On Speech Intelligibility in Tonal and Non-Tonal Languages

Von der Fakultät für Medizin und Gesundheitswissenschaften der Carl von Ossietzky Universität Oldenburg zur Erlangung des Grades und Titels

Doctor of Philosophy (Ph.D.)

angenommene Dissertation von

Maximilian Karl Scharf

geboren am 27.05.1994 in Düsseldorf

08.07.2026

Gutachter:

Prof. Dr. Dr. Birger Kollmeier

Prof. Dr. Damir Kovačić

Dr. Anna Warzybok-Oetjen

Prof. Dr. Steven van de Par

Tag der Disputation: 08.07.2026

Imagine that you are explaining
your ideas to your former smart,
but ignorant self, at the
beginning of your studies!

Richard P. Feynman, *Lectures on Computation*

Abstract

The central objective of this thesis is a model-understanding of empirical, normal hearing speech recognition thresholds (SRTs) in Mandarin Chinese, Cantonese, and English for plain and Lombard speech.

State-of-the-art speech intelligibility (SI) models reach high predictive power for SRTs of normal-hearing listeners with prediction errors in the range of the measurement uncertainty. Thus, special emphasis was laid on measurement uncertainties and their implications for interpreting both empirical data and comparing model predictions. This required a thorough understanding of the entire measurement process, from calibration to the estimation of the threshold (i.e., a point on the psychometric function) with adaptive procedures.

This book consists of an introduction to the methods of this thesis in chapter 1, followed by five main chapters. Chapter 2 describes a procedure to estimate the calibration of a microphone with resonating bottles, followed by chapter 3, where a model of consistency for psychometric tracks is introduced. These chapters provide the foundation for the following analysis of the influence of measurement uncertainties on the informative value of SI models.

The following three chapters are concerned with the SI prediction of plain and Lombard speech in Mandarin Chinese, Cantonese, and English. Chapter 4 provides evidence that the band importance function (BIF) for the extension of the speech intelligibility index (SII) to Mandarin Chinese in background noise does not generalize between comparable speech material and that generic BIFs lead to equivalently good predictions compared to published gender-specific BIFs in Mandarin.

Chapter 5 shows for the Lombard gain (LG), i.e., the improvement in SRT between plain and Lombard speech, in Mandarin Chinese, that spectral information is sufficient to predict the empirical data to within the measurement uncertainty. This was done with two different modeling approaches, i.e., the framework of auditory discrimination experiments (FADE) and the SII. The predictions of the absolute SRT, however, require spectro-temporal information, only accessible by FADE.

The last chapter 6 is a discussion similar to chapter 5 for the Lombard effect of bilingual speakers in Cantonese and English with a smaller dataset. This chapter confirms the findings of chapter 5. For all languages considered, spectral changes in the speech features fully explain the LG, while spectro-temporal information is necessary for absolute SRT predictions.

Zusammenfassung

Das Ziel dieser Arbeit ist ein modellbasiertes Verständnis empirischer Sprachverständlichkeitsschwellen (SRTs) bei normalhörenden Versuchspersonen in Mandarin-Chinesisch, Kantonesisch und Englisch bei Normal- und Lombard-Aussprache.

Moderne Sprachverständlichkeits- (SI-) Modelle erreichen für SRTs Normalhörender eine hohe Vorhersagegenauigkeit, mit Fehlern im Bereich der Messunsicherheit. Daher wurde besonderer Wert auf Messunsicherheiten und deren Implikationen für die Interpretation empirischer Daten sowie den Vergleich von Modellvorhersagen gelegt. Dies erforderte ein umfassendes Verständnis des gesamten Messprozesses, von der Kalibrierung bis zur Schätzung der Schwelle (d.h. eines Punktes auf der psychometrischen Funktion) mittels adaptiver Verfahren.

Dieses Buch besteht aus einer Einführung in die Methoden dieser Arbeit in Kapitel 1, gefolgt von fünf Hauptkapiteln. Kapitel 2 beschreibt ein Verfahren zur Abschätzung der Kalibrierung eines Mikrofons mit resonierenden Flaschen, gefolgt von Kapitel 3, in dem ein Konsistenzmodell für psychometrische Messreihen eingeführt wird. Diese Kapitel bilden die Grundlage für die folgende Diskussion der SI-Vorhersagen.

Die anschließenden drei Kapitel befassen sich mit der SI-Vorhersage von Normaler und Lombard-Aussprache in Mandarin-Chinesisch, Kantonesisch und Englisch. Kapitel 4 diskutiert ein Beispiel dafür, dass die Bandgewichtungsfunktionen (BIFs) für die Erweiterung des "Speech Intelligibility Index" (SII) auf Mandarin im Störgeräusch nicht auf vergleichbares Sprachmaterial verallgemeinert werden können. Ferner wird gezeigt, dass generische BIFs ebenso gute Vorhersagen liefern wie veröffentlichte geschlechtsspezifische BIFs für Mandarin.

Kapitel 5 zeigt, dass spektrale Information ausreicht, um den Lombard-Schwellengewinn (LG), d.h. die Verbesserung der SRTs zwischen Normal- und Lombard-Aussprache, für Mandarin-Chinesisch innerhalb der Messunsicherheit vorherzusagen. Dies wurde mithilfe zweier unterschiedlicher SI-Modelle, nämlich des "Framework of Auditory Discrimination Experiments" (FADE) und des SII, gezeigt. Vorhersagen der absoluten SRTs erfordern jedoch spektral-zeitliche Information, die von den beiden Modellen nur mit FADE zugänglich ist.

Das letzte Kapitel 6 diskutiert, analog zu Kapitel 5, den Lombard-Effekt zweisprachiger Sprecherinnen und Sprecher in Kantonesisch und Englisch auf Basis eines kleineren Datensatzes von SRTs. Dieses Kapitel bestätigt die Ergebnisse aus Kapitel 5. Für alle untersuchten Sprachen erklären spektrale Änderungen den LG vollständig, während für absolute SRT-Vorhersagen spektral-zeitliche Information erforderlich ist.

Contents

List of Figures

List of Tables

Chapter 1:

Introduction

In 1911, the French physician Étienne Lombard (1869–1920) made a groundbreaking discovery: speech production changes in the presence of background noise (Lombard 1911). More specifically, when placing a person in audible background noise, he or she automatically raises the voice when speaking. This effect is known as the Lombard effect and is central to this book. 46 years later, Dreher and O'Neill 1957 reported that Lombard speech requires a lower signal to noise ratio (SNR) than plain, unaltered speech for the same speech intelligibility (SI), which is called Lombard gain (LG).

The question of which changes in Lombard speech are mainly responsible for the LG has been largely unexplored. Reason for this is that we can not simply remove relevant speech features in hearing experiments, and that SI models did not reach the necessary predictive power. However, with the framework of auditory discrimination experiments (FADE) of Schädler, Warzybok, Hochmuth, et al. 2015 it became possible to predict empirical speech recognition thresholds (SRTs) with high accuracy and control the speech features that are used for the prediction. The precision of this SI model required a rigorous treatment of the empirical measurement uncertainties, which, in the past, has never been necessary.

This book traces the sources of these uncertainties in chapters 2 and 3. Chapters 5 and 6 apply a statistical analysis method, standard in the field of physics, to the effect of measurement uncertainties on the informative value of SI models for the LG in two tonal and one non-tonal language.

Scientific interest in quantitative SI started with the invention of telecommunication devices. Nowadays, we take realistic, high-fidelity[1] audio recording, storage, transmission, and reproduction for granted. However, the challenges that engineers had to overcome were numerous and essentially an optimization problem. Initially, this optimization was approached through trial and error, aided by SI tests, where the circuit, which was to be optimized, processed speech (Castner and Carter 1933).

As with many technological advances, rapid improvements in telecommunication were highly beneficial for military applications during the last world war. For this purpose, American scientists at the Bell Laboratories and Harvard University designed several SI tests in the years from 1919-1945 (Fletcher and Galt 1950).

With these early SI tests, it became evident that measurements may be inconsistent if the experimenter does not appropriately control the tests. It was found that responses may be inconsistent if there are sudden changes in the circuit's performance or if participants become tired during the experiment. A second problem has been to control the level of the target signal, i.e., speech. The second problem was solved by a visual feedback loop of the level to the speaker while speaking (Castner and Carter 1933). It required adequate calibration of the recording microphone in a controlled laboratory environment.

The problem of calibration persists today with handheld devices outside the laboratory and is generally unsolved for mobile health (mHealth) products on uncontrolled hardware.

[1] High fidelity, or short "HIFI", is a standardized description of audio equipment in the European Union. Since 1996, there have been no minimal requirements for this predicate (DIN EN 61305-1:1996-02); only the complete list of specifications, such as transmission range and distortion factor, has to be published.

Chapter 2 presents an approximation to a calibration of microphones that is exact enough for meaningful audiological measurements.

In the past, the standard approach to address the problem of inconsistent responses to SI tests has been to optimize the measurement procedures. These procedures were made robust against inconsistent response behavior to some degree. However, the issue of distinguishing between inconsistent and consistent measurement results has never been fully resolved. Chapter 3 proposes the first model-understanding of this process and a way to quantify the inconsistency of a measurement.

With reliable SI tests in the 1950s, there was then a strong interest in inventing a model-based understanding of human speech recognition that could replace slow empirical measurements when optimizing audio devices (Fletcher and Galt 1950). Derivatives of these early SI models, such as the speech intelligibility index (SII), are still used for automatic optimization of audio processing algorithms. However, as discussed in chapter 4, not all parameterizations of this model generalize between comparable speech tests.

Since 1945, the focus of the hearing sciences has shifted to fundamental research and medical applications[2]. Although the medical aspect of hearing science was influenced by wars too[3], hearing science of today is focused on reconnecting people with people.

The deaf-blind writer Helen Keller aptly expressed the importance of this effort with her famous quote: *"Blindness separates people from things; deafness separates people from people"*. As chapters 5 and 6 explore the Lombard gain (LG) in two mutually unintelligible representatives of non-tonal (i.e., English) and tonal (i.e., Mandarin or Cantonese) languages, my hope is also to connect these peoples.

[2]In fact, Alexander Graham Bell(1847-1922), a teacher for deaf children, who patented the first telephone in the USA, had his primary interest in how to help hearing-impaired people to communicate with normal-hearing people (Bruce 1990). However, in their article about the Lombard Gain, Dreher and O'Neill 1957 still highlighted the usefulness of the LG for military applications.

[3]Lombard had been researching the damage of air blasts on pilots' hearing during the First World War shortly before his death (Lermoyez 1921).

1.1 Definition of Key Terms

First, we have to define the ability of a human (referred to as the test subject or listener) to understand speech during a listening experiment. To make it a measurable quantity, we define the speech recognition threshold for 50 % correctly understood words (SRT_{50}) as the signal to noise ratio (SNR) of the average power of a signal with speech P_S and noise P_N at which a listener understands 50% of words correctly. The speech signal is also referred to as the target, while the noise is referred to as the masker. In signal processing, the power P of a scalar signal $x(t)$ of length τ and a constant c is defined as:

$$P = \frac{1}{\tau}\int_0^\tau c \cdot x^2(t)dt \tag{1.1}$$

For the physical property, the power of sound waves with units of energy per second, the calculation is more complex. Still, it simplifies to the same scaling factor c after a few approximations that are valid for everyday speech communication. The power of a pressure wave $p(t)$ together with the resulting volume flow $Q(t) = dV/dt$ is defined as:

$$P = \frac{1}{\tau}\int_0^\tau p(t) \cdot Q(t)dt \tag{1.2}$$

This is because the energy of a sound wave can do mechanical work[4] W by integration of the pressure p with the displaced volume V:

$$W = \int pdV \tag{1.3}$$

$$= \int p \cdot Qdt \tag{1.4}$$

The relation between volume flow Q and pressure p can be approximated by a linear time invariant system (LTI):

$$\mathscr{F}[Q](\omega) = Y(\omega) \cdot \mathscr{F}[p](\omega) \tag{1.5}$$

$$Q(t) = \mathscr{F}^{-1}[Y](t) * p(t) \tag{1.6}$$

$$= G(t) * p(t) \tag{1.7}$$

Where $\mathscr{F}[\dots]$ is the Fourier transformation[5] of the argument in brackets $[\dots]$, Y is called acoustic admittance (=1/impedance) in full analogy to electrical circuit theory, $*$ is the convolution operator, and $G(t)$ is the acoustic conductance, which is a function of time. This way, the power of a pressure wave can be expressed only as a function of $p(t)$:

$$P = \frac{1}{\tau}\int_0^\tau p(t) \cdot (G(t) * p(t))\, dt \tag{1.8}$$

If $Y(\omega)$ is constant for all frequencies ω, which is a good approximation for plane waves of the acoustic signals of speech communication, $G(t)$ is a Dirac δ-function multiplied by a scalar G, and the convolution turns into a multiplication by G.

$$\boxed{P \approx \frac{1}{\tau}\int_0^\tau G \cdot p^2(t) \cdot dt} \tag{1.9}$$

We can express power on a logarithmic decibel (dB) scale and refer to it as level L:

$$L = 10 \cdot \log_{10}\left(\frac{P}{P_0}\right) \tag{1.10}$$

Where P_0 is a divisor to make the argument of the logarithm dimensionless. For acoustical signals, $P_0 = G \cdot (20\mu\mathrm{Pa})^2$. To distinguish levels with different divisors P_0, the level

[4] Assuming an adiabatic process, which is a good approximation for the frequencies of the sound waves considered here (Pierce 2019, Sec.1.10).

[5] Depending on if the unitary Fourier transform ($\mathscr{F} = \mathscr{F}^{-1}$) is used, definitions may differ in factors of 2π.

of sound waves with $P_0 = G \cdot (20\mu\text{Pa})^2$ is given the suffix [6] sound pressure level (SPL). A sound wave with a level of $L = 0$dB-SPL corresponds to the smallest pressure wave that is still audible for the average normal-hearing person (see Figure 1.4).

If the signal is digital, which means that it is a discretized, scalar signal, we compute the power according to Equation 1.1. P_0 is then defined as the square of the maximum amplitude. The level L in dB is then marked with the suffix full scale (FS). FS is commonly a power of two.

The name "decibel" was chosen to honor Alexander Graham Bell, and since 1928, it has been the de facto standard for describing the power of a signal (Morat and Ziemer 2018).

Because of the properties of the logarithm, we can rewrite Equation 1.10 into its other common form:

$$L/_{\text{dB}} = 20 \cdot \log_{10}\left(\sqrt{\frac{P}{P_0}}\right) \tag{1.11}$$

$$L/_{\text{dB-SPL}} = 20 \cdot \log_{10}\left(\frac{\sqrt{\frac{1}{\tau}\int_0^\tau p^2(t) \cdot dt}}{20\mu\text{Pa}}\right) \tag{1.12}$$

Expressing levels in dB is useful to compute SNRs elegantly. The SRT_{50} is defined as a SNR, so we can express it simply as the difference of the level of speech L_S and level of noise L_N in dB-SPL at the eardrum of the listener, at which he or she can correctly recognize 50% of speech:

$$\text{SRT}_{50} = L_{S,50\%} - L_{N,50\%} \tag{1.13}$$

The SRT_{50} depends both on L_S and L_N. The following distinction is made according to the level of the masking noise L_N:

- **Listening in quiet** means that L_N is below the hearing threshold of the listener.
- **Supra-threshold listening** means that L_N is audible to the listener.

In the following, all SRT_{50} measurements with speech tests are supra-threshold experiments with a constant level of the masking noise $L_N = 65$ dB-SPL at the eardrum. This convention means that the masker is audible to the normal-hearing listeners of the studies, and L_S is varied until 50% of speech is recognized correctly.

There exists another measure for speech in noise with the same abbreviation: the speech reception threshold. It is sometimes confused with the speech recognition threshold (SRT) because it is also a SNR and shares the same abbreviation. However, the speech reception threshold describes the SNR at which speech is identifiable as such, even though no word is necessarily recognizable and can be repeated by the listener.[7] For that reason, the term speech detection threshold (SDT) is sometimes used to avoid the ambiguous abbreviation. In the following, SRT will only refer to speech recognition thresholds.

[6] It is a suffix as, strictly speaking, levels in dB do not have actual units like angles and probabilities.

[7] This may seem complicated to remember, but if you think about automatic speech recognizer (ASR)-systems that you find as a service on most operating systems, it becomes clear that recognition includes processing the signal.

1.2 Psychometric Experiments

This book focuses on the prediction of speech intelligibility (SI)[8], which was measured with matrix speech tests. Many such psychometric tests exist, which all share the same assumption: a psychometric function (PF) models the response of a test subject to a stimulus.

1.2.1 The Psychometric Function (PF)

The psychometric function (PF) is a mathematical model of the response probability during a psychometric test. Psychometric tests are experiments that quantify the response r of a test subject (i.e., person or animal) to a stimulus s, such as an acoustic signal. An example of a PF is shown in Figure 1.1. Typically, PFs have a sigmoidal shape with an inflection point referred to as the detection threshold. For speech intelligibility (SI) the speech recognition threshold for 50 % correctly understood words (SRT_{50}) is often used (see Equation 1.13). In the following, this 50% threshold is only referred to as SRT without the subscript.

We can model the PF with four parameters (Doire, Brookes, and Naylor 2017):

$$\Phi(s) = \gamma + \frac{1-\gamma-\lambda}{1+\exp\left(-\beta(s-\alpha)\right)} \tag{1.14}$$

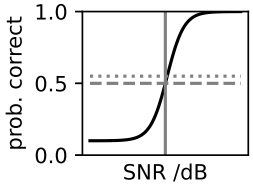


Figure 1.1: Example of a psychometric function (PF): The solid black PF connects the probability of a correct response to the SNR. In SI measurements, the vertical line corresponds to the SRT_{50}. The dashed horizontal line is the 50% threshold. The dotted horizontal line crosses the PF at the inflection point.

The lapse rate (λ) describes the probability of answering incorrectly at high SNR, which a lapse in attention could cause (therefore the name). The guess rate (γ) is the probability of answering correctly at the chance level. For the closed matrix test, the γ for the word-specific PF is 1/10 (see section 3.7.1). α and β control the threshold and slope, respectively.

Other thresholds, such as the SRT_{80} for 80% correctly identified items, may also be used. However, the advantage of defining the threshold at 50% is that for small γ and λ, the threshold is close to the inflection point of the PF (see Figure 1.1), where the slope is the highest, and accordingly the measurement uncertainty is lowest. This point is also referred to as the "sweetpoint" (Thomas Brand and Kollmeier 2002). The advantage of higher percentages of correctly identified items, i.e., SRT_{80} is that participants in listening experiments enjoy the experiments more and do not tire as quickly, which may lead to inconsistent response behavior.

1.2.2 Speechtests

Speech is essentially an encoding of information. And as there are many ways to encode the same information with more or less redundancy, every speech material has its own SRT, which also depends on the type of masker.

To resolve this ambiguity, several speech tests have been defined to make SRTs measurements comparable. However, one difficulty that even a well-defined speech test cannot solve is that humans are aware of how to outsmart the hearing part of the test. If the listener is well-informed about the words or phrases he or she has to expect during the experiment, the chances of guessing are improved. This changes the SRT. One prominent example of this effect is the Göttinger speech test (GÖSA), a speech test of grammatically and semantically correct sentences with 3-7 words (Kollmeier and Wesselkamp 1997). Test subjects can memorize the sentences of the GÖSA after a few repetitions of the same test list because of the predictability of the sentences (K. Wagener, Brand, and Kollmeier 1999) and perform better and better (lowering the SRT) with each test.

This problem has been partially solved with the matrix-sentence test, which is a test with 20 syntactically fixed, random five-word sentences, called test lists. Hagerman 1982 first described it for Swedish, which served as a blueprint for the Oldenburger matrix sentence test (OLSA)[9] in German (K. Wagener, Brand, and Kollmeier 1999) which has since been extended to other languages (Kollmeier, Warzybok, et al. 2015). The word corpus of the Mandarin matrix-sentence test is shown in Figure 1.2. A sentence is constructed by randomly selecting one of ten options each for name, verb, numeral, adjective, and noun. This

[8] Speech intelligibility and speech recognition are closely related concepts. Intelligibility is a property of a speech test. Recognition is the task of decoding the audio of a speech test to labels, i.e., text, or single words.

[9] K. Wagener, Brand, and Kollmeier 1999 included the coarticulation when test lists were generated, which was not done in Hagerman 1982. See section 1.2.2.

randomization has the effect that it is much harder to alter the SRT through memorization, even if we provide the listener with a table of all possible words as a reference. [10]

Listeners with no experience repeat the matrix-sentence test at least three times with different test lists from the same corpus of words. In the first two tests, listeners get familiar with the test and the possible options for each word, which also leads to a gradual improvement of the SRT. The results of these first two tests are discarded, as the SRT improvement is significantly larger than zero (K. Wagener, Brand, and Kollmeier 1999). For the following tests, the SRT improvement is smaller than the measurement uncertainty, as the random nature of the test makes it very difficult to memorize all sentences.

