ of BKB speech material recorded with normal speaking style, for both the single-talker and the two-talker speech, was found in alignment with the values of naturalistic speech with similar $\overline{F_0}$. However, for both the single-talker and the two-talker speech, the $\sigma(F_0)$ of the flat speaking style was found at the lower bound of naturalistic values, at a value that is not typical of voices with that $\overline{F_0}$, while the $\sigma(F_0)$ of the exaggerated style was unrealistically high.

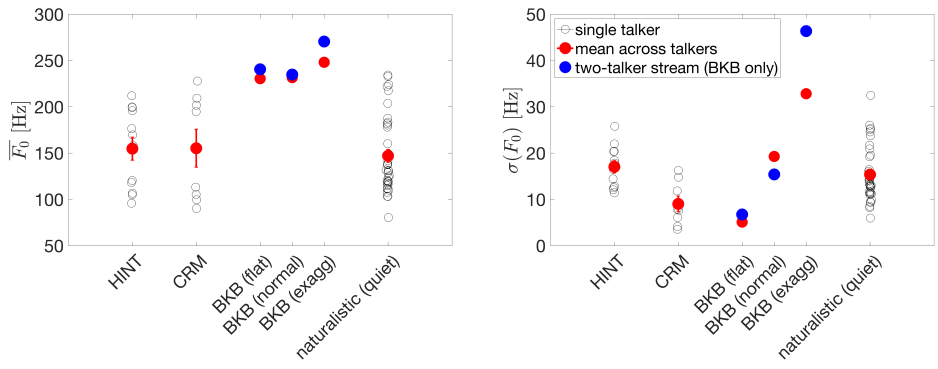


Figure 2.4: Overview of F_0 median values (left panel) and F_0 dynamic range values (right panel) in each speech material analyzed. The empty black circles represent the data for individual talkers. The red full circles represent the mean of the individual-talker F_0 medians, with the error bars representing standard errors. For the BKB material, the data from the two-talker streams are shown with blue circles.

Figure 2.5 shows the analysis of $\overline{F_0}$ and $\sigma(F_0)$ on a single-sentence level for the HINT, CRM and single-talker BKB speech materials. As in the talker-based analysis, also at the level of the single sentence, F_0 trajectories with higher $\overline{F_0}$ were characterized by wider F_0 variations, i.e., larger $\sigma(F_0)$ values (see HINT

and CRM data shown in Figure 2.5, panel A and B, respectively). For both HINT and CRM, the data of each talker appeared grouped in clusters. However, the sentences from CRM talkers formed smaller clusters than the sentences spoken by HINT talkers, indicating a small variability of F_0 statistics within a specific talker in the CRM speech material. Therefore, the CRM speech material offers a limited variability in F_0 statistics, both across talkers and across the sentences spoken by a given talker.

As for the BKB speech material, single-sentence recordings were available only from talker A. For the normal speaking style, the F_0 variability offered by this corpus is equivalent to that of the HINT talkers with a similar $\overline{F_0}$. However, the flat and the exaggerated speaking style produced F_0 trajectories that are not in line with the statistics of realistic speech. For the flat speaking style, the sentences have an unrealistically low $\overline{F_0}$ and exhibit very little variation in both $\overline{F_0}$ and $\sigma(F_0)$. In the exaggerated speaking style, the single-sentence data are more spread over a wider range of values, with many sentences having unrealistically high $\overline{F_0}$ and $\sigma(F_0)$, probably because, in some sentences, the talker modifies the $\overline{F_0}$ when forcing her speaking style to have wider-than-normal F_0 variations. It is possible that these large $\overline{F_0}$ and $\sigma(F_0)$ values might be a consequence of erroneous F_0 estimates (i.e., octave jumps), but no explanation could be found as to why this would happen more often with the exaggerated speaking style than with the flat or normal ones.

It can be concluded that, from the point of view of the F_0 statistics considered here, the HINT speech material offers a wider variability and a better representation of naturalistic speech compared to the CRM and BKB speech recordings analyzed here. Both the single-sentence and the talker-specific data from HINT are more continuously distributed along the $\overline{F_0}$ and $\sigma(F_0)$ dimensions, and better replicate the F_0 of naturalistic speech in both $\overline{F_0}$and $\sigma(F_0)$ values, their variability and their relationship. On the contrary, the CRM and BKB materials have either unnatural F_0 statistics or represent only specific cases that occur at the boundaries of the range of values found in naturalistic speech.

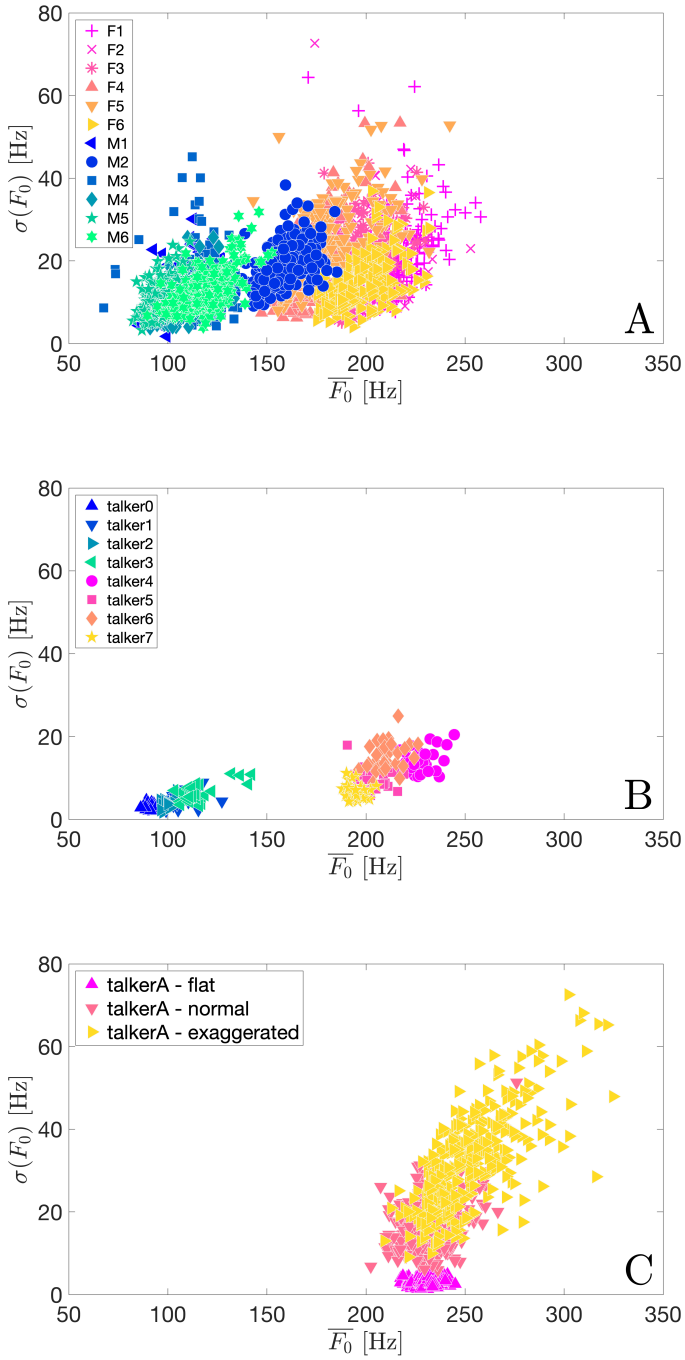


Figure 2.5: Single-sentence F_0 statistics of the HINT (panel A), CRM (panel B) and BKB (panel C) speech materials. For the BKB speech material, single-sentence recordings were available only for Talker A. Each point in the figure represents the F_0 dynamic range of a sentence as a function of its F_0 median. Sentences from different talkers (or with different speaking styles, in the case of BKB speech material) are represented with different colors and symbols.

3

The role of average fundamental frequency difference on the intelligibility of real-life competing sentences[a]

Abstract

The average fundamental frequency separation ($\Delta\overline{F_0}$) between competing voices has been shown to provide an important cue for target-speech intelligibility. However, some of the previous investigations used speech materials with linguistic properties and F_0 characteristics that may not be typical of realistic acoustic scenarios. The present study investigated to what extent the effect of $\Delta\overline{F_0}$ generalizes to more real-life situations. Real-life sentences and a well-controlled method for manipulating the acoustic stimuli were employed. Fifteen young normal-hearing native Danish listeners were tested in a competing-voices sentence recognition task at several target-to-masker ratios (TMRs) and $\Delta\overline{F_0}$s. Compared to previous studies, the present results showed only a moderate effect of $\Delta\overline{F_0}$ at negative TMRs and a negligible effect at positive TMRs. An analysis of the employed stimuli showed that a large $\Delta\overline{F_0}$ effect on the target speech intelligibility is only observed when the competing sentences have highly synchronous F_0 trajectories, which is typical of the artificial speech materials employed in some previous studies. Overall, the present results suggests a relatively small effect of $\Delta\overline{F_0}$ on the intelligibility of real-life speech, as compared to artificial speech, in competing-speech conditions.

[a] This chapter is based on Mesiano et al. (2022c), Journal of Speech, Language, and Hearing Research (under revision).

3.1 Introduction

Complex acoustic scenarios with several sound sources are ubiquitous in daily life and can be challenging for successful communication, particularly for people that are older and/or affected by a hearing impairment. The perceptual task of attending to a speech signal in the presence of competing sound sources, often referred to as the 'cocktail-party' problem (Cherry, 1953), represents a main focus in auditory research. A particular case of interest are auditory scenes with several talkers speaking at the same time. In such competing-talker situations, the perception of the target speech signal that the listener wants to attend to is hindered by the presence of one or several interfering speech signals. Normal-hearing (NH) listeners show a remarkable ability to separate the target speech from the interfering sources (Bramsløw et al., 2015; Humes et al., 2006). However, the auditory sensory and cognitive processes underlying the robust representation of the attended speech in the healthy auditory system are yet to be fully understood.

When facing competing-talker scenarios, the auditory system performs an analysis of the auditory scene, utilizing auditory cues that facilitate the identification of the target-speech signal and its segregation from the speech mixture. For example, spatial cues, such as interaural level differences (ILDs) and interaural time differences (ITDs), have been shown to facilitate speech intelligibility (e.g., Culling et al., 1994; Freyman et al., 1999; Hawley et al., 2004; Lőcsei et al., 2016; Plomp, 1976). ITDs and ILDs can be exploited to segregate the competing signals into separate streams. However, in complex acoustic scenarios, spatial cues can be disrupted due to reverberation (e.g., Culling et al., 1994; Culling et al., 2003; Darwin and Hukin, 2000b; Freyman et al., 1999; Plomp, 1976). Nonetheless, even when spatial information is disrupted or absent, the listener can utilize non-spatial cues to disentangle the target signal from the speech mixture, enabling robust speech perception. In particular, acoustic and prosodic features, such as fundamental frequency (F_0), formant structure and intensity, with their perceptual counterparts pitch, timbre and loudness, have been shown to provide relevant information that helps the listener attend to the target speaker (Brungart, 2001; Brungart et al., 2001; Darwin and Hukin, 2000a,b; Darwin et al., 2003; Festen and Plomp, 1990). Similarly to spatial cues, the representation of F_0 and the formant structure can be degraded due to reverberation (Darwin and Hukin, 2000b). However, F_0 and formant structure

have been shown to provide more salient cues for selective attention (Darwin and Hukin, 2000a) and to be more robust to reverberation than spatial cues (Culling et al., 1994; Darwin and Hukin, 2000b).

Several studies with NH listeners have demonstrated that the difference between the mean F_0 of two competing voices provides an effective cue for their perceptual segregation (Başkent and Gaudrain, 2016; Brokx and Nooteboom, 1982; Darwin et al., 2003; Summers and Leek, 1998). In these studies, the competing-voice scenarios were created by pairing two speech signals that were processed to obtain a certain separation between their average F_0. The studies differed in the details of their experimental design, but their results all yielded consistent evidence that a separation in F_0 between the competing voices, measured as an average over the entire stimulus duration, was beneficial for the intelligibility of the target speech. For example, Summers and Leek (1998) investigated the effect of F_0 separation between competing voices by pairing either synthetic vowels or sentences, in both cases with a flat F_0 trajectory (i.e., a constant F_0 along the entire duration of the stimulus). In the case of concurrent vowels, the recognition of both competing speech signals increased by 18 percentage points when an F_0 separation of just one semitone was introduced, compared to a 0-semitones separation, and performance plateaued for larger F_0 separations. Darwin et al. (2003) used sentences from the coordinate response measure (CRM; Bolia et al., 2000) with naturally-varying F_0 trajectories. The CRM speech corpus represents highly time-aligned, closed-set sentences with a fixed structure ("ready *call-sign* go to *color number* now"), where call-sign, color and number are words selected from a closed set of alternatives. Also in their study, a large effect of F_0 separation on speech intelligibility was observed: target-word recognition improved by 12 percentage points for a two-semitone separation between the competing sentences, compared to a zero-semitones separation. This improvement increased to 20 percentage points for a separation of nine semitones.

Some of the speech materials employed in the mentioned studies, like the competing CRM sentences used by Darwin et al. (2003) or the concurrent vowels used by Summers and Leek (1998), were designed with specific constraints and limitations imposed either on their linguistic variability (e.g., syntax and context) or on the variability of their F_0 over time (e.g., monotonized intonation). Such speech materials ensured control of the experimental scenario but differed considerably from real-life speech. In particular, the high levels of speech

synchrony between the competing signals might have enhanced the effect of F_0 separation on speech segregation. In fact, when introducing more 'realistic' syntax variability in the study of Summers and Leek (1998) with the use of Harvard sentences (Rothauser, 1969), the speech intelligibility improvement induced by a F_0 separation between competing voices was considerably reduced, compared to the improvement found with the use of concurrent vowels. Similarly, Assmann (1999) used realistic sentences in a competing speech task and reported a difference of only 13 percentage points for an average F_0 separation of six semitones (as compared to 20 percentage points measured for the same F_0 separation and target-to-masker ratio by Darwin et al., 2003). The fact that the use of meaningful sentences produced smaller F_0-separation effects suggests that the more variable structure of real-life speech facilitates the segregation of the competing signals and limits the actual benefit provided by F_0 separation. However, Başkent and Gaudrain (2016) observed speech-intelligibility improvements of up to 24 percentage points for an F_0 separation of eight semitones applied to realistic sentences. In their experiment, while the target was a meaningful natural sentence, the masker was created by using an excerpt of a sentence or several concatenated excerpts taken from different sentences. Since the excerpts were cut starting from the ending of the original sentences, without retaining first words or specific portions of them, the resulting masker signal might have sounded unnatural and were not, strictly speaking, intelligible speech. In light of the available findings, it is not clear if the large F_0-separation benefit measured in the mentioned studies can be also observed when using real-life speech targets and maskers, or whether the F_0-separation benefit is reduced in this case.

Additionally, some of the previous results might have been influenced by the level of numerical control of the experimental variables. For example, Brokx and Nooteboom (1982), who reported speech intelligibility improvements for F_0-separated voices, used sentences that were naturally spoken and not synthetically manipulated. This approach has the advantage of avoiding potential signal-processing artifacts in the acoustic stimuli but reduces the numerical control of the F_0 values. In fact, they created speech stimuli by instructing the talker to imitate a higher pitched female voice, but the resulting speech signal showed unrealistically large F_0 variations over time, as stated by the authors. Darwin et al. (2003) generated F_0 separations between sentences through a controlled signal processing method, but they assumed that all sentences had

the same mean F_0 corresponding to the talker's overall mean. However, the mean-F_0 values can vary substantially across sentences spoken by the same talker (Darwin et al., 2003), such that the resulting mean-F_0 separation created with this approach might differ from the desired one.

The present study aimed to advance the understanding of how average F_0 differences between competing voices influence the intelligibility of real speech. It was hypothesized that the prominent effect of the mean-F_0 separation observed in earlier studies has been enhanced by the choice of speech material and stimulus processing and might not generalize to real-life speech stimuli. The effect of average F_0 separation was studied using a speech material consisting of open-set sentences closer to real-life speech than the CRM corpus used in Darwin et al. (2003). A rigorous method was used for estimating and generating the F_0 separation between the sentences with accurate numerical control. The employed experimental approach was similar to the one used by Darwin et al. (2003): pairs of sentences spoken by the same talker were generated with different F_0 separations at several target-to-masker ratios (TMRs). The results were compared to those reported in Darwin et al. (2003) and analyzed using a metric reflecting the amount of speech synchrony and its variability in the employed speech material. Furthermore, the differences in experimental method between the present study and previous studies were assessed with the aim to determine what aspects of the speech stimuli contribute to the effect of the F_0-separation cue under study.

3.2 Methods

3.2.1 Participants

Fifteen NH native Danish listeners (8 females), aged between 21 and 32 (mean 25) years participated in the study. All participants provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). The listeners had normal hearing, with pure-tone thresholds below 20 dB Hearing Level between 125 Hz and 8 kHz. The listeners completed the speech intelligibility experiment in a single experimental session that lasted no more than two hours.

3.2.2 Stimuli

The experimental stimuli were generated by pairing two sentences (a target and a masker) from the Danish HINT speech corpus (Nielsen and Dau, 2011). The Danish HINT consists of 200 open-set, five-word natural sentences split into ten phonetically-balanced test lists, intended to resemble simple, every-day speech, spoken by a male talker (later labelled as 'M1'). Additional recordings of the speech material were made by Eriksholm Research Centre (Bramsløw et al., 2019) to create a total of twelve different talkers (six males and six females, all native Danish speakers) which were used in the present study. The stimulus generation required a prior analysis of the F_0 information contained in the speech corpus, conducted as follows. The F_0 trajectory of each of the 200 sentences in the speech corpus for each talker was extracted using the software PRAAT (Boersma et al., 1993) version 6.0.49. The F_0 trajectories were sampled with a time step of 10 ms, using 100-ms time windows (i.e., 90% overlap). In each time window, the F_0 candidate values were searched within the range 30-550 Hz. The time average and dynamic range of the individual F_0 trajectories were computed. It was observed that PRAAT is rather prone to potential F_0 extraction errors, like F_0 variations of the order of an octave and above[b]. In order to reduce the impact of such errors on the measurements, the F_0 statistics for the individual sentences were quantified using median moments: F_0 median (indicated as $\overline{F_0}$) for the time average and F_0 median absolute deviation (MAD) for the dynamic range, defined as $\sigma(F_0) = median(|F_0 - \overline{F_0}|)$, with F_0 indicating the array of fundamental frequency values computed along the stimulus. Additionally, F_0 median and F_0 MAD were computed over the entire speech corpus for each of the 12 talkers, henceforth referred to as talker median F_0 ($\overline{F_0}^{talker}$), and talker F_0 MAD ($\sigma(F_0^{talker})$). The obtained values are indicated in Table 3.1.

These values served as a reference for the F_0 information of each talker in the available speech material. However, the $\overline{F_0}$ of individual sentences spoken by a given talker was observed to vary substantially in the Danish HINT speech

[b] Since no ground-truth data exists for the F_0, it is not possible to assess if the excessively large F_0 variations actually occurred in the speech material or if they were F_0-extraction errors. Several approaches were considered to resolve this issue, including the detection and omission or correction of potential octave jumps. However, given the negligible frequency of their occurrence (less than 1% on the entire HINT recordings employed in the current study), it was decided to keep the extracted trajectories unmodified and consider such F_0 extraction errors as 'noise' in the data.

corpus (by up to 4 semitones). Furthermore, it was observed that the F_0 dynamic range, $\sigma(F_0)$, tends to increase with increasing median F_0, $\overline{F_0}$. To illustrate this, Figure 3.1 shows $\sigma(F_0)$ as a function of the $\overline{F_0}$ for the individual sentences of the female (red squares) and the male (blue triangles) talkers.

Table 3.1: Median F_0 and F_0 dynamic range values (measured as median absolute deviations, MAD) for each talker computed over the entire HINT speech corpus.

Talker ID	$\overline{F_0}^{talker}$	$\sigma(F_0^{talker})$
F1	214	22
F2	201	18
F3	200	20
F4	171	18
F5	178	26
F6	197	16
M1	107	13
M2	159	20
M3	122	12
M4	106	11
M5	97	12
M6	120	14

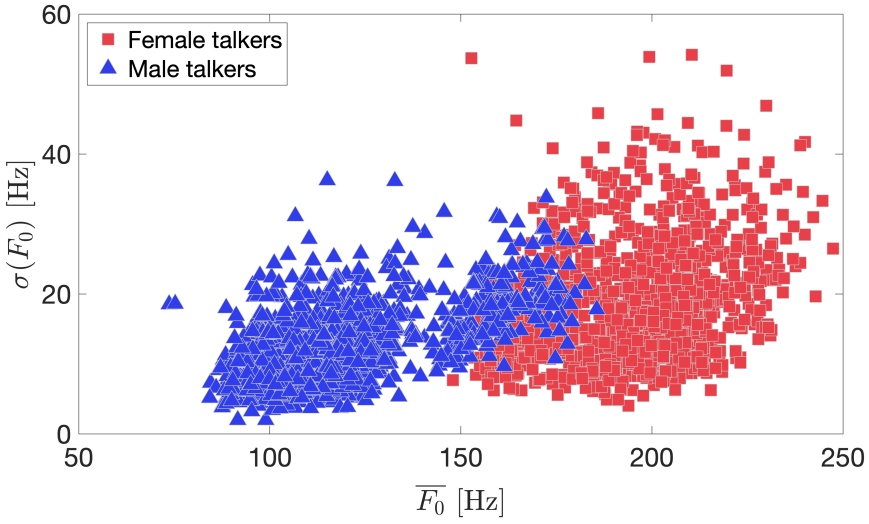


Figure 3.1: F_0 dynamic range (measured as median absolute deviation, MAD) as a function of the median F_0 for each sentence in the Danish HINT speech corpus. Data are shown separately for male talkers (blue triangles) and female talkers (red squares).

The sentence pairs presented during the experiment were created by mixing two sentences spoken by the same talker, randomly taken from different lists and processed with PRAAT to obtain a difference between their $\overline{F_0}$ values ($\Delta\overline{F_0}$) of 0, 3, 6 or 12 semitones. The same talker was used in any given pair to reduce the influence of other cues such as differences in vocal-tract properties and long-term speech spectrum. The four $\Delta\overline{F_0}$ values were applied to sentence pairs mixed at TMRs of -12, -8, -4, 0 and 4 dB, resulting in 20 testing conditions (4 $\Delta\overline{F_0}$s x 5 TMRs).

3.2.3 F_0 processing of the sentences

The processing method applied to the speech material was inspired by the one used by Darwin et al. (2003). For each pair of sentences, the desired $\Delta\overline{F_0}$ was obtained by separating the original F_0 trajectories. The $\Delta\overline{F_0}$ was split across the two sentences: one F_0 trajectory was shifted upward and the other downward, by a number of semitones relative to the $\overline{F_0}^{talker}$ of the talker used in the pair. In order to avoid unnatural F_0 values, a larger F_0 shift was applied towards lower frequencies when $\overline{F_0}^{talker}$ was higher than the average F_0 computed across all talkers (indicated as $\overline{F_0}^{HINT}$), and towards higher frequencies otherwise. The largest (upward or downward) shift that was applied to a sentence was eight semitones. The shifts applied to the sentences in a pair for the different $\Delta\overline{F_0}$ conditions are listed in Table 3.2. In each sentence, the F_0 shift was achieved by multiplying the F_0 trajectory by a positive factor s, defined as $s = \widehat{\overline{F_0}}/\overline{F_0}$, where $\widehat{\overline{F_0}}$ is the desired median F_0 above or below $\overline{F_0}^{talker}$ and $\overline{F_0}$ is the median F_0 of the unprocessed sentence. Instead of assuming that $\overline{F_0} = \overline{F_0}^{talker}$ for all sentences spoken by a given talker (as done in Darwin et al., 2003) and potentially generating $\Delta\overline{F_0}$s that differed from the desired ones, this method allowed to account for the median-F_0 variability observed across all sentences spoken by a given talker and thus obtain a precise measure of $\Delta\overline{F_0}$ for a specific sentence pair. Besides shifting $\overline{F_0}$ to $\widehat{\overline{F_0}}$, the multiplication expanded ($s > 1$) or compressed ($s < 1$) the F_0 trajectory and therefore preserved the natural increase of the F_0 dynamic range with increasing median F_0 observed for the speech corpus (see Figure 3.1). Finally, the modified F_0 trajectory was applied to the sentence by means of the Pitch Synchronous Overlap and Add (PSOLA) resynthesis algorithm (Moulines and Charpentier, 1990) in PRAAT.

Table 3.2: Combination of F_0 shifts applied to the sentences in a pair for the desired $\Delta\overline{F_0}$ conditions. All values are shown in semitones. $\overline{F_0}^{talker}$ refers to the talker median F_0 computed across all sentences; $\overline{F_0}^{HINT}$ refers to the median F_0 computed across all sentences and talkers. s indicates the multiplication factor applied to the F_0 trajectory to shift its median F_0.

$\Delta\overline{F_0}$	$\overline{F_0}^{talker} > \overline{F_0}^{HINT}$		$\overline{F_0}^{talker} < \overline{F_0}^{HINT}$	
	Shift upward ($s > 1$)	Shift downward ($s < 1$)	Shift upward ($s > 1$)	Shift downward ($s < 1$)
0	0	0	0	0
3	1	-2	2	-1
6	2	-4	4	-2
12	4	-8	8	-4

3.2.4 Procedure and apparatus

The experimental design followed the competing-voices test (CVT) framework developed by Bramsløw et al. (2019). The target sentence was visually pre-cued to the listener by providing its first word on a screen prior to the stimulus playback. Target and masker sentences were aligned at the onset, while the offsets of the sentences were not necessarily aligned. The mixture of two sentences was presented to the listeners who were asked to repeat as many words as possible from the target sentence. Each of the 20 testing conditions was tested using 20 sentence pairs. To avoid any effect of presentation order or sentence repetition in the group results, the test conditions ($\Delta\overline{F_0}$ and TMR) were balanced across listeners using a Latin square design, while sentence-list and talker were randomized across conditions. In each pair, the target was randomly assigned to either the sentence with the higher or lower $\overline{F_0}$.

The stimuli were presented diotically over headphones, which were free-field equalized to the entrance of the ear canal. The target sentence was presented at an average sound pressure level of 65 dB SPL, randomly roved over a ±5 dB range. The level of the masker sentence was adjusted according to the desired TMR. Level adjustment, sentence mixing and stimulus playback were performed with MATLAB on an Apple computer. The mixture of two sentences was played back at a sampling rate of 16 kHz through a Fireface UCX soundcard and presented via Sennheiser HDA-200 headphones to the listener seated in a sound-proof booth. All listener's responses were scored by the same native Danish audiologist.

Speech intelligibility was quantified as the percentage of correctly repeated words from the target sentence (excluding the initial cue word), computed over

each sentence pair. Since the difficulty of the task could vary across sentence pairs depending on the specific combination of sentences (for example depending on the presence of linguistic context in the target sentence or syntax similarity between the two competing sentences), the performance was averaged across the 20 sentence pairs presented in each experimental condition.

3.2.5 Measures of speech synchrony

A quantitative assessment of the effects of speech synchrony on speech intelligibility and on the benefit produced by $\Delta\overline{F_0}$ was conducted as follows. Speech synchrony between competing sentences was assumed to be reflected at the level of the F_0 trajectories, as the simultaneous periodic activation of target and masker signals, here referred to as 'periodicity synchrony'. When the periodicity synchrony is low, periodic time frames of the target will often occur simultaneously with non-periodic time frames of the masker, due to the misalignment of the competing F_0 trajectories. Such time frames where the target is the only periodic signal are here referred to as 'periodicity glimpses'[c]. The amount of periodicity glimpses was quantified for each pair of sentences by introducing the 'periodicity-glimpse index' (PGI), a metric defined as the number of time frames where the target is the only periodic signal in the mixture, normalized with respect to the number of time frames in the entire stimulus, N ($\text{PGI} = \frac{|\exists F_0^{target} \wedge \nexists F_0^{masker}|}{N}$. The PGI therefore takes values between zero and one. Values close to zero indicate limited availability of target-periodicity glimpses due to a high degree of synchrony between target and masker F_0 trajectories, while values closer to one indicate a dominance of target-periodicity glimpses along the stimulus duration, resulting from a substantial misalignment between the competing F_0 trajectories. The time windows used for computing the PGI were 100-ms long and separated by 10-ms (corresponding to the durations employed to extract the F_0 trajectory using PRAAT).

The relationship between PGI and speech intelligibility was analyzed for the different TMR and $\Delta\overline{F_0}$ tested. To do this, the PGI values obtained for the individual stimuli were binned into twenty equally spaced intervals between 0 and 1 and average speech intelligibility and standard error were computed in

[c] The use of terminology that is typically used to describe theories of processes in the auditory domain (e.g., 'energetic masking' and 'energy glimpses') is intended to facilitate the description of the concept, without necessarily implying the existence of any auditory mechanism.

each interval. To avoid excessively large standard errors on speech intelligibility, only bins with more than 50 observations were included in the analysis. For comparison, PGI values were also computed for 6000 pairs generated by mixing random sentences from the CRM corpus, using the same talker in each pair and replicating the call sign, color and number combinations used by Darwin et al. (2003).

Periodicity glimpses may occur when the target voice is periodic and the masker voice is either active but non-periodic or simply inactive (i.e., silent). To quantify the relative contribution of the latter case to the overall number of periodicity glimpses, an energy-based voice-activity-detection (VAD) algorithm was applied to the masker signals in all sentence pairs used in the experiment, identifying time frames where the masker voice was silent. The VAD algorithm was implemented as follows: each signal was divided into 15-ms segments, separated by a time step of 10 ms. To allow a comparison with periodicity-glimpse detection, the outcome of the VAD algorithm was time-aligned to the F_0 trajectory in terms of the midpoints of the evaluated segments, whereas the segment duration differed (VAD: 15 ms; F_0 extraction: 100 ms) due to the different technical requirements. The root mean square of each segment was computed as a measure of local energy of the signal. If the energy in a segment was more than 20 dB lower than the maximum energy in all the segments in the signals, the segment was considered silent, and active otherwise. A comparison between periodicity glimpses and the masker-VAD results showed that 52% of periodicity glimpses occurred when the masker was silent. The same proportion was found when this analysis was conducted on the CRM sentence pairs.

