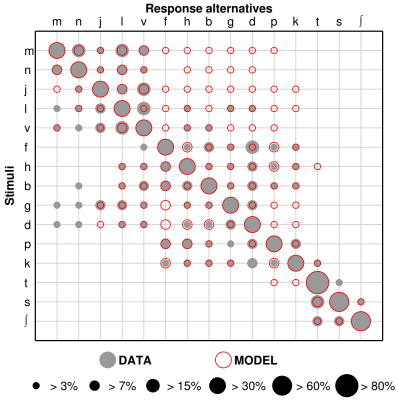
Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 23

Johannes Zaar

Measures and computational models of microscopic speech perception



Hearing Systems
Department of Electrical Engineering

Measures and computational models of microscopic speech perception

PhD thesis by
Johannes Zaar

Preliminary version: September 14, 2016

HEARING SYSTEMS

Technical University of Denmark

2016

This PhD dissertation is the result of a research project carried out at the Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark.

The project was financed by the European Commission under Contract No. FP7-PEOPLE-2011-290000.

Supervisor

Prof. Torsten Dau
Hearing Systems Group
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Abstract

Speech is a crucial part of the way we communicate with each other. We have the ability to convey complex information via speech sounds and rely on our hearing sense to decode and interpret this information. The auditory system is adapted to extracting target speech sounds in adverse acoustic conditions and high-level cognitive processing furthermore allows us to make sense of what we hear, even if the acoustic information is severely degraded or sparse. In order to determine speech intelligibility in a given acoustic condition, sentences are typically presented to listeners and the amount of correctly recognized speech items is counted, providing an overall measure of speech intelligibility. However, such a *macroscopic* speech intelligibility measure is rather coarse as it represents a combination of (i) effects of the salience of the perceived speech and (ii) effects related to linguistic processing (e.g., using context information).

This thesis presents an alternative approach reflecting a *microscopic* measure of speech perception that is solely related to the salience of the perceived speech, i.e., without effects of linguistic context. Here, the perception of individual consonants is investigated using nonsense syllables like /ta, ba/ as stimuli and evaluating the responses in terms of consonant recognition and consonant confusions. This approach allows to investigate the effects of acoustical transmission channels (e.g., rooms, mobile phones), as well as effects of hearing impairment and hearing-instrument signal processing on the fundamental speech sounds. Here, the effects of different sources of variability in consonant-in-noise perception are analyzed, such as differences in the stimuli and differences in the normal-hearing listeners. Based on the experimental data, a computational model of microscopic speech perception is proposed that consists of a model of the auditory periphery and a template-matching based decision stage. Model predictions of the consonant-in-noise data were obtained and shown to account for the perceptual effects both in terms of consonant recognition and confusions. Furthermore, effects of hearing-instrument signal processing on consonant perception were studied and shown to lead to distinct consonant confusions. The corresponding model predictions showed a large agreement with the perceptual data, both in terms of consonant recognition and confusions.

The experimental results of this thesis have implications for the design of consonant perception experiments. Furthermore, the proposed model framework could be useful for the evaluation of hearing-instrument processing strate-

gies, particularly when combined with simulations of individual hearing impairment.

Resumé

Tale er en central del af daglig kommunikation. Vi kan bruge talelyde til at formidle kompleks information som kan afkodes og fortolkes gennem hørelsen. Det auditive system er tilpasset evnen til at udtrække bestemte lyde i vanskelige akustiske omgivelser og yderligere kognitiv bearbejdning gør os i stand til at afkode hvad vi hører, selv når den akustiske information er degraderet eller knap. Taleforståelighed i en given akustik bestemmes traditionelt ved at præsentere lyttere for hele sætninger hvorefter antal korrekt genkendte ord anvendes som globalt mål for taleforståelighed. Et sådant makroskopisk mål for taleforståelighed er forholdsvist uspecifikt idet det repræsenterer en kombination af både (i) virkninger, der er knyttet til talens saliens, og (ii) virkninger, der er knyttet til den lingvistiske bearbejdning i sig selv (fx brug af kontekstuel information).

I denne afhandling præsenteres en alternativ tilgang i form af et mikroskopisk mål for taleperception, der udelukkende er baseret på talens saliens, dvs. uden konteksteffekter. Perception af individuelle konsonanter undersøges ved brug af stimuli i form af stavelser uden betydningsindhold, som fx /ta, ba/, der undersøges eksperimentelt i form af konsonant-genkendelse eller mønstre af konsonant-forvekslinger. Denne tilgang gør det muligt at undersøge, hvilken virkning den akustiske transmissions-kanal (fx forskellige rum eller mobiltelefoner), hørenedsættelse eller signalbehandling i et høreapparat har på de grundlæggende talelyde. Forskellige kilder til variabilitet i perceptionen af konsonanter i støj undersøges, som fx variabilitet i de akustiske stimuli samt variabilitet i lyttere med normal hørelse. På baggrund af eksperimentelt data fremsættes en computationel model for mikroskopisk taleperception, der består af en model af det perifere auditive system samt et mønster-baseret genkendelses-modul. Det vises at modelforudsigelser af data for konsonanter i støj kan gøre rede for de perceptuelle data, både i forhold til konsonant-genkendelse og -forveksling. Modelforudsigelserne Virkninger af forskellige former for signal behandling i høreapparater på konsonant-perception blev undersøgt og resulterede i distinkte mønstre af konsonant-forveksling. Tilhørende model-forudsigelser viste god overensstemmelse med de perceptuelle data, både i forhold til konsonant-genkendelse og -forveksling.

De eksperimentelle resultater der præsenteres i nærværende afhandling har betydning for udformningen af eksperimenter, der undersøger konsonant-perception. Derudover vil den foreslåede modellering kunne anvendes til evaluering af processerings-strategier for høreinstrumenter, især når de kombineres med simuleringer af det individuelle høretab.

Acknowledgments

This thesis represents the outcome of more than 3 years of research within the Hearing Systems Group at the Technical University of Denmark. The PhD project was furthermore embedded in the EU initial training network INSPIRE (Investigating Speech Processing in Realistic Environments). These past years have been very challenging as well as very rewarding, both on a professional and on a personal level.

I would like to thank my supervisor Torsten Dau for giving me the opportunity and the confidence to conduct this challenging research project and for finding the right motivational or critical words to keep me on track. I am grateful for his excellent support and his enthusiasm throughout the entire project, which I benefitted greatly from in terms of scientific progress as well as regarding my ability to convey scientific content.

I would furthermore like to thank Søren Jørgensen, who helped me enormously with his modeling skills and his clear thinking during the first half of the PhD project. Moreover, I would like to thank Nicola Schmitt and Ralph-Peter Derleth as well as Mishaela DiNino and Julie Bierer for fruitful and insightful collaborations.

I also want to thank my colleagues from the Hearing Systems Group, in particular Christoph Scheidiger, Tobias May, and Jens Hjortkjær for many scientific discussions, as well as Caroline van Oosterhout for her fantastically competent support regarding various organizational aspects.

Finally, I want to thank my wife Miranda for her love and support.

Related publications

Journal papers

- Zaar, J. and Dau, T. (**2015**). "Sources of variability in consonant perception of normal-hearing listeners," J. Acoust. Soc. Am. **135**, 1253–1267.

- Zaar, J. and Dau, T. (**2016**). "Predicting consonant recognition and confusion in normal-hearing listeners," J. Acoust. Soc. Am., under review.

- Zaar, J., Schmitt, N., Derleth, R.-P., DiNino, M., Bierer, J.M., and Dau, T. (**2016**). "Predicting effects of hearing-instrument signal processing on consonant perception," in preparation.

Conference papers

- Zaar, J., Jørgensen, S., and Dau, T. (**2014**). "Exploring the physical correlates of consonant recognition and confusions," Proc. of the 17. Jahrestagung der Deutschen Gesellschaft für Audiologie, Oldenburg, Germany.

- Zaar, J., Jørgensen, S., and Dau, T. (**2014**). "Modeling consonant perception in normal-hearing listeners," Proc. of the Forum Acusticum, Krakow, Poland.

- Zaar, J. and Dau, T. (**2015**). "Auditory correlates of stimulus-induced variability in consonant perception," Proc. of the 41. Jahrestagung für Akustik (DAGA), Nürnberg, Germany.

- Zaar, J. and Dau, T. (**2016**). "Auditory features in consonant perception - a modeling perspective," Proc. of the Speech Processing in Realistic Environments (SPIRE) workshop, Groningen, The Netherlands.

Book chapters

- Zaar, J. and Dau, T. (**2016**). "Sources of Variability in Consonant Perception and Implications for Speech Perception Modeling," in *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing*, edited by van Dijk, P. et al. (Springer Science+Business Media, New York), 437–446.

Conference posters and published abstracts

- Zaar, J., Jørgensen, S., and Dau, T. (**2014**). "Consonant confusions in frozen and random white noise," 6th Workshop on Speech in Noise: Intelligibility and Quality, Marseille, France.

- Zaar, J. and Dau, T. (**2015**). "Sources of variability in consonant perception and their auditory correlates," J. Acoust. Soc. Am. **137**, 2306.

Contents

1

General introduction

The human auditory system is a highly developed sensory organ that is capable of detecting and interpreting the large variety of acoustic information that surrounds us. We have the ability to convey immensely complex information via speech sounds and rely strongly on our hearing sense to decode and interpret this information. The auditory system is adapted to speech understanding and allows us to hear out the desired speech signal when the speech is masked by other sounds. We are also able to adapt to changes in the speech signals we hear, for example to another talker's articulation or to changes induced by a transmission channel like a mobile phone. However, our ability to understand speech does not only originate from our auditory processing, but also from analyzing the acoustic information using high-level cognitive processing. This allows us to make use of our linguistic knowledge to make sense of what we hear, even if the acoustic information is severely degraded. The field of psychoacoustics provides the means to measure the detectability and recognizability of acoustic signal features in various acoustic conditions. Similarly, psychoacoustic methods have been applied to measure the recognition of speech *information* in terms of the intelligibility of sentences, as well as to examine the recognition of speech *sounds* in terms of recognition and confusions of individual phonemes.

To measure the recognition of speech information, researchers have tested the intelligibility of sentences in the presence of, e.g., stationary noise, fluctuating noise, competing talkers, and reverberation. Various speech tests have been designed, some of which consist of syntactically diverse meaningful sentences, as in the "hearing in noise test" (e.g., HINT; Nilsson et al., 1994; Nielsen and Dau, 2011) and the "conversational language understanding evaluation" (CLUE; Nielsen and Dau, 2009). In these tests, the listener is presented with a given sentence, e.g., mixed with background noise, and asked to repeat the sentence to an experimenter who evaluates whether the sentence was correctly understood. As this procedure is time consuming and requires large speech corpora as well as the presence of an experimenter, matrix sentence tests (e.g., Hagerman, 1982; Wagener et al., 2003) have been proposed as an alternative. In these tests, semantically unpredictable sentences (i.e., without meaning) are presented within a fixed syntactical structure, using only a limited number of words for each noun, verb, etc. While this allows computer-based self-scoring of the listeners and limits the required size of the speech corpus, it also facilitates guessing and can lead to an overestimation of speech intelligibility in adverse conditions.

Commonly, the speech reception threshold (SRT) has been used to quantify speech intelligibility, which reflects the speech-to-masker energy at which 50% of the presented speech items have been correctly identified. The use of long-term (meaningful) speech units provides a *macroscopic* perspective on speech perception, as listeners can exploit information obtained from various stages of speech processing. For instance, missing acoustic information can be restored using lexical, semantic and/or syntactic information (e.g., Miller and Licklider, 1950; Kashino, 2006). Thus, while macroscopic speech intelligibility tests may provide useful global measures for, e.g., comparing different speech transmission channels, the measures do not necessarily provide information about the transmission of the actual acoustic speech cues, which may be affected by the transmission channel (e.g., noise, reverberation, non-linear processing in a phone or a hearing aid) and/or by the receiver (e.g., due to hearing impairment).

Speech perception can alternatively also be studied at a more basic level using a *microscopic* approach. Many studies have focused on investigating the perception of consonants and vowels embedded in nonsense syllables. The perception of consonants has attracted special attention, as many consonants exhibit short duration and high-frequency energy, which makes them perceptually more vulnerable and thus more "critical" than the vowels in most conditions (Phatak and Allen, 2007). Typically, combinations of consonants and vowels (e.g., /ta/, /ba/, etc.), mixed with stationary noise, have been considered and the perceptual data have been analyzed in terms of consonant recognition (i.e., the percentage of correctly identified consonants) as well as in terms of consonant confusions. In contrast to the macroscopic approach, the microscopic approach (i) uses short-term speech stimuli, (ii) analyzes recognition as well as confusions, and (iii) employs nonsense speech stimuli, thus excluding the contribution of effects related to lexicon, meaning, and syntax. In this sense, microscopic speech perception tests are strongly related to the "original" psychoacoustic measures of signal detection and identification, while they at the same time test the integrity of the considered speech categories (e.g., the consonants).

Interestingly, one of the very first speech studies (e.g., Fletcher and Galt, 1950), conducted in the context of research on telephone speech transmission quality, focused on the recognition of consonants and vowels to assess the amount of correctly transmitted *articulation* under conditions of noise and spectral filtering. These investigations eventually resulted in the Articulation Index (AI) model (ANSI, 1969, see further below). In a famous study, Miller and Nicely (1955) investigated perceptual confusions among consonants in conditions of white noise at various signal-to-noise ratios (SNRs) and spectral filtering. Their study suggested that distinct perceptual confusions among consonants may have a major effect on speech intelligibility in noise. Many studies followed, investigating consonant perception with respect to articulatory features (Wang and Bilger, 1973), the influence of the noise spectrum (Phatak and Allen, 2007; Phatak et al., 2008), and the spectro-temporal "footprints" of specific consonant cues (Li et al., 2010; Li et al., 2012). While the earlier studies used various speech tokens to represent a given consonant, recent studies indicated substantial perceptual differences across different speech tokens of the same type (e.g., Toscano and Allen, 2014).

However, a systematic investigation of the factors that influence consonant perception in noise (e.g., different speech tokens, different masking-noise waveforms, listener effects) has so far not been undertaken. This may be particularly relevant given that consonant tests have been shown to be informative for assessing individual hearing impairment (Phatak et al., 2009; Trevino and Allen, 2013), evaluating hearing-aid amplification schemes (Scheidiger and Allen, 2013), and examining effects of highly non-linear hearing-aid signal processing strategies (Schmitt et al., 2016).

To simulate the behaviour of human listeners and to better understand the acoustic features that are crucial for speech intelligibility, a variety of computational models has been proposed. These speech intelligibility models have been based on the assumption that speech intelligibility is related to the SNR after simplified simulations of the signal processing in the auditory system. These simulations typically include the well-established frequency-selective processing in the peripheral auditory system, while some models also consider a modulation-frequency selective process, inspired by findings from psychoacoustic amplitude-modulation detection studies (Dau et al., 1997). Modeling approaches like the AI (ANSI, 1969) and the Speech Intelligibility Index (ANSI, 1997; Rhebergen et al., 2006) take the speech and noise signals as separate inputs and predict speech intelligibility based on a weighted average of the SNR across a range of spectral bands. The widely used Modulation Transfer Function (MTF) based modeling approaches, like the Speech Transmission Index (STI, Houtgast et al., 1980) and the speech-based STI (sSTI, Payton and Braida, 1999), predict speech intelligibility based on both a frequency-selective and a modulation-frequency selective analysis of the signal processed through a transmission channel. The recently proposed speech-based Envelope Power Spectrum Model (sEPSM, Jørgensen and Dau, 2011; Jørgensen et al., 2013) applies a frequency-selective and a modulation-frequency selective analysis to the noisy speech and the noise alone and relates speech intelligibility to the signal-to-noise ratio in the envelope domain ($\mathrm{SNR}_{\mathrm{env}}$). The STI and, in particular, the sEPSM have been shown to yield a larger predictive power as compared to the AI/SII, which are solely based on spectral signal analysis.

Only a few studies have attempted to predict microscopic speech perception data, which may be particularly insightful with respect to understanding effects of differences in the sensory processing (e.g., induced by a hearing impairment) and effects of hearing-aid compensation strategies on fundamental speech cues. These studies combined elaborate models of the auditory periphery with a template-matching speech recognition back end to predict nonsense syllable perception in terms of recognition and confusions. In particular, the auditory model of Dau et al. (1996), which consists of a linear auditory filterbank, an envelope extraction stage, a nonlinear adaptation stage, and a low-pass filter, was used in combination with a template matcher by Holube and Kollmeier (1996) to predict consonant recognition. Jürgens and Brand (2009) used a later version of the model (Dau et al., 1997), which contains a modulation filterbank and also constitutes the basis of the sEPSM model, to predict consonant recognition in normal-hearing (NH) listeners, as well as in hearing-impaired (HI) listeners (Jürgens et al., 2014). Another related auditory model by Jepsen et al. (2008), which includes nonlinear amplification in the auditory filterbank and thus accounts for active processes in the cochlea, was used for predicting consonant perception in HI listeners using template matching (Jepsen et al., 2014).

However, while the proposed "normal-hearing" microscopic models were shown to largely account for average consonant recognition scores (Holube and

Kollmeier, 1996) and consonant-specific recognition scores (Jürgens and Brand, 2009) measured at different SNRs in stationary noise, they did not account well for the consonant confusions made by the listeners. Furthermore, it remained unclear whether these models would be able to capture the perceptual differences observed on the level of individual speech tokens of the same type (e.g., Toscano and Allen, 2014).

The present thesis addresses two main challenges. First, an in-depth experimental investigation of the factors that influence consonant-in-noise perception in NH listeners is described. This investigation was conducted to clarify the role of various potential sources of variability and, thus, gain essential insights regarding the perceptual reference for modeling consonant perception. Second, a computational model of microscopic speech perception is proposed, designed as an extension of the model of Dau et al. (1997) towards predicting consonant recognition and confusions. The model differs considerably from the model of Jürgens and Brand (2009) in that it maintains the crucial decision-stage mechanisms of Dau et al. (1997). The predictive power of the model was evaluated using the detailed data set measured in the experimental investigation of consonant-in-noise perception as well as consonant perception data obtained in conditions of hearing-instrument signal processing.

Chapter 2 describes an extensive investigation of the factors that influence consonant perception in NH listeners. In particular, two consonant perception experiments were conducted and analyzed with respect to perceptual differences that arise from variations in the *source* (i.e., in the stimulus) and differences that are related to the *receiver* (i.e., the listeners). The source-related variability comprises effects of differences in speech tokens of the same type (spoken by different talkers or the same talker) as well as effects of differences in the noise waveforms. The receiver-related variability reflects differences across different listeners as well as within individual listeners in terms of test-retest reproducibility. The different factors are compared by means of graphical examples (confusion patterns and confusion matrices) and using an analysis scheme based on the perceptual distance between responses.

Chapter 3 proposes a computational model of microscopic speech perception based on the auditory processing model of Dau et al. (1997). The model consists of a temporally dynamic template matching back end, combined with a cross-correlation based decision metric and an internal-noise term. The predictive power of the model in terms of consonant recognition and consonant confusions is evaluated based on the data obtained in Chapter 2. The evaluation is conducted by means of graphical comparisons of data and model predictions, as well as using correlation analyses.

Chapter 4 presents experimental investigations of effects of hearing-instrument signal processing on consonant perception and assesses the predictive power of the proposed microscopic model for the considered conditions. In particular, effects of strong nonlinear frequency compression and impulse-noise suppression on consonant perception in NH listeners are considered experimentally and in the model. Furthermore, consonant perception data from a study by DiNino et al. (2016), obtained with simulations of cochlear-implant processing in NH listeners, are used to test the model. The model performance for the two data sets is evaluated using comparisons of confusion matrices as well as correlation analyses.

Finally, *Chapter 5* summarizes the main findings and discusses the implications of the experimental results, the limitations and perspectives of the experimental method, as well as the role of the model components and the limitations and perspectives of the proposed model framework.

2

Sources of variability in consonant perception of normal-hearing listeners[a]

Abstract Responses obtained in consonant perception experiments typically show a large variability across stimuli of the same phonetic identity. The present study investigated the influence of different potential sources of this response variability. It was distinguished between source-induced variability, referring to perceptual differences caused by acoustical differences in the speech tokens and/or the masking noise tokens, and receiver-related variability, referring to perceptual differences caused by within- and across-listener uncertainty. Consonant-vowel combinations (CVs) consisting of 15 consonants followed by the vowel /i/ were spoken by two talkers and presented to eight normal-hearing listeners both in quiet and in white noise at six different signal-to-noise ratios. The obtained responses were analyzed with respect to the different sources of variability using a measure of the perceptual distance between responses. The speech-induced variability across and within talkers and the across-listener variability were substantial and of similar magnitude. The noise-induced variability, obtained with time-shifted realizations of the same random process, was smaller but significantly larger than the amount of within-listener variability, which represented the smallest effect. The results have implications for the design of consonant perception experiments and provide constraints for future models of consonant perception.

[a] This chapter is based on Zaar and Dau (2015).

2.1 Introduction

Speech intelligibility is often characterized in terms of the percentage of correctly identified meaningful words or sentences presented to the listener, either in quiet or in the presence of a noise masker or interfering talker(s). For instance, a common measure of speech intelligibility is the speech reception threshold (SRT), which reflects the speech-to-masker/interferer energy at which 50% of the presented speech items have been correctly identified. The SRT measure may be considered as reflecting a *macroscopic* view on speech perception. The term macroscopic is threefold in the sense that (i) long-term speech units are used, such as words or sentences, (ii) only speech recognition is considered while confusions of words are not investigated, and (iii) meaningful speech is used, typically consisting of common words in a syntactically correct sentence structure. In this type of experimental setting, listeners can exploit information obtained from various stages of speech processing. For instance, missing acoustic information can be extrapolated using lexical, semantic and/or syntactic information. Approaches for measuring macroscopic speech intelligibility range from presenting syntactically diverse meaningful sentences as in the "hearing in noise test" (e.g., HINT, Nilsson et al., 1994; Nielsen and Dau, 2011) and the "conversational language understanding evaluation" (CLUE, Nielsen and Dau, 2009) to using matrix sentence tests (e.g., Hagerman, 1982; Wagener et al., 2003), where semantically unpredictable sentences are presented within a fixed syntactical structure. Therefore, macroscopic speech intelligibility tests differ in the semantic and syntactic predictability provided, while lexical effects play a considerable role in any of these tests.

Addressing speech intelligibility at a more fundamental level, many studies have focused on investigating the perception of smaller units of speech, such as syllables or phones (i.e., consonants and vowels). The perception of vowels has been shown to be more robust in the presence of steady-state noise than the perception of many consonants (Phatak and Allen, 2007). Therefore, the most "critical" or vulnerable phones in this context are consonants. Combinations of consonants and vowels (e.g., /ta/, /ba/, etc.) have typically been considered and the perceptual data have been analyzed in terms of consonant recognition (i.e., the percentage of correctly identified consonants) as well as in terms of consonant confusions. This type of approach may be considered as *microscopic* as it (i) uses short-term speech stimuli, (ii) analyzes recognition as well as confusions, and (iii) employs nonsense speech stimuli, thus excluding the contribution of effects related to lexicon, meaning, and syntax. The microscopic approach therefore allows for an analysis of the mapping from the acoustical stimulus to the associated phone percept by minimizing the biases induced by higher-level speech processing. This could be relevant, for example when analyzing the effects of acoustical transmission channels (e.g., mobile phones), hearing impairment, and hearing-aid signal processing algorithms on the perception of the fundamental building blocks of speech. However, in order to fully exploit the microscopic approach it seems crucial to understand the factors that contribute to consonant perception.

The first investigations of nonsense syllable perception were conducted by Fletcher and colleagues in the context of their pioneering research on telephone speech transmission quality at the Bell Laboratories between 1919 and 1945 (e.g., Fletcher and Galt, 1950; see also Allen, 1994). Nonsense consonant-vowel-consonant (CVC), consonant-vowel (CV), and vowel-consonant (VC) combinations were used to assess the amount of correctly transmitted articulation under conditions of noise and spectral filtering. These investigations resulted in the definition of the articulation index (AI), a technical measure to determine the quality of speech transmission channels (French and Steinberg, 1947). Although Fletcher and Galt (1950) did not directly address phonetic confusions, their work provided the basis for further research on nonsense speech perception.

Miller and Nicely (1955) conducted the first study that focused on perceptual confusions among consonants. CVs consisting of the sixteen most common English consonants followed by the vowel /a/ (as in father) were spoken by five talkers and presented to four listeners. In a set of experimental conditions, white noise was added at various signal-to-noise ratios (SNRs) and different band-pass filters were applied to the speech. After each presentation, listeners had to indicate the consonant they had heard. The responses were pooled across listeners and displayed as confusion matrices (CMs). Several perceptual confusion groups of consonants (e.g., /p, t, k/) were observed and the data were investigated in terms of the information transmitted by different articulatory features (voicing, nasality, affrication, duration, and place of articulation). Wang and Bilger (1973) considered CVs and VCs consisting of 25 consonants and the vowels /a, i, u/ and applied a sequential information analysis in an attempt to derive an ideal set of relevant articulatory features. However, their results suggested that an articulatory feature-based analysis might be inappropriate to account for the data. Furthermore, Wang and Bilger (1973) found that the accompanying vowel had an influence on the consonant detection performance as well as the type of consonant-vowel combination (CV or VC), demonstrating that consonant perception does not solely depend on the consonant but also on the vowel context the consonant is embedded in.

In a related more recent study, Allen (2005) re-analyzed the Miller and Nicely (1955) data and related them to the AI. Allen proposed that confusion matrices should be analyzed in terms of perceptual events rather than in terms of articulatory features. He introduced the confusion pattern (CP) which, for a given speech stimulus, depicts the proportions of the different response alternatives as a function of the experimental conditions (e.g., SNRs). The CP was shown to be more appropriate than the confusion matrix for characterizing perceptual confusion groups and other trends in the data since it provides an overview of the data across experimental conditions. Phatak et al. (2008) reproduced the main results obtained in the Miller and Nicely (1955) study and demonstrated considerable noise-type specific perceptual differences (comparing white noise and speech-weighted noise). They also showed that different speech tokens of the same phonetic identity induced strong differences in consonant recognition and confusions. Li et al. (2010; 2012) developed a psychoacoustic method named “three-dimensional deep search” which was designed to identify the spectro-temporal cue regions of consonants based on experimental consonant recognition data obtained with noise masking, spectral filtering, and time truncation. As this kind of microscopic speech investigation relies heavily on the characteristics of the individual speech tokens, the perceptual differences across different speech tokens of the same phonetic identity came more into focus.

Consistent with the findings of Phatak et al. (2008), Singh and Allen (2012) demonstrated the occurrence of major within-consonant speech-token specific differences in the recognition of stop consonants. Toscano and Allen (2014) investigated across- and within-consonant recognition errors for CVs consisting of the sixteen consonants used by Miller and Nicely (1955) followed by four different vowels. Each of the CVs was spoken by fourteen different talkers and presented at six SNRs in speech-weighted noise. The results suggested that consonant recognition greatly varies across consonants as well as within consonants (i.e., across talkers and accompanying vowels). This implies talker-dependent effects, which have also been shown for spoken word recognition (e.g., Mullennix et al., 1989) and represent a major challenge in automatic speech recognition (Benzeghiba et al., 2007).

While the above studies provided major insights into consonant perception from various perspectives, it has remained unclear (i) to what extent the reported speech-token dependence of consonant perception is related to articulatory differences across talkers or to differences in the accompanying vowel, (ii) how articulatory differences across different utterances of a given talker affect consonant perception, and (iii) whether spectro-temporal details of the individual masking-noise waveform affect consonant perception. Furthermore, perceptual differences across and within individual listeners have not yet been addressed systematically, apart from individual studies considering hearing-impaired (HI) listeners (e.g., Phatak et al., 2009; Trevino and Allen, 2013) or groups of listeners with different language background (e.g., Cutler et al., 2004).

The present study was undertaken in an attempt to quantify the relative importance of some of the factors that influence consonant perception both in terms of stimulus-related ("source") and listener-related ("receiver") effects. Here, it was distinguished between *source-induced* variability and *receiver-related* variability. Source-induced variability refers to perceptual differences that arise due to variations in the acoustic properties of the stimulus and is subdivided into (i) *speech-induced* variability (perceptual differences arising from articulatory differences in speech tokens of the same phonetic identity, categorized as across-talker and within-talker variability) and (ii) *noise-induced* variability (perceptual differences arising from differences in the waveform of the masking noise). Receiver-related variability refers to the uncertainty/variation of the perceptual response due to encoding/resolution differences and limits in the listeners and is subdivided into (i) *across-listener* variability and (ii) *within-listener* variability. Additional well-known sources of variability like the position and type of the accompanying vowel, as well as the long-term spectral characteristics of the noise (e.g., white vs. speech-weighted) were not considered here.

Fifteen Danish consonants combined with the vowel /i/ as CVs were used in the present study, spoken by non-professional native Danish talkers and presented to NH native Danish listeners. Two experiments were conducted using white noise maskers at six SNRs. Experiment 1 investigated the effect of variations in the speech stimulus using several speech tokens for each CV presented in deterministic white noise maskers. Experiment 2 addressed the effect of noise variability using only a single speech token per CV and presenting it in different deterministic realizations of white noise maskers. The experimental data were analyzed with respect to source-induced variability using the data obtained in experiment 1 for analyzing the speech-induced variability, and using the data obtained in experiment 2 for analyzing the noise-induced variability. Furthermore, the data obtained in the two experiments were analyzed with respect to receiver-related variability by comparing the responses to physically identical stimuli across and within listeners. The analyses were performed by comparing example confusion patterns and confusion matrices. The entire set of the collected data was furthermore analyzed using a perceptual distance measure and the entropy of responses to quantify the contributions of the different considered sources of variability.

2.2 Method

2.2.1 Experiment 1: Effects of variations in the speech stimulus

Listeners

Eight native Danish listeners (one female, seven male) with audiometric thresholds of 20 dB hearing level (HL) or less at the measured frequencies between 125 Hz and 8 kHz participated in the experiment. The age of the listeners ranged from 19 years to 27 years, except for one listener who was 38 years old. The average age was 26 years. Listeners were paid for their participation in the experiment.