	Name a	Verb b	Numeral c	Adjective d	Noun e
0	郭毅	带走	一个	彩色的	板凳
	Guō Yì	DàiZǒu	Yī Gè	CǎiSè de	BǎnDèng
1	李锐	借来	两个	大号的	茶杯
	Lǐ Ruì	JièLái	Liǎng Gè	DàHào de	CháBei
2	沈悦	看见	三个	很旧的	灯笼
	Shěn Yuè	KànJiàn	Sān Gè	HěnJiù de	DēngLong
3	王石	留下	四个	便宜的	饭盒
	Wáng Shí	LiúXià	Sì Gè	PiányYi de	FànHé
4	徐敏	买回	五个	漂亮的	花瓶
	Xú Mǐn	MǎiHuí	Wǔ Gè	PiàoLiang de	HuāPíng
5	杨硕	拿起	六个	普通的	戒指
	Yáng Shuò	NáQǐ	Liù Gè	PǔTōng de	JièZhi
6	张伟	弄丢	七个	奇怪的	闹钟
	Zhāng Wěi	NòngDiū	Qī Gè	QíGuài de	NàoZhōng
7	郑贤	收好	八个	全新的	书包
	Zhèng Xián	ShōuHǎo	Bā Gè	QuánXīn de	ShūBāo
8	周明	需要	九个	特别的	水壶
	Zhōu Míng	XūYào	Jiǔ Gè	TèBié de	ShuǐHú
9	朱婷	找出	十个	用过的	玩具
	Zhū Tíng	ZhǎoChū	Shí Gè	YòngGuò de	WánJù

Figure 1.2: Corpus of the Mandarin matrix-sentence test:
adopted from Hu, Xi, et al. 2018.

There are two ways to construct a test list, depending on the required quality or homogeneity of the test. The homogeneity of a test has a direct impact on the measurement uncertainty of the test result.

Unoptimized Test Lists

If the sentences of the test list are read out with no postprocessing, the individual words of a sentence are not guaranteed to be equally understandable. A normal speaking style does not guarantee that each word in a sentence leads to the same SRT. This results in a larger variability of the word-specific SRTs, which can be described with a shallower PF at the sentence-specific SRT (Kollmeier 1990). A shallower slope generally leads to a larger measurement uncertainty of the SRT.

The uncertainty σ of a test can be reduced by taking the average SRT of N test subjects, assuming healthy, normal-hearing listeners:

$$\sigma_{\overline{\mathrm{SRT}}} = \frac{\sigma_{\mathrm{SRT}}}{\sqrt{N}} \tag{1.15}$$

[10] The effect of memorization may be reduced further with the Oldenburg phrase test (OLPHRA) of Ibelings et al. 2024, which contains about 700 sentences that do not contain the same words (excluding articles). However, this test still shows a training effect of 1.5dB.

This means that if the average SRT of the speaker for normal-hearing listeners is of interest, unoptimized tests are suitable, as the measurement uncertainty scales with $1/\sqrt{N}$. Furthermore, this approach removes proficiency effects of the listeners, which is important for the discussion in chapter 4. For chapters 5 and 6 on the Lombard effect in Cantonese and Mandarin Chinese, F. Chen, Pan, et al. 2025 recorded unoptimized matrix-sentence tests with male and female speakers. Thus, the measurement uncertainties of the resulting SRT measurements follow Equation 1.15.

Optimized Test Lists

For medical applications of the matrix test, small measurement uncertainties of the SRT are favorable, as no statistics over N test sessions with the same listener (see Equation 1.15) is available because of time constraints. This time constraint requires the test-retest variability to be as low as possible.

For the optimization, the test corpus is recorded such that all possible transitions between words are realized. Generally, the articulation (and thereby SRT) of a word is not independent of the preceding word. This effect is called coarticulation. The smallest number of sentences that realize all possible coarticulations of the matrix test does not depend on the length of the sentence but on the maximum number of transitions between sections of a sentence, e.g., subject transition to verb. For the matrix test with a fixed grammar and 10 options for each word, the maximum number of transitions between two sections is $10 \times 10 = 100$.[11] These recordings are then split into individual words and can be used to resynthesize any of the 10^5 combinations of the matrix sentence test. K. Wagener, Brand, and Kollmeier 1999 proposed this splitting method, which is not included in the Swedish matrix sentence test of Hagerman 1982.

This corpus of 500 coarticulated words is used for the test optimization. This optimization aims to change the level of each word such that words with a comparatively high word-specific SRT are amplified and words with a relatively low word-specific SRT are attenuated. This way, the intelligibility of the words is equalized, which leads to a steeper slope of the PF for the entire sentence (**kollmeier1990**; Kollmeier, Warzybok, et al. 2015) and thus smaller measurement uncertainty.

1.2.3 Consistency of Psychometric Test Results

The conclusions of this work critically depend on empirical SI measurements. Therefore, it is crucial to closely inspect the empirical SRT for measurement uncertainties. One source of uncertainty is the unsuccessful termination of the measurement procedure.

We can do psychometric tests either with fixed stimuli or according to an adaptive procedure. Fixed stimuli have the advantage that the PF is scanned at predefined values, and measurement uncertainties scale with the square root of the number of repetitions, like Equation 1.15. However, if the threshold or slope of the PF is of interest, predefined stimuli do not necessarily measure at the most optimal stimulus, i.e., sweetpoints according to Thomas Brand and Kollmeier 2002. Thus, predefined stimuli require a comparatively large number of responses for the same measurement uncertainty as adaptive procedures.

Adaptive procedures, on the other hand, require less time because the stimulus is changed adaptively according to the response of the test subject. The adaptive procedure aims to converge to the threshold of the PF[12]. However, this comes at the cost of a potentially failing experiment when the adaptive procedure does not converge. Failed convergence may be caused by an incorrect initialization of the adaptive procedure, i.e., the test subject responds "not heard" even though the stimulus was "heard", tricking the adaptive procedure into a search for the threshold at too high levels.

The check of consistency is an essential manual processing step in the analysis of empirical psychometric data. Manual consistency checks are problematic in the context of automated psychometric measurements with smartphone apps, like in the work of Schell-Majoor, Büchner, and Kollmeier 2023 or the quality screening of large databases for machine

[11] A matrix sentence test with 13 instead of 10 adjectives would require 130 recordings to capture all possible coarticulations. etc.

[12] The adaptive procedure should converge to the SRT_{50} if a precise measurement of the threshold is of interest. The SRT_{50} corresponds to the sweetpoint for a PF with zero λ and γ. For an optimal measurement of the slope, two sweetpoints are required (Thomas Brand and Kollmeier 2002).

learning approaches, like in Buhl 2022; Saak et al. 2022, because of timing or scalability, respectively. Chapter 3 proposes a model-based approach to this challenge.

The primary focus of chapter 3 is not to define a better adaptive procedure but to define a model understanding of inconsistency, which applies to the outcome of any measurement procedure. See Treutwein 1995; Leek 2001; Xu et al. 2024 for an introduction to current adaptive measurement procedures and alternatives that are more robust to certain kinds of inconsistent response behavior.

1.3 Models of Speech Intelligibility

SI models can be classified into two categories. The first, older category is for index-based models of speech intelligibility. These models produce a scalar intelligibility measure. Prominent examples are the speech intelligibility index (SII) (ANSIS3.5 1997), speech transmission index (STI) (Houtgast et al. 2002), and short time objective intelligibility measure (STOI) (Andersen et al. 2017). Central to all these models is that the scalar intelligibility measure is monotonously related to SI. This way, we can predict empirical measures of SI like the SRT with a so-called transfer function, or a reference condition (see section 1.3.5).

The second type of SI model is based on an automatic speech recognizer (ASR). The ASR is trained and tested with speech material for which the SI prediction is required (see Figure 1.3). Using the performance of an ASR as a predictor of SI has the advantage that no reference condition is needed if the encoded message is known. This way, the model can directly predict the SRT, but this comes at the cost of large and potentially slow models. Examples of this type of SI model are the model of Holube and Kollmeier 1996, the model of Barker and Cooke 2007, framework of auditory discrimination experiments (FADE) (Schädler, Warzybok, Hochmuth, et al. 2015), the model of Fontan et al. 2017,phone-based binaural intelligibility model (PHOBI) (Roßbach, Kirsten C. Wagener, and Meyer 2023), and the model of Saddler and McDermott 2024.

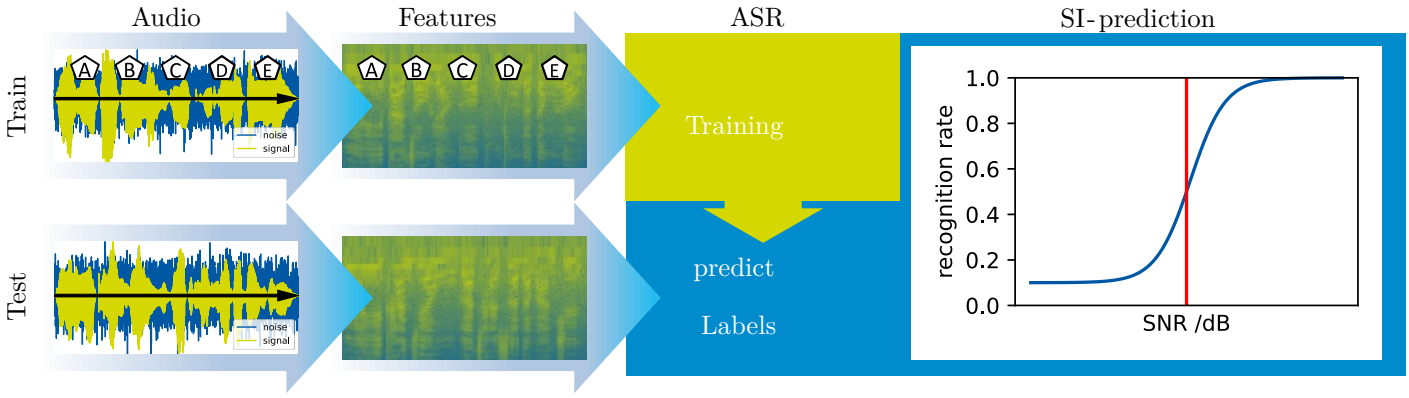


Figure 1.3: Concept of automatic speech recognizer (ASR)-based models of speech intelligibility:
Labelled training audio is converted to features and used to train an ASR model. In a second step, the speech audio for which the SRT is to be predicted is fed into the trained ASR to predict the labels. The number of correctly predicted labels results in the recognition rate as a function of the SNR of the testing audio. The SRT is the threshold indicated in red.

Central to all these models is the intermediate representation of the speech signal. This representation of data is called speech features. These speech features can be altered, like removing or distorting parts, to influence the prediction results. The advantage of this intermediate, interpretable data representation is that it is accessible to and interpretable by humans. State-of-the-art ASR do not offer this possibility because they do not need an interpretable inner structure as long as the ASR functions correctly. These models are typically end-to-end systems, which are trained on the raw audio and final transcription. However, the feature extraction stages that end-to-end models develop can be shown to represent phonetic concepts (Dieck et al. 2022).

1.3.1 Representation of Speech

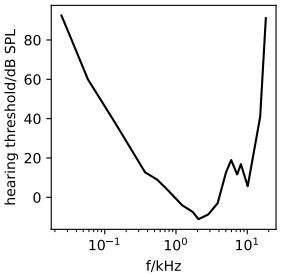


Figure 1.4: Absolute hearing threshold for pure tones (normal hearing):
Data from Vogten 1974.

Audio signals in digital computers are often processed and stored as ".wav" data. These are simple vectors of the microphone displacement at discrete time steps. It is common to store each value as a 16-bit integer.[13] The human perception of sound is different from that. Because of the physiology of the middle and inner ear, we have a spectral selectivity for frequencies above 30Hz and below about 13kHz (see Figure 1.4).

In addition to that, because of dynamic compression of the auditory system, our perception of the sound level is not a linear function of the level. Dynamic compression is the result of several processes, such as the stapedial reflex, which changes the impedance of the middle ear; the outer hair cells, which act as an amplifier in our inner ear; and three types of afferent nerve fibers that innervate the inner hair cell with different rate-level functions (Heil and Peterson 2015).

This nonlinear perception of the sound level is sometimes approximated with a logarithmic transformation. This transformation is derived from the Weber-Fechner law, which is an empirical law about the just noticeable difference (JND)[14]. It states that the JND in stimulus intensity s is proportional to s and that this JND (i.e., just noticeable change Δs) elicits a constant perceived change Δr:

$$JND \propto s \tag{1.16}$$

$$\frac{\text{JND}}{s} = c_0 \cdot \frac{\Delta s}{s} = \Delta r \tag{1.17}$$

For small Δs and Δr, and constants c integration yields:

$$r = \int dr = \int \frac{c_0}{s} ds = c_1 + c_2 \log(s) \tag{1.18}$$

This nonlinear transformation has the typical implications of nonlinear distortion (see appendix D). To roughly approximate this physiological representation, the following transformations of the acoustic data are commonly used:

Spectrogram

Digital signal processing, and especially visualization, often involves a transformation that highlights both spectral and temporal structures. This is done with consecutive Fourier transformations $\mathscr{F}$ of short segments of the signal $s(t)$, called short time Fourier transform (STFT):

$$\text{STFT} = \mathscr{F}\left[s(t) \cdot \text{wf}(\tau_n, \delta\tau, t)\right] \tag{1.19}$$

where $\text{wf}(\tau_n, \Delta\tau, t)$ is called the time window, or window function. It removes all parts of the signal except for a duration $\Delta\tau$ at a discrete time step τ_n. Depending on the application, different time windows with specific cutoffs are used. This is because multiplication in the time domain is equal to convolution in the frequency domain, which changes the spectrum of the segment. For the simulations with FADE, a Hamming window with $\Delta\tau = 25$ms was used.

The sequence of spectra can then be concatenated to a matrix and displayed as an image (see Figure 1.5). Typically, the time windows overlap. For the simulations with FADE, the overlap was such that the center of the time windows was placed 10ms apart. This number is sometimes referred to as window shift.

The STFT is invertible:

$$\text{STFT}^{-1}\left(\text{STFT}s(t)\right) = s(t) \tag{1.20}$$

This makes it a commonly used feature space for digital signal processing.

[13] This encoding leads to the so-called quantization noise. The continuous value of the microphone displacement and its discrete approximation with a digital signal differ whenever the analog-to-digital converter has to round up or down to the next closest digital value for the amplitude. The quantization noise has an equal distribution, so it is not Gaussian white noise. It is mathematically similar to the implementation of the distortion component (Plomp 1978) in Hülsmeier, Buhl, et al. 2021, who used a normal distribution.

[14] The Weber-Fechner law is not always valid, such as for stimuli just above the hearing threshold and some temporal structures of the stimulus (Carlyon and B. C. J. Moore 1986). It is an optimal sensing strategy for maskers with a log-normal distribution in the time domain (Pednekar et al. 2023).

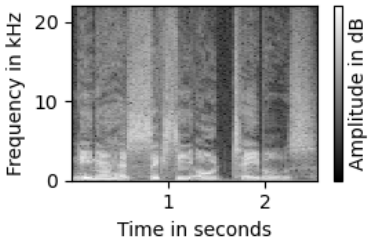


Figure 1.5: Spectrogram of a matrix test sentence:
The amplitude was logarithmically scaled for visibility. short time Fourier transform (STFT) with 30 ms non overlapping 0.25 Tukey time window.

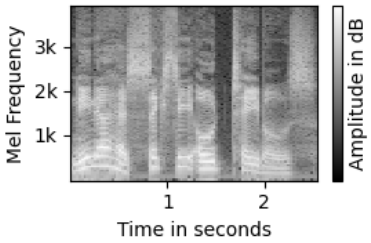


Figure 1.6: Log-mel-spectrogram of a matrix test sentence:
The spectrogram is not discretized. Note the change of the frequency axis.

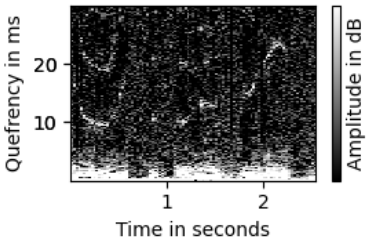


Figure 1.7: Mel-Frequency-Cepstrum of a matrix test sentence:
The low quefrencies correspond to the impulse response of the vocal tract, while the high quefrencies correspond to the glottal pulse train. The amplitude was logarithmically scaled and cropped to enhance the contrast.

Log-Mel-Spectrogram

The log-mel-spectrum (LMS) of a signal $s(t)$, according to ETSI 2003, p.10[15], is defined as:

$$\mathrm{LMS}(\nu) = \ln M(\nu) \tag{1.21}$$

$$\nu = 2595 \cdot \log_{10}(1 + f/700\mathrm{Hz}) \tag{1.22}$$

$$M(f) = |\mathrm{STFT}(s(t))| \tag{1.23}$$

ν is called the mel frequency, and f is the frequency in Hz. M is called the magnitude and is equal to the absolute value of the spectrum. The LMS is typically a discrete list of $N = 23$ values (not shown in Figure 1.6). For this, $M(f)$ (equation 1.23) is integrated with N triangular, normalized kernels placed at equidistant mel frequencies below the Nyquist frequency (see Figure 1.8).

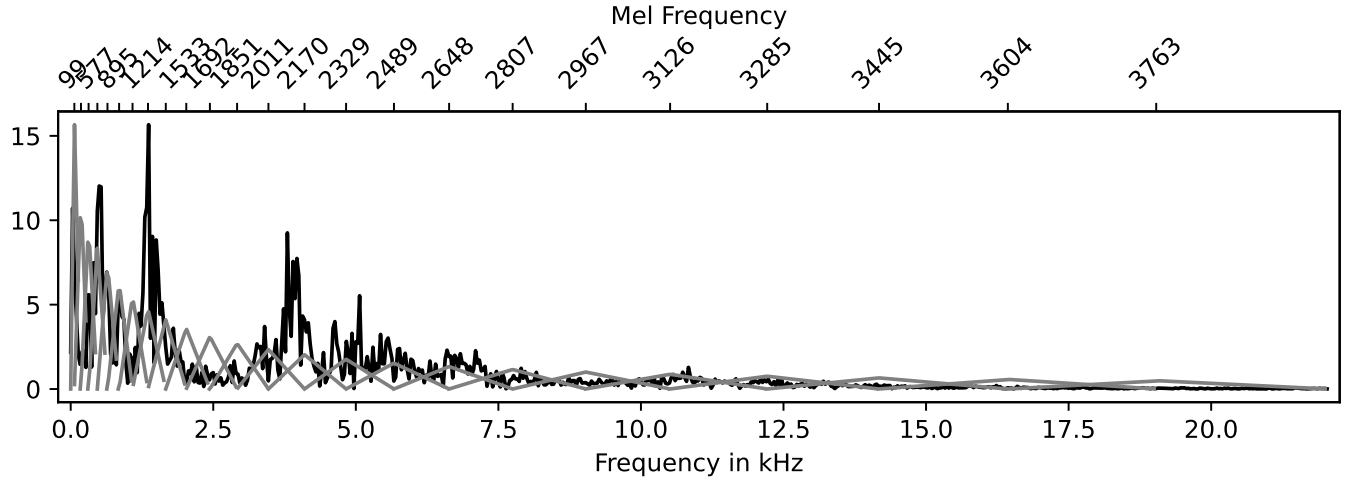


Figure 1.8: log-mel-spectrum (LMS) filterbank:
A time frame of the spectrogram (black) is integrated with normalized, triangular kernels (grey). The center frequencies of the kernels are equally spaced on the mel frequency scale. Some labels for lower mel frequencies are omitted.

The mel frequency[16] is simply a logarithmic transformation of the frequency in Hz with an offset, like Equation 1.18. This can be seen with the equivalent formulation of Equation

[15] It is part of the standard implementation of FADE.

[16] Another commonly used scale is the Bark scale, which is based on the number of critical bands that cover the audible range. The Bark scale is proportional to the mel scale (Fastl and Eberhard Zwicker 2007). The proportionality constant depends on equation 1.22.

1.22:

$$\nu = \alpha + \log_\beta(700 + f/\text{Hz}) \tag{1.24}$$

$$\alpha = -\log_\beta(700) \quad , \quad \beta = 10^{1/2595} \tag{1.25}$$

The discrete LMS according to Equation 1.21 is the basis of the feature extraction stage of FADE.