3.3 Results

3.3.1 Measured speech intelligibility scores

Figure 3.2 shows the speech intelligibility scores for the individual listeners (1-15), indicated as the percentage of correctly repeated target words (P_C), averaged over the 20 sentence pairs presented in each experimental condition. The median speech intelligibility varied substantially across listeners, between 50% and about 80% P_C. At the level of the single sentence pair, some listeners (e.g., listeners 2 and 4) showed a large variability in their performance, ranging from 10% to perfect recognition. Others (e.g., listeners 9 and 14) showed a very small

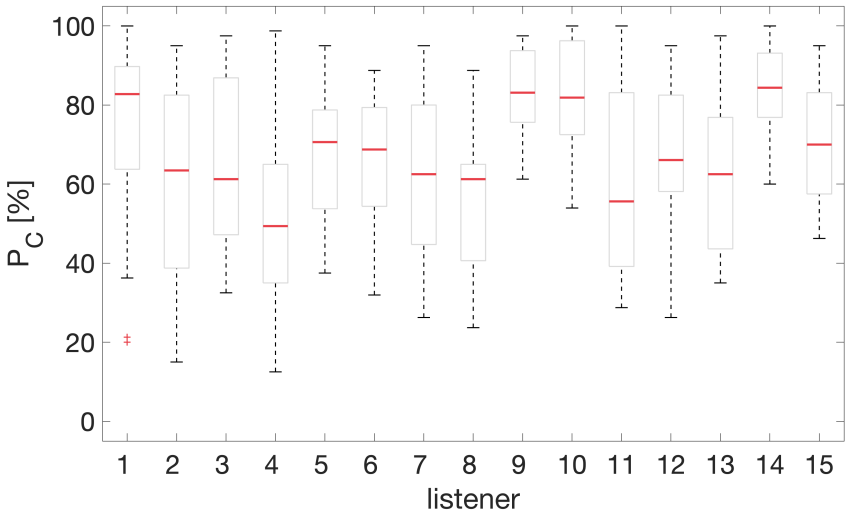


Figure 3.2: Boxplot of speech intelligibility scores for each of the 15 listeners, averaged across the 20 sentence pairs within each experimental condition. The plot includes data obtained in all experimental conditions. Central red marks indicate median performance values, the boxes indicate the range of data between the 25th and the 75th percentiles, the whiskers indicate the utmost non-outlier values and the red crosses indicate outliers (defined as values that are more than 1.5 times the interquartile range beyond the upper or lower edge of the box).

variability in performance, ranging from 60% to 100%.

The left panel of Figure 3.3 shows the mean results as a function of TMR, averaged across all listeners, with each $\Delta\overline{F_0}$ condition indicated by a different symbol and color. Overall, speech intelligibility was found to increase monotonically with increasing TMR for all $\Delta\overline{F_0}$ values. The strongest effect of $\Delta\overline{F_0}$ was observed at a TMR value of -8 dB. At this TMR, P_C increased by 15 percentage points from $\Delta\overline{F_0} = 0$ semitones (red circles) to $\Delta\overline{F_0} = 12$ semitones (blue diamonds). For TMRs at and above -4 dB, the four curves all converged to similar values. At the limits of the TMR range tested, only a minor effect of $\Delta\overline{F_0}$ was found.

A mixed-model analysis of variance (ANOVA) on the rationalized arcsine unit (RAU)-transformed data was conducted, including the factors listener (treated as random factor), $\Delta\overline{F_0}$ and TMR (treated as fixed factors). Two-way interactions were included to analyze the interactions between these three main experimental factors. The results of the ANOVA revealed that all main factors were statistically significant ($p<0.01$), indicating that speech intelligibility differed significantly across listeners and that both $\Delta\overline{F_0}$ and TMR were significant

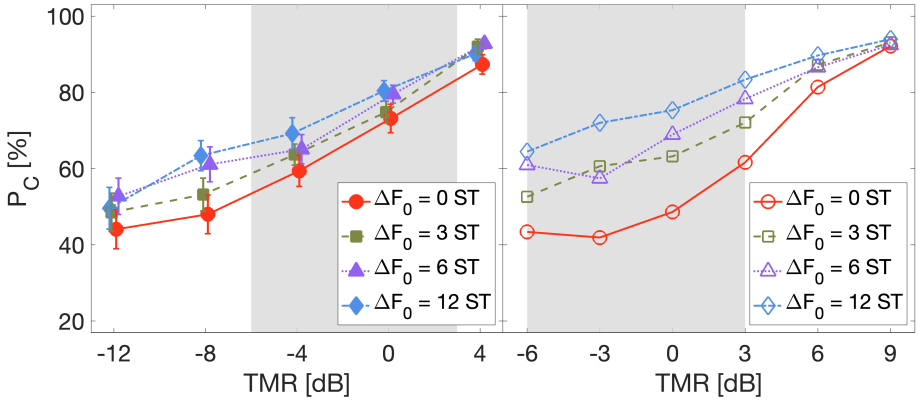


Figure 3.3: Speech intelligibility scores, shown as proportion of correctly identified target words (P_C) as a function of TMR, averaged across listeners. The different curves show the results for different $\Delta\overline{F_0}$s, indicated by different symbols and colors. Left panel: Data from the current study; error bars represent standard errors. Right panel: Data from Darwin et al. (2003) for comparison. The grey areas indicate the TMR range in common between the two studies.

sources of variation in the experiment. However, no significant interactions between the main factors were observed, with the exception of an interaction between listener and TMR.

For direct comparison, the data from the reference study by Darwin et al. (2003) are shown in the right panel of Figure 3.3, using the same color and symbol coding as in the left panel. The grey area in the two panels of Figure 3.3 indicates the range of TMRs values that were in common between the two studies. Darwin et al. (2003) also observed a monotonically increasing speech intelligibility with increasing TMR, for all F_0 separations except $\Delta\overline{F_0} = 6$ semitones, where a local minimum was found at TMR=-3 dB. In their experiment, the largest benefit of $\Delta\overline{F_0}$ on speech intelligibility was obtained for TMRs in the range from -6 dB to 3 dB and decreased for increasing TMR, becoming negligible at the higher applied TMRs where a ceiling effect was observed. A maximum improvement in speech intelligibility of 30 percentage points was observed in their study at a TMR of -3 dB, between $\Delta\overline{F_0} = 0$ (red circles) and $\Delta\overline{F_0} = 12$ semitones (blue diamonds). In the range of TMRs in common with the current study (grey area), the data from Darwin et al. (2003) showed the largest effect of $\Delta\overline{F_0}$ on P_C, whereas the data from the present study did not show significant differences.

In Figure 3.4, the filled circles indicate the overall effect of $\Delta\overline{F_0}$ on speech intelligibility found in the present study, averaged across TMRs within the interval [-8, 0] dB. Across this range of TMRs, the average percentage of correct words in-

creased by about ten percentage points across the $\Delta\overline{F_0}$ tested. For $\Delta\overline{F_0} = 0$ semitones, an average P_C of 60% was obtained. P_C increased by about 4 percentage points for $\Delta\overline{F_0} = 3$ semitones and 9 percentage points for $\Delta\overline{F_0} = 6$ semitones. Larger $\Delta\overline{F_0}$s did not provide additional improvements as speech intelligibility saturated at P_C =70% for F_0 separations beyond 6 semitones. A post-hoc pairwise comparison analysis of the data showed that the difference between $\Delta\overline{F_0} = 0$ semitones and $\Delta\overline{F_0} = 6$ or 12 semitones was statistically significant at the $p < 0.001$ level. For comparison, the data from Darwin et al. (2003) are also indicated (open triangles), averaged across their three lowest TMRs ranging from -6 to 0 dB. In contrast to the results from the present study, the data from Darwin et al. (2003) showed a larger average improvement in terms of percentage of correct words (25 percentage points over the same range of $\Delta\overline{F_0}$s), with a minimum speech intelligibility of 45% for $\Delta\overline{F_0} = 0$ semitones. In Darwin et al. (2003), P_C increased monotonically with increasing $\Delta\overline{F_0}$ and an average F_0 separation of three semitones was sufficient to induce an increase of P_C by 14 percentage points. The overall level of P_C in Darwin et al. (2003) was lower than the one measured in the present study. The largest difference in terms of P_C between the data of the two studies was found at $\Delta\overline{F_0} = 0$ semitones (16 percentage points). The difference was only 5 and 6 percentage points at $\Delta\overline{F_0}$ of 3 and 6 semitones, respectively, and vanished at $\Delta\overline{F_0} = 12$ semitones.

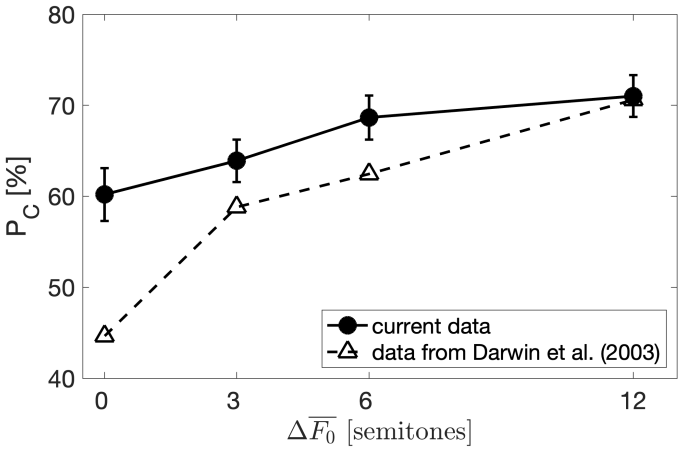


Figure 3.4: Performance as a function of $\Delta\overline{F_0}$, averaged across TMRs within the interval [-8,0] dB, shown with filled circles. Error bars represent standard errors. Data from Darwin et al. (2003), averaged across TMRs within the interval [-6,0] dB, are shown with open triangles. The TMR intervals were chosen to maximally cover the TMR values in common between the two studies.

Analysis of the effects of speech synchrony

The PGI, as defined in Section 3.2.5, was calculated for the stimuli employed in the experiment and was related to the measured speech intelligibility data. Panel A of Figure 3.5 shows P_C, averaged across all experimental conditions, as a function of the obtained PGI values. Mean and standard error of P_C were computed in each of the twenty bins over which the PGI values were partitioned. It can be seen that P_C and PGI are strongly correlated (ρ=0.97, p<0.01). Panels B and C of Figure 3.5 show a separate analysis for portions of the data that represented the most extreme $\Delta\overline{F_0}$ conditions, i.e., at TMRs of -8 and +4 dB, respectively. The results obtained for $\Delta\overline{F_0} = 0$ semitones are indicated by the red symbols and those obtained for $\Delta\overline{F_0} = 12$ semitones are shown with the blue symbols. The dashed lines indicate linear fits to the data. At TMR=-8 dB (panel B), a large spread of P_C was observed between the two $\Delta\overline{F_0}$ conditions at low PGI values (PGI < 0.1), where P_C improved by 21 percentage points when a $\Delta\overline{F_0}$ of 12 semitones was introduced, compared to the baseline condition with $\Delta\overline{F_0} = 0$ semitones. This difference decreases with increasing PGI and becomes negligible at the highest PGI. For $\Delta\overline{F_0} = 0$ semitones, P_C increases with higher PGI values, while for $\Delta\overline{F_0} = 12$ semitones P_C is independent of PGI. In contrast, at TMR=4 dB (panel C), ceiling performance was observed both for $\Delta\overline{F_0} = 0$ and for $\Delta\overline{F_0} = 12$ semitones and no dependency of P_C on PGI was found.

Panel D of Figure 3.5 shows a probability histogram of the PGI occurrences distributed across the stimuli presented during the experiment (6000 observations in total), indicated by the grey bars. For comparison, the corresponding probability histogram of the PGI values obtained for the sentence pairs from the CRM corpus (as used in Darwin et al., 2003) is indicated by the orange bars. Since the PGI is a metric based only on the periodic activation of the target and the masker signals and not on their intensity, the distributions are the same for all TMRs. In the case of the experimental stimuli employed in the current study, the PGI values ranged from a minimum of 0 to a maximum of 0.58 and the largest probability of occurrence was found for PGI values between 0.1 and 0.3 (almost 80% of the observations fell in this interval). In contrast, in the case of the CRM corpus, the PGI values (ranging between 0 and 0.34) showed a peak in the histogram at lower values, with more than 80% of the observations falling below 0.15. Overall, the HINT sentence pairs contain more periodicity glimpses as a consequence of the substantially smaller overlap between competing F_0

trajectories, as compared to the CRM sentences.

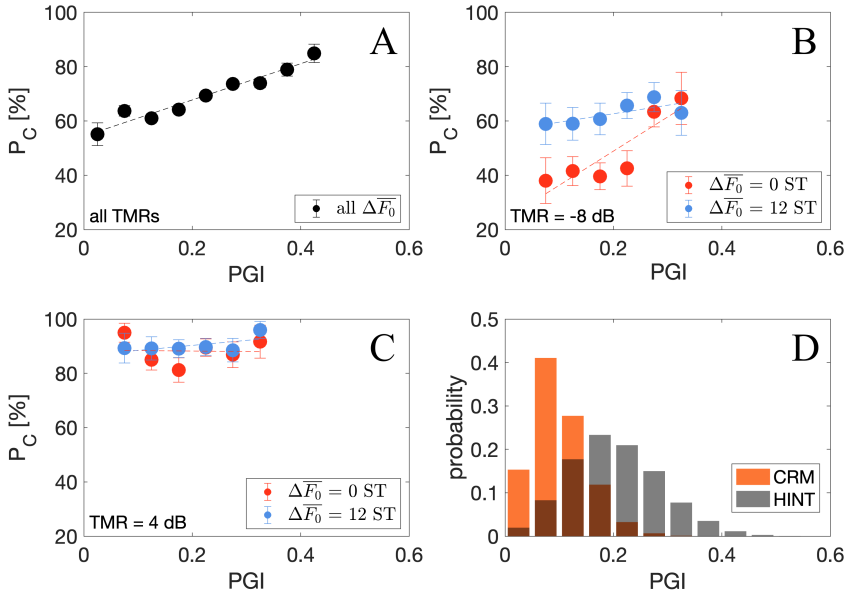


Figure 3.5: Results of periodicity-glimpse index (PGI) analysis. Panel A: Speech intelligibility as a function of PGI, averaged across all experimental conditions. Panels B and C: Speech intelligibility as a function of PGI for TMR=-8 dB and TMR=4 dB, respectively, for the two most extreme $\Delta\overline{F_0}$ condition of 0 semitones (red circles) and 12 semitones (blue circles). Panel D: Probability histogram of occurrence for PGI values obtained with HINT sentences (grey bars) and CRM (orange bars). In all panels, the PGI values are binned into 20 equally-spaced intervals between 0 and 1. In panels A, B and C, average speech intelligibility and standard errors computed within PGI bins are shown for bins with more than 50 observations; the dashed lines represent linear fits to the data.

3.4 Discussion

The main finding of the present study was the very moderate effect of the average-F_0 separation ($\Delta\overline{F_0}$) between competing sentences on target-speech intelligibility found with realistic speech material. An effect of $\Delta\overline{F_0}$ was observed (i) only at small TMRs (-4 and -8 dB) that are well below TMR values reflecting typical conversational conditions (Smeds et al., 2015) and (ii) only for $\Delta\overline{F_0}$s of six semitones or above. In contrast, the results from previous studies that employed less realistic speech materials (e.g., Darwin et al., 2003) showed speech intelligibility improvements that were at least twice as large as those found in the present study and that were obtained at smaller $\Delta\overline{F_0}$s and higher TMRs. In the experimental conditions ($\Delta\overline{F_0}$s and higher TMRs) that were in common with

Darwin et al., 2003, the results of the present study showed higher intelligibility scores.

The PGI metric, introduced to quantify the access to 'clean' target-periodicity information resulting from the level of asynchrony between competing F_0 trajectories, was found to be a good predictor of speech intelligibility (Figure 3.5, panel A). Furthermore, PGI and $\Delta\overline{F_0}$ appeared to act as two counterbalancing factors: when the PGI was small (i.e., when the competing F_0 trajectories were highly synchronous and only a small amount of the periodic information of the target occurred in the non-periodic or silent segments of the masker), the target-speech intelligibility was lowest in absence of average-F_0 differences ($\Delta\overline{F_0} = 0$ semitones) and increased substantially when a non-zero $\Delta\overline{F_0}$ between the competing voices was introduced. In contrast, when the PGI was high (i.e., when the competing F_0 trajectories where largely asynchronous and a substantial part of the periodic information of the target occurred in the non-periodic or silent segments of the masker), the intelligibility of the target speech in absence of average-F_0 differences was higher and was not further improved by the introduction of a non-zero $\Delta\overline{F_0}$. This relation between PGI and $\Delta\overline{F_0}$ and their effect on speech intelligibility was observed at low TMRs (e.g., -8 dB, Figure 3.5, panel B), where target speech intelligibility increased with increasing PGI and, for low PGI values, benefited from a non-zero $\Delta\overline{F_0}$. At the highest TMR tested (4 dB), performance reached ceiling level and was independent of PGI and $\Delta\overline{F_0}$ (Figure 3.5, panel C). At this TMR, the higher level of the target was sufficient to make it intelligible, with speech intelligibility levels as high as 90% even in the condition without F_0 separation ($\Delta F_0 = 0$). It thus seems that at positive TMR values, reflecting real-life conversational conditions, the access to target-speech information is sufficient for good speech understanding and additional cues (provided either as a high PGI or a $\Delta\overline{F_0}$) become redundant, whereas they are beneficial when more detrimental TMRs limit the intelligibility of the target sentence. Indeed, the largest effect of $\Delta\overline{F_0}$ in this study was found at TMR=-8 dB.

The realistic, unconstrained structure of the HINT sentences yields a large variation in the amount of synchrony between the target and the masker speech across sentence pairs, resulting in a wider range of PGI values. In contrast, the fixed structure of the CRM sentences used by Darwin et al. (2003) produces high levels of speech synchrony between competing sentences and thus a substantial time alignment of their F_0 trajectories. Therefore, the PGI values obtained with CRM sentences are limited to low values and a narrower range compared to

HINT sentences (Figure 3.5, panel D). These lower PGI values are less likely to occur with HINT sentences and are not representative of realistic competing-talker scenarios but yielded by far the largest effect of $\Delta\overline{F_0}$ in the current study. The analysis based on the PGI thus offers a possible explanation as to why differences in average F_0 yield greater changes in intelligibility for CRM sentences than for HINT sentences.

The F_0 processing used in the present study preserved the dynamics naturally present in the F_0 trajectories of the HINT sentences. F_0 dynamics are known to convey important information about the speech message and to aid speech intelligibility in adverse listening conditions (Binns and Culling, 2007; Calandruccio et al., 2019). When such dynamics are completely removed, as in Summers and Leek (1998) and one of the experiments reported by Assmann (1999), or strongly limited, as in Darwin et al. (2003), speech intelligibility is decreased and the F_0 separation becomes a dominating segregation cue. In fact, the effect of $\Delta\overline{F_0}$ on speech intelligibility observed by Darwin et al. (2003, their Figs. 2 and 3) was substantially smaller for one talker that had stronger F_0 dynamics than the other talkers used in that experiment. Similarly, Assmann (1999) observed a larger benefit of $\Delta\overline{F_0}$ for monotonous F_0 trajectories than for naturally varying ones.

The constraints imposed on the structure of the CRM sentences also removed several linguistic cues related to context, syntax and semantics that are otherwise normally present in real-life speech and considerably aid the segregation and perception of the target speech (Boothroyd and Nittrouer, 1988). Furthermore, in the experiment by Darwin et al. (2003), only two words from the target sentence had to be repeated (color and number) and chosen within a closed set of possibilities. Due to the similarity in syntax and structure in the CRM sentences, the color and number words appeared almost simultaneously in the target and masker sentences. Thus, in the absence of any context cue, the color and number words may be easily assigned to the wrong sentence, especially when pairing sentences from the same talker and with a zero-semitone F_0 separation. This might have lowered the chance of correct responses for the smallest $\Delta\overline{F_0}$ and therefore increased the difference in speech intelligibility between conditions of small and large $\Delta\overline{F_0}$. In contrast, in experiments employing the HINT material, the listener's task is to repeat the entire target sentence (four out of five words in the case of the present experiment, with the first word being the cue) by choosing any possible combination of Danish words without any

constraints on the sentence structure, provided its syntactical and grammatical correctness.

Therefore, the syntax and context present in the HINT sentences may have provided linguistic cues to understand the target speech, thus limiting the utility of the $\Delta\overline{F_0}$ cue. With the competing vowels used by Summers and Leek (1998), the reduction of both linguistic and periodicity cues was even more exaggerated. In fact, neither context nor syntax were available in those conditions and the PGI is expected to be zero for all stimuli, since both target and masker had continuous, monotonous, F_0 trajectories, hence not providing periodicity glimpses or F_0 dynamics cues. As a consequence, Summers and Leek (1998) observed beneficial effects for $\Delta\overline{F_0}$s as small as fractions of a semitone.

While the contribution of previous investigations has been essential to the knowledge of the role of F_0 differences in speech perception, their findings do not seem to generalize to more realistic speech materials. In realistic competing-talker scenarios, various cues contribute to the perception of target speech. The experimental methods employed in the previous studies eliminated several of these cues and thus promoted the contribution of $\Delta\overline{F_0}$ as the predominant available cue. In particular, high PGI values (due to asynchronous periodic activation of target and masker speech) and fast dynamic changes in F_0 (and thus in the instantaneous F_0 separation between the competing speech signals) are typical of more realistic speech stimuli, but their contributions were limited substantially by the choice of speech materials in most of the previous studies.

When creating pairs of sentences with either the HINT or the CRM speech corpus, about half of the periodicity glimpses occur in the silences of the masker signal (see Section 3.2.5). These time frames might be described more generally as "voice glimpses", i.e., as time segments where the target is active (and periodic) while the masker is silent. The remaining half of periodicity glimpses occurs when the masker is active but not periodic. Even though it is not possible to make a direct comparison between periodicity glimpses and voice glimpses (since the two measures are calculated over time windows that differ in duration), there is a possibility that the results of the PGI analysis might be a consequence of voice-activity asynchrony between competing signals, rather than periodicity asynchrony per se. The proposed PGI analysis is thus not meant to represent any hypothesis related to the auditory processes involved in competing-voice separation, but rather provides a description of the properties of the competing speech signals. Nonetheless, the PGI is correlated with speech intelligibility,

accounts for the improvements in speech intelligibility when a $\Delta\overline{F_0}$ is introduced between competing voices and offers a potential explanation for the differences in results observed across the different studies. Furthermore, since the focus of this study was on periodicity cues, it was considered most appropriate to interpret the results from the perspective of periodicity information. However, it remains an open question which auditory processes can account for the observed effects in the data. What appears evident from this study is that the level of (a)synchrony between competing speech signals is a feature of the speech corpus that highly impacts the occurrence and size of $\Delta\overline{F_0}$-related effects on the speech intelligibility results and should be carefully considered in the design of speech-intelligibility experiments.

The notion that periodicity synchrony is detrimental for the perception of target speech (see Figure 3.5) seems to be in contrast with previous studies showing that the simultaneous periodic activation of target and masker signals aids speech intelligibility via a mechanism of "harmonic cancellation" (e.g., Cheveigné et al., 1995; Prud'Homme et al., 2020; Steinmetzger and Rosen, 2015). However, a direct comparison of these studies with the present one is not possible since the competing synthetic vowels of Cheveigné et al. (1995), the non-speech maskers employed by Steinmetzger and Rosen (2015) or the stimuli used in the computational model simulations conducted by Prud'Homme et al. (2020) differ substantially from the competing sentences employed in the present study.

The small $\Delta\overline{F_0}$ benefit found in the present study at first glance appears to be in disagreement with some previous research that also employed realistic speech materials (Assmann, 1999; Başkent and Gaudrain, 2016; Flaherty et al., 2021). However, important differences should be noted between their approaches to provide more realistic experimental paradigms and the approach adopted in the present study. For example, the masker voice used in Başkent and Gaudrain (2016) was created using excerpts of one (or more) sentences, with each excerpt cut from the ending of the sentences, producing a concatenated masker voice that lacked natural prosody, meaning, and context. A similar absence of linguistic cues is expected also in the masker signals used by Flaherty et al. (2021), which were created by mixing two speech signals consisting of concatenated sentences from the target talker. Furthermore, Flaherty et al. (2021) edited the two-talker masker by removing silences longer than 200 ms and likely increased the probability of simultaneous periodic (and therefore

voice) activity between the competing speech stimuli, leading to lower PGI values, with less periodicity-glimpsing opportunities and therefore potentially stronger effects of $\Delta\overline{F_0}$. Another crucial aspect of competing-talker experiments that can largely influence the results is the method used for cueing the target voice. This aspect may provide an additional explanation for the difference in magnitude of the $\Delta\overline{F_0}$ effect found between the present results and the ones from some of the previous studies. In the experiment proposed by Başkent and Gaudrain (2016), Flaherty et al. (2021) and Assmann (1999), the target cue was either based on a short target-voice-recognition training prior to the experiment or not provided at all. In either case, since these three studies used the same talkers for target and masker, the two competing voices were very similar for the $\Delta\overline{F_0} = 0$ semitones condition and therefore likely difficult to segregate given the absence of an F_0-separation cue and the effectively absent target cue, resulting in low speech intelligibility. When a non-zero $\Delta\overline{F_0}$ was introduced the target and masker voices became presumably easier to segregate. As a result, a large difference in performance between a condition with $\Delta\overline{F_0} = 0$ semitones and a condition with positive $\Delta\overline{F_0}$ conditions could be observed. In contrast, the target pre-cue used in the present experiment produced the same advantage regardless of the $\Delta\overline{F_0}$ condition tested, that can potentially explain the smaller magnitude of the $\Delta\overline{F_0}$ effect observed.

The present study aimed to assess the importance of a difference in F_0 between two competing voices employing speech materials that are more realistic compared to the speech materials employed in some of the previous studies. However, the experimental approach in its entirety is far from being fully realistic. One aspect that limited the realism of the present experimental stimuli was the onset alignment between target and masker sentences, since misaligned onsets may be considered more typical of realistic situations. It appears more likely in daily life that a target talker speaks in the presence of one (or several) continuous background voices. In such situations, the listener can integrate the F_0 information of the background masker, compare it with that of a target once the target occurs as a new auditory event, and make use of potential F_0 differences between the two streams. This was not possible in the present experiment as the competing sentences were aligned at their onsets. Previous studies that took this aspect into account (Assmann, 1999; Başkent and Gaudrain, 2016; Flaherty et al., 2021) obtained larger $\Delta\overline{F_0}$ effects on speech intelligibility, indicating that this might be a relevant aspect in realistic situations. However, due

to the other methodological differences discussed above, a clear conclusion cannot be drawn regarding the impact of same/different onset times of target and masker.

Finally, to maintain a sufficient level of control and focus on the experimental variable of interest, the proposed experimental design dispensed with other aspects that are typical of real-life competing-voice scenarios. This concerns, for example, the absence of spatial cues due to the co-located talker condition, the lack of talker-specific acoustic attributes like vocal-tract differences due to the use of the same talker in target and masker signals, as well as the absence of visual cues. Furthermore, realistic competing-talker scenarios are usually characterized by the presence of background noise which was not used in the present study. For these reasons, the present experimental paradigm cannot be considered fully realistic, but it nevertheless may contribute to a better understanding of the role of F_0-related cues to speech intelligibility along the continuum between unnatural but highly-controlled listening conditions and the more variable and realistic ones that the listener experiences in daily life.

3.5 Summary and conclusions

The present study investigated the effect of long-term average F_0 separation ($\Delta\overline{F_0}$) on the intelligibility of real-life target sentences in competing-talker conditions. The effect of $\Delta\overline{F_0}$ was found to be moderate at negative target-to-masker ratios (TMRs) and negligible at positive TMRs. This result is in contrast with previous studies that employed less realistic speech materials and reported a large speech-intelligibility benefit induced by $\Delta\overline{F_0}$. A detailed analysis of the F_0 trajectories of the competing sentences revealed a considerable impact of their time alignment on the effectiveness of the $\Delta\overline{F_0}$ cue on speech intelligibility. It was found that the effect of $\Delta\overline{F_0}$ on speech intelligibility at low TMRs was substantial when the competing F_0 trajectories showed high synchrony (i.e., when target and masker were simultaneously periodic), whereas the effect was absent when they did not. This observation may explain the difference between the current and previous studies, as previously employed speech materials led to very high levels of periodicity synchrony in the sentence pairs, which may have amplified the effect of $\Delta\overline{F_0}$. Additionally, realistic speech offers a multitude of linguistic and auditory cues (e.g., syntax, context, and F_0-trajectory dynamics)

that may aid speech intelligibility but were limited or completely removed in some of the previous studies, therefore contributing to an exaggerated effect of the average F_0 separation.

The results presented in this study cast new light on the importance of long-term average F_0 separation for the intelligibility of realistic competing speech, suggesting that this cue is beneficial only in certain listening conditions. Other aspects related to the periodicity of speech signals might play a more important role. For example, further research might be directed to assess the role of instantaneous F_0 separation and of the difference in F_0 dynamics between competing talkers. It furthermore remains an open question to be investigated whether other aspects that are typical of more realistic stimulus presentation (e.g., an offset misalignment between competing voices, the number of masker voices or the degree of linguistic cues available in the masker) contribute to the importance of the $\Delta\overline{F_0}$ cue on speech intelligibility in everyday life.

Acknowledgments

I would like to thank Rikke Skovhøj Sørensen for her help in collecting the speech intelligibility measurements reported in this study.

4

Effects of fundamental-frequency dynamics on sentence intelligibility in competing-talker scenarios[a]

Abstract

Differences in the dynamic range of the fundamental frequency (F_0) between competing voices (i.e., an F_0 dynamic range contrast) have recently been shown to facilitate the segregation of the target speech signal from the interfering speech. Previous studies investigated the F_0-dynamic-range contrast by pairing voices speaking with different levels of intonation and using a fixed combination of target and masking talkers. The present study aimed at extending the previous findings by using a larger variety of talkers, levels of F_0 dynamic range and F_0-dynamic-range contrast. To isolate F_0-related effects, naturally produced speech recordings were manipulated using a signal-processing method that allows to control the F_0 information without affecting other acoustic features of speech. Target speech intelligibility was measured in young normal-hearing listeners as a function of the F_0-dynamic range contrast between two competing sentences (spoken by the same talker), which had either the same or different average F_0. Speech intelligibility (i) was only moderately affected by the F_0-dynamic-range contrast, both in presence and absence of a difference in average F_0, (ii) was lowest when both sentences had a small F_0 dynamic range and (iii) increased when a moderate level of F_0 dynamics was introduced in at least one of the sentences, regardless of the F_0-dynamic-range contrast between them. These findings suggest that the overall dissimilarity between

[a] This chapter is based on Mesiano et al. (2022b), in preparation for submission to the Journal of Speech, Language, and Hearing Research.

the F_0 trajectories of the competing sentences, rather than their F_0-dynamic-range contrast, may be utilized by the listeners when performing the speech segregation task.

4.1 Introduction

In daily life, we are often faced with the complex auditory task of attending to speech from a target voice in the presence of one or several interfering voices. In such competing-talker scenarios, to segregate the target speech signal from the mixture, the healthy auditory system can utilize several cues that are related to the properties of the signals in the acoustic scenario, such as the spatial configuration of the sound sources, their spectral differences, or the amplitude modulations of the individual signals (Bronkhorst, 2000; Cherry, 1953). The fundamental frequency (F_0) is another signal property that characterizes the individual sources and provides relevant cues in support of the auditory challenges in competing-talker scenarios. In fact, the F_0 trajectory of a speech signal (i.e., its evolution over time) carries important information about certain features of the voice, such as the talker's sex (Honorof and Whalen, 2010), the talker's intentions and emotions (Arnott, 1993), as well as prosodic aspects of the speech signal that can serve as cues for syntax and grammatical structure (Ladd, 2008). The F_0 information and the differences in F_0 between competing voices can thus help segregate competing speech signals and aid speech understanding in complex acoustic scenarios.