Stimuli

CVs consisting of the 15 consonants /b, d, f, g, h, j, k, l, m, n, p, s, ʃ, t, v/ followed by the vowel /i/ were used throughout this study. For experiment 1, six recordings of each CV were taken from the Danish nonsense syllable speech material collected by Christiansen and Henrichsen (2011). For each CV, three of these speech tokens were spoken by one particular male talker, the other three speech tokens were spoken by one particular female talker. A total of 90 speech tokens was used in the experiment (15 CVs × 3 speech tokens × 2 talkers). The individual speech tokens were cut and faded in and out manually. Their levels were equalized using VUSOFT, a software implementation of an analog VU-meter developed by Lobdell and Allen (2007), which was also used for level equalization in Phatak et al. (2008). The level equalization was performed such that all CVs showed the same VUSOFT peak value. This equalization strategy is based on the vowel levels, thus ensuring realistic relations between the levels of the individual consonants. Therefore, the vowel levels across the equalized CVs were similar while the consonant levels differed, much like in natural speech. After equalization, the reference speech level for the SNR calculation was defined as the overall root-mean-square level of all speech tokens.

For the masking noise generation, a “half-frozen” noise approach was taken in order to avoid a potential blur in the perceptual data that might arise from an effect of differences in the noise waveforms. Specifically, the noise waveform was fixed (“frozen”) for a given speech token in a given SNR condition and the waveform of the presented mixture of speech and noise was thus exactly the same across (i) repeated presentations of a speech token in a given SNR condition and (ii) across different listeners. For each speech token and each SNR condition, one white Gaussian masking noise token with a duration of 1 s was generated and faded in and out using raised cosine ramps with a duration of 50 ms.

SNR conditions of 12, 6, 0, -6, -12, and -15 dB were created by fixing the noise level and adjusting the level of the speech tokens based on the reference speech level according to the desired SNR. The sound pressure level of the noise was set to 60 dB, while the overall stimulus level differed depending on the level of the speech (i.e., on the SNR). This fixed noise level approach was chosen instead of the commonly used fixed speech level approach in order to avoid extremely high noise levels at low SNRs, which can lead to annoyance and fatigue in listeners. The speech tokens were mixed with the respective noise tokens such that the speech token onset was temporally positioned 400 ms after the noise onset. The clean speech at the respective levels, the noise tokens, and the mixture of speech and noise tokens were individually stored in “.wav” format at a sampling rate of 44.1 kHz with a resolution of 16 bits per sample.

Experimental design

The experiment was split into two sessions, one using the 45 male talker speech tokens and the other one using the 45 female talker speech tokens. The listeners performed the individual sessions on different days. One session lasted approximately 2.5 hours including instruction, training, and breaks. Two control conditions with speech in quiet were defined, “Q60” and “Q45”. Q60 was designed to evaluate whether the speech tokens were sufficiently identifiable in ideal listening conditions; the speech was presented in quiet at the same presentation level at which the fixed-level noise was presented in the SNR conditions [60 dB sound pressure level (SPL)]. Since the noise level was fixed and the speech level was adapted to generate the individual SNR conditions, Q45 was designed to investigate whether the speech tokens were still sufficiently intelligible in quiet at the lowest speech level occurring in the SNR conditions; the speech was therefore presented at 45 dB SPL, corresponding to the speech level in the -15 dB SNR condition.

The experimental sessions were split into eight consecutive blocks corresponding to the eight experimental conditions. In order to get the listeners accustomed to the task, the first condition was the “easy” control listening condition Q60, followed by the slightly more challenging control condition Q45. The third to eighth conditions were the six SNR conditions ranked from easy to difficult, i.e., with SNR = 12, 6, 0, -6, -12, and -15 dB. Each block consisted of a training run followed by the experiment run. In the training run, all 45 stimuli (depending on the condition speech tokens or speech tokens mixed with the predefined noise tokens) were presented once in random order to familiarize the listener with the respective condition. In the experiment run, each of the 45 stimuli was presented 3 times, resulting in a total of 135 presentations. The order of presentation was again randomized. One experimental block therefore comprised 180, a whole session 1440 stimulus presentations.

Procedure and apparatus

Listeners were seated in a sound attenuating listening booth in front of a computer and listened to the stimuli monaurally through Sennheiser HD580 headphones. For headphone equalization, a 256-tab finite impulse response filter designed to equalize the third-octave smoothed version of the headphone transfer function in the range between 40 Hz and 21 kHz was applied. The test software was run under Matlab on a Windows-based PC. The stimuli were played at a sampling rate of 44.1 kHz. After each stimulus presentation, listeners had to choose one of the response alternatives displayed on a graphical user interface (GUI). The task was identical in training and experiment with no feedback provided. When in doubt, the listeners could repeat the sound playback up to two times using a “repeat” button included in the GUI. The response alternatives consisted of 15 buttons displaying the consonants in the corresponding Danish spelling (b, d, f, g, h, j, k, l, m, n, p, s, Sj, t, v) and one button labeled “I don’t know”. Listeners could respond to the stimulus using a computer mouse. After a decision was made, the next stimulus was played after a 500-ms pause. Prior to the experiment, the listeners were instructed to make use of the “repeat” button whenever they were uncertain about their percept and to use the “I don’t know” button instead of guessing when they had only heard the vowel. The proportions of responses for each speech token and condition were calculated via division of the obtained occurrences of responses by the number of stimulus presentations. The “I don’t know” responses were evenly distributed across all response alternatives. The conversion was performed both for the pooled responses across listeners and the individual listeners’ responses. Three observations per stimulus and individual listener were obtained; the number of pooled observations per stimulus was thus 24 (3 observations × 8 listeners).

2.2.2 Experiment 2: Effects of variations in the noise

Listeners

Eight native Danish listeners (one female, seven male) with audiometric thresholds of 20 dB hearing level (HL) or less at the measured frequencies between 125 Hz and 8 kHz participated in the experiment. Four of these listeners had also participated in experiment 1. The age of the listeners ranged from 20 years to 28 years, with a mean age of 24 years. Listeners were paid for their participation in the experiment. To obtain test-retest data, a subset of the original listener panel (four of the eight listeners) conducted a retest approximately one month after the first test.

Stimuli

Experiment 2 addressed perceptual differences induced by different “frozen-noise” masker waveforms. For each type of CV from experiment 1, only one recording was used, resulting in 15 speech tokens. The recordings were a subset of the speech material used in experiment 1, spoken by the male talker. The level of the speech tokens was equalized according to the VUSOFT peak value (cf. Sec. 2.2.1). Three masking-noise conditions (frozen noise A, frozen noise B, and random noise) were considered. For each speech token, one particular white noise waveform with a duration of one second was generated and labeled “frozen noise A”; the same noise token was then circularly shifted in time by

100 ms to obtain “frozen noise B”. The noise tokens were faded in and out using raised cosine ramps with a duration of 50 ms. The noise waveforms for the random noise condition were newly generated for each presentation and faded in and out in the same manner during the experimental procedure. The responses obtained in the random noise condition were not considered in the analysis as this condition was only included to prevent listeners from noise learning. Note that, for a given speech token, the frozen-noise tokens in this experiment were the same across all SNR conditions (“frozen”), in contrast to experiment 1 where different noise tokens were used across SNR conditions (“half-frozen”). SNR conditions of 12, 6, 0, -6, -12, and -15 dB were created by fixing the noise level to 60 dB SPL and adjusting the level of the speech tokens (cf. Sec. 2.2.1). Each speech token was mixed with the two respective frozen-noise tokens such that the speech token onset was temporally positioned 400 ms after the noise onset. For the random-noise condition, the same was done during the experiment using randomly generated noise waveforms. The clean speech at the respective levels, the frozen-noise tokens, and the mixture of speech and frozen noise tokens were individually stored in “.wav” format at a sampling rate of 44.1 kHz with a resolution 16 bits per sample.

Experimental design

As in experiment 1, two control conditions were defined (“Q60” and “Q45”), in which the speech was presented in quiet at 60 dB SPL and 45 dB SPL, respectively (see also Sec. 2.2.1). The experiment was split into eight consecutive blocks corresponding to the eight experimental conditions (in this order: Q60, Q45, SNR = 12, 6, 0, -6, -12, and -15 dB). Each block consisted of a training run followed by the experiment run. For the quiet conditions Q60 and Q45, the training run comprised one presentation of each of the 15 speech tokens; in the experiment run, each of the speech tokens was presented 5 times, amounting to 75 presentations. The order of presentation was randomized. For the SNR conditions, the training run consisted of 3 presentations of each of the 15 speech tokens, i.e., 45 presentations. The masking noise was newly generated for each presentation during the training run. In the experiment run, each speech token was presented 5 times in each masking-noise condition, i.e., 5 times in frozen noise A, 5 times in frozen noise B, and 5 times in random noise, resulting in a total of 225 presentations. The order of presentation was randomized. One entire experimental block comprised 90 (quiet conditions Q60 and Q45) or

270 (main conditions, SNR: -15. -12, -6, 0, 6, 12 dB), the whole experiment 1800 stimulus presentations. The full experiment lasted approximately 3 hours including instruction, training, and breaks.

Procedure and apparatus

The listening situation, instructions, interface, and further technical details were the same as in experiment 1, described in Sec. 2.2.1. The data were converted to proportions of responses in the same manner as described in section Sec. 2.2.1. Five observations per stimulus and individual listener had been obtained; the number of pooled observations was thus 40 (5 observations × 8 listeners).

2.2.3 Perceptual distance calculation

To quantify the size of the different source-induced and receiver-related effects, a measure of perceptual distance was applied. Following an approach suggested by Scheidiger and Allen (2013), each response alternative (i.e., each consonant) was considered to represent one dimension in an R-dimensional space (with R denoting the number of response alternatives). In this space, each response pattern was considered as a vector. The perceptual distance between two such response patterns was calculated as the normalized angular distance between two R-dimensional response vectors $\mathbf{x}$ and $\mathbf{y}$,

$$\mathrm{D}[\mathbf{x},\mathbf{y}] = \cos^{-1}\left(\frac{\langle\mathbf{x},\mathbf{y}\rangle}{||\mathbf{x}||\cdot||\mathbf{y}||}\right)\cdot\frac{100\%}{\frac{\pi}{2}} \tag{2.1}$$

where $\langle\mathbf{x},\mathbf{y}\rangle$ denotes the scalar product and $||\mathbf{x}||$ and $||\mathbf{y}||$ represents the Euclidean norm of the response vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. The response vectors contain the proportions of responses for all R response alternatives (with $\mathrm{R} = 15$); thus, the values of the individual coordinates range from 0 to 1 and the angular distance between the two vectors therefore ranges from 0 to $\pi/2$. Normalization by $\pi/2$ and multiplication by 100% yields the normalized angular distance in percent.

The perceptual distance measure was used to describe the amount of perceptual variability induced by the different sources of variability considered in this study. It was calculated between all pairwise combinations of individual listeners' responses that are representative of each factor. For instance, the perceptual influence of across-talker variability can be described using the perceptual distances between all pairs of response vectors obtained with pairs of speech tokens of the same phonetic identity that were spoken by different talkers.

The calculations were performed for each SNR condition separately based on the response vectors obtained with the individual listeners. Depending on the factor, the number of considered response pairs and thus the number of individual distance values varied. For each factor and each SNR condition, a distribution of perceptual distance values (across the considered response pairs) was obtained. As a reference for maximal distance, the perceptual distance *across CVs* (D_{acrCV}) was calculated from the data obtained in experiment 1. To quantify the source-induced variability, the perceptual distances across talkers, within talkers (D_{acrTalk} and D_{wtnTalk}, both based on experiment 1), and across noise tokens (D_{acrNoise}, based on experiment 2: frozen noise A vs. frozen noise B condition) were calculated using response vectors obtained with physically different stimuli of the same phonetic identity. To assess the receiver-related variability, the perceptual distances of the responses across listeners (D_{acrList}, based on experiment 1 and 2) and within listeners (D_{wtnList}, based on experiment 2 test vs retest) were calculated by comparing response vectors obtained with physically identical stimuli. The perceptual distance within listeners represents the listener uncertainty and was thus considered as a baseline for minimal perceptual distance. A detailed description of the perceptual distance calculation is provided in Sec. 2.7.2.

A more common descriptor of response variability used in related studies (e.g., Miller and Nicely, 1955; Phatak et al., 2008) is the entropy of responses. For comparison with the results obtained based on the perceptual distance measure, the data were also analyzed using the normalized entropy (see Sec. 2.7.3 for details). The perceptual distance may provide an intuitive approach for investigating the perceptual effects of the different sources of variability, whereas the application of the normalized entropy for this purpose is formally less straightforward (see Sec. 2.7.2 and Sec. 2.7.3).

2.3 Results

2.3.1 Consonant recognition in quiet

The average recognition rate across all 90 speech tokens used in experiment 1 was found to be 99.2% with a standard deviation of 2.7% across CVs for Q60 (at 60 dB SPL presentation level), while the average recognition rate was 96.1% with a standard deviation of 8.5% across CVs for Q45 (at 45 dB SPL). Regarding experiment 2, the analysis showed that the average recognition rate across all 15 speech tokens was 98.8% with a standard deviation of 3.4% across CVs for Q60, while the average recognition rate was 98.2% with a standard deviation of 4.8% across CVs for Q45. All speech tokens used in the two experiments were thus considered sufficiently recognizable in quiet and taken into account for the further analyses.

2.3.2 Source-induced variability

To illustrate the source-induced variability in consonant perception, i.e. perceptual differences that occur for physically different stimuli of the same phonetic identity, selected example confusion patterns (CPs) are shown for the "average listener", representing the average proportions of responses obtained with eight listeners. The examples illustrate the large observed effect of the considered source variations on consonant recognition and confusions. An analysis of the complete data set follows further below (Sec. 2.4).

Speech-induced variability

Figure 2.1 shows average CPs obtained in experiment 1 for the CVs /di/ (left panels), /hi/ (middle panels), and /pi/ (right panels), spoken by the male talker A (top panels) and the female talker B (bottom panels), respectively. The figure illustrates the perceptual effect of *across-talker variability*. For a given speech token, the CPs show the proportions of the four predominant responses as a function of SNR. The proportions of correct responses, denoted as recognition curves, are depicted as thick black lines. The thinner colored lines indicate confusions. The Q60 quiet condition is included as a reference (the rightmost value on the abscissae).

It can be seen that /di/ spoken by talker A (top left panel) is far more confusable (mainly with /gi/) and hence far less recognizable than /di/ spoken by talker B (bottom left panel), particularly at SNRs between -12 and 0 dB. In contrast, an utterance of /hi/ spoken by talker A (top middle panel) was perfectly recognized by the listeners at SNRs down to 0 dB while the same CV spoken by talker B (bottom middle panel) yielded pronounced confusions (with /pi/, /ki/, and /fi/) and thus recognition rates of less than 50% at the same SNRs (0, 6, and 12 dB). Comparably large differences were observed between /pi/ spoken by talker A (top right panel) and /pi/ spoken by talker B (bottom right panel), especially at SNRs of 0 and 6 dB.

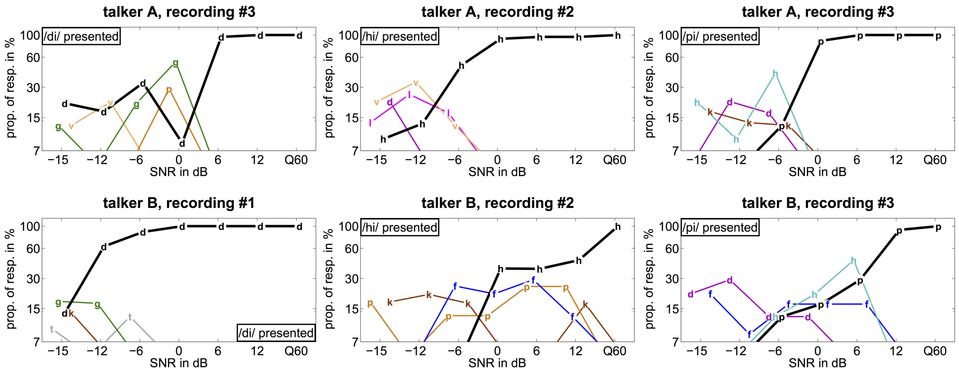


Figure 2.1: Across-talker comparison of average confusion patterns for /di/ (left panels), /hi/ (middle panels), and /pi/ (right panels). The upper and lower panels show the results for talker A (male) and talker B (female), respectively. The correct responses are indicated as thick black lines and confusions are shown as thinner lines in different colors; the data points are labeled with the corresponding consonants. Only the four predominant responses are depicted for clarity. A slight horizontal jitter was introduced to the data for better readability. The ordinate is scaled logarithmically to emphasize the confusions. The 7% minimum of the ordinate represents chance level.

Figure 2.2 shows average CPs obtained in experiment 1 for two different recordings of the CVs /gi/ spoken by the same male talker A (left panels), /ki/ spoken by the same female talker B (middle panels), and /ʃi/ spoken again by female talker B (right panels). The figure thus illustrates the perceptual effect of *within-talker variability*.

The two different recordings of /gi/ spoken by talker A (left panels) caused large differences in the recognition curves and the confusions, particularly at SNRs of 6 and 12 dB. Similarly, the two recordings of /ki/ spoken by talker B (middle panels) yielded substantially different recognition rates and confusions, in particular at SNRs of 0, 6, and 12 dB. Regarding the two recordings of /ʃi/ spoken by talker B (right panels), it can be seen that this CV was generally detected quite robustly. However, large differences in the recognition rates obtained with the two different recordings were observed for SNRs of -6 and -12 dB.

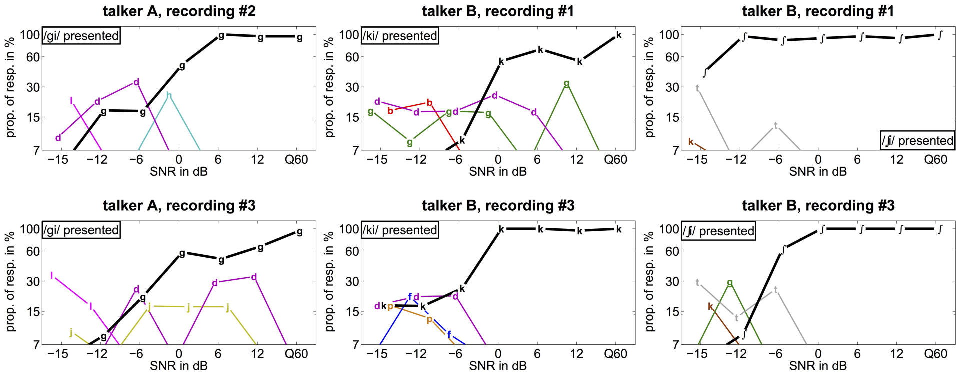


Figure 2.2: Within-talker comparison of average confusion patterns for /gi/ (left panels), /ki/ (middle panels), and /ʃi/ (right panels). The upper and lower panels show the results for two different recordings of the same CV, spoken by the same talker (male talker A in the case of /gi/ and female talker B in the cases of /ki/ and /ʃi/). The confusion patterns were obtained as described in Fig. 2.1.

Noise-induced variability

Figure 2.3 shows average CPs obtained in experiment 2 for the speech tokens /fi/ (left panels), /gi/ (middle panels), and /ni/ (right panels), each presented in frozen noise A (top) and frozen noise B (bottom), respectively. All speech tokens were spoken by the male talker A. Thus, the only difference in the acoustic waveforms of the considered stimulus pairs was a 100-ms temporal shift in the masking-noise waveform.

For the same recording of /fi/ (left panels), the two different noise waveforms led to different CPs. Noise A (top) caused a steeply sloping recognition curve due to a major confusion with /di/ while noise B (bottom) produced a shallower recognition curve as the /di/ confusion was less pronounced. In the case of /gi/ (middle panels), noise A (top) and noise B (bottom) led to substantial differences in the recognition rate at most SNRs as noise A produced different and more pronounced confusions and thus lower recognition rates as compared to noise B. Regarding the results for /ni/ (right panels), noise A (top) yielded a more steeply sloping recognition curve than noise B (bottom). The confusions obtained with the two noise waveforms were the same but much more pronounced for noise A than for noise B.

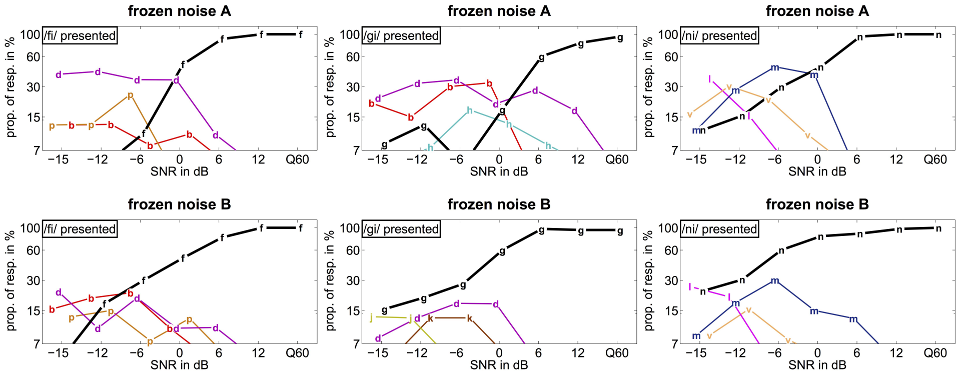


Figure 2.3: Across-noise token comparison of average confusion patterns for /fi/ (left panels), /gi/ (middle panels), and /ni/ (right panels). The upper and lower panels show the confusion patterns for the same speech token mixed with different waveforms of frozen noise (top: frozen noise A, bottom: frozen noise B). All speech tokens were spoken by male talker A. The only difference between frozen noise A and frozen noise B was a 100-ms temporal shift. The confusion patterns were obtained as described in Fig. 2.1.

2.3.3 Receiver-related variability

Here, examples of receiver-related variability are shown in terms of selected confusion matrices (CMs). The results demonstrate the observed effect of perceptual differences that occur across listeners and within listeners when no source-induced variability is present, i.e., for physically identical stimuli. An overall analysis of the results follows further below (Sec. 2.4).

Across-listener variability

Figure 2.4 shows the across-SNR average of CMs obtained in experiment 2 for four individual listeners. Only the responses obtained in noise B were considered here; thus, the speech and noise waveforms of the stimuli were identical across repeated stimulus presentations and across listeners. The left and right panels show two examples, each comparing the data obtained with two listeners. Each row in the CM reflects the across-SNR average of the proportions of responses obtained for a given speech token mixed with a given noise waveform. The circles indicate the proportions of responses; the filled gray circles show the data obtained with listeners 1 (left) and 3 (right) and the open red circles represent the data obtained with listeners 2 (left) and 4 (right). Thus, the amount of overlap between the filled gray circles and the open red circles indicates the agreement between the responses of listeners 1 and 2 (left) and listeners 3 and 4 (right). The figure hence illustrates the effect of across-listener variability.

Comparing the results of listeners 1 and 2 (left panel), considerable differences can be seen. For example, for /di/ and /ni/, listener 1 showed a larger recognition rate than listener 2, reflected along the diagonal where the filled gray circles exceed the open red circles in size. In contrast, for /fi/ and /mi/, listener 1 showed a smaller recognition rate than listener 2. Regarding confusions, represented by the off-diagonal circles in the CMs, a large variability was found. Some of the major confusions occurred in both listeners (e.g., /di/ confused with /gi/). However, the proportions of the individual confusions mostly differed, as indicated by the differences in the size of the overlapping off-diagonal filled gray and open red circles. Furthermore, many distinct confusions made by listener 1 (e.g., /fi/ confused with /bi/) were not made by listener 2 and vice versa.

Comparing the results of listeners 3 and 4 (right panel), the inter-individual differences in the results become even more apparent. The recognition rates (diagonal entries in the CMs) differed for /di/, /fi/, /gi/, /hi/, /li/, /mi/, /ni/, /pi/, and /vi/ (i.e., for nine out of fifteen CVs). Particularly in the case of /di/, listener 3 showed a high recognition rate, while listener 4 selected /gi/ instead of /di/ in about the same number of presentations. Some of the confusions were observed in both listeners, indicated by the overlapping off-diagonal filled gray and open red circles (e.g., /li/ confused with /vi/). However, the proportions of the shared confusions differed and most of the confusions made by listener 3 were not made by listener 4, as indicated by the non-overlapping off-diagonal filled gray and open red circles.

Within-listener variability

Figure 2.5 shows the across-SNR average of CMs obtained in test and retest of experiment 2 for two individual listeners. As above, only the responses obtained in frozen noise B were considered here; the speech and noise waveforms of the stimuli were therefore identical across repeated stimulus presentations and across test and retest. The illustration of the CMs is equivalent to the one used above. However, while Fig. 2.4 compared results across two pairs of listeners (listener 1 vs. 2 and listener 3 vs. 4), the left and right panels of Fig. 2.5 show the comparison of results obtained in test and retest for two individual listeners (listener 1 and listener 3). The figure therefore illustrates the effect of within-listener variability.

Listener 1 (left panel) showed fairly similar recognition rates in test and

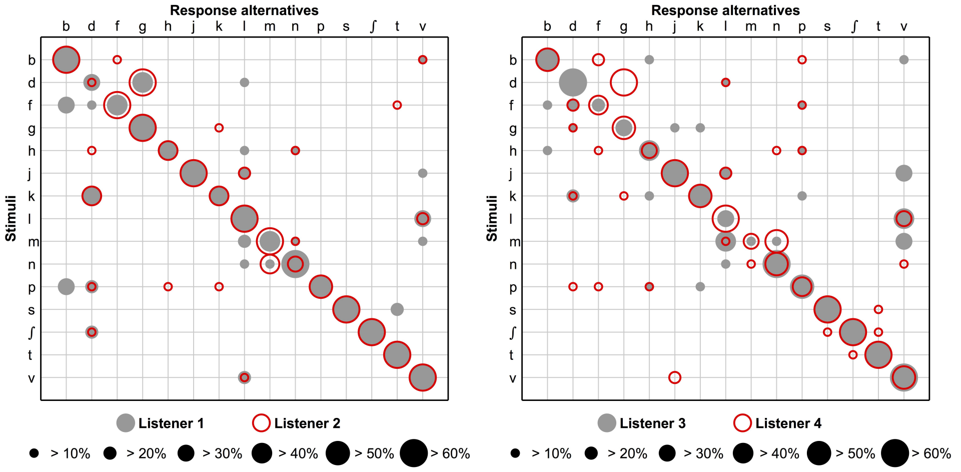


Figure 2.4: Across-listener comparison of confusion matrices for the 15 speech tokens used in experiment 2, mixed with frozen noise B. The speech and noise waveforms presented to the individual listeners were identical. For visual clarity, the responses were averaged across SNR conditions. For each stimulus (in each row of the matrix), the size of the circles indicates the proportions of responses for the individual response alternatives (columns of the matrix). Left: Responses of listener 1 (gray filled circles) vs. listener 2 (red open circles). Right: Responses of listener 3 (gray filled circles) vs. listener 4 (red open circles).

retest, as can be seen from the overlap of the filled gray and the open red circles along the diagonal. However, for /fi/, /gi/, /ki/, /pi/, and /vi/, the recognition rates were found to be slightly larger in the test (filled gray circles) than in the retest (open red circles). Regarding confusions, it can be seen that most of the major confusions were reproducible since most of the large off-diagonal circles overlap. However, the proportions of the confusions differed slightly across test and retest results, indicated by the differences in the sizes of the filled gray and the open red off-diagonal circles.

Listener 3 (right panel) also showed a large similarity of results obtained in test and retest. The recognition rates were found to be virtually identical, indicated by the perfect overlap of the on-diagonal filled gray and open red circles. Two exceptions were the results for /li/ and /mi/, where the recognition rate in the retest (open red circles) exceeded the recognition rate in the test (filled gray circles). As observed for listener 1, most of the major confusions were reproducible since most of the large off-diagonal circles share a large overlap while the proportions of the confusions partly differed between the test and retest results.

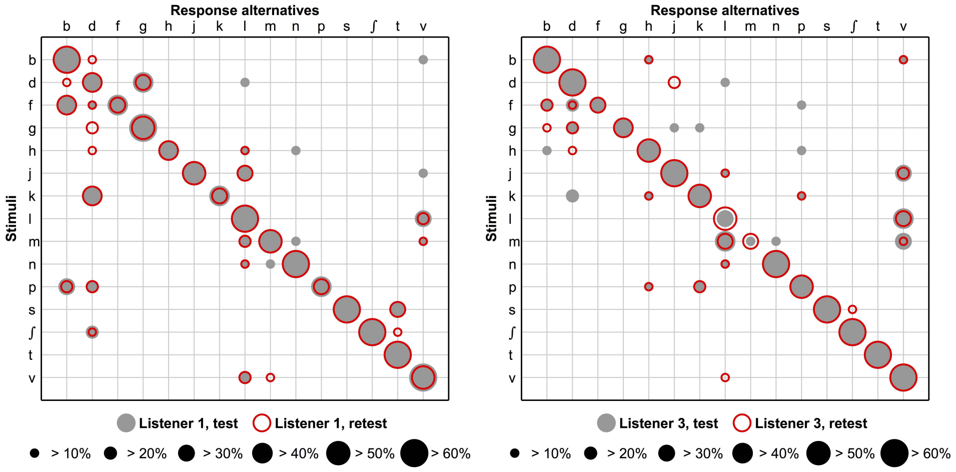


Figure 2.5: Within-listener comparison of confusion matrices for the 15 speech tokens used in experiment 2, mixed with frozen noise B. The speech and noise waveforms presented to the listeners in test and retest were identical. For visual clarity, the responses were averaged across SNR conditions. The confusion matrix depiction was obtained as in Fig. 2.4. Left: Responses of listener 1 obtained in test (gray filled circles) and retest (red open circles). Right: Responses of listener 3 obtained in test (gray filled circles) and retest (red open circles).

2.4 Analysis

The entire data set of the present study was analyzed in terms of source-induced and receiver-related effects using perceptual distance distributions as defined in Sec. 2.2.3. Figure 2.6 shows the mean perceptual distances, in percent, derived from the response variability across CVs (black), across talkers (blue), within talkers (green), across noise tokens (red), across listeners (light gray), and within listeners (dark gray), respectively, as a function of SNR. On the left, the average across SNR is shown. The error bars indicate the standard error across the underlying distributions of perceptual distance values obtained with the individual response pairs. The standard errors are proportional to the number of the respective considered response pairs, which varied greatly across the individual sources of variability (across CVs: 30240; across talkers: 1080; within talkers: 720; across noise: 120; across listeners: 3360; within listeners: 120).