Mel-Frequency Cepstral Coefficients

The Mel-Frequency-Cepstrum (MFC) is defined as the absolute value of the inverse Fourier transform of the LMS:

$$\text{MFC} = |\mathscr{F}^{-1}[\text{LMS}(\nu)]| \tag{1.26}$$

$$= |\mathscr{F}^{-1}[\ln M(\nu)]| \tag{1.27}$$

The logarithm turns multiplication into addition, and the Fourier transformation is a linear operator. This means that modulated signals can be deconstructed into the sum of a carrier M_c and a modulating signal M_m:

$$\text{MFC} = |\mathscr{F}[\ln(M_c(\nu) \cdot M_m(\nu))]| \tag{1.28}$$

$$= |\mathscr{F}[\ln M_c(\nu)] + \mathscr{F}[\ln M_m(\nu)]| \tag{1.29}$$

This way we can distinguish the impulse response of the vocal tract from the glottal pulse train (see Figure 1.7) because convolution of a glottal pulse with the impulse response of the vocal tract is equivalent to multiplication in the Fourier domain. [17]

Mel frequency cepstral coefficients (MFCC), as they are used in the following, are the real part of the Cepstrum of a discrete LMS. Different from the implementation in Schädler, Warzybok, Hochmuth, et al. 2015, no discrete temporal derivative was used, such that the MFCC features only contained information about their specific time window $\text{wf}(\tau_n, \delta\tau, t)$ (see Equation 1.19).

The name "Cepstrum" is an anagram of the word "spectrum", with the first four letters reversed. The argument for the MFCC is called the quefrency. Bogert, M. Healy, and Tukey 1963, ch.15 invented this analysis. See Oppenheim and Schafer 2004 for an introduction and the history of the cepstrum.

Gabor Filter Banks

Separable gabor filter bank (SGBFB) are part of the standard implementation of FADE. Like Gabor filter banks, these extract spectral (frequency axis of the spectrogram) and temporal (time axis of the spectrogram) modulations from a spectrogram. Typically, LMS is used as input. The operation of SGBFB is very similar to pattern recognition in images, only that

[17] That mel frequency cepstral coefficients (MFCC) contain some information about temporal structure is also evident from the relation between the magnitude $M(\omega)$ of a spectrum and the minimum phase Φ_0 (Bechhoefer 2011).

$$\Phi_0 \approx -\frac{1}{4} \cdot \frac{d}{d\ln(f)} \ln M(f)$$

Together with the Fourier transform of the derivative operator of a test-function g:

$$\mathscr{F}\left[\frac{dg}{dt}\right] = i\omega \cdot \mathscr{F}[g]$$

The relation between the minimum phase Φ_0 and the MFCC, with index ν (reciprocal space of the mel-frequency), holds approximately:

$$\mathscr{F}(\Phi_0) \propto \mathscr{F}[\ln M] \cdot \nu$$

$$|\mathscr{F}[\Phi_0]/\nu| \propto \text{MFCC}$$

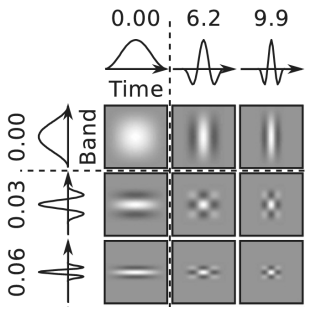


Figure 1.9: Separable gabor filter bank (SGBFB):
The SGBFB works as a filter of a LMS (see Figure 1.6). Adopted from Schädler and Kollmeier 2015.

the computation can be separated into two independent subproblems on each dimension of the LMS.

This is done by separating the filter into a filter for the time and frequency axis and the subsequent outer product of the features (Schädler and Kollmeier 2015). SGBFB form an orthogonal basis of conventional Gabor filters and can be used as a fast alternative to Gabor filters. An example of SGBFBs for the lowest modulation frequencies is shown in Figure 1.9.

1.3.2 Framework of Auditory Discrimination Experiments

The framework of auditory discrimination experiments (FADE) was developed in 2015 in Oldenburg by Schädler, Warzybok, Hochmuth, et al. 2015. For this book, FADE was used to simulate the SRT of matrix sentence tests in Mandarin Chinese, Cantonese, and English. FADE approximates human listening performance with an optimum detector. This approximation assumes that neurologically healthy humans can make optimal use of their sensory input to decode a message, i.e., that every available acoustic information that can be received by the auditory system is evaluated for a response in a hearing experiment. This is done with the following model, which takes as input any of the speech features introduced in section 1.3.1:

ASR Backend

The ASR backend of FADE is using the "Hidden Markov Toolkit software" (HTK) (see Young et al. 2002; Schädler, Warzybok, Hochmuth, et al. 2015), which is an implementation of a hidden Markov model in conjunction with a Gaussian mixture model (HMM-GMM). Central to a HMM-GMM are two components:

1) The hidden Markov model (HMM) is a model of an evolving system. The system can be in several (indirectly observable, i.e., hidden) states that control how the system behaves. An important assumption of HMMs is the Markov assumption, which states that the transition from one state to the next only depends on the previous state. This way, transitions between states are modeled as a statistical process with a transition matrix A.

The transition matrix can include boundary conditions (e.g., transitions between some states are forbidden and fixed to a probability of zero). In the case of language models, this would be determined by the so-called grammar, so that, e.g., subjects always follow verbs. For matrix sentence tests (see Figure 1.10), a word with the label ⟨A⟩ is made up of several states (indicated by black boxes) that can only persist or transition to the next state. The transition between words in Figure 1.10 is limited by the grammatical structure of the matrix sentence test.

Likelihoods for HMMs can be computed efficiently, which makes them a versatile tool to decode noisy messages such as speech. A different application of a HMM will also be introduced in chapter 3 for the implementation of the single psychometric function goodness of fit measure (SiPGOF). Refer to Rabiner 1989 or Bishop 2019 for an introduction to HMM algorithms such as the forward-backward algorithm, Baum-Welch algorithm, or Viterbi algorithm.

2) The Gaussian mixture model (GMM) is a statistical model of the acoustic features at a specific time frame (see Figure 1.10). The multidimensional feature space is modeled with a set of Gaussian distributions. The mean and standard deviation (STD) of the Gaussian distributions change over time with each transition of the states of the HMM.

The working principle of the GMM inside the HMM can be understood as a fully independent sub-problem in the optimization of the HMM. For this we parametrize the Gaussian distributions of state $z = i$ with a vector $\theta_i = (\mu_1, \mu_2, \ldots, \sigma_1, \sigma_2, \ldots)$. μ and σ are the mean vector and covariance matrix in the multidimensional feature space. One function of the GMM is to compute the likelihood l to be in state z at every time frame 1 up to N with a weight $w_z \in (0, 1)^N$:

$$\mathrm{GMM}(z, w_z) : \mathrm{l}_z \tag{1.30}$$

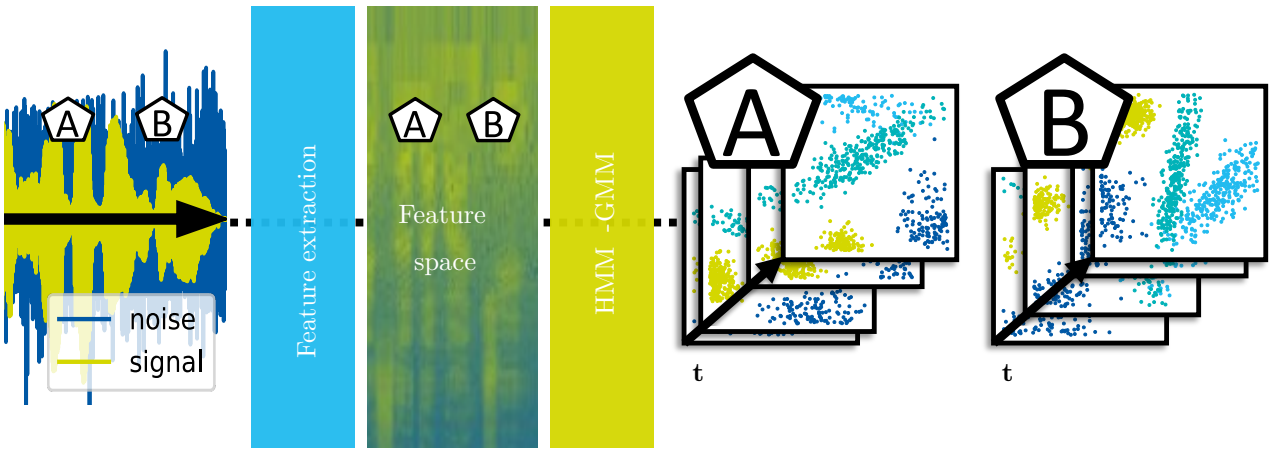


Figure 1.10: Concept of a hidden Markov model with a Gaussian mixture model (HMM-GMM) for automatic speech recognition (ASR):
Audio files with labels Ⓐ , Ⓑ are transformed to a labeled feature space. The HMM-GMM models the changes of the features (only the two principal components of the multidimensional feature vector are shown in the black boxes) over time t with discrete states, which are assigned to the labels Ⓐ , Ⓑ . Transitions between labels are only allowed according to the transition rules of the speech, called grammar.

The open parameter w_z is a vector of the probability to be in state z during each time frame of the evolution of the HMM. A second function of the GMM is to optimize the internal parameters θ_i of every state $z = i$ for a higher likelihood l_z:

$$\theta_{z,opt} = \max_{\theta_z} (\mathrm{l}_z) \tag{1.31}$$

It is important to note that θ_i never leaves the scope of the GMM, which decouples its optimization from the optimization of the HMM.

3) The expectation maximization algorithm (EM) is the optimizing procedure, which ensures that the information in the features is optimally used. This step is central to the optimum detector approximation of FADE. The EM can be shown to always converge to a local optimum[18]. It uses the likelihood (equation 1.30) that the entire observation (i.e., the list of feature vectors of an audio file) was emitted by a state z. For this computation, it is not relevant what statistical model is used or how the state z is parametrized as long as Equation 1.30 can be computed.

For the optimization of the HMM, only the likelihood from the GMM is used, while for the optimization of the GMM, only the weights w_z from the HMM are used. The weights follow from the standard forward-backward algorithm for the optimization of HMMs (see Rabiner 1989, or Young et al. 2002, sec.9.4). The EM is simply a cooptimization of the HMM together with a GMM.

To give an example, the optimization process with the EM will first initialize the HMM and GMM with default parameters. Then, with the first iteration, the weights w_z are computed with the likelihoods provided by the GMM. Thereafter, the GMM is optimized to maximize l_z (equation 1.31). Then, the next iteration of the forward-backward algorithm for the HMM starts, and the process repeats until the convergence criterion is met. Once the optimization procedure is finished, the HMM-GMM is fully trained.

A fully trained HMM-GMM-ASR consists of the parameters of the Gaussian mixture for each state, the transition matrix, and the most likely distribution of the first state of the Markov chain. With these, it is possible to calculate the most likely state sequence and corresponding labels of an unlabeled sound file, which is the ASR prediction. An advantage of this model is that it is small, follows a specific grammar, and does not require much training data compared to state-of-the-art ASR. This makes it a strong model to predict SRTs of matrix sentence tests, but not for open-set speech corpora.

[18]This is one of the reasons why the training data needs to be sufficiently large and mixed with random noise, such that a global optimum can be found through the variation of the training data.

The implementation of the EM is not trivial. A small example of this algorithm can be found under M. K. Scharf 2025. The EM of the FADE backend is implemented in the commercial HTK software (Young et al. 2002), which is freely available for research purposes.

Training Phase

Like all ASR-based SI prediction models, FADE starts with a training phase of the ASR backend. As the performance of the HMM-GMM backend is dependent on the SNR of the training data, N instances of the ASR are trained, each for a specific value of SNR (see left-hand side of Figure 1.11). For the training data, entire labeled sentences of the matrix test are mixed with random sections of the masker.

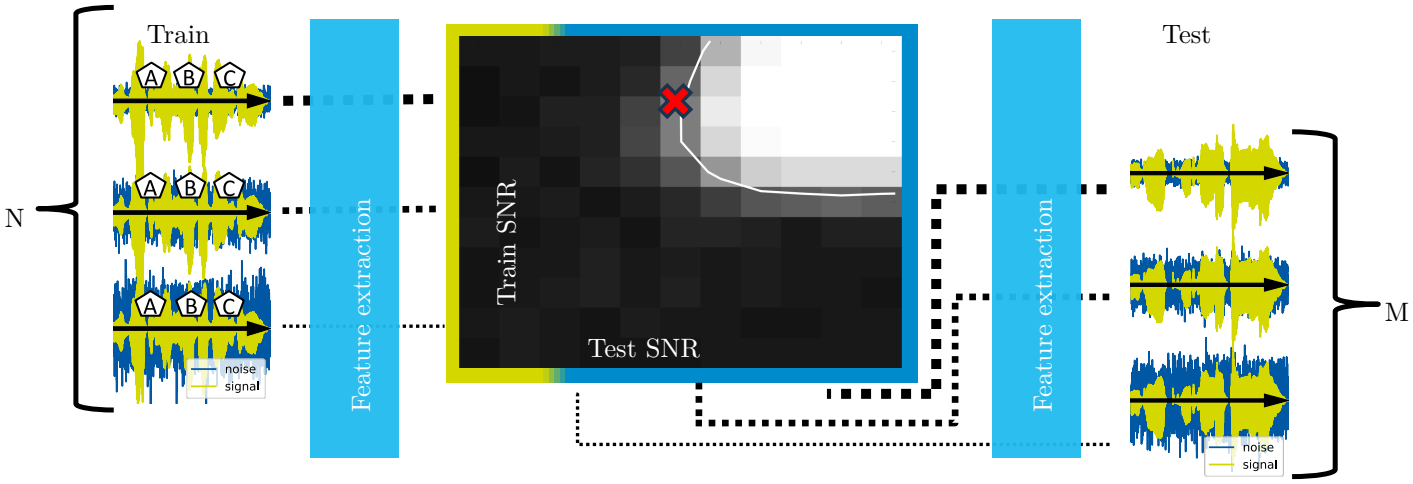


Figure 1.11: Concept of finding the optimum performance of the automatic speech recognizer (ASR) in framework of auditory discrimination experiments (FADE):
Labeled training audio at N discrete values of SNR is transformed to a feature space and used to train N instances of the ASR. Each of the ASRs is tested at M discrete values of SNR. The resulting performance matrix is colored white for 100% correctly identified labels Ⓐ, Ⓑ, Ⓒ, and black if no label was correctly identified. The lowest possible testing SNR for 50% correct labels is the predicted SRT and marked with a red cross.

For other psychometric tests with a binary response (heard or not heard), such as tone in noise detection, the training data would contain a random section of the masker mixed with the target and assigned to a binary label (target present or absent). An example of labeled matrix sentence training data is depicted on the left of Figure 1.3 with labels Ⓐ, Ⓑ, Ⓒ, Ⓓ, Ⓔ for each word.[19] The raw audio is always transformed to a feature space on which the training and testing of the ASR is performed (see section 1.3.1).

Testing Phase

In a second step, new testing data is generated over a range of M values of SNR, which is typically the same range as the training data (see right-hand side of Figure 1.11). Although both training data and testing data are generated from the same audio corpus, the data itself is not the same. This is because the masker is always a random section of the original, much longer masking sound file. It is important to stress this point, as we do not want to train and test with identical data, which leads to overfitting effects (see chapter 4).

Evaluation Phase

The N previously trained ASRs are then used to predict the labels of the M testing data, which results in an $N \times M$ performance matrix (see center of Figure 1.11). The predicted SRT is defined as the lowest possible testing SNR that can achieve the desired word error rate. Typically, the SNR at 50% correctly understood items is used. However, as the

[19] The labels of the audio files for FADE do not need a time stamp, as the position of the words are determined by the most likely sequence of the Markov chain of the HMM.

performance matrix would only allow discrete values for the testing SNR, an iso-word-error-rate is interpolated (white curve in Figure 1.11), and its minimum test SNR is used instead. This is the predicted SRT of the FADE model.

1.3.3 Speech Intelligibility Index (SII)

The SII is defined under ANSIS3.5 1997 and is a widely used measure of SI. Its widespread use partly owes to the fact that it quickly maps the level of the long-term spectrum L[20] of signal and masker to a value on the standard interval [0,1] that is highly correlated with SI.

$$\mathrm{SII} : L_{\mathrm{signal}}, L_{\mathrm{masker}} \rightarrow [0, 1] \tag{1.32}$$

Therefore, the SII can be used as an objective function for machine learning algorithms.

1.3.4 Band Importance Function (BIF)

When modeling normal hearing listeners, the SII takes one set of free parameters: the band importance function (BIF). The BIF is a spectral weighting of the audibility function A_i.

$$\mathrm{SII} = \sum_{i=1}^{N} \mathrm{BIF}_i A_i \tag{1.33}$$

Where A_i in the frequency band[21] i is a function of the level of the entire signal and masker[22].

$$A_i : L_{\mathrm{signal}}, L_{\mathrm{masker}} \rightarrow [0, 1] \tag{1.34}$$

Accordingly, A_i is invariant to changes of the spectral phase, which determines the temporal structure of the signals.

The BIF assigns a weight to each frequency band according to its contribution to SI. The BIF is normalized to one[23].

$$\sum_{i=1}^{N} \mathrm{BIF} = 1 \tag{1.35}$$

In addition to the BIFs published alongside the definition in ANSIS3.5 1997, BIFs have been published for various languages (e.g., Wong et al. 2007; Du et al. 2019; J. Chen, Huang, and Wu 2016), speech tests (Yoho et al. 2018), and speakers (Shen and Kern 2018) to improve the SII even further. This way, machine learning algorithms for digital signal processing that aim to optimize SI for a specific type of speech can use the SII with an appropriate BIF as an objective function for the optimization process.

Unfortunately, as will be discussed in chapter 4, the BIF generalizes poorly from the speech for which it was measured to comparable speech. This means that there might not be as much predictive value in publishing BIF for a specific language, test, and speaker as has been suggested by the literature.

The SII can be improved if temporal changes of the signals are considered in the computation, which has led to other index-based SI-models (see Feng and F. Chen 2022) such as the speech transmission index (STI) (Houtgast et al. 2002), extended SII (eSII)(Rhebergen, Versfeld, and Wouter. A. Dreschler 2006), STOI (Taal et al. 2011, earlier proposed by Ludvigsen et al. 1990 and Kollmeier 1990), hearing aid speech perception index (HASPI) (Kates and Arehart 2022), and, more recently, deep learning models with speech foundation models (SFM) frontends (Schneider et al. 2019; Chiang et al. 2024; Saddler and McDermott 2024). Some ASR-based models, such as the phone-based binaural intelligibility model (PHOBI) (Roßbach, Kirsten C. Wagener, and Meyer 2023) or FADE, directly predict the

[20] ANSIS3.5 1997 uses the variable E to denote the level. See Equation 1.10.

[21] ANSIS3.5 1997 defines four calculation methods that differ in the resolution of the filter banks with 6 up to 21 bands.

[22] All frequency bands of signal and masker are necessary to compute the audibility function of a specific band because of masking.

[23] The extension of the SII with proficiency factors (Pavlovic 1987 and section 4.4) can also be seen as a different normalization constant of the BIF.

mean temporal distance of the phoneme posteriorgram, or the SRT_{50} respectively. ASR-based approaches do not rely on predefined tabulated weighting functions. Instead, they automatically determine the most appropriate weighting of speech features during the processing. This allows a more flexible and data-driven weighting that can adapt to varying acoustic conditions.

1.3.5 Mapping between SII and SRT

To convert the SII into a SRT, a reference SII (SII_{ref}) is used. The SNR of a new speech test that we want to predict is changed, while keeping the masking level fixed, until its SII matches the reference SII_{ref}:

$$SRT_{pred} = \underset{SNR}{\operatorname{argmin}} \left[SII_{ref} - SII(SNR) \right] \tag{1.36}$$

A graphical example of this method is shown in figure 1.12, where the exemplary reference SII_{ref} is indicated by a horizontal line.