When two competing voices differ in the time average of their F_0, the listener can more easily attend to a desired target voice and separate it from the other voice, compared to a situation where the two voices have a similar average F_0 (Assmann, 1999; Darwin et al., 2003; Flaherty et al., 2021; Summers and Leek, 1998; Chapter 3 of this thesis). The mentioned studies used different types of speech material that varied in their similarity to real-life speech, ranging from synthetic vowels with monotonous F_0 trajectories or meaningless sentences to naturally produced everyday sentences. The largest speech-intelligibility improvements due to an average F_0 separation between competing voices were found for speech materials that differed substantially from real-life speech. For example, experiments using synthetic vowels or materials lacking context and linguistic coherence and having unrealistic similarity in structure, as in the case of the coordinate response measure (CRM; Bolia et al., 2000), reported

substantial benefits of introducing average F_0 differences between target and masker speech (Darwin et al., 2003; Summers and Leek, 1998). Real-life speech provides a multitude of cues, such as context, prosody and grammatical coherence, which can also support speech intelligibility of competing sentences, but that may have been removed or highly constrained by the use of unrealistic speech materials (see Chapter 3). Indeed, studies that employed more realistic speech materials, such as meaningful sentences, have found smaller effects for the same average F_0 separations (Assmann, 1999; Flaherty et al., 2021; Summers and Leek, 1998; Chapter 3) than those using rather artificial speech materials. Therefore, the use of such artificial speech materials may have enhanced the effect of the average F_0 separation between competing voices on speech intelligibility, by removing or limiting several other cues that aid speech intelligibility in realistic competing-talker scenarios.

The F_0 variations of realistic speech signals represents one of the potential cues for speech intelligibility that has been removed or limited in some of the studies that investigated the effect of average F_0 separation. In fact, both the synthetic vowels used by Summers and Leek (1998) and the CRM sentences used by Darwin et al. (2003) had F_0 trajectories that were either flat (i.e., monotonous) or had a dynamic range smaller than that found in realistic speech. Assmann (1999) observed that the effect of average F_0 separation on speech intelligibility was stronger when the F_0 variations were removed from the speech signals, compared to the effect found when leaving the F_0 variations at their natural magnitude. Furthermore, speech perception in noise and in the presence of competing speech has been shown to decrease when the natural F_0 variations are reduced or removed from the target speech signal (Binns and Culling, 2007; Laures and Bunton, 2003; Laures and Weismer, 1999; Miller et al., 2010). These findings suggest that the F_0 dynamics and the F_0 dynamic range that characterize real-life speech may be crucial for understanding speech in complex acoustic scenarios.

Not only the F_0 dynamics and the F_0 dynamic range of an individual speech signal provide important cues for speech perception, but also the difference in F_0-dynamic range between competing speech signals seems to aid speech segregation. This was shown by Calandruccio et al. (2019) who measured speech reception thresholds (SRTs) of real-life sentences masked by a two-talker interferer as a function of the contrast in F_0-dynamic-range between the target and the masker signals (i.e., the difference in the range of their F_0 trajectories). To

create the F_0-dynamic-range contrast, Calandruccio et al. (2019) paired speech signals with F_0 dynamic ranges of different magnitudes, produced by recording speech with 'normal', 'flat' and 'exaggerated' speaking styles. These speaking styles were obtained by asking the talkers to respectively speak naturally, with "a monotone voice pitch (as if they were sad)" or with "wide variations in voice pitch (as if they were happy and excited)". The resulting mixtures of speech were assumed to have no contrast in F_0-dynamic range (i.e., the competing speech signals had the same speaking style), a contrast with stronger F_0 variations in the target or a contrast with stronger F_0 variations in the masker. Calandruccio et al. (2019) measured higher SRTs (i.e., lower speech intelligibility) when target and masker speaking styles were matched (i.e., no F_0-dynamic-range contrast) and lower SRTs (i.e., higher speech intelligibility) when target and masker were spoken with different speaking styles. In their study, speech intelligibility was not only influenced by matching or unmatching speaking styles, but also by the specific combination of the individual target and masker speaking styles. In fact, for speech stimuli with matched speaking style, speech intelligibility was higher when both sentences were spoken 'normally' compared to the conditions of flat-versus-flat or exaggerated-versus-exaggerated speech, which instead led to the worst speech intelligibility scores overall. Furthermore, in the unmatched speaking style conditions, the highest speech intelligibility was obtained by pairing a flat target with an exaggerated masker, whereas the inverse condition (i.e., an exaggerated target paired with a flat masker) led to slightly lower speech intelligibility scores. In summary, the results from Calandruccio et al. (2019) indicated that (i) a F_0-dynamic-range contrast between competing voices can be beneficial for speech intelligibility and (ii) the magnitude of this benefit depends on the specific F_0 dynamic ranges of target and masker that are used to create the contrast.

The present study extended the investigation of Calandruccio et al. (2019) by (i) using a wide variability of F_0 dynamic ranges that were combined at different levels of F_0-dynamic-range contrast and (ii) using a processing method of the speech stimuli that targeted only their F_0 trajectories. It was hypothesized that (i) an F_0-dynamic-range contrast between competing voices would improve target speech intelligibility and that (ii) such improvement would depend on the specific F_0-dynamic-range combination of target and masker sentences used for generating the contrast. Speech intelligibility was measured in normal-hearing listeners by presenting pairs of daily-life sentences (both spoken by

the same talker) with varying degrees of F_0-dynamic-range contrast that were generated with a variety of target and masker F_0-dynamic-range combinations and employing recordings of the speech material from several male and female talkers. The F_0-dynamic-range combinations were obtained by modifying the F_0 trajectories of the individual speech signals with a numerically-accurate signal-processing method, preserving the range of natural F_0 variations observed in the unmodified speech material. The results were compared with those from Calandruccio et al. (2019) and discussed in relation to the differences in the experimental design, the speech stimuli and the methods for controlling the magnitude of the F_0 dynamic range in the speech stimuli.

4.2 Methods

4.2.1 Participants

Nineteen NH listeners (12 females), aged between 19 and 33 (mean 25) years participated in the study. All participants were native Danish listeners, with pure-tone thresholds below 20 dB Hearing Level between 125 Hz and 8 kHz. Informed consent was provided by all participants and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). All participants completed the speech intelligibility experiment in a single experimental session that lasted no more than two hours.

4.2.2 Stimuli

The experimental stimuli were pairs of sentences from the Danish HINT corpus (Nielsen and Dau, 2011). The Danish HINT consists of 200 open-set, five-words, daily-life sentences, divided into ten phonetically balanced lists. Recordings of the speech material were available from twelve different talkers (six males and six females, all native Danish speakers): the original recordings of a male talker (later labelled as 'M1') from Nielsen and Dau (2011) and eleven additional recordings provided by Eriksholm Research Centre (Bramsløw et al., 2019).

The stimulus generation required a prior analysis of the F_0 information contained in the speech corpus, which was analyzed by extracting the F_0 trajectory of each of the 200 sentences in the speech corpus for each talker. The F_0 trajectories were extracted using the software PRAAT (Boersma et al., 1993) version 6.1.37, using a time step of 10 ms, 100-ms time windows (i.e., 90% overlap) and

searching for F_0 candidate values within the range 50-450 Hz. For each sentence, the time average and dynamic range of the F_0 trajectory were computed. The average F_0 was quantified with the F_0 median (indicated with the symbol $\overline{F_0}$), while the F_0 dynamic range was quantified with the F_0 median absolute deviation (MAD, indicated with the symbol σ and defined as $\sigma(F_0) = \text{median}\left(\left|F_0 - \overline{F_0}\right|\right)$, with F_0 indicating the array of fundamental frequency values representing the trajectory). The advantage of median moments in measuring the statistics of the F_0 trajectories is that they underweight the influence of potential outliers (such as erroneous F_0 estimates that PRAAT can generate by assigning the F_0 values to the wrong octave, i.e., producing unrealistic 'octave jumps') and that they are more suitable for data that may not be normally distributed, such as the F_0 of the voice. For each of the 12 talkers, the F_0 median was also computed over the entire speech corpus (referred to as talker median F_0 and indicated as $\overline{F_0}^{talker}$). The analysis of the extracted information showed that the range of $\sigma(F_0)$ values of the sentences is talker dependent, with talkers characterized by a higher $\overline{F_0}^{talker}$ showing wider variations in the $\sigma(F_0)$ of their sentences, compared to talkers with a lower $\overline{F_0}^{talker}$. Additionally, the lower and upper limit of the individual $\sigma(F_0)$ ranges increased with increasing $\overline{F_0}^{talker}$. This is illustrated in Figure 4.1, which shows $\sigma(F_0)$ as a function of $\overline{F_0}$ of each sentence for the different talkers, together with the individual talker's $\sigma(F_0)$ ranges, represented as vertical bars. For each talker, the upper and lower limit of the $\sigma(F_0)$ range were obtained as follows. The lower limit $\sigma_{min}(F_0)$ was set to the minimum $\sigma(F_0)$ value across the sentences spoken by the talker. The upper limit $\sigma_{max}(F_0)$ was estimated as the 98$^{\text{th}}$ percentile of the $\sigma(F_0)$ distribution of each talker. The individual values of $\overline{F_0}^{talker}$ and upper and lower limit of the $\sigma(F_0)$ range of each talker are indicated in Table 4.1.

The experimental stimuli were generated by mixing two sentences that were randomly selected from different sentence lists, spoken by the same talker. The sentences were processed with PRAAT to obtain the desired average F_0 separation and F_0-dynamic-range contrast. The average F_0 separation (indicated as $\Delta\overline{F_0}$) was quantified in terms of the absolute value of the difference between the $\overline{F_0}$ of the competing trajectories ($\Delta\overline{F_0} = \left|\overline{F_0}^{target} - \overline{F_0}^{masker}\right|$). $\Delta\overline{F_0}$ of 0 and 6 semitones were used in the experiment. The F_0-dynamic-range contrast (indicated as R) was quantified as the natural logarithm of the ratio between target and masker F_0-dynamic ranges ($R = \log\frac{\sigma(F_0^{target})}{\sigma(F_0^{masker})}$). The F_0 trajectories of target and masker sentences were processed such that the resulting log-ratio of their

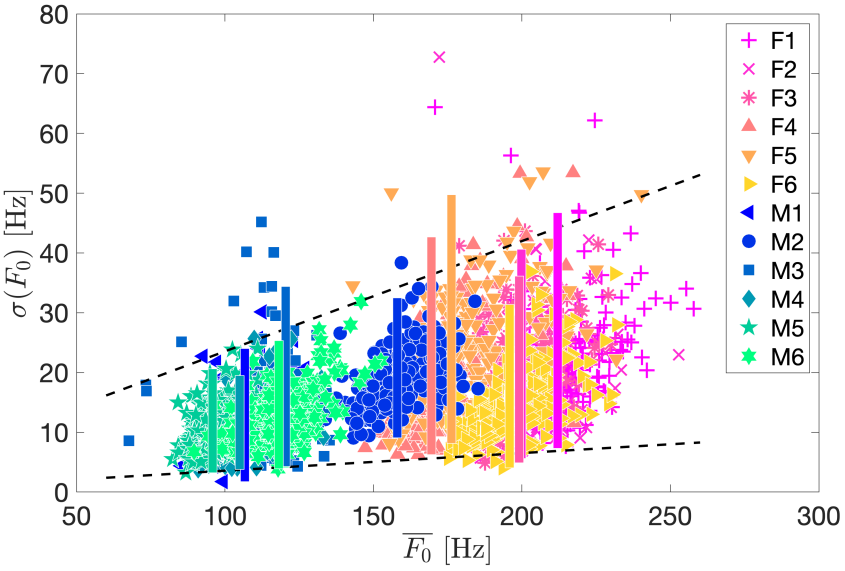


Figure 4.1: F_0 statistics of the 200 sentences in the HINT corpus for each of the 12 talkers. Each point represents the F_0 dynamic range ($\sigma(F_0)$) of a single sentence as a function of its F_0 median ($\overline{F_0}$). Sentences spoken by different talkers are shown with different colors and symbols. The vertical colored bars represent the $\sigma(F_0)$ range found over the sentences spoken by each talker. The dashed straight lines represent linear fits to the lower and upper limits of the $\sigma(F_0)$ ranges over the 12 talkers.

$\sigma(F_0)$s matched the desired R. Seven different R values were used, ranging from $R=-1.8$ to $R=1.8$ in 0.6 steps, and were combined with the two $\Delta\overline{F_0}$s, resulting in a total of 14 experimental conditions.

4.2.3 F_0 processing of the sentences

A specific R condition can be generated with infinite combinations of target and masker $\sigma(F_0)$s. Since the purpose of this study was to explore the effects of all these three variables (R, $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$), it was decided to generate the R conditions by varying $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ over a range of values. To represent the natural range of $\sigma(F_0)$ across the sentences and the increase of the lower and upper limits of such range with increasing $\overline{F_0}^{talker}$, the F_0-trajectory manipulations were limited such that all processed sentences had a $\sigma(F_0)$ within a given interval that was talker dependent and that reflected the talker's original range of $\sigma(F_0)$. These intervals were obtained as follows: linear fits of the lower and upper limits of $\sigma(F_0)$ for each talker ($\sigma_{min}(F_0)$ and $\sigma_{max}(F_0)$ shown in Table 4.1) were computed to obtain the overall lower and upper boundaries of the

Table 4.1: F_0 statistics of each talker in the HINT corpus. For each talker, the table shows the median F_0 computed over the entire HINT speech corpus ($\overline{F_0}^{talker}$) together with the lower and upper limits of the F_0 dynamic range values of the individual-sentences ($\sigma_{min}(F_0)$ and $\sigma_{max}(F_0)$, respectively) and the lower and upper limits imposed on the F_0 dynamic range of the manipulated stimuli ($\sigma^*_{min}(F_0)$ and $\sigma^*_{max}(F_0)$, respectively). The values in bold indicate the overall minimum and maximum values used for stimulus manipulations.

Talker ID	$\overline{F_0}^{talker}$ [Hz]	$\sigma_{min}(F_0)$ [Hz]	$\sigma_{max}(F_0)$	$\sigma^*_{min}(F_0)$ [Hz]	$\sigma^*_{max}(F_0)$ [Hz]
F1	212	7.3	46.7	6.9	**44.2**
F2	200	5.6	40.6	6.5	42.0
F3	199	4.9	36.1	6.5	41.8
F4	170	6.3	42.7	5.6	36.4
F5	176	8.1	49.8	5.8	37.6
F6	196	4.0	31.4	6.4	41.2
M1	107	1.7	24.0	3.8	24.8
M2	158	9.1	32.5	5.3	34.2
M3	120	4.3	34.4	4.2	27.3
M4	105	3.8	19.5	3.7	24.5
M5	96	3.2	20.6	**3.5**	22.8
M6	118	3.9	25.3	4.1	26.9

$\left(\overline{F_0}, \sigma(F_0)\right)$ distribution over the HINT sentences, indicated as straight dashed lines in Figure 4.1. For each of the 12 talkers, the lower and upper processing limits of the sentences, expressed in terms of maximal possible compression and expansion of the $\sigma(F_0)$ and indicated as $\sigma^*_{min}(F_0)$ and $\sigma^*_{max}(F_0)$, were computed as the intercepts of these lines with the respective $\overline{F_0}^{talker}$ and are also reported in Table 4.1.

The experimental stimuli can be represented in a 2-dimensional cartesian space, in terms of the experimental variables R, $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$. As illustrated in Figure 4.2 (left panel), each pair of sentences can be described by a cartesian pair $(x, y) = \left(\sigma(F_0^{target}), \sigma(F_0^{masker})\right)$. Pairs with a given R value fall along straight lines (iso-R lines) passing through the origin of the plane. For each talker, the stimulus space is bounded along both axes by the talker-dependent F_0-processing limits ($\sigma^*_{min}(F_0)$ and $\sigma^*_{max}(F_0)$). The red dashed square represents the overall boundaries of the stimulus space, resulting from the F_0-processing limits imposed on each talker. The figure also shows the probability of occurrence (indicated by the color scale) of each $\left(\sigma(F_0^{target}), \sigma(F_0^{masker})\right)$ combination for the sentence pairs used in the experiment.

For each pair of sentences, the desired combination of $\Delta\overline{F_0}$ and F_0-dynamic-range contrast was generated by compressing or expanding the original F_0 trajectories of the sentences to obtain the desired F_0 dynamic range values ($\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$), and subsequently shifting them to the desired median F_0 values

($\overline{F_0}^{target}$ and $\overline{F_0}^{masker}$) that were $\Delta\overline{F_0}$ apart. The specific F_0 dynamic range values of the sentences in a pair were assigned in the following way. For a given R value, a random point was drawn from a uniform distribution along the intersection between the iso-R line and the talker's stimulus boundary. The corresponding cartesian coordinates were used as $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$. As a consequence of this experimental design, the stimuli were not uniformly distributed along iso-R lines in the stimulus space (see color scale in Figure 4.2, left panel). In fact, the contribution of different talkers to the stimulus space was limited to certain regions of the space that reflected the different manipulation limits imposed on the sentences spoken by each talker (see Table 4.1): while sentences with low $\sigma(F_0)$ could be generated for all talkers, the highest $\sigma(F_0)$ values were used only for sentences spoken by talkers with the highest $\overline{F_0}^{talker}$.

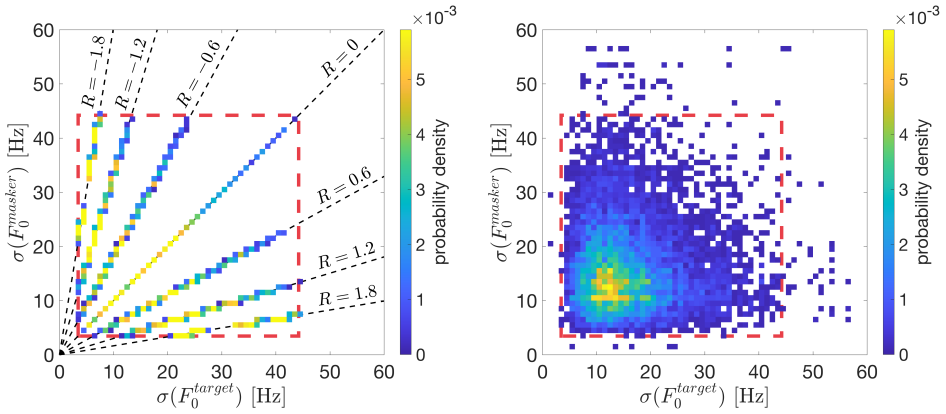


Figure 4.2: Two-dimensional representation of the stimulus space, where each pair of sentences can be represented as a point described by the cartesian pair $(x, y) = \left(\sigma(F_0^{target}), \sigma(F_0^{masker})\right)$. Left panel: probability density histogram of $\left(\sigma(F_0^{target}), \sigma(F_0^{masker})\right)$ pairs generated for the experimental stimuli. The black dashed lines passing through the origin of the space represent pairs with a specific R value (iso-R lines). Right panel: probability density histogram of $\left(\sigma(F_0^{target}), \sigma(F_0^{masker})\right)$ pairs occurring with original HINT sentences, unmodified in their F_0 trajectories, with different target and masker talkers in each pair. In both panels, the red dashed square represents the space boundaries, defined by the $\sigma(F_0)$ manipulation limit imposed on the experimental stimuli to reflect the $\sigma(F_0)$ variability found in the original HINT sentences of each talker. In both panels, the histogram bins are 1-Hz wide.

The two $\Delta\overline{F_0}$ conditions were generated by assigning $\overline{F_0}$ values to the sentences as follows. For $\Delta\overline{F_0} = 0$ semitones, both sentences had $\overline{F_0} = \overline{F_0}^{talker}$. For $\Delta\overline{F_0} = 6$ semitones, the average F_0 separation was split across the two sentences: one F_0 trajectory was shifted upward and the other was shifted downward, by a number of semitones relative to $\overline{F_0}^{talker}$ of the talker used in the pair. In order to

avoid unnatural F_0 values, a larger F_0 shift of four semitones was applied towards lower frequencies when $\overline{F_0}^{talker}$ was higher than the average F_0 computed across all talkers in the HINT corpus, and towards higher frequencies otherwise. The other sentence was shifted in the opposite direction by two semitones, with respect to $\overline{F_0}^{talker}$.

The F_0 manipulation was done in Matlab by applying the following formula to the original F_0 trajectories, previously extracted with PRAAT: $F_0^* = s(F_0 - \overline{F_0}) + \overline{F_0}^*$, where F_0^* is the desired F_0 trajectory, $s = \frac{\sigma^*(F_0)}{\sigma(F_0)}$ is the multiplication factor for expanding ($s > 1$) or compressing ($s < 1$) the trajectory to the desired F_0 dynamic range, F_0 the original trajectory, $\overline{F_0}$ the original F_0 median and $\overline{F_0}^*$ the desired F_0 median.

4.2.4 Procedure and apparatus

The competing-voices test (CVT) framework developed by Bramsløw et al. (2019) was used to conduct the experiment. In the experimental procedure, the target sentence was pre-cued to the listener by presenting its first word on a screen before the stimulus playback. After the mixture of two sentences was played back, the listeners were asked to repeat all the words they believed belonged to the target sentence. Target and masker sentences were aligned at their onsets. Each of the 14 testing conditions was tested with 20 pairs of sentences. In each pair, the target was randomly assigned to either the sentence with the higher or the lower F_0. To avoid any effect of presentation order or sentence repetition in the group results, a Latin square design was used to balance the test conditions ($\Delta\overline{F_0}$ and R) across listeners, while sentence-list and talker were randomized across conditions.

Based on the results from the study presented in Chapter 3, where the same speech material was used for a competing-speech task, the sentences were mixed at a TMR of -4 dB chosen to avoid ceiling and flooring effects (which would occur at higher and lower TMRs, respectively). The target sentence was presented at an average sound pressure level (SPL) of 65 dB, randomly roved over a ±5 dB range. The stimuli were presented diotically over headphones, which were free-field equalized to the entrance of the ear canal. The level of the masker sentence was adjusted according to the desired TMR. The experimental procedure (level setting, sentence mixing and stimulus playback) was controlled through a Matlab script on an Apple computer. The stimuli were presented at a sampling rate of 16 kHz using a Fireface UCX soundcard and Sennheiser

HDA-200 headphones, while the listener was seated in a sound-proof booth. All listener's responses were scored by the same native Danish audiologist. Speech intelligibility was measured for each sentence pair as the percentage of correctly repeated target words (excluding the initial cue word). Since the difficulty of the task could vary across sentence pairs depending on the specific combination of sentences (for example depending on the presence of linguistic context in the target sentence or syntax similarity between the two competing sentences), the performance was averaged across the 20 sentence pairs presented in each experimental condition.

4.2.5 Assessment of real-life fidelity of the stimulus space

To assess how faithfully the experimental stimulus design represented the variability of F_0 dynamics found in the unmodified speech material (which, given the number of talkers and variety of voices, was considered a good representation of real-life speech), a 'reference' stimulus space was generated by creating 10.000 random pairs of HINT sentences. The unmodified sentences were paired with different target and masker talkers without any control/manipulation of their F_0-dynamic-range nor of their F_0-dynamic-range contrast. This stimulus space was considered to provide a reference of the realistic probability of occurrence of target and masker $\sigma(F_0)$ combinations that can be obtained with HINT sentences with the 12 available talkers. Figure 4.2 (right panel) shows the generated reference stimulus space, together with the probability of occurrence of $\left(\sigma(F_0{}^{target}), \sigma(F_0{}^{masker})\right)$ pairs obtained with unmodified HINT sentences, indicated by the color scale.

When comparing the left and right panels of Figure 4.2, it can be seen that the $\left(\sigma(F_0{}^{target}), \sigma(F_0{}^{masker})\right)$ pairs and the R values used in the experiment reflected the range of values found in realistic speech. However, the distribution of the values in the two spaces differed, since pairs of the unmodified sentences showed a higher probability of occurrence in the bottom-left region of the space than the manipulated sentence pairs used in the experiment. For this reason, an additional analysis of speech intelligibility was conducted on a subset of the stimuli that was considered the most realistically relevant and defined as the region of highest probability of occurrence in the reference stimulus space. This stimulus subset was extracted as follows: from the reference stimulus space, probability distributions of occurrence of $\left(\sigma(F_0{}^{target}), \sigma(F_0{}^{masker})\right)$ pairs were obtained along each iso-R line used in the experiment. From such distributions,

the upper and lower 10%-tails were removed and the remaining portions of distributions were used to obtain an estimate of the region where 80% of the realistic stimuli with a given R value would occur. The analysis of speech intelligibility was restricted to the scores obtained from the experimental stimuli falling within these portions of iso-R lines and was compared to the analysis conducted over the entire stimulus space.

4.2.6 Measures of target and masker F_0 dynamic ranges effects

Speech intelligibility was also analyzed as a function of the F_0 dynamic range of the individual sentences in a pair (target and masker). Since $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ were not fixed to specific values in the experiment but randomly varied over a range of values, this analysis was conducted as follows. The experimental stimuli were grouped by binning the $\sigma(F_0^{target})$ values into 5-Hz wide intervals, spanning the range of $\sigma(F_0)$ values used across all sentences and talkers in the experimental stimuli (from 3.5 Hz to 44.2 Hz, see Table 4.1). Average speech intelligibility and standard errors were computed over the stimuli falling in each interval and were related to the corresponding $\sigma(F_0^{target})$ values. This was done separately for the each R and $\Delta\overline{F_0}$ conditions. The same analysis was conducted on speech intelligibility as a function of $\sigma(F_0^{masker})$. Because of the non-uniform distribution of the stimuli along each iso-R line (see Section 4.2.3), the histogram bins were populated differently depending on the $\sigma(F_0)$ values they covered. To avoid excessively large error bars in the less populated bins, bins with standard errors larger than 10 percentage points were excluded from the analysis.

4.2.7 Statistical analysis

A mixed-effects analysis of variance (ANOVA) and post-hoc tests were performed on the speech intelligibility scores that were averaged over the 20 pairs in each condition and transformed to rationalized arcsine units (RAU). The ANOVA included the listeners as a random factor, $\Delta\overline{F_0}$ and R as fixed factors, as well as two-way interactions between all main factors.

Additionally, a linear regression analysis was performed. A mixed-effects regression model was obtained using a stepwise backward selection procedure, starting with an initial model that included $\Delta\overline{F_0}$, R, $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ as fixed effects and all pairwise interactions between them. The listener was in-

cluded as a random effect. As the effects of $\overline{F_0}^{target}$ and $\overline{F_0}^{masker}$ were not explicitly investigated in this study these factors were not included in the model, despite their contribution to improving the model's predictive power, in a trade-off between predictive power and interpretability. In the model-selection procedure, likelihood ratio tests were used for model comparisons and non-significant effects were sequentially removed until all remaining effects were significant. The analysis was performed using the software R (R Core Team, 2021) and the `lme4` package (Bates et al., 2014).

The ANOVA and the regression model offer different analyses since the ANOVA considered P_C averaged across the 20 pairs in each experimental condition (therefore allowing an analysis of the effects of the main experimental variables listener, $\Delta\overline{F_0}$ and R), whereas the regression analysis considered P_C at the level of each individual sentence pair presented to the listener, allowing to include the $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ variables, which varied randomly from pair to pair.

4.3 Results

4.3.1 Measured speech intelligibility scores

A boxplot of the speech intelligibility scores measured for each listener (1-19) is shown in Figure 4.3, with speech intelligibility indicated as the percentage of correctly recognized words from the target sentence (P_C), averaged over the 20 sentence pairs presented in each experimental condition. Speech intelligibility varied substantially across listeners, with median values ranging from 38% to 87%. Speech intelligibility also varied across conditions for each listener, with some listeners showing a very large performance variability (e.g., listeners 1 and 11, with P_C ranging from about 20% to 80%) and other listeners showing a rather small variability (e.g., listeners 2, 5 and 13, who had P_C varying over a range of about 30 percentage points).

The ANOVA analysis conducted on the speech intelligibility scores and main experimental variables (listener, TMR, $\Delta\overline{F_0}$ and R) revealed that only the variables listener and $\Delta\overline{F_0}$ had a significant effect on speech intelligibility ($p < 10^{-4}$ and $p < 10^{-5}$, respectively).

Figure 4.4 (filled symbols) shows speech intelligibility, averaged across listeners, as a function of R, for the two $\Delta\overline{F_0}$ conditions of 0 semitones (blue circles)

and 6 semitones (red squares). Overall, no significant effect of R was observed, in either of the $\Delta\overline{F_0}$ conditions tested. When averaged across R conditions, speech intelligibility was found to be 10 percentage points higher for $\Delta\overline{F_0} = 6$ semitones than for $\Delta\overline{F_0} = 0$ semitones, and this difference was significant at the $p < 10^{-5}$ level.

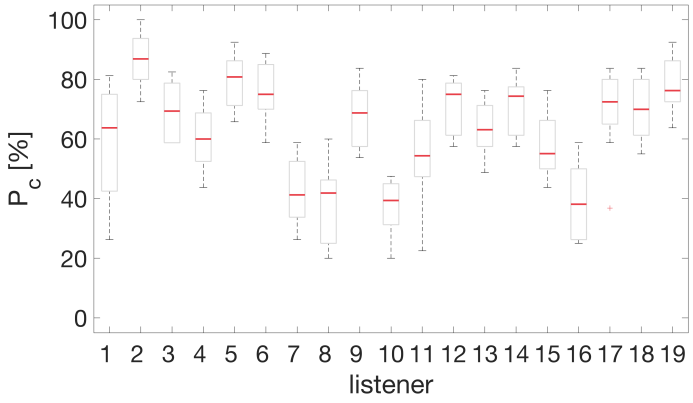


Figure 4.3: Boxplot of speech intelligibility scores for each of the 19 listeners, averaged across the 20 sentence pairs within each experimental condition. The red horizontal lines indicate median values, the boxes indicate the range of data between the 25th and the 75th percentiles, the whiskers indicate the utmost non-outlier values and the red cross indicates an outlier (defined as a value that is more than 1.5 times the interquartile range above the upper or lower edge of the box).

4.3.2 Analysis of speech intelligibility scores on a stimulus subspace

Figure 4.4 (open symbols) shows the average speech intelligibility scores obtained when only data belonging to a subset of the stimulus space were considered, corresponding to the region with highest frequency of occurrence for pairs of unmodified HINT sentences (see Section 4.2.5). When the analysis was restricted to this stimulus subset, P_C at $R = -1.8$ and $\Delta\overline{F_0} = 0$ semitones was significantly higher than in the other R conditions measured for the same $\Delta\overline{F_0}$ ($p < 0.02$). None of the other conditions showed a statistically significant effect of R nor indicated a speech intelligibility score that differed significantly from the one measured over the entire dataset.

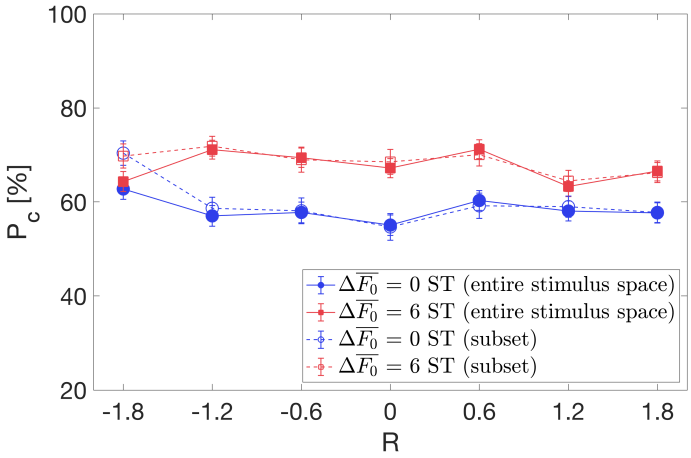


Figure 4.4: Speech intelligibility scores, representing the proportion of correctly recognized words from the target sentence, as a function of F_0-dynamic-range contrast (R), averaged across listeners. The blue circles represent the results for $\Delta\overline{F_0} = 0$ semitones and the red squares for the $\Delta\overline{F_0} = 6$ semitones. The filled symbols are the speech intelligibility scores calculated using the data from the entire stimulus space. The open symbols represent scores calculated using the data from the subset of stimulus space that showed the highest frequency of occurrence for HINT pairs with unmodified F_0 trajectories. The error bars represent standard errors.