The averages of the perceptual distances across SNR (leftmost bars) provide a good approximation of the size of the perceptual effects induced by the considered sources of variability. The reference for maximal perceptual distance, the perceptual distance across CVs (black bars), confirmed the expected large effect of consonant identity (91%). Regarding the source-induced perceptual distances across stimuli of the same phonetic identity, the largest perceptual distance of 51% was obtained for the across-talker condition (blue bar), followed by the perceptual distance of 47% obtained for the within-talker condition (green bar). This indicates that articulatory differences in utterances of a given talker had a perceptually comparable effect to articulatory differences in utterances of different talkers of different gender. The perhaps most striking observation was that even a slight temporal shift in the waveform of the noise masker mixed with the same speech token produced a considerable effect and led to a perceptual distance of 39% (red bar). Regarding the receiver-related effects, a substantial across-listener effect was found, corresponding to a perceptual distance of 46% for physically identical stimuli (light gray bar). This indicates a large variability in the consonant perception across NH listeners with similar language background. The across-listener effect was found to be as large as that resulting from within-talker variability (47%, green bar). In other words, the perceptual variability *across listeners* presented with physically identical stimuli was in the range of the perceptual variability in *individual listeners* induced by different speech tokens of the same phonetic type. In contrast, the relatively low perceptual distance within listeners of 30% (dark gray bar) indicated that the individual listeners were able to reproduce their responses fairly reliably.

Two-tailed paired-sample t-tests were performed to verify the statistical significance of the differences observed across the considered conditions. The across-CV reference condition was not considered here. The tests were conducted based on the across-SNR average of the obtained distance distributions. As the sample sizes for the individual conditions differed, with 120 being the minimum sample size, 120 observations were randomly chosen from each sample. The procedure was iterated 10.000 times for convergence and the resulting p-values and t-values were then averaged. A significance level of $\alpha = 0.05$ was assumed and divided by 10 to correct for the ten considered comparisons between the five remaining conditions ($\alpha_{\text{corr}} = 0.005$). The results are given in Table 2.1 and indicate that all considered conditions were significantly different from each other ($p < 0.005$) except the across-talker, within-talker and across-listener conditions.

The SNR-specific results show that the within-listener perceptual distance (dark gray bars in Fig. 2.6) increased with decreasing SNR. The lower the SNR, the more challenging was the task and the less reproducible were the responses of the individual listeners obtained with identical stimuli in test and retest.

Table 2.1: T-test results obtained for across-SNR average perceptual distance distributions. Bold numbers indicate p-values < 0.005, with 0.005 being the significance level after Bonferroni correction.

Conditions	t(119)	p
D_{wtnList} vs D_{acrTalk}	9.9448	**0.0000**
D_{wtnList} vs D_{wtnTalk}	8.2776	**0.0000**
D_{wtnList} vs D_{acrNoise}	4.6621	**0.0000**
D_{wtnList} vs D_{acrList}	8.2242	**0.0000**
D_{acrNoise} vs D_{acrTalk}	5.3434	**0.0000**
D_{acrNoise} vs D_{wtnTalk}	3.6554	**0.0031**
D_{acrNoise} vs D_{acrList}	3.6789	**0.0040**
D_{acrTalk} vs D_{wtnTalk}	-1.6744	0.2038
D_{acrTalk} vs D_{acrList}	-1.6130	0.2210
D_{wtnTalk} vs D_{acrList}	0.0441	0.5194

The within-listener response variability represents an intrinsic limitation and was therefore considered as the baseline ("internal noise"). The perceptual distances obtained across talkers (blue bars), within talkers (green bars), across noise tokens (red bars), and across talkers (light gray bars) increased along with – but were well above – the within-listener distance (as indicated by the shaded regions in Fig. 2.6). The perceptual distance across CVs, reflecting the maximal perceptual distance, showed the opposite trend since responses obtained with stimuli of different phonetic identity were compared: when the task was easy, the perceptual distance across responses obtained with stimuli of different phonetic identity was at ceiling (e.g., /bi/ and /di/ correctly recognized at large SNRs, response vectors thus orthogonal); with decreasing SNR the perceptual distance across these responses decreased as the recognition dropped and the number of confusions increased. Thus, while the perceptual distance across CVs (black bars) represented the largest contribution at all SNRs, it almost reached the level of the CV-specific perceptual distances for an SNR of -15 dB. Disregarding the influence of the listener uncertainty (within-listener distance, dark gray bars), the relation between the CV-specific perceptual distances remained almost the same at all SNRs. Thus, the across-SNR average distances described earlier capture the main effects observed at all SNRs.

Figure 2.7 shows the relation between the perceptual distance (abscissa) and the normalized entropy (ordinate) by means of a scatter plot. The respective

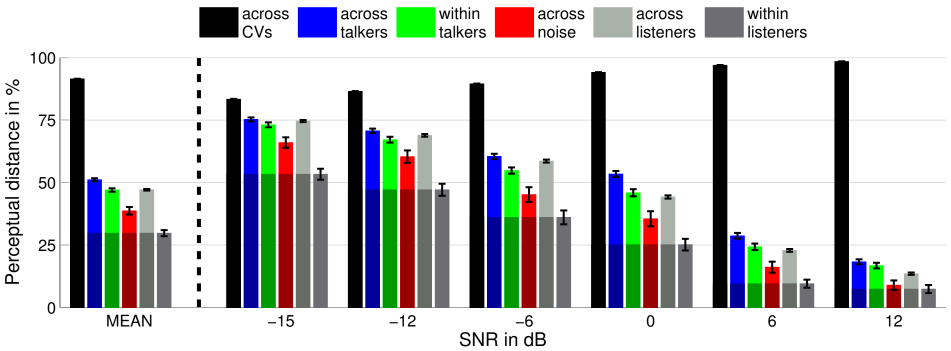


Figure 2.6: Mean perceptual distances as a function of SNR and averaged across SNR (left cluster). The error bars represent the standard error across the considered response pairs. As a reference for the maximum occurring perceptual distance, the perceptual distance across different CVs is shown (black bars). Comparing responses to physically different stimuli that share the same phonetic identity, the perceptual distances across talkers (blue bars), within talkers (green bars), and across frozen masking-noise tokens mixed with the same speech token (red bars) are depicted. Comparing responses across physically identical stimuli, the perceptual distances across listeners (light gray bars) and within listeners (dark gray bars) are shown. The shaded areas represent values below the within-listener distance, i.e., below the internal-noise baseline.

conditions are indicated by different colors and symbols, whereas the different SNRs are indicated by the size of the symbols. For large SNRs, the normalized entropy and the perceptual distance were almost fully correlated, i.e., the large symbols lie on top of the diagonal. For lower SNRs, the normalized entropy slightly exceeded the perceptual distance and the correlation between the two measures thus slightly decreased. Still, the overall correlation was 0.99. The SNR-specific correlation coefficients decreased with decreasing SNR but were all above 0.98. The results obtained based on the perceptual distance are thus supported by the concept of entropy.

2.5 Discussion

2.5.1 Summary of main findings

The present study investigated the effects of different sources of variability in NH listeners' perception of consonants presented in steady-state masking noise. Two main categories of perceptual variability were defined: source-induced and receiver-related variability. The former describes perceptual differences caused by acoustical differences in stimuli of the same phonetic identity and was subdivided into speech-induced variability (across talkers and within talkers)

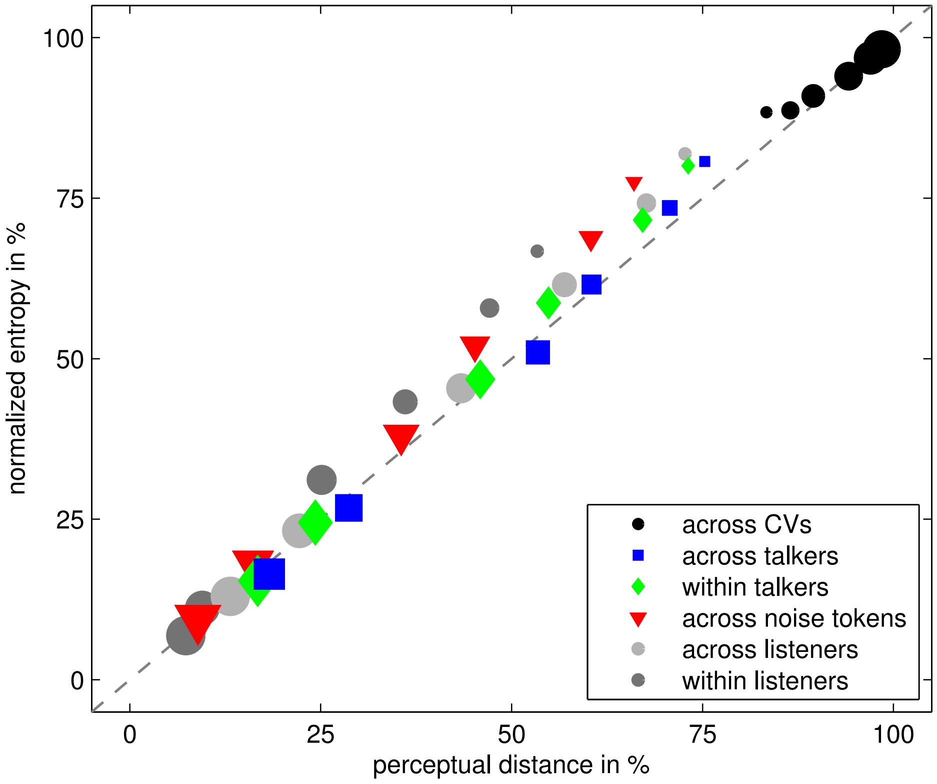


Figure 2.7: Scatter plot of perceptual distance in percent (abscissa) versus normalized entropy in percent (ordinate) for the different considered conditions: across CVs (black circles), across talkers (blue squares), within talkers (green diamonds), across noise tokens (red triangles), across listeners (light gray circles), and within listeners (dark gray circles). The perceptual distance and normalized entropy values obtained for the six different SNR conditions (12, 6, 0, -6, -12, and -15 dB) are plotted against each other. The sizes of the respective symbols are proportional to the SNR values. The gray diagonal dashed line represents perfect correlation of perceptual distance and normalized entropy.

and noise-induced variability. A special case of source-induced variability is the variability across consonants, which has been considered here as a reference for maximal variability. The latter comprises perceptual differences across listeners and within listeners. To quantify the relative influence of the individual sources of variability, the responses obtained in two experiments were analyzed in terms of example comparisons using a subset of the data and by means of a perceptual distance measure and the entropy of responses using the entire data set.

Regarding the source-induced variability for stimuli of the same phonetic identity, it was shown that the largest perceptual variability was induced by across-talker articulatory differences, closely followed by the effect of within-talker articulatory differences. Furthermore, even a slight temporal shift in the waveform of the steady-state masking noise was found to produce a smaller, yet clearly measurable and statistically significant perceptual effect. Regarding receiver-related variability, the analysis showed that, for physically identical stimuli, the perceptual differences across the NH listeners were very large (in the range of the speech-induced differences). In contrast, the within-listener variability (listener uncertainty) was found to be much smaller, indicating that the reproducibility of the responses for individual listeners was much larger than the agreement between the responses of different listeners. The within-listener variability depended inversely on the SNR, i.e., the “internal noise” (listener uncertainty) was proportional to the “external noise” (acoustic noise).

2.5.2 Relation to other studies

In the present study, a large perceptual effect of across-talker articulatory differences was found for identical CVs. This is consistent with other recent studies on consonant perception (e.g., Phatak et al., 2008; Singh and Allen, 2012; Toscano and Allen, 2014), which demonstrated that different speech tokens of the same phonetic identity spoken by different talkers elicit largely different percepts. In contrast, early studies on consonant perception (e.g., Miller and Nicely, 1955; Wang and Bilger, 1973) pooled the responses obtained with different speech tokens of the same phonetic identity spoken by different talkers, thus neglecting the talker-specific perceptual details.

The effect of within-talker articulatory differences was in the present study found to be almost as large as the effect of across-talker articulatory differences. This has otherwise not been reported yet since related studies on consonant perception typically used only one speech token from a given talker for each CV. A within-talker effect was expected given the natural within-talker articulatory variability; however, the authors of the present study did not expect such a prominent effect.

A significant perceptual effect of a temporal shift in the masking noise waveform was found, demonstrating that different white-noise waveforms, mixed with the same speech token, can elicit different speech percepts. Thus, the common assumption in various previous studies (e.g., Miller and Nicely, 1955; Phatak and Allen, 2007; Phatak et al., 2008) of an invariance of consonant perception across steady-state noise realizations cannot be supported by the present study. In fact, the results obtained here suggest that the interaction between a given speech token and the spectro-temporal details of the "steady-state" masking noise waveform matter in the context of microscopic consonant perception. When analyzing responses obtained with individual speech tokens (as in Li et al., 2010; Li et al., 2012; Singh and Allen, 2012; Toscano and Allen, 2014), averaging responses across noise realizations thus appears problematic.

Furthermore, the results of the present study showed that, even for physically identical stimuli, the across-listener perceptual variability is large. The within-listener perceptual variability was found to be clearly smaller. Studies on consonant perception in NH listeners (e.g., Miller and Nicely, 1955; Phatak and Allen, 2007; Phatak et al., 2008; Toscano and Allen, 2014) relied solely on across-listener average data without assessing deviations from the across-listener average due to inter-individual perceptual differences. Toscano and Allen (2014) stated that listeners were highly consistent without providing explicit evidence for this claim. Their analysis was based on consonant recognition only while the analysis performed in the present study also took consonant confusions into account, which may yield different results. Nevertheless, the assumption that consonant perception of NH listeners with similar language background is as consistent across listeners as within listeners is in contrast to the results of the present study. Therefore, across-listener average data should be treated as a population response that is not representative of individual listeners (and vice versa).

2.5.3 Implications for the design of consonant perception experiments

The present study demonstrated that all considered differences in the speech token and/or in the noise token led to different consonant percepts. Further, the perceptual variability across NH listeners with the same language background was found to be large. The implications of these findings for the design of consonant experiments largely depend on the goal of the respective study.

If the goal is to "globally" assess consonant perception as a function of consonant identity and SNR, it should be ensured that the described sources of variability (source-induced and receiver-related) do not bias the resulting data. Thus, (i) many speech tokens spoken by different talkers should be considered for each consonant to cover the speech-induced variability, (ii) randomly generated masking noise should be employed to cover the noise-induced variability, and (iii) many listeners should be tested to cover the across-listener perceptual variability. The responses may then be averaged across different speech tokens of the same phonetic identity, different noise waveforms, and different listeners, yielding an overall pattern of consonant perception as a function of consonant identity and SNR. A more realistic description of the data obtained in such an experiment may be achieved by interpreting the responses obtained with each considered CV as multi-dimensional probability distributions across speech tokens, noise tokens, listeners, and SNR.

In contrast, if the purpose of the study is to investigate which acoustic cues determine a specific confusion pattern, (i) the responses need to be evaluated for each speech token separately (since different speech tokens can elicit different speech percepts), (ii) the combination of speech token and masking-noise token needs to be unique (since the use of randomly generated masking noise mixed with identical speech tokens can elicit different speech percepts in each trial), and (iii) the responses need to be evaluated in individual listeners (due to the substantial perceptual differences across listeners). This level of detail is also needed when assessing effects of individual hearing impairment and hearing-aid signal processing via consonant perception tests. If the above constraints are not respected, the observed results may well be blurred by speech-induced variability, noise-induced variability, and across-listener perceptual variability. Recent detailed studies of consonant cues (Li et al., 2010; Li et al., 2012) indeed analyzed the data for each speech token individually. However, random realizations of steady-state masking noise were used for each trial in the masking experiments and the analyses were performed based on across-listener average data.

2.5.4 Implications for consonant perception modeling

So far, no model has been proposed that is able to predict consonant perception in terms of recognition and confusions. The results of the present study may provide some general constraints for microscopic models of speech perception.

If the goal is to predict the average responses for a given consonant and SNR obtained with many speech tokens, many noise realizations, and many listeners, the model's responses should reflect the same *average* outcome measures obtained with the same set of stimuli. Such a "global" model would not be designed to account for the sources of variability considered in the present study, but may account for the effects of consonant identity and SNR. For such an approach, the observations from the present study motivated that the decision-making process in the model back end should incorporate an internal-noise term that scales with the amount of external noise represented in the stimulus.

If a model of consonant perception is targeted towards more details in the consonant perception results, the model needs to reflect all sources of variability. For instance, the fact that a temporal shift in the noise waveform can lead to substantial perceptual differences indicates that a suitable model front end should be sensitive to signal-to-masker phase relations probably already in the peripheral processing of the stimuli. Furthermore, similar to the considerations regarding the "global" model, the observed relation between SNR and within-listener variability suggests that an internal-noise term which scales with the amount of external noise in the stimulus should be incorporated in the model back end. The observed large across-listener perceptual variability represents a major challenge for modeling, since this variability can either arise from differences in the sensory processing in the individual listeners, or from differences at higher-level processes, or both. Such differences may occur even in the case of NH listeners (as considered in the present study), since (i) the applied "criterion" for NH was not very strict, (ii) the audiogram may not be a sufficient descriptor for sensory processing, and (iii) higher-level speech processing may differ across listeners independent of their sensory capabilities, e.g., due to different cognitive abilities. How these inter-individual differences across NH listeners can be quantified and eventually integrated into a modeling framework remains a major challenge.

2.5.5 Limitations of the approach

In the present study, several parameters that are known to play a perceptual role were *fixed* to solely focus on specific sources of variability, as it is not feasible to test all possible factors at once. Specifically, the vowel (/i/), the type of consonant-vowel combination (CV) and the spectral shape of the noise (white) were fixed. The same holds for the choice of response alternatives, response method, and the instructions given to the test subjects. The influence of these parameters, which also represent sources of variability, was thus neglected.

The claims made in the present study were based on a set of 90 speech tokens, spoken by two talkers, and presented to two different panels of eight listeners (experiment 1 and experiment 2) and a panel of four listeners (retest of experiment 2), respectively. Therefore, the results may be biased by the choice of speech tokens, talkers, and listeners. Furthermore, the relative size of the reported effects may be different when considering speech samples obtained from natural speech utterances as opposed to isolated syllable productions.

The individual sources of variability investigated in the present study represent *categories* and only provide indications about the relative contributions of these categories (e.g., across-talker articulatory differences) to consonant perception. Thus, it remains unanswered here which specific acoustical properties of the stimuli caused the observed perceptual differences. The three-dimensional deep search method introduced by Li et al. (2010) might be one way of addressing this question in terms of a spectro-temporal analysis. Another approach might be to identify the importance of consonant cue regions as a function of audio frequency and modulation frequency, as suggested by Christiansen et al. (2007). Further investigations are required to provide an understanding of the relationship between acoustic features in the noisy speech waveform, their internal representation in the auditory system, and the contribution of the different features to robust phoneme recognition.

2.6 Summary and conclusions

An experimental approach to investigate the influence of various sources of variability in consonant perception was presented. The study focused on the consonant perception of NH listeners presented with CVs in white noise at different SNRs. The perceptual variability was split into two main categories, source-induced and receiver-related variability. Using example-based comparisons for the different conditions, and quantifying the observations by means of a measure of perceptual distance, the relative importance of the individual sources of variability was described. Regarding the source-induced variability, the largest effect was found for across-talker articulatory differences, followed by within-talker articulatory differences. Furthermore, even the waveform of the masking noise was shown to induce a significant perceptual effect. In terms of receiver-related effects, a large variability of the responses across listeners was found, whereas the within-listener variability was rather small. Furthermore, the within-listener variability (i.e., the “internal noise”) was found to be proportional to the amount of masking noise (“external noise”) in the stimulus.

The results from the present study complement current knowledge on consonant perception. It is suggested that, in addition to speech-induced variability, also noise-induced variability as well as across-listener perceptual variability should be taken into account, which has implications for the design of consonant perception experiments and models of consonant perception.

Acknowledgments

We would like to thank Søren Jørgensen for various discussions about the design of the study and the data analysis, and Christoph Scheidiger for discussions regarding the perceptual distance measure. We also thank two anonymous reviewers for their very helpful and supportive comments. J. Z. is a fellow in the initial training network INSPIRE, which has been funded with support from the European Commission under Contract FP7-PEOPLE-2011-290000.

2.7 Appendix

2.7.1 Description of the data set

The data set considered for the analysis comprised various factors. Table 2.2 provides an overview of the dimensionality of the data set. For clarity, the Q45 and Q60 quiet conditions were not considered here. In experiment 1, 3 recordings from each of the 2 talkers of each of the 15 CVs were used. All speech tokens were mixed with white noise at 6 different SNRs and presented to 8 listeners. The listeners had to select one of 16 response alternatives (15 consonants and "I don't know"). Each speech token was presented 3 times at each SNR ($\mathrm{N}_{\mathrm{trial1}} = 3$). In experiment 2, only 1 speech token was used for each of the 15 CVs. All speech tokens were mixed with 2 different white noise waveforms at 6 different SNRs and presented to 8 listeners. The random masking-noise condition from this experiment was neglected here since it was not used in the analysis. Again, the listeners had to select one of the same 16 response alternatives. Each combination of speech token and noise waveform was presented 5 times at each SNR ($\mathrm{N}_{\mathrm{trial2}} = 5$). The retest of experiment 2 was conducted with only $\mathrm{N}_{\mathrm{list,retest}} = 4$ of the 8 listeners.

The responses obtained in experiment 1 are denoted as a function $\mathbf{R_I}(c, \tau, \rho, s, l, v)$. Accordingly, the responses obtained in experiment 2 are denoted $\mathbf{R_{II}}(c, \eta, s, l, v)$. Table 2.2 describes the function variables. For each feasible combination of variables, the functions $\mathbf{R_I}$ and $\mathbf{R_{II}}$ return the vectors

Table 2.2: Overview of the entire data set along with the mathematical notation used for the individual factors.

Factors	Variable name	Experiment 1	Experiment 2
CVs	c	$N_{CV} = 15$	$N_{CV} = 15$
Talkers	τ	$N_{talk} = 2$	– (1)
Recordings	ρ	$N_{rec} = 3$	– (1)
Masking-noise conditions	η	– (1)	$N_{noise} = 2$
SNRs	s	$N_{SNR} = 6$	$N_{SNR} = 6$
Listeners	l	$N_{list} = 8$	$N_{list} = 8$
Response alternatives	–	$N_{resp} = N_{CV} + 1$	$N_{resp} = N_{CV} + 1$
Trials	ν	$N_{trial1} = 3$	$N_{trial2} = 5$

$\mathbf{r_I} = [r_{I,1}, r_{I,2}, ..., r_{I,N_{resp}}]$ and $\mathbf{r_{II}} = [r_{II,1}, r_{II,2}, ..., r_{II,N_{resp}}]$, respectively, which have a length of N_{resp}, a value of "1" for the element corresponding to the chosen response, and a value of "0" for all other elements. The last element of these vectors corresponds to the "I don't know" response and all other elements correspond to the 15 consonants provided as response alternatives. The proportions of responses were obtained by distributing the "I don't know" responses evenly across the 15 other response alternatives, summing the responses across all trials, and finally dividing by the number of trials. For the data $\mathbf{R_I}$ obtained in experiment 1, the conversion of the responses obtained with a given CV, talker, recording, SNR, and listener is expressed as:

$$p_{I,i} = \frac{1}{N_{trial1}} \sum_{\nu=1}^{N_{trial1}} r_{I,i}(\nu) + \frac{r_{I,N_{resp}}(\nu)}{N_{CV}}, \qquad i = 1,\ 2,\ ...,\ N_{CV}. \tag{2.2}$$

For the data $\mathbf{R_{II}}$ obtained in experiment 2, the conversion of the responses obtained with a given CV, masking-noise waveform, SNR, and listener is provided by:

$$p_{II,i} = \frac{1}{N_{trial2}} \sum_{\nu=1}^{N_{trial2}} r_{II,i}(\nu) + \frac{r_{II,N_{resp}}(\nu)}{N_{CV}}, \qquad i = 1,\ 2,\ ...,\ N_{CV}. \tag{2.3}$$

The resulting vectors $\mathbf{p_I} = [p_{I,1}, p_{I,2}, ..., p_{I,N_{CV}}]$ and $\mathbf{p_{II}} = [p_{II,1}, p_{II,2}, ..., p_{II,N_{CV}}]$ contain the respective proportions of responses and are summarized as the functions $\mathbf{P_I}(c,\tau,\rho,s,l)$ and $\mathbf{P_{II}}(c,\eta,s,l)$, representing the proportions of responses obtained in the two experiments.

2.7.2 Calculation of perceptual distance

The perceptual distance measure is defined in Sec. 2.2.3. The perceptual distance *across CVs* was calculated across all response pairs obtained with speech tokens of different phonetic identity based on $\mathbf{P_I}$,

$$\mathrm{D}_{\mathrm{acrCV}}(s,l,\tau,\tau',\rho,\rho',\delta) = \mathrm{D}[\mathbf{P_I}(c,\tau,\rho,s,l),\ \mathbf{P_I}(c',\tau',\rho',s,l)] \tag{2.4}$$

where $c = [1,\ \mathrm{N_{CV}}-1]$, $c' = [c+1,\ \mathrm{N_{CV}}]$, $\tau = [1,\ \mathrm{N_{talk}}]$, $\tau' = [1,\ \mathrm{N_{talk}}]$, $\rho = [1,\ \mathrm{N_{rec}}]$, and $\rho' = [1,\ \mathrm{N_{rec}}]$. The variable $\delta = [1,\ \mathrm{N}_\delta]$ describes all $\mathrm{N}_\delta = \sum_{n=1}^{\mathrm{N_{CV}}-1} n$ possible combinations of CV identities. For each SNR, the across-CV distances between $\mathrm{N_{list}} \cdot \mathrm{N_{talk}^2} \cdot \mathrm{N_{rec}^2} \cdot \mathrm{N}_\delta = 30240$ response vector pairs were calculated.

The perceptual distance *across talkers* was calculated across all response pairs obtained with speech tokens of the same phonetic identity spoken by different talkers, based on $\mathbf{P_I}$,

$$\mathrm{D}_{\mathrm{acrTalk}}(s,l,c,\rho,\rho') = \mathrm{D}[\mathbf{P_I}(c,\tau,\rho,s,l),\ \mathbf{P_I}(c,\tau',\rho',s,l)] \tag{2.5}$$

where $\tau = 1$ and $\tau' = 2 = \mathrm{N_{talk}}$, $\rho = [1,\ \mathrm{N_{rec}}]$, and $\rho' = [1,\ \mathrm{N_{rec}}]$. For each SNR, the across-talker distances between $\mathrm{N_{list}} \cdot \mathrm{N_{CV}} \cdot \mathrm{N_{rec}^2} = 1080$ response vector pairs were calculated.

The perceptual distance *within talkers* was calculated across all response pairs obtained with different speech tokens of the same phonetic identity, spoken by the same talker, based on $\mathbf{P_I}$,

$$\mathrm{D}_{\mathrm{wtnTalk}}(s,l,c,\tau,\delta) = \mathrm{D}[\mathbf{P_I}(c,\tau,\rho,s,l),\ \mathbf{P_I}(c,\tau,\rho',s,l)] \tag{2.6}$$

where $\rho = [1,\ \mathrm{N_{rec}}-1]$ and $\rho' = [\rho+1,\ \mathrm{N_{rec}}]$. The variable $\delta = [1,\ \mathrm{N}_\delta]$ describes all $\mathrm{N}_\delta = \sum_{n=1}^{\mathrm{N_{rec}}-1} n$ possible combinations of recordings from a given talker. For each SNR, the within-talker distances between $\mathrm{N_{list}} \cdot \mathrm{N_{CV}} \cdot \mathrm{N_{talk}} \cdot \mathrm{N}_\delta = 720$ response vector pairs were calculated.

The perceptual distance *across noise tokens* was calculated across all response pairs obtained with identical speech tokens mixed with two different frozen noise tokens based on $\mathbf{P_{II}}$,

$$\mathrm{D}_{\mathrm{acrNoise}}(s,l,c)=\mathrm{D}[\mathbf{P_{II}}(c,\eta,s,l),\ \mathbf{P_{II}}(c,\eta',s,l)] \tag{2.7}$$

where $\eta = 1$ and $\eta' = 2 = \mathrm{N}_{\mathrm{noise}}$. For each SNR, the across noise-token distances between $\mathrm{N}_{\mathrm{list}} \cdot \mathrm{N}_{\mathrm{CV}} = 120$ response vector pairs were calculated.

The perceptual distance *across listeners* was calculated across all response pairs obtained with physically identical stimuli but different listeners based on $\mathbf{P_I}$ and $\mathbf{P_{II}}$,

$$\mathrm{D}_{\mathrm{acrList,I}}(s,c,\tau,\rho,\delta)=\mathrm{D}[\mathbf{P_I}(c,\tau,\rho,s,l),\ \mathbf{P_I}(c,\tau,\rho,s,l')] \tag{2.8}$$

$$\mathrm{D}_{\mathrm{acrList,II}}(s,c,\eta,\delta)=\mathrm{D}[\mathbf{P_{II}}(c,\eta,s,l),\ \mathbf{P_{II}}(c,\eta,s,l')] \tag{2.9}$$

where $l=[1,\ \mathrm{N}_{\mathrm{list}}-1]$ and $l'=[l+1,\ \mathrm{N}_{\mathrm{list}}]$. The variable $\delta=[1,\ \mathrm{N}_{\delta}]$ describes all $\mathrm{N}_{\delta}=\sum_{n=1}^{\mathrm{N}_{\mathrm{list}}-1} n$ possible combinations of listeners. For each SNR, the across-listener distances between $\mathrm{N}_{\mathrm{CV}} \cdot \mathrm{N}_{\mathrm{talk}} \cdot \mathrm{N}_{\mathrm{rec}} \cdot \mathrm{N}_{\delta} = 2520$ response vector pairs were calculated from $\mathbf{P_I}$ and between $\mathrm{N}_{\mathrm{CV}} \cdot \mathrm{N}_{\mathrm{noise}} \cdot \mathrm{N}_{\delta} = 840$ response vector pairs from $\mathbf{P_{II}}$. $\mathrm{D}_{\mathrm{acrList,I}}$ and $\mathrm{D}_{\mathrm{acrList,II}}$ were combined to $\mathrm{D}_{\mathrm{acrList}}$, (comprising $2520+840=3360$ distance values per SNR).