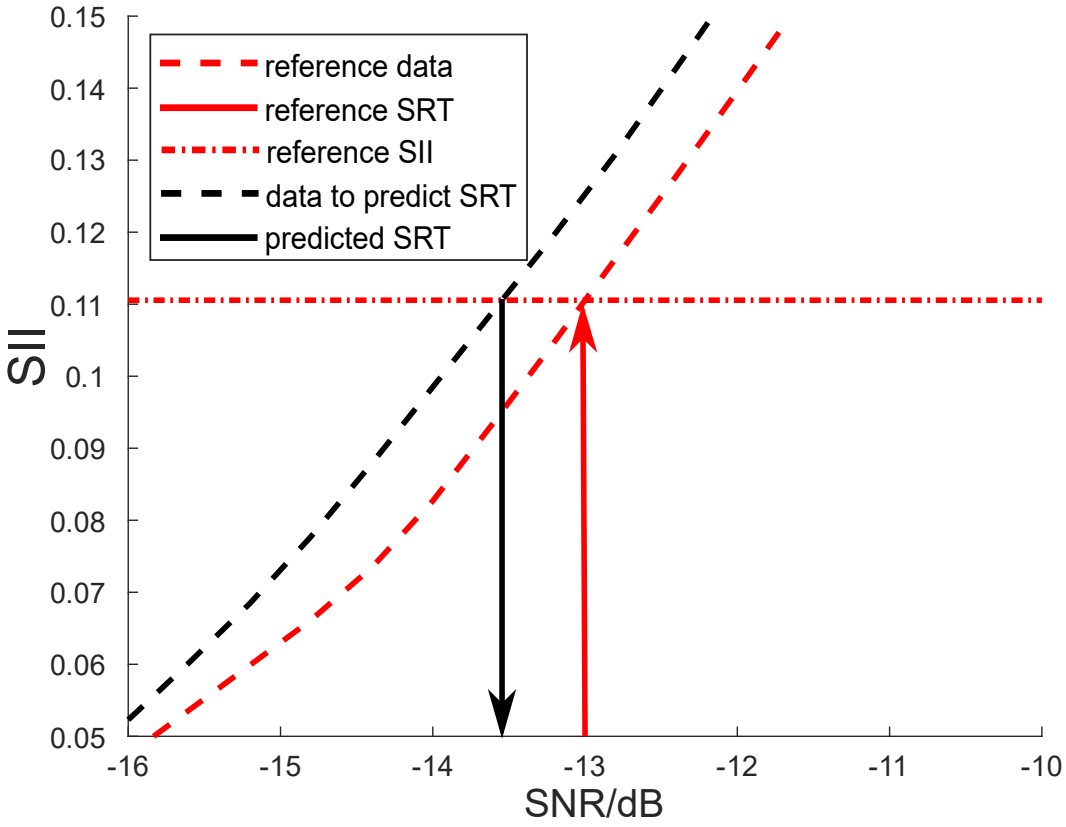


Figure 1.12: Map between SII and SRT:
With a reference signal and masking noise at a known SRT, a reference SII is calculated (red arrow, $SII_{ref} \approx 0.11$). The prediction is made by matching the SII of the prediction data to the reference SII_{ref}. The corresponding SNR is the prediction (black arrow). Note that the SII is no linear function of the SNR at low SRTs.

A reference SII can be derived from a different speech test with comparable information content, language, style, and speaker. In the studies of chapters 4, 5, and 6, the reference condition was defined using SRT of the original Mandarin matrix test (Hu, Xi, et al. 2018) in unmodulated speech noise with a long-term spectrum of a male speaker at normal vocal effort (Wouter A. Dreschler et al. 2001) (ICRA1) of −14.8dB corresponding to SII_{ref} of 0.0408.

Alternatively, the mean SII of all (to be predicted) SRT measurements can be used as a reference if a sufficiently large dataset is available. This, however, may lead to a kind of circular inference for small data sets and requires cautious interpretation. More specifically, using information about the mean of the empirical data will reduce the degrees of freedom

of the χ^2-distribution ν by one. This measure of goodness of a model (χ^2) is introduced as follows.

1.3.6 Fair Model Comparison

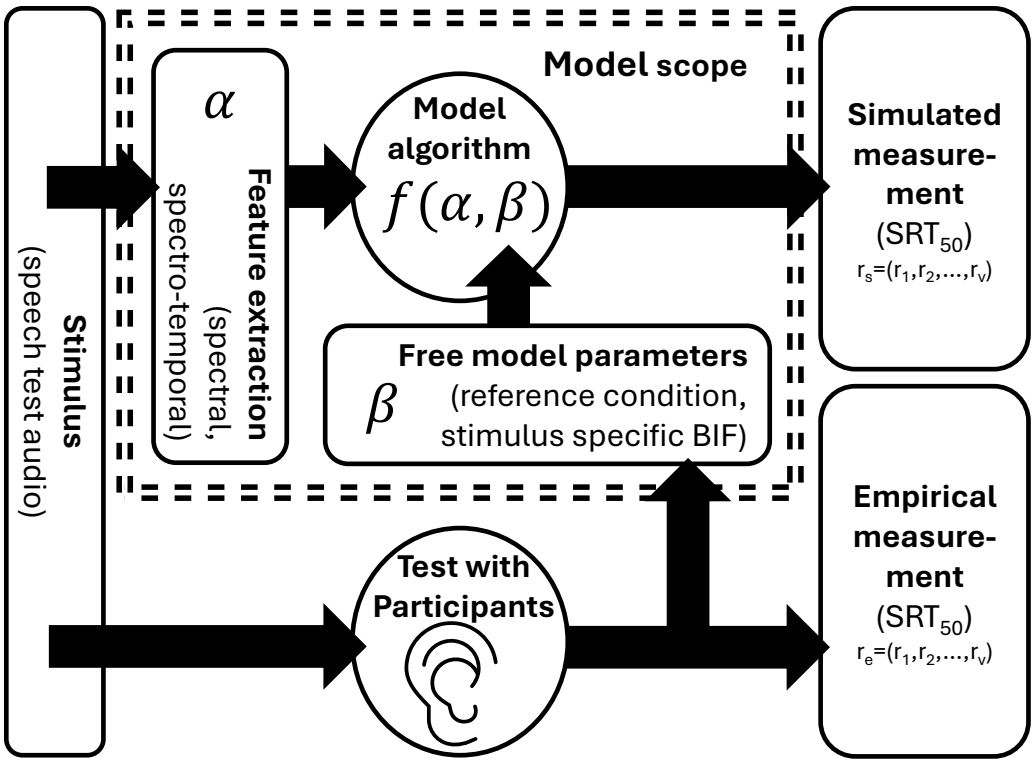


Figure 1.13: Information flow in the comparison between measurements and simulations: Participants and the model of the sentence test receive the stimulus information as speech test audio signals. The test procedure for the participants produces ν empirical SRT measurements r_e. The model frontend extracts features α from the stimulus and feeds these (highlighted arrow) together with tunable, free parameters β to the model algorithm, which returns ν simulated responses r_s. The free parameters β can be tuned with information about the empirical response. Examples of β are reference conditions, but also BIFs that were measured for the stimulus beforehand.

When comparing different model frameworks, care has to be taken with which information enters and leaves the model (see figure 1.13). All models considered here receive mixed or unmixed signal and masker audio signals, as well as information about the type of test (here, a matrix test). Based on this data, the models return a prediction of the SRT. However, additional free model parameters may be used to optimize the model prediction. If these free parameters depend on empirical measurements with the same speech material and masker, they change the degrees of freedom ν of the model prediction. This effect can be treated with the measure of goodness of a model (χ^2) (see section 4.2.4):

$$\chi^2 = \sum_i^{\mathrm{N}} \frac{(a_i - b_i)^2}{\alpha_i^2 + \beta_i^2} \tag{1.37}$$

a_i are the empirical values with their measurement uncertainty α_i. b_i are the predicted values with prediction uncertainty β_i. χ^2 can be used to quantify the significance of the agreement between the model and empirical data as a function of sample size. However, measurement uncertainties are often neglected in published comparisons of SI models such that χ^2 cannot be calculated.

One advantage of χ^2 is that it includes a correction if the model requires a reference, which is the case for the SII. Traditional measures are not sufficient for a comparison of SII-based predictions with FADE predictions. Pearson correlation coefficient ($\mathrm{R_p}$), bias, and the root mean square error (RMS) cannot discriminate between prediction and information

that entered the model as a reference condition. The R_p is not a fair statistical measure for a comparison between two models if one requires a reference condition, which is derived from the empirical data itself. This is evident from the definition:

$$R_p = \sqrt{\frac{\sigma^2_{AB}}{\sigma_A \sigma_B}} \tag{1.38}$$

σ^2_{AB} stands for the covariance of a dataset A with a dataset B, and σ_A, σ_B stand for the standard deviation of datasets A and B, respectively. A model that utilizes some information from A to produce a prediction B may report the same covariance as a second model that utilizes no information about A for a similar prediction B. However, the second model may be superior when such an advantage is corrected for. This can be corrected with the χ^2. A reference condition may reduce the number ν of independent predictions of empirical data if it was derived from the same empirical data.

χ^2 is commonly used in physics to measure the quality of a model fit to empirical data (Barlow 1989). Its properties are a characteristic of quadratic forms of random variables, specifically sums of squared errors, which is a consequence of Cochran's theorem. χ^2 was first proposed by Pearson 1900. See Hughes 2010, Chapter 8.2 for an introduction.

1.4 Tonal Languages

Chapter 6 and 5 discuss SRT predictions in the two tonal languages, Cantonese and Mandarin. Tonal languages exist not only in Asia but are also quite common in Africa, the Pacific, and the Americas, totaling more than half of all spoken languages (Yip 2002). In contrast to non-tonal languages such as English and German, tonality can best be described as a distinct encoding of syllables with changes in the fundamental frequency F_0 and formants. F_0 is the dominant frequency in the spectrum. The formants are the higher frequencies that make up the timbre of a voice. These can be seen as horizontal lines in the spectrogram of Figure 1.14.

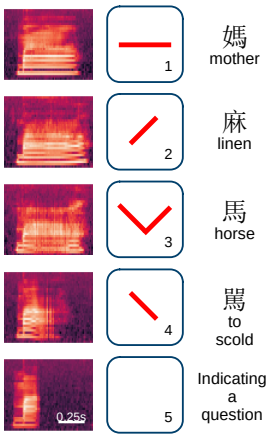


Figure 1.14: Spectrogram and tones of the syllable "ma" in Mandarin:
Possible meanings are indicated and can be used to construct a sentence that mainly contains the same syllable:
mā mā qí mǎ
媽媽騎馬,
mǎ màn
馬慢,
mā mā mà mà
媽媽罵馬
This translates to "Mother is riding a slow horse, so she is scolding the horse".

Mandarin Chinese defines four tones and the concept of a missing tone, resulting in a total of five distinctions between tones[24]. These five tones are depicted in Fig. 1.14 for the example syllable "ma". It shows the spectrogram of the syllable together with the commonly used nomenclature, which indicates the change in pitch. The first tone describes a constant high-pitched syllable with no temporal changes of the spectrum, while the second tone describes an upward shift in pitch. The third tone follows a downward slope, with a subsequent upward slope of the spectro-temporal structure. A downward slope in the spectrogram characterizes the fourth tone. The fifth option, which describes a missing tone, is sometimes counted as an additional tone. This tone has the same onset as all other tones, but misses a temporal characteristic due to its short duration. Also, not every combination of tones in successive words is allowed. In Mandarin, for two consecutive 3rd tones (down-up), the first 3rd tone changes to the 2nd (up). This type of rule is called a sandhi rule.

The SGBFB feature representation of speech (see section 1.3.1) is well-suited to capture tonal cues. For the simulations with FADE, the maximum size of the temporal modulation filter was 40 frames, which captures temporal changes of up to 0.4s for the window shift of 10ms. Tonal cues fall well within this range (see Figure 1.14).

Tonal cues carry a significant amount of speech information, which becomes evident from a popular children's play in Mandarin-speaking countries, where a sentence is guessed from the tonal structure alone. Tones in Mandarin Chinese must have already existed before the Middle-Chinese period, which is evident from the 601AD Qiè-Yùn (切韻) rhyme dictionary (see Baxter 1992, chapt.8).

1.5 The Lombard Effect

The discovery of the Lombard effect started with an invention. In 1908, Robert Bárány[25] (1876–1936) reported and patented a deafening device called the Bánáry noise drum (see

[24] Cantonese has a total of six tones.

[25] A later Nobel Prize winner, who linked the caloric response to the physiology of the inner ear, see Bracha and Tan 2015 for a short biography and his portrait.

Figure 1.15). Quickly, electromechanical versions of this device had been developed and were used for medical testing by otorhinolaryngologists all around Europe. One of them, a French physician by the name of Étienne Eugéne Joseph Lombard (1869–1920, see Lermoyez 1921), noticed an effect with his patients: As soon as they were subject to the intense noise of the device, they would speak with an altered voice. Lombard published his results in 1911 in French (Lombard 1911). The German-speaking community, however, learned of this discovery through the English physician Donald Schaerer, who brought word from Paris to Vienna such that the discovery was falsely assigned to Bárány. The resulting dispute between Bárány and Lombard was eventually settled in favor of Lombard a few months later (Lane and Tranel 1971).

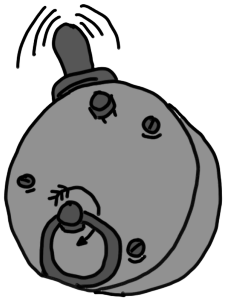


Figure 1.15: Illustration of Bánáry's noise drum: The earpiece (top) of the noise drum was inserted into the ear of a patient. It worked similarly to an egg clock and was meant to diagnose severe deafness.

Upon its discovery, it was suspected that the Lombard effect in humans is a reflex. However, speakers can actively control and even learn to suppress the Lombard effect. But even for trained speakers, it is difficult to maintain this suppression (Pick et al. 1989). After the discovery of the Lombard effect with Japanese quails (Potash 1972), the Lombard effect has been identified over a broad taxonomic range, even manifesting for some frogs (Love and Bee 2010). This way, it is estimated that the Lombard effect may have evolved more than 450 million years ago (Luo, Hage, and Moss 2018). The Lombard effect in humans depends on the background noise for the speaker. It is strongest for broadband noise (Stowe and Golob 2013), which masks the entire speech spectrum. The ICRA1 masker of the studies in chapters 4,5, and 6 fulfills this condition (see Figure 4.1a).

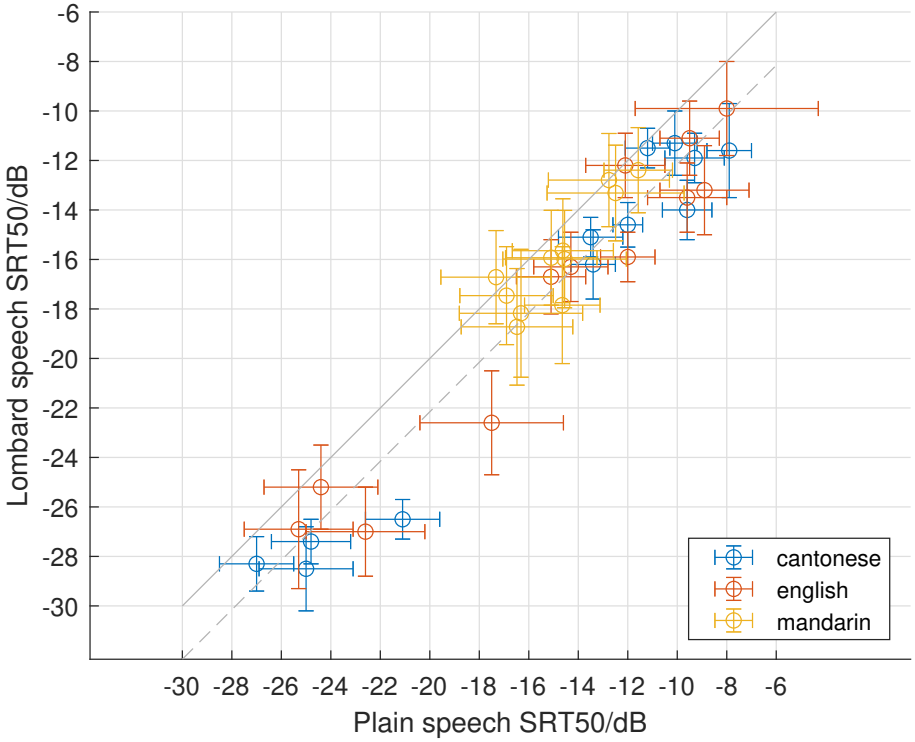


Figure 1.16: Comparison of empirical SRT for Lombard- and plain speech: Lombard speech has a lower SRT than plain speech for the three languages that were modeled in this book. The Lombard gain (LG) is the bias (dashed line, see section 4.2.4) in this Figure, according to Equation 1.39. Pooled data from chapters 6 and 5.

Lombard speech differs from plain speech in the following aspects:

- raised sound level (Lombard 1911; Hanley and Steer 1949; Pittman and Wiley 2001; Lu and Cooke 2009; Trujillo et al. 2021; F. Chen, Pan, et al. 2025)
- upward shift of the fundamental frequency together with a change of the overall spectrum (spectral tilt) (Summers et al. 1988; Pittman and Wiley 2001; Lu and Cooke 2009; F. Chen, Pan, et al. 2025)
- upwards and downwards shifts of the formants (Summers et al. 1988)

- longer duration of syllables and slower speaking rate (Hanley and Steer 1949; Pittman and Wiley 2001)
- changes in gestures and facial cues (Alghamdi et al. 2018; Trujillo et al. 2021)

These findings have been applied to speech enhancement algorithms that emulate Lombard speech (Skowronski and Harris 2006; Saba and Hansen 2022). For applications in ASR, Lombard speech requires different processing than plain speech (Marxer et al. 2018), suggesting distinct coding strategies.

Lombard speech has a lower SRT than normal, plain speech (Dreher and O'Neill 1957; Summers et al. 1988; Pittman and Wiley 2001; Lu and Cooke 2008), which is also the case for the empirical data of this book (see Figure 1.16). The SRT difference between plain and Lombard speech is called the Lombard gain (LG):

$$\mathrm{LG} = \mathrm{SRT}_{\mathrm{plain}} - \mathrm{SRT}_{\mathrm{Lombard}} \tag{1.39}$$

1.6 Calibration

The calibration of a measurement setup is always a comparison to at least one known reference (e.g., a reference sound source in the case of acoustic measurements). A calibration aims to make the numerical output of the measurement independent of the measurement setup used.

Calibration is different from gauging (German "Eichung"). The former is a necessary step in every measurement procedure and can be done by the experimenter. Gauging in the context of calibration requires an additional trusted authority, which certifies that the numerical output of a measurement device adheres to a standard. These authorities (German "Eichbehörden") are typically operated by states[26], as gauged measurement devices are, among other things, necessary for trade and taxation. Thus, gauging provides an additional layer of trust in a calibration. Devices that are used as references for calibration are typically standardized and certified by these authorities.

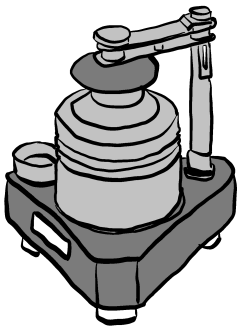


Figure 1.17: Illustration of a headphone coupler:
A headphone is placed on top of the coupler and held in place with a lever (top, black) to ensure a consistent downward force. The coupler consists of a capsule microphone and an air chamber that matches the impedance of an average human ear. Thus, the colloquial name "artificial ear".

Depending on the requirements of the experiment, different levels of calibration accuracy are commonly used. Standards often report uncertainties of a quantity x as a confidence interval:

$$\Delta x = \sigma_x \cdot k \tag{1.40}$$

σ_x is the standard deviation (STD) of a series of measurements of x (modeled with a Gaussian distributed random variable). The coverage or expansion factor k (German "Erweiterungsfaktor", see ISO GUIDE 98-3), which is often chosen to be in the range of $2-3$, controls the confidence in this uncertainty estimate.

1.6.1 Calibration of Headphones

A common calibration approach for headphones in hearing tests is to use a standardized microphone setup, sometimes called an artificial ear, to record the sound level produced by the headphones with a specific signal. Tabulated data for the headphones, called reference equivalent threshold sound pressure level (RETSPL)[27], is then used to derive the level in units of decibel relative to the hearing threshold of a normal hearing person (dB HL), which is then independent of the measurement setup. This approach makes it possible to compare the results of different hearing experiments and to reproduce measurements with entirely different equipment.

Several companies have specialized in the production of certified artificial ears, which are sold worldwide. However, the cost of these devices is too high for the calibration of mobile

[26] Even though reproducible measurements for people within a state are guaranteed by this method, it does not guarantee that the measurement is compatible with standards of other states, e.g., standard metric (SI) units. The SI units are an outcome of the French Revolution, when the meter as the standard of length and the gram as a standard of weight were defined for the first time. The historic fact that the royal English neighbor did not readily adopt this revolutionary French idea remains an extremely costly cause of confusion today (see Mishap Investigation board 1999).

[27] These tables are created by repeated measurements of the hearing thresholds of normal-hearing people of both genders with the specified headphones.

health (mHealth) applications by the patients. A calibration method that is accurate enough for meaningful audiological measurements with mHealth applications on unknown hardware is discussed in chapter 2. As a first step, this procedure requires a calibrated microphone.