4.3.3 Effects of target and masker F_0 dynamic range

Figure 4.5 shows the analysis of speech intelligibility data as a function of $\sigma(F_0)$ of the individual sentences in a pair. Each panel represents the analysis for a specific R value (between $R = -1.8$ and $R = 1.8$). For $R \geq 0$, speech intelligibility is shown as a function of $\sigma(F_0^{target})$ as for these R values $\sigma(F_0^{masker})$ only varied over a very narrow frequency range (of a few Hz in the case of $R = 1.8$). Accordingly, for $R \leq 0$, speech intelligibility is shown as a function of $\sigma(F_0^{masker})$ because of the very limited range of variation of $\sigma(F_0^{target})$ for these R values. The blue circles show the results for $\Delta\overline{F_0} = 0$ semitones and the red squares show the results for $\Delta\overline{F_0} = 6$ semitones. The open symbols indicate the P_C values averaged across the $\sigma(F_0)$ values shown in each panel.

At certain R values, a substantial increase in speech intelligibility was observed over the range of target or masker $\sigma(F_0)$ used in the experimental stimuli, especially when $\Delta\overline{F_0} = 0$ semitones. Examples (all relating to the $\Delta\overline{F_0} = 0$ semitones condition, shown in blue in Figure 4.5) are $R = 0$ (top panel of the figure), $R = 0.6$ (right panel of second row), or $R = -1.2$ (left panel of third row). In all these cases, speech intelligibility increased for more than 30 percentage points across the range of $\sigma(F_0^{target})$ or $\sigma(F_0^{masker})$ used in the experiment. Similar ef-

fects were also observed for $\Delta\overline{F_0} = 6$ semitones in certain fixed R conditions, but with more moderate increases (see left panel of third row, the increase in P_C as a function of $\sigma(F_0^{masker})$). However, these increments were not observed at every R and $\Delta\overline{F_0}$ (see for example the right panel in the third row, with $R = 1.2$ and $\Delta\overline{F_0} = 0$ semitones).

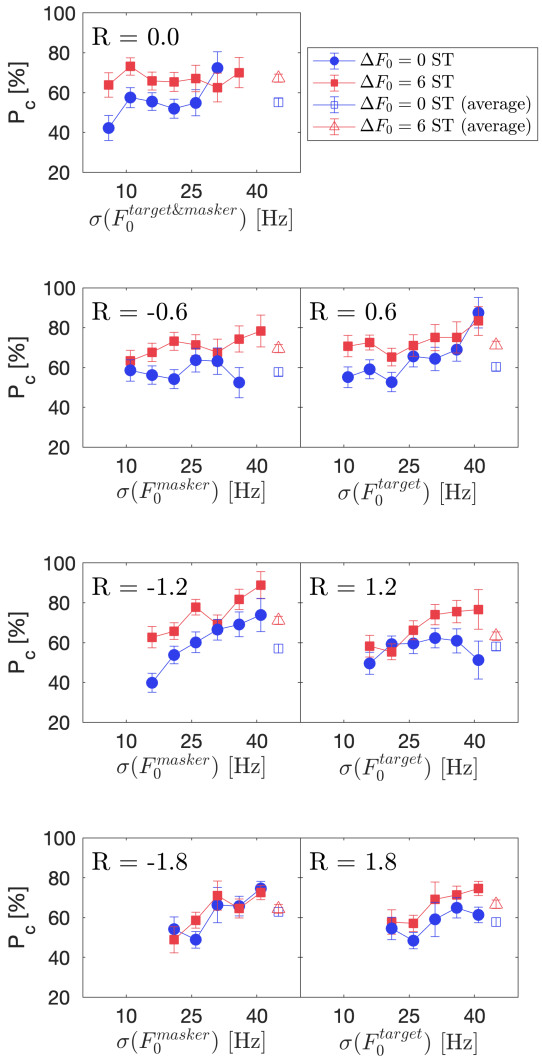


Figure 4.5: Speech intelligibility as a function of single-sentence $\sigma(F_0)$ (target or masker), for different R values. For $R \geq 0$, speech intelligibility is shown as a function of $\sigma(F_0^{target})$, while for $R \leq 0$ speech intelligibility is shown as a function of masker $\sigma(F_0^{masker})$. The blue and red curves represent the data for the $\Delta\overline{F_0}$ of 0 and 6 semitones, respectively. The open symbols represent the average values calculated across all $\sigma(F_0)$ values. The error bars represent standard errors.

Figure 4.6 (left panel) illustrates the interaction of the effects of R, $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$. The figure shows the stimulus space divided into nine squared regions, by selecting equally-spaced intervals on both axes. Average speech intelligibility was calculated over the stimuli belonging to each region. For each region, the figure shows the average speech intelligibility change calculated with respect to the region where both target and masker had relatively low F_0 fluctuations, i.e., the bottom-left corner region (baseline condition), with lighter gray tones indicating improvements in speech intelligibility and darker gray tones indicating reductions in speech intelligibility. This analysis was only conducted for the $\Delta\overline{F_0} = 0$ semitones condition for illustration. In this representation, target and masker sentences are essentially classified as having low, medium and high F_0 fluctuations, a distinction similar to that of the 'flat', 'normal' and 'exaggerated' speaking styles used by Calandruccio et al. (2019). The region with the lowest average speech intelligibility score (52%) was the one corresponding to the baseline condition. Speech intelligibility improvements with respect to this region were observed in correspondence of an increase of either $\sigma(F_0^{target})$, $\sigma(F_0^{masker})$ or both. The upper-right region, representing stimuli with a high amount of F_0 fluctuations both in the target and the masker

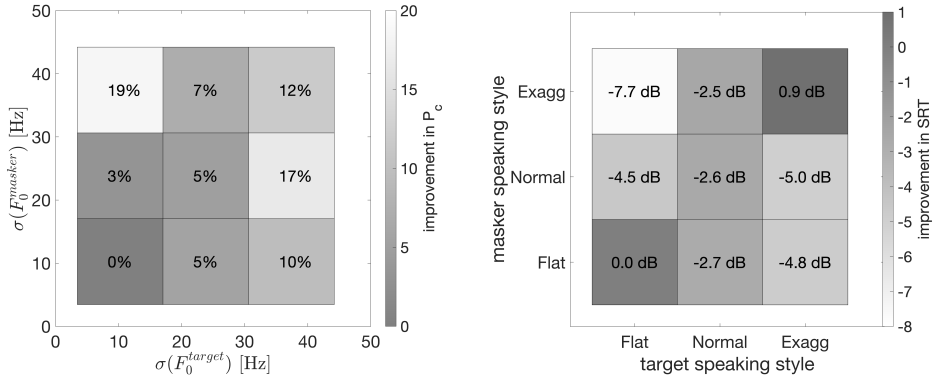


Figure 4.6: Left panel: Analysis of average change in speech intelligibility score for different regions of the stimulus space, in the $\Delta\overline{F_0} = 0$ semitones condition. The changes are calculated with respect to the average score in the region with lowest $\sigma(F_0)$ for both target and masker (bottom-left region), with lighter gray tones indicating improvements in speech intelligibility and darker tones indicating reductions in speech intelligibility. Right panel: corresponding analysis conducted on the data from Calandruccio et al. (2019), represented in a stimulus space divided in terms of the speaking style and their combinations. Improvements in speech intelligibility are expressed as SRT difference in dB with respect to the region with flat target and flat masker (bottom-left region).

F_0 trajectories (i.e., with $R = 0$), showed speech intelligibility improvements that were larger than those found for some of the largest R values tested, where high-fluctuation targets were paired with low-fluctuation maskers (bottom-right region of the stimulus space).

The right panel of Figure 4.6 shows a similar analysis conducted on the stimuli and the speech intelligibility data from Calandruccio et al. (2019). In their results, speech intelligibility was also lowest when both target and masker had small F_0 fluctuations (their 'flat-versus-flat' condition) and improved when target or masker increased in intonation. However, speech intelligibility remained low in the conditions with matched speaking styles, even when both target and masker had strong F_0 fluctuations (i.e., their 'exaggerated-versus-exaggerated' condition).

4.3.4 Regression analysis

Table 4.2 shows the results of the mixed linear model applied to the speech intelligibility scores (in words correct) for each individual sentence pair. The estimates represent unstandardized mean differences, calculated with respect to the baseline condition of $R = -1.8$ and $\Delta\overline{F_0} = 0$ semitones. The regression model resulting from the stepwise selection procedure described in Section 4.2.7, was defined as follows:

$$P_C \sim \Delta\overline{F_0} + \sigma(F_0^{target}) + R + \Delta\overline{F_0} * R + \sigma(F_0^{target}) * R + (1\,|\,\text{listener})^{\text{b}}.$$

The model showed significant differences across R for $R = -0.6$ ($p < 0.01$), $R = 0$ ($p < 0.001$), $R = 0.6$ ($p < 0.001$) and $R = 1.2$ ($p < 0.05$) but no significant differences for $R = -1.2$ ($p = 0.404$) and $R = 1.8$ ($p = 0.378$). A main effect of $\Delta\overline{F_0}$ was found to be non-significant ($p = 0.631$), whereas the interaction between $\Delta\overline{F_0}$ and R was significant ($p < 0.05$), indicating differences in P_C

[b] Note that the final statistical model did not include any effects of $\sigma(F_0^{masker})$, as the variable and its interactions were dropped during model selection due to their non-significance. Given the results showed in Figure 4.5, this is likely due to the strong overlap in the variance explained by $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$. Indeed, a statistical model that included $\sigma(F_0^{masker})$ but not $\sigma(F_0^{target})$, revealed similar trends for the two variables (i.e., a significant main effect of $\sigma(F_0^{masker})$ and an interaction between R and $\sigma(F_0^{masker})$). However, the $\sigma(F_0^{masker})$-based model showed less overall predictive power as measured using Akaike's Information Criterion Akaike, 1998 and thus was rejected.

for the two $\Delta\overline{F_0}$ conditions for R values of −1.2, −0.6, 0 (all $p < 0.01$), and 0.6 ($p < 0.05$). Additionally, the model revealed a significant ($p < 0.001$) main effect of $\sigma(F_0^{target})$ indicating that overall higher P_C values were measured for higher $\sigma(F_0^{target})$. Additionally, a significant ($p < 0.001$) negative interaction between R and $\sigma(F_0^{target})$ was found for all R values, indicating different rates of P_C changes with $\sigma(F_0^{target})$ for different R values.

Table 4.2: Outputs for the mixed-effects model on the speech intelligibility scores. The model formula follows: $P_C \sim \Delta\overline{F_0} + \sigma(F_0^{target}) + R + \Delta\overline{F_0} * R + \sigma(F_0^{target}) * R + (1|\text{listener})$. The interaction between $\Delta\overline{F_0}$ and $\sigma(F_0^{target})$ was not included in the final model after model selection due to lack of significance. Thus, estimates for the $R * \sigma(F_0^{target})$ interaction are independent of $\Delta\overline{F_0}$ and only reported once (redundant values are shown as '–'). The significance levels are indicated as $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	$\Delta\overline{F_0} = 0$				$\Delta\overline{F_0} = 6$			
	Estimate	Std error	t	p	Estimate	Std error	t	p
R								
Baseline condition ($R = -1.8$, $\Delta\overline{F_0} = 0$)	0.215	0.071			0.013	0.028	0.479	0.631
$R = -1.2$	0.068	0.082	0.834	0.404	0.128	0.039	3.266	**
$R = -0.6$	0.303	0.075	4.034	**	0.102	0.039	2.590	**
$R = 0$	0.293	0.073	4.014	***	0.108	0.039	2.750	**
$R = 0.6$	0.286	0.076	3.758	***	0.099	0.039	2.524	*
$R = 1.2$	0.173	0.082	2.097	*	0.045	0.039	1.151	0.249
$R = 1.8$	0.079	0.090	0.882	0.378	0.074	0.039	1.894	0.582
$\sigma(F_0^{target}) * R$								
Baseline condition ($R = -1.8$)	0.077	0.011	6.780	***	-	-	-	-
$R = -1.2$	-0.039	0.013	-3.030	**	-	-	-	-
$R = -0.6$	-0.071	0.012	-6.130	***	-	-	-	-
$R = 0$	-0.074	0.011	-6.515	***	-	-	-	-
$R = 0.6$	-0.072	0.011	-6.297	***	-	-	-	-
$R = 1.2$	-0.069	0.011	-6.034	***	-	-	-	-
$R = 1.8$	-0.068	0.011	-5.932	***	-	-	-	-

4.4 Discussion

4.4.1 Summary of main results

The present study investigated the effect of the F_0 dynamic range contrast (R) on the intelligibility of competing sentences, using speech stimuli that offered a variety of voices and levels of F_0 dynamics. The F_0 information of competing

sentences was manipulated to generate several degrees of F_0-dynamic-range contrasts, obtained with varying levels of F_0 dynamic ranges of the individual sentence ($\sigma(F_0)$), in presence and in absence of an average F_0 separation ($\Delta\overline{F_0}$). Overall, only moderate speech intelligibility improvements (within 10 percentage points) were found in presence of an F_0 dynamic range contrast between competing sentences, compared to when no contrast was present (see Figure 4.4). This limited effect did not change when the data analysis was restricted to a region of the stimulus space characterized by the most frequent F_0 dynamics that were found in the unmodified speech material. Only for certain R conditions, the $\Delta\overline{F_0}$ between the competing sentences was found to have an effect on target speech intelligibility. The magnitude of the effect of $\Delta\overline{F_0}$ was consistent with the findings from the study described in Chapter 3, where a similar experimental paradigm was employed to measure the effect of various levels of $\Delta\overline{F_0}$ on the intelligibility of competing sentences. The $\sigma(F_0)$ of the individual sentences in a pair was found to have strong effects on the target speech intelligibility (see Figure 4.5 and left panel of Figure 4.6). These results demonstrated that the presence of relatively large F_0 dynamics in at least one of the two competing sentences is sufficient for good speech intelligibility, even in absence of an average F_0 separation.

4.4.2 Interaction between $\Delta\overline{F_0}$ and R

The present study explored the interaction between $\Delta\overline{F_0}$ and R, which was found to be a predictor of speech intelligibility (see Section 4.3.4). In fact, the largest effect of $\Delta\overline{F_0}$ was found for the central values of R, between -1.2 and 0.6, whereas the difference in speech intelligibility between $\Delta\overline{F_0}$ conditions was reduced or absent for the most extreme R conditions (see Figure 4.4). An explanation of this result may be that, e.g., in the case of $R = 0$ (i.e., in the absence of a dynamic range contrast) and when the F_0 fluctuations of the competing sentences are sufficiently low, the separation in F_0 is constantly present along the entire duration of the stimulus, not only in terms of a time average. On the contrary, when at least one of the two sentences contains sufficiently strong F_0 dynamics, the instantaneous difference in F_0 along the trajectories can largely differ from its long-term average, which therefore becomes less relevant for speech segregation. Not surprisingly, the largest effects of $\Delta\overline{F_0}$ measured in previous studies were found when the competing voices had F_0 trajectories that were monotonous or had limited F_0 dynamics (Assmann, 1999; Darwin et al.,

2003). It is therefore likely that, in presence of relatively strong F_0 fluctuations (which occur rather often in unmodified HINT sentences, as shown in Figure 4.1 and in the right panel of Figure 4.2), the $\Delta\overline{F_0}$ is a descriptor that does not capture the characteristics of the F_0 trajectories that are relevant for the intelligibility of realistic speech.

4.4.3 Effects of the individual sentence's F_0 dynamic range

Overall, the F_0 dynamic range of the individual sentences in a pair had a substantial effect on the intelligibility of target speech. In the regression analysis (see Section 4.3.4), both $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ were found to be predictors of target-speech intelligibility. The effects of the individual sentences' F_0 dynamic range differed slightly in each R condition, as was shown by the interaction between $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ with R in the regression analysis (see also Figure 4.5).

The most difficult listening condition with the lowest measured speech intelligibility scores was the one where both sentences had low F_0 fluctuations (i.e., the lower-left corner of the two-dimensional stimulus space shown in Figure 4.6, left panel). The introduction of F_0 variations in at least one of the two competing sentences was sufficient for improving target-speech intelligibility from this baseline condition, regardless of the F_0 dynamic range contrast between the sentences. These results thus suggest that the F_0 dynamic range contrast might not be the correct metric for revealing the effects of F_0 dynamics on the intelligibility of competing voices. Instead, it seems that the dissimilarity between the F_0 trajectories, rather than their contrast in dynamic range, can be utilized by the listeners to segregate the competing speech signals. Therefore, the following new hypothesis can be formulated: if the two F_0 trajectories are relatively flat, they are hardly distinguishable (unless a $\Delta\overline{F_0}$ is present); when a certain level of F_0 fluctuations is introduced (even in absence of a $\Delta\overline{F_0}$), the dissimilarity in the time-frequency pattern of the F_0 becomes sufficiently salient for the listeners to discriminate the F_0 trajectories and better segregate the target speech. To verify this theory, further research is necessary to assess if the dissimilarity between competing F_0 trajectories can influence their discrimination, what levels of F_0 dynamics in the competing signals are necessary for this task and how the ability of the listeners to discriminate the F_0 trajectories is related to speech intelligibility.

4.4.4 Comparison with previous findings

The results from a previous study by Calandruccio et al. (2019), who found a significant effect of the F_0-dynamic-range contrast on the intelligibility of competing voices, could not be replicated in the present study. The different outcomes might have been caused by differences in the employed stimuli and experimental paradigms, which are discussed in the following.

First, the different levels of F_0 dynamics in the stimuli employed by Calandruccio et al. (2019) were obtained by instructing the talkers to modify their speaking style to produce a monotonous, normal or exaggerated intonation. An analysis of the experimental stimuli used by Calandruccio et al. (2019) is provided in appendix to this chapter. The analysis showed that the method used by Calandruccio et al. (2019) produced F_0 dynamics that, when measured as median absolute deviations ($\sigma(F_0)$), were beyond the range of values used in the present experiment: their flat speech showed $\sigma(F_0)$s that were lower than the minimum value used in the present experiment, while their exaggerated speech showed $\sigma(F_0)$s well above the highest value used in the present experimental stimuli. As a consequence, Calandruccio et al. (2019) may have generated stronger F_0-dynamic range contrasts that exceeded the range tested in the present experiment. The analysis of the stimuli by Calandruccio et al. (2019) also showed that the different speaking styles could not always be discriminated in terms of their F_0 dynamic range. As a consequence, it was not possible to establish a one-to-one relationship between a contrast in speaking style and a numerically-determined F_0 dynamic range contrast. Furthermore, the modifications of the speaking style may have influenced other acoustic and prosodic features of speech beyond the F_0, which may have an impact on speech intelligibility, such as the magnitude of variations in intensity (i.e., amplitude modulations), changes in formant structure or spectral properties of the voice. For example, when flattening or exaggerating the intonation of a sentence, it is likely that, together with F_0 dynamics, the amplitude modulations in the speech signal are also reduced or enhanced, respectively, compared to the normal speaking style. Therefore, it is possible that the effect of speaking style and speaking style contrast on speech intelligibility found by Calandruccio et al. (2019) was not related only to the F_0 dynamic range and F_0-dynamic-range contrast, but potentially also to other features of speech that were altered by the speaking style concurrently with the F_0 dynamics. However, further analysis

of the stimuli by Calandruccio et al. (2019) is required to understand if such other features of speech varied substantially with speaking style and if these variations are correlated with the effect observed on speech intelligibility.

A second aspect of the experimental paradigm that might be responsible for the difference in the results between the present study and Calandruccio et al. (2019) is the number of interfering talkers. Since the masker signals used by Calandruccio et al. (2019) were randomly extracted excerpts of a two-voices stream, they did not contain meaningful linguistic and prosodic information. By contrast, the single-talker maskers used in the present study carry linguistic content that is coherent with the prosodic information encoded in the F_0 trajectory.

Other aspects of the experimental design that may have influenced the results of the different studies are related to the alignment of target and masker onsets and the way used to cue the target to the listener. In the study by Calandruccio et al. (2019), the masker started 500 ms before the target and ended 500 ms after it. It is possible that the later occurrence of the target onset provided a strong cue indicating the presence of the target voice. Especially when the target differed in F_0 dynamic range from the masker, its onset may have produced a more salient change in the properties of the stimulus signal than in the case of target and masker with matched speaking styles. Therefore, the difference in target and masker onsets may have had different effects in the different F_0-dynamic-range contrast conditions. Furthermore, in the experiment by Calandruccio et al. (2019), the listener was instructed to listen to a female voice that was the same throughout the entire experiment. The characteristics of the target voice were likely learned by the listener during the initial phase of the SRT-tracking procedure, where an SNR of 5 dB ensured that the target voice was well above the two interfering ones. In the present experiment, target and masker were aligned at their onsets, the same talker was used in both speech signals and changed at every trial. Except for the first target word (provided via visual text cue), the listener had no prior access nor time to adapt to any of the characteristics of the target or masker speech that could be exploited for speech segregation. However, the difference in overall level between target and masker (maintained constant through the entire experiment due to the fixed TMR = −4 dB), may have been 'learned' by the listener and used as a cue for identifying the target voice as the softer one. This 'level effect' has been observed in previous studies that investigated the intelligibility of competing

voices (Brungart, 2001), showing that a difference in level between target and masker sentences can be utilized by the listener to segregate the target, even at rather adverse TMRs. However, it is difficult to determine to what extent the differences in the experimental methods between the present study and Calandruccio et al. (2019) can account for their different findings.

Both the present experimental approach and the one by Calandruccio et al. (2019) may present advantages and disadvantages to the investigation of real-life competing-talker scenarios. The approach used by Calandruccio et al. (2019) has the advantage of (i) offering an assessment of auditory scenarios where the focus is on a familiar voice masked by more than one interfering voices, a situation that can be considered quite common in daily life, and (ii) avoiding potential stimulus artifacts that are usually generated by signal processing manipulations of the stimuli. However, Calandruccio et al. (2019) used a method for generating the experimental stimuli that did not allow an accurate control of the F_0 dynamics and may thus have influenced other features of speech that can affect speech intelligibility in competing-talker scenarios. Furthermore, they limited their exploration to a fixed combination of target and masker voices (all females and with similar average F_0 values as shown in Section 2.3, Figure 2.5, panel C). On the other hand, the approach used in the present study has the advantage of (i) focusing the investigation on auditory scenarios with unfamiliar voices and single-talker maskers, (ii) offering a tool to isolate the effects of F_0 dynamics on speech intelligibility, without introducing other differences between the characteristics of the competing voices and (iii) utilizing a variety of male and female voices covering a range of F_0 dynamics that may be considered a good representation of the natural variability of voices. However, the F_0 manipulation method used in the present study, has the disadvantage of potentially introducing signal-processing artifacts in the experimental stimuli that may compromise the naturalness of the speech signals. Furthermore, the experimental paradigm employed here presents some aspects that limit the realism of the auditory scene, for example, the use of the same co-located talker as target and masker in each pair of sentences.

The approaches and the results of the present study and Calandruccio et al. (2019) may be considered complementary in advancing the understanding of the role of F_0 dynamics in competing-talker scenarios, which still remains incomplete. To further extend the research on F_0 dynamics in real-life competing-talker scenarios, future studies may consider extending the approach presented

here by including certain aspects of the method employed by Calandruccio et al. (2019), such as (i) the use of continuous, running maskers, with the purpose to test the effect of target and masker onset differences and (ii) the combination of more than one masking voice to assess how the number of interfering talkers influences the intelligibility of the target speech.

4.5 Summary and conclusion

This study investigated how the F_0-dynamic-range contrast (R) between two competing sentences can affect speech intelligibility for normal-hearing listeners. The experimental method utilized sentences with meaning and syntactical coherence typical of realistic speech, spoken by a variety of talkers and with different levels of F_0 dynamics. From the obtained findings it can be concluded that: (i) the effect of R on speech intelligibility is negligible, regardless of the presence of an average F_0 separation between competing sentences; (ii) speech intelligibility is lowest when both sentences have small amounts of F_0 dynamics; and (iii) speech intelligibility is positively affected by the presence of higher levels of F_0 dynamics in at least one of the two competing sentences.

4.6 Appendix: Analysis of speech stimuli from a previous study

The experimental stimuli employed in the study described in this chapter were compared to those from Calandruccio et al. (2019) in terms of their F_0 dynamics. Calandruccio et al. (2019) used the Bamford-Kowal-Bench (BKB; Bench et al., 1979) speech material, which consists of 336 meaningful English sentences, characterized by simple syntax and grammar and with a number of words varying between three and seven (for example, “The clown had a funny face”).

In their experimental design, Calandruccio et al. (2019) used single BKB sentences spoken always by the same female talker as target speech and excerpts of a two-talker speech stream as masker. The two-talker streams were built by concatenating 50 different sentences spoken by two other female talkers. The same target-masker talker combination was used in each stimulus presentation. All sentences from each talker were recorded with three different speaking styles, obtained by instructing the talkers to speak with a flat, normal and exaggerated intonation in order to obtain three different degrees of F_0 dynamics.

A simulation of the stimuli employed by Calandruccio et al. (2019) was generated by combining the 336 BKB single talker sentences in every speaking style (1008 sentences in total) with 3-s long excerpts randomly extracted from the two-talker streams in each speaking style. The resulting set of stimuli comprised each BKB target sentence, spoken in each speaking style, combined with a masker excerpt also in each speaking style, for a total of 3024 stimuli. These stimuli were interpreted using the two-dimensional representation described in Section 4.2.3 and compared to the stimulus space used in the present experiment, shown in the left panel of Figure 4.2.

4.6.1 Results of the analysis

Figure 4.7 shows a two-dimensional representation of $\left(\sigma(F_0^{target}), \sigma(F_0^{masker})\right)$ pairs obtained in the simulation of the experimental stimuli employed by Calandruccio et al. (2019). In the figure, each of the nine combinations of target-masker speaking styles is indicated with a different color. Marginal probability density histograms of $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ also shown in correspondence of the x- and y-axis, respectively, for each speaking style (flat, normal and exaggerated), indicated by different shades of gray. The red dashed square in the two-dimensional space and the red dashed lines in the marginal histograms represent the $\sigma(F_0)$ range estimated from the HINT speech material that were used as F_0-manipulation limits in the experimental stimuli used in the present study (see Section 4.2.3). As shown by the marginal histograms for both target and masker speech, the three speaking styles were found to have increasing $\sigma(F_0)$ values for 'increasing' speaking style, i.e., from flat to exaggerated. The range of values spanned by each speaking style increased for increasing F_0 dynamics: for both target and masker speech, the flat speaking style (shown by white histograms) covered the lowest and smallest range of $\sigma(F_0)$, while the exaggerated speaking style (shown by dark gray histograms) covered the highest and widest range. The $\sigma(F_0)$ ranges of the three speaking styles were well separated, with the exception of the normal and exaggerated target sentences, as shown by the overlap of the corresponding marginal histograms, with exaggerated sentences showing $\sigma(F_0^{target})$ as low as 10 Hz. The $\sigma(F_0)$ values of some speech stimuli were found outside the range of HINT F_0 dynamics (represented by the red dashed vertical lines in Figure 4.7). This is the case of the flat target sentences, more than half of which showed $\sigma(F_0)$ below the lower boundary used in the present experiment (i.e., 3.5 Hz), and the exaggerated speech from both target

and masker, which showed $\sigma(F_0)$ well above the upper boundary used in the present experiment (i.e., 44.2 Hz) with values as high as 60 Hz.

As a consequence of the $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ distributions described above, when the target sentences and masker excerpts used by Calandruccio et al. (2019) were combined to simulate their experimental stimuli and represented as $\left(\sigma(F_0^{target}), \sigma(F_0^{masker})\right)$ pairs in the two-dimensional space of Figure 4.7, the stimuli from different speaking style combinations were grouped in clusters that

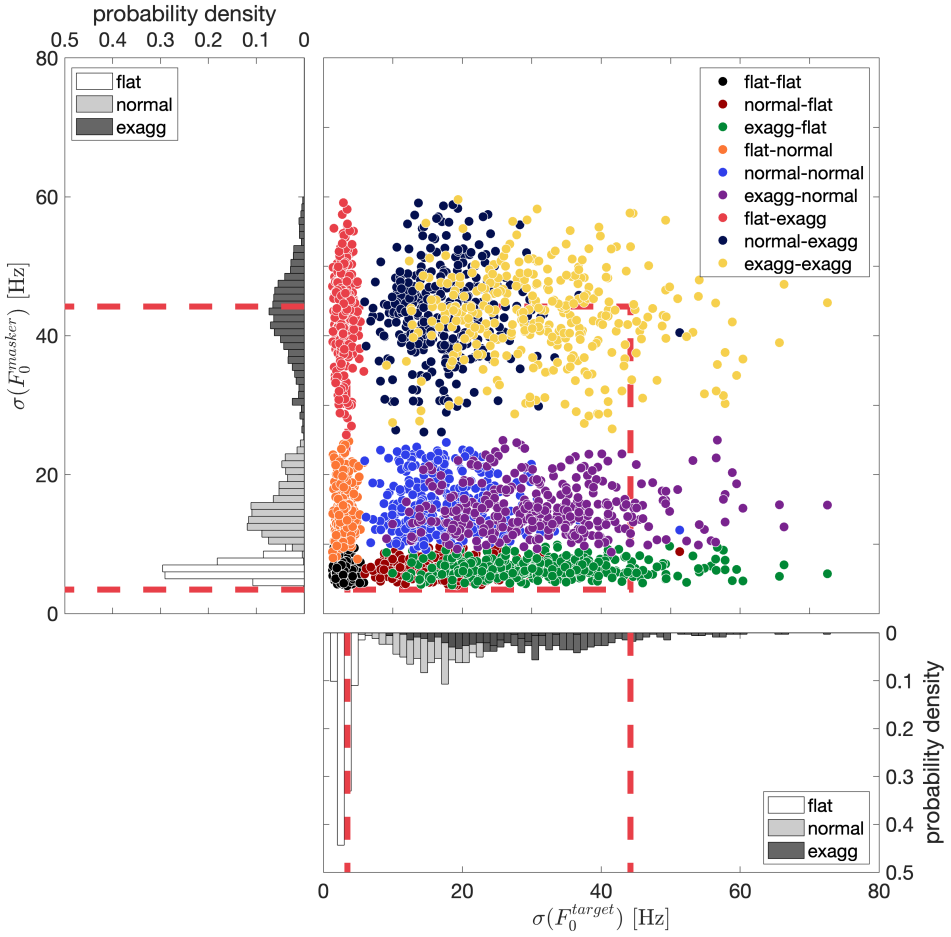


Figure 4.7: Two-dimensional representation of $\left(\sigma(F_0^{target}), \sigma(F_0^{masker})\right)$ pairs obtained from a simulation of the experimental stimuli from Calandruccio et al. (2019). Each combination of target-masker speaking styles is represented with different colors. Marginal histograms of $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ probability densities are shown on the x- and y-axis, respectively, for the three different speaking styles (flat, normal and exaggerated), indicated with different shades of gray. The histogram bins are 1-Hz wide. The red dashed square in the two-dimensional histogram and the red dashed lines in the marginal histograms represent the $\sigma(F_0)$ manipulation limit imposed on the experimental stimuli used in the present study (see Section 4.2.3).

overlapped with each other in some cases. An example is the overlap between the normal-target versus normal-masker combination (light blue points) and the exaggerated-target versus normal-masker combination (purple points). Therefore, the different speaking styles combinations could not be associated to regions of the two-dimensional space with a one-to-one correspondence.