The average perceptual distance *within listeners* was calculated across all response pairs obtained with physically identical stimuli and identical listeners in test and retest of experiment 2, i.e., based on $\mathbf{P_{II,test}}$ and $\mathbf{P_{II,retest}}$,

$$\mathrm{D}_{\mathrm{wtnList}}(s,c,\eta,l)=\mathrm{D}[\mathbf{P_{II,test}}(c,\eta,s,l),\ \mathbf{P_{II,retest}}(c,\eta,s,l)] \tag{2.10}$$

where $l=[1,\ \mathrm{N}_{\mathrm{list,retest}}]$. For each SNR, the within-listener distances between $\mathrm{N}_{\mathrm{list,retest}} \cdot \mathrm{N}_{\mathrm{CV}} \cdot \mathrm{N}_{\mathrm{noise}} = 120$ response vector pairs were calculated.

2.7.3 Calculation of normalized entropy

In analogy to the calculation of the perceptual distance, the data were also analyzed in terms of entropy. The entropy specifies the amount of variability in a given response vector. Here, the normalized entropy was used, which was defined as:

$$H_{norm}(\mathbf{p})=\frac{100\%}{\log_2(\min[\mathrm{R},\mathrm{N}])}\cdot\sum_{i=1}^{R}\mathrm{p}_i\log_2\left(\frac{1}{\mathrm{p}_i}\right), \qquad \forall\ \mathrm{p}_i>0 \tag{2.11}$$

where $\mathbf{p} = [\mathrm{p}_1, \mathrm{p}_2, ..., \mathrm{p}_\mathrm{R}], \mathrm{p}_\mathrm{i}$ is the proportion of response alternative i, R denotes the number of response alternatives, and N represents the number of observations. The denominator is the theoretical entropy maximum $H_{max} = \log_2(\min[\mathrm{R}, \mathrm{N}])$. Division by H_{max} thus normalizes the entropy to a range from 0 to 1; multiplication by 100% yields the normalized entropy in percent.

The normalized entropy describes the perceptual variability *for a given* response vector whereas the perceptual distance measure represents a comparison between *a pair* of response vectors. To quantify the effect of the different sources of variability based on the normalized entropy, the perceptual variability induced by the different sources of variability therefore had to be contained in individual response vectors. To obtain such response vectors, the raw data $\mathbf{R_I}$ and $\mathbf{R_{II}}$ were converted to proportions of responses obtained (on a trial-by-trial basis) with (1) different CVs, (2) identical CV but different talkers, (3) identical CV and talker but different recordings, (4) identical speech token but different masking-noise waveforms, (5) identical stimulus but different listeners, (6) identical stimulus and listener but test and retest results. The “I don’t know” responses were here attributed to randomly chosen response alternatives. The entropy is sensitive to differences in N, the number of observations. To avoid an obscured across-condition comparison, it was therefore necessary to have the same basic number of observations for all the considered conditions that were to be compared. As this was not the case in the data set (e.g., $\mathrm{N_{CV}} = 15$ but $\mathrm{N_{talk}} = 3$), a “fair” entropy comparison could only be obtained using random processes and iterating them many times for the resulting entropy to converge to its true value. The effective number of observations was set to $\mathrm{N} = 3$, which corresponds to $\mathrm{N_{trial1}}$ and $\mathrm{N_{rec}}$. To illustrate the procedure, three examples are given that represent the cases (1) $\mathrm{N_{factor}} > \mathrm{N}$, (2) $\mathrm{N_{factor}} < \mathrm{N}$, and (3) $\mathrm{N_{factor}} = \mathrm{N}$.

(1) $\mathrm{N_{factor}} > \mathrm{N}$. The normalized entropy *across CVs* ($\mathrm{N_{CV}} = 15$), $\mathrm{E_{acrCV}}$, was calculated based on $\mathbf{R_I}$. For each SNR condition and for each listener, 3 vectors $\mathbf{r_I}$ obtained in 3 randomly chosen trials with 3 different randomly chosen CVs were collected, summed, and converted to a proportions of response vector $\mathbf{p}$ via division by $\mathrm{N} = 3$. In order for the random processes to converge, the procedure was iterated 1000 times. The normalized entropy was calculated and eventually averaged across listeners and iterations.

(2) $\mathrm{N_{factor}} < \mathrm{N}$. The normalized entropy *across talkers* ($\mathrm{N_{talk}} = 2$), $\mathrm{E_{acrTalk}}$, was calculated based on $\mathbf{R_I}$. For each SNR condition, for each CV, and for each

listener, 3 vectors $\mathbf{r_I}$ obtained in 3 randomly chosen trials with (1) one randomly chosen recording of talker A, (2) one randomly chosen recording of talker B, and (3) one recording of talker A or talker B (randomly chosen from the residual recordings) were collected, summed, and converted to a proportions of response vector **p** via division by N = 3. The procedure was iterated 1000 times for convergence. The normalized entropy was calculated and averaged across CVs, listeners, and iterations.

(3) $\mathrm{N_{factor}} = \mathrm{N}$. The normalized entropy *within talkers* ($\mathrm{N_{rec}} = 3$), $\mathrm{E_{wtnTalk}}$, was calculated based on $\mathbf{R_I}$. For each SNR condition, for each CV, for each talker, and for each listener, 3 vectors $\mathbf{r_I}$ obtained in 3 randomly chosen trials with the 3 different recordings spoken by the respective talker were collected, summed, and converted to a proportions of response vector **p** via division by N = 3. The procedure was iterated 1000 times for convergence. The normalized entropy was calculated and averaged across CVs, talkers, listeners, and iterations.

Similar calculations were performed to obtain the normalized entropy *across noise tokens* ($\mathrm{E_{acrNoise}}$), *across listeners* ($\mathrm{E_{acrList}}$), and *within listeners* ($\mathrm{E_{wtnList}}$).

3

Predicting consonant recognition and confusions in normal-hearing listeners[b]

Abstract

The perception of consonants in background noise has been investigated in various studies and was shown to critically depend on fine details in the stimuli. In this study, a microscopic speech perception model is proposed that represents an extension of the auditory signal processing model by Dau et al. [(1997). J. Acoust. Soc. Am. 102, 2892–2905]. The model was evaluated based on the extensive consonant perception data set provided by Zaar and Dau [(2015). J. Acoust. Soc. Am. 138, 1253–1267], which was obtained with normal-hearing listeners using 15 consonant-vowel combinations (CVs) mixed with white noise. Accurate predictions of the consonant recognition scores were obtained across a large range of signal-to-noise ratios. Furthermore, the model yielded convincing predictions of the consonant confusion scores, such that the predicted errors were clustered in perceptually plausible confusion groups. The large predictive power of the proposed model suggests that adaptive processes in the auditory preprocessing in combination with a cross-correlation based template-matching back end can account for some of the processes underlying consonant perception in normal-hearing listeners. The proposed model may provide a valuable framework, e.g., for investigating the effects of hearing impairment and hearing-aid signal processing on phoneme recognition.

[b] This chapter is based on Zaar and Dau (2016).

3.1 Introduction

The way how humans decode speech has been investigated from various perspectives. Most commonly, the percentage of correctly identified words or sentences is assessed in the presence of some acoustical interference or degradation, such as additive noise and/or reverberation. The speech reception threshold (SRT), i.e., the signal-to-noise ratio (SNR) at which, e.g., 50% correct responses are obtained, has often been used to describe the properties of the transmission channel and/or the receiver (cf. Hagerman, 1982; Nilsson et al., 1994; Wagener et al., 2003; Nielsen and Dau, 2009; Nielsen and Dau, 2011). Such speech tests provide some useful *macroscopic* information about limiting effects induced by the acoustic conditions or the global speech reception ability of listeners. However, the SRT measure is rather coarse as it reflects responses averaged across many speech tokens. Furthermore, the listeners' performance may be strongly influenced by cognitive effects as listeners can restore missing acoustic information using semantic predictability and lexical information (e.g., Miller and Licklider, 1950; Warren, 1970; Bashford et al., 1992; Kashino, 2006).

Speech perception has also been studied at a more basic level using a *microscopic* approach. Several studies have reported consistent misperceptions of isolated words (e.g. Cooke, 2009; Tóth et al., 2015), typically collected in conditions of speech-on-speech masking using an open response set. Such an approach excludes semantic predictability while taking the language-specific lexical possibilities for misperceptions into account. Various other studies have focused on the perception of consonants embedded in nonsense syllables (e.g., Miller and Nicely, 1955; Wang and Bilger, 1973; Phatak and Allen, 2007; Phatak et al., 2008; Zaar and Dau, 2015), e.g. in the form of consonant-vowel combinations (CVs like /ba/, /ta/, etc.), typically presented in steady-state noise at various SNRs in the context of a closed response set. This approach has the advantage that (i) the contribution of higher-level semantic *and* lexical effects is eliminated due to the nonsense nature of the stimuli and that (ii) the importance of the critical[1] high-frequency speech cues is emphasized as many consonants contain high-frequency energy (cf. Li et al., 2010; Li et al., 2012). These aspects make consonant perception measurements an interesting tool for assessing

[1] Many hearing-impaired listeners suffer from a loss of sensitivity at high-frequencies, which affects their speech perception but is difficult to measure using macroscopic tests with meaningful, low-frequency dominated speech.

the effects of acoustical transmission channels as well as the effects of hearing impairment and hearing-aid signal processing on fundamental speech cues.

Miller and Nicely (1955) investigated consonant perception in terms of consonant recognition and confusions, such that not only the amount of errors but also the patterns of confusions were analyzed. Their study suggested that distinct perceptual confusions among consonants may have a major effect on speech intelligibility in noise. Miller and Nicely (1955) and related studies (e.g., Wang and Bilger, 1973) used many speech tokens to represent each consonant. The obtained responses were averaged across tokens such that the data were represented as a function of consonant identity. This analysis approach was later shown to misrepresent the data since substantial perceptual differences across different speech tokens of the same phonetic identity were observed (Phatak et al., 2008; Singh and Allen, 2012; Toscano and Allen, 2014). Zaar and Dau (2015) employed a measure of the perceptual distance between responses obtained with CVs presented in noise to investigate the influence of various sources of perceptual variability on consonant perception. Consistent with the aforementioned studies, different speech tokens of the same phonetic identity were found to induce a large perceptual variability. Moreover, even a slight temporal shift in the steady-state masking noise waveform was shown to induce a perceptual effect when presented along with the same speech token. On the receiver side, it was found that different normal-hearing (NH) listeners with the same language background showed large perceptual differences when presented with identical stimuli, whereas the individual listeners could reproduce their responses fairly reliably in a retest. Overall, the listeners' sensitivity to fine differences in the stimuli suggests that measures of consonant perception represent a detailed descriptor of the listeners' sensory processing.

To better understand how specific effects in consonant perception are related to differences in sensory processing, computational models of speech perception may be insightful. Various *macroscopic* speech intelligibility models have been presented, which are all based on simulations of the auditory periphery in terms of frequency selectivity (e.g., ANSI, 1969; ANSI, 1997; Rhebergen et al., 2006), while some models also consider modulation-frequency selective processing (e.g., Houtgast et al., 1980; Payton and Braida, 1999; Jørgensen and Dau, 2011; Jørgensen et al., 2013). Based on the assumption that speech intelligibility is monotonically related to the speech-to-noise power ratio in the considered domain, these macroscopic models have been shown to account well for average SRTs in various acoustic conditions. Only a few modeling studies have addressed *microscopic* speech perception, where typically elaborate models of the auditory periphery have been combined with a speech recognition back end to predict nonsense syllable perception. As "blind" automatic speech recognition (ASR) systems perform much worse than human listeners in terms of phoneme recognition (e.g., Sroka and Braida, 2005; Meyer et al., 2011), all microscopic speech perception models presume some kind of *a-priori* information about the stimuli to reduce the gap to human recognition performance.

Messing et al. (2009) used a non-linear model of the auditory periphery with a feedback mechanism in combination with a simplistic template matching back end (using "frozen speech", i.e., a-priori knowledge about the presented speech token) to predict results of a diagnostic rhyme test (DRT) obtained with NH listeners. The predictions matched the data quite well in terms of the errors as a function of phonetic attributes. Jepsen et al. (2014) applied a similar approach to model DRT results in hearing-impaired (HI) listeners, using a different non-linear auditory model that includes an adaptation process and a modulation filterbank (Jepsen et al., 2008). However, the two studies used highly controlled *synthetic* consonant-vowel-consonant (CVC) syllables mixed with speech-shaped noise (SSN). Thus, it remained unclear to what extent these models could generalize to the less controlled case of *natural* speech stimuli.

Cooke (2006) predicted NH listeners' consonant perception obtained with natural vowel-consonant-vowel (VCV) syllables in SSN based on the spectro-temporal excitation pattern (Moore, 2003). Speech-dominated spectro-temporal "glimpses" in the speech and noise mixture were fed to a Hidden-Markov Model (HMM) based missing-data speech recognizer trained on talker-specific speech samples. While the model accounted reasonably well for the consonant-specific recognition scores, the predicted consonant confusions differed strongly from those observed in the measured data.

Holube and Kollmeier (1996) used an auditory model (Dau et al., 1996) in combination with a template-matching back end to predict the recognition of CVCs in SSN in NH and HI listeners. The auditory model by Dau et al. (1996) consists of a linear auditory filterbank, an envelope extraction stage, a nonlinear adaptation stage, and a low-pass filter, such that the internal representation (IR) is a function of time and frequency. In order to compensate for the temporal differences in the CVC test signals and the CVC templates, Holube and Kollmeier

(1996) applied a dynamic time warping (DTW) algorithm (Sakoe and Chiba, 1978) as a back end. The DTW algorithm temporally warps (i.e., locally stretches and compresses) two signals such that they ideally align in time according to some distance measure. The templates were mixed with noise at the same SNR as the test signal and the decision was based on the minimum distance between the test signal and the templates after DTW. Assuming a-priori knowledge, the speech signal contained in the correct template was identical to the test speech token such that the distance between the two signals resulted only from the differences in the noise waveforms. The model by Holube and Kollmeier (1996), fitted to account for psychoacoustic data of the individual NH and HI listeners using the original "optimal detector" back end from Dau et al. (1996), was shown to predict CVC-in-noise recognition data of the individual listeners (averaged across all considered speech tokens) with good accuracy while confusions were not considered.

Focusing on consonant- and vowel-specific recognition and confusion data (measured in NH listeners using CVC and VCV syllables in SSN), Jürgens and Brand (2009) applied a modeling approach largely comparable to that of Holube and Kollmeier (1996). The difference in the model front end was mainly the use of a modulation filterbank (Dau et al., 1997) instead of an envelope low pass filter, which is supported by several studies arguing that temporal modulations play a crucial role in consonant perception (e.g., Christiansen et al., 2007; Gallun and Souza, 2008). In the back end, Jürgens and Brand (2009) considered different distance measures for the DTW and investigated model configurations with and without a-priori knowledge. Their study concluded that (i) a-priori knowledge was necessary to obtain realistic consonant recognition performance, (ii) the Lorentzian distance measure yielded the best predictions when a-priori knowledge was used, (iii) consonant- and vowel-specific recognition scores were generally well predicted (although the model tended to overestimate the recognition performance for many consonants at large SNRs), and (iv) the confusion predictions were inaccurate.

Thus, while the above microscopic speech perception models yielded reasonable predictions in terms of consonant-specific recognition scores, consonant confusions have not yet been predicted successfully. Moreover, it has been demonstrated that consonant perception depends on individual speech tokens and, to some extent, even on the specific choice of the masking noise waveforms (Zaar and Dau, 2015). The discussed models have been either evaluated with respect to the grand average recognition performance across phonemes or on phoneme-specific data that still represent averages across many speech tokens of the same type. In contrast, modeling consonant perception on a token-by-token basis has not been considered yet.

The present study considers another microscopic speech perception model that was evaluated on the basis of the extensive data set provided by Zaar and Dau (2015), obtained with 15 CVs (each represented by six speech tokens) in conditions of white masking noise at six SNRs. A similar auditory model front end as the one employed by Jürgens and Brand (2009) was used and a template-matching process was applied in the back end. In contrast to Jürgens and Brand (2009), the IR of the noise alone was subtracted from the IRs of the test signals and the templates prior to template matching (as in the models by Dau et al., 1996; 1997). Furthermore, while a DTW algorithm was applied to temporally align test signals and templates, a maximum-correlation based approach was chosen in the decision stage (cf. Dau et al., 1996; 1997), as opposed to the minimum-distance based approach by Jürgens and Brand (2009). As proposed by Dau et al. (1996; 1997), a constant-variance internal noise was added in the decision stage. Finally, the speech and noise materials used in the present study (CVs in white noise) largely differed from the material used in Jürgens and Brand (2009), where CVCs and VCVs in SSN were considered. Average consonant recognition scores, consonant-specific recognition and confusion scores, as well as speech-token specific consonant recognition and confusion scores were considered to evaluate the model. Additionally, the response behavior of the listeners and the model was investigated by means of an entropy-based analysis.

3.2 Model framework and experimental conditions

3.2.1 Front-end processing

As in Jürgens and Brand (2009), the auditory preprocessing stages from Dau et al. (1997) were used. The model is shown in Fig. 3.1 (“auditory model”). The first stage of the model simulates the frequency selectivity of the human auditory system by means of a linear filterbank, consisting of 15 fourth-order gammatone filters with center frequencies logarithmically spaced between 315 Hz and 8 kHz. The outputs of the gammatone filters were shifted in time to time-align the peak delay of the individual gammatone filters. The second stage represents a rough approximation of the transformation of the basilar membrane vibrations into inner hair cell potentials and is realized as an envelope extraction mechanism. Each gammatone filter output signal is half-wave rectified and then filtered using a low pass filter with at a cut-off frequency of 1 kHz. The third stage consists of a chain of five adaptation loops that were designed to mimic adaptive properties of the auditory periphery and to account for perceptual forward masking in human listeners (Kohlrausch and Püschel, 1988; Kohlrausch et al., 1992; Dau et al., 1996). For stationary signals, the adaptation loops provide an approximately logarithmic compression, whereas faster fluctuations are transformed more linearly. Therefore, the adaptation loops effectively perform an onset enhancement of the individual subband envelope representations. The time constants chosen for the five adaptation loops were $\tau_1 = 5$ ms, $\tau_2 =$ 20 ms, $\tau_3 = 129$ ms, $\tau_4 = 253$ ms, and $\tau_5 = 500$ ms (taken from Dau et al., 1996). The fourth stage of the model is a low-frequency modulation filterbank consisting of a third-order low pass filter with a cut-off frequency of 2 Hz in parallel with three second-order band pass filters with a constant Q of 1 and center frequencies of 4, 8, and 16 Hz, respectively. After being fed through the adaptation loops, each subband envelope is thus further decomposed into four modulation bands. The output of the model front end obtained for any given input signal $x(t)$ is denoted as $R_x(t, f_g, f_m)$, where t denotes the temporal samples, f_g represents the gammatone filter center frequency, and f_m refers to the modulation frequency. CV speech tokens mixed with white noise were considered in this study (see Sec. 3.2.3). As in the original auditory model (Dau et al., 1997), and in contrast to Holube and Kollmeier (1996) and Jürgens and Brand (2009), the noisy speech token $(s+n)$ and the noise alone (n) were separately passed through the model front end, yielding the respective temporal patterns R_{s+n} and R_n. As an input to the back end, the difference between these temporal patterns was obtained as the model's signal representation: $R_s = R_{s+n} - R_n$.

3.2.2 Speech recognition back end

The model predictions were obtained using a template-matching approach. An overview of the modeling approach is depicted in Fig. 3.1. In order to compare a given test signal (stimulus) with a given template, the corresponding signal representations $R_{test}(t, f_g, f_m)$ and $R_{temp}(t, f_g, f_m)$ were time aligned using a DTW algorithm as proposed by Sakoe and Chiba (1978). The DTW algorithm locally compresses and expands the time axes of two signal representations such that the temporal alignment is ideal according to the chosen distance measure. In the present study, the Euclidean distance[2] measure was used and defined as

$$D(t_i,t_j)=\sqrt{\sum_{f_g}\sum_{f_m}\left[R_{test}(t_i,f_g,f_m)-R_{temp}(t_j,f_g,f_m)\right]^2}\ , \tag{3.1}$$

where t_i and t_j denote arbitrary temporal samples. Traditionally, the chosen distance measure has also been used as a decision metric, i.e., the template showing the smallest distance to the test signal was chosen as the model response (e.g., Holube and Kollmeier, 1996; Jürgens and Brand, 2009). In the present study, however, the DTW algorithm was solely applied to obtain time aligned versions[3] of the test-signal and template representations, $\widehat{R}_{test}$ and $\widehat{R}_{temp}$, respectively. Inspired by the original auditory model (Dau et al., 1996; Dau et al., 1997), the correlation coefficient between these time-aligned representations was then calculated as the model's decision metric as:

$$C(\widehat{R}_{test},\widehat{R}_{temp})=\frac{\sum_{t,f_g,f_m}\left[\widehat{R}_{test}(t,f_g,f_m)-\overline{\widehat{R}_{test}}\right]\cdot\left[\widehat{R}_{temp}(t,f_g,f_m)-\overline{\widehat{R}_{temp}}\right]}{N_{t,g,m}\cdot\sigma_{test}\cdot\sigma_{temp}}, \tag{3.2}$$

where $\overline{\widehat{R}_{test}}$ and $\overline{\widehat{R}_{temp}}$ represent the mean values and σ_{test} and σ_{temp} the standard deviations of $\widehat{R}_{test}$ and $\widehat{R}_{temp}$, respectively, and $N_{t,g,m}$ denotes the number of elements (number of samples × number of gammatone filters × number of modulation filters). A constant-variance Gaussian noise was added to the correlation coefficients, reflecting the listeners' uncertainty (internal noise). The variance of the noise was kept the same across experimental conditions. Eventually, the consonant corresponding to the template that yielded the largest correlation with the test signal was chosen as the model response (see Sec. 3.2.4).

[2] The Euclidean distance was used for DTW as it yielded far more plausible time alignment results as compared to the Lorentzian distance suggested by Jürgens and Brand (2009).

[3] The DTW algorithm from Sakoe and Chiba (1978) was applied without any path limitations, such that any local time-axis warping was in principle allowed.

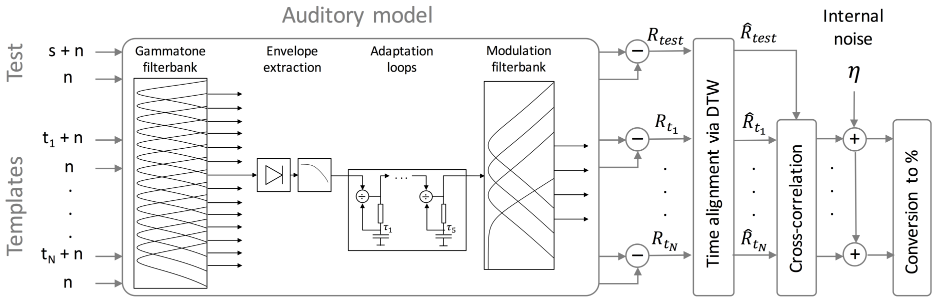


Figure 3.1: Scheme of the proposed consonant perception model. For the test signal and a set of templates, the noisy speech and the noise alone were passed separately through the auditory model, consisting of a gammatone filterbank, an envelope extraction stage, a chain of adaptation loops, and a modulation filterbank. The difference between the temporal patterns of the noisy speech and the noise alone was obtained. The resulting representations of the test signal and the templates were time-aligned using a dynamic time warping (DTW) algorithm. Finally, the cross-correlation coefficients between the test signal and each template were calculated and, after addition of a constant-variance internal noise, converted to percent.

3.2.3 Simulated conditions

The model was evaluated using the experimental conditions described in Zaar and Dau (2015; experiment 1). 15 CVs consisting of the 15 consonants /b, d, f, g, h, j, k, l, m, n, p, s, ʃ, t, v/ followed by the vowel /i/ were used whereby six recordings of each CV were taken from a Danish nonsense syllable speech material (Christiansen and Henrichsen, 2011). For each CV, three of these speech tokens were spoken by one particular male talker, the other three speech tokens were spoken by one particular female talker, amounting to a total of 90 speech tokens (15 CVs × 3 speech tokens × 2 talkers).

The speech tokens were equalized based on the peak level of an analog VU-meter simulation that responds sluggishly to the input signal (VUSOFT; Lobdell and Allen, 2007), such that they exhibited similar vowel levels while the consonant levels differed (cf. Zaar and Dau, 2015). White Gaussian noise was mixed with the speech tokens at different SNRs. SNR conditions of 12, 6, 0, -6, -12, and -15 dB were created by fixing the noise at a sound pressure level of 60 dB and adjusting the level of the speech tokens (based on the overal root-mean-square level of all speech tokens) according to the desired SNR. Each speech token was paired with one particular noise token in a given SNR condition. The noise tokens had a duration of one second and were faded in and out using raised cosine ramps with a duration of 50 ms. The speech tokens were mixed with the respective noise tokens such that the speech token onset was temporally positioned 400 ms after the noise onset. Eight NH native Danish listeners were presented three times with each speech token at each SNR and asked to vote for the consonant they heard. Thus, 24 responses (8 listeners × 3 repetitions) were collected per speech token and SNR, while 144 responses (8 listeners × 3 repetitions × 3 speech tokens × 2 talkers) were obtained per CV and SNR. The occurrences of responses were divided by the number of stimulus presentations to obtain the proportions of responses. The above described stimuli and the corresponding consonant perception data of Zaar and Dau (2015) were used throughout this study as inputs to the model and as reference data, respectively.

3.2.4 Simulation procedure

The same experimental stimuli the listeners had been presented with were fed to the model. While Jürgens and Brand (2009) added threshold-equalizing noise to the signals, audibility thresholds were not explicitly considered in the present study since the fixed-level masking noise was above the NH listeners' thresholds in the considered frequency range. Each test signal (i.e., each experimental stimulus) was compared to a talker-specific template set. The speech token contained in the correct template was identical to the speech token contained in the test signal (assumption of a-priori information); the other 14 consonants were each represented by the three available talker-specific speech tokens, such that, overall, 43 speech tokens were used as templates $(1{\times}1 + 14{\times}3)$. The masking noise waveforms in the test signals were the same as in the experiment. The templates were mixed with randomly generated white noise at the test-signal SNR in analogy to the stimulus generation described in Sec. 3.2.3. Five different templates were obtained from each considered speech token by mixing the speech token with five randomly generated noise waveforms. Thus, for a given test signal, the correct response alternative was represented by 5 templates (1 speech token × 5 noise tokens), whereas the other response alternatives were each represented by 15 templates (3 speech tokens × 5 noise tokens), amounting to 215 templates overall.

All test signals and templates and the corresponding noise signals were fed through the model front end, as described in Sec. 3.2.1, to obtain the respective signal representations R_{test} and R_{temp}. The signal representations were cut such that the noise-only parts at the beginning and the end were omitted and only the speech-containing portions[4] of the test signals and templates were further processed. For computational efficiency, the temporal resolution was reduced from a sampling rate of 44.1 kHz to 100 Hz by buffering R_{test} and R_{temp} into 10-ms time frames and taking the mean value across all samples within each frame. Time aligned versions of the signal representations – $\widehat{R}_{test}$ and $\widehat{R}_{temp}$ – were obtained for each combination of test signals and templates using DTW and the correlation coefficients between them were calculated (as described in Sec. 3.2.2). As a result, correlation coefficients between each test signal and each of the respective 215 templates were obtained. Internal Gaussian noise was added to the correlation coefficients with a constant variance of $\sigma^2_{int} = 0.05$. The variance of the internal noise was chosen such that it yielded the best possible agreement of the predicted and measured grand average consonant recognition scores, i.e., the noise globally calibrated the model but did not change across SNRs, stimuli, or templates.

To convert the noisy correlation coefficients obtained for a specific test signal to proportions of responses, multiple subsets of templates were drawn from the available 215 templates. Model responses were obtained based on each template subset and finally averaged across the considered subsets. Each subset consisted of 15 templates, each representing a different response alternative (i.e., one consonant). To ensure an unbiased comparison, all feasible combinations of templates were considered as subsets. As the 14 incorrect response alternatives were each represented by 15 different templates and the correct response alternative was represented by 5 different templates (see above), the number of combinations (i.e., the number of template subsets) was $15^{14} \cdot 5$. For each subset, the template that showed the largest correlation with the test signal was selected as the model response. The occurrences of model responses were then divided by the number of considered template subsets to obtain the modeled proportions of responses. The procedure described above was iterated 100 times with randomly generated internal noise in each iteration and the results obtained in the individual iterations were finally averaged.

[4] The start and end times of the speech-containing portions were defined as the first and last sample of the corresponding clean speech token's power (in dB) that were less than 40 dB below the speech token's power maximum.

3.3 Results and analysis

3.3.1 Consonant recognition

Figure 3.2 depicts the grand average consonant recognition scores, i.e., the average recognition scores across all considered speech tokens, as a function of SNR. The open circles represent the average consonant recognition scores measured in NH listeners (Zaar and Dau, 2015). The filled black circles show the model predictions from the present study, obtained with the calibrated model (with internal noise variance σ^2_{int} = 0.05), while the small gray circles and dashed gray lines represent model predictions obtained with a range of internal noise variances ranging from $\sigma^2_{int} = 0$, i.e., no internal noise, to $\sigma^2_{int} = 0.5$. It can be observed that the predictions obtained with the calibrated model at this global level were very close to the perceptual data. This was the case for both the SRTs (data: -3 dB / predictions: -3.4 dB) and the slopes of the recognition curves. Thus, the correlation between the two curves was at ceiling (Pearson's $r = 0.998$) and the root-mean-squared error (RMSE) between them was small (RMSE = 1.68%).