1.6.2 Calibration of Microphones

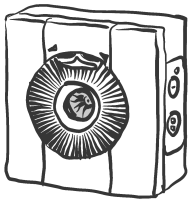


Figure 1.18: Device to produce a reference sound level:
Brüel & Kjær© Sound Calibrator Type 4231. The microphone capsule for which the calibration is required is placed inside the opening, facing the internal microphone.

To calibrate microphones, we require a device that can produce a reference sound of known frequency f and level at the microphone[28]. For reproducibility, the microphone has to couple consistently to the reference sound source in a way that is comparable to later arrangements of the microphone. Only then can we map the SPL at the microphone to the numerical output of the measurement device with a scalar proportionality constant for every f.

These frequencies f and corresponding levels define the map between numerical and calibrated output. Assuming linearity and time invariance, this map is the magnitude of the transfer function or frequency response of the microphone. It may most elegantly be measured with spectral sweeps (Müller and Massarani 2001). For higher precision, the transfer function is measured for all possible directions of incidence in the far field. This describes the directionality of the microphone (see appendix B).

To calibrate microphones that violate time invariance on large timescales, regular recalibration is required. If the frequency response is known except for a constant offset, it is common to calibrate with a single scalar proportionality constant for a reference frequency. In audiology, this reference frequency is often a 1kHz tone with 94dB-SPL or 114dB-SPL at the microphone (see Figure 1.18). The proportionality constant is also referred to as the calibration offset.

While the uncertainties, which are caused by the directionality and a varying frequency response of the microphone over time, are not trivial to quantify, the uncertainties of the reference sound source are easy to control. The measurement uncertainties of a microphone are equal to the uncertainties of the calibration offset if the uncertainty of the reference sound level is the dominating source of uncertainty.

Uncertainties of the calibration offset are systematic, which means that the uncertainties of two consecutive measurements with the same device correlate. This way, differences between two measurements are not influenced by uncertainties of the reference (but, for example, by assumptions about linearity between sound pressure level and numerical outcome of the measurement).

1.6.3 Helmholtz resonators

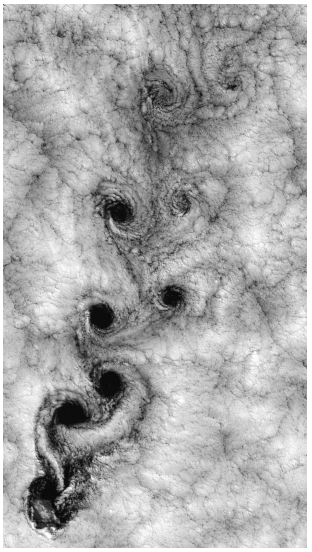


Figure 1.19: Kármán vortex street off the Robinson Crusoe Islands:
Turbulent streams produce repeating vortices behind an obstacle (here an island). The frequency scales with the speed of the stream. adopted from NASA 1999.

Chapter 2 proposes an estimate of a calibration offset with resonating bottles. These bottles are played like a flute and can be described with a jet-driven Helmholtz resonator. In the following, these whistles are called bottle-whistles.

Bottle-whistles make use of the periodic pressure fluctuations that are inherent to turbulent air streams behind obstacles. If a laminar stream of air is disturbed by an obstacle, repeating vortices may detach from the volume behind the obstacle. The size and frequency of the vortices depend on the speed of the air relative to the obstacle and the size of the obstacle (see Equation 2.1). Figure 1.19 shows an example of air vortices in a slowly moving cloud behind the Robinson Crusoe Islands. If the turbulent stream matches the resonance frequency f_r of a Helmholtz resonator next to the stream, a standing wave builds up inside the volume of the resonator. This resonance is audible as whistling.

The resonance frequency f_r of a bottle-whistle of volume V, effective length l', area A of the opening, and speed of sound c can be approximated by:

$$f_r = \frac{1}{2\pi} c \sqrt{\frac{A}{l'V}} \tag{1.42}$$

To derive this equation, you consider a piston of mass m levitating on the opening with area A of a large container of massless air. The piston at position x can oscillate on the

[28] Introducing a microphone into a sound field will produce scattering if the dimension of the microphone is in the same range as the wavelength λ, which requires special attention. The resonance frequency of bottle-whistles is $161.5 - 176$ Hz (see chapter 2). It follows:

$$\lambda = c_{air}/f \approx 343 \frac{\mathrm{m}}{\mathrm{s}} / 176\mathrm{Hz} \approx 2\mathrm{m} \tag{1.41}$$

This means that the effect of scattering can be neglected for calibration with bottle-whistles

trapped air like a spring with the spring constant k:

$$m\frac{d^2x}{dt} = -kx$$

Accordingly, the resonance frequency ω is (F is the force acting on the piston):

$$\begin{aligned}\omega^2 &= \frac{k}{m} \\ &= -\frac{1}{m}\frac{dF}{dx} \\ &= -\frac{A}{m}\frac{dp}{dx} \\ &= -\frac{A^2}{m}\frac{dp}{dV} \\ &= \frac{\rho^2}{m}\frac{A^2}{m}\frac{dp}{d\varrho}\end{aligned}$$

For an ideal gas, the speed of sound is defined by the rate of change of pressure p with density ϱ:

$$\frac{dp}{d\varrho} = c^2$$

With the approximation that the piston can be thought of as the air with mass m and density ϱ moving inside and close to the opening of the container: $m^2/\varrho^2 = VAl'$, where l' can be interpreted as a correction factor for this approximation, we arrive at (equation 1.42):

$$f_r = \frac{1}{\pi}\omega_r = \frac{1}{\pi}c\sqrt{\frac{A}{l'V}}$$

In chapter 2 330 ml Vichy shape-bottles[29] are discussed, which have a resonance frequency $f_r \approx 170$ Hz in agreement with Equation 1.42.

1.6.4 Directionality and Transfer Function of Smartphones

In general, smartphone microphones that run mHealth applications come in large variations. Like for human hearing, the angle of incidence has a profound influence on the measured sound field. This effect is used in some aspects of spatial hearing, but is problematic for reproducible experiments with microphones. Thus, after calibration, the angle of incidence may not be changed to keep the setup calibrated. For smartphones, this means that the orientation of the smartphone to a sound source has to match the direction for which the smartphone was calibrated.

As preparation for the study in chapter 2, the transfer functions of a small sample of smartphones were measured for multiple directions. These are discussed briefly in the appendix B.

Distance Dependence

The sound pressure level at the microphone depends on the distance between the sound source and receiver, i.e., the microphone. In the far field, for which sound propagation can be approximated by spherical waves and for lossless sound transmission, the energy E passing through a closed surface A around the source stays constant. The energy is proportional to the mean squared pressure $\overline{p^2}$ or power of the signal P (see Equation 1.9), which results in the well-known 1/distance dependence of the mean pressure of a spherical sound wave:

$$E = \int P \quad dAdt \tag{1.43}$$

$$E \propto \int \overline{p^2} 4\pi r^2 dt \tag{1.44}$$

[29] 0.5 l Vichy shape-bottles are known as NRW-bottles, or Sudpfanne. Although very common in Germany these are not standardized as DIN in contrast to their smaller 330 ml version.

Doubling the distance between the emitter and receiver leads to an additional factor of four in Equation 1.10:

$$10 \cdot \log_{10}\left(\frac{P(2r)}{P_0}\right) = 10 \cdot \log_{10}\left(\frac{P(r)}{4P_0}\right) \tag{1.45}$$

Expressed in dB, this factor is approximately six. This means that doubling the distance to an emitter in a free field leads to an attenuation of 6dB-SPL.

In the case of bottles held at arm's length from the microphone in chapter 2, the variation in distance can be approximated with 10% or 1dB, which is much smaller than the variation between whistling skills. Because of this, the method in chapter 2 includes the arm length uncertainties in the final measurement uncertainty.

1.6.5 Frequency Stability of Smartphones

In general, we can assume frequency measurements with microelectronic devices to be exact for the purpose of audible frequencies. Two effects are mainly responsible for imprecise frequency measurements with quartz clocks: the thermal drift, which leads to a systematic error, and jitter, which leads to a statistical error.

The thermal frequency drift for a temperature change ΔT of a typical quartz clock is in the range of (Gerber and Sykes 1966):

$$\frac{\Delta f}{f} \propto 10^{-6} \cdot \Delta T^2 \tag{1.46}$$

In comparison, according to Wier, Jesteadt, and Green 1977 human JND for frequencies between 0.4 kHz and 2 kHz is:

$$\frac{\Delta f}{f} = 2 \cdot 10^{-3} \tag{1.47}$$

The frequency drift in normal listening conditions is small enough for audiological applications with the resonance frequency of a typical quartz clock $f = 32768$ Hz[30].

The jitter is a random deviation of the frequency from the average frequency. As this deviation has a random distribution, sufficiently many (N) measurements can reduce the error like in Equation 1.15. Audio recordings with the typical sampling rate of 44100 Hz can be assumed to be unaffected by the jitter.

In conclusion, for audiological measurements on smartphones, the frequency does not have to be calibrated. This fact is critical for pure tone audiograms and the algorithm proposed in 2, which checks if the resonance frequency of the Helmholtz resonator is in the expected range.

1.7 Computing in Natural Science

A considerable part of the work that led up to the chapters of this book consisted of the "mechanics" behind the models. The methods that are discussed here cannot be calculated simply with pen and paper. As this typically happens unseen and unmentioned in the final publications, but makes up a considerable portion of the work, this section is meant to sketch the current tools and methods.

The analysis in the following chapters was done with the scripting languages: Python, MATLAB, and R. The model FADE consists of a binary, i.e., machine code, ASR backend that was compiled from C. The signal processing scripts were written in the MATLAB-style GNU Octave scripting language. Bash shell scripts handled files, environment variables, and processes. The two procedures of chapters 2 and 3 were implemented in Python.

The advantage of scripting languages that are interpreted and not compiled, in some fields of natural science, may be attributed to:

[30] This interesting number is a direct consequence of our hearing range: the frequency is the smallest power of two that we cannot hear (2^{15}). It is technically very simple to gear down a frequency with flip-flop-gates if it is a power of two.

- **Shallow learning curve:** Many scientific projects require computer resources. Not every scientist has a solid background in computer science. The community readily adopts solutions that require minimal preparation. A novice can learn the basics of scripting languages quickly.
- **Cross-platform capability:** Code can be run on different hardware with minimal effort if the interpreter is available. Interoperability accelerates the development process in everyday university life.
- **Community:** The community of users of some scripting languages is a major driving factor for tutorials and the development of high-quality, free libraries.

Although still widely used, MATLAB has been superseded by Python[31] as the de facto standard scripting language in the scientific community at the time of writing.

Hardware-optimized programs such as the ASR backend of FADE are not scripted and run comparably faster[32]. These programs have to be compiled for a given system. In the case of FADE, the system was a GNU Linux distribution, which comes with Bash shell scripting. Although Turing-complete, Bash is most powerful for automation. A serious downside of Bash scripts is that they are difficult to debug and maintain[33]. FADE relies on Bash scripts.

A relatively new development in the scientific community is the emphasis on the findable, accessible, interoperable, and reusable (FAIR) principle. While scripting languages adhere to some of these requirements, databases of scientific data and code often lack the FAIR properties. In my view, this may have several reasons:

- **Lifetime of Scientific Projects:** In the German scientific community, projects have a typical lifetime of 3 years. Employment on PhD projects is limited to 6 years.[34] After this period, data and corresponding scripts may lose interoperability and reusability as they are not maintained.
- **Incentive:** Well-written, documented, and reusable databases have only recently received recognition in scientific publications.
- **Education:** Although basic programming skills are taught and applied in most lab courses for undergraduate students, databases are not commonly part of the exercises. The reason for this may be that instructors are not up-to-date with appropriate solutions, because of rapid changes in computing infrastructure.

The data and programs published alongside this work adhere to the FAIR principle (Hu, Warzybok, et al. 2022; M. K. Scharf 2023; M. K. Scharf 2025).

In my view, young scientists in this field should get familiar with the necessary tools and try to follow the FAIR principle early on. FAIR is not only crucial for published data and tools but also for data and tools that are not primarily intended for publication (i.e., for internal use or to be part of a publication later on). FAIR starts with a proper version control like GIT for every project, consistent naming conventions, small working examples, and ends with appropriate comments for future reference. Ideally, this should become part of the daily routine of a PhD student.

[31] Followed by R, for statistics, especially in the social and behavioral sciences.

[32] The distinction between scripting languages and compiled languages has been blurred in the last few years with, e.g., Cython and MATLAB coder.

[33] This is a significant problem for ASR frameworks such as KALDI (Daniel Povey et al. 2011).

[34] Some exceptions apply, for example, to parental leave or other care work. The German law behind this is called "Wissenschaftszeitvertragsgesetz" (WissZeitVG).

Chapter 2:

Microphone Calibration Estimation for Mobile Audiological Tests with Resonating Bottles

reprint from
International Journal of Audiology

Maximilian Karl Scharf[1,2], *Rainer Huber*[1,3], *Michael Schulte*[1,4], *Birger Kollmeier*[1,2]

[1] Cluster of Excellence "Hearing4all", Oldenburg, Germany
[2] Medizinische Physik, Carl von Ossietzky Universität Oldenburg, Germany
[3] Fraunhofer-Institut für Digitale Medientechnologie (IDMT), Oldenburg, Germany
[4] Hörzentrum Oldenburg gGmbH, Oldenburg, Germany
maximilian.scharf@uni-oldenburg.de

DOI: 10.1080/14992027.2024.2395416

Abstract

Objective: Audiological tests on smartphones require consistent microphone recordings across device types with a reasonable standard uncertainty (2-3dB) of the sound pressure level at the microphone. However, the calibration of smartphone microphones by the non-expert user is still an unsolved issue. We show that whistling on standardized glass bottles permits a coarse sound level calibration with an uncertainty that is smaller than the standard uncertainty of clinical audiograms (4.9dB) and enough for mHealth products.

Design: We define and test a calibration procedure with bottle-whistles for smartphones. The empirical sound pressure levels are used to calculate the mean and standard deviation of a single measurement.

Study Sample: Two uncalibrated studies with a total of 30 participants, one calibrated study with 11 participants.

Results: The mean maximal sound pressure level of 330ml Vichy-shape bottle-whistles at 50 cm distance is 92.8 ±1.6dB-SPL. The sound pressure level variation of a single measurement is 3.0dB-SPL.

Conclusions: In comparison to other possible ways of level calibration estimates for smartphones (e.g., level of own voice, level of common environmental sounds), the current method appears to be robust in background noise and easily reproducible with glass bottles of defined dimensions.

Keywords: Smartphone, Acoustic, Calibration, Mobile Health, Citizen Science

2.1 Introduction

Audiological tests on mobile devices, commonly referred to as mobile health (mHealth), are gaining popularity due to the widespread availability of devices equipped with advanced sensors and powerful processors (Swanepoel and Hall 2010). Despite this trend, these devices frequently face challenges with acoustical calibration and background noise control, significantly impacting their reliability and usefulness (Swanepoel, Clark, et al. 2010).

In general, acoustic sensors in smartphones cannot be considered as calibrated because, unlike frequency, the sound pressure level in standard international units at the microphone cannot be deduced from a recording. The reason for this is a large variation between manufacturers, protective covers, and the general condition of the recording device (Serpanos et al. 2018; Fava et al. 2016; Kardous and Shaw 2014; McLennon et al. 2019). Historically, specialized equipment with a certified calibration is used when some degree of trustworthiness of measurement data is required. This, however, is not always feasible: low socioeconomic status or disproportionate costs can make the recording of calibrated data a difficult challenge. This issue is particularly critical in applications where accurate measurements are vital, such as noise monitoring and audiological measurements with mHealth applications.

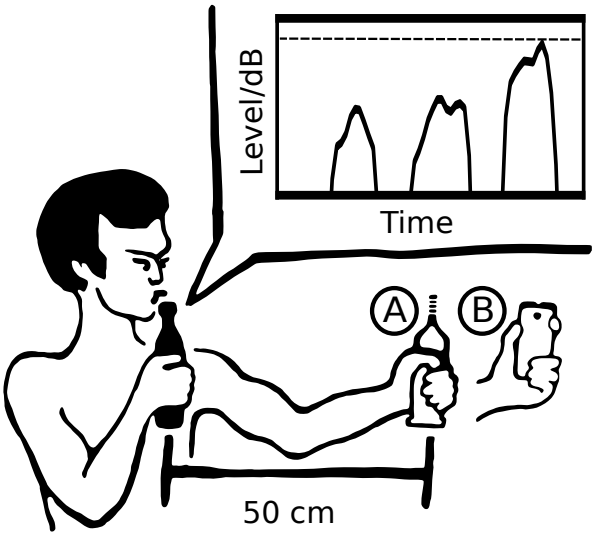


Figure 2.1: Recording setup of the study:
Participants whistle on an empty glass bottle (black) and record the level with a calibrated level meter A (class 1) or mobile device B (white). The inset shows the maximum sound pressure level around the resonance frequency of the bottle-whistle during one recording. The dashed line indicates the global maximum, which is close to the maximum possible level, independent of whistler.

This article explores the lowest-order approximation to a calibration of acoustic sensors with a single, frequency-independent level correction, suitable for microphones with flat frequency responses. We introduce an innovative calibration technique utilizing simple physical principles of "bottle-whistles", which are small glass bottles, typically under 500ml—like those used by international brewers in Vichy-shaped beer bottles. Played similarly to panpipes, these bottles provide a practical calibration tool (see figure 2.1). We define and implement the procedure, explain why it works, and report an estimate of the achievable measurement uncertainty for use in audiological measurements.

2.1.1 State of the Art

Kisić et al. 2022 reported a calibration estimation method involving speech production with a reported uncertainty of 3-4 decibel (dB) standard deviation (STD), also found by Weisser and Buchholz 2019. The relatively small uncertainty reported by Kisić et al. 2022 is only caused by variation between speakers and does not consider background noise, as speech production is dependent on background noise via the Lombard Effect (M. K. Scharf et al. 2022). Kisić et al. 2022 also showed that human rating of loudness has an uncertainty of 6–7dB STD for normal hearing subjects. In a comparison of uncalibrated level meter applications with reference sounds, McLennon et al. 2019 found deviations of the reported level from calibrated equipment of up to 15dB and concludes that uncalibrated smartphone microphones cannot be used for sound level measurements (see also Serpanos et al. 2018; Fava et al. 2016; Kardous and Shaw 2014). Gardner and Zacharias 2016 define a calibration procedure with coins in a metal can, but only report a level uncertainty estimate, which does not consider the effects of different shaking methods. Edwards, Menzel, and Puria 2005 proposed to rub two sheets of paper on a table to produce a broadband rubbing noise as a reference. However, the patent does not report the expected level uncertainty.

All calibration methods that have been proposed so far are sensitive to environmental noise, as they rely on broadband signals. In general, reference sounds for the microphone calibration are mixed with background noise. For these signals, no reliable noise suppression is available that eliminates all possible types of background noise and their influence on the calibration. This may invalidate the calibration estimate if prospective users of the method do not have access to silent environments.

Depending on the application, requirements on accuracy may vary over orders of magnitude. Uncertainties are commonly reported as a multiple of the STD σ to define the confidence interval. In this contribution, the STD is reported without any mul-

tiplication (also referred to as expansion factor [1]), corresponding to a ≈ 68% confidence interval. In audiology, the calibration with class 1 sound level meters has a STD of 0.5 dB (DIN EN 61672-1:2014-07) while a clinical audiogram has a standard error of 4.9dB (DIN EN ISO 8253-1:2011-04). This justifies a calibration precision requirement for mobile hearing health application of 3 dB std.

2.1.2 Maximum Achievable Sound Pressure Level on Bottle-Whistles

The proposed method for the definition of a reference sound level is based on a fundamental limit for the maximum sound pressure level of a bottle-whistle. For a more thorough treatment of turbulent air-stream driven Helmholtz resonators, refer to Dequand et al. 2003; Boujo et al. 2020; Auvray et al. 2016. Qualitatively, the bottle-whistle can be described by a Helmholtz resonator with resonances of finite frequency width due to damping. The resonator is driven with an independent air stream by blowing on the edge of the bottle. The air-stream produces periodic vortices that travel as pressure fluctuations inside the bottle.