Acknowledgments

I would like to thank Lauren Calandruccio and Peter Wasiuk from Case Western Reserve University, Cleveland, OH, USA, for sharing their experimental stimuli and results. I would also like to thank Andrea Clara Dich Jensen for her help in collecting the speech intelligibility measurements reported in this study.

5

Effects of fundamental-frequency differences between competing sentences on speech intelligibility for listeners with hearing impairment[a]

Abstract

Hearing-impaired (HI) listeners often report difficulties in understanding speech in competing-talker scenarios. Differences in fundamental frequency (F_0) between competing voices (such as differences in average F_0 and in F_0 dynamic range) have been shown to be beneficial for speech intelligibility in normal-hearing (NH) listeners but seem to yield strongly reduced benefits for HI listeners. However, the findings from previous studies may be limited to the specific, partially highly selective, experimental methods and speech materials used. The present study thus aimed at extending the available knowledge on how HI listeners can exploit F_0-related cues by employing a variety of voices with F_0 characteristics that well represent realistic speech and by manipulating the F_0 information of the stimuli without modifying other acoustic features of the speech signals. Target-speech intelligibility was measured in thirteen HI listeners using pairs of competing sentences that were manipulated in terms of their difference in average F_0, F_0 dynamic range, and level (considering target-to-masker ratios, TMRs, of 0 and 4 dB). Overall, the results confirmed that HI listeners have limited ability in exploiting F_0-related cues in competing-talker scenarios, as compared to NH listeners. In particular, it was found that (i) the separation in average F_0 was beneficial for speech intelligi-

[a] This chapter is based on Mesiano et al. (2022a), in preparation for submission to the Journal of Speech, Language, and Hearing Research.

bility, but this benefit was only present at the lowest TMR and was smaller than that previously measured in NH listeners, (ii) the difference in F_0 dynamic range between competing sentences did not affect speech intelligibility, (iii) the F_0 dynamic range of the individual sentences had no effect on speech intelligibility, in contrast to the substantial effect previously observed in NH listeners, and (iv) across the experimental variables, the TMR had the largest effect on speech intelligibility. Overall, the results of this study suggest that the challenges that HI listeners face in competing-talker scenarios may be related to their inability to resolve F_0 variations in normally produced speech.

5.1 Introduction

Complex auditory environments where several talkers speak simultaneously are pervasive in daily life and can pose a challenge for the listener. In such competing-talker scenarios, the ability to listen to a target voice out of many others is essential for social interactions and represents a highly complex auditory and cognitive task that has attracted the attention of the scientific community for decades (Bronkhorst and Plomp, 1992; Brungart et al., 2001; Cherry, 1953; Freyman et al., 2004; Kidd Jr et al., 2016; Miller, 1947; Rosen et al., 2013; Yost et al., 1996). Young listeners with a healthy auditory system have remarkable abilities when it comes to separating the voice of interest from the interfering one(s), even in the most detrimental auditory situations that may include several interfering talkers and negative target-to-masker ratios (TMRs). These listeners have access to a large set of auditory cues that support the perception of the target speech, such as energetic cues, spatial cues or cues related to the qualities of the voices in the scenario (Balakrishnan and Freyman, 2008; Brungart, 2001; Brungart et al., 2001; Darwin and Hukin, 2000a; Festen and Plomp, 1990; Freyman et al., 2001, 2004).

In contrast, hearing-impaired people report severe difficulties in attending to speech and in engaging in conversations in presence of competing-talkers, even when using hearing-loss compensation strategies (Bramsløw et al., 2018; Kochkin, 2002; Neher et al., 2007). Hearing loss and cognitive decline due to aging often limit the access to auditory cues and degrade the auditory and cognitive mechanisms that normally allow to listen to a target voice in the

presence of one or many interfering ones (Ezzatian et al., 2015; Helfer and Freyman, 2008, 2014; Koelewijn et al., 2014; Rossi-Katz and Arehart, 2009). For example, hearing loss can hinder the ability to glimpse information of the target-speech signal in the silences of the masker (Duquesnoy, 1983; Festen and Plomp, 1990; George et al., 2006; Kidd Jr et al., 2019; Summers and Molis, 2004) and can limit the speech-intelligibility benefit deriving from the spatial separation of sound sources, compared to that observed in NH listeners (Bronkhorst and Plomp, 1992; Duquesnoy, 1983; Kidd Jr et al., 2019).

Among the auditory cues that NH listeners utilize for segregating competing speech signals, the cues deriving from differences in fundamental frequency (F_0) have been shown to be rather effective. For example, when two competing voices have different average F_0s, target-speech intelligibility can be improved substantially, compared to the case when the voices have similar average F_0 (Assmann, 1999; Brokx and Nooteboom, 1982; Darwin et al., 2003; Flaherty et al., 2021; Summers and Leek, 1998). Speech-intelligibility improvements have also been observed when the F_0 variations of two competing voices differ in dynamic range (i.e., when there is a F_0-dynamic-range contrast between two competing voices), compared to when the competing voices have similar magnitude of F_0 variation (Calandruccio et al., 2019). However, as shown in the studies presented in Chapter 3 and 4, the previously reported effects of the average F_0 difference and F_0-dynamic-range contrast on speech intelligibility for NH listeners might have been enhanced by the specific choice of the speech stimuli and experimental scenarios and may not apply universally to realistic speech and auditory situations. The study reported in Chapter 3 suggests that even at negative TMRs and in absence of spatial cues, when using speech materials containing meaning, context and levels of linguistic and prosodic variability that are typical of realistic speech, the average F_0 separation has a more moderate effect than previously reported in studies that used less realistic speech materials. As for the effect of F_0-dynamic-range contrast between competing speech signals, the study presented in Chapter 4 of this thesis showed that its effects on the intelligibility of realistic sentences are negligible, at least in presence of a single interfering talker. However, in the same study, a substantial beneficial effect of the F_0 dynamic range of the individual competing sentences was found, regardless of the F_0 dynamic range contrast between the two sentences.

The role of F_0-related cues in speech intelligibility has also been investigated in hearing-impaired persons. Much attention has been focused on the effect

of the average F_0 separation between competing voices and several findings indicated that HI listeners can benefit from average F_0 separation to segregate the target speech from the mixture, but that this ability is reduced compared to what is normally observed in NH listeners (Arehart et al., 1997; Hee Lee and Humes, 2012; Mackersie et al., 2011; Summers and Leek, 1998). Several studies also investigated how HI listeners can benefit from other F_0-related cues, such as the magnitude of the F_0 dynamic range (Grant, 1987; Shen and Souza, 2017) and its contrast between competing voices (Wasiuk et al., 2020). Grant (1987) investigated the ability of young NH and young HI listeners to discriminate the frequency-variation pattern of frequency- and amplitude-modulated sinusoids as a function of the magnitude of these variations. Three of the five HI listeners they tested had performance levels comparable to the NH listeners, whereas the other two HI listeners required larger frequency variations to achieve such performance levels. Shen and Souza (2017) found that enhanced F_0 dynamics led to an increase in speech recognition in noise for older HI listeners, but the presence and magnitude of this increase varied across individuals. Wasiuk et al. (2020) investigated the effect of F_0-dynamic-range contrast on the intelligibility of speech in the presence of a two-talker masker, testing HI listeners with the experimental method and stimuli previously used for testing NH listeners in the study by Calandruccio et al. (2019). By pairing speech signals with three different speaking styles (characterized by different degrees of F_0 variations) they found that HI listeners had little benefit when the two competing speech signals differed in the magnitude of their F_0 dynamics (compared to a baseline condition with no contrast in F_0-dynamic-range), whereas a much larger benefit was observed in the NH listeners tested by Calandruccio et al. (2019).

The previous research on how the F_0 information may be used by HI listeners in competing-talker scenarios has employed experimental stimuli that may not be fully representative of realistic speech or experimental methods that did not allow an accurate control of the F_0 information in isolation from other auditory cues. For example, the competing vowels used by Arehart et al. (1997) and in one of the experiments by Summers and Leek (1998), the coordinate response measure (CRM; Bolia et al., 2000) sentences used by Hee Lee and Humes (2012) and Mackersie et al. (2011), or the monotonous sentences (i.e., with a flat F_0 trajectory) used in the sentence-recognition experiment by Summers and Leek (1998), are all highly constrained speech materials, with unrealistic time-alignment and lack of prosodic or linguistic cues such as intonation, context and meaning.

As previously described, the use of unrealistic speech materials and the removal of many auditory and linguistic cues that are usually available in the most common realistic scenarios might not be ideal for understanding the effects of the F_0-related cues on speech intelligibility, as these aspects might enhance the F_0 effects beyond their realistic importance. Meaningful realistic speech such as that used by Wasiuk et al. (2020) may be more suitable for measuring the effect of F_0-information in the presence of other cues. However, Wasiuk et al. (2020) limited their investigation to a fixed combination of talkers (a target talker and two interfering talkers, all females and with similar average F_0). While their results have been obtained with realistic speech, it is not clear if they can be generalized to the wider variety of voices and F_0 characteristics that is available in real life. Furthermore, Wasiuk et al. (2020) created the F_0-dynamic-range contrast between target and masker by combining in pairs three different levels of F_0 dynamic range that were obtained by instructing the talkers to modify their speaking style in order to obtain speech with reduced, normal or enhanced intonation (namely "flat", "normal" and "exaggerated" speaking style, respectively). Generating different levels of intonation by modifying the speaking style may not allow a precise numerical control of the F_0 dynamic range values used in the experiment (as shown in the analysis of the stimuli used by Calandruccio et al., 2019, reported in Chapter 4, Section 4.6). Also, the modifications produced by varying the speaking style may have affected other spectro-temporal features of the speech signals beyond the F_0, such as their amplitude modulations and formant structure, that might also have contributed to the observed effects of speaking-style contrast on speech intelligibility, as also mentioned in Chapter 4 in relation to the results from Calandruccio et al. (2019) on NH listeners.

The purpose of this study was to expand the available knowledge of the effects of hearing loss on the use of F_0 differences for the intelligibility of competing voices. F_0 differences were investigated in terms of average F_0 separation and F_0-dynamic-range contrast between competing sentences, employing speech recordings from several talkers and a variety of levels of F_0 dynamic range and F_0-dynamic-range contrast, generated by manipulating the speech stimuli with a numerically controlled method that targets the F_0 information without modifying other acoustic features of the speech signal. Target-speech intelligibility was measured for a cohort of older HI listeners by means of a two-competing-voices experimental paradigm, using pairs of sentences spoken by the same co-located talker, mixed at conversational TMRs. The speech-intelligibility performance

was related to the differences in F_0 (average separation and dynamic-range contrast), to the level of F_0 dynamics present in the individual sentences and to the TMR at which the competing sentences were presented. The limited ability to utilize the average F_0 difference and the F_0-dynamic-range contrast (compared to NH listeners) observed in previous studies was expected to be confirmed. The obtained results were compared to those from previous studies and discussed in relation to the properties of the experimental stimuli and methods employed.

5.2 Methods

5.2.1 Participants

Thirteen HI native Danish listeners (3 females), aged between 65 and 78 (mean 72.5) years participated in the study. All participants provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). All listeners had symmetric hearing loss with sloping pure-tone hearing thresholds, measured for

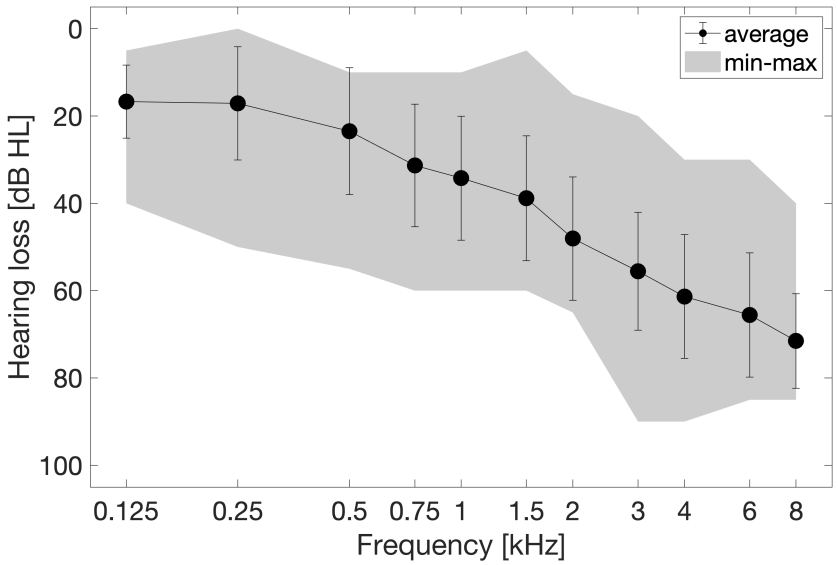


Figure 5.1: Summary of the audiometric profiles of the 13 hearing-impaired listeners that participated in the study. The black circles represent mean pure-tone thresholds averaged over left and right ears of all listeners. Error bars represent standard deviations. The dashed gray area represents the minimum and maximum audiometric thresholds observed across the group of listeners at each frequency.

audiometric frequencies between 125 Hz and 8 kHz. The average audiometric profile, with minimum and maximum hearing losses, calculated over the left- and right-ear audiogram of each listener, is shown in Figure 5.1. The listeners completed all tests in two visits that lasted no more than two hours each.

5.2.2 Speech material, F_0 processing and stimulus generation

The speech material, F_0 processing and stimulus generation were the same as used in the study described in Chapter 4. Pairs of sentences (a target and a masker) from the Danish HINT corpus (Nielsen and Dau, 2011) were manipulated in their average F_0 (quantified as F_0 median and indicated as $\overline{F_0}$) and F_0-dynamic range (quantified as F_0 median absolute deviation, MAD, and indicated as $\sigma(F_0)$) to obtain a desired average F_0 difference ($\Delta\overline{F_0} = \left|\overline{F_0}^{target} - \overline{F_0}^{masker}\right|$) and a F_0-dynamic-range contrast (quantified as the natural logarithm of the ratio between target and masker F_0-dynamic ranges and indicated as $R = \log\frac{\sigma(F_0^{target})}{\sigma(F_0^{masker})}$). The details of the F_0 manipulation method are reported in Section 4.2.3. In the present study, $\Delta\overline{F_0}$s of 0 and 6 semitones were used in combination with five R values ranging between −1.8 and 1.8 in 0.9 steps. The $\Delta\overline{F_0}$ and R conditions were combined with two target-to-masker ratios (TMRs) of 0 and 4 dB, resulting in a total of 20 experimental conditions. As in the study described in Chapter 4, in each pair, the same co-located talker was used in both target and masker sentences of a pair.

5.2.3 Procedure and apparatus

As in the experiments described in Chapters 3 and 4, the experimental design followed the competing-voices test (CVT) framework developed by Bramsløw et al. (2019). The target sentence was visually cued to the listener by providing its first word on a screen prior to the stimulus playback. The mixture of two sentences was presented to the listeners who were asked to repeat as many words as possible from the target sentence. Each of the 20 testing conditions was tested using 20 sentence pairs. To avoid any effect of presentation order or sentence repetition in the group results, the test conditions ($\Delta\overline{F_0}$, R and TMR) were balanced across listeners using a Latin square design, while sentence-list and talker were randomized across conditions. In each pair, the target was randomly assigned between the two sentences.

The target sentence was set at an average sound pressure level of 65 dB

SPL, randomly roved over a ±5 dB range, and the masker sentence level was adjusted depending on the TMR tested. The stimuli were presented diotically over headphones, which were free-field equalized to the entrance of the ear canal. Level setting and free field equalization were followed by individual ear-specific hearing loss compensation, provided by means of linear gain amplification according to the Cambridge formula (CAMEQ; Moore and Glasberg, 1998). Level adjustment, hearing-loss compensation, sentence mixing and stimulus playback were performed with Matlab on a Microsoft Windows computer. The mixture of two sentences was played back at a sampling rate of 16 kHz through a Fireface UCX soundcard and presented diotically via Sennheiser HDA-200 headphones to the listener seated in a sound-proof booth. All listener's responses were scored by the same native Danish audiologist.

Speech intelligibility was quantified as the percentage of correctly repeated words from the target sentence (excluding the initial cue word), computed over each sentence pair. Since the difficulty of the task could vary across sentence pairs depending on the specific combination of sentences (for example depending on the presence of linguistic context in the target sentence or syntax similarity between the two competing sentences), the performance was averaged across the 20 sentence pairs presented in each experimental condition.

5.2.4 Measures of target and masker F_0 dynamic ranges effects

For each R value tested, speech intelligibility was analyzed as a function of the F_0 dynamic range of the individual sentences in a pair (i.e., the target and the masker). This analysis was conducted in the same way described in Section 4.2.6, by grouping the experimental stimuli based on their $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ values into equally spaced bins and by computing average speech intelligibility and standard errors within each bin for the two $\Delta\overline{F_0}$s. This analysis was conducted only for the TMR = 0 dB condition.

5.2.5 Statistical analysis

A mixed-effects analysis of variance (ANOVA) and post-hoc tests were conducted on the speech-intelligibility scores (transformed to rationalized arcsine units, RAU) averaged over the 20 pairs in each experimental condition. The analysis included the factors listener (treated as random factor), TMR, $\Delta\overline{F_0}$ and R (treated as fixed factors) and all two-way interactions between these factors.

Additionally, a linear regression analysis was performed on the speech intelligibility at the level of the single sentence pair. A mixed-effects model was obtained using a stepwise backward selection procedure, starting with an initial model that included the listener as random factors, the variables TMR, $\Delta\overline{F_0}$, $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ and R as fixed factors and all pairwise interactions between factors. Since the model aimed to clarify the combined effects of the experimental variables only, listener information such as, e.g., their amount of hearing loss or age was not included in the model and was assumed to be captured by the variance of the random structure. In the stepwise procedure, likelihood ratio tests were used for model comparisons and non-significant effects were sequentially removed until all remaining variables in the final model were significant. The analysis was performed using the software R (R Core Team, 2021) and the `lme4` package (Bates et al., 2014).

The ANOVA and the regression model offer different analyses since the ANOVA considered speech intelligibility averaged across the 20 pairs in each experimental condition (therefore allowing an analysis of the effects of the main experimental variables listener, TMR, $\Delta\overline{F_0}$ and R), whereas the regression analysis considered speech intelligibility at the level of each individual sentence pair presented to the listener, allowing to include the $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ variables, which varied randomly from pair to pair.

5.3 Results

5.3.1 Measured speech intelligibility scores

Figure 5.2 shows a boxplot of the speech intelligibility scores of each listener (1-13), indicated as percentage of correctly identified words from the target sentence (P_C). The boxplot was obtained from the P_C measured in all experimental conditions. Across listeners and experimental conditions, speech intelligibility varied widely, from a minimum of $P_C = 11\%$ (e.g., listeners 1 and 6) up to a maximum of 100% (e.g., listener 7). The individual average performance across all conditions (indicated by the red marks in the boxplot) varied substantially across listeners, ranging from a minimum of 39% for listener 1 to a maximum of 81% for listener 10.

The ANOVA conducted on the speech intelligibility scores and main experimental variables (listener, TMR, $\Delta\overline{F_0}$ and R) revealed listener, TMR and $\Delta\overline{F_0}$

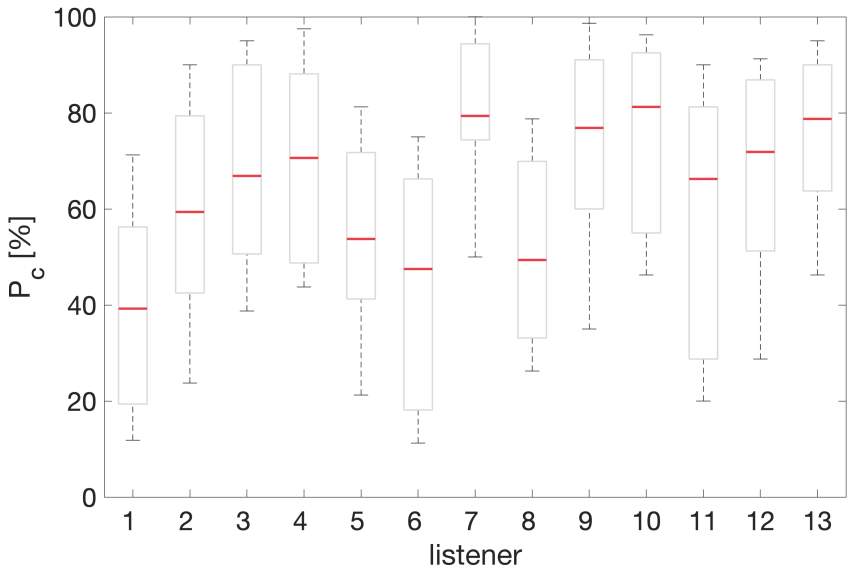


Figure 5.2: Boxplot of speech intelligibility scores for each of the 13 hearing-impaired listeners, averaged across the 20 sentence pairs within each experimental condition. The plot includes data obtained in all experimental conditions. Central red marks indicate median performance values, the boxes indicate the range of data between the 25$^{\text{th}}$ and the 75$^{\text{th}}$ percentiles and the whiskers indicate the utmost non-outlier values.

as significant factors ($p < 10^{-4}$, $p < 10^{-4}$ and $p < 10^{-3}$, respectively), as well as the interaction TMR$*\Delta\overline{F_0}$ ($p < 0.05$). Figure 5.3 shows the speech-intelligibility scores as a function of R, averaged over listeners for each experimental condition ($\Delta\overline{F_0}$ and TMR). The two $\Delta\overline{F_0}$ conditions of zero and six semitones are indicated by blue and red curves, respectively, while the two TMRs of 0 and 4 dB are indicated by solid lines with filled symbols and dashed lines with open symbols, respectively. The figure also shows P_C averaged across R for each $\Delta\overline{F_0}$ and TMR condition. Overall, no significant effect of R was found in any of the TMR and $\Delta\overline{F_0}$ conditions tested. When averaged across the $\Delta\overline{F_0}$ and R values tested, speech intelligibility differed significantly between TMR conditions ($p < 10^{-6}$), with average P_C of 46% at TMR = 0 dB and of 82% at TMR = 4 dB. Speech intelligibility differed significantly ($p < 10^{-4}$) between $\Delta\overline{F_0} = 0$ semitones and $\Delta\overline{F_0} = 6$ semitones. This difference, averaged across R conditions, amounted to 7 percentage points for TMR = 0 dB and was reduced to one percentage point at TMR = 4 dB.

Given the large variability in P_C across listeners, speech intelligibility was also analyzed by selecting the TMR condition that, for each listener, resulted

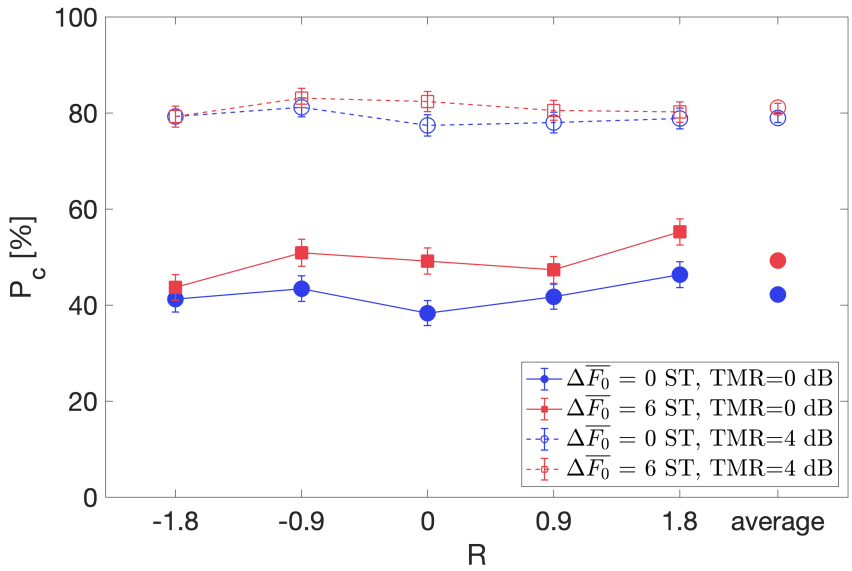


Figure 5.3: Speech intelligibility scores, shown as proportion of correctly recognized words from the target sentence, as a function of F_0-dynamic-range contrast, averaged across listeners. The figure shows data for the TMR = 0 dB condition (solid lines and filled symbols) and for the TMR = 4 dB condition (dashed lines and open symbols). Results for the two $\Delta\overline{F_0}$ conditions of zero and six semitones are shown with blue and red curves, respectively. For each TMR and $\Delta\overline{F_0}$ condition, results averaged across R values are also shown. The error bars represent standard errors.

in the speech-intelligibility score closer to 50% (indicated as TMR*). A TMR* of 0 dB was selected for all listeners except for listeners 1 and 6, who had the overall lowest average performance (see boxplot in Figure 5.2). An additional mixed-model ANOVA and post-hoc pairwise comparison analysis were conducted on the data for TMR*, by including the listener as random factor, $\Delta\overline{F_0}$ and R as fixed factors as well as their pairwise interactions. All main factors were found to have a statistically significant effect on P_C ($p < 10^{-4}$ for the factor listener, $p < 10^{-3}$ for the factor $\Delta\overline{F_0}$ and $p < 0.05$ for the factor R). None of the interactions had a statistically significant effect. Figure 5.4 shows P_C averaged across listeners as a function of F_0-dynamic-range contrast for the TMR* condition, with speech-intelligibility scores for the two $\Delta\overline{F_0}$ conditions of zero and six semitones shown with blue and red curves, respectively. In the TMR* condition, the two $\Delta\overline{F_0}$ conditions differed by 7 percentage points and this difference was significant at the $p < 10^{-4}$ level. When averaged across $\Delta\overline{F_0}$ conditions, P_C at $R = 1.8$ was found significantly higher than the value at $R = -1.8$ and $R = 0$ (8 percentage points, $p < 0.05$). No other R conditions differed significantly from one another.

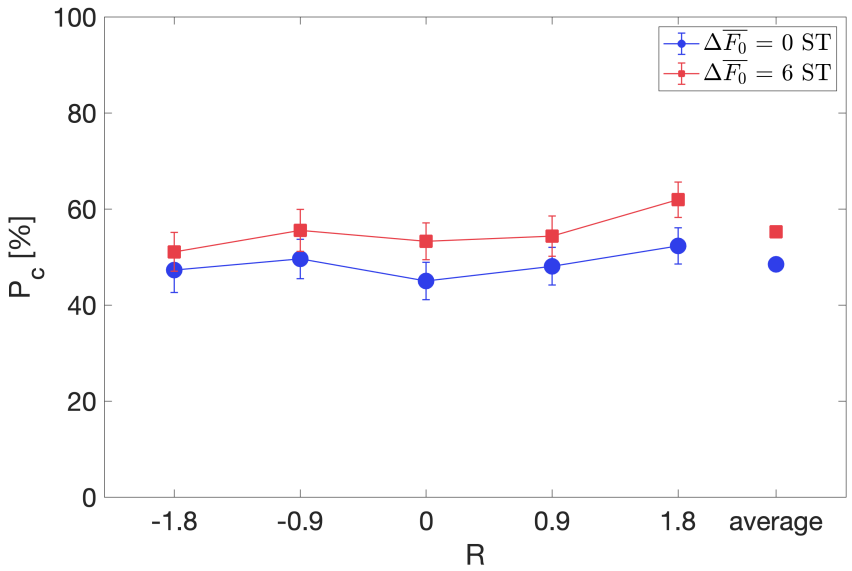


Figure 5.4: Speech intelligibility scores, shown as proportion of correctly recognized words from the target sentence, as a function of F_0-dynamic-range contrast, averaged across listeners in the TMR condition that, for each listener, resulted in the speech intelligibility score closer to 50%. Results for the two $\Delta\overline{F_0}$ conditions of zero and six semitones are shown with blue and red curves, respectively. For each $\Delta\overline{F_0}$ condition, results averaged across R values are also shown. The error bars represent standard errors.

5.3.2 Effects of target and masker F_0 dynamic range

Figure 5.5 shows the analysis of speech intelligibility as a function of the F_0 dynamic range of the individual sentences in a pair, for the TMR = 0 dB condition. Each panel shows the analysis for different R values, with speech intelligibility shown as a function of the $\sigma(F_0)$ of the sentence that spans the wider range of values (the target sentence for $R \geq 0$, the masker sentence for $R \leq 0$). This was done because, for non-zero R values, the $\sigma(F_0)$ of the less fluctuating sentence varied on a limited range (only few Hz for the most extreme R values) which did not allow to explore its effect on speech intelligibility. Points with standard errors larger than 10 percentage points are not shown. For reference, the P_C values averaged across $\sigma(F_0)$ values are shown for each condition. Overall, no systematic trend in the effects of $\sigma(F_0)$ on speech intelligibility could be observed, even though large variations in P_C were measured for specific conditions, such as the effect of $\sigma(F_0^{target})$ for $\Delta\overline{F_0} = 0$ semitones with $R = 0.9$, or the effect of $\sigma(F_0^{masker})$ for $\Delta\overline{F_0} = 6$ semitones with $R = -1.8$.

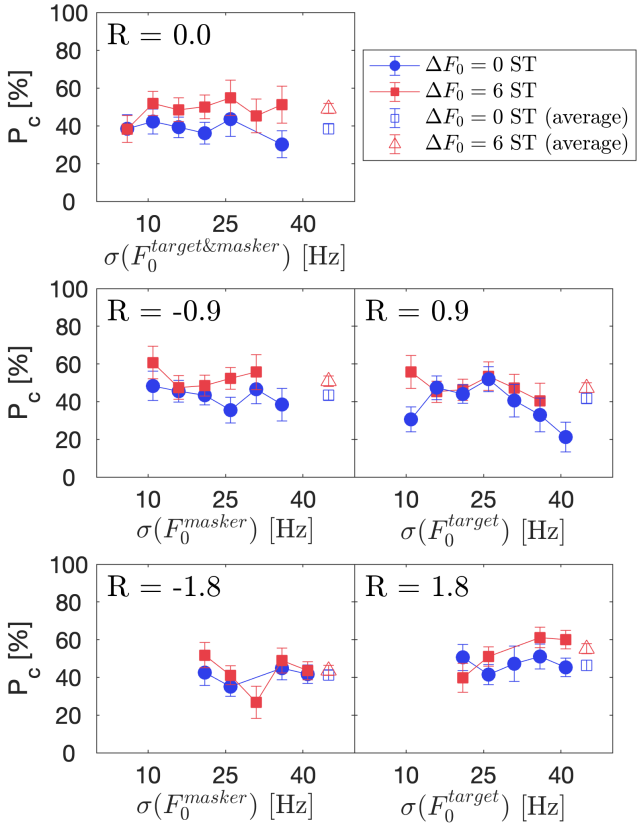


Figure 5.5: Speech intelligibility as a function of single-sentence $\sigma(F_0)$ (target or masker), for different R values, at TMR = 0 dB. For $R \geq 0$, speech intelligibility is shown as a function of $\sigma(F_0^{target})$, while for $R \leq 0$ speech intelligibility is shown as a function of masker $\sigma(F_0^{masker})$. The blue and red curves represent the data for the $\Delta\overline{F_0}$ of 0 and 6 semitones, respectively. The open symbols represent the average values calculated across all $\sigma(F_0)$ values. The error bars represent standard errors. Points with error bars larger than 10 percentage points are not shown.