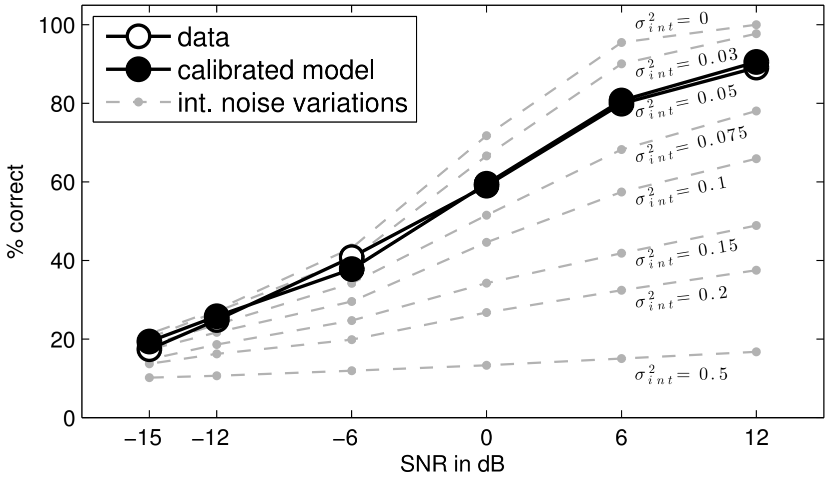


Figure 3.2: Grand average consonant recognition scores in percent as a function of SNR.The open black circles represent the perceptual data and the filled black circles show the model predictions obtained with the calibrated model (internal noise variance $\sigma^2_{int} = 0.05$). The small gray circles and dashed gray lines represent model predictions obtained with a range of internal-noise variances σ^2_{int}, which are indicated next to the corresponding curves.

Regarding the role of the internal noise, the upper dashed gray lines ($\sigma^2_{int} = 0$ and $\sigma^2_{int} = 0.03$) reveal that the model overestimated consonant recognition at

SNRs of 0, 6, and 12 dB when no or not enough internal noise was considered, resulting in overly steep slopes. In contrast, internal noise variances $\sigma^2_{int} > 0.05$ led to an underestimation of consonant recognition and thus to too shallow slopes. For the following figures and analyses only the calibrated model was considered.

Figure 3.3 shows the consonant-specific recognition scores, i.e., the consonant recognition scores averaged across speech tokens of the same phonetic identity (e.g., /bi/). The consonants are indicated in the upper left corners of the respective figure panels. Comparing the measured recognition scores (open circles) across panels, it can be observed that the individual consonants exhibited drastic differences with respect to their perceptual robustness to the influence of the masking noise. For instance, the consonant /t/ (bottom middle panel in Fig. 3.3) was, on average, almost perfectly recognized by listeners down to an SNR of -6 dB and still recognized about 50% of the times at -15 dB SNR. This noise robustness can also be observed for /s/ (right panel in fourth row) and /ʃ/ (bottom left panel). In contrast, some of the consonants were perceptually much more vulnerable. For example, /v/ (bottom right panel in Fig. 3.3) shows a recognition score of only about 80% at the large SNRs of 12 and 6 dB, followed by a sudden drop to around 30% at 0 dB SNR, from where the recognition scores approached chance-level (6.7%) performance towards lower SNRs. Equally low recognition scores can also be observed for /b/, /f/, /h/, /l/, /m/, and /p/.

The recognition scores predicted by the model are indicated as filled circles in Fig. 3.3. Overall, the model predictions of the consonant-specific recognition scores fit the perceptual data very well. In particular, the noise robustness of /s/, /ʃ/, and /t/ was well reflected in the predictions, as indicated by the overlap of the corresponding measured and simulated recognition curves. Furthermore, the predicted recognition curves for most of the other consonants provided an almost exact match with the measured ones (e.g., /f, g, h, k, n, v/). In the case of /b/, /l/, /m/, and /p/, the model performed slightly better than the listeners, particularly for large SNRs. For /d/ and /j/, however, the model slightly underestimated the listeners' performance. The predicted recognition scores in these cases showed an offset across all SNRs while the predicted recognition curves were qualitatively quite similar to the measured ones.

To quantify the agreement between predictions and measurements, Pearson's r was calculated at each SNR condition between the measured and the

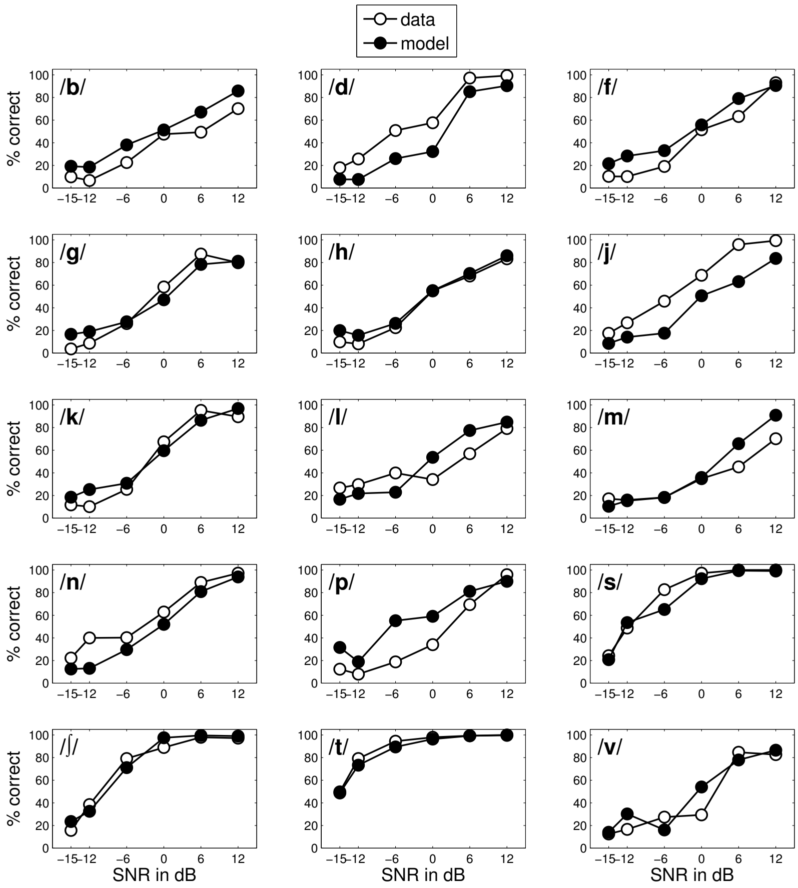


Figure 3.3: Consonant-specific recognition scores in percent as a function of SNR (averaged across speech tokens of the same type). The open circles represent the perceptual data and the filled circles show the corresponding model predictions. The consonants are indicated in the upper left corners of the panels.

predicted recognition scores (i) across the consonant-specific recognition scores (averaged across different speech tokens of the same type) and (ii) across the speech-token specific recognition scores. Table 3.1 summarizes the results. It can be seen that the measured and predicted recognition scores were significantly ($p < 0.05$) correlated across consonants; for SNRs of 6, 0, -6 and -12 dB the correlations were highly significant ($p < 0.01$). Correspondingly, the correlations were large particularly at medium SNRs (maximum: $r = 0.76$ at 0 dB SNR / minimum: $r = 0.55$ at 12 dB SNR). Furthermore, Table 3.1 shows that the measured and predicted recognition scores were highly significantly ($p < 0.01$) correlated even for individual speech tokens. Again, the largest cor-

Table 3.1: Correlation between perceptual and predicted consonant recognition scores in terms of Pearson's correlation coefficients r and the corresponding p-values. p-values indicating significant correlation ($p < 0.05$) are given in bold font. For each SNR condition, the correlation analysis was performed across consonants (left) and across individual speech tokens (right).

	Across consonants		Across speech tokens	
SNR	r	p	r	p
12 dB	0.55	**0.017**	0.35	**0.000**
6 dB	0.65	**0.004**	0.39	**0.000**
0 dB	0.76	**0.001**	0.43	**0.000**
-6 dB	0.75	**0.001**	0.57	**0.000**
-12 dB	0.75	**0.001**	0.56	**0.000**
-15 dB	0.57	**0.013**	0.31	**0.001**

relation was observed at medium SNRs (maximum: $r = 0.57$ at -6 dB SNR / minimum: $r = 0.31$ at -15 dB SNR). As expected, the correlation coefficients across the speech-token specific recognition scores were generally lower than the correlation coefficients across the consonant-specific recognition scores. However, the p-values for the speech-token specific correlations were also lower, indicating higher significance than in the consonant-specific case. This was due to the difference in the number of data points considered for the individual correlations (15 for the consonant-specific case vs. 90 for the speech-token specific case).

3.3.2 Consonant confusions

Figure 3.4 provides an overview of the entire measured and predicted data in terms of a confusion matrix (CM). The perceptual data and the model predictions were averaged across speech tokens of the same identity and across the six considered SNRs to obtain the CM. The vertical axis indicates the presented consonants, while the horizontal axis represents the consonants provided as response alternatives. Therefore, the full response patterns obtained for the individual consonants (consisting of the average consonant recognition as well as consonant confusion scores) are reflected in the individual rows of the CM and the average recognition scores are represented by the diagonal elements of the CM. The perceptual data and the predictions are depicted as circles, the size of which indicates the underlying proportions of responses according to the six categories shown in the figure's legend.

A complete overlap of circles indicates a large agreement between the respective measured (filled gray circles) and predicted (open red circles) average response scores. Such complete overlap can be observed along the CM's diagonal, which reflects the average consonant recognition scores. This is another view of the good agreement of measured and predicted consonant-specific recognition scores demonstrated in Table 3.1 and Fig. 3.3. The off-diagonal CM elements represent the average consonant confusions. Certain groups of consonants that were likely to be confused with each other (*confusion groups*) can be observed in the perceptual data (filled gray circles). Most notably, three groups can easily be identified: $/\mathrm{m, n, j, l, v}/$, $/\mathrm{f, h, b, g, d, p, k}/$, and $/\mathrm{s}, \int, \mathrm{t}/$. Additionally, there was some overlap between the first and the second group. In general, the confusion predictions of the model (open red circles) captured the measured confusions (filled gray circles) quite well, as can be seen from the overlap of the off-diagonal circles. In particular, the vast majority of the measured confusions was reflected in the predictions (70 out of 81 measured confusions "hit" by the model according to the categories used in Fig. 3.4), i.e., the model's errors were, on average, very similar to the errors made by the listeners. This was also reflected in the clustering of the model predictions, which, to a large extent, followed the confusion group clustering discussed above for the perceptual data. However, the model tended to underestimate the extent of the confusions (i.e., there are many red circles that are smaller than their gray counterparts) and, instead, predicted additional confusions (e.g. $/\mathrm{m, n, j}/$ confused with $/\mathrm{f, h, b, g, d, p}/$) that were not reflected in the perceptual data (36 "false alarms" predicted by the model according to the categories used in Fig. 3.4).

Figure 3.5 shows three example confusion patterns (first introduced by Allen, 2005) for $/\mathrm{m}/$, $/\mathrm{s}/$, and $/\mathrm{k}/$, respectively, each reflecting the average responses obtained with six different speech tokens. While $/\mathrm{m}/$ and $/\mathrm{s}/$ represent two examples with highly correlated measured and predicted confusions (Pearson's r of 0.76 and 0.96, respectively; cf. Table 3.2), $/\mathrm{k}/$ showed the least correlation between the measured and the predicted confusions (Pearson's r of 0.43). In the top row, the perceptual data are depicted in terms of consonant recognition (black line) and consonant confusions (colored lines) as a function of SNR. In the bottom row, the corresponding model predictions are shown. It can be observed that the model predictions captured the types of confusions made by

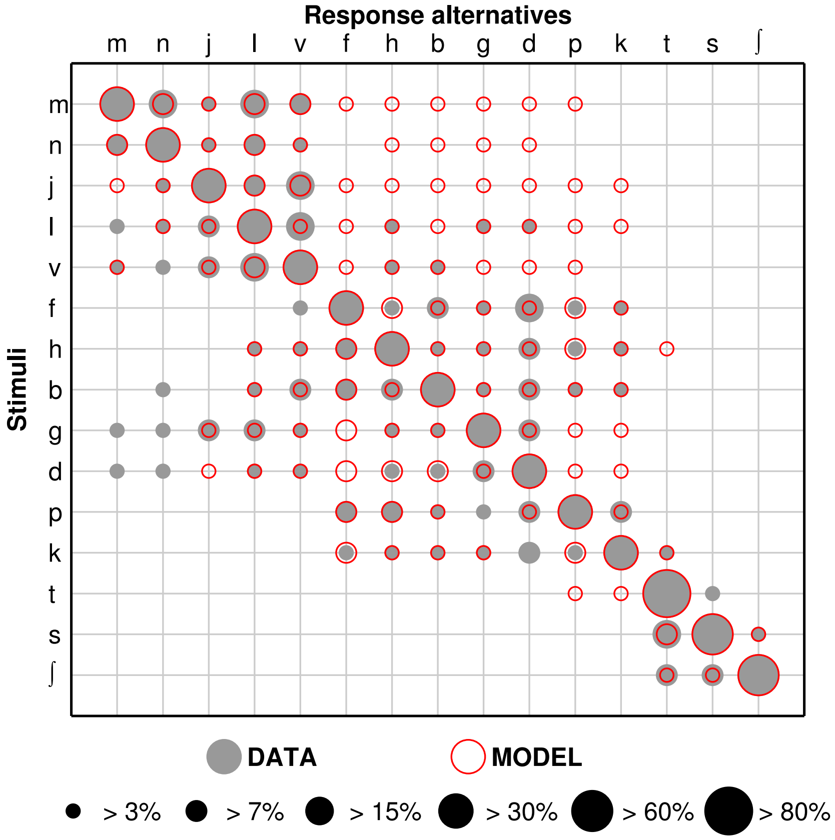


Figure 3.4: Data and predictions averaged across SNR and across speech tokens of the same type, depicted as a confusion matrix. The presented consonants are shown on the vertical axis and the response alternatives on the horizontal axis. The filled gray circles represent the perceptual data while the open red circles show the model predictions. The size of the circles indicates the proportions of responses according to the six categories provided in the legend.

the listeners to a large extent. /m/ was confused with /n, l, v, j/ both by the listeners and the model (left panel). /s/ (middle panel) was confused with /t, ʃ, f/ by the listeners and the model, while the fourth confusion at the lowest SNR of -15 dB differed (listeners: /d/; model: /v/). In the case of /k/, it can be seen that there still was some agreement, as the model and the listeners showed confusions with /h/ and /p/. However, the other measured confusions (/d, g/) were not reflected in the model predictions, which instead showed confusions with /f, b/. Nevertheless, the overall agreement between measured and predicted confusions was large (mean Pearson's *r* across consonants: 0.66;

cf. Table 3.2).

As already seen in the CM (Fig. 3.4), the perceptual confusions were more pronounced than the predicted ones, i.e., the listeners were more consistent in their errors than the model. This is reflected in the generally lower confusion scores obtained in the model predictions as compared to the perceptual data. For instance, in the case of /m/ (top left panel), the listeners showed a very pronounced confusion with /n/, which reached up to 44% at 6 dB SNR. In the model predictions (bottom left panel), however, the maximum confusion with /n/ reached only 17% (at 0 dB SNR). Similar underestimations of the confusions can be observed for the consonant /s/ (middle panel), as well as for many other consonants that exhibited large perceptual confusions (not shown here).

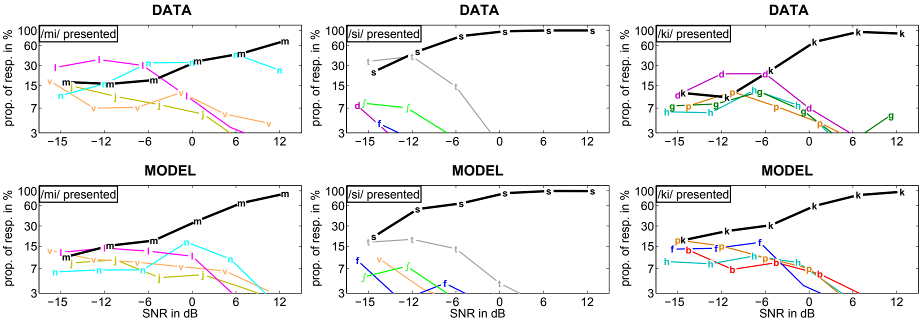


Figure 3.5: Measured (top) and predicted (bottom) confusion patterns obtained for /m/ (left), /s/ (middle), and /k/ (right). The data were averaged across different speech tokens of the same type. The correct responses are indicated as thick black lines and the confusions are shown as thinner lines in different colors; the data points are labeled with the corresponding consonants. Maximally five responses are depicted for clarity, which were chosen based on their extent. A slight horizontal shift was introduced to the data for better readability. The ordinate is scaled logarithmically to emphasize the confusions.

To evaluate the significance of the observed agreement between the confusions in the perceptual data and in the model predictions, Pearson's r was calculated between the measured and predicted across-SNR average response patterns using (i) consonant-specific data (i.e., data averaged across different speech tokens of the same type) and (ii) speech-token specific data. Only the erroneous responses obtained for each CV/speech token (i.e., only the off-diagonal elements of the CM) were correlated; the recognition scores (on-diagonal elements of the CM), which would otherwise strongly dominate the correlations, were excluded in order to evaluate the qualitative agreement of the measured and predicted confusions irrespective of the recognition score agreement. This *confusion correlation* was only taken into account if the cumulative error P_e (i.e., the sum of all perceptual confusions averaged across SNR) exceeded 20%.

The left part of Table 3.2 summarizes the results obtained with the consonant-specific data in terms of a correlation coefficient r and a corresponding p-value for each stimulus consonant. The analysis revealed that the predicted confusions were strongly correlated with the measured confusions when considered at the consonant level (maximum: $r = 0.96$ for /s/; minimum: $r = 0.43$ for /k/; average: $r_{avg} = 0.66$). Almost all (12 out of 15) consonant-specific confusion correlations were significant ($p < 0.05$, in bold font), except for /l/ and /k/, which exhibited p-values just above 0.05. For /t/, no correlation was obtained as the error was too small ($P_e \leq 20\%$).

For the speech-token specific case, correlation coefficients and p-values were obtained for each of the 90 speech tokens. For the sake of compactness, the right side of Table 3.2 shows a collapsed version of the results obtained with the speech-token specific data in terms of the average correlation coefficients $\overline{r}$ and the average p-values $\overline{p}$ for each stimulus consonant (i.e., averaged across speech tokens of the same type). Additionally, the number of significantly correlated confusion patterns ($p < 0.05$), $\mathrm{N_s}$, and the number of considered speech tokens (with $P_e > 20\%$), $\mathrm{N_c}$, are provided in the rightmost column of Table 3.2. The speech-token specific confusion correlation analysis revealed that the confusion correlations were significant only for 43 of the 83 eligible speech tokens (7 of the 90 speech tokens showed $P_e \leq 20\%$ and were thus not considered). The maximum average confusion correlation at the speech-token level was $\overline{r} = 0.89$ for /ʃ/. All other correlations were much smaller, with a minimum at $\overline{r} = -0.02$ for /t/. The average confusion correlation coefficient across all considered 83 speech tokens was $\overline{r}_{avg} = 0.47$.

In addition to the confusion correlation analysis of the across-SNR average data described above, the consonant-specific and speech-token specific confusion correlations were also evaluated for the individual SNR conditions. The left side of Table 3.3 shows the average correlation coefficients and p-values obtained based on the consonant-specific data. The number $\mathrm{N_s}$ of consonants exhibiting significant confusion correlation ($p < 0.05$) and the number $\mathrm{N_c}$ of considered consonants (with $P_e > 20\%$) are given in parentheses. It can be observed that the model captured most of the measured confusions well at the consonant- and SNR-specific level. Average confusion correlations ranged between 0.44 and 0.66 and the highest correlation values were obtained for SNRs

Table 3.2: Correlation between perceptual and predicted consonant confusion scores as a function of the presented consonant (only obtained if the overall error $P_e > 20\%$). The Pearson's correlation coefficients r and the corresponding p-values were obtained across the response alternatives (excluding the recognition scores) based on the consonant-specific and on the speech-token specific across-SNR average data, respectively. The speech-token specific correlation results were then averaged across the different speech tokens of the same type (averages $\overline{r}$ and $\overline{p}$). p-values indicating significant confusion correlation ($p < 0.05$) are given in bold font. The rightmost column additionally contains the number $\mathrm{N_s}$ of tokens showing significant confusion correlation ($p < 0.05$) and the number $\mathrm{N_c}$ of considered tokens (with error $P_e > 20\%$). The consonants are ordered as in Fig. 3.4.

	Consonant-specific data		Speech-token specific data	
Consonant	r	p	$\overline{r}$	$\overline{p}$ ($\mathrm{N_s}/\mathrm{N_c}$)
/m/	0.76	**0.001**	0.56	**0.045** (4/6)
/n/	0.76	**0.001**	0.58	**0.042** (5/6)
/j/	0.68	**0.004**	0.51	0.095 (4/6)
/l/	0.45	0.053	0.24	0.332 (2/6)
/v/	0.60	**0.012**	0.37	0.222 (2/6)
/f/	0.60	**0.011**	0.42	0.098 (3/6)
/h/	0.69	**0.003**	0.45	0.117 (2/6)
/b/	0.65	**0.006**	0.52	0.055 (3/6)
/g/	0.60	**0.012**	0.55	**0.048** (4/6)
/d/	0.49	**0.038**	0.31	0.217 (2/6)
/p/	0.82	**0.000**	0.55	**0.047** (4/6)
/k/	0.43	0.060	0.33	0.176 (2/6)
/t/	N/A	N/A	-0.02	0.520 (0/2)
/s/	0.96	**0.000**	0.89	**0.000** (4/4)
/ʃ/	0.80	**0.000**	0.54	0.058 (3/5)

of 0 and 6 dB. The model showed significant confusion correlations for almost all (19 out of 22) considered consonants at SNRs ≥ 0 dB and for more than half (24 out of 43) of the considered consonants at negative SNRs. When considering the speech-token specific data per SNR (right side of Table 3.3), the average confusion correlations were substantially lower, ranging from 0.24 to 0.44. The largest average correlations were again found for SNRs ≥ 0 dB, with significant confusion correlations obtained for about half (57 out of 113) of the considered speech tokens. For negative SNRs, the confusions were significantly correlated for only 30% (75 out of 250) of the considered speech tokens. This substantial decrease of the model performance at the level of individual speech tokens

Table 3.3: Correlation between perceptual and predicted consonant confusion scores as a function of SNR (only obtained if the overall error $P_e > 20\%$). For each SNR condition, the Pearson's correlation coefficients r and the corresponding p-values were obtained across the response alternatives (excluding the recognition scores) based on the consonant-specific and on the speech-token specific data, respectively. The SNR-specific results were then averaged across the different consonants and the different speech tokens, respectively (averages $\overline{r}$ and $\overline{p}$). p-values indicating significant confusion correlation ($p < 0.05$) are given in bold font. The p-values are accompanied by the number $\mathrm{N_s}$ of consonants/tokens showing significant confusion correlation ($p < 0.05$) and the number $\mathrm{N_c}$ of considered consonants/tokens with error $P_e > 20\%$ (maximally 15 consonants/90 speech tokens).

	Across consonants		Across speech tokens	
SNR	$\overline{r}$	$\overline{p}$ ($\mathrm{N_s/N_c}$)	$\overline{r}$	$\overline{p}$ ($\mathrm{N_s/N_c}$)
12 dB	0.47	0.097 (2/4)	0.44	0.156 (11/20)
6 dB	0.66	**0.014** (6/6)	0.37	0.226 (14/33)
0 dB	0.66	**0.019** (11/12)	0.43	0.165 (32/60)
-6 dB	0.44	0.118 (7/13)	0.27	0.258 (21/77)
-12 dB	0.45	0.157 (9/15)	0.26	0.283 (27/85)
-15 dB	0.49	0.104 (8/15)	0.24	0.294 (27/88)

and SNRs was probably caused by the extremely low number of observations[5] considered in this case, which resulted in noisy reference data.

3.3.3 Entropy-based analysis

The above analysis demonstrated that while the model mostly accounted for the types of measured confusions, it showed a tendency to underestimate the amount of these confusions and, instead, additionally selected other confusions that were not reflected in the perceptual data. This suggests that the model responded more randomly than the listeners. To analyze the overall response behavior of the listeners and the model in terms of the randomness of the responses, the entropy of responses was calculated (cf. Miller and Nicely, 1955; Phatak et al., 2008; Zaar and Dau, 2015). In particular, the *normalized entropy* for a given response vector $\mathbf{p} = [\mathrm{p}_1, \mathrm{p}_2, ..., \mathrm{p}_\mathrm{R}]$, with $\mathrm{p}_1, ..., \mathrm{p}_\mathrm{R}$ denoting the proportions of responses for the individual response alternatives, was defined

[5] 24 observations were available per speech token and SNR condition; in case of the error just exceeding the 20% threshold, the considered confusion patterns therefore consisted of only 5 observations.

as:

$$H_{norm}(\mathbf{p}) = \frac{100\%}{\log_2(\mathrm{R})} \cdot \sum_{i=1}^{R} \mathrm{p}_i \log_2\left(\frac{1}{\mathrm{p}_i}\right), \qquad \forall\ \mathrm{p}_i > 0 \tag{3.3}$$

with $\log_2(\mathrm{R})$ representing the theoretical entropy maximum. The normalized entropy is therefore confined to the interval [0%, 100%]. When the randomness in the response vector is minimal, i.e., one element has a value of 1 and the other elements are 0, the normalized entropy is 0%. When the randomness in the response vector is maximal, i.e., all elements have the same value of 1/R, the normalized entropy is 100%. The normalized entropy was calculated per SNR condition (i) for each response vector in the consonant-specific perceptual data and predictions and (ii) for each response vector in the speech-token specific perceptual data and predictions and, finally, averaged across consonants and speech tokens, respectively.

Figure 3.6 shows the normalized entropy obtained from the perceptual data (white bars) and from the model predictions (black bars) as a function of SNR for the consonant-specific case (left panel) and for the speech-token specific case (right panel). The entropy generally increased with decreasing SNR as the task became more challenging and the consonant percept became more uncertain due to the increased masking effect of the noise, such that more errors and less systematic errors occurred. Furthermore, the entropy in the consonant-specific perceptual data (left panel, white bars) was around 10% larger than the entropy in the speech-token specific perceptual data (right panel, white bars), except at the largest SNR of 12 dB (5% difference). This indicates that averaging across speech tokens of the same type increases the randomness in the responses, implying perceptual differences across the considered speech tokens. This effect has already been shown for the considered data set on a listener-by-listener basis (Zaar and Dau, 2015) and is here confirmed for the across-listener average data, highlighting the importance of considering the data (and predictions) at the speech-token level.

Regarding the comparison between the perceptual data and the model predictions, the entropy analysis revealed that the model predictions showed a larger entropy than the perceptual data. This was the case both for the entropy analysis at the consonant level (left panel of Fig. 3.6, black bars vs. white bars) and at the speech-token level (right panel, black bars vs. white bars), with differences of up to 13% in both cases. Thus, the entropy-based analysis showed that the model's response behavior was indeed more random than that of the listener panel.

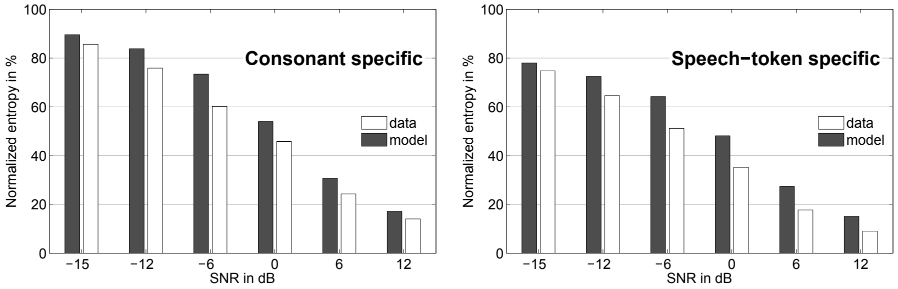


Figure 3.6: Normalized entropy in percent as a function of SNR calculated from the perceptual data (white bars) and the model predictions (black bars). Left: normalized entropy obtained from consonant-specific data and predictions; right: normalized entropy obtained from speech-token specific data and predictions. The normalized entropy was calculated for each consonant/speech token and SNR and then averaged across consonants/speech tokens.

3.4 Discussion

3.4.1 Relation to other studies

The model proposed in the present study represents an extension of the auditory detection model by Dau et al. (1997) towards predicting microscopic speech perception data. The main references for comparison of the model performance are the related modeling work of Jürgens and Brand (2009), which partly inspired the present study, and the Glimpse-model approach by Cooke (2006). However, it should be noted that these models were evaluated on different stimuli and data, such that a direct comparison is difficult. In particular, Jürgens and Brand (2009) used VCVs in steady-state SSN and Cooke (2006) employed VCVs in N-talker babble modulated SSN, while the present study used CVs in steady-state white noise (cf. Zaar and Dau, 2015). In terms of the grand average consonant recognition as a function of SNR, the proposed model showed an almost perfect fit with the perceptual data, whereas the model by Jürgens and Brand (2009) showed an overly steep recognition curve in their study (see their Fig. 3); this is mainly attributable to the calibration of the proposed model using internal noise (as shown in Fig. 3.2, see also Sec. 3.4.2), which had not been performed by Jürgens and Brand (2009). Cooke (2006) only considered one SNR condition, such that no comparison is feasible here. Regarding consonant-specific recognition scores, Jürgens and Brand (2009) showed a good agreement between their perceptual data and the corresponding predictions at medium to low SNRs, whereas their model predicted perfect recognition irrespective of the considered consonant at large SNRs, which was not reflected in their perceptual data (see their Fig. 4). Cooke (2006) obtained reasonable predictions of the consonant-specific trends in the recognition scores for the considered SNR of -6 dB (see his Fig. 10). The model presented in the current study, however, provided significantly correlated recognition scores across consonants at *all* considered SNR conditions, including large positive SNRs of 6 and 12 dB. Furthermore, the proposed model yielded highly significantly correlated recognition score predictions even at the speech-token level, which has so far not been reported in the related literature. Finally, while Jürgens and Brand (2009) and Cooke (2006) concluded that their respective models did not account well for consonant confusions, the present study demonstrated that the proposed model predicted the perceptual consonant confusions to a large extent (at the consonant-specific level).