The frequency of the fluctuations f is connected to the velocity of the air-stream U via the dimensionless Strouhal number ($\mathrm{S_r}$):

$$f = \frac{S_r \cdot U}{W} \tag{2.1}$$

Here, W is the width of the bottle-mouth. The excitation amplitude for driving the Helmholtz resonator is approximately proportional to the air-stream velocity as well. This means that for a given $\mathrm{S_r}$, only one air-stream velocity U and thus excitation amplitude matches the resonance condition of the bottle. At resonance, the maximum sound pressure level is produced. $\mathrm{S_r}$ depends on the geometry of the bottle mouth as well as the shape of the air-stream. Since the sound pressure level has a maximum for all $\mathrm{S_r}$ and assuming that $\mathrm{S_r}$ can only vary over a finite range with the shape of the air stream, there must exist a maximum possible sound pressure level that can be produced by whistling on a bottle (Dequand et al. 2003; Boujo et al. 2020; Auvray et al. 2016). The existence of this maximum level is enough to define a consistent reference level.

The remaining question is whether humans who blow on a bottle can optimize their whistling technique sufficiently well to achieve this maximum? It is not necessary to produce stable sounds to generate the maximum possible sound pressure level, as short durations of optimal whistling conditions are sufficient to produce the required sound pressure level (see figure 2.2).

2.1.3 Resonance Frequency of Bottle-Whistles

The Resonance Frequency f_r of a bottle-whistle can be approximated by [2]:

$$f_r = \frac{1}{2\pi} c \sqrt{\frac{A}{l' V}} \tag{2.2}$$

c is the speed of sound in air, A is the area of the bottle opening, and V is the volume of the bottle. The effective length of the bottleneck l' depends on the shape of the bottle and the opening, and has been calculated for some geometries. Here, we use $l' = l + 0.2 \cdot \sqrt{A}$ where l is an estimate of the bottleneck length and $0.2 \cdot \sqrt{A}$ is the end correction term of an open pipe with thin walls according to Levine and Schwinger 1948. For an introduction to the end correction, see Pierce 2019. For calculations that take the shape of the bottle into account, see Alster 1972.

2.2 Materials and Methods

The diagram in Figure 2.2 shows the working principle of the proposed method. An implementation of this procedure in the scripting language Python can be found under Maximilian Karl Scharf 2023 together with examples.

During the recording, the bottle has to be held in one hand. The microphone is held in the other hand at an arm's length distance (≈50 cm) in the same horizontal plane. The influence of variations in arm length is thus also included in our reported uncertainty. To account for the directionality of the microphone, the microphone should be directed in the same way during calibration and measurement for which the calibration is required. Occlusion by the hand has to be minimized. Asking for three continuous sounds is an incentive for the participant to practice before the measurement.

The full spectrum of the recorded signal is analyzed, and the resonance frequency is calculated. If it matches the expected resonance frequency of the bottle (see Equation 2.2), the procedure continues. A mismatch between measured and expected frequency can be caused by remaining liquid in the bottle, an unexpected bottle shape, or higher harmonics of the bottle volume. The recording should be repeated in case of a frequency mismatch. Further processing is done with the short time Fourier transform (STFT) of the recording, to resolve the time-dependent level of the bottle-flute,

[1] see Equation 1.40

[2] see derivation in section 1.6.3

which changes over time and eventually reaches the optimum, like in figure 2.1.

Next, the power in the frequency interval around the resonance frequency is calculated from the STFT. The filter width and the time window of the STFT are open parameters of the procedure. Their influence on the outcome of the calculation (which should be small) is discussed in the results section. One effect of the band-pass filter is that background noise is mostly removed by this procedure. Accordingly, the filter width should be as small as possible.

Finally, the maximum value of the resonance level is subtracted from the empirical 92.8dB-sound pressure level (SPL). The resulting number is an estimate of the calibration offset in dB. The 92.8 ±1.6dB-SPL are the mean value of the empirical maximum resonance levels of studies with human subjects, which are presented in the results section.

The procedure was tested with the pooled data of three studies using the 330ml Vichy-shape bottle according to DIN 6075-1:2021-06; DIN 6075-2:2021-06 with a 26 H 180 crown finish according to DIN EN ISO 12821:2020-02:

1. Preliminary study of 11 participants (mixed gender, 20-30 years old) with an uncalibrated recording device with a flat frequency response (DR-100 MK2) at a distance of 75±5 cm between bottle opening and microphone in an outside environment (average background noise was 76dB-SPL).

2. Follow-up study of 19 participants (mixed gender, 20-65 years old) with an uncalibrated smartphone microphone (participant facing backside of device) at arm-length distance between bottle opening and microphone in an outside environment (average background noise was 75dB-SPL).

3. Study of 11 participants with a calibrated microphone setup, with measurements in the living room of the participants (average background noise was 57dB-SPL) at a distance of 50 cm between the bottle opening and the microphone. The calibration error of the measured sound pressure level with the calibrated microphone was conservatively estimated to be 0.5dB STD (standard for class 1 level meter (DIN EN 61672-1:2014-07)). The participants were three women and eight men aged between 23 and 28 years (mean=25).

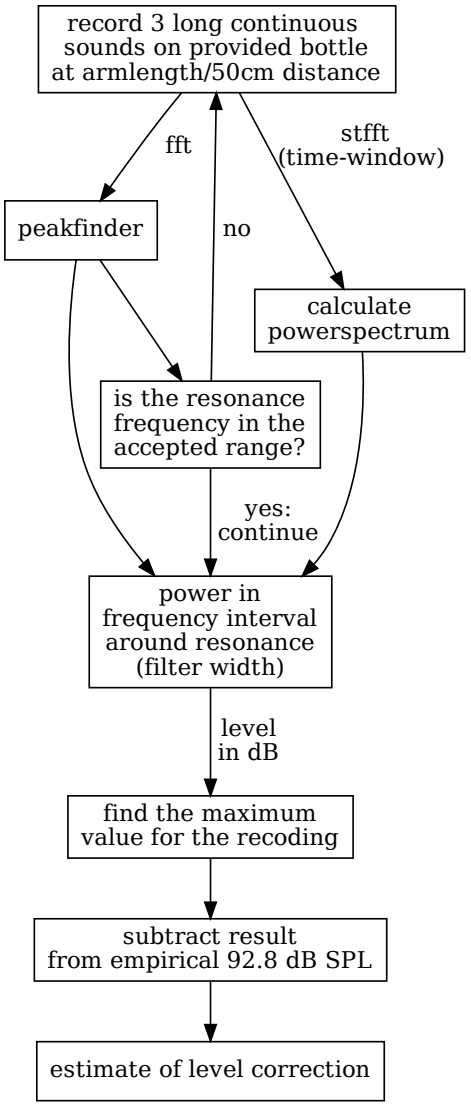


Figure 2.2: Working principle of the proposed method

For the measurements with uncalibrated equipment, the microphone gain was kept constant during the experiment to make a post-hoc calibration possible. For the post-hoc calibration, the calibration offset was calculated for the calibrated study and applied to the uncalibrated measurements. This also removed any influence of the different distances (75cm in study 1 vs. 50cm in the calibrated study) of the microphone to the bottle between the studies. The ensemble of calibrated and post-hoc calibrated measurements (figure 2.3) was used to calculate the STD of the distribution. This way, the mean of 92.8dB-SPL was calculated only from the third, calibrated study, while the STD of 3.0dB was calculated from all measurements.

2.3 Results

The pooled results over the three studies are shown in Figure 2.3. The mean peak level of the calibrated study (3) was 92.8 ±1.6dB-SPL. With this, the means of the uncalibrated studies (1 & 2) were scaled with a post-hoc calibration to match the mean of the calibrated data. The STD of the pooled distribution (1, 2 & 3) was used to calculate the uncertainty of the calibration offset for the 330ml Vichy-shape bottle. The generated frequencies were in the range of 161.5-176 Hz. These agree with the approximated resonance-frequency according to Equation 2.2 for the bottleneck length $l = 1/3 \cdot 23$ cm and an opening diameter of 18mm (see DIN 6075-1:2021-06; DIN 6075-2:2021-06; DIN EN ISO 12821:2020-02).

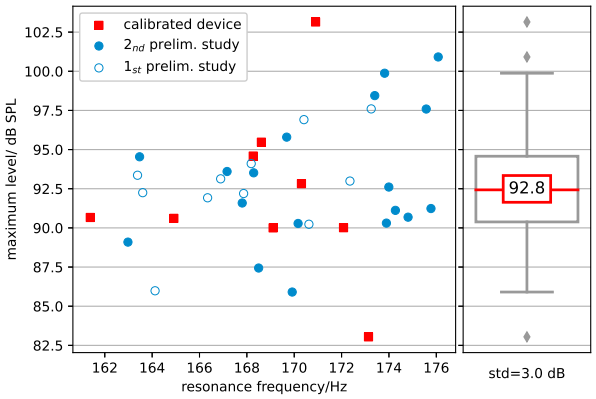


Figure 2.3: Maximum sound pressure level of bottle-whistle resonances for all studies:
left panel: Level as a function of resonance frequency for a 330ml Vichy-shape bottle with a time window of 35 ms and 120 Hz filter width. Each point corresponds to a recording of a different whistler. The calibrated study (red squares) was used to adjust the level of the uncalibrated studies (blue circles) with a post-hoc calibration. **right panel:** Box plot of the level across all studies.

The open parameters of the calibration procedure (filter width and time window of the STFT) should not influence the measurement outcome. The influence of the two parameters on the data of the calibrated study is shown in Figure 2.4. The maximum level was calculated for several time window sizes and filter widths according to our procedure, with a feed-forward of 1/2 time window. The maximum level was invariant to changes in filter width above 60 Hz for time windows above 35 ms. Longer time-windows led to smaller levels, as the time-window was longer than the duration of the sounds of maximum sound pressure levels.

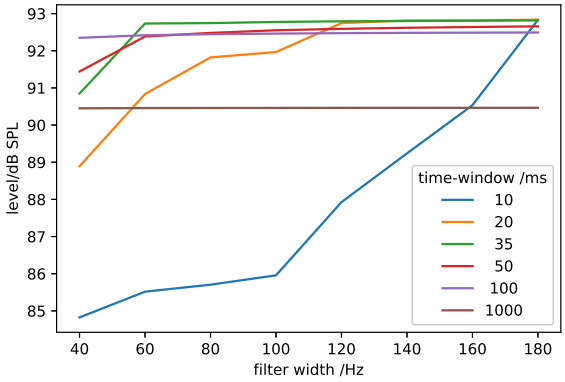


Figure 2.4: Invariance of the mean maximum sound pressure level:
The result does not depend on the filter width for time-windows above 35 ms. Longer integration times reduce the maximum level as expected.

Testing the distribution of maximum levels in figure 2.3 with D'Agostino and Pearson's normality test showed that the zero hypothesis of a normal distribution cannot be rejected. Under the assumption that the distribution is indeed normal, we calculated the pooled uncertainty of the average calibration offset σ_{calib} for N independent measurements with:

$$\sigma_{\text{calib}} = \frac{3.0 \text{ dB}}{\sqrt{N}}$$

This means that for 4 recordings of different people whistling on the same shape of bottle, the 68% confidence interval for the calibration offset is equal to 1.6dB.

2.4 Discussion

A method for a first-order approximation of a calibration offset with a bottle-whistle was presented and evaluated with 41 participants. It was shown that one single measurement leads to a standard uncertainty of 3.0dB STD, which is smaller than the uncertainty for entirely uncalibrated measurements (McLennon et al. 2019). A summary of microphone calibration uncertainties is shown in Table 2.1. These calibrated microphones can then be used to further calibrate loudspeakers of mobile audiological tests (see section 2.4.3). The presented method is not suitable to replace a calibration with proper equipment, if available.

2.4.1 Uncertainties

The directivity of the bottle-whistle has not been measured within this study, as it was fixed to the forward direction with no elevation (figure 2.1). Additionally, the directivity of the microphone may have an effect if not controlled properly. Any recording after the calibration has to maintain the

calibration uncertainty type	how to mitigate	handled by procedure
Different reverberation time and modes between rooms	Choose a semi-anechoic environment for the recording, e.g, a large patch of grass, rooftop, fresh snow	No
Broadband noise during the recording, e.g, wind, talking individuals, open window	Apply a filter, that cannot interfere with the calibration signal	Yes
Unknown shape of the frequency response of the microphone	Use several reference frequencies	Possible
Directionality of the microphone	The microphone direction to the target has to stay fixed after the calibration	No
Variation in skill for whistling on a bottle	The calibration estimate has a STD of 3dB. Multiple measurements with different individuals reduce the uncertainty of the mean.	Yes
Liquid is left in the bottle	Provide a clear explanation of the procedure or check if the resonance frequency is in the expected range	Yes

Table 2.1: sources of calibration uncertainties and ways of mitigation with this procedure

direction of the microphone to the target.

Moreover, we do not provide any measurement uncertainty, which is caused by the transfer function of the microphone. The microphones of smartphones are not guaranteed to have a flat frequency response and may differ substantially for low frequencies. One way to control this may be to record the bottle-whistle at the fundamental frequency as well as the higher harmonics, or to use small bottles with higher resonance frequencies, for which the same principle of maximum possible sound pressure level applies. However, to enable a high applicability of the current method, a bottle size and shape have been selected that have a very high usage among German breweries. Internationally, there is no protected trademark on the Vichy-shape, and standards may differ from the standards that were cited in this study. In an application of this method, the name of the beverage and country of origin would have to be specified and compared to a database of reference levels. Hence, a compromise has to be made between a broad applicability and a loss in precision for frequencies other than the resonance frequency.

2.4.2 Robustness Against Noise

In comparison to other methods (see section 2.1.1), the presented calibration procedure is more robust against background noise due to the narrowband reference sound. Frequencies other than the resonance frequency are not considered for the calibration estimation. Thus, the method may be more suitable for the application in everyday environments, such as an outside patch of grass, which is favorable because of the low reverberation. The reported reference value of 92.8 $\pm$1.6dB-SPL for the average maximum sound pressure level is only valid for average living room conditions. For highly reverberant environments, this value may be larger and depend on the resonance modes of the room.

2.4.3 Prerequisite for Calibrated Headphones

We describe a method to calibrate microphones with a measurement uncertainty below the 4.9dB uncertainty of clinical audiograms (DIN EN ISO 8253-1:2011-04). However, most audiological experiments require calibrated sound sources, i.e., headphones instead of microphones. The calibration of headphones with calibrated microphones can be seen as a secondary step and was not the scope of this study.

A simple calibration method for headphones using a calibrated microphone could involve subjective loudness matching of narrowband signals in both free field and headphone presentation, similar to measuring free-field equalization curves (DIN EN 60268-7:2021-08). The microphone measures the sound pressure level of the free-field narrowband signal, while a person has to match the loudness of the headphones outputting the same signal. Denk et al. 2021 reported a STD of 2.5-5.5dB for loudness matching of microphone and free-field presentation with third octave band noise. The just noticeable difference (JND) in sound pressure level for pure tones is about 1dB according to Zwicker 1952. Broadband signals, such as speech, should not be used, as this may lead to a bias in the loudness comparison, known as "Effect of missing 6dB" (Völk and Fastl 2011; Denk et al. 2021; Kisić et al. 2022).

If measurement uncertainties of microphone-free-field equalization and headphone calibration are uncorrelated, we can estimate the total level uncertainty of a headphone after calibration with

quadratic addition of the STDs:

$$\sigma^2_{\text{headphone}} = \sigma^2_{\text{levelestimate}} + \sigma^2_{\text{vichybottle}} + \sigma^2_{\text{JND}} \quad (2.3)$$

$\sigma_{\text{vichybottle}} = 1.6\text{dB}$ is our measurement uncertainty of the empirical maximum sound pressure level of 92.8dB for the 300ml Vichy-shape bottle shape in this study. If we assume flat transfer functions and use the sound level estimate with bottle-flutes with an uncertainty of $\sigma_{\text{levelestimate}} = 3.0\text{dB}$, we can expect a level uncertainty of the headphone of $\sigma_{\text{headphone}} \approx 4.2 - 6.5\text{dB}$.

2.5 Conclusion

This study demonstrates the feasibility of employing standardized glass bottles, termed "bottle-whistles," for calibrating microphones in mobile devices. This method provides an alternative to traditional, expensive calibration techniques.

The procedure not only adheres to the required precision for clinical audiological tests but also incorporates everyday objects in scientific measurement, making it both accessible and practical. With a standard uncertainty of 3.0 decibel (dB)-sound pressure level (SPL) for individual measurements, this method falls well within the acceptable range for basic diagnostic purposes.

Furthermore, the robustness of this method in noisy environments highlights its potential for widespread application. This technique could significantly broaden the scope of reliable, mobile health diagnostics, reaching populations in varied socioeconomic settings without the need for expensive equipment.

Future studies should include a calibration of headphones with calibrated microphones. Furthermore, our assumption of an approximately flat frequency response of smartphone microphones for frequencies above the resonance frequency should be validated with a representative sample of smartphones. The maximum sound pressure level with other bottle shapes should be collected to apply this method in regions with different bottle shape standards.

We conclude that it is possible to estimate a microphone calibration offset with standardized glass bottles that are used as whistles. The method can handle various sources of measurement error during calibration (see table 2.1). It is a precise and playful calibration procedure for acoustic measurements with smartphones. The level uncertainty with a standard deviation (STD) of 3.0dB-SPL for a single measurement is small enough for meaningful audiological tests in mobile health (mHealth) products.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

MKS, RH, MS, and BK contributed to the development of the procedure. RH and MS organized the study with calibrated equipment. MKS is responsible for the idea and implementation. MKS wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was funded by: the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy –EXC 2177/1 - Project ID 390895286 - "Hearing4all" and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) –Project ID 352015383 – "SFB 1330 A5 Model-based diagnostics and hearing aid fitting in complex acoustic scenarios: Perceptive Principles, Algorithms and Applications".

Acknowledgments

The authors are thankful for the support of their colleagues and the Excellence Cluster "Hearing4all". The authors thank the two anonymous reviewers and Vani Sundaram for helpful comments on the manuscript.

Data Availability Statement

The procedure of this study is available as a command-line tool under Maximilian Karl Scharf 2023.

Ethics Declarations

The study with calibrated equipment was performed by the Hoerzentrum Oldenburg, supervision of ethics committee Drs.EK/2021/031-03. All participants consented to being recorded while whistling.

Chapter 3:

A Consistency Measure for Psychometric Measurements

reprint from
Journal of the Acoustical Society of America

Maximilian Karl Scharf[1], *Anna Warzybok*[1], *Sabine Hochmuth*[2], *Birger Kollmeier*[1]

[1] Medical Physics and Cluster of Excellence "Hearing4all",
Carl von Ossietzky Universität Oldenburg, Germany.
[2]Department of Otolaryngology, Head and Neck Surgery, Carl von Ossietzky Universität Oldenburg,
Germany. maximilian.scharf@uni-oldenburg.de

DOI: 10.1121/10.0039426

Abstract

Adaptive tracking procedures in psychophysics may produce erroneous, "untypical" results and non-converging tracks due to, e.g., inattention of the test subject or external disturbances. This paper presents a multi-state psychometric model, which is used to rate the outcome of psychometric measurement procedures with a consistency measure. The consistency measure may be used for a post-hoc, automated consistency estimation for any psychometric measurement procedure that can be modeled with a sigmoid psychometric function. The model calculates the log likelihood difference between single and two interleaved psychometric functions, potentially underlying a recorded adaptive track.

A binary classifier was tested with a range of candidates for consistency measures with simulated, inconsistent tracks and expert ratings of empirical tracks. The proposed consistency measure was identified as the best candidate to classify inconsistent tracks, while expert ratings were best predicted with the spectrum of the stimulus level, which is shown to be a suboptimal predictor of consistency. A threshold of the proposed measure for the German matrix sentence test is 10 to test for inconsistency, with a sensitivity of 60% and a specificity of 80%.

Keywords: psychometrics, attention, adaptive procedures, multi-state listener, quality control, consistency measure

3.1 Introduction

The assessment of psychophysical thresholds (such as, e.g., pure tone detection or the speech recognition threshold (SRT)) is a fundamental task in psychometrics with many applications in hearing science or audiology. Numerous psychometric measurement procedures have been defined with the goal of optimizing the reliability of the measurement for a given time spent on the task (Treutwein 1995; Leek 2001).

In the following, we define the term "consistency" of a psychometric test to describe how much the experimental results match the underlying assumptions of the experimenter. Traditionally, the consistency of the outcome of a psychometric test is considered a direct consequence of the applied procedure, i.e., the occurrence of inconsistent measurements is minimized with the most appropriate measurement procedure. However, the consistency is typically not checked and quantified during a listening experiment. It remains the task of the experimenter to assess the consistency of the outcome of the procedure he or she has chosen. If the consistency is unsatisfactory, the measurement procedure is altered, e.g., with a retest, break, termination, etc. However, this consistency rating remains highly subjective and prone to mistakes, depending on, e.g., the visual representation of the data or the experience of the experimenter with the testing procedure.