Figure 5.6 (left panel) further illustrates the interaction of the effects of R, $\sigma(F_0^{target})$ and $\sigma(F_0^{masker})$ on speech intelligibility, by representing the stimulus space as divided into nine squared regions, defined by equally-spaced intervals on both x- and y-axis. For each region, the figure shows the average speech intelligibility change calculated with respect to the region where both target and masker had relatively low F_0 fluctuations, i.e., the bottom-left corner region (baseline condition), with lighter gray tones indicating improvements in speech intelligibility and darker gray tones indicating reductions in speech intelligibility. This analysis was conducted using the data from the $\Delta\overline{F_0} = 0$ semitones and TMR = 0 dB condition only. Because of the R values used in the present exper-

iment and the non-uniform distribution of the stimuli along iso-R lines (see Figure 4.2), two regions did not contain any stimuli and therefore no speech-intelligibility data are reported for these regions. The largest change in speech intelligibility with respect to the baseline condition (10 percentage points) was found for stimuli where both target and masker sentences were strongly fluctuating in F_0 (top-right corner of the space). However, this region contained only 3% of the experimental stimuli and therefore this result may be inaccurate. In all other regions, which all contained larger amounts of experimental stimuli, only small changes (all within 5 percentage points) in speech intelligibility from the baseline condition were measured.

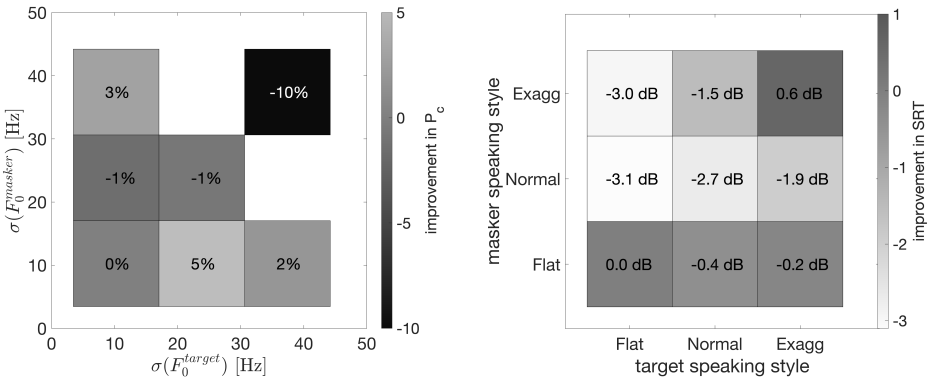


Figure 5.6: Left panel: Analysis of average changes in speech intelligibility scores for different regions of the stimulus space, in the $\Delta\overline{F_0} = 0$ semitones condition. The changes are calculated with respect to the average score in the region with lowest $\sigma(F_0)$ for both target and masker (bottom-left region), with lighter gray tones indicating improvements in speech intelligibility and darker tones indicating reductions in speech intelligibility. Right panel: corresponding analysis conducted on the data from Wasiuk et al. (2020), represented in a stimulus space divided in terms of the speaking style and their combinations. Changes in speech intelligibility are expressed as SRT difference in dB with respect to the region with flat target and flat masker (bottom-left region).

A similar analysis, shown in the right panel of Figure 5.6, was conducted on the data from Wasiuk et al. (2020), by representing their stimuli in a two-dimensional space also divided into nine regions defined by the target-masker speaking style combinations used in their study. Also in this case, changes in average speech intelligibility were computed for each region using the bottom-left region (corresponding to flat-versus-flat speaking style) as baseline condition. In their experiment, the largest changes in speech intelligibility with respect to the baseline condition were observed as a result of an increase in the F_0 dynamic

range of the masker sentence and were no larger than 3 dB. No other major effects were observed in the other regions of the stimulus space.

5.3.3 Regression analysis

Table 5.1 shows the results of the mixed linear model applied to the speech intelligibility scores (in percentage of correct target words, P_C) for each individual sentence pair. The estimates represent unstandardized mean differences with respect to the baseline condition of $R = -1.8$ with $\Delta\overline{F_0} = 0$ semitones and TMR = 0 dB. The final regression model resulting from the stepwise selection procedure was:

$$P_C \sim \Delta\overline{F_0} + \mathrm{TMR} + R + \Delta\overline{F_0} * TMR + (1\,|\,\mathrm{listener}).$$

The model showed that the strongest effect on P_C was induced by a change in TMR (mean difference 36.7 percentage points between TMR = 0 dB and TMR = 4 dB, $p < 0.001$), followed by a change in $\Delta\overline{F_0}$ (mean difference 7 percentage points between $\Delta\overline{F_0} = 0$ and $\Delta\overline{F_0} = 6$ semitone conditions, $p < 0.001$). The model showed significant differences from baseline R value only for $R = -0.9$ ($p < 0.05$) and $R = -1.8$ ($p < 0.01$), with an effect size of 4 percentage points for both R values. The model also showed a significant negative interaction for TMR and $\Delta\overline{F_0}$ ($p < 0.05$), indicating that the effect of $\Delta\overline{F_0}$ is reduced for higher TMR values (i.e., easier listening conditions). The outcome of the regression analysis confirmed the results of the ANOVA (see Section 5.3.1) and provided the additional result that, despite being included in the maximal model, the factors $\sigma(F_0^{target})$ nor $\sigma(F_0^{masker})$ did not have a significant effect on speech intelligibility.

5.4 Discussion

This study investigated the effects of average F_0 separation ($\Delta\overline{F_0}$) and F_0-dynamic-range contrast (R) on the intelligibility of competing sentences in listeners affected by hearing impairment. The intelligibility (P_C) of a target sentence masked by an interfering sentence was measured as a function of $\Delta\overline{F_0}$, R and target-to-masker ratio (TMR), by employing recordings of Danish sentences from several talkers and exploring different levels of F_0 dynamic range of the individual sentences and of F_0-dynamic-range contrast between competing sentences. The average F_0 and the F_0 dynamic range of the sentences were

Table 5.1: Outputs for the mixed effect model on the speech intelligibility scores. The model formula follows: $P_C \sim \Delta\overline{F_0} + \text{TMR} + R + \Delta\overline{F_0} * TMR + (1|\text{listener})$. The significance levels are indicated as $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Estimate	Std error	t	p
Baseline condition ($R = -1.8$, $\Delta\overline{F_0} = 0$ semitones, TMR = 0)	0.402	0.037	-	-
$R = -0.9$	0.038	0.016	2.345	*
$R = 0$	0.009	0.016	0.558	0.578
$R = 0.9$	0.010	0.016	0.640	0.523
$R = 1.8$	0.043	0.016	2.641	**
$\Delta\overline{F_0}$	0.070	0.015	4.843	***
TMR	0.367	0.015	25.221	***
$\Delta\overline{F_0} *$ TMR	-0.049	0.021	-2.375	*

manipulated with a numerically controlled method that modified the F_0 information without affecting other properties of speech. On average, the results revealed (i) an improvement in speech intelligibility when the two sentences were separated in their average F_0, compared to a condition where the competing sentences had the same average F_0; (ii) no effects on speech intelligibility as a function of F_0-dynamic-range contrast; (iii) no effect of the F_0 dynamic range of the individual sentences; (iv) an effect of TMR larger than the effect of any other experimental variable.

In line with what has been reported in previous research (Arehart et al., 1997; Hee Lee and Humes, 2012; Mackersie et al., 2011; Summers and Leek, 1998), the presence of a $\Delta\overline{F_0}$ of six semitones was found to be beneficial for target speech intelligibility, compared to a condition with $\Delta\overline{F_0} = 0$ semitones. This benefit (7 percentage points) was present at TMR = 0 dB but not at TMR = 4 dB (see Figure 5.3), indicating that, on average, hearing-impaired listeners rely on the $\Delta\overline{F_0}$ cue only when energetic cues are limited. At TMR = 4 dB the $\Delta\overline{F_0}$ cue might be redundant, likely due to good opportunities for energetic glimpsing of the target signal information, resulting in good speech intelligibility even without F_0 separation. This is also indicated by the overall good levels of performance at TMR = 4 dB (with average speech intelligibility above 80%), which was substantially higher than the performance at TMR = 0 dB (where average speech intelligibility was 46%). It is also possible that at positive TMR, the higher level of the target may have acted as a cue for the listener to more robustly identify the target between the competing sentences, compared to a condition with

TMR = 0 where both sentences had the same level and such level cue was not present. These effects of $\Delta\overline{F_0}$ and TMR on speech intelligibility can be compared to the effects observed in a study conducted with NH listeners, where a similar experimental method as the one used here was employed (reported in Chapter 4). For NH listeners, a 6-semitone average F_0 separation was beneficial for speech intelligibility for TMR ≤ -4 dB (with improvements of 13 and 6 percentage points at TMR of -8 dB and -4 dB, respectively) and did not have any effect for higher TMRs. Therefore, at TMRs that can be considered closer to conversational conditions (such as TMR = 0 dB), it seems that the $\Delta\overline{F_0}$ is a redundant cue for NH listeners, but is advantageous for the speech perception of HI listeners, even if moderately.

The improvement in speech intelligibility induced by $\Delta\overline{F_0}$ observed in the present experiment was smaller than that found in previous research that employed highly constrained and unrealistic speech materials (Arehart et al., 1997; Hee Lee and Humes, 2012; Mackersie et al., 2011; the experiment with competing vowels by Summers and Leek, 1998), but was instead comparable to the findings of previous studies that used more realistic speech materials, containing meaningful linguistic information such as syntax and context (Summers and Leek, 1998, in their experiment with competing sentences). A similar observation was reported in Chapter 3 in relation to the investigation of the $\Delta\overline{F_0}$ with NH listeners. However, the overall effect of $\Delta\overline{F_0}$ on speech intelligibility measured on NH listeners is reduced in HI listeners. Therefore, the present study is consistent with the previous findings that $\Delta\overline{F_0}$ is beneficial for speech intelligibility in HI listeners, but that the benefit of this cue is reduced compared to what has been previously observed in NH listeners.

The measured effects of R on speech intelligibility were negligible. However, when the results were analyzed by selecting the TMR that, for each listener, provided the speech intelligibility closer to 50% (indicated as TMR*), speech intelligibility was found to improve for $R = 1.8$ compared to a condition with $R = 0$ or $R = -1.8$. This approach allowed to measure the effect of R on a listening condition that, in terms of difficulty of the auditory task, can be considered uniform across listeners. This analysis revealed that, once the HI listeners have partial access to energetic cues (that alone is not sufficient for good speech intelligibility), their speech perception can be improved when the target has stronger F_0 dynamics than the masker (i.e., when there is a positive F_0-dynamic-range contrast). This effect was modest (8 percentage points) and observable

only between the extreme R values tested ($R = 1.8$ versus $R = 0$ or $R = -1.8$).

The present experiment represents a parallel investigation to the one conducted with NH listeners using the same experimental paradigm and speech material (see Chapter 4), where no effect of R on speech intelligibility was observed. Thus, in terms of their ability to utilize an F_0-dynamic-range contrast, HI listeners do not appear to differ substantially from NH listeners. However, for NH listeners speech intelligibility was found to be strongly influenced by the individual F_0 dynamic range of target and masker, especially when $\Delta\overline{F_0} = 0$ semitones (see Figures 4.5 and 4.6 in Section 4.3.3): with respect to a baseline condition where both sentences had small F_0 dynamic range, steep increases of speech intelligibility were observed in NH listeners when the F_0 dynamic range of at least one of the competing sentences increased, regardless of the F_0-dynamic-range contrast created. It was hypothesized in Chapter 4 that an F_0-dynamic-range increase of at least one of the competing sentences increases the dissimilarity between their F_0 trajectories and can thus be utilized by the NH listener to disentangle the target from the masker. However, this effect was not found in HI listeners in the present study. Their speech intelligibility performance was not influenced by the amount of F_0 fluctuations of the individual sentences, as shown by the regression analysis reported in Section 5.3.3. A negative effect on speech intelligibility was found when both sentences had wide F_0 fluctuations, but this finding was supported only by a small portion of the collected results and may not be sufficiently accurate. From these outcomes it seems that, contrary to what has been observed for NH listeners, HI listeners are not able to utilize the differences in the temporal pattern of the competing F_0 trajectories and that wide F_0 fluctuations in both sentences may even have the opposite effect and be a source of confusion for them.

The findings of the present study, where no effect of R could be observed, are in disagreement with the results from Wasiuk et al. (2020), who instead showed a moderate improvement in speech intelligibility as a function of R. In particular, Wasiuk et al. (2020) found that, compared to a baseline condition with $R = 0$, speech intelligibility increased when $R < 0$ but not when $R > 0$. Similarly to what reported in the discussion of Chapter 4 in relation to the differences with the results by Calandruccio et al. (2019) (whose paradigm was replicated by Wasiuk et al. (2020), the difference between the present findings and the results by Wasiuk et al. (2020) might be explained by the differences between the experimental methods and speech materials. In particular, the different

levels of F_0 dynamic range used by Calandruccio et al. (2019) and Wasiuk et al. (2020) were generated by asking the talkers to alter their speaking style such as to produce monotonous, normal and exaggerated intonations. The modification of the speaking style may have influenced other acoustic features of speech than F_0 that may have affected speech intelligibility. Therefore, the effect of R observed by Wasiuk et al. (2020) may not be the consequence of the contrast in F_0 dynamic range only. Furthermore, the auditory task in the present study differed substantially from that used by Calandruccio et al. (2019) and Wasiuk et al. (2020). In fact, in their experiment the target speech (always spoken by the same female voice, which was learned by the listeners during a training procedure) was presented 500 ms after the onset of a two-talker masker, which was obtained by mixing random excerpts of continuous speech from two other talkers and therefore did not contain meaningful linguistic information. By contrast, in the present experiment, the target sentence was (i) spoken by a talker that varied from pair to pair, (ii) cued by presenting its first word to the listener prior to playback, and (iii) masked by a single sentence spoken by the same target talker. Therefore, the two experimental approaches represent auditory situations that differ in several aspects and a more elaborate comparison of their results is thus not possible.

This study offers a tool to quantitatively assess how F_0-related cues, manipulated in isolation from other auditory cues, can affect the intelligibility of competing speech for HI listeners. The understanding of how distinct acoustic properties of the stimulus can affect auditory perception is important as it can contribute to inspiring novel hearing-aid technologies aimed at manipulating and enhancing such properties for improving speech intelligibility. From the obtained findings, it seems that HI listeners provided with linear-gain amplification benefit from an average F_0 separation between competing voices but not from the dynamic variations in the F_0 of the voices. However, the investigation presented here utilized a range of average F_0 and F_0 dynamics that replicated the range of values of natural voices and could be extended to explore the effects of F_0 trajectory manipulations that exceed the natural values (of both average F_0 and F_0 dynamic range), to verify if HI listeners can benefit from F_0 information that is enhanced beyond its natural range. At the same time, such additional investigation should be combined with an assessment of the acoustic naturalness of the manipulated stimuli, to ensure that sound quality is not affected by the F_0 manipulation. Finally, additional research is required to understand the

effects of F_0-related cues on speech intelligibility for HI listeners when they are provided with state-of-the-art hearing-aid technologies that can include, for example, non-linear amplification strategies.

5.5 Summary and conclusion

The present study investigated the effects of average F_0 separation and F_0-dynamic-range contrast on the intelligibility of competing sentences mixed at two different TMRs and presented to hearing-impaired (HI) listeners provided with linear-gain amplification. The experimental method was based on a speech material that provides a variety of voices and F_0 characteristics typical of realistic speech and on a manipulation method of the F_0 information of the voice in isolation from any other acoustic feature of speech. The results were compared to results previously measured in normal-hearing (NH) listeners tested with a similar experimental method and speech material. It was observed that:

(i) The separation in average F_0 between competing sentences had a moderate benefit on speech intelligibility, compared to the case where the sentences had the same average F_0. Compared to previous findings in NH listeners, this effect was smaller and present at higher TMRs.

(ii) The F_0-dynamic-range contrast between competing sentences had a negligible effect on speech intelligibility, regardless of the TMR at which the sentences were presented.

(iii) The F_0 dynamic range of the individual sentences had no effect on speech intelligibility, whereas it was previously shown to be largely beneficial for NH listeners.

Directions for future research were suggested to further explore the effects of F_0-related cues on speech intelligibility in competing-talker scenarios for HI listeners.

Acknowledgments

I would like to thank Andrea Dich Jensen for her valuable help in contacting the participants to the experiment and collecting the speech intelligibility measurements reported in this study.

6

On the influence of the auditory environment on fundamental-frequency production

Abstract

The fundamental frequency (F_0) is one of the many features of the voice that a talker can alter in adverse communicative situations, such as in presence of background noise or of a hearing-impaired (HI) interlocutor, with the intent to produce 'clear speech' and increase the intelligibility of the transmitted message. Compared to 'conversational speech', clear speech has been shown to be characterized by higher average F_0 and larger F_0 dynamic range. However, these changes in F_0 were often measured in laboratory simulations where normal-hearing (NH) listeners were speaking in the absence of an interlocutor. The present study further explored changes in F_0 occurring in clear speech by analyzing the F_0 statistics of naturalistic dialogues conducted in quiet, in presence of background noise, and/or in presence of a HI interlocutor. It was found that (i) in presence of background noise, both NH and HI talkers increased their average F_0 and F_0 dynamic range, and (ii) when speaking with a HI interlocutor, NH talkers exhibited a higher average F_0 as well as a larger F_0 dynamic range compared to when they were speaking with a NH interlocutor. However, when analyzed in the individual talker, the changes in average F_0 and F_0 dynamic range were not always produced concurrently. Overall, these results provide further evidence that, in adverse communicative conditions, talkers change their voice and speaking style with the intent to facilitate the transmission of the speech message.

6.1 Introduction

Oral communication can be challenged by factors related to the acoustic environment, such as background noise or the presence of interfering talkers (Bronkhorst, 2000; Cherry, 1953), as well as to characteristics of the interlocutors that may affect the quality of the speech signal and its perception, such as hearing impairment or poor language proficiency (e.g., in non-native speakers or infants). The exchange of oral information between a talker and a listener is an environment-adaptive task that is performed with joint efforts from both interlocutors. While the listener's task is to focus their attention on the available auditory and linguistic cues that support speech perception, whose processing can require considerable cognitive resources (Peelle, 2018), the intent of the talker is to promote speech intelligibility while maintaining the attention of the listener and minimizing their cognitive effort. To do that, the talker may adapt aspects of their speech production to enhance the salience of auditory and linguistic cues that can help overcome the specific communication obstacles in the acoustic scene, i.e., by producing the so-called 'clear speech' (Cooke et al., 2014; Uchanski, 2005). It has been demonstrated that clear speech has a positive impact on speech understanding in challenging acoustic situations as it is more intelligible than conversational speech, especially for hearing-impaired listeners (Payton et al., 1994; Picheny et al., 1985).

The speech adaptations leading to clear speech can be targeted to the linguistic content of the speech message, for example by reducing its semantic and syntactical complexity when talking to a person with poor proficiency in the language used in the conversation, such as an infant (i.e., when using 'infant-directed speech') or when talking to a foreigner. Other adaptations can be targeted to spectro-temporal features of the speech signal, such as its overall intensity, amplitude modulations or long-term spectral shape.

Adaptations of the acoustic properties of speech production to the acoustic environment were first reported by Lombard (1911), who observed that talkers increased the intensity of their voice when speaking in the presence of noise (i.e., the so-called Lombard effect). Since then, many studies focused on the changes in speech production that occur in adverse communicative conditions, extending the research to aspects other than the overall intensity of speech and to environmental factors other than background noise. A thorough overview of the research on the topic can be found in Brumm and Zollinger (2011) and

Cooke et al. (2014).

One of the acoustic features of speech that talkers can alter to facilitate the listening task for their interlocutors is the fundamental frequency (F_0). It has been shown that, compared to conventional speech, clear speech has a higher average F_0 (Garnier and Henrich, 2014; Krause and Braida, 2004; Picheny et al., 1986; Summers et al., 1988) and a wider F_0 dynamic range (Fernald and Simon, 1984; Grieser and Kuhl, 1988; Krause and Braida, 2004; Stern et al., 1983). However, the available research is not exhaustive in describing the changes in F_0 in clear speech. Some of these studies limited their analysis to a small amount of talkers of the same gender (Picheny et al., 1986; Summers et al., 1988) and in some cases the changes in F_0 were not observed consistently across talkers (Krause and Braida, 2004; Picheny et al., 1986; Summers et al., 1988). In general, all available evidence of clear speech production seems to be obtained from normal-hearing (NH) talkers, whereas no data are available in relation to hearing-impaired (HI) talkers. Furthermore, the previous investigations have been focused on specific communicative situations, such as mothers speaking to infants (Fernald and Simon, 1984; Grieser and Kuhl, 1988; Stern et al., 1983). Finally, the speech samples analyzed in some of these studies may not be representative of realistic situations, since they were recorded in absence of an interlocutor, by asking the talker to speak as if (or literally "imagining that") they were in specific communicative situations that could limit the intelligibility of their speech to their interlocutors (Krause and Braida, 2004; Picheny et al., 1985; Summers et al., 1988). Therefore, the knowledge about the adaptations in F_0 that talkers produce to overcome certain communication obstacles in realistic situations can be expanded toward the investigation of speech produced in a wider set of more realistic communicative situations, for example situations involving hearing-impaired individuals, not only as listeners but also as talkers. The understanding of the natural speech production changes applied when a hearing-impaired person participates in the conversation (either as a talker or a listener) can be inspiring for the development of signal-processing strategies targeted to overcome hearing disabilities and improve the active participation in conversations by people with hearing loss.

This study presents a comparative analysis of the F_0 information measured in speech obtained from laboratory recordings of naturalistic dialogues conducted in quiet and in different noisy conditions, between NH interlocutors as well as between NH and HI ones. These recordings were obtained from previ-

ous studies (Sørensen et al., 2021; Sørensen et al., 2019) that investigated the turn-taking dynamics occurring between interlocutors engaging in dialogues. These two studies found that NH talkers, when conversating with NH interlocutors in background noise, compared to when they were speaking in quiet, (i) increased the overall level of their voice, (ii) increased the average duration of their speaking turn, (iii) produced speech with faster articulation rates (i.e., a higher number of syllables per second), (iv) did not change the duration of the turn-taking transition between interlocutors. In contrast, in dialogues between NH and HI people, both interlocutors attempted to ease the effort of the conversation in presence of background noise (compared to a quiet condition) by (i) increasing the overall level of their voice, (ii) decreasing the articulation rate, (iii) taking longer durations in their speaking turn, (iv) increasing the durations of turn-taking transition and (v) reducing the occurrence of talking-turn overlap. Furthermore, they found that NH talkers speaking with HI persons overall spoke less and produced speech with lower articulation rate than their HI interlocutors, likely in the attempt to facilitate speech comprehension for their interlocutors.

The analysis presented in this study complements the findings of Sørensen et al. (2021) and Sørensen et al. (2019) by offering an overview of the changes in F_0 statistics occurring when speaking in the presence of noise and/or in situations with HI interlocutors. Compared to previous studies investigating the changes in F_0 that occur in challenging communicative situations, this analysis had the advantage of (i) including a large number of talkers, (ii) observing the natural (i.e., not simulated) adaptations in F_0 production that the talkers apply in presence of their interlocutors and (iii) measuring how such adaptations are produced for and by HI interlocutors. Measures of average F_0 and F_0 dynamic range of speech were compared across different communicative conditions between pairs of talkers, i.e., for different levels of background noise and in presence or absence of hearing impairment in one of the interlocutors.

6.2 Methods

The speech material consisted of laboratory-recordings of dialogues in the Danish language conducted by 19 pairs of NH interlocutors (indicated as NH1) and 12 other pairs consisting of a NH interlocutor (belonging to a different group, indicated as NH2) speaking with a HI interlocutor (indicated as HI), in quiet

or with different levels of noise, while they performed a Diapix task (Baker and Hazan, 2011). The recordings of the voice of each talker from the dialogues were labelled as talker-listener combination, i.e., NH1-NH1, NH2-HI and HI-NH2, where the first acronym indicated the type of talker and the second acronym indicated their interlocutor in the dialogue, followed by a label indicating if the dialogue was produced in quiet or in background noise. For the pairs of NH1 interlocutors, the background noise consisted of 6-talker speech-shaped noise (ICRA 7; Dreschler et al., 2001) at 70-dBA sound pressure level. For the pairs of NH2 and HI interlocutors, the background noise consisted of 20-talker babble noise at 60-, 65- and 70-dBA sound pressure level. Each pair of interlocutors recorded three dialogue sessions. More details about the dialogue recordings are reported in Section 2.1.2 and in Appendix A.

The F_0 of the speech material was analyzed as follows. For each talker-listener combination and background noise condition, the F_0 trajectories from the speech signal of the talker in the different recording sessions were extracted with the software PRAAT (using the autocorrelation method by Boersma et al., 1993) and concatenated to obtain a unique F_0 trajectory. The details of the F_0-extraction algorithm in PRAAT are the same used in Section 2.3 and reported in Table 2.2. The F_0 trajectories were not processed to remove silences (naturally present due to the turn-taking between the interlocutors) or non-speech sounds such as coughing or laughing. The long-term statistics of the F_0 of each talker in each condition were quantified in terms of average F_0 (calculated as the median value of the trajectory and indicated as $\overline{F_0}$) and F_0 dynamic range (calculated as the median absolute deviation, i.e., MAD, of the trajectory and indicated as $\sigma(F_0)$). The $\overline{F_0}$ and $\sigma(F_0)$ were compared across the different talker-listener combinations and background noise conditions.

6.3 Results

Figure 6.1 shows the $\overline{F_0}$ of the individual talkers for each dialogue condition (talker-listener combination and background noise condition), indicated by open black circles, as well as the corresponding group average $\overline{F_0}$ and standard errors, indicated by filled red circles. A two-way analysis of variance (ANOVA) and a post-hoc pairwise comparison analysis were conducted on the $\overline{F_0}$ data, including the type of talker-listener combination (NH1-NH1, NH2-HI and HI-NH2) and the background noise condition (only the quiet and the 70-dBA con-

ditions, i.e., the only conditions that were available for all groups of talkers) as fixed factors, as well as their interaction. The ANOVA indicated both talker and background noise as significant factors ($p < 10^{-3}$), but not their interaction. The post-hoc analysis revealed that the $\overline{F_0}$ was significantly higher in presence of 70-dBA background noise than in quiet ($p < 10^{-3}$). On average, the $\overline{F_0}$ difference between the two conditions was 26 Hz (2.8 semitones) for NH1-NH1, 36 Hz (3 semitones) for NH2-HI and 35 Hz (3.6 semitones) for HI-NH2 talkers. For NH2-HI and HI-NH2 (whose recordings were available at three different levels of background noise), the increase in $\overline{F_0}$ was monotonic with increasing in noise level. With very few exceptions, the increases in $\overline{F_0}$ for increasing level of background noise were consistent across talkers within each group. NH1-NH1 and HI-NH2 had similar (i.e., not significantly different) $\overline{F_0}$ on a group average level both in the quiet and the 70-dBA noise condition. However, both groups had $\overline{F_0}$ values that were significantly lower than for NH2-HI ($p < 10^{-3}$).

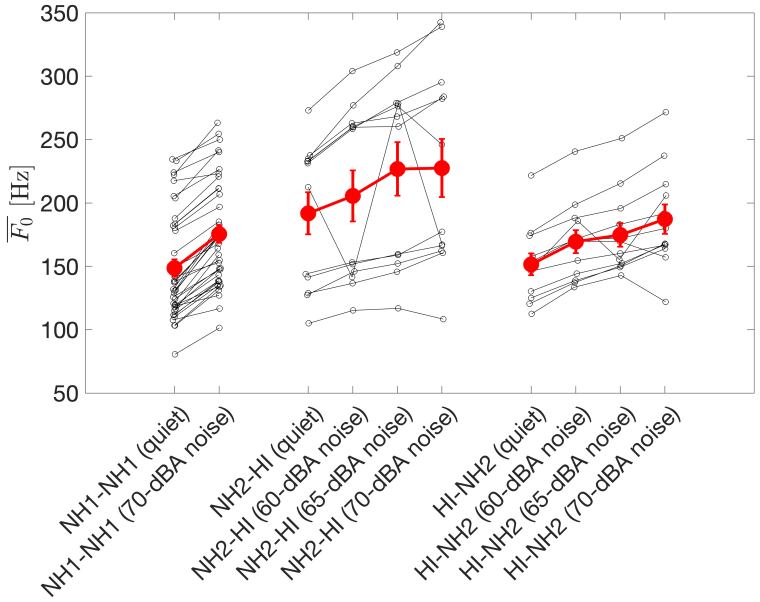


Figure 6.1: F_0 median for speech recordings of different talkers engaged in dialogues in different environmental conditions. The recordings are grouped as *talker-listener* combination: NH speaking with NH (NH1-NH1), NH speaking with a HI (NH2-HI, where NH2 is a different group from NH1) and the same HI speaking with NH2 (HI-NH2). The NH1-NH1 dialogues were recorded in quiet and in 6-talker modulated speech-shaped noise presented at 70 dB SPL(A). The NH2-HI dialogues were recorded in quiet and in 20-talker babble noise presented at 60, 65 and 70 dBA SPL(A). Open black circles indicate individual talker data, while filled red circles indicate group averages with error bars showing standard errors.