3.4.2 Significance of the model components

During the development of the proposed model, many decisions were taken regarding the model design. This section lays out the reasons for including the individual model components and how they influence the predictions.

The auditory model used as a front end (Dau et al., 1997) was adapted for consonant perception modeling in a similar way as in Jürgens and Brand (2009). The low-frequency bands (between 50 and 300 Hz), typically considered in the gammatone filterbank, were omitted in order to mitigate the effect of differences in the low-frequency vowel portions of the stimuli and the templates, which may otherwise result in undesired effects that are independent of the consonant cues (e.g., prediction biases based on vowel-portion similarity[6]). The envelope extraction stage and the adaptation loops were parametrized as suggested by Dau et al. (1997). The onset enhancement performed by the adaptation loops provided realistic predictions as, e.g., the onset of the high-frequency frication

[6] Since the low-frequency energy of the CVs is large compared to the mid- and high-frequency bands but does not contribute substantially to consonant perception, slight differences in the vowel pronunciation can induce biases that are independent of consonant-cue similarity and thus detrimental to the predictive power of the model.

noise of an /s/ was enhanced such that it became more similar to the high-frequency burst of a /t/, which led to a perceptually plausible confusion at low SNRs (see Fig. 3.5, middle panels). Finally, four low-frequency modulation filters were applied, as also proposed by Jürgens and Brand (2009). It should be noted that simulations obtained using a simple low-pass filter with a cut-off frequency of 8 Hz (Dau et al., 1996) instead of a modulation filterbank led to comparably accurate results. However, the modulation-filterbank model is expected to generalize to a broader range of conditions as (i) the corresponding Dau et al. (1997) model accounts for more psychoacoustic conditions than the Dau et al. (1996) model and (ii) modulation-domain based macroscopic speech intelligibility models (e.g., Houtgast et al., 1980; Jørgensen et al., 2013) have been shown to account for a large variety of acoustic conditions.

While Jürgens and Brand (2009) directly fed the outputs of the model front end obtained with the noisy speech tokens to the back end, the present study followed the original model from Dau et al. (1997) in that the difference between the front-end outputs obtained with the noisy speech and the noise alone was considered in the back end. This assumption of a-priori knowledge about the masking noise was necessary to correctly predict the robustness of high-frequency cues (observed in the perceptual data for /s, ʃ, t/). In contrast, Jürgens and Brand (2009) could partly predict the robustness of high-frequency cues (/t, s, ts, ʃ/, see their Fig. 4) without this assumption. However, they used masking noise with a speech-shaped spectrum (sloping down towards high frequencies), such that the masking in the relevant high-frequency region was much less effective than in the present study, where white masking noise with a flat spectrum was employed. Thus, it can be concluded that if all the relevant consonant cues are masked to a comparable extent, the assumption of a-priori knowledge about the masking noise appears to be necessary for realistic predictions, at least when using the auditory model of Dau et al. (1997) as a front end. The need for such a mechanism in the model is consistent with the results from a study by Mesgarani et al. (2014), which showed that spectrograms reconstructed from neural representations of noisy phonemes measured in ferret primary auditory cortex were more similar to the clean phonemes than to the noisy ones. This implies the existence of a de-noising mechanism at higher stages of auditory processing, which the auditory model considered in the present study does not capture. Using a-priori knowledge about the noise may thus be considered as a simplistic way of simulating a de-noising mechanism.

The model's decision was based on the maximum cross-correlation (as in Dau et al., 1997; see also Gallun and Souza, 2008) of the time-aligned IRs of the test signal and the templates, as opposed to the minimum distance used by Jürgens and Brand (2009). The cross-correlation has the advantage that it is insensitive to level differences (i.e., solely describes *covariation*), which may be more closely related to the perceptual decision-making process than any distance measures (be it Euclidean or Lorentzian distance), which are typically sensitive to level differences. An earlier distance-based version of the model indeed yielded less convincing predictions of the perceptual data, partly due to biases that were presumably induced by this level sensitivity. A similarly biased behavior can be observed in the Jürgens and Brand (2009) predictions (see their Fig. 6, panel 2). The correlation-based back end alleviated this problem to a large extent and, thus, yielded realistic predictions in terms of consonant recognition and confusion scores.

Finally, the constant-variance internal noise in the model's decision stage (representing the listeners' uncertainty, cf. Dau et al., 1997) provided a realistic amount of uncertainty at medium to large SNRs, where the predicted recognition scores otherwise exceeded the measured ones, leading to overly steep recognition curves (see upper gray curve in Fig. 3.2). This result has also been reported by Jürgens and Brand (2009), who did not include an explicit calibration mechanism in their model. Although the internal noise affected the model predictions differently at different SNRs (cf. Fig. 3.2), the internal noise used in the present study merely calibrated the model as a whole, i.e., it did not change across SNRs, stimuli, or templates. The entropy-based analysis showed that the model responded slightly more randomly than the listeners did. It might seem intuitive to reduce the internal-noise variance in order to mitigate this mismatch; however, this is not feasible as it would considerably worsen the model's prediction accuracy with respect to the consonant recognition scores.

3.4.3 Limitations of the approach

Despite its large predictive power for the considered data/stimuli, the consonant perception model proposed in the present study needs to be tested for generalizability using other data sets that differ with respect to the speech tokens (e.g., VCVs instead of CVs), the native language of the talkers and listeners (e.g., English instead of Danish), and/or the noise type (e.g., SSN instead of white noise). Furthermore, as all stimuli were above NH audibility thresholds in the considered frequency bands, no audibility thresholds were considered in the model. Therefore, the model is bound to fail for partly or fully inaudible stimuli due to low presentation levels or hearing impairment. This could be overcome by adding threshold-simulating noise (cf. Jürgens and Brand, 2009; Jürgens et al., 2014) or by excluding the frequency bands below threshold from further processing (cf. Jørgensen and Dau, 2011). Moreover, it has been shown in Zaar and Dau (2015), based on the data set considered in the present study, that different NH listeners with the same language background can exhibit large perceptual differences for identical stimuli. The current study, however, focused on the across-listener average data, thus neglecting the across-listener perceptual variability. The proposed model has, in its current form, no means of explaining such listener-specific effects, which may be attributable to individual biases or supra-threshold processing deficits that were not captured by the audiometric test.

3.4.4 Perspectives

The most common acoustic condition that has been considered in consonant perception studies is additive stationary noise (e.g., Miller and Nicely, 1955; Wang and Bilger, 1973; Phatak and Allen, 2007; Phatak et al., 2008; Zaar and Dau, 2015). While this condition has provided valuable insights in the cues underlying consonant perception, it does not reflect realistic acoustic scenarios, in which most competing sound sources are strongly modulated and reverberation is typically present. An experimental investigation of consonant perception in such conditions and a subsequent evaluation of the proposed model's predictive power for the corresponding data may therefore be a crucial next step.

The present study focused on modeling consonant perception data obtained with NH listeners. However, consonant perception measurements may be particularly insightful when used as a tool to identify specific problems experienced by HI listeners. To better understand the cause of these problems, a version of the model that is conceptually capable of explaining effects of hearing impairment may be useful. To that end, sensitivity, compression, and frequency selectivity should be adjustable in the model front end. Furthermore, a model version that simulates the effects of hearing-aid signal processing in combination with the effects of certain types of hearing impairment may be a powerful tool for parametrizing hearing aid algorithms. A comparable model extension may be conceived for simulating the effects of cochlear-implant phoneme transduction and adjusting the corresponding algorithms (e.g., regarding channel selection).

The proposed model predicts consonant perception from an auditory modeling perspective, i.e., using a-priori information where necessary to predict the data. A "blind" model that bases its predictions only on the stimulus, just like listeners give their responses solely based on the stimulus, would represent a more elegant approach. Such a model requires a massive ASR back end that reaches human performance, which has so far not been feasible (Meyer et al., 2011). However, recent advances in ASR using HMMs in combination with Deep Neural Networks (DNNs, e.g. Hinton et al., 2012; Dahl et al., 2012) suggest that the gap between human and machine speech recognition is decreasing substantially. When blind ASR-based models become technically feasible, the present study may serve as a reference with respect to the front end features that should be considered to obtain realistic predictions. Furthermore, the reported predictive power of the assumption of a-priori knowledge about the masking noise motivates the use of suitable source separation algorithms prior to the speech recognition process.

3.5 Summary and conclusions

A consonant perception model was presented and evaluated with respect to consonant recognition and consonant confusions at different levels of detail. The model consists of an auditory modeling front end in combination with a correlation-based template-matching back end and represents an extension of the auditory processing model by Dau et al. (1997) towards predicting microscopic speech perception data. The model was evaluated based on the extensive CV-in-noise data from Zaar and Dau (2015), obtained with NH listeners. Overall, a good agreement between the perceptual data and the model predictions was demonstrated. The measured grand average consonant recognition scores as a function of SNR were almost perfectly accounted for by the model. Furthermore, the predicted consonant-specific recognition scores were highly correlated with the measured ones. Even at the speech-token level, large correlations between the predicted and the perceptual recognition scores were obtained. Regarding consonant confusions, the model predictions showed a strong similarity with the measured confusions at the consonant-specific level. However, the model tended to underestimate the extent of the main confusions in this scenario and showed only partially satisfactory confusion predictions at the speech-token level. It was shown in an additional entropy-based analysis that the model generally responded slightly more randomly than the listener panel did, which explains the observed shortcomings.

Overall, the large predictive power of the proposed model suggests that adaptive processes in the auditory preprocessing in combination with a cross-correlation based template-matching back end functionally account for some of the processes underlying consonant perception in normal-hearing listeners. The modeling framework may serve as a normal-hearing baseline for future microscopic models of speech perception that can account for effects of hearing-impairment and hearing-aid signal processing on phoneme perception.

Acknowledgments

We would like to thank two anonymous reviewers for their helpful and supportive comments. This research was funded with support from the European Commission under Contract No. FP7-PEOPLE-2011-290000.

4

Predicting effects of hearing-instrument signal processing on consonant perception[c]

Abstract

This study investigates the influence of hearing-aid (HA) and cochlear-implant (CI) processing on consonant perception in normal-hearing (NH) listeners. Measured data were compared to predictions obtained with a speech perception model [Zaar and Dau (2016). J. Acoust. Soc. Am., under review] that combines an auditory processing front end with a correlation-based template matching back end. In terms of HA processing, effects of strong nonlinear frequency compression and impulse-noise suppression were measured in 10 NH listeners using consonant-vowel stimuli. Regarding CI processing, the consonant perception data from DiNino et al. [(2016). J. Acoust. Soc. Am., under review] were considered, which were obtained with noise-vocoded vowel-consonant-vowel stimuli in 12 NH listeners. The inputs to the model were the same stimuli as were used in the corresponding experiments. The model predictions obtained for the two data sets showed a large agreement with the perceptual data both in terms of consonant recognition and confusions, demonstrating the model's sensitivity to supra-threshold effects of hearing-instrument signal processing on consonant perception. The results could be useful for the evaluation of hearing-instrument processing strategies, particularly when combined with simulations of individual hearing impairment.

[c] This chapter is based on Zaar et al. (2016).

4.1 Introduction

Speech perception is commonly tested by assessing the percentage of correctly identified words or sentences in the presence of some acoustical interference or degradation, such as additive noise and/or reverberation (cf. Hagerman, 1982; Nilsson et al., 1994; Wagener et al., 2003; Nielsen and Dau, 2009; Nielsen and Dau, 2011). While such speech tests provide some useful "macroscopic" information about the effects of different acoustic conditions on intelligibility, the typically used speech reception threshold measure (SRT; representing the signal-to-noise ratio at which 50% intelligibility is obtained) is rather coarse as it reflects responses averaged across many speech tokens. Furthermore, the listeners can "restore" missing acoustic information using semantic predictability and lexical information (e.g., Miller and Licklider, 1950; Warren, 1970; Bashford et al., 1992; Kashino, 2006), such that linguistic processing ability may strongly influence the listeners' performance. Moreover, the frequency importance function for the intelligibility of sentences is strongly dominated by the low-frequency speech content (Pavlovic, 1987), such that macroscopic speech intelligibility tests are not very sensitive to effects in the mid and high frequency ranges (e.g., due to high-frequency masking noise, filtering, or nonlinear speech processing). Therefore, such tests seem to be suboptimal for investigating effects of hearing impairment (which is typically most pronounced at high frequencies) and hearing-instrument signal processing on speech perception.

Instead, it can be insightful to examine the perception of individual phonemes, sometimes referred to as a "microscopic" approach of studying speech perception. Various studies have focused on the perception of consonants embedded in nonsense syllables in normal-hearing (NH) listeners (e.g., Miller and Nicely, 1955; Wang and Bilger, 1973; Phatak and Allen, 2007; Phatak et al., 2008; Zaar and Dau, 2015), e.g. in the form of consonant-vowel combinations (CVs like /ba/, /ta/, etc.), typically presented in steady-state noise at various signal-to-noise ratios (SNRs). In such tests, the contribution of high-level restoration effects is eliminated due to the nonsense nature of the stimuli and the importance of the critical high-frequency speech cues is taken into account as many consonant cues contain high-frequency energy (cf. Li et al., 2010; Li et al., 2012). Furthermore, not only the correct consonant recognition can be evaluated, but also consonant confusions, i.e., the type of error that occurred.

Several studies have investigated the effects of hearing impairment on consonant perception (e.g., Phatak et al., 2009; Trevino and Allen, 2013). Scheidiger and Allen (2013) studied the influence of different amplification schemes on consonant perception in hearing-impaired (HI) listeners and demonstrated that consonant perception tests may be more informative for hearing-aid (HA) fitting than pure-tone audiometry. Schmitt et al. (2016) presented a consonant perception test specifically designed for high-frequency HA fitting, which determines (i) the audibility thresholds of high-pass filtered representations of /s/ and /ʃ/ and (ii) the recognition thresholds of these consonants in a vowel-consonant-vowel (VCV) context (i.e., /asa, aʃa/). Testing HI listeners with and without HAs, they demonstrated that the test was sensitive to effects of high-frequency amplification as well as to effects of nonlinear frequency compression (NLFC). NLFC (Simpson et al., 2005) is designed to restore high-frequency acoustic information in listeners with pronounced high-frequency hearing loss by compressing the high-frequency signal content and shifting it to lower frequencies, as HAs typically cannot provide sufficient gain at frequencies above 5 kHz (Kimlinger et al., 2015). Glista et al. (2009) showed that NLFC can substantially improve high-frequency consonant recognition scores in listeners with a high-frequency hearing loss. However, NLFC with "too strong" settings can result in a drastic reduction of consonant recognition, as demonstrated by Schmitt et al. (2016). This is consistent with the strongly frequency dependent acoustic cues that lead to different consonant percepts (cf. Li et al., 2010; Li et al., 2012), as frequency-compressed high-frequency consonants may perceptually "morph" into other consonants that exhibit a temporally similar cue in a lower frequency region. For example, /s/ and /ʃ/ are represented by frication noise at very high and slightly lower frequencies, respectively, such that a too strong NLFC leads to /s/ being perceived as /ʃ/. However, such perceptual morphs may disappear after an acclimatization period due to re-learning of the modified consonant cues (cf. Wolfe et al., 2011). Consonant perception depends not only on the spectral characteristics of the signal but also on its temporal properties. Temporal signal modifications due to the highly nonlinear processing schemes typically applied in HAs (e.g., impulse-noise suppression, INS) may thus also affect consonant perception.

An alternative compensation strategy is represented by cochlear implant (CI) processing, applied in more severe cases of hearing impairment. CIs yield great improvements in terms of speech intelligibility by transmitting individual frequency bands of a signal directly to different places in the cochlea using an implanted electrode array. However, CIs are limited with respect to spectral resolution, as the number of electrodes in an implanted array is limited and channel interactions typically occur (White et al., 1984; Stickney et al., 2006). Furthermore, spectral resolution may be further degraded due to poor electrode-neuron interfaces – defined by regions of poor neural survival or large distance between the CI electrodes and the auditory neurons (for review see Bierer, 2010). DiNino et al. (2016) investigated the effect of CI processing with poor electrode-neuron interfaces on the perception of consonants and vowels in NH listeners using VCV and consonant-vowel-consonant (CVC) syllables, respectively, noise-vocoded to simulate CI processing. A reference CI simulation condition using all available channels was considered along with conditions where low-, middle-, and high-frequency channels were either set to zero ("Zero") simulating neural dead regions or re-distributed to neighboring channels ("Split") simulating poor electrode positioning. While listeners exhibited considerable perceptual differences across the considered frequency regions (but not across the Zero and Split conditions) in the vowel perception test, the consonant perception test showed less variability across frequency regions, as all CI processing conditions induced largely similar effects on consonant perception.

To better understand how various aspects of HA and CI processing affect consonant perception, computational models of speech perception may serve as valuable tools. If such a model can account for the effects of specific HA/CI processing strategies on consonant perception, it may provide useful information about the auditory cues that contribute to the recognition of a specific consonant or its confusion with another consonant. Several approaches for modeling consonant perception in NH listeners (Cooke, 2006; Jürgens and Brand, 2009) and in HI listeners (Holube and Kollmeier, 1996; Jürgens et al., 2014; Jepsen et al., 2014) have been proposed. While the mentioned models were shown to account for consonant recognition scores in masking noise (or in quiet at low signal levels), they did not account well for the consonant confusions, i.e., the predicted errors were different from the listeners' errors. However, effects of the described hearing-instrument signal processing approaches on consonant recognition are induced by specific strong confusions (cf. Schmitt et al., 2016; DiNino et al., 2016), which result from consonant-cue morphs/ambiguities due to the applied signal processing. In contrast to masking-noise conditions, the consonant cues are in such conditions not *masked* but rather *changed*; a model that can account for such effects therefore needs to be sensitive not only to the presence of a consonant cue, but also to its perceptual similarity with other consonant cues. Recently, Zaar and Dau (2016) proposed a consonant perception model that appears to provide such sensitivity. It combines an auditory model (Dau et al., 1996; Dau et al., 1997) that includes adaptive processes and modulation-frequency selective processing with a temporally dynamic correlation-based template-matching back end. The model was evaluated on the extensive data set by Zaar and Dau (2015), obtained in NH listeners with CVs presented in white noise at various SNRs. The model was shown to account well for consonant recognition even on the level of individual speech tokens. Moreover, a good agreement of the model predictions with the perceptual consonant confusions was demonstrated, albeit with some underestimation of the perceptual confusions' extent.

To evaluate the potential of the modeling approach of Zaar and Dau (2016) for predicting and exploring the effects of different hearing-instrument processing strategies on consonant perception, the present study investigated the model's predictive power in several HA and CI processing conditions. In particular, an experimental investigation of the effects of NLFC and INS on NH listeners' consonant perception was conducted using speech material from Schmitt et al. (2016). Strong settings were selected for the considered algorithms and no training was provided to the NH listeners to exclude effects of acclimatization. To test the model, these experimental data, which represent effects of HA processing on consonant perception, were used along with the consonant perception data from DiNino et al. (2016), representing effects of CI processing on consonant perception. Model predictions were obtained for the two data sets by feeding the respective stimuli to the consonant perception model of Zaar and Dau (2016). The model performance was evaluated by means of confusion matrix (CM) comparisons, as well as on the basis of correlation analyses of the perceptual and predicted consonant recognition and confusion scores.

4.2 Method

4.2.1 Experiment 1: Effects of HA signal processing

Stimuli and experimental conditions

The audio material was taken from the speech material recorded by Schmitt et al. (2016) and consisted of the VCVs /aba, aga, ada, apa, aka, ata, asa, aʃa, afa, atsa/[1], spoken by a female native German speaker. The speaker was trained to speak all VCVs with similar speed and pitch. Schmitt et al. (2016) used two versions of /asa/ and /aʃa/, respectively, filtered to have different spectral peaks: /ʃ/ exhibited a spectral peak at 4.6 kHz and was spectrally shaped to show spectral peaks at 3 kHz and 5 kHz, resulting in /aʃa3/ and /aʃa5/. /s/ exhibited a spectral peak at 7.2 kHz and was spectrally shaped to show spectral peaks at 6 kHz and 9 kHz, resulting in /asa6/ and /asa9/. For evaluating effects of impulse-noise suppression on consonant perception, the stimuli need to start with the consonant. Thus, the initial vowels of the considered twelve VCV tokens /aba, aga, ada, apa, aka, ata, asa6, asa9, aʃa3, aʃa5, afa, atsa/ were manually removed to obtain the CVs /ba, ga, da, pa, ka, ta, sa6, sa9, ʃa3, ʃa5, fa, tsa/.

Five conditions were considered: *unaided*, *default*, *NLFC*, *INS*, and *NLFC&INS*. The unaided condition was a natural listening situation without HA processing. For the other four conditions, Phonak Naida V90-RIC HAs were employed, assuming a moderate to severe hearing loss with 55 dB hearing level (HL) at frequencies of 1 kHz and below, 65 dB HL at 2 kHz, 75 dB HL at 4 kHz, and 80 dB HL at 8 kHz. The *default* condition was defined as the default HA settings suggested by the fitting software. In the *NLFC* condition, the strongest possible setting of the provided nonlinear frequency compression algorithm (Phonak SoundRecover) was selected, such that the frequency content in the range between 1.5 and 10 kHz was compressed by a factor of 4 to the range between 1.5 and 2.41 kHz. In the *INS* condition, the strongest possible setting of the provided impulse-noise suppression (Phonak SoundRelax) was selected. In the *NLFC&INS* condition, NLFC and INS were combined using the settings described above for the *NLFC* condition and the *INS* condition, respectively. For all HA settings, omni-directional microphone directivity was selected.

[1] Only the subset /ada, aha, ama, aka, asa, aʃa, afa/ of the recorded VCVs were eventually used in Schmitt et al. (2016). The present study used a different subset.

One sound file with all CVs was obtained by concatenating the CVs with 500-ms pauses between them. Steady-state speech-shaped noise (SSN) with a long-term average spectrum of female speech was added at an effective SNR of 8 dB (in the speech-containing portions). 10 seconds of noise alone preceded the first CV. The mixture of CVs and noise was played back from a loudspeaker, positioned at a distance of 1.5 m and 0°azimuth relative to a KEMAR dummy head in a sound-attenuating room. The speech level at the position of the dummy head was set to 70 dBA. The signals were recorded at a sampling rate of 48 kHz at the position of the dummy head's left tympanic membrane either without HA (unaided condition) or with HA using the condition-specific HA setting. The recordings were equalized to compensate for the applied amplification (half-gain rule) and cut into individual CV stimuli with 350 ms of noise at the beginning and 50 ms of noise at the end, using 50-ms raised-cosine ramps for fade in/out.

Listening test

Ten adult NH native German listeners (mean age: 29.5 years; standard deviation: 3.6 years) were tested. The listeners were seated in a sound-insulated booth in front of a computer screen and binaurally presented with the diotic stimuli via Sennheiser HD 650 headphones at 60 dB sound pressure level (SPL). They were asked to select the consonants they heard on a graphical user interface (GUI), which displayed the considered response alternatives /b, g, d, p, k, t, s, ʃ, f, ts/ in the corresponding German spelling (b, g, d, p, k, t, s, sch, f, z). Each of the 60 stimuli (12 CVs in five conditions) was presented 8 times to each listener, amounting to a total of 480 stimulus presentations per listener. The order of presentation was randomized across CVs and conditions. After the listener had made a decision, the next stimulus was played after a pause of 500 ms. The experiment duration was about 25 minutes per listener. No training or feedback was provided to the listeners. As some stimuli sounded rather ambiguous, listeners were instructed to select the response alternative that most closely resembled what they heard. The frequencies of responses obtained for each stimulus were summed across listeners and divided by the overall number of presentations (80; 8 presentations × 10 listeners) to obtain the proportions of responses.

4.2.2 Experiment 2: Effects of CI signal processing

DiNino et al. (2016) considered sixteen VCVs, consisting of consonants embedded in an /aCa/ context (/p/, “apa”; /t/, “ata”; /k/, “aka”; /b/, “aba”; /d/, “ada”; /g/, ”aga“; /f/, ”afa“; /θ/, “atha”; /s/, “asa”; /ʃ/, ”asha“; /v/, ”ava“; /z/, ”aza“; /dʒ/, “aja”; /m/, “ama”; /n/, “ana”; /l/, “ala”). All VCVs were naturally spoken by a male talker (native speaker of American English). Vocoder processing was applied to the stimuli to simulate CI processing in combination with regions of poor neural survival. The processing was designed to simulate the fidelity 120 processing strategy with the same frequency band allocations as Advanced Bionics devices and realized in Matlab using CI simulation software developed by Litvak et al. (2007). 15 vocoder bands with logarithmic spacing in the frequency range between 250 Hz and 8.7 kHz and a slope of 30 dB/octave were considered for the simulations. The subband envelopes of the VCVs were extracted, lowpass-filtered at 68 Hz, and used to modulate noise bands with the same center frequencies. As a control condition, the VCVs were processed using all vocoder bands (*AllChannels*). For the other six conditions, the spectral information in three frequency regions (*Apical* / 421 – 876 Hz; *Middle* / 877 – 1826 Hz; *Basal* / 1827 – 3808 Hz) was degraded by either (i) setting the corresponding channels to zero (*Zero*) or (ii) setting them to zero and adding half of the envelope energy from the zeroed channels to the neighboring lower-frequency channels and the other half to adjacent higher-frequency channels (*Split*). The noise bands were summed and the resulting vocoded stimuli were stored at a sampling rate of 17.4 kHz.

Twelve adult NH listeners with a mean age of 25.2 years participated in the study of DiNino et al. (2016). All listeners were native speakers of American English. All 112 VCV stimuli (15 VCVs × 7 conditions) were presented 3 times in random order to each listener at 60 dBA via a loudspeaker positioned one meter from the subject at 0°azimuth in a sound-insulated booth. Listeners were asked to select the consonant they heard on a computer screen. Two such experimental blocks were run, such that 6 responses per listener were obtained for each stimulus. Prior to the test run, listeners completed a practice run with feedback using the stimuli in the *AllChannels* condition only. The frequencies of responses were summed across listeners and divided by the overall number of stimulus presentations (72; 6 presentations × 12 listeners) to obtain the proportions of responses.

4.2.3 Model simulations

Model description

The consonant perception model of Zaar and Dau (2016) was considered for predicting the perceptual data obtained with the HA-processed CVs as well as with the CI-processed VCVs. Figure 4.1 shows the model, which combines the auditory model front end of Dau et al. (1996; 1997) with a temporally dynamic correlation-based back end. The auditory model consists of (i) a bank of 15 fourth-order gammatone filters with center frequencies logarithmically spaced between 315 Hz and 8 kHz, (ii) an envelope extraction stage (realized by half-wave rectification and lowpass filtering at 1 kHz), (iii) a chain of five adaptation loops (designed to mimic adaptive properties of the auditory periphery), and (iv) a bank of 4 modulation filters, implemented as a 2-Hz lowpass filter in parallel with three second-order bandpass filters with a Q-factor of 1 and center frequencies of 4, 8, and 16 Hz, respectively. For a given noisy speech signal, the temporal pattern of the noise alone (after the preprocessing stages) is subtracted from the corresponding temporal pattern of the noisy speech. The resulting model representations of the test signal (R_{test}) and of a set of templates (R_{t_1}, R_{t_2}, ..., R_{t_N}) are then aligned in time using a dynamic time warping (DTW) algorithm (Sakoe and Chiba, 1978). Finally, the cross-correlation coefficients between the time-aligned test-signal representation ($\widehat{R}_{test}$) and the time-aligned template representations ($\widehat{R}_{t_1}$, $\widehat{R}_{t_2}$, ..., $\widehat{R}_{t_N}$) are calculated and, after adding a constant-variance internal noise to limit the model's resolution, converted to response percentages.

Simulation procedure

To predict the data from experiment 1, the recorded HA-processed CVs that were used as experimental stimuli were fed to the model. Portions of the respective dummy-head recordings that contained only noise where considered as "noise alone" signals in the model (depending on the condition of the considered stimulus). The CV recordings obtained in the unaided condition were considered as templates since they had not been passed through a HA but still contained the effects of the noise, the room, and the KEMAR dummy head on the CV speech tokens. 9 templates were generated from each noisy CV recording by using 9 randomly selected samples of the noise alone, such that the template-matching procedure could be iterated 9 times. After obtaining the correlation coefficients

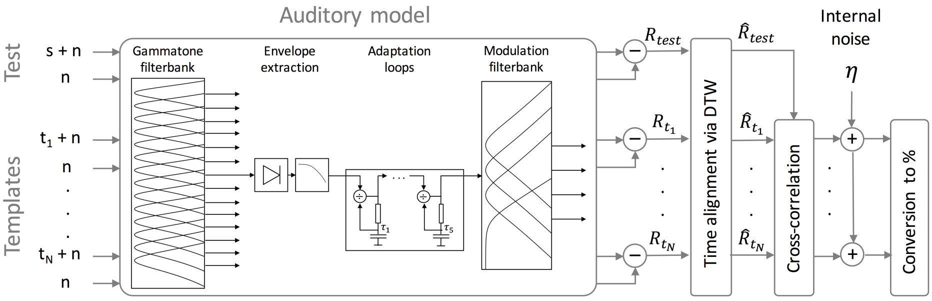


Figure 4.1: Scheme of the consonant perception model (reprint from Zaar and Dau, 2016). For the test signal and a set of templates, the noisy speech and the noise alone were passed separately through the auditory model, consisting of a gammatone filterbank, an envelope extraction stage, a chain of adaptation loops, and a modulation filterbank. The difference between the temporal patterns of the noisy speech and the noise alone was obtained. The resulting representations of the test signal and the templates were time-aligned using a dynamic time warping (DTW) algorithm. Finally, the cross-correlation coefficients between the test signal and each template were calculated and, after addition of a constant-variance internal noise, converted to percent.

between each test signal and all templates, the internal noise was added and the model response for each iteration was defined as the template showing the largest correlation with the test signal. As proposed in Zaar and Dau (2016), the model was calibrated by adjusting the variance of the internal noise based on the average consonant recognition scores obtained in the considered conditions. Here, a variance of $\sigma^2_{int,1} = 0.15$ was found to be optimal, which is larger than the variance of 0.05 used in Zaar and Dau (2016). This larger internal noise was necessary to account for the higher difficulty that the listeners experienced due to the HA signal processing conditions considered here (as compared to the additive white noise conditions considered in Zaar and Dau, 2016). However, the internal noise was held constant across the considered conditions. For each test signal, the numbers of occurrences of the model responses were divided by their sum to obtain the modeled proportions of responses.