We describe a post-hoc model that can be used to assess the consistency of measurement data (in the following called track) with the common assumption of a single psychometric function. The model can be used to monitor the consistency of a psychometric measurement with an objective, easy-to-interpret scalar value. The aim of this paper is not to define a new type of adaptive procedure that is more robust in difficult listening conditions (see, for instance, Xu et al. 2024), but rather to quantify the consistency of the respective test results. The model is implemented for, but not limited to, the matrix sentence test (Kollmeier, Warzybok, et al. 2015) as a Python package and supplements this work.

3.1.1 Suboptimal Measurement Conditions

Fully remote and self-administered e-health systems with audiological tests (Swanepoel, Clark, et al. 2010; Ooster et al. 2020; Swanepoel and Hall 2010) require robust and fast psychometric measurements. In uncontrolled environments, the outcome of these measurements is subject to internal and external variation:

- Internal variations are changes in attention of the test subject, e.g., messaging services on the same device causing a slip in attention (short-term inattention according to Xu et al. 2024) or exhaustion with permanently decreased attention (long-term inattention, according to Xu et al. 2024).

- External variations are sudden changes in the performance of the measurement equipment during an experiment, like a forced change in volume or a change of background noise. These continuously alter the measurement condition.

The two types of variation suggest that under suboptimal measurement conditions, the psychometric function, connecting response probability to the stimulus, differs from the standard sigmoid function. This may result in a shift of the threshold after a lasting change in attention during the measurement (Bianco et al. 2021) or a nonstationary psychometric function Frund, Haenel, and Wichmann 2011.

3.1.2 Problem Statement

In the following, we define the outcome of a psychometric measurement as a sequence of stimuli ($\mathbf{s}$) together with a sequence of responses ($\mathbf{r}$), called track, with N numbered values of presentation and corresponding response, called trial, during the test:

$$\mathbf{s} = \{s_0, s_1, \ldots, s_{N-1}\}, \quad s_n \in \mathbf{s} \tag{3.1}$$

$$\mathbf{r} = \{r_0, r_1, \ldots, r_{N-1}\}, \quad r_n \in \mathbf{r} \tag{3.2}$$

Every trial may consist of several test items in immediate succession. Each corresponds to a binary decision together with its evaluation (e.g., correctly heard/classified (hit) vs. not heard or incorrectly classified (miss)).

The reason for this distinction between trial and test item is that the responses to the test items inside a trial are given together, and that the time between test items is much shorter than the time between trials. The transition probability for the change of the psychometric function may differ substantially between trials and between the test items inside a trial. Our definition assumes that the psychometric function only changes between trials and that the transition between test items can be neglected.

The general assumption is that the response as a function of the corresponding stimulus is scattered around a sigmoid function. This function is referred to as the psychometric function.

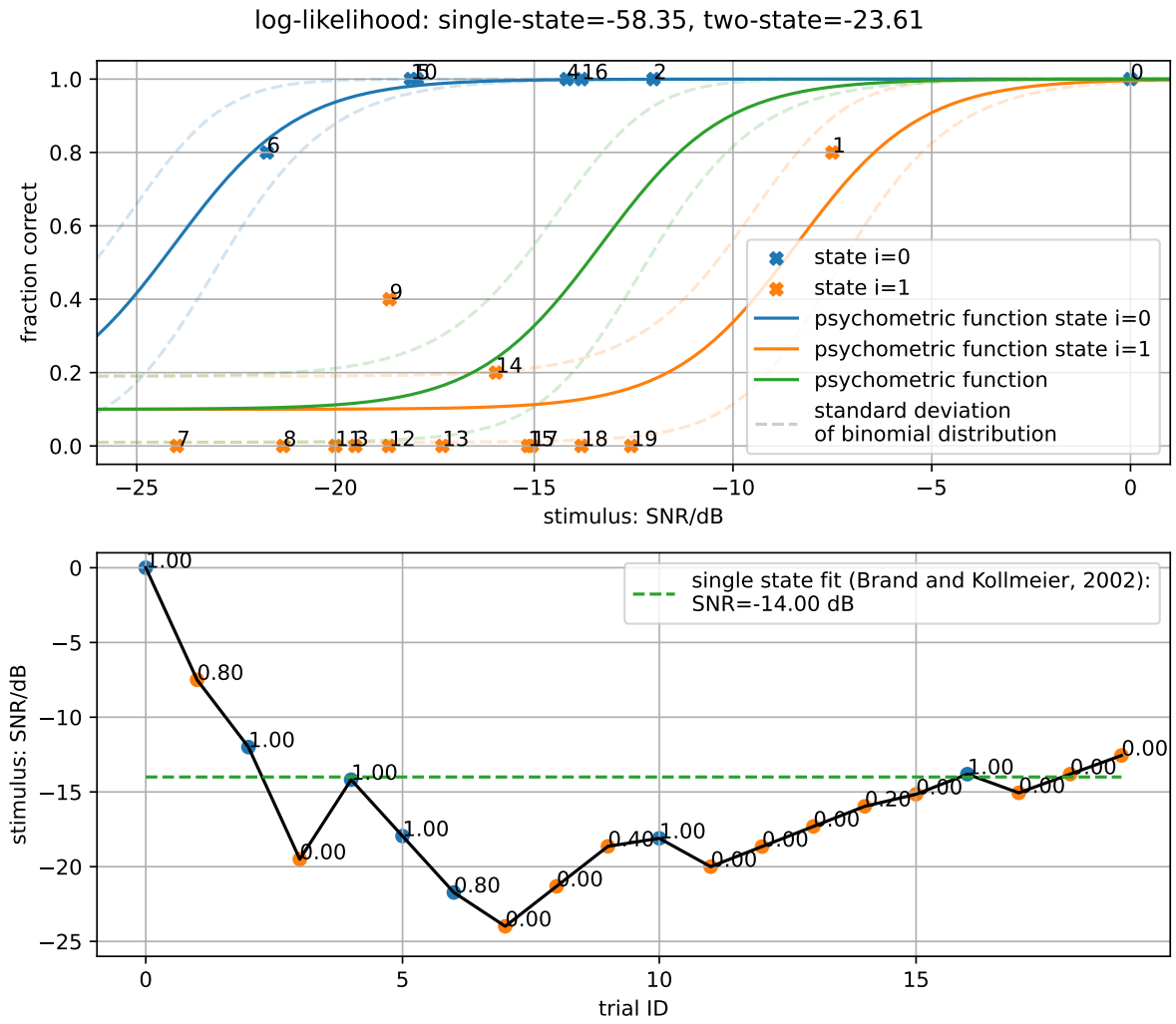


Figure 3.1: Adaptive track of an inconsistent psychometric measurement:
The most likely single- and two-state models are indicated in green and blue/orange, respectively. The **upper panel** plots the sequence of responses (**r**) of the test subject against the sequence of stimuli (**s**). The index n of the trial is written next to each entry. Psychometric functions are indicated with solid lines; the standard deviation of the binomial distribution is indicated with broken lines. The **lower panel** shows the same measurement as a function of n with the response indicated next to the entry. Here, the broken line shows the estimated threshold (SRT) according to the single-state psychometric function (Thomas Brand and Kollmeier 2002).

As an example, consider the track of a matrix sentence test (Kollmeier, Warzybok, et al. 2015) in Figure 3.1. The closed matrix sentence test contains 20 sentences and a test-specific masking noise. Here, a trial with index n is a sentence in noise at a signal-to-noise level of s_n. The trial consists of five syntactically fixed, random words (i.e., test items of the sentence), which have ten alternatives each. This way, the rate of correctly answered trials **r** is measured in discrete steps of 0.2. The guess rate, which is the chance of a random, correctly identified trial, is equal to 0.1 (see appendix). Because r_n is the average of five Bernoulli experiments, it is a sample of a binomial distribution around the value of the psychometric function at the corresponding s_n (see upper panel of Figure 3.1).

At the beginning of a matrix test ($n = 0$), the first trial is presented to the test subject with the stimulus level s_0, and the initial response r_0 is recorded. The measurement procedure defines the next stimulus s_{n+1}, conditioned on all previous stimuli and responses. The procedure repeats until an ending criterion is met (see upper panel of Figure 3.2). The track (**s** and **r**) of the matrix test is stored for later analysis.

We need a consistency measure of the track that fulfills the following requirements:

1. It has to be a scalar measure and thereby easy to interpret.

2. The measure should indicate how well the track matches the assumption of a single psy-

chometric function.

3. The measure should be fast and easy to use
4. If properties of the expected psychometric function (e.g., guess rate, average slope) are known, these should be included in the calculation to improve the results.

Because of the way this measure is defined, we will refer to it as the single psychometric function goodness of fit measure (SiPGOF).

3.2 Methods

3.2.1 Model Definition

In this study, the subject of a psychometric measurement is modeled by a Markov process. Each Markov state is associated with a psychometric function (see lower panel of Figure 3.2). The psychometric function is commonly described with the four-parameter transformed logistic model (see, for example, Doire, Brookes, and Naylor 2017). These four parameters are the lapse rate (λ), guess rate (γ), and α as well as β, which determine the threshold (SRT for matrix sentence tests) and slope at the threshold:

$$\Phi(s) = \gamma + \frac{1-\gamma-\lambda}{1+\exp\left(-\beta(s-\alpha)\right)} \tag{3.3}$$

The simplest model of this type is a single-state with only one psychometric function and no state transitions, with a total of 4 degrees of freedom. We refer to this as the single-state model. This model is equivalent to the model of Thomas Brand and Kollmeier 2002.

With two states, the model has two psychometric functions (one for each state $i = 0$, $i = 1$), the probability distribution of the first state π, and the 2x2 transition matrix A. A describes the probability $p(i \to i')$ for all possible transitions of the current state i to the next state i' (see lower panel of Figure 3.2):

$$\mathrm{A} = \begin{pmatrix} p(0\to 0) & p(0\to 1) \\ p(1\to 0) & p(1\to 1) \end{pmatrix} = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} \tag{3.4}$$

These two states may be interpreted as a state of concentration and a state of distraction, and we refer to this in the following as a two-state model. It has 8 degrees of freedom if the two psychometric functions only differ in threshold (i.e., two thresholds, one slope, γ, λ, transition matrix with two degrees of freedom, π with one degree of freedom). A maximum of 11 degrees of freedom is possible if the two psychometric functions differ in all four parameters.

Note that it is crucial to keep the number of free parameters of the model as small as possible, as overfitting might occur for short tracks. For the matrix sentence test considered here, a track consists of $N = 20$ trials. For this reason, we chose the following strict boundary conditions: the slope was fixed to the average slope of the considered matrix test, λ was set to zero, and γ was set to the theoretical 0.1 for closed tests. This led to a two-state model with five degrees of freedom and a single state model with one degree of freedom.

Model Fit to Tracks

The single-state and two-state versions of the model were used to calculate the maximum likelihood across the model parameters for a given track. For this, we initialized the free parameters of both models with default values and optimized for the likelihood. The optimization worked as follows:

For each state i of the model, with corresponding psychometric function Φ_i: given a response r_k to a stimulus s_k, we compute the likelihood, or emission probability, $p_{i,k}$ that r_k was the result of a Bernoulli experiment with $H = r_k \cdot T$ hits out of T test items ($T = 5$ words in case of the matrix test):

$$p_{i,k} = \binom{T}{H} \left(\Phi_i(s_k)\right)^H \cdot \left(1-\Phi_i(s_k)\right)^{T-H} \tag{3.5}$$

The multi-state model parameters are optimized using the expectation maximization algorithm (EM)/Baum-Welch algorithm Bishop 2019, Ch. 13.2; Rabiner 1989. The likelihood of the model is calculated with the forward algorithm. The free parameters of the psychometric functions (model states) are updated using a standard function in the Scipy optimization library Virtanen et al. 2020. If properties of the expected psychometric function are known, these are included as a boundary condition. All computational details are documented in the code supplement M. K. Scharf 2025. The resulting maximum log-likelihood ($\mathrm{l}_{\max}$) quantifies how well the model can describe the track, assuming the local optimum is equal to the global optimum:

$$l_{\max} = \max_{\gamma,\lambda,\alpha,\beta,\ldots} (l) \tag{3.6}$$

We define SiPGOF as the $\mathrm{l}_{\max}$ difference between the single- and two-state models:

$$\mathrm{SiPGOF} = l_{\max,\mathrm{two-state}} - l_{\max,\mathrm{single-state}} \tag{3.7}$$

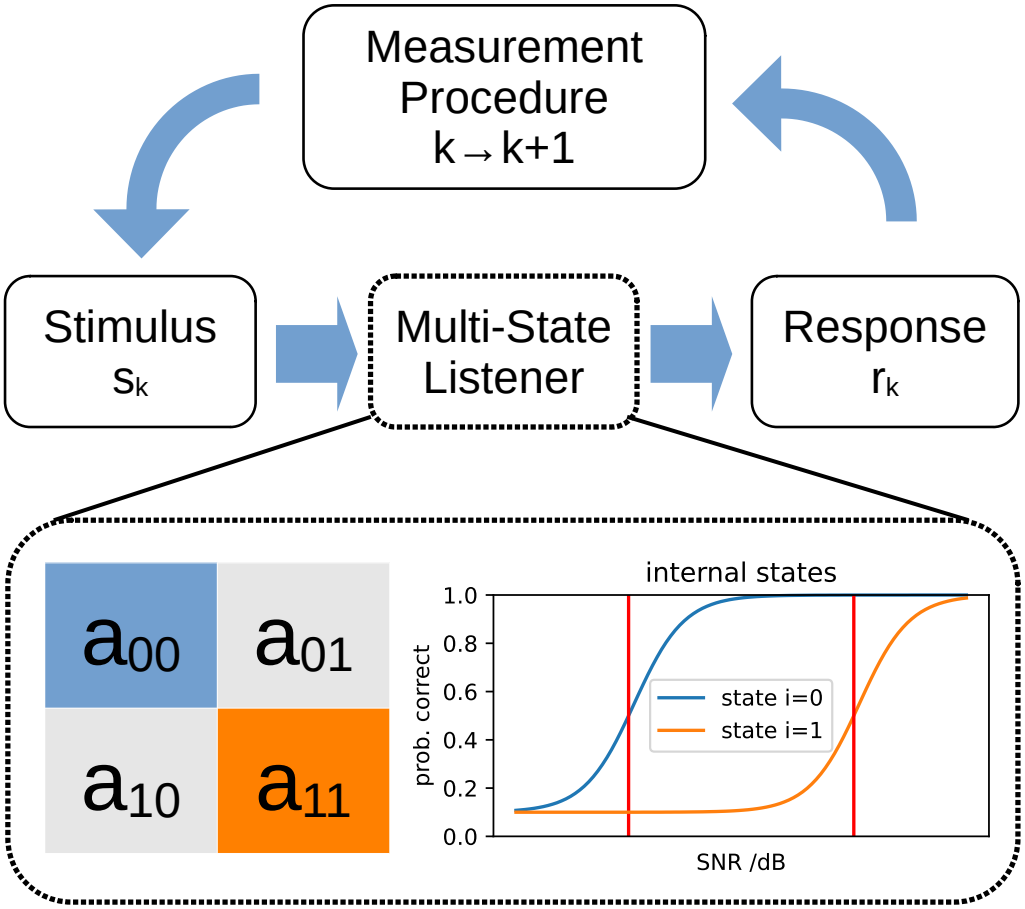


Figure 3.2: Model of responses to an adaptive measurement procedure:
top: Workflow of the simulation of inconsistent tracks. From the previous stimuli s_k and responses $r_0, r_1, r_{...}, r_k$, the adaptive procedure defines a new stimulus s_{k+1}, which is presented to the subject to trigger the next response r_{k+1}. The loop repeats until an ending criterion is met. **bottom:** A subject may have several internal states with different psychometric functions (bottom right example for the matrix sentence test). A Markov process models the state transitions with a transition matrix (bottom left). The row index is the current state, and the column index is the next state (see Equation 3.4). Depending on the current state, the probability of a correct response is given by the psychometric function of that state.

This means that SiPGOF assigns a value to a track:

$$\text{SiPGOF} : (\mathbf{s}, \mathbf{r}) \to \mathbb{R} \tag{3.8}$$

Small values indicate agreement with a single-state model, and large values indicate good agreement with the two-state model.

Monte Carlo Simulation of Tracks

For the first part of the study, a simplified model (with fewer free parameters than for fitting) was used to simulate inconsistent adaptive tracks of the matrix sentence test. These tracks were used to test candidates for consistency measures. A fully controllable model with known internal states made it possible to quantify consistency (i.e., the mismatch between the estimated psychometric threshold and the threshold of the model, referred to as known inconsistency (KIC), see definition below) and compare it with candidates for a measure of consistency.

We initialized the model with psychometric functions Φ_i from literature values. A was parameterized by a single variable $a = a_{00} = a_{11}$ (i.e., the probability to stay in the same state):

$$\mathrm{A} = \begin{pmatrix} a & 1-a \\ 1-a & a \end{pmatrix} \tag{3.9}$$

This was done to reduce the number of model parameters. It reflects the assumption that transitions between states are equally likely, an effect that is commonly observed during measurements as ”losing the cue”, i.e., a situation where a subject continuously stays distracted with probability

$a_{01} = a_{10} = 1 - a$ after a seldom occurring distraction. The initial state i of Eqs. 3.4 was initialized randomly to remove any bias towards one of the states in the Markov chain.

Provided with an initial stimulus s_0, the model returned the probability of a correct response for each item of a trial according to the psychometric function Φ_i of the current internal state i (see Eqs. 3.3).

With this probability, the response r_0 of a subject was simulated (Monte Carlo). r_0 was then fed as input to the measurement procedure to compute the next stimulus s_1 (see upper panel of Figure 3.2). This repeated until the procedure terminated with $N-1 = 20$ trials. The result was a simulation of an inconsistent track for which the hidden states were known.

$50 \cdot 10^3$ artificial tracks were generated this way with average psychometric slopes for Cantonese and US-English matrix sentence tests of 14.25 %/ decibel (dB) (Kollmeier, Warzybok, et al. 2015; Hu, Hochmuth, et al. 2022), λ=0, threshold differences of the two psychometric functions (in steps of 1dB in the range from 0-4dB) and parametrization of the transition matrix a (in 10 equidistant steps for values between 0-1) of the transition matrix. The initial stimulus value of the adaptive procedure was set to $s_0 = 0$dB.

The generated database contained fully consistent (threshold difference = 0dB or $a = 1$) and highly inconsistent tracks (threshold difference > 0dB and $a \neq 1$) to test candidates for consistency measures. The internal states of every simulated track were used to define the known inconsistency.

Known Inconsistency of simulated tracks

We define the scalar known inconsistency (KIC) of a matrix test as the smallest absolute threshold difference $\Delta\mathrm{SRT}_i$ between the SRT of either one of the two states i and the estimated $\mathrm{SRT}_{\mathrm{estim}}$ of a simulated track:

$$\Delta\mathrm{SRT}_i = \mathrm{SRT}_i - \mathrm{SRT}_{\mathrm{estim}} \tag{3.10}$$

$$\mathrm{KIC} = \min_{|\Delta\mathrm{SRT}|} \Delta\mathrm{SRT}_i \tag{3.11}$$

The estimated threshold is calculated according to Thomas Brand and Kollmeier 2002, which fits a single psychometric function to the entire track.

Identifying the smallest absolute threshold difference with the KIC was important, as the simulation included consistently distracted tracks (large a in Eqs. 3.9). This way, all consistent tracks had a small absolute KIC, while larger values corresponded to a higher impact of two psychometric functions on the track.

KIC can take on both positive and negative values. The sign indicates whether the SRT is over- or underestimated because of an inconsistency in the track. If we find a consistency measure that is a good predictor of the signed KIC, we can correct inconsistent measurements. To find out if this is possible, we compared several candidates for consistency measures with both signed and unsigned (i.e., the absolute value of the KIC).

3.2.2 Candidates for a Consistency Measure

A set of candidates for consistency measures was defined and compared with the KIC:

- Standard deviation σ after the second reversal point n_{rev}

$$\begin{aligned} \sigma_r &= \mathrm{SD}\left(\{r_{n_{\mathrm{rev}}}, \ldots, r_{N-1}\}\right) \\ \sigma_s &= \mathrm{SD}\left(\{s_{n_{\mathrm{rev}}}, \ldots, s_{N-1}\}\right) \end{aligned} \tag{3.12}$$

- slope m of a linear interpolation of the track after the second reversal point

$$\begin{aligned} m_r &= \mathrm{slope}\left(\{r_{n_{\mathrm{rev}}}, \ldots, r_{N-1}\}\right) \\ m_s &= \mathrm{slope}\left(\{s_{n_{\mathrm{rev}}}, \ldots, s_{N-1}\}\right) \end{aligned} \tag{3.13}$$

- resonance frequency f of the track after the second reversal point.

$$\begin{aligned} f_r &= \operatorname*{argmax}_f \left(|\mathscr{F}\{r_{n_{\mathrm{rev}}}, \ldots, r_{N-1}\}|\right) \\ f_s &= \operatorname*{argmax}_f \left(|\mathscr{F}\{s_{n_{\mathrm{rev}}}, \ldots, s_{N-1}\}|\right) \end{aligned} \tag{3.14}$$

  $\mathscr{F}$ stands for the discrete Fourier transformation. $|\ldots|$ is the absolute value.