Figure 6.2 shows the $\sigma(F_0)$ for each talker-listener combination and condition, as well as group averages for each condition, with the same colors and symbols as in Figure 6.1. The same two-way analysis of variance (ANOVA) and a post-hoc pairwise comparison analysis conducted on the $\overline{F_0}$ were conducted on the $\sigma(F_0)$ data. Similarly to what was found for the $\overline{F_0}$, both talker and background-noise condition were significant factors ($p < 10^{-3}$ and $p < 10^{-2}$, respectively). On average, the $\sigma(F_0)$ showed trends in line with those observed for the $\overline{F_0}$, both across groups of talkers and across the background noise conditions within a group of talkers. First, $\sigma(F_0)$ increased in presence of background noise, with larger increases at higher noise levels, compared to the quiet condition. The increment in $\sigma(F_0)$ measured between the quiet and the 70-dBA noise conditions was statistically significant ($p < 10^{-2}$) but was smaller for NH1-NH1 and HI-NH2 (2.5 Hz and 2.1 Hz, respectively) than for NH2-HI (5.3 Hz). In contrast to what was observed for the $\overline{F_0}$ at the level of the individual talker, the increment in $\sigma(F_0)$ induced by the increase in noise level was not coherent across all talkers within a group. The $\sigma(F_0)$ of NH2-HI was significantly higher than that

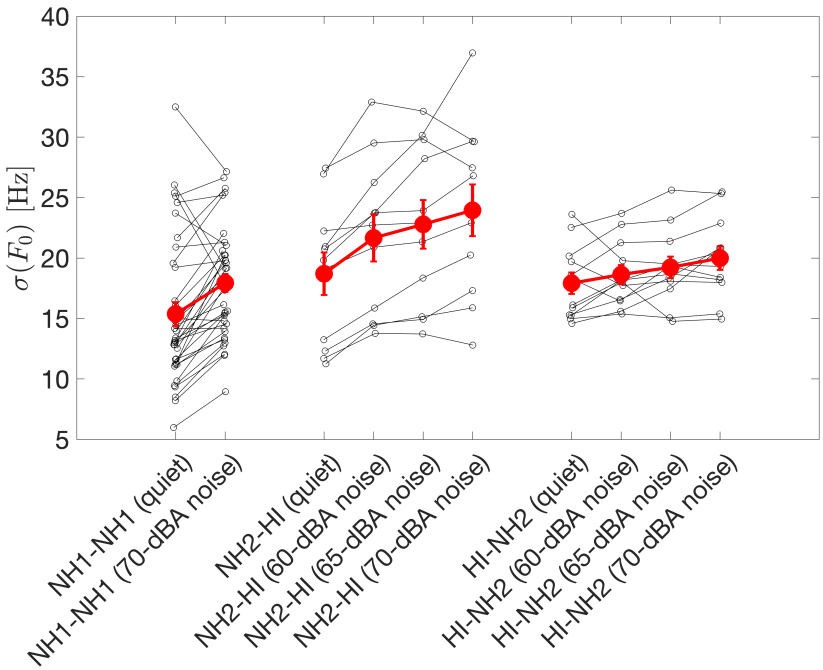


Figure 6.2: F_0 median absolute deviation for speech samples spoken by different talkers engaged in dialogues in different environmental conditions. The symbols are the same as defined in Figure 6.1.

of NH1-NH1 ($p < 10^{-2}$), while $\sigma(F_0)$ of HI-NH2 did not differ significantly from the other two groups of talkers.

Figure 6.3 illustrates the relationship between $\sigma(F_0)$ and $\overline{F_0}$ for each talker. For visualization purposes, the figure shows the data only for the quiet (open symbols) and the 70-dBA background noise (filled symbols) conditions, which were available for all groups of talkers. The general trend observed was that higher $\sigma(F_0)$ corresponded to higher $\overline{F_0}$. When considering the data from all conditions (including also the 60- and 65-dBA SPL noise conditions, available only for NH2-HI and HI-NH2), the two statistical metrics were found to be linearly correlated ($\rho = 0.8$). The changes in $\overline{F_0}$ and $\sigma(F_0)$ were calculated for each talker between the quiet and the 70-dBA background noise condition (i.e., the conditions in common between all talkers).

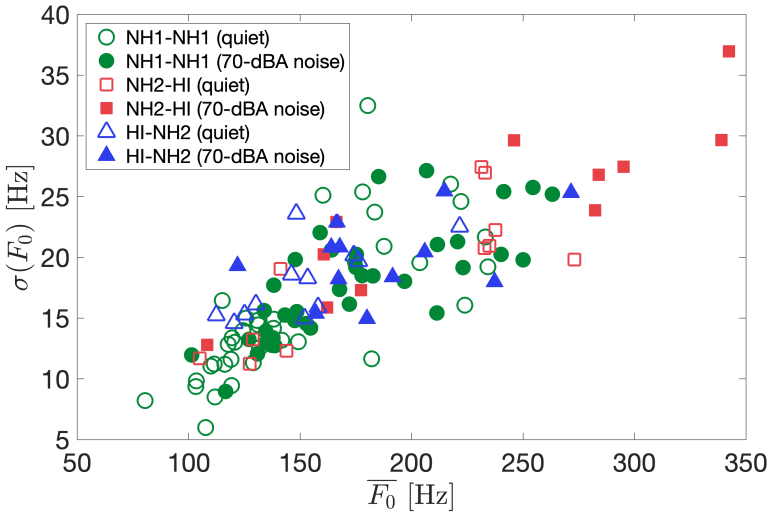


Figure 6.3: F_0 dynamic range as a function of median F_0 for each talker in the quiet condition (open symbols) and 70-dBA background noise (filled symbols) conditions. Talkers from different groups are represented by different symbols.

Figure 6.4 shows the change in $\sigma(F_0)$ ($\Delta\sigma(F_0)$) as a function of the change in $\overline{F_0}$ ($\Delta\overline{F_0}$) between these two conditions. Data from talkers from different groups are distinguished by different colors and symbols. Despite the group average increase of $\overline{F_0}$ and $\sigma(F_0)$ with increasing background noise level and the correlation between $\overline{F_0}$ and $\sigma(F_0)$ found across all conditions (shown in Figure 6.3), this analysis showed that not all talkers modified the average F_0 and the F_0 dynamic range concurrently. In fact, a correlation between $\Delta\overline{F_0}$ and $\Delta\sigma(F_0)$ was only found for the NH2-HI talker-listener group ($\rho = 0.72$), and not

for the other two groups ($\rho = 0.46$ and $\rho = -0.2$ for NH1-NH1 and for HI-NH2, respectively).

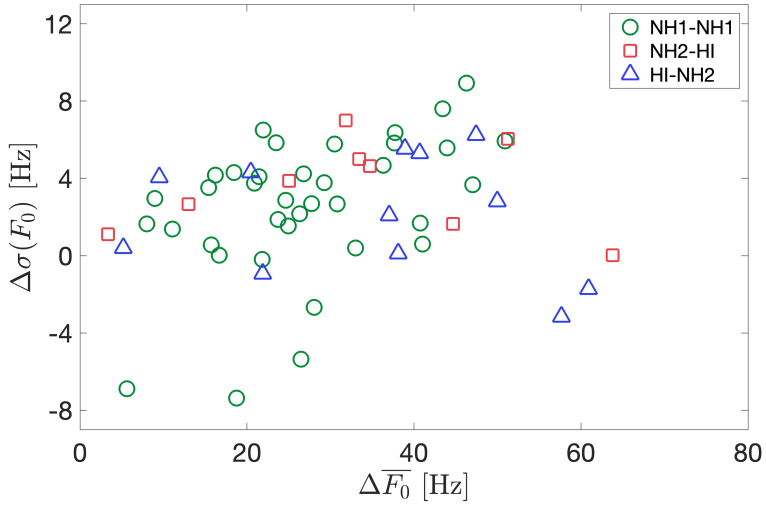


Figure 6.4: Change in F_0 dynamic range as a function of the change in median F_0 between quiet and background-noise (70-dBA SPL) conditions, for each talker. Talkers from different groups are indicated with different colors and symbols.

6.4 Discussion

The analysis of the speech materials presented in this study offers an overview of the changes in F_0 (described in terms of its long-term statistics) occurring in the speech production (i) of NH talkers when they speak in presence of background noise to a NH or a HI interlocutor and (ii) of HI talkers when they speak to NH interlocutors in presence of background noise.

On average, the presence of background noise during the dialogue prompted both NH and HI talkers to increase their average F_0 (quantified as median F_0, $\overline{F_0}$) and F_0 dynamic range (quantified with the F_0 median absolute deviation, $\sigma(F_0)$). These increments were monotonic with increasing noise level. The increase in $\overline{F_0}$ and $\sigma(F_0)$ when speaking in presence of background noise observed in the present study is in agreement with previous research that investigated clear speech. However, most of the previous studies focused their investigation on the changes occurring in $\overline{F_0}$ while the changes in $\sigma(F_0)$ were investigated only for speech produced by NH talkers that were talking in specific situations (such as mothers speaking to infants; Fernald and Simon, 1984; Stern et al., 1983) or

in unrealistic communicative situations (e.g., where the talkers were asked to mimic clear speech from previous recordings in the absence of an interlocutor; Krause and Braida, 2004). Furthermore, the available studies did not investigate the $\sigma(F_0)$ changes occurring in presence of background noise. The present study provided additional evidence of the changes in $\overline{F_0}$ and $\sigma(F_0)$ occurring in adverse communicative conditions, expanding the research to realistic dialogues that include HI interlocutors and various levels of background noise.

Larger $\overline{F_0}$ and $\sigma(F_0)$ observed in background noise compared to the quiet condition indicate an increased vocal effort of the talkers to improve the intelligibility of their speech in adverse communicative conditions since voices with higher average F_0 and higher F_0 dynamics are known to be more intelligible in both background noise and competing speech listening conditions (Assmann, 1999; Binns and Culling, 2007; Mackersie et al., 2011; Chapter 4 of this thesis). In conversational speech, the $\overline{F_0}$ and the $\sigma(F_0)$ of the voice are positively correlated, with larger F_0 dynamics usually produced by voices with a higher average F_0, as shown for example in Chapter 4, Figure 4.1. The $\overline{F_0}$ and the $\sigma(F_0)$ data obtained in the present study seem to confirm this trend also in the case of clear speech (see Figure 6.3 and the high correlation between $\overline{F_0}$ and the $\sigma(F_0)$ measured for each talker in all conditions). However, at the level of the individual talker, $\overline{F_0}$ and $\sigma(F_0)$ were not always adapted concurrently, as shown in Figure 6.4. This result indicates that different talkers may apply different strategies of F_0 enhancement to increase the intelligibility of their speech for their interlocutors.

The NH and HI talkers were found to behave similarly when speaking in noise, in terms of the changes in F_0 statistics. In fact, on average, in all talker-listener groups (NH1-NH1, NH2-HI and HI-NH2), the talkers increased their $\overline{F_0}$ and $\sigma(F_0)$ monotonically with increasing level of background noise. Additionally, both in quiet and in noisy conditions, the NH talkers showed a considerable increase of both $\overline{F_0}$ and $\sigma(F_0)$ when they were talking with HI listeners (NH2-HI) compared to when they were talking with other NH listeners (NH1-NH1). However, the NH talkers participating in conversations with HI interlocutors (i.e., the NH2 -HI group) were not the same talkers that conducted the dialogues with the NH interlocutors (i.e., the NH1-NH1 group). Furthermore, the NH2-HI group conducted dialogues in presence of a background noise that lacks F_0 information (namely, the ICRA-7 noise, consisting of a broadband noise carrier modulated by the envelopes of a 6-talker babble), whereas the NH1-NH1 group conducted the dialogues in presence of a 20-talker speech mixture,

characterized by all the acoustic features of speech. Therefore, the differences observed in the F_0 statistics between the groups of talkers may have been due to the specific group composition and background noise conditions. A more accurate analysis of how NH modify their $\overline{F_0}$ and $\sigma(F_0)$ would require testing the same group of NH talkers when speaking with both NH and HI interlocutors in the same background noise. Since this study analyzed speech recordings available from previous research, such accurate comparison was not possible.

A distinction must be made between Lombard speech and clear speech. Lombard speech pertains to the voice modifications actuated (potentially involuntarily) by the talker when speaking in loud environments in response to difficulties in talking, regardless of the potential difficulties that their interlocutor may exhibit in perceiving speech. In fact, the available evidence of Lombard speech has been obtained in laboratory settings where a talker was speaking without any interlocutor. On the contrary, clear speech may contain a set of voice modifications that are triggered by the need (or actuated with the intent) to overcome potential perception barriers for the interlocutor. Since the speech recordings analyzed in this study were obtained from realistic dialogues (with a talker speaking in presence of an interlocutor), it is difficult to disentangle the modifications induced by challenges for the talker to speak and those induced by challenges for the listener to perceive speech.

Overall, the results of the present study support the findings by Sørensen et al. (2021) and Sørensen et al. (2019) that, in the presence of communication barriers such as background noise and/or hearing-impairment, participants in a conversation actuate changes to their voice and speaking style with the aim to facilitate the transmission of the speech messages.

6.5 Summary and conclusion

This study analyzed the modification in the F_0 of the voice occurring in normal-hearing (NH) and hearing-impaired (HI) talkers when conducting dialogues in adverse conversational environments, such as characterized by the presence of background noise or of hearing-impaired participants in the conversation. The F_0 information from recordings of realistic dialogues conducted at different levels of background noise between pairs of NH and/or HI talkers was quantified in terms of average F_0 ($\overline{F_0}$) and F_0 dynamic-range ($\sigma(F_0)$). Changes in $\overline{F_0}$ and $\sigma(F_0)$ of NH and HI talkers were measured in correspondence of absence or

presence of background noise and (for NH talkers only) in presence of a hearing-impaired interlocutor. It was found that:

(i) In reaction to the presence of background noise, NH and HI talkers behaved similarly, by increasing $\overline{F_0}$ and $\sigma(F_0)$. This increase was monotonic with increasing noise level. However, at the level of the individual talker, $\overline{F_0}$ and $\sigma(F_0)$ were not always adapted concurrently.

(ii) When speaking with HI interlocutors, NH talkers increased their $\overline{F_0}$ and $\sigma(F_0)$ considerably, compared to when they were speaking with another NH person. However, this result may be biased by the different composition of the two NH groups that conducted the dialogues with NH and HI interlocutors as well as differences in noise type across studies.

These results contribute to the knowledge of how talkers modify their voices to overcome auditory barriers to speech perception in conversation, extending the available findings to measurements on HI interlocutors (considered both as talkers and as listeners). The obtained findings may provide useful knowledge for future studies of speech perception in relation to the F_0 and for the development of speech-enhancement strategies.

Acknowledgments

I would like to thank Josefine Munch Sørensen for sharing the speech recordings from Sørensen et al. (2019) and Sørensen et al. (2021) analyzed in the study presented in this chapter.

7

General discussion

The work presented in this thesis focused on assessing the role of the fundamental frequency (F_0) on speech intelligibility in competing-talker scenarios. Speech intelligibility was investigated in normal-hearing (NH) and hearing-impaired (HI) listeners as a function of the difference in average F_0 and as a function of the difference in F_0 dynamic range (i.e., an F_0-dynamic-range contrast) between two competing sentences. The investigation of how these auditory cues influence speech intelligibility in both NH and HI listeners is not a novelty in auditory research. Using different experimental approaches, many studies have shown that F_0 cues are beneficial for speech intelligibility in NH listeners, while their effects are severely reduced in HI listeners. However, realistic competing-talker scenarios represent complex acoustic situations that may be difficult to replicate in the laboratory environment, as they comprise a multitude of aspects related to, for example, the acoustics of the surroundings, the number of talkers involved in the scenario, their distribution in space with respect to the listener, the features of their voices and the content of the spoken messages, let alone visual aspects like face movements and gestures. All these aspects are difficult to capture altogether in a laboratory simulation of competing-talker scenarios, which has to be necessarily selective as to what features of the realistic settings are being reproduced. The previous research on the effects of F_0-related cues in competing-talker scenarios is no exception to that rule and thus employed experimental methods that reproduced only certain aspects of real-life competing-talker scenarios. In the attempt to isolate the effects of F_0-related cues, several studies utilized speech materials with highly constrained acoustic and linguistic characteristics (including the F_0 itself) that may not be representative of the heterogeneity that is found in realistic speech. Furthermore, in some cases, the experimental methods did not allow an accurate and isolated control of the experimental variables, such that the cue under investigation (the F_0) was not very well controlled and/or co-varied with other acoustical features of the speech.

This thesis aimed at extending the available knowledge of the effects of F_0-related cues on speech intelligibility by using speech materials that are well representative of the acoustic and linguistic variety of real-life speech and by offering an experimental method that is based on a numerically accurate technical approach to describe and manipulate the naturally produced F_0.

The utilized speech materials were sentences from the Danish Hearing In Noise Test (HINT; Nielsen and Dau, 2011), spoken by a variety of male and female talkers. To assess the realism of the F_0 information contained in the utilized speech materials, a comparative analysis of the F_0 information for the HINT corpus and other speech corpora used in previous studies was presented in Chapter 2, along with an analysis of Danish dialogues (used as a highly naturalistic reference). The analysis showed that the HINT speech material used in the experiments presented in this thesis well reproduces the values and variety of the long-term F_0 statistics that are found in realistic voices. By contrast, speech materials utilized in previous studies contained F_0 information that, in terms of its long-term statistics, was either not found in realistic voices or was only partially representative of them. Equipped with such F_0-realistic speech material, three experiments were conducted that investigated the effects of a difference in average F_0 and of an F_0-dynamic-range contrast between competing sentences on speech intelligibility in NH and HI listeners. In all experiments, speech intelligibility was measured by presenting pairs of sentences, a target and a masker, both spoken by the same talker (which varied from pair to pair), in the absence of spatial cues (i.e., in a co-located talker condition) and mixed at various target-to-masker ratios (TMRs).

7.1 Summary of main results

In Chapter 3, the effects of the long-term average F_0 separation on speech intelligibility for NH listeners were reported. The experiment confirmed the results of previous studies showing that the average F_0 separation is beneficial for speech intelligibility. However, compared to what had previously been reported, the effect was more moderate and observable only when the target sentence was substantially lower in level than the masker sentence (i.e., at negative TMRs). It was shown that the synchrony between the competing speech signals (measured as the time alignment of their F_0 trajectories, i.e., their 'periodicity synchrony') strongly influenced the magnitude of the benefit induced by the average F_0

separation. In fact, the improvement in speech intelligibility induced by the average F_0 separation was present only when the two competing sentences had high levels of periodicity synchrony. Such levels of periodicity synchrony rarely occur between HINT sentences (and likely also in realistic speech) but are typical of the speech stimuli used in some previous studies that found a much stronger effect of the average F_0 separation on speech intelligibility (likely enhanced by the choice of highly synchronous speech material). Periodicity synchrony might be a consequence of the synchrony in voice activity between competing sentences, but despite extensive analyses conducted on the speech stimuli it was not possible to disentangle these two aspects. However, the low levels of periodicity and voice-activity synchrony found in HINT are a consequence of the wide linguistic and syntactical variety that is typical of realistic speech, which was limited by the strict constraints imposed on the linguistic content of the speech material employed in earlier studies. These constraints also limited the availability of linguistic cues, such as context and syntax, which are naturally present in realistic speech, but whose absence in previous studies may have contributed to enhancing the effects of the average F_0 separation.

In Chapter 4, the effects of F_0-dynamic-range contrast on speech intelligibility for NH listeners were investigated. The study revealed that, when using the range of F_0 variety that is typical of realistic speech, the F_0-dynamic-range contrast between competing sentences has a negligible effect on speech intelligibility. However, target speech was found less intelligible when both competing sentences had relatively small F_0 dynamic ranges and became easier to understand when at least one of the sentences exhibited stronger F_0 dynamics, regardless of the contrast between them. In fact, the F_0 dynamic range of the individual target and masker sentences was found to have a major effect on target-speech intelligibility, even stronger than that of an average F_0 separation. This suggested that, when the F_0 dynamics of at least one of the sentences is sufficiently large, the listener is able to discriminate the temporal patterns of the competing F_0 trajectories and exploit their differences for disentangling them. The results of the study reported in Chapter 4 are in disagreement with the only previous study (Calandruccio et al., 2019) that investigated the role of the F_0-dynamic-range contrast in competing-talker scenarios and found a beneficial effect of this auditory cue on speech intelligibility. However, a direct comparison between the results obtained in the two studies is difficult because they utilized different experimental paradigms: in the study presented in Chap-

ter 4, the interfering speech was a single-sentence whose onset was aligned to the onset of the target and the employed F_0 manipulation allowed to reproduce the F_0 statistics found in real speech, whereas the study by Calandruccio et al. (2019) employed a two-talker masker without coherent syntax or meaning that started before and ended after the target and, in particular, their method for controlling the F_0 trajectories produced speech with F_0 statistics beyond the range found in realistic voices and may have affected other properties of speech beyond the F_0.

Chapter 5 described an investigation of the effects of average F_0 separation and F_0-dynamic-range contrast on speech intelligibility in HI listeners, with the same speech material and experimental paradigm as used in the corresponding studies conducted with NH listeners (Chapters 3 and 4). The study confirmed the findings from previous studies suggesting that these F_0-related cues have limited effects on speech intelligibility in HI listeners as compared to the effects observed in NH listeners. The average F_0 separation was found to be beneficial also for HI listeners, albeit with a more moderate effect than for NH listeners. However, for HI listeners, this effect was observed at a TMR of 0 dB (which may often occur in a conversational real-life situation), whereas for NH listeners, a benefit of the average F_0 separation cue was only found at negative TMRs (which are less likely to occur in daily conversational situations). The F_0-dynamic-range contrast, as for NH listeners, had no effect on speech intelligibility in HI listeners. As for the F_0 dynamic range of the individual sentences, no effect was found in HI listeners, indicating that hearing impairment might hinder the ability to perceive and exploit the F_0 variations found in realistic speech, whereas such F_0 variations appear to be highly beneficial for NH listeners (as shown in Chapter 4).

Finally, Chapter 6 presented an investigation of how talkers engaged in dialogues modify the F_0 of their voice in the presence of communication barriers such as background noise and hearing-impairment in one of the interlocutors. The study showed that both NH and HI listeners increase the average and the dynamic range of their F_0 when speaking in background noise. Furthermore, a group of NH talkers speaking with HI interlocutors showed a higher average F_0 and also a higher F_0 dynamic range than another group of NH talkers speaking with NH interlocutors. These results indicate that NH talkers, in response to the hearing difficulty of their interlocutors, increase their vocal effort to improve the intelligibility of their voice. However, on an individual basis, not every

talker modified the average F_0 and the F_0 dynamic range concurrently. It seems therefore that, in response to the presence of communication barriers that may limit the comprehension of the message to be conveyed, different talkers apply different strategies, such as (i) increasing their average F_0, (ii) enhancing the variations in F_0 (i.e., the F_0 dynamic range), or (iii) doing both. The changes in F_0 production generated by the talker in adverse communicative situations (in response to the presence of background noise and/or a hearing-impairment of the interlocutor) are done towards higher values of F_0 dynamic range (in some cases accompanied by higher average F_0) that were shown to provide higher levels of speech intelligibility in Chapter 4. It seems that the talkers, who themselves experience the effects of enhanced F_0 information when holding the role of the listener in the communication scenario, naturally tend to modify their F_0 to produce speech which they would perceive as more intelligible.

7.2 Pros and cons of the experimental approach

In comparison to the previous research, this work offers two main advantages. First, the use of a speech material that is characterized by (i) a wide variety of voices and F_0 characteristics (quantified in terms of its long-term statistics) that have been shown to be a good representation of the F_0 characteristics occurring in real-life speech and (ii) the presence of linguistic cues that are typically present in realistic speech. Second, an experimental design where the main experimental variables (i.e., the F_0 information of the speech signals) are numerically controlled and their manipulation does not affect other attributes of speech. While these aspects were not always considered in previous research, in the experiments presented in this thesis they revealed that F_0-related cues, such as average F_0 separation and F_0-dynamic-range contrast, have a moderate or negligible effect in competing-talker scenarios, especially when other auditory and linguistic cues that are typical of many realistic conversational situations are also available. Rather than the differences in average F_0 and F_0 dynamic range between the competing sentences, the magnitude of the F_0 dynamics of the individual sentences provided the most beneficial effect on speech intelligibility, at least in NH listeners.

One of the aims of this thesis was to extend the investigation of F_0-related cues towards more realistic speech materials and experimental scenarios. Between a highly controlled laboratory settings and real-life competing-talker

scenarios there exists a continuum that can be explored along different dimensions by selecting which aspects of realistic scenarios to replicate experimentally and the constraints applied to them. The approach proposed in this thesis allowed to explore parts of this continuum that have not been explored before and the reported experimental findings contribute to the body of knowledge of how F_0-related cues can aid speech intelligibility in competing-talker scenarios.

However, the experimental approach presented here, despite presenting potential advancements compared to previously employed methods, also lacks certain aspects of realistic auditory situations and cannot be considered a faithful representation of naturalistic competing-talker scenarios. In particular, the use of the same co-located talker in each pair of sentences removed aspects of realistic acoustic scenarios that may influence speech perception considerably, such as the spatial separation between competing talkers and differences between the voices other than the F_0 (for example, differences in vocal tract properties and long-term average speech spectra). Furthermore, the employed paradigm covers only a subset of realistic situations, where the target is a sentence spoken by a voice that is unfamiliar to the listener and that is masked by a single sentence that begins simultaneously.

7.3 Perspectives

Inspired by the results presented in this thesis, future research could be conducted in several directions. The experimental paradigm employed in the three experiments can be modified to include continuous, running maskers instead of single interfering sentences that have the same onset as the target. This may be of interest given that a continuous interfering speech signal might represent a more typical situation in real life competing-talker scenarios. Furthermore, the number of masking speech signals can be varied to measure how multiple interfering voices (and their F_0 trajectories) affect speech intelligibility. As suggested in Chapter 4, experiments can be designed to explore how the (dis)similarity in the temporal pattern of the competing F_0 trajectories can influence speech intelligibility. Such investigation would require the development of a metric that quantifies this dissimilarity (beyond average separation and dynamic range difference) between F_0 trajectories that can be experimentally controlled to create stimuli with different levels of such metric and that can be related to their speech intelligibility. The results of such an experiment may also be com-

bined with measures of the listener's discrimination abilities of F_0 trajectories for different degrees of dissimilarity.

In relation to the investigation of HI listeners, future research might assess how individual supra-threshold deficits influence the ability to utilize F_0-related cues. Additional research may be directed to explore how state-of-the-art hearing-aid technology influences the availability and utility of F_0 information. Furthermore, as reported in Chapter 6, NH individuals may tend to enhance the average F_0 and F_0 dynamic range of their voice well above their conversational values when speaking with HI interlocutors, likely because such extreme values facilitate speech perception. This hypothesis requires more solid experimental evidence than that presented in Chapter 6. However, if confirmed, it may inspire future research directed to assessing how values of average F_0, F_0 dynamic range and their differences between competing voices can influence speech intelligibility in HI listeners, if manipulated to exceed the range of values found in realistic conversational voices, measured in NH talkers when speaking with NH interlocutors in quiet. Such investigations, combined with an assessment of how extreme manipulations of the F_0 affect the perceived sound quality, may provide valuable insights for the development of novel hearing-aid processing strategies aimed at enhancing speech perception.

A

Appendix: Details of the speech materials

For each speech material described in Section 2.1, the number of talkers for which the recordings were available and the time durations of the recordings are reported in Table A.1. The time durations were calculated as average over the sentences for HINT, CRM, BKB (Talker A only, since for talker B and C, recordings of separate sentences were not available) or over the recording sessions of naturalistic dialogues for talkers NH1-NH1, NH2-HI and HI-NH2. Average durations for each talker and overall duration of the recordings in each speech material were also calculated.

Table A.1: Time durations of the different speech materials analyzed, described in Section 2.1. Durations are shown as: averages across sentences (for HINT, CRM, 1-talker BKB materials) or per recording session (for recordings of naturalistic dialogues) from all talkers; averages across talkers; overall durations computed over the entire speech material. Values in parenthesis indicate standard deviations.

Speech material	**Number of talkers**	**Average per sentence/rec session**	**Average per talker**	**Overall duration**
HINT	12	1.5" (0.2")	4':51" (20")	58':17"
CRM	8	1.8" (0.1")	7':42" (25")	1h:1':32"
BKB Talker A (flat)	1	1.4" (0.2")	7':46" (0")	7':46"
BKB Talker A (normal)	1	1.3" (0.2")	7':18" (0")	7':18"
BKB Talker A (exaggerated)	1	1.7" (0.2")	9':28" (0")	9':28"
BKB Talker BC (flat)	Mixture of 2	N/A	1':18" (0")	1':18"
BKB Talker BC (normal)	Mixture of 2	N/A	1':18" (0")	1':18"
BKB Talker BC (exaggerated)	Mixture of 2	N/A	1':22" (0")	1':22"
NH1-NH1 dialogues (quiet)	38	5':8" (1':17")	15':7" (3':15")	9h:34':20"
NH1-NH1 dialogues (70-dBA noise)	38	5':29" (1':33")	16':28" (3':18")	10h:25':56"
NH2-HI dialogues (quiet)	12	5':50" (1':53")	17':31" (4':47")	3h:30':14"
NH2-HI dialogues (60-dBA noise)	12	6':11" (2':16")	18':32" (5':15")	3h:42':20"
NH2-HI dialogues (65-dBA noise)	12	6':27" (2':6")	19':20" (5':40")	3h:52':3"
NH2-HI dialogues (70-dBA noise)	12	6':48" (2':12")	20':21" (6':17")	4h:4':37"
HI-NH2 dialogues (quiet)	12	5':50" (1':53")	17':31" (4':47")	3h:30':14"
HI-NH2 dialogues (60-dBA noise)	12	6':11" (2':16")	18':32" (5':15")	3h:42':20"
HI-NH2 dialogues (65-dBA noise)	12	6':27" (2':6")	19':20" (5':40")	3h:52':3"
HI-NH2 dialogues (70-dBA noise)	12	6':48" (2':12")	20':21" (6':17")	4h:4':37"

B

Appendix: Alternative measures of fundamental frequency

The F_0 of the voice can convey a lot of information about the speech signal. Not only its average value, but also the dynamics of the F_0, provide information on the linguistic and prosodic content of the speech message, the intention, the emotional state and the gender of the talker (Arnott, 1993; Honorof and Whalen, 2010; Ladd, 2008; O'Shaughnessy, 1979). Metrics that quantify the dynamics of the F_0 can be useful for a variety of applications and purposes that include scientific research in speech production (Arnott, 1993; Baker and Hazan, 2011; Picheny et al., 1986) and perception (Assmann, 1999; Brokx and Nooteboom, 1982; Calandruccio et al., 2019; Darwin et al., 2003), clinical applications such as the diagnosis and prevention of speech pathologies and brain or cognitive disorders (Harel et al., 2004; Lieberman, 1963; Meilán et al., 2012), as well as forensic purposes (Jessen et al., 2005; Künzel et al., 1995).

This thesis focused on how differences in F_0 (measured as long-term averages) between competing voices can aid speech intelligibility. Long-term descriptors of the F_0 of speech, namely its time average and dynamic range, were used for this purpose. Further measurements of F_0 dynamics and its statistics, that can be useful for quantifying and studying aspects of speech production and perception, were considered during the work of this thesis, but not fully developed due to lack of time and eventually not utilized. Nevertheless, these measures may be useful for future studies involving the production and/or perception of the F_0 and its properties as they can characterize differences between talkers, their speaking styles and their emotional states, and are therefore reported in this appendix.

A simple measurement of F_0 dynamics can be, e.g., the average change of F_0 between consecutive time frames in a F_0 trajectory. However, this measure is largely affected by the sampling period used for the estimate of the F_0 trajectory (the lower the sampling period, the smaller the F_0 changes measured between

consecutive time frames). An alternative measure that would overcome this limitation is the average speed of F_0 changes, which considers the change in F_0 relatively to the time frame within which it occurs. Both the average change in F_0 and the average speed of F_0 change are metrics that can be measured either on a linear (Hertz-based) or on a logarithmic (e.g., semitone-based) scale. While a Hertz-based scale would allow to measure the physical change in F_0, a logarithmic scale would allow to measure the magnitude of these changes in relation to the value at which they occur. Considering that the perception of pitch (i.e., the perceptual counterpart of the F_0) operates along such a relative, quasi-logarithmic, frequency-axis (Bianchi et al., 2016), the logarithmic measures of the change in F_0 would provide a measure that may be more relevant when considering perceptual aspects.