The data from experiment 2, collected by DiNino et al. (2016), were predicted in a similar fashion, using the vocoded VCVs in the considered vocoder conditions as test signals and the unprocessed VCVs as templates. The model's gammatone filterbank was modified to comprise 20 filters with center frequencies logarithmically spaced between 100 Hz and 8 kHz to take the entire spectral content of the vocoded signals into account, which showed increasing energy above 100 Hz. This low-frequency extension was particularly relevant to cover the re-distributed channels in the *ApicalSplit* condition. In contrast to experiment 1, the experimental stimuli contained no additive noise. Therefore, the temporal patterns of the stimuli and the templates were here directly considered in the model back end, as no "noise alone" pattern could be obtained. Nine iterations of the model simulation were run using newly generated noise-vocoded stimuli in each iteration. As before, internal noise was added and the model response for each iteration was defined as the template showing the largest correlation with the test signal. An internal-noise variance of $\sigma^2_{int,2} = 0.071$ was found to be optimal based on the average recognition scores obtained in the considered conditions, which is in the same range as the variance of 0.05 used in Zaar and Dau (2016) and reflects the relatively low difficulty of the task (cf. Sec. 4.3.2). The internal noise was held constant across the considered conditions of experiment 2. For each VCV in each condition, the numbers of occurrences of the model responses were divided by their sum to obtain the modeled proportions of responses.

4.3 Results and analysis

4.3.1 Effects of HA signal processing

The grand average consonant recognition scores obtained in the five experimental conditions considered in experiment 1 are shown in Table 4.1. The recognition was at ceiling for the *unaided* condition (96%), the *default* HA condition (94%), and the *INS* condition (92%). In contrast, largely reduced recognition scores were observed in the conditions with NLFC, namely *NLFC* (55%) and *NLFC&INS* (56%). The large standard deviations across consonants (36% and 34%, respectively) indicate that the perception of specific consonants was strongly affected by the HA processing while other consonants remained perceptually unaffected. As only the results obtained in the *NLFC* and *NLFC&INS* conditions showed substantial perceptual effects of the applied HA processing, the remainder of this section focuses solely on these two conditions.

On average, the model predicted a slightly larger recognition score (59%) than observed in the listeners (55%) for the *NLFC* condition, whereas it predicted a slightly lower recognition score (51%) than observed in the listeners (56%) for the *NLFC&INS* condition. To inspect the data more closely in terms of the stimulus-specific recognition and confusion scores, Fig. 4.2 shows the measured

Table 4.1: Grand average consonant recognition scores measured in experiment 1 for each condition along with the standard deviations across stimuli.

Condition	% correct	Std in %
Unaided	95.9	8.1
Default	93.7	7.3
NLFC	55.3	36.2
INS	92.3	10.8
NLFC&INS	56.2	34.3

and predicted confusion matrices (CMs) obtained in the *NLFC* and *NLFC&INS* conditions. The vertical axes indicate the 12 presented consonants (including the different realizations considered for /s/ and /ʃ/, cf. Sec. 4.2.1), while the horizontal axes represent the 10 consonants provided as response alternatives. The perceptual data (filled gray circles) and the predictions (open red circles) are depicted as circles of different sizes that correspond to the percentage categories shown in the figure's legend.

In the *NLFC* condition (left panel), the listeners exhibited distinct consonant confusions. Most notably, the gray filled circles indicate that /d/ was confused with /b/, /t/ was confused with /k/, /s6, s9/ were confused with /ʃ/, /ʃ3, ʃ5/ were confused with /f/, and /ts/ was confused with /ʃ, f/. The recognition scores for the mentioned stimuli were thus reduced, with particularly low scores for /s6, s9, ts/. The model provided convincing predictions of the stimulus-specific recognition scores, as indicated by the good agreement of the red and gray circles on the "diagonal" of the CM (which has two "steps" as two representations of /s, ʃ/ were considered as stimuli). Furthermore, the model predicted some of the confusions remarkably well (particularly for /d, s6, s9, ts/), although the extent of the confusions was partly underestimated (consistent with the observations in Zaar and Dau, 2016). However, some distinct confusions were not accounted for by the model (/t/ confused with /k/) or predicted to a lesser extent such that they are not visible in Fig 4.2. For example, /ʃ3, ʃ5/ were confused with /f/, but the predicted response probabilities for /f/ were just below 7%. Moreover, the model predicted some additional confusions that were not observed in the perceptual data, in particular /ts/ confused with /t/.

The perceptual data obtained in the *NLFC&INS* condition (right panel of Fig. 4.2) were largely comparable to the data obtained in the *NLFC* condition (left panel). However, some clear differences can be observed (gray circles), as in the *NLFC&INS* condition /k/ was confused with /p, f/ and the confusion of /t/ with /k/ observed in the *NLFC* condition disappeared. Furthermore, /ts/ was not recognized at all in the *NLFC* condition, but was recognized to some extent in the *NLFC&INS* condition. The model predictions (red circles) captured these perceptual changes between the *NLFC* and the *NLFC&INS* condition well, apart from the confusion of /k/ with /f/, which was not accounted for by the model.

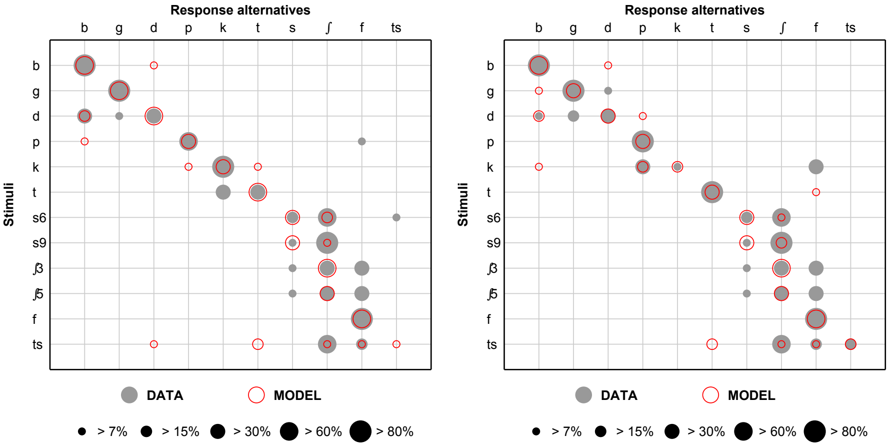


Figure 4.2: Measured and predicted confusion matrices obtained in experiment 1 with CVs processed with *NLFC* (left panel) *NLFC&INS* (right panel). The presented consonants are shown on the vertical axis and the response alternatives on the horizontal axis. The filled gray circles represent the perceptual data while the open red circles show the model predictions. The size of the circles indicates the proportions of responses according to the five categories provided in the legend.

To evaluate the significance of the agreement between the measured and the predicted stimulus-specific consonant recognition scores (on-diagonal elements of the CMs), a correlation analysis was conducted. Table 4.2 summarizes the results, which revealed that the measured and predicted recognition scores were significantly ($p < 0.05$) correlated across stimuli for both the *NLFC* ($r = 0.56$) and the *NLFC&INS* ($r = 0.67$) condition.

To also quantify the agreement between the measured and predicted confusions, a correlation analysis of the consonant confusions was performed. For each stimulus, the correlation between the erroneous part of the measured and predicted response patterns (off-diagonal elements of the CMs) was obtained across response alternatives. This analysis was only performed for the stimuli that showed an error of $P_e > 20\%$ in the perceptual data. Table 4.3 shows the

Table 4.2: Pearson's correlation coefficients across stimuli between measured and predicted consonant recognition scores obtained in the *NLFC* and *NLFC&INS* conditions of experiment 1 along with the corresponding p-values. p-values indicating significant correlation ($p < 0.05$) are given in bold font.

Condition	% Pearson's r	p-value
NLFC	0.5588	**0.0295**
NLFC&INS	0.6651	**0.0091**

results of the confusion correlation analysis, which revealed that the confusions were positively correlated for all considered stimuli, with most correlations being significant. In the *NLFC* condition, large correlations ($r > 0.88$) were obtained for /d, s6, s9, ʃ3, ʃ5/ but not for /p, t, ts/, i.e., the confusions were highly correlated for 5 out of the 8 stimuli with $P_e > 20\%$. In the *NLFC&INS* condition, large correlations ($r > 0.62$) were obtained for /k, s6, s9, ʃ3, ʃ5/ but not for /d, ts/, indicating highly correlated confusion patterns for 5 out of the 7 stimuli with $P_e > 20\%$. This is consistent with the observations made based on the CMs in Fig. 4.2, apart from the large confusion correlations found in the two conditions for /ʃ3, ʃ5/, for which the model predicted confusions below 7%, which are thus not displayed in Fig. 4.2. The patterns of predicted confusions were in these cases merely scaled down but qualitatively similar to the measured ones, resulting in large confusion correlations.

4.3.2 Effects of CI signal processing

Table 4.4 shows the grand average measured and predicted consonant recognition scores obtained in the seven experimental conditions of experiment 2 along with the standard deviations across stimuli. As reported by DiNino et al. (2016), the measured recognition scores, including the *AllChannels* condition, were below ceiling and showed little variability across conditions (73% ± 5%) and a large variability across stimuli (with standard deviations of about 30%). The predicted recognition scores exhibited a similar behavior, albeit with a somewhat smaller variability across stimuli (with standard deviations of about 18.5%).

Figure 4.3 shows the measured (filled gray circles) and predicted (open red circles) CMs obtained in the *AllChannels* control condition. The main measured confusions were /g/ with /d/, /p/ with /t/, /k/ with /t/, and /th/ with /v/,

Table 4.3: Pearson's correlation coefficients across response alternatives between measured and predicted consonant confusion patterns obtained in the *NLFC* and *NLFC&INS* conditions of experiment 1 along with the corresponding p-values. p-values indicating significant correlation ($p < 0.05$) are given in bold font. The confusion correlation was only obtained for stimuli with a measured error of $P_e > 20\%$.

	NLFC		*NLFC&INS*	
Consonant	r	p	r	p
/b/	–	–	–	–
/g/	–	–	–	–
/d/	0.9676	**0.0000**	0.2478	0.2601
/p/	0.1561	0.3442	–	–
/k/	–	–	0.6239	**0.0363**
/t/	0.1181	0.3811	–	–
/s6/	0.9365	**0.0001**	0.9314	**0.0001**
/s9/	0.9662	**0.0000**	0.9671	**0.0000**
/ʃ3/	0.8891	**0.0007**	0.6490	**0.0293**
/ʃ5/	0.8829	**0.0008**	0.7779	**0.0068**
/f/	–	–	–	–
/ts/	0.2473	0.2606	0.0538	0.4454

Table 4.4: Grand average consonant recognition scores measured and predicted for each condition of experiment 2 along with the standard deviations across stimuli.

	Perceptual data		Model predictions	
Condition	% correct	Std in %	% correct	Std in %
AllChannels	77.9	28.9	74.8	15.9
ApicalZero	70.1	29.9	74.3	18.7
ApicalSplit	73.7	29.1	71.3	21.0
MiddleZero	70.7	32.7	71.8	21.1
MiddleSplit	73.1	32.6	72.3	19.0
BasalZero	69.5	30.9	72.2	15.8
BasalSplit	74.0	25.1	74.8	18.3

which resulted in low recognition scores for these stimuli. The main confusions were well accounted for but slightly underestimated by the model, except for /th/ confused with /v/, where the model predicted a perfect recognition of /th/. Thus, the predicted stimulus-specific recognition scores (along the CM's diagonal) showed a similar trend as their measured counterparts, except for the recognition score for /th/. However, the model also predicted some confusions that were not represented in the data. These "false alarms" were typically made within the consonant categories voiced stops (/b, g, d/), unvoiced stops (/p, k, t/), fricatives (/f, v, th, s, z, sh, j/), and nasals (/m, n/).

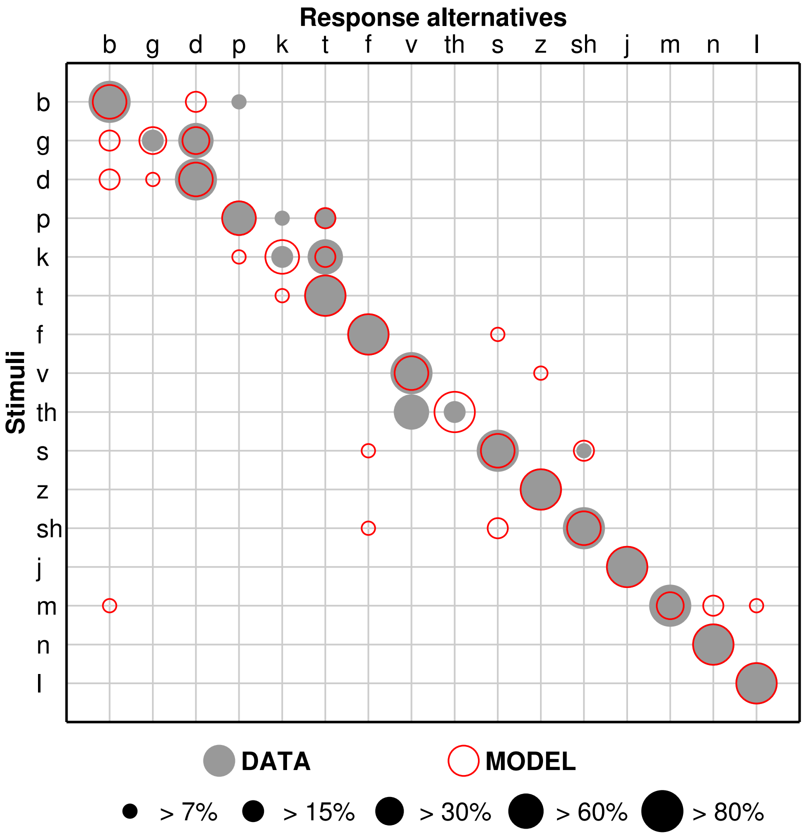


Figure 4.3: Measured and predicted confusion matrices obtained in the *AllChannels* condition of experiment 2. The data is presented in a similar manner as in Fig. 4.2.

Figure 4.4 shows the measured and predicted CMs obtained in the condition with the overall best fitting (*MiddleSplit*, left panel) and least fitting predictions

(*BasalZero*, right panel), in terms of the recognition score correlation across stimuli (cf. Table 4.5). The main confusions observed in the *MiddleSplit* condition (left panel, filled gray circles) were the same as the ones measured in the *AllChannels* condition, namely /g/ with /d/, /p/ with /t/, /k/ with /t/, and /th/ with /v/. The model predictions were also very similar to the *AllChannels* condition, capturing the main measured confusions except for /th/ confused with /v/. Accordingly, the predicted stimulus-specific recognition scores showed a similar trend as the measured ones, again except for the /th/, for which the model predicted a too high recognition score. The data measured in the *BasalZero* condition (right panel) also followed the same main trends. However, an additional large perceptual confusion of /sh/ with /s/ can be observed along with a reduction in the recognition score for /sh/. This additional confusion was correctly predicted by the model.

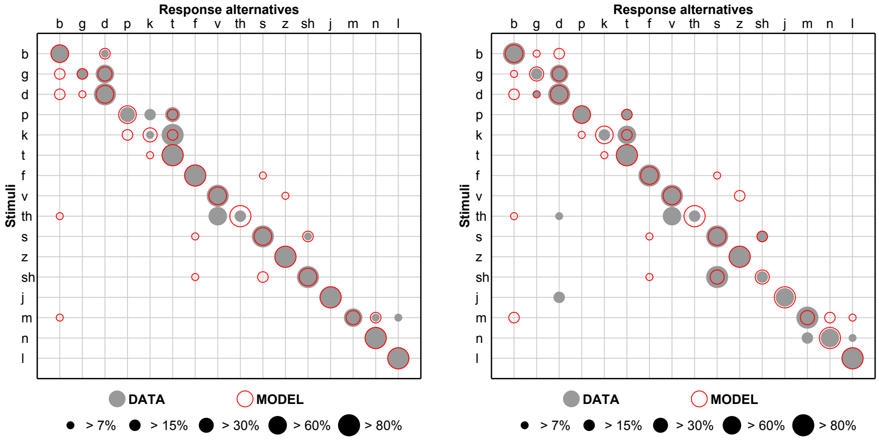


Figure 4.4: Measured and predicted confusion matrices obtained in the *MiddleSplit* (left panel) and *BasalZero* (right panel) conditions of experiment 2. The data is presented in a similar manner as in Fig. 4.2 and 4.3.

To evaluate the significance of the agreement between the measured and the predicted stimulus-specific consonant recognition scores, a correlation analysis was conducted. Table 4.5 summarizes the results, which revealed that the measured and predicted recognition scores (on-diagonal elements of the CMs) were significantly ($p < 0.05$) correlated across stimuli for all but the *AllChannels* and *BasalZero* conditions. As the results obtained for /th/, which showed a strong confusion with /v/ and a correspondingly low recognition score in all conditions, seemed to be strongly biased by the low phoneme frequency

Table 4.5: Pearson's correlation coefficients across stimuli between measured and predicted consonant recognition scores obtained in each condition of experiment 2 along with the corresponding p-values. p-values indicating significant correlation ($p < 0.05$) are given in bold font. The values in parentheses represent the analysis results obtained when omitting the recognition score for /th/.

Condition	% Pearson's r	p-value
AllChannels	0.4249 (0.6373)	0.0505 **(0.0053)**
ApicalZero	0.5233 (0.7453)	**0.0187 (0.0007)**
ApicalSplit	0.4509 (0.7055)	**0.0398 (0.0016)**
MiddleZero	0.5967 (0.7897)	**0.0073 (0.0002)**
MiddleSplit	0.6120 (0.7532)	**0.0059 (0.0006)**
BasalZero	0.3087 (0.4303)	0.1223 (0.0547)
BasalSplit	0.4314 (0.6628)	**0.0476 (0.0066)**

of /th/ in the English language, an additional analysis was conducted omitting the /th/ recognition scores. The analysis results without /th/ are presented in parentheses in Table 4.5 and indicate significant recognition score correlations for all conditions, apart from the *BasalZero* condition, for which a p-value slightly greater than 0.05 was obtained.

A correlation analysis of the consonant confusions was performed to also quantify the relation between the measured and the predicted confusions using only the erroneous part of the response patterns (off-diagonal elements of the CMs). As in Sec. 4.3.1, this analysis was conducted only for the stimuli that showed a perceptual error of $P_e > 20\%$. Table 4.6 summarizes the results, which revealed that the confusion correlations for the considered stimuli were very large (mostly above $r = 0.8$) and significant ($p < 0.05$) for the majority of the considered stimuli. However, as observed in the CMs (Fig. 4.3 and 4.4), the /th/ confusions were not well predicted by the model (presumably because they originated from a phoneme-frequency effect rather than from the signal characteristics) and the measured and predicted confusions obtained for /b, d/ in the two *Apical* conditions and for /j/ in the *BasalZero* condition showed either weak correlations or none at all.

Table 4.6: Pearson's correlation coefficients across response alternatives between measured and predicted consonant confusion patterns obtained in each condition of experiment 2. Correlation coefficients indicating significant correlation ($p < 0.05$) are given in bold font. The confusion correlation was only obtained for stimuli with a measured error of $P_e > 20\%$.

	AllChannels	*Apical*		*Middle*		*Basal*	
Consonant		*Zero*	*Split*	*Zero*	*Split*	*Zero*	*Split*
/b/	–	0.0006	-0.0360	–	**0.9645**	–	–
/g/	**0.9151**	**0.8733**	**0.9013**	**0.8810**	**0.9310**	**0.9525**	**0.8565**
/d/	–	0.2116	0.3831	–	–	–	**0.5133**
/p/	**0.9622**	**0.9313**	**0.9825**	**0.9736**	**0.9361**	**0.9328**	**0.8711**
/k/	**0.8967**	**0.7110**	**0.8196**	**0.7883**	**0.8178**	**0.8631**	**0.8416**
/t/	–	–	–	–	–	–	–
/f/	–	–	–	–	–	–	–
/v/	–	–	–	–	–	–	–
/th/	0.0604	-0.1052	-0.0302	0.3798	-0.0526	0.0814	0.0182
/s/	–	–	–	–	–	–	–
/z/	–	–	–	–	–	–	–
/sh/	–	–	–	–	–	**0.9510**	**0.9595**
/j/	–	–	–	–	–	0.1086	–
/m/	–	–	–	**0.4965**	**0.6813**	–	–
/n/	–	**0.8277**	**0.8055**	**0.7611**	–	**0.9025**	**0.8063**
/l/	–	–	–	–	–	–	–

4.4 Discussion

4.4.1 Relation to other studies

The detrimental effects of NLFC on consonant perception observed in experiment 1 are consistent with the results reported by Schmitt et al. (2016) for HI listeners provided with "too strong" NLFC. The present study, which used a modified version of the speech material from Schmitt et al. (2016), showed similarly strong detrimental effects of NLFC on the recognition of the consonants /s6/ and /s9/ (see their Fig. 7). This loss of recognition was shown here to result from a strong confusion of /s/ with /ʃ/, as also discussed in Schmitt et al. (2016). These findings do *not* contradict studies showing large improvements of high-frequency consonant perception with NLFC in HI listeners (e.g., Glista et al., 2009), as (i) NH listeners were tested in the present study such that no benefit was expected, (ii) strong NLFC settings were used, and (iii) effects of increasing performance/acclimatization over time (cf. Wolfe et al., 2011) were not considered.

The results from experiment 1 revealed that consonant confusions induced by strong NLFC only occurred within the categories voiced stops, unvoiced stops, and fricatives. Li et al. (2010; 2012) demonstrated that the consonant cues within each of these categories exhibit a similar temporal structure but differ with respect to their spectral energy distributions. The observed effects can therefore be assumed to be caused by spectral changes, resulting from a too strong frequency compression applied to the high-frequency consonant cues. The only substantial confusion that did not fall within the above mentioned categories (/k/ confused with /f/) resulted from combining NLFC with INS, which suppresses sharp onsets and thus produces nonlinear changes *over time*. The CI processing applied in experiment 2 (DiNino et al., 2016) induced confusions within the categories voiced stops, unvoiced stops, fricatives, and nasals. According to Li et al. (2010; 2012), this again indicates a main perceptual effect of the spectral changes caused by the CI processing.

The present study suggests that the considered model is not limited to conditions of stationary noise but also accounts to a large extent for highly nonlinear signal modifications, implying a versatility that has not been reported so far. The prediction performance was found to be comparable to that reported by Zaar and Dau (2016) for CVs in stationary masking noise in that (i) the predicted stimulus-specific recognition scores were, overall, strongly correlated with the measured recognition scores, (ii) the consonant confusions were mostly well accounted for by the model even though the extent of the confusions was slightly underestimated, and (iii) the model predicted some additional confusions incorrectly ("false alarms"), which mostly fell within perceptually plausible confusion groups. In contrast to Zaar and Dau (2016), effects of masking played a negligible role in the present study as the variability in the perceptual data was mainly induced by changes in the consonant cues of the processed stimuli, resulting in strong confusions. Thus, the results of the present study suggest that the model is sensitive to modifications of consonant cues and the resulting perceptual changes/ambiguities.

4.4.2 Limitations of the approach

The model tended to slightly underestimate the extent of the measured consonant confusions and partly predicted additional confusions that were not reflected in the perceptual data. This resulted most likely from a bias induced by similarities/dissimilarities in the vowel portions of the CV/VCV stimuli and templates, which are only partly related to the consonant percept. However, a separation of the signals into consonant and vowel portions is not feasible, particularly for speech tokens containing voiced consonants, which rely on formant transitions in the adjacent portions of the accompanying vowels.

Furthermore, the model does not take any linguistic processing into account, which can play a certain role in the perceptual results despite the nonsense nature of the stimuli. For example, the consistent perceptual morph of /th/ to /v/ in experiment 2 presumably[2] resulted from a perceptual bias induced by the large phoneme frequency of /v/ and the low phoneme frequency of /th/ in the English language. This was not accounted for by the model, which is based solely on the similarity of the signals.

4.4.3 Perspectives

An important extension of the model would be to incorporate aspects of hearing impairment, such as elevated audiometric thresholds, reduced frequency selectivity, loss of compression and other supra-threshold deficits (cf. Jürgens et al., 2014; Jepsen et al., 2014). The results of the present study suggest that, if a version of the model that can account for consonant perception in unaided HI listeners was established, the effects of hearing-instrument compensation strategies might be well-represented in the model predictions. Thus, the model could provide guidance regarding HA fitting, e.g., by suggesting specific fitting parameters based on a listener's auditory profile. Furthermore, it might become feasible to predict the *long-term* effects (including acclimatization) of specific HA or CI processing strategies on consonant perception by applying the respective signal processing not only to the stimuli, but also to the templates (instead of using unprocessed speech tokens as templates).

4.5 Summary and conlusion

The present study evaluated the predictive power of the model of Zaar and Dau (2016) regarding effects of HA and CI signal processing on consonant

[2] This assumption is based on informal listenining. The perception of the authors matched the confusions observed in the data for all stimuli except the /atha/ stimuli, which seemed well recognizable.

perception. Experiment 1 considered consonant perception in NH listeners after HA processing in terms of nonlinear frequency compression and impulse-noise suppression using CVs. Experiment 2 considered consonant perception in NH listeners after CI processing with different simulations of poor electrode-neuron interfaces using VCVs. The model was shown to account for most perceptual effects observed in the data from experiment 1. In particular, the stimulus-specific predicted recognition scores were significantly correlated with the measured ones, as well as most of the stimulus-specific confusion patterns. Furthermore, the model accounted to a large extent for the data from experiment 2, i.e., for the effects of the CI signal processing on consonant perception. Specifically, the simulated stimulus-specific recognition scores were significantly correlated with the measured ones in most conditions. Moreover, the vast majority of the stimulus-specific predicted confusion patterns was highly significantly correlated with the perceptual data.

The results indicate that the large predictive power of the model, previously demonstrated for consonant recognition and confusions obtained with CVs in stationary noise (Zaar and Dau, 2016), also extends to supra-threshold effects of hearing-instrument signal processing on consonant perception. This suggests a large potential of the model for evaluating and adjusting such processing schemes, in particular when extended to account for individual hearing impairment.

Acknowledgments

This research was funded with support from the European Commission under Contract No. FP7-PEOPLE-2011-290000. The authors wish to thank Leonid Litvak for providing the vocoder simulation code used in experiment 2.

5 Overall discussion

5.1 Summary of main results

This thesis described an extensive experimental investigation of some of the factors that influence consonant-in-noise perception in NH listeners as well as the development and evaluation of a computational model of consonant perception. The experimental investigation, described in Chapter 2, was conducted to clarify the level of detail and the methods required for measuring and analyzing consonant perception data. While representing a valuable contribution in itself, the results of this investigation were also essential for the development and evaluation of the model framework. The model framework proposed in this thesis was designed as an extension of the auditory processing model of Dau et al. (1997) towards predicting consonant perception by means of a temporally dynamic template-matching process, maintaining the crucial aspects of the original model's decision stage. The proposed model was evaluated based on the experimental data from Chapter 2 (CVs in stationary noise) as well as based on experimental data obtained in conditions of hearing-aid and cochlear-implant signal processing (Chapter 4).

Chapter 2 (Zaar and Dau, 2015) investigated the role of various sources of variability in consonant perception in steady-state masking noise. The investigation was motivated by previous studies (Phatak et al., 2008; Singh and Allen, 2012; Toscano and Allen, 2014) that demonstrated large perceptual differences for different speech tokens of the same type, spoken by different talkers. However, a systematic analysis of the factors that influence consonant perception had so far not been undertaken. Two categories of perceptual variability were considered: (i) *source-induced* variability, which comprises perceptual differences resulting from acoustical differences in stimuli of the same phonetic identity (caused by different speech tokens or different masking-noise waveforms) and (ii) *receiver-related* variability, which refers to perceptual differences across listeners and within listeners. The data were analyzed by means of a perceptual distance measure (cf. Scheidiger and Allen, 2013), which was shown to be related to (and more straightforward than) the commonly used entropy of responses (cf. Miller and Nicely, 1955; Phatak et al., 2008).

Consistent with Phatak et al. (2008), Singh and Allen (2012), and Toscano and Allen (2014), a large perceptual effect of across-talker articulatory differences was demonstrated, which has been neglected in earlier studies (Miller and Nicely, 1955; Wang and Bilger, 1973). Additionally, the study showed that different utterances spoken by the same talker induced equally large perceptual differences. Furthermore, even a slight temporal shift in the waveform of the steady-state masking noise was found to produce a significant perceptual effect. Regarding the receiver-related variability, it was demonstrated that the perceptual differences across the NH listeners, obtained with identical stimuli, were large (in the range of the effect induced by different speech tokens). In contrast, the perceptual differences measured in individual listeners in test and retest ("within-listener variability", reflecting the uncertainty of individual listeners) were small, indicating a large reproducibility of the responses on a listener-by-listener basis. This within-listener variability was found to depend inversely on the SNR, i.e., the "internal noise" (listener uncertainty) was proportional to the "external noise" (acoustic noise).

The source-induced perceptual effects, caused by differences in the speech tokens and even slight differences in the masking noise, suggest that consonant perception strongly depends on very fine details in the stimuli, which has so far not been sufficiently taken into account in related studies. The receiver-related effects indicate that not only the exact waveform of the stimulus, but also the individual listener plays a major role in consonant perception, even when considering NH listeners with the same language background. It is concluded that the complex interaction between source and receiver revealed here should either be *averaged out* by considering many representations of each factor (speech tokens, noise waveforms, and listeners), or measured *in detail* by employing a well-controlled set of stimuli (consisting of a limited amount of speech tokens and noise waveforms) and evaluating the results in individual listeners. Given the sensitivity of listeners to fine details in the stimuli, the latter may represent a valuable approach for auditory profiling and investigating effects of hearing-instrument processing (as demonstrated in Chapter 4 as well as by Schmitt et al., 2016 and DiNino et al., 2016).