- slope m' of a linear interpolation to the absolute value of the spectrum of the track after the second reversal point.

$$\begin{aligned} m'_r &= \mathrm{slope}\left(|\mathscr{F}\{r_{n_{\mathrm{rev}}}, \ldots, r_{N-1}\}|\right) \\ m'_s &= \mathrm{slope}\left(|\mathscr{F}\{s_{n_{\mathrm{rev}}}, \ldots, s_{N-1}\}|\right) \end{aligned} \tag{3.15}$$

- Bootstrapped consistency measure: The standard deviation of the threshold distribution generated by a bootstrapping procedure.

- Single psychometric function goodness of fit measure (SiPGOF)

All candidates for a consistency measure, except SiPGOF and the bootstrapped consistency measure, were calculated for the part of the track that came after the second reversal point of the track. This was done to make the consistency measure less sensitive to the initial stimulus level of the measurement procedure.

For the bootstrapped consistency measure, the threshold distribution of a maximum likelihood fit with a single psychometric function according to Thomas Brand and Kollmeier 2002 was calculated from 100 random samples of the original track. Each sample consisted of 20 random draws (with replacement) from the 20 trials in the original track. The standard deviation of the resulting distribution of thresholds was evaluated. It indicates how consistently a single psychometric function can describe the track.

3.2.3 Assessment of Consistency Measures with Expert-labeled Tracks

The candidates for a consistency measure were compared to expert-labeled tracks of adaptive matrix sentence tests. Two independent experts with long experience in psychometric measurements with the matrix test were asked to rate the outcome of US-English and Cantonese matrix sentence tests with a traffic light color coding. Green stood for consistent (group 0), orange resembled questionable consistency (group 1), and red stood for a clearly inconsistent measurement (group 2). The matrix sentence tests included Lombard and plain speech, i.e., speech with raised and normal vocal effort (see M. K. Scharf et al. 2022), in the following noise conditions (ascending order of difficulty):

- unmodulated speech noise with a long-term spectrum of a male speaker at normal vocal effort (Wouter A. Dreschler et al. 2001) (ICRA1)
- modulated ICRA1 noise with a duration of pauses no longer than 250ms (Kirsten Carola Wagener, Thomas Brand, and Kollmeier 2006) (ICRA5-250)
- multi-talker babble noise
- ICRA1 noise in reverberation ($t_{60} = 2.69$s)

The reverb condition had the effect that several subjects performed well at the beginning but lost the cue, after which they continuously performed worse than at the beginning of the test.

The Spearman correlation coefficient (R_s) of the expert rating with the consistency measures was calculated alongside the area under the ROC-curve (AUC) of the receiver operating characteristic (ROC), which was mirrored on the diagonal if the true positive rate was smaller than the false positive rate. This way $AUC \geq 0.5$. For the ROC, the expert rating was defined as ground truth. For this, the three levels of expert rating were merged into two groups (0 and 1+2). This was done because groups 1&2 were an order of magnitude smaller than group 0, and there was no significant difference in SiPGOF between 1 and 2 (see Figure 3.4).

For the calculation of SiPGOF for the expert-labeled tracks, the following boundary conditions were used for all word-specific psychometric functions (see Eqs. 3.3): $\gamma = 0.1$, $\lambda = 0$, slope = 14.25%/dB (average of US-English and Cantonese matrix sentence test (Kollmeier, Warzybok, et al. 2015; Hu, Hochmuth, et al. 2022)).

3.2.4 Consistency of Unlabeled Tracks

Unrated matrix sentence tracks from Schädler, Hülsmeier, Warzybok, et al. 2020 and Hülsmeier 2022 were used to derive a classification threshold for the consistency of tracks of the matrix sentence test. The tracks were for German, male matrix sentence tests in ICRA1, ICRA5-250, and babble talker noise with the adaptive procedure of Thomas Brand and Kollmeier 2002. The boundary condition for the psychometric function of SiPGOF was adjusted to the average slope of the German, male matrix sentence test of 17.1%/dB Kollmeier, Warzybok, et al. 2015.

The classification threshold was estimated with a histogram of SiPGOF for a target classification rate. The SiPGOF threshold has been selected such that the cumulative probability of SiPGOF is equal to the target true positive rate of the test (i.e., truly consistent classified as consistent, see appendix).

3.3 Results

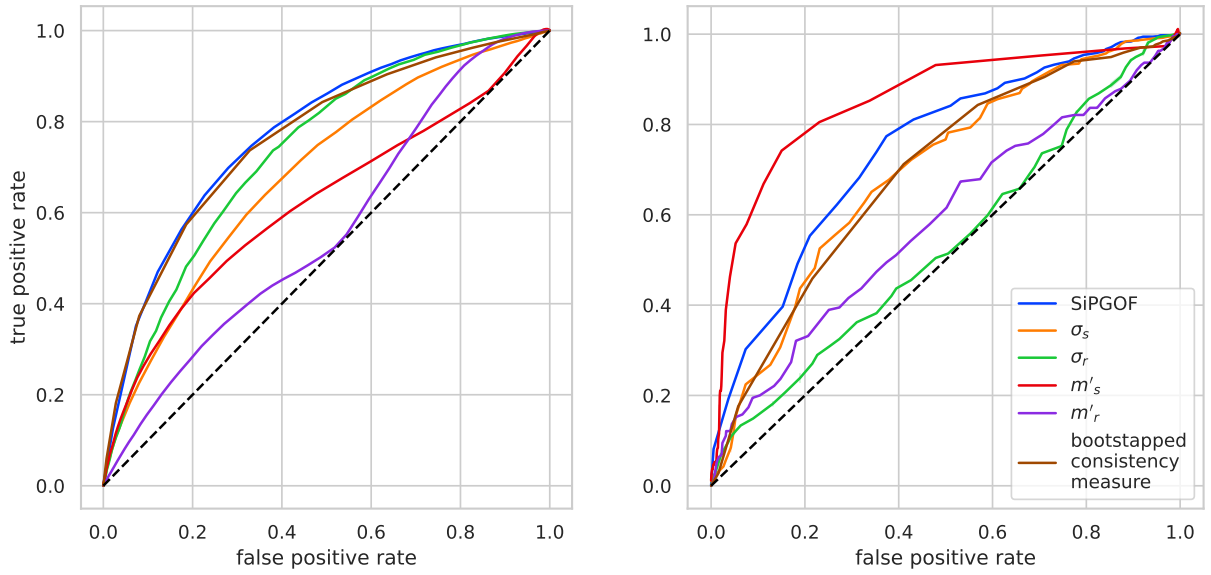
3.3.1 Best Consistency Measure to Predict Simulated Inconsistent Tracks

The upper panel of Table 3.1 contains the Spearman correlation coefficient (R_s) of the consistency measure candidates, the area under the ROC-curve (AUC) of a classification with the known inconsistency (KIC) for simulated, inconsistent tracks for both signed and unsigned KIC. A track with $|\text{KIC}| < 1$ decibel (dB) was classified as consistent. No candidate for the consistency measures showed a high correlation with the signed KIC.

Table 3.1: Spearman correlation coefficient (R_s) and area under the ROC-curve (AUC) of consistency measures with known inconsistency (KIC) and expert rating

see Eqs. 3.12-3.15	SiPGOF	standard deviation		slope		resonance frequency		slope of spectrum		bootstrapped consistency measure
		σ_s	σ_r	m_s	m_r	f_s	f_r	m'_s	m'_r	
					signed KIC:					
R_s	0.16	0.07	0.17	-0.06	-0.05	-0.00	0.06	-0.00	0.08	0.13
					unsigned KIC:					
R_s	0.39	0.27	0.32	-0.04	-0.04	-0.09	0.02	-0.17	0.01	0.38
AUC	0.78	0.69	0.74	0.56	0.54	0.53	0.55	0.62	0.56	0.76
					ground truth from expert rating[a]:					
R_s	0.29	0.23	0.04	-0.09	0.11	-0.22	-0.04	-0.42	-0.10	0.23
AUC	0.74	0.69	0.53	0.60	0.59	0.67	0.55	0.85	0.58	0.69

[a] Prediction of expert ratings for empirical tracks



(a) classification of the known inconsistency (KIC) for a simulation of inconsistent tracks

(b) classification of expert rating for empirical tracks

Figure 3.3: ROC for a classification of psychometric tracks of the matrix sentence test: σ is the standard deviation, and m' is the slope of the spectrum (see Eqs. 3.15 and 3.12). Subscripts s and r stand for stimulus and response, respectively. A consistent track is counted as positive.

3.3.2 Best Consistency Measure to Predict Expert Ratings

The experts agreed with a Kendall τ_B correlation of 0.77. Figure 3.4 shows the distribution of expert ratings and SiPGOF, which is the most promising consistency measure for simulated, inconsistent tracks. Tracks that were labeled as consistent (group 0) showed a significant difference in SiPGOF to tracks that were not labeled as consistent (group 1+2). No significant difference existed between groups 1 and 2 (questionable and unsuccessful).

The prediction results of expert ratings with consistency measures are displayed in the lower panel of Table 3.1 and the right panel of Figure 3.3. Of all consistency measures considered, the spectral slope m'_s was the best predictor of expert ratings, outperforming SiPGOF, the standard deviation of sequence of stimuli (**s**), and the bootstrapped consistency measure. The standard deviation of sequence of responses (**r**) was not a predictor of expert ratings.

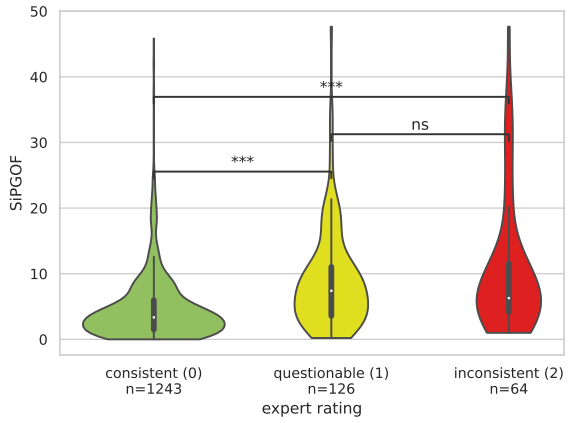


Figure 3.4: Violin plot of the single psychometric function goodness of fit measure (SiPGOF) for all expert rating groups:
Relative frequency of occurrence as a vertical histogram including median, interquartile range, and empirical standard deviation. The expert rating is indicated by a traffic light color coding. Green (left) stands for consistent, orange resembles uncertain consistency, and red (right) stands for inconsistent tracks. Ratings are grouped 0-2. The significance level (Turkey's test) is indicated with brackets, "⋆⋆⋆" is significant with $p<10^{-3}$, "ns" is not significant.

3.3.3 Consistency of Unlabeled Tracks

SiPGOF was calculated for every unlabeled track of the German, male matrix sentence test of publicly available tracks (Schädler, Hülsmeier, Warzybok, et al. 2020; Hülsmeier 2022). The histogram of the distribution of SiPGOF is displayed in Figure 3.5.

From the expert-labeled database, which contained an elevated amount of inconsistent tracks, we can expect 15% or fewer inconsistent tracks in an unscreened database.

A retest of an inconsistent track doubles the duration of the experiment. Therefore, to keep the number of retests small, we propose a target of 80% correctly classified consistent tracks. For this, we can expect to correctly classify 60% of all inconsistent tracks in the simulation and estimate a threshold SiPGOF ≈ 10 (see dashed line in figure 3.5 and appendix).

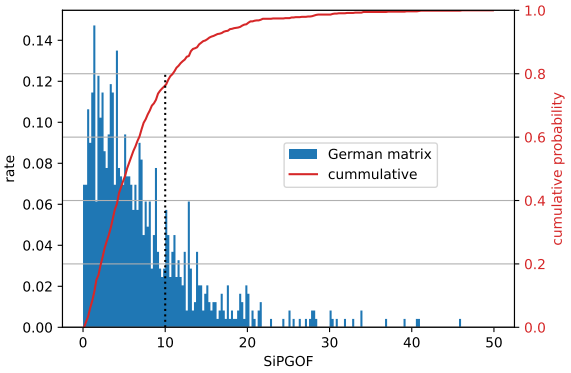


Figure 3.5: Histogram of SiPGOF for unlabeled German, male matrix tests:
For the calculation of SiPGOF, both slopes of the psychometric functions were fixed to the average value for the German, male matrix sentence test. The dotted line points to the cumulative probability for SiPGOF = 10.

3.4 Discussion

3.4.1 Best Consistency Measure

SiPGOF was the best consistency measure among our tested candidates for simulated, inconsistent tracks. It had a slightly better true positive rate than a bootstrapping approach (see Figure 3.3a) with equal AUC.

This is because the two measures share a related definition of consistency: a track is consistent if it agrees with a single psychometric function. However, the bootstrapped consistency measure can mislead in some cases because it cannot discriminate between random, but consistent variations and inconsistent response behavior: the response r_k is binomial distributed around the value of the psychometric function $\Phi(s_k)$ for the stimulus s_k (see Figure 3.1). r_k can scatter substantially around the expected mean for a single underlying psychometric function. The bootstrapping approach will rate any deviation from $\Phi(s_k)$ as inconsistent. For SiPGOF only those deviations of

r_k are counted as detrimental, which are in agreement with the effect of a distraction. Accordingly, the bootstrapped consistency measure is expected to have a reduced true positive rate (i.e., it rates more consistent tracks as inconsistent) in comparison to SiPGOF. This effect is observable for our simulation of inconsistent tracks and the prediction of expert ratings (see Figure 3.3).

The slope and resonance frequency of **s** and **r** had no predictive value. However, measures of the fluctuations of the track, e.g., the standard deviation, or spectral slope of **s** and **r**, were above chance level. Still, the correlation with the KIC of the simulated tracks was smaller than for SiPGOF. This is in line with Smits et al. 2022, who found that measures of fluctuations of a track, like the standard deviation, are not sufficient to measure the consistency of the one-up one-down adaptive procedure.

The spectral slope of **s** was not the best consistency measure for simulated, inconsistent tracks. Valid tracks, although less common, may contain low-frequency fluctuations of **s**. However, we found the spectral slope to be the best predictor of expert ratings (see Figure 3.3b). One possible explanation for this outcome is that large fluctuations in **s** are generally identified as inconsistency of a track (Smits et al. 2022). The spectral slope, or the standard deviation after the second reversal, measures this property.

However, the following reasons speak against the spectral slope of **s** as a consistency measure:

- It cannot generalize. Any measure of the fluctuations of the stimulus in a track will be misleading if the procedure exhibits oscillations that are not detrimental for the purpose of the measurement, e.g., experiments with constant stimuli where a sequential response bias may cause a quasi-periodic cycle.
- It does not consider all information. The spectral slope of **s** ignores information about **r**.
- Experts may be biased. The expert rating may be biased to rate fluctuating tracks as inconsistent. Our simulation suggests this, since the spectral slope **s** was no good predictor of the KIC.

The standard deviation of **r** was no predictor of the expert ratings, even though it was shown in our simulation to contain useful information about the consistency of the track. This indicates that experts may overly rely on certain features (like the spectral slope of **s**), while other aspects are not considered. This is different from a model-based measure of consistency, like SiPGOF. SiPGOF weights all aspects of the track based on a model understanding of inconsistency. This may suggest the advantage of SiPGOF as support for expert decisions.

3.4.2 SiPGOF as Local Optimum

To compute the maximum log-likelihood (l_{max}), the free parameters of the model were optimized with the expectation maximization algorithm (EM) algorithm. The EM/Baum-Welch algorithm is guaranteed to converge at a local and not a global optimum. In this application, the optimization was started at the 25th and 75th percentiles of the signal to noise ratio (SNR) values in the track. Even if a track allows more than one local optimum, we believe a result near this starting point is an acceptable solution. Other starting points might be used in other applications.

3.4.3 Consistency Measure as Correction Factor

For the simulated tracks, no consistency measure is correlated with the signed KIC (indicating over- or underestimation of the speech recognition threshold (SRT); see Table 3.1). This is because there is no defined order of states: a continuously distracted subject may produce relatively consistent tracks with high SRT. The tested consistency measures were unable to point out which state is the unwanted state of the subject; only the existence of two rivaling states can be indicated. This means that if an inconsistent track is identified, it cannot be corrected without further assumptions, e.g., the assumption that the true SRT is below or above the estimated threshold.

SiPGOF had the highest correlation with the KIC (see Table 3.1). The KIC of the simulated, inconsistent tracks is defined as the mismatch between the estimated threshold and the threshold of the underlying psychometric function (see methods section). Out of all tested consistency measures, SiPGOF is the most promising correction factor for inconsistent threshold estimates.

3.4.4 Definition of a Classification Threshold from Unlabeled Data

To use SiPGOF as a binary classifier, a rejection criterion needs to be known. We can estimate this threshold with a histogram of SiPGOF for unrated measurements (see Figure 3.5).

The threshold is estimated from the cumulative distribution of SiPGOF (figure 3.5) and the ROC of the simulation (left panel of Figure 3.3). From this, we estimate a lower bound for the true positive rate of 80% (i.e., truly consistent rated as consistent) for a threshold of SiPGOF = 10 (dotted line in Figure

3.5) because the dataset was not guaranteed to only contain consistent tracks. From the simulation, we can expect SiPGOF to classify 6 out of 10 truly inconsistent tracks as inconsistent for this threshold (see appendix). Therefore, we recommend using the rejection criterion of:

$$\text{SiPGOF} > 10$$

for the German male matrix sentence test.

Threshold values for SiPGOF with other psychometric tests are derived in the same way. However, if no simulation with KIC is available, only the true positive rate (i.e., truly consistent rated as consistent) can be calculated.

3.5 Conclusions

We conclude that single psychometric function goodness of fit measure (SiPGOF) can rate simulated tracks correctly and encode important information about the consistency of a track to support experts. SiPGOF should be preferred over a bootstrapping approach because of the theoretically more efficient calculation and a larger true positive rate (i.e., truly consistent rated as consistent), which is favorable for rare occurrences of inconsistent tracks.

SiPGOF encodes a different definition of inconsistency than the spectral slope of the sequence of stimuli ($\mathbf{s}$), which is a better predictor of expert ratings. The predictability of expert ratings with the spectral slope of $\mathbf{s}$ indicates that experts may be biased in their rating towards large fluctuations in the track. This emphasizes the need for a model-based consistency measure.

We present a method to estimate a suitable SiPGOF threshold with unlabeled tracks and find a suitable value to be 10 for the German, male matrix sentence test for a 60% chance to correctly classify inconsistent tracks as inconsistent and a false alarm rate of 20%.

A clear limitation of this study is that for any empirical data, there is no well-defined ground truth, i.e., some indication about how consistent the subjects actually were able to respond to the psychometric measurement. Monitoring the subjects' performance with EEG, pupil dilation, or purposely exhaustive measurement sessions might help to further verify the presented consistency measures.

In conclusion, while it is challenging to establish a meaningful ground truth of consistency, SiPGOF appears to be a promising candidate for a model-based measure of the consistency of psychometric measurements.

3.6 Outlook

The proposed consistency measure could be applied to the following use cases:

3.6.1 Indicator for Breaks

Exhaustion of test subjects during extensive tests can increase the rate of inconsistent responses due to a continuous lapse in attention. Modern measurement procedures should be robust against inconsistent response behavior or suggest an option to overcome this problem, i.e., repeat the measurement, take a break, or terminate the session (to be resumed later). A consistency measure might therefore indicate that a break is beneficial.

3.6.2 Quality Screening

Machine learning approaches with psychometric data (Saak et al. 2022) require a large amount of data from high-quality measurements. Crowdsourcing with self-administered tests is a possible approach to generating the necessary amount of data. However, to be useful, a quality control has to exist (Allahbakhsh et al. 2013). Presently, trained specialists are required for monitoring tests. Furthermore, if the monitoring results are not available at run time, the likelihood of a consistent retest is low. For automated procedures, an objective definition of consistency and a range of acceptable values of this rating have to be defined. We have reported an appropriate value of the classification threshold for the German matrix sentence test.

3.6.3 Expert Support System

In clinical practice of psychometric measurements, as performed by, e.g., audiologists, an objective quality control may be beneficial. This has been discussed by Smits et al. 2022 for the one-up, one-down adaptive