A more complex metric that can describe the dynamics of the F_0 is the 'length' of the F_0 trajectory: the F_0 trajectory is a discrete time series and its length can be measured as the sum of absolute values of the F_0 variations, that is the length of the path 'roved' by the F_0 during the speech signal ($length = \sum_{i=2}^{N}\left[\left(F_{0_i} - F_{0_{i-1}}\right)^2 + (t_i - t_{i-1})^2\right]^{1/2}$, where F_{0_i} is the F_0 value measured at the time sample t_i and N the number of samples in the trajectory). This measure can provide more information about the F_0 dynamics than a simpler measure of the dynamic range, such as the standard deviation or the median absolute deviation, as it takes into consideration all the instantaneous variations of the F_0 rather than its overall spread from the median value. However, this measure is influenced by the duration of the speech signal and thus requires to be normalized. A possible normalization method that is suggested for this measure consists in dividing the measured length by the time duration of the F_0 trajectory, that is the length of a hypothetical constant-F_0, fully-voiced trajectory (i.e., with no dynamics and uninterrupted by unvoiced segments), corresponding to the shortest path that can be 'roved' by the F_0 within the same time duration. While the suggested 'length' measure seems ideal for capturing the 'fine structure' of F_0 variations that are not captured by long-term statistics such as the median absolute deviation or the standard deviation, it can be strongly affected by erroneous estimates of F_0, like the octave jumps described in Section 2.2.1, and would require a reliable method for identifying such errors so that they can be excluded from the measure of F_0-trajectory length. In absence of such method, the impact of octave jumps on the 'length' measure can be limited by excluding from the sum F_0 variations that occur too fast (for example, faster

than a number of semitones per second, as described in Section 2.2.1).

Bibliography

Akaike, H. (1998). "Information theory and an extension of the maximum likelihood principle". In: *Selected papers of hirotugu akaike*. Springer, pp. 199–213.

Arehart, K. H., C. A. King, and K. S. McLean-Mudgett (1997). "Role of fundamental frequency differences in the perceptual separation of competing vowel sounds by listeners with normal hearing and listeners with hearing loss". In: *Journal of Speech, Language, and Hearing Research* 40.6, pp. 1434–1444.

Arnott, M. (1993). "Towards the simulation of emotion in synthetic speech". In: *Journal Acoustical Society of Speech* 93.2, pp. 1097–1108.

Assmann, P. F. (1999). "Fundamental frequency and the intelligibility of competing voices". In: *Proceedings of the 14th International Congress of Phonetic Sciences*. University of California Press Oakland, CA, pp. 179–182.

Baker, R. and V. Hazan (2011). "DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs". In: *Behavior research methods* 43.3, pp. 761–770.

Balakrishnan, U. and R. L. Freyman (2008). "Speech detection in spatial and nonspatial speech maskers". In: *The Journal of the Acoustical Society of America* 123.5, pp. 2680–2691.

Başkent, D. and E. Gaudrain (2016). "Musician advantage for speech-on-speech perception". In: *The Journal of the Acoustical Society of America* 139.3, EL51–EL56.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2014). "Fitting linear mixed-effects models using lme4". In: *arXiv preprint arXiv:1406.5823*.

Bench, J., Å. Kowal, and J. Bamford (1979). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children". In: *British journal of audiology* 13.3, pp. 108–112.

Best, V., C. R. Mason, and G. Kidd Jr (2011). "Spatial release from masking in normally hearing and hearing-impaired listeners as a function of the tem-

poral overlap of competing talkers". In: *The Journal of the Acoustical Society of America* 129.3, pp. 1616–1625.

Bianchi, F., S. Santurette, D. Wendt, and T. Dau (2016). "Pitch discrimination in musicians and non-musicians: Effects of harmonic resolvability and processing effort". In: *Journal of the Association for Research in Otolaryngology* 17.1, pp. 69–79.

Binns, C. and J. F. Culling (2007). "The role of fundamental frequency contours in the perception of speech against interfering speech". In: *The Journal of the Acoustical Society of America* 122.3, pp. 1765–1776.

Boersma, P. et al. (1993). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound". In: *Proceedings of the institute of phonetic sciences*. Vol. 17. 1193. Citeseer, pp. 97–110.

Bolia, R. S., W. T. Nelson, M. A. Ericson, and B. D. Simpson (2000). "A speech corpus for multitalker communications research". In: *The Journal of the Acoustical Society of America* 107.2, pp. 1065–1066.

Boothroyd, A. and S. Nittrouer (1988). "Mathematical treatment of context effects in phoneme and word recognition". In: *The Journal of the Acoustical Society of America* 84.1, pp. 101–114.

Bramsløw, L., G. Naithani, A. Hafez, T. Barker, N. H. Pontoppidan, and T. Virtanen (2018). "Improving competing voices segregation for hearing impaired listeners using a low-latency deep neural network algorithm". In: *The Journal of the Acoustical Society of America* 144.1, pp. 172–185.

Bramsløw, L., M. Vatti, R. K. Hietkamp, and N. H. Pontoppidan (2015). "Binaural speech recognition for normal-hearing and hearing-impaired listeners in a competing voice test". In: *Poster presented at the Speech in Noise workshop, Copenhagen, Denmark*.

Bramsløw, L., M. Vatti, R. Rossing, G. Naithani, and N. Henrik Pontoppidan (2019). "A Competing Voices Test for hearing-impaired listeners applied to spatial separation and ideal time-frequency masks". In: *Trends in hearing* 23, p. 2331216519848288.

Brokx, J. P. L. and S. G. Nooteboom (1982). "Intonation and the perceptual separation of simultaneous voices". In: *Journal of Phonetics* 10.1, pp. 23–36.

Bronkhorst, A. and R. Plomp (1992). "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing". In: *The Journal of the Acoustical Society of America* 92.6, pp. 3132–3139.

Bronkhorst, A. W. (2000). “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions”. In: *Acta Acustica united with Acustica* 86.1, pp. 117–128.

Brumm, H. and S. A. Zollinger (2011). “The evolution of the Lombard effect: 100 years of psychoacoustic research”. In: *Behaviour* 148.11-13, pp. 1173–1198.

Brungart, D. S. (2001). “Informational and energetic masking effects in the perception of two simultaneous talkers”. In: *The Journal of the Acoustical Society of America* 109.3, pp. 1101–1109.

Brungart, D. S., B. D. Simpson, M. A. Ericson, and K. R. Scott (2001). “Informational and energetic masking effects in the perception of multiple simultaneous talkers”. In: *The Journal of the Acoustical Society of America* 110.5, pp. 2527–2538.

Calandruccio, L. et al. (2019). “The effect of target/masker fundamental frequency contour similarity on masked-speech recognition”. In: *The Journal of the Acoustical Society of America* 146.2, pp. 1065–1076.

Cherry, E. C. (1953). “Some experiments on the recognition of speech, with one and with two ears”. In: *The Journal of the acoustical society of America* 25.5, pp. 975–979.

Cheveigné, A. de, S. McAdams, J. Laroche, and M. Rosenberg (1995). “Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement”. In: *The Journal of the Acoustical Society of America* 97.6, pp. 3736–3748.

Cooke, M., S. King, M. Garnier, and V. Aubanel (2014). “The listening talker: A review of human and algorithmic context-induced modifications of speech”. In: *Computer Speech & Language* 28.2, pp. 543–571.

Culling, J. F., K. I. Hodder, and C. Y. Toh (2003). “Effects of reverberation on perceptual segregation of competing voices”. In: *The Journal of the Acoustical Society of America* 114.5, pp. 2871–2876.

Culling, J. F., Q. Summerfield, and D. H. Marshall (1994). “Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels”. In: *Speech Communication* 14.1, pp. 71–95.

Darwin, C. and R. Hukin (2000a). “Effectiveness of spatial cues, prosody, and talker characteristics in selective attention”. In: *The Journal of the Acoustical Society of America* 107.2, pp. 970–977.

Darwin, C. and R. Hukin (2000b). “Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention”. In: *The Journal of the Acoustical Society of America* 108.1, pp. 335–342.

Darwin, C. J., D. S. Brungart, and B. D. Simpson (2003). “Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers”. In: *The Journal of the Acoustical Society of America* 114.5, pp. 2913–2922.

De Looze, C. and D. Hirst (2008). “Detecting changes in key and range for the automatic modelling and coding of intonation”. In: *Speech Prosody*, pp. 135–138.

Dreschler, W. A., H. Verschuure, C. Ludvigsen, and S. Westermann (2001). “ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment: Ruidos ICRA: Señates de ruido artificial con espectro similar al habla y propiedades temporales para pruebas de instrumentos auditivos”. In: *Audiology* 40.3, pp. 148–157.

Duquesnoy, A. (1983). “Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons”. In: *The Journal of the Acoustical Society of America* 74.3, pp. 739–743.

Ezzatian, P., L. Li, K. Pichora-Fuller, and B. A. Schneider (2015). “Delayed stream segregation in older adults: More than just informational masking”. In: *Ear and Hearing* 36.4, pp. 482–484.

Fernald, A. and T. Simon (1984). “Expanded intonation contours in mothers’ speech to newborns.” In: *Developmental psychology* 20.1, p. 104.

Festen, J. M. and R. Plomp (1990). “Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing”. In: *The Journal of the Acoustical Society of America* 88.4, pp. 1725–1736.

Flaherty, M. M., E. Buss, and L. J. Leibold (2021). “Independent and combined effects of fundamental frequency and vocal tract length differences for school-age children’s sentence recognition in a two-talker masker”. In: *Journal of Speech, Language, and Hearing Research* 64.1, pp. 206–217.

Freyman, R. L., U. Balakrishnan, and K. S. Helfer (2001). “Spatial release from informational masking in speech recognition”. In: *The Journal of the Acoustical Society of America* 109.5, pp. 2112–2122.

Freyman, R. L., U. Balakrishnan, and K. S. Helfer (2004). “Effect of number of masking talkers and auditory priming on informational masking in speech

recognition". In: *The Journal of the Acoustical Society of America* 115.5, pp. 2246–2256.

Freyman, R. L., K. S. Helfer, D. D. McCall, and R. K. Clifton (1999). "The role of perceived spatial separation in the unmasking of speech". In: *The Journal of the Acoustical Society of America* 106.6, pp. 3578–3588.

Garnier, M. and N. Henrich (2014). "Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?" In: *Computer Speech & Language* 28.2, pp. 580–597.

George, E. L., J. M. Festen, and T. Houtgast (2006). "Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners". In: *The Journal of the Acoustical society of America* 120.4, pp. 2295–2311.

Grant, K. W. (1987). "Identification of intonation contours by normally hearing and profoundly hearing-impaired listeners". In: *The Journal of the Acoustical Society of America* 82.4, pp. 1172–1178.

Grieser, D. L. and P. K. Kuhl (1988). "Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese." In: *Developmental psychology* 24.1, p. 14.

Harel, B., M. Cannizzaro, and P. J. Snyder (2004). "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study". In: *Brain and cognition* 56.1, pp. 24–29.

Hawley, M. L., R. Y. Litovsky, and J. F. Culling (2004). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer". In: *The Journal of the Acoustical Society of America* 115.2, pp. 833–843.

Hee Lee, J. and L. E. Humes (2012). "Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background". In: *The Journal of the Acoustical Society of America* 132.3, pp. 1700–1717.

Helfer, K. S. and R. L. Freyman (2008). "Aging and speech-on-speech masking". In: *Ear and hearing* 29.1, p. 87.

Helfer, K. S. and R. L. Freyman (2014). "Stimulus and listener factors affecting age-related changes in competing speech perception". In: *The Journal of the Acoustical Society of America* 136.2, pp. 748–759.

Honorof, D. N. and D. Whalen (2010). "Identification of speaker sex from one vowel across a range of fundamental frequencies". In: *The Journal of the Acoustical Society of America* 128.5, pp. 3095–3104.

Humes, L. E., J. H. Lee, and M. P. Coughlin (2006). "Auditory measures of selective and divided attention in young and older adults using single-talker competition". In: *The Journal of the Acoustical Society of America* 120.5, pp. 2926–2937.

Jessen, M., O. Koster, and S. Gfroerer (2005). "Influence of vocal effort on average and variability of fundamental frequency". In: *International Journal of Speech, Language and the Law* 12.2, pp. 174–213.

Kidd Jr, G., C. R. Mason, J. Swaminathan, E. Roverud, K. K. Clayton, and V. Best (2016). "Determining the energetic and informational components of speech-on-speech masking". In: *The Journal of the Acoustical Society of America* 140.1, pp. 132–144.

Kidd Jr, G. et al. (2019). "Determining the energetic and informational components of speech-on-speech masking in listeners with sensorineural hearing loss". In: *The Journal of the Acoustical society of America* 145.1, pp. 440–457.

Kochkin, S. (2002). "Consumers rate improvements sought in hearing instruments". In: *Hear Rev* 9.11, pp. 18–22.

Koelewijn, T., A. A. Zekveld, J. M. Festen, and S. E. Kramer (2014). "The influence of informational masking on speech perception and pupil response in adults with hearing impairment". In: *The Journal of the Acoustical Society of America* 135.3, pp. 1596–1606.

Krause, J. C. and L. D. Braida (2004). "Acoustic properties of naturally produced clear speech at normal speaking rates". In: *The Journal of the Acoustical Society of America* 115.1, pp. 362–378.

Künzel, H. J., H. Masthoff, and J.-P. Köster (1995). "The relation between speech tempo, loudness, and fundamental frequency: An important issue in forensic speaker recognition." In: *Science & Justice: Journal of the Forensic Science Society* 35.4, pp. 291–295.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

Laures, J. S. and K. Bunton (2003). "Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions". In: *Journal of communication disorders* 36.6, pp. 449–464.

Laures, J. S. and G. Weismer (1999). "The effects of a flattened fundamental frequency on intelligibility at the sentence level". In: *Journal of Speech, Language, and Hearing Research* 42.5, pp. 1148–1156.

Lieberman, P. (1963). “Some acoustic measures of the fundamental periodicity of normal and pathologic larynges”. In: *The Journal of the Acoustical Society of America* 35.3, pp. 344–353.

Lőcsei, G., J. H. Pedersen, S. Laugesen, S. Santurette, T. Dau, and E. N. MacDonald (2016). “Temporal fine-structure coding and lateralized speech perception in normal-hearing and hearing-impaired listeners”. In: *Trends in Hearing* 20, p. 2331216516660962.

Lombard, E. (1911). “Le signe de l’elevation de la voix”. In: *Ann. Mal. de L’Oreille et du Larynx*, pp. 101–119.

Mackersie, C. L., J. Dewey, and L. A. Guthrie (2011). “Effects of fundamental frequency and vocal-tract length cues on sentence segregation by listeners with hearing loss”. In: *The Journal of the Acoustical Society of America* 130.2, pp. 1006–1019.

Marrone, N., C. R. Mason, and G. Kidd Jr (2008). “The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms”. In: *The Journal of the Acoustical Society of America* 124.5, pp. 3064–3075.

Meilán, J. J., F. Martínez-Sánchez, J. Carro, J. A. Sánchez, and E. Pérez (2012). “Acoustic markers associated with impairment in language processing in Alzheimer’s disease”. In: *The Spanish journal of psychology* 15.2, pp. 487–494.

Mesiano, P. A., J. Zaar, H. Relaño Iborra, L. Bramsløw, and T. Dau (2022a). “Effects of fundamental-frequency differences between competing sentences on speech intelligibility for listeners with hearing impairment”. Unpublished manuscript.

Mesiano, P. A., J. Zaar, H. Relaño Iborra, L. Bramsløw, and T. Dau (2022b). “Effects of fundamental-frequency dynamics on sentence intelligibility in competing-talker scenarios”. Unpublished manuscript.

Mesiano, P. A., J. Zaar, H. Relaño Iborra, L. Bramsløw, and T. Dau (2022c). “The role of average fundamental frequency difference on the intelligibility of real-life competing sentences”. Manuscript submitted for publication.

Miller, G. A. (1947). “The masking of speech.” In: *Psychological bulletin* 44.2.

Miller, S. E., R. S. Schlauch, and P. J. Watson (2010). “The effects of fundamental frequency contour manipulations on speech intelligibility in background noise”. In: *The Journal of the Acoustical Society of America* 128.1, pp. 435–443.

Moore, B. C. and B. R. Glasberg (1998). “Use of a loudness model for hearing-aid fitting. I. Linear hearing aids”. In: *British journal of audiology* 32.5, pp. 317–335.

Moulines, E. and F. Charpentier (1990). “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”. In: *Speech communication* 9.5-6, pp. 453–467.

Neher, T., T. Behrens, L. Kragelund, and A. S. Petersen (2007). “Spatial unmasking in aided hearing-impaired listeners and the need for training”. In: *Proceedings of the International Symposium on Auditory and Audiological Research*. Vol. 1, pp. 515–522.

Nielsen, J. B. and T. Dau (2011). “The Danish hearing in noise test”. In: *International journal of audiology* 50.3, pp. 202–208.

O’Shaughnessy, D. (1979). “Linguistic features in fundamental frequency patterns”. In: *Journal of Phonetics* 7.2, pp. 119–145.

Payton, K. L., R. M. Uchanski, and L. D. Braida (1994). “Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing”. In: *The Journal of the Acoustical Society of America* 95.3, pp. 1581–1592.

Peelle, J. E. (2018). “Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior”. In: *Ear and hearing* 39.2, p. 204.

Picheny, M. A., N. I. Durlach, and L. D. Braida (1985). “Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech”. In: *Journal of Speech, Language, and Hearing Research* 28.1, pp. 96–103.

Picheny, M. A., N. I. Durlach, and L. D. Braida (1986). “Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech”. In: *Journal of Speech, Language, and Hearing Research* 29.4, pp. 434–446.

Plomp, R (1976). “Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise)”. In: *Acta Acustica united with Acustica* 34.4, pp. 200–211.

Prud’Homme, L., M. Lavandier, and V. Best (2020). “A harmonic-cancellation-based model to predict speech intelligibility against a harmonic masker”. In: *The Journal of the Acoustical Society of America* 148.5, pp. 3246–3254.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rosen, S., P. Souza, C. Ekelund, and A. A. Majeed (2013). "Listening to speech in a background of other talkers: Effects of talker number and noise vocoding". In: *The Journal of the Acoustical Society of America* 133.4, pp. 2431–2443.

Rossi-Katz, J. and K. H. Arehart (2009). "Message and talker identification in older adults: Effects of task, distinctiveness of the talkers' voices, and meaningfulness of the competing message". In.

Rothauser, E. (1969). "IEEE recommended practice for speech quality measurements". In: *IEEE Trans. on Audio and Electroacoustics* 17, pp. 225–246.

Shen, J. and P. E. Souza (2017). "Do older listeners with hearing loss benefit from dynamic pitch for speech recognition in noise?" In: *American Journal of Audiology* 26.3S, pp. 462–466.

Smeds, K., F. Wolters, and M. Rung (2015). "Estimation of signal-to-noise ratios in realistic sound scenarios". In: *Journal of the American Academy of Audiology* 26.02, pp. 183–196.

Sørensen, A. J. M., M. Fereczkowski, and E. N. MacDonald (2021). "Effects of Noise and Second Language on Conversational Dynamics in Task Dialogue". In: *Trends in Hearing* 25, p. 23312165211024482.

Sørensen, A. J. M., E. N. MacDonald, and T. Lunner (2019). "Timing of turn taking between normal-hearing and hearing-impaired interlocutors". In: *Proceedings of the International Symposium on Auditory and Audiological Research*. Vol. 7, pp. 37–44.

Steinmetzger, K. and S. Rosen (2015). "The role of periodicity in perceiving speech in quiet and in background noise". In: *The Journal of the Acoustical Society of America* 138.6, pp. 3586–3599.

Stern, D. N., S. Spieker, R. Barnett, and K. MacKain (1983). "The prosody of maternal speech: Infant age and context related changes". In: *Journal of child language* 10.1, pp. 1–15.

Summers, V. and M. R. Leek (1998). "F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss". In: *Journal of Speech, Language, and Hearing Research* 41.6, pp. 1294–1306.

Summers, V. and M. R. Molis (2004). "Speech recognition in fluctuating and continuous maskers". In.

Summers, W. V., D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes (1988). "Effects of noise on speech production: Acoustic and perceptual analyses". In: *The Journal of the Acoustical Society of America* 84.3, pp. 917–928.

Uchanski, R. (2005). *Clear speech. The handbook of speech perception, ed. by DB Pisoni and R. Remez, 207–35. Malden, MA.*

Viveros Muñoz, R., L. Aspöck, and J. Fels (2019). "Spatial release from masking under different reverberant conditions in young and elderly subjects: Effect of moving or stationary maskers at circular and radial conditions". In: *Journal of Speech, Language, and Hearing Research* 62.9, pp. 3582–3595.

Wasiuk, P. A., M. Lavandier, E. Buss, J. Oleson, and L. Calandruccio (2020). "The effect of fundamental frequency contour similarity on multi-talker listening in older and younger adults". In: *The Journal of the Acoustical Society of America* 148.6, pp. 3527–3543.

Yost, W. A., R. H. Dye, and S. Sheft (1996). "A simulated "cocktail party" with up to three sound sources". In: *Perception & psychophysics* 58.7, pp. 1026–1036.

Contributions to Hearing Research

Vol. 1: *Gilles Pigasse*, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.
External examiners: Mark Lutman, Stefan Stenfeld

Vol. 2: *Olaf Strelcyk*, Peripheral auditory processing and speech reception in impaired hearing, 2009.
External examiners: Brian Moore, Kathrin Krumbholz

Vol. 3: *Eric R. Thompson*, Characterizing binaural processing of amplitude-modulated sounds, 2009.
External examiners: Michael Akeroyd, Armin Kohlrausch

Vol. 4: *Tobias Piechowiak*, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.
External examiners: Jesko Verhey, Steven van de Par

Vol. 5: *Jens Bo Nielsen*, Assessment of speech intelligibility in background noise and reverberation, 2009.
External examiners: Björn Hagerman, Ejnar Laukli

Vol. 6: *Helen Connor*, Hearing aid amplification at soft input levels, 2010.
External examiners: Inga Holube, Birgitta Larsby

Vol. 7: *Morten Løve Jepsen*, Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.
External examiners: Birger Kollmeier, Ray Meddis

Vol. 8: *Sarah Verhulst*, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.
External examiners: David Kemp, Stephen Neely

Vol. 9: *Sylvain Favrot*, A loudspeaker-based room auralization system for auditory research, 2010.
External examiners: Bernhard Seeber, Michael Vorländer

Vol. 10: *Sébastien Santurette*, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.
External examiners: Christopher Plack, Christian Lorenzi

Vol. 11: *Iris Arweiler*, Processing of spatial sounds in the impaired auditory system, 2011.
External examiners: Joost Festen, Jürgen Tchorz

Vol. 12: *Filip Munch Rønne*, Modeling auditory evoked potentials to complex stimuli, 2012.
External examiners: Bob Burkard, Stephen Neely

Vol. 13: *Claus Forup Corlin Jespersgaard*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.
External examiners: Stuart Rosen, Christian Lorenzi

Vol. 14: *Rémi Decorsière*, Spectrogram inversion and potential applications for hearing research, 2013.
External examiners: Michael Stone, Oded Ghitza

Vol. 15: *Søren Jørgensen*, Modeling speech intelligibility based on the signal-to-noise envelope power ration, 2014.
External examiners: John Culling, Martin Cooke

Vol. 16: *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.
External examiners: Lawrence Rosenblum, Matthias Gondan

Vol. 17: *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.
External examiners: Shihab Shamma, Guy Brown

Vol. 18: *Márton Marschall*, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.
External examiners: Sascha Spors, Ville Pulkki

Vol. 19: *Jasmina Catic*, Human sound externalization in reverberant environments, 2014.
External examiners: Bernhard Seeber, Steven van de Par

Vol. 20: *Michał Fereczkowski*, Design and evaluation of individualized hearing-aid signal processing and fitting, 2015.
External examiners: Christopher Plack, Enrique Lopez-Poveda

Vol. 21: *Alexandre Chabot-Leclerc*, Computational modeling of speech intelligibility in adverse conditions, 2015.
External examiners: Steven van de Par, John Culling

Vol. 22: *Federica Bianchi*, Pitch representations in the impaired auditory system and implications for music perception, 2016.
External examiners: Ingrid Johnsrude, Christian Lorenzi

Vol. 23: *Johannes Zaar*, Measures and computational models of microscopic speech perception, 2016.
External examiners: Judy Dubno, Martin Cooke

Vol. 24: *Johannes Käsbach*, Characterizing apparent source width perception, 2016.
External examiners: William Whitmer, Jürgen Tchorz

Vol. 25: *Gusztáv Löcsei*, Lateralized speech perception with normal and impaired hearing, 2016.
External examiners: Thomas Brand, Armin Kohlrausch

Vol. 26: *Suyash Narendra Joshi*, Modelling auditory nerve responses to electrical stimulation, 2017.
External examiners: Laurel Carney, Bob Carlyon

Vol. 27: *Henrik Gerd Hassager*, Characterizing perceptual externalization in listeners with normal, impaired and aided-impaired hearing, 2017.
External examiners: Volker Hohmann, Piotr Majdak

Vol. 28: *Richard Ian McWalter*, Analysis of the auditory system via synthesis of natural sounds, speech and music, 2017.
External examiners: Maria Chait, Christian Lorenzi

Vol. 29: *Jens Cubick*, Characterizing the auditory cues for the processing and perception of spatial sounds, 2017.
External examiners: Ville Pulkki, Pavel Zahorik

Vol. 30: *Gerard Encina-Llamas*, Characterizing cochlear hearing impairment using advanced electrophysiological methods, 2017.
External examiners: Roland Schaette, Ian Bruce

Vol. 31: *Christoph Scheidiger*, Assessing speech intelligibility in hearing-impaired listeners, 2018.
External examiners: Enrique Lopez-Poveda, Tim Jürgens

Vol. 32: *Alan Wiinberg*, Perceptual effects of non-linear hearing aid amplification strategies, 2018.
External examiners: Armin Kohlrausch, James Kates

Vol. 33: *Thomas Bentsen*, Computational speech segregation inspired by principles of auditory processing, 2018.
External examiners: Stefan Bleeck, Jürgen Tchorz

Vol. 34: *François Guérit*, Temporal charge interactions in cochlear implant listeners, 2018.
External examiners: Julie Arenberg, Olivier Macherey

Vol. 35: *Andreu Paredes Gallardo*, Behavioral and objective measures of stream segregation in cochlear implant users, 2018.
External examiners: Christophe Micheyl, Monita Chatterjee

Vol. 36: *Søren Fuglsang*, Characterizing neural mechanisms of attention-driven speech processing, 2019.
External examiners: Shihab Shamma, Maarten de Vos

Vol. 37: *Borys Kowalewski*, Assessing the effects of hearing-aid dynamic-range compression on auditory signal processing and perception, 2019.
External examiners: Brian Moore, Graham Naylor

Vol. 38: *Helia Relaño Iborra*, Predicting speech perception of normal-hearing and hearing-impaired listeners, 2019.
External examiners: Ian Bruce, Armin Kohlrausch

Vol. 39: *Axel Ahrens*, Characterizing auditory and audio-visual perception in virtual environments, 2019.
External examiners: Pavel Zahorik, Piotr Majdak

Vol. 40: *Niclas A. Janssen*, Binaural streaming in cochlear implant patients, 2019.
External examiners: Tim Jürgens, Hamish Innes-Brown

Vol. 41: *Wiebke Lamping*, Improving cochlear implant performance through psychophysical measures, 2019.
External examiners: Can Gnasia, David Landsberger

Vol. 42: *Antoine Favre-Félix*, Controlling a hearing aid with electrically assessed eye gaze, 2020.
External examiners: Jürgen Tchorz, Graham Naylor

Vol. 43: *Raul Sanchez Lopez*, Clinical auditory profiling and profile-based hearing-aid fitting, 2020.
External examiners: Judy R. Dubno, Pamela E. Souza

Vol. 44: *Juan Camilo Gil Carvajal*, Modeling audiovisual speech perception, 2020.
External examiners: Salvador Soto-Faraco, Kaisa Maria Tippana

Vol. 45: *Charlotte Amalie Emdal Navntoft*, Improving cochlear implant performance with new pulse shapes: a multidisciplinary approach, 2020.
External examiners: Andrew Kral, Johannes Frijns

Vol. 46: *Naim Mansour*, Assessing hearing device benefit using virtual sound environments, 2021.
External examiners: Virginia Best, Pavel Zahorik

Vol. 47: *Anna Josefine Munch Sørensen*, The effects of noise and hearing loss on conversational dynamics, 2021.
External examiners: William McAllister Whitmer, Martin Cooke

Vol. 48: *Thirsa Huisman*, The influence of vision on spatial localization in normal-hearing and hearing-impaired listeners, 2021.
External examiners: Steven van de Par, Christopher Stecker

Vol. 49: *Florine Lena Bachmann*, Subcortical electrophysiological measures of running speech, 2021.
External examiners: Samira Anderson, Tobias Reichenbach

Vol. 50: *Nicolai Pedersen*, Audiovisual speech analysis with deep learning, 2021.
External examiners: Zheng-Hua Tan, Hani Camille Yehia

Vol. 51: *Aleksandra Koprowska*, Auditory Training Strategies to Improve Speech Intelligibility in Hearing-Impaired Listeners, 2022.
External examiners: Ulrich Hoppe, David Jackson Morris

Vol. 52: *Hyojin Kim*, Physiological correlates of the audibility of masked signals at supra-threshold levels, 2022.
External examiners: Laurel Carney, Jesko L. Verhey

Vol. 53: *Chiara Casolani*, Electrophysiological characterization of tinnitus in listeners with normal audiogram and supra-threshold hearing deficits, 2022.
External examiners: Pim van Dijk, Holger Schulze

Vol. 54: *Mie Lærkegård Jørgensen*, Exploring innovative Hearing Aid Techniques for Tinnitus Treatment, 2022.
External examiners: Pim van Dijk, Tobias Kleinjung

Vol. 55: *Mihaela-Beatrice Neagu*, Evaluation of pupillometry as a diagnostic tool, 2022.
External examiners: Adriana Zekveld, William McAllister Whitmer

The end.

To be continued...

Competing-talker scenarios are challenging auditory situations that are pervasive in daily social life. Hearing loss can represent a major obstacle in these situations, as it can severely limit the participation in conversations. In order to develop engineering solutions that can improve speech intelligibility for hearing-impaired people, it is necessary to understand what acoustic features of the speech signals are relevant for speech intelligibility in such auditory situations.

This thesis focused on how the dynamics of the fundamental frequency of two competing voices can influence the intelligibility of the target speech. Three experiments were conducted on normal-hearing and hearing-impaired listeners. It was shown that the fundamental frequency dynamics are useful for speech intelligibility in competing-talker scenarios, but hearing loss can severely limit their effects. Additionally, an analysis of the fundamental frequency of naturalistic speech was presented, showing how the talkers, when in presence of communication barriers such as background noise or hearing loss, can adapt the fundamental frequency of their voice to promote the intelligibility of their speech for the interlocutor.

Overall, the findings of this thesis can provide useful insights on how potential signal-processing strategies targeted to the fundamental frequency of the speech signals can be implemented to improve speech intelligibility in hearing-impaired listeners.

DTU Health Tech
Department of Health Technology

Ørsteds Plads
Building 352
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 39 50
www.dtu.dk