Chapter 3 (Zaar and Dau, 2016) proposed a computational model of microscopic speech perception, which is based on the auditory processing model of Dau et al. (1997) and combines a front end that represents peripheral auditory processing with a temporally dynamic correlation-based template-matching decision stage. The front end consists of a frequency-selective process, an envelope extraction stage, a chain of adaptation loops (performing envelope-onset enhancement), and a modulation-frequency selective process. Using a-priori knowledge about the noisy stimulus and the noise alone, the decision stage temporally aligns the stimulus and the templates using DTW and defines the model response as the template that shows the maximum correlation with the stimulus. Motivated by the findings from Chapter 2, a constant-variance "internal-noise" process was applied that globally calibrates the model. While the model front end is similar to that of a related model by Jürgens and Brand (2009), the decision stage differs considerably and represents a rather straightforward extension of the original model's decision stage (Dau et al., 1997) towards predicting consonant perception.

The proposed model was evaluated using the experimental data and stimuli from Chapter 2. Motivated by the experimental findings, the model performance was analyzed for different levels of detail, down to the level of individual combinations of speech and noise tokens. However, as no audiometric differences had been measured in the listeners, the described across-listener perceptual variability could not be simulated and the data were instead averaged across listeners. The model showed highly accurate predictions of the grand average recognition scores (as a function of SNR), and also accounted well for the consonant-specific recognition scores and even for the speech-token specific recognition scores. In terms of consonant confusions, the model predictions showed a substantial similarity with the measured confusions at the consonant-specific level, while the confusion predictions were less convincing for some of the stimuli at the token-specific level. Nonetheless, the predictive power of the proposed model exceeded the performance of the related models by Jürgens and Brand (2009) and Cooke (2006) substantially, at least as far as a comparison based on the different sets of data is feasible. While largely accurate recognition score predictions were demonstrated for these related models, the proposed model's recognition score predictions generalized to a larger range of SNRs and to the level of individual speech tokens. Moreover, the proposed model's confusions were the same as, or perceptually related to, the perceptual confusions, while the related models (Cooke, 2006; Jürgens and

Brand, 2009) did not account well for confusions. This indicates that the model is not only sensitive to the presence of the consonant cues (recognition), but also to their similarities with other consonant cues that may arise from noise masking, resulting in perceptual confusions. The fact that the model made similar errors as the listeners might represent a crucial step towards predicting effects of severe speech signal *modifications* on consonant perception (e.g., due to hearing-instrument signal processing), as opposed to effects of additive masking-noise.

Chapter 4 (Zaar and Dau, 2016) evaluated the predictive power of the proposed model with respect to effects of HA and CI signal processing on consonant perception. The perceptual effects of strongly parametrized nonlinear frequency compression (NLFC) and impulse-noise suppression (INS) on consonant perception were measured in NH listeners. The experimental results showed the expected detrimental effects on consonant recognition for strong NLFC as well as for NLFC combined with INS, whereas INS alone and a default hearing-aid setting yielded recognition at ceiling. The loss of consonant recognition was found to result from strong confusions of specific consonants with other consonants (e.g. /s/ with /ʃ/), some of which substantially exceeded the corresponding recognition score (i.e., some consonants perceptually "morphed" into others). This is consistent with the results from Schmitt et al. (2016) obtained in HI listeners provided with "too strong" NLFC; the results were furthermore in agreement with the findings from Li et al. (2010; 2012) regarding the spectro-temporal consonant cue regions.

The data obtained in the conditions with NLFC and the corresponding stimuli were considered in the model framework. Additionally, effects of CI processing on consonant perception, measured with noise-vocoded VCVs in NH listeners by DiNino et al. (2016), were simulated based on the corresponding stimuli. The model was demonstrated to yield remarkably accurate predictions for the two data sets, both in terms of consonant recognition and consonant confusions. As noise masking played a negligible role in the considered experimental stimuli, the predictive power of the model in these conditions arises mainly from its ability to predict consonant confusions. Thus, the proposed model is not limited to conditions of additive stationary noise (as considered in Chapter 3) but also accounts to a large extent for highly nonlinear signal modifications.

5.2 The role of the model's decision stage

The model proposed in this thesis was partly inspired by a related approach of Jürgens and Brand (2009), which extended the same auditory processing model (Dau et al., 1997) as used in the present thesis by modifying various components in its decision stage. The auditory processing front ends of the two models are thus essentially identical and the reasons for the larger predictive power[1] of the proposed model must be connected to the model's decision stage. The model front end was chosen because of the extensive previous modeling efforts (not reported here) and the large predictive power of the original model of Dau et al. (1997). Three fundamental differences in the decision-making process are discussed here.

First, Jürgens and Brand (2009) directly fed the IRs of the noisy speech tokens to the back end whereas the present study followed the original model from Dau et al. (1997) in that the output of the front end was the difference between the IR of the noisy speech and the IR of the noise alone, which is furthermore related to the concept of the $\mathrm{SNR}_{\mathrm{env}}$ applied in the sEPSM model (Jørgensen and Dau, 2011; Jørgensen et al., 2013). Additionally, the measured perceptual effect of slight differences in the noise waveforms (Chapter 2) indicates that the noise waveform should explicitly be taken into account. Furthermore, recent findings from animal studies (e.g. Mesgarani et al., 2014) suggest the existence of a de-noising mechanism at higher stages of auditory processing. The use of a-priori knowledge about the noise may be interpreted as a simplistic way of simulating such a mechanism. In the considered conditions with white masking noise, the assumption of perfect a-priori knowledge about the noise (and thus ideal streaming) was necessary to correctly predict the perceptual robustness of high-frequency cues with respect to noise masking. In contrast, Jürgens and Brand (2009) could partly predict the robustness of high-frequency cues without this assumption, which was presumably due to the differences in the considered noise spectra (white vs speech-shaped).

Second, the model's decision was based on the maximum cross-correlation (as in Dau et al., 1997; see also Gallun and Souza, 2008) of the time-aligned IRs

[1] It should be taken into account that the proposed model was evaluated on CVs in white noise while the related model of Jürgens and Brand (2009) was tested with VCVs in speech-shaped noise, such that differences in the predictive power may also arise from differences in the considered stimuli and data.

(via DTW) of the test signal and the templates, as opposed to the minimum distance used by Jürgens and Brand (2009). The cross-correlation is insensitive to level differences, i.e., it solely describes covariation between the test-signal and template activation patterns obtained in the auditory model front end. This may be more closely related to the perceptual decision-making process than a distance measure, which is by definition sensitive to level differences and may thus result in prediction biases. A strongly biased behavior of a distance-based decision stage was observed in earlier versions of the proposed model as well as in the study of Jürgens and Brand (2009). The correlation-based back end alleviated this problem to a large extent and, thus, yielded realistic predictions in terms of consonant recognition and confusion scores.

Third, a constant-variance internal noise was added in the model's decision stage, which represents the listeners' uncertainty (cf. Dau et al., 1997) and was also reflected in the experimental results obtained in Chapter 2 ("within-listener distance"). This provided a realistic amount of uncertainty at medium to large SNRs, where the predicted recognition scores otherwise exceeded the measured ones. A similar trend was also reported by Jürgens and Brand (2009), who did not include uncertainty in the decision stage. The findings from Chapter 2 suggested to use internal-noise that increases with decreasing SNR (i.e., along with the effect of the "external" noise). However, from a modeling perspective, this assumption is only valid if the test-signal SNR is always large and the external-noise induced uncertainty is thus solely represented by the internal-noise term. As the predictions were obtained with test signals at all SNRs considered in the experiment, the constant-variance internal noise, which globally calibrates the model, was sufficient.

5.3 Limitations of the considered approaches

Regarding the experimental investigation of sources of variability in consonant perception described in Chapter 2, the considered potential influencing factors by no means represent any "complete picture". To keep the experiment duration in a feasible range, several parameters that are known to play a perceptual role were *fixed* to solely focus on specific sources of variability. Specifically, the vowel (/i/), the type of consonant-vowel combination (CV) and the spectral shape of the noise (white) were fixed. The same holds for the choice of response alternatives, the response method, and the instructions given to the test subjects.

The influence of these parameters, which also represent sources of variability, was thus neglected. Furthermore, the experimental results were based on a set of 90 speech tokens, spoken by two talkers, and presented to two different panels of eight listeners. The results may therefore be biased by the choice of speech tokens, talkers, and listeners. Moreover, the sources of variability investigated here represent *categories* and, thus, only provide indications of the relative contributions of these categories (e.g., across-talker articulatory differences) to consonant perception. Which specific properties of the stimuli caused the observed perceptual effects remains an open question, which may be addressed in further investigations and compared to related studies on consonant cues (e.g., Li et al., 2010; Li et al., 2012; Christiansen et al., 2007). Furthermore, the model framework presented in Chapter 3, which connects the perceptual and the acoustic domain, may serve as a tool to gain a better understanding of the contribution of specific signal characteristics to robust phoneme recognition.

The model framework has – despite its large predictive power in the considered conditions – several systematic limitations. For example, the model could not account for the large perceptual differences across NH listeners shown in Chapter 2, as they did not show any audiometric differences. These listener-specific effects may be attributable to individual biases or supra-threshold processing deficits that were not captured by the audiometric test. Furthermore, even if there were audiometric differences, the model could not account for them since audibility thresholds were not included so far. Thus, the model is also bound to fail in the case of partly or fully inaudible stimuli due to low presentation levels or hearing impairment. This could be overcome by adding threshold-simulating noise (cf. Jürgens and Brand, 2009; Jürgens et al., 2014) or by excluding the frequency bands below threshold from further processing (cf. Jørgensen and Dau, 2011). Finally, the model showed a tendency to slightly underestimate the extent of the measured consonant confusions and partly predicted additional confusions that were not reflected in the perceptual data. This resulted most likely from a bias induced by similarities/dissimilarities in the vowel portions of the CV/VCV stimuli and templates, which are only partly related to the consonant percept.

5.4 Perspectives

The experimental investigation of sources of perceptual variability (Chapter 2) was conducted with NH listeners using CVs presented in additive white noise. However, the applied methodology may easily be adapted to quantify the effects of other influencing factors. For example, effects of different accompanying vowels, other types of nonsense syllables (e.g., VCVs, CVCs), and other acoustic conditions (e.g., fluctuating additive noise, speech interferers, reverberation, hearing-instrument processing) could be considered. Furthermore, the perceptual-distance based experimental method is also suitable for analyzing vowel perception. DiNino et al. (2016) used the perceptual distance along with other analysis methods to evaluate effects of CI processing on consonant and vowel perception and demonstrated that it was more informative than the sequential information analysis (SINFA) proposed by Wang and Bilger (1973). Apart from the above mentioned stimulus-related ("source-related") factors, it could be insightful to also focus on receiver-related effects in terms of individual hearing impairment. To that end, an experimental investigation with HI listeners that uses the same stimuli as the study described in Chapter 2 may reveal differences between the contributions of the considered factors to consonant perception in HI vs NH listeners. This could provide insights with respect to the main problems HI listeners face in terms of speech perception and how to compensate for them.

Regarding the model framework, many further investigations and applications are conceivable. For example, it should be clarified whether the predictive power of the model extends to acoustic conditions that are more realistic than stationary noise (e.g., fluctuating noise, speech interferers, reverberation). Furthermore, the model has only been tested on consonant perception data but could, in principle, also be used to predict vowel perception data. This could be useful particularly with respect to evaluating CI processing strategies, as vowel perception tests are considered more informative than consonant perception tests in this field.

Consonant perception measurements have been shown to be particularly insightful when used as a tool to identify specific problems experienced by HI listeners with hearing aids (Scheidiger and Allen, 2013; Schmitt et al., 2016) or without hearing aids (Phatak et al., 2009; Trevino and Allen, 2013). A crucial next step would therefore be to incorporate aspects of hearing impairment, such as elevated audiometric thresholds, reduced frequency selectivity, loss of compression and other supra-threshold deficits (cf. Jürgens et al., 2014; Jepsen et al., 2014). If a version of the model that can account for consonant perception in *unaided* HI listeners were established, the results from Chapter 4 suggest that the effects of hearing-instrument compensation strategies might be well accounted for by the model. Such a "hearing-impaired" model could be informative for HA fitting, for example by suggesting specific fitting parameters based on a listener's auditory profile. In addition, such a model may be able to predict long-term effects of acclimatization for HA or CI processing strategies by using processed instead of unprocessed templates, representing the listener' adaptation to the provided processing.

The proposed model predicts consonant perception using a-priori information about the stimuli to predict the data. In contrast, the listeners do not need to know the exact signal to perceive a given consonant. In modeling terms, this would correspond to a much more elegant "blind" modeling approach that does not presume any a-priori information. While NH listeners can deal remarkably well with the large natural variability of speech utterances and robustly assign them to specific phonetic categories, blind automatic speech recognizers exhibit a much poorer performance, particularly in the presence of acoustical interference (cf. Meyer et al., 2011). However, as automatic speech recognition approaches are currently improving rapidly using HMMs in combination with deep neural networks (e.g. Hinton et al., 2012; Dahl et al., 2012), the gap between human and machine speech recognition seems to be decreasing substantially. When (and if) blind automatic speech recognizers reach human recognition performance, a perceptually adapted blind model may become feasible, using an auditory-inspired processing model as a front end and a state-of-the-art speech recognition system as a back end.

Bibliography

ANSI (1969). *American National Standard Methods for the Calculation of the Articulation Index.* ANSI S3.5. Acoustical Society of America, New York.

ANSI (1997). *Methods for the Calculation of the Speech Intelligibility Index.* ANSI S3.5. Acoustical Society of America, New York.

Allen, J. B. (1994). “How do humans process and recognize speech?” In: *Speech and Audio Processing, IEEE Transactions on* 2.4, pp. 567–577.

Allen, J. B. (2005). “Consonant recognition and the articulation index”. In: *The Journal of the Acoustical Society of America* 117.4, p. 2212.

Bashford, J. A., K. R. Riener, and R. M. Warren (1992). “Increasing the intelligibility of speech through multiple phonemic restorations”. In: *Perception & Psychophysics* 51.3, pp. 211–217.

Benzeghiba, M. et al. (2007). “Automatic speech recognition and speech variability: A review”. In: *Speech Communication* 49.10-11, pp. 763–786.

Bierer, J. A. (2010). “Probing the electrode-neuron interface with focused cochlear implant stimulation”. In: *Trends in amplification* 14.2, pp. 84–95.

Christiansen, T. U. and P. J. Henrichsen (2011). “Objective evaluation of consonant-vowel pairs produced by native speakers of Danish”. In: *Forum Acusticum 2011*.

Christiansen, T. U., T. Dau, and S. Greenberg (2007). “Spectro-temporal processing of speech–An information-theoretic framework”. In: *Hearing–From sensory processing*… Pp. 517–523.

Cooke, M. (2006). “A glimpsing model of speech perception in noise”. In: *The Journal of the Acoustical Society of America* 119.3, pp. 1562–1573.

Cooke, M. (2009). “Discovering consistent word confusions in noise”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1887–1890.

Cutler, A., A. Weber, R. Smits, and N. Cooper (2004). “Patterns of English phoneme confusions by native and non-native listeners”. In: *The Journal of the Acoustical Society of America* 116.6, pp. 3668–3678.

Dahl, G. E., D. Yu, L. Deng, and A. Acero (2012). “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition”. In: *IEEE Transactions on Audio, Speech and Language Processing* 20.1, pp. 30–42.

Dau, T., D. Püschel, and A. Kohlrausch (1996). “A quantitative model of the "effective" signal processing in the auditory system. I. Model structure.” In: *The Journal of the Acoustical Society of America* 99.6, pp. 3615–3622.

Dau, T., B. Kollmeier, and A. Kohlrausch (1997). “Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers”. In: *The Journal of the Acoustical Society of America* 102.5, pp. 2892–2905.

DiNino, M., R. A. Wright, M. B. Winn, and J. A. Bierer (2016). “Vowel and consonant confusions from spectrally-manipulated stimuli designed to simulate poor cochlear implant electrode-neuron interfaces”. In: *The Journal of the Acoustical Society of America (under review)*.

Fletcher, H. and R. Galt (1950). “The perception of speech and its relation to telephony”. In: *The Journal of the Acoustical Society of…* 22.2, pp. 89–151.

French, N. and J. Steinberg (1947). “Factors governing the intelligibility of speech sounds”. In: *The journal of the Acoustical society of…* 19.1, pp. 90–119.

Gallun, F. and P. Souza (2008). “Exploring the role of the modulation spectrum in phoneme recognition.” In: *Ear and hearing* 29.5, pp. 800–813.

Glista, D., S. Scollie, M. Bagatto, R. Seewald, V. Parsa, and A. Johnson (2009). “Evaluation of nonlinear frequency compression: clinical outcomes”. In: *International Journal of Audiology* 48.9, pp. 632–644.

Hagerman, B (1982). “Sentences for testing speech intelligibility in noise”. In: *Scandinavian audiology* 11, pp. 79–87.

Hinton, G. et al. (2012). “Deep Neural Networks for Acoustic Modeling in Speech Recognition”. In: *IEEE Signal Processing Magazine*, pp. 82–97.

Holube, I. and B. Kollmeier (1996). “Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model.” In: *The Journal of the Acoustical Society of America* 100.3, pp. 1703–1716.

Houtgast, T., H. Steeneken, and R. Plomp (1980). “Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics”. In: *Acta Acustica united with Acustica* 46.1, pp. 60–72.

Jepsen, M. L., S. D. Ewert, and T. Dau (2008). “A computational model of human auditory signal processing and perception”. In: *Journal of the Acoustical Society of America* 124.1, pp. 422–438.

Jepsen, M. L., T. Dau, and O. Ghitza (2014). “Refining a model of hearing impairment using speech psychophysics”. In: *The Journal of the Acoustical Society of America* 135.4, EL179–EL185.

Jørgensen, S. and T. Dau (2011). “Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing”. In: *The Journal of the Acoustical Society of America* 130.3, pp. 1475–1487.

Jørgensen, S., S. D. Ewert, and T. Dau (2013). “A multi-resolution envelope-power based model for speech intelligibility.” In: *The Journal of the Acoustical Society of America* 134.1, pp. 436–446.

Jürgens, T. and T. Brand (2009). “Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model”. In: *The Journal of the Acoustical Society of America* 126.5, pp. 2635–2648.

Jürgens, T., S. D. Ewert, B. Kollmeier, and T. Brand (2014). “Prediction of consonant recognition in quiet for listeners with normal and impaired hearing using an auditory model”. In: *The Journal of the Acoustical Society of America* 135.3, pp. 1506–1517.

Kashino, M. (2006). “Phonemic restoration: The brain creates missing speech sounds”. In: *Acoustical Science and Technology* 27.6, pp. 318–321.

Kimlinger, C., R. McCreery, and D. Lewis (2015). “High-Frequency Audibility: The Effects of Audiometric Configuration, Stimulus Type, and Device”. In: *J. Am. Acad. Audiol.* 26.2, pp. 128–137.

Kohlrausch, A. and D. Püschel (1988). “Interrelations between a psychoacoustical model of temporal deficits in hearing and neurophysiological observations”. In: *Sense Organs*. Ed. by N. Elsner and F. G. Barth. Thieme, Stuttgart, p. 39.

Kohlrausch, A., D. Püschel, and H. Alphei (1992). “Temporal resolution and modulation analysis in models of the auditory system”. In: *The Auditory Processing of Speech*. Ed. by M. E. H. Schouten. Mouton de Gruyter, Berlin, New York, pp. 85–98.

Li, F., A. Menon, and J. B. Allen (2010). “A psychoacoustic method to find the perceptual cues of stop consonants in natural speech”. In: *The Journal of the Acoustical Society of America* 127.4, pp. 2599–2610.

Li, F., A. Trevino, A. Menon, and J. B. Allen (2012). “A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise”. In: *The Journal of the Acoustical Society of America* 132.4, pp. 2663–2675.

Litvak, L. M., A. J. Spahr, A. A. Saoji, and G. Y. Fridman (2007). “Relationship between perception of spectral ripple and speech recognition in cochlear implant and vocoder listeners”. In: *The Journal of the Acoustical Society of America* 122.2, pp. 982–991.

Lobdell, B. E. and J. B. Allen (2007). “A model of the VU (volume-unit) meter, with speech applications”. In: *The Journal of the Acoustical Society of America* 121.1, pp. 279–285.

Mesgarani, N., S. V. David, J. B. Fritz, and S. A. Shamma (2014). “Mechanisms of noise robust representation of speech in primary auditory cortex”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111, pp. 6792 –6797.

Messing, D. P., L. Delhorne, E. Bruckert, L. D. Braida, and O. Ghitza (2009). “A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise”. In: *Speech Communication* 51, pp. 668–683.

Meyer, B. T., T. Brand, and B. Kollmeier (2011). “Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes”. In: *The Journal of the Acoustical Society of America* 129.1, pp. 388–403.

Miller, G. A. and J. C. R. Licklider (1950). “The Intelligibility of Interrupted Speech”. In: *The Journal of the Acoustical Society of America* 22.2, pp. 167–173.

Miller, G. A. and P. E. Nicely (1955). “An analysis of perceptual confusions among some English consonants”. In: *The Journal of the Acoustical Society of America* 27.2, pp. 338–352.

Moore, B. C. J. (2003). “Temporal integration and context effects in hearing”. In: *Journal of Phonetics* 31, pp. 563–574.

Mullennix, J. W., D. B. Pisoni, and C. S. Martin (1989). “Some effects of talker variability on spoken word recognition.” In: *The Journal of the Acoustical Society of America* 85.1, pp. 365–378.

Nielsen, J. B. and T. Dau (2009). “Development of a Danish speech intelligibility test”. In: *International journal of audiology* 48.10, pp. 729–741.

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

Nilsson, M., D. Soli, Sigfrid, and J. A. Sullivan (1994). “Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise”. In: *The Journal of the Acoustical* ... 95.June 1993, pp. 1085–1099.

Pavlovic, C. V. (1987). “Derivation of primary parameters and procedures for use in speech intelligibility predictions”. In: *The Journal of the Acoustical Society of America* 82.2, pp. 413–422.

Payton, K. L. and L. D. Braida (1999). “A method to determine the speech transmission index from speech waveforms”. In: *The Journal of the Acoustical Society of America* 106.6, pp. 3637–3648.

Phatak, S. A. and J. B. Allen (2007). “Consonant and vowel confusions in speech-weighted noise”. In: *The Journal of the Acoustical Society of America* 121.4, pp. 2312–2326.

Phatak, S. A., A. Lovitt, and J. B. Allen (2008). “Consonant confusions in white noise”. In: *The Journal of the Acoustical Society of America* 124.2, pp. 1220–1233.

Phatak, S. A., Y.-S. Yoon, D. M. Gooler, and J. B. Allen (2009). “Consonant recognition loss in hearing impaired listeners.” In: *The Journal of the Acoustical Society of America* 126.5, pp. 2683–2694.

Rhebergen, K. S., N. J. Versfeld, and W. a. Dreschler (2006). “Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise”. In: *The Journal of the Acoustical Society of America* 120.6, pp. 3988–3997.

Sakoe, H. and S. Chiba (1978). “Dynamic programming algorithm optimization for spoken word recognition”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-26.1, pp. 43–49.

Scheidiger, C. and J. B. Allen (2013). “Effects of NALR on consonant-vowel perception”. In: *4th International Symposium on Auditory and Audiological Research (ISAAR 2013), Nyborg, Denmark*, pp. 1–8.

Schmitt, N., A. Winkler, M. Boretzki, and I. Holube (2016). “A Phoneme Perception Test Method for High-Frequency Hearing Aid Fitting”. In: *J. Am. Acad. Audiol.* 27.5, pp. 367–379.

Simpson, A., A. A. Hersbach, and H. J. McDermott (2005). “Improvements in speech perception with an experimental nonlinear frequency compression hearing device”. In: *International journal of audiology* 44.5, pp. 281–292.

Singh, R. and J. B. Allen (2012). “The influence of stop consonants’ perceptual features on the Articulation Index model”. In: *The Journal of the Acoustical Society of America* 131.4, pp. 3051–3068.

Sroka, J. J. and L. D. Braida (2005). “Human and machine consonant recognition”. In: *Speech Communication* 45, pp. 401–423.

Stickney, G. S., P. C. Loizou, L. N. Mishra, P. F. Assmann, R. V. Shannon, and J. M. Opie (2006). “Effects of electrode design and configuration on channel interactions”. In: *Hearing Research* 211, pp. 33–45.

Toscano, J. C. and J. B. Allen (2014). “Across- and Within-Consonant Errors for Isolated Syllables in Noise”. In: *Journal of Speech, Language, and Hearing Research* 57, pp. 2293–2307.

Tóth, M. A., M. L. García Lecumberri, Y. Tang, and M. Cooke (2015). “A corpus of noise-induced word misperceptions for Spanish”. In: *The Journal of the Acoustical Society of America* 137.2, EL184–EL189.

Trevino, A. and J. B. Allen (2013). “Within-consonant perceptual differences in the hearing impaired ear”. In: *The Journal of the Acoustical Society of America* 134.1, pp. 607–617.

Wagener, K., J. L. Josvassen, and R. Ardenkjaer (2003). “Design, optimization and evaluation of a Danish sentence test in noise”. In: *International Journal of Audiology* 42.1, pp. 10–17.

Wang, M. D. and R. C. Bilger (1973). “Consonant confusions in noise: a study of perceptual features”. In: *The Journal of the Acoustical Society of America* 54.5, pp. 1248–1266.

Warren, R. M. (1970). *Perceptual restoration of missing speech sounds.*

White, M. W., M. M. Merzenich, and J. N. Gardi (1984). “Multichannel cochlear implants”. In: *Archives of Otolaryngology* 110.8, pp. 493–501.

Wolfe, J. et al. (2011). “Long-term effects of non-linear frequency compression for children with moderate hearing loss”. In: *International Journal of Audiology* 50.6, pp. 396–404.

Zaar, J. and T. Dau (2015). “Sources of variability in consonant perception of normal-hearing listeners”. In: *Journal of the Acoustical Society of America* 138.3, pp. 1253–1267.

Zaar, J. and T. Dau (2016). “Predicting consonant recognition and confusions in normal-hearing listeners”. In: *The Journal of the Acoustical Society of America (under review).*

Zaar, J., N. Schmitt, R.-P. Derleth, M. DiNino, J. A. Bierer, and T. Dau (2016). "Predicting effects of hearing-instrument signal processing on consonant perception". in preparation.

Contributions to Hearing Research

Vol. 1: *Gilles Pigasse,* Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.

Vol. 2: *Olaf Strelcyk,* Peripheral auditory processing and speech reception in impaired hearing, 2009.

Vol. 3: *Eric R. Thompson,* Characterizing binaural processing of amplitude-modulated sounds, 2009.

Vol. 4: *Tobias Piechowiak,* Spectro-temporal analysis of complex sounds in the human auditory system, 2009.

Vol. 5: *Jens Bo Nielsen,* Assessment of speech intelligibility in background noise and reverberation, 2009.

Vol. 6: *Helen Connor,* Hearing aid amplification at soft input levels, 2010.

Vol. 7: *Morten Løve Jepsen,* Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.

Vol. 8: *Sarah Verhulst,* Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.

Vol. 9: *Sylvain Favrot,* A loudspeaker-based room auralization system for auditory research, 2010.

Vol. 10: *Sébastien Santurette,* Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.

Vol. 11: *Iris Arweiler,* Processing of spatial sounds in the impaired auditory system, 2011.

Vol. 12: *Filip Munch Rønne,* Modeling auditory evoked potentials to complex stimuli, 2012.

Vol. 13: *Claus Forup Corlin Jespersgaard*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.

Vol. 14: *Rémi Decorsière*, Spectrogram inversion and potential applications for hearing research, 2013.

Vol. 15: *Søren Jørgensen*, Modeling speech intelligibility based on the signal-to-noise envelope power ration, 2014.

Vol. 16: *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.

Vol. 17: *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.

Vol. 18: *Márton Marschall*, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.

Vol. 19: *Jasmina Catic*, Human sound externalization in reverberant environments, 2014.

Vol. 20: *Michał Feręczkowski*, Design and evaluation of individualized hearing-aid signal processing and fitting, 2015.

Vol. 21: *Alexandre Chabot-Leclerc*, Computational modeling of speech intelligibility in adverse conditions, 2015.

Vol. 22: *Federica Bianchi*, Complex-tone pitch representations in the human auditory system, 2016.

Vol. 23: *Johannes Zaar*, Measures and computational models of microscopic speech perception, 2016.

The end.

To be continued…

The human auditory system is well-adapted to extracting target speech sounds in adverse acoustic conditions. Furthermore, high-level cognitive processing allows us to make sense of what we hear, even if the acoustic information is severely degraded or sparse. Therefore, commonly used *macroscopic* speech intelligibility tests that typically use sentences as stimuli measure a combination of (i) effects of the salience of the perceived speech and (ii) effects related to linguistic processing (e.g., using context information). This thesis presents an alternative approach reflecting a *microscopic* measure of speech perception that is solely related to the salience of the perceived speech. Here, the perception of individual consonants is investigated using nonsense syllables like /ta, ba/ as stimuli and the responses are evaluated in terms of consonant recognition and confusions. This approach allows to investigate the effects of acoustical transmission channels (e.g., rooms, mobile phones), as well as effects of hearing impairment and hearing-instrument signal processing on the fundamental speech sounds. The factors that are perceptually relevant for consonant-in-noise perception are analyzed, such as differences in the stimuli and differences in the normal-hearing listeners. Moreover, a computational model of microscopic speech perception is proposed that consists of a model of the auditory periphery and a template-matching based decision stage. The model is shown to account well for effects of masking noise as well as effects of hearing-instrument signal processing on consonant recognition and confusions.
The experimental results of this thesis have implications for the design of consonant perception experiments. Furthermore, the proposed model framework could be useful for the evaluation of hearing-instrument processing strategies, particularly when combined with simulations of individual hearing impairment.

DTU Electrical Engineering
Department of Electrical Engineering

Ørsteds Plads
Building 348
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 38 00
Fax: (+45) 45 93 16 34
www.elektro.dtu.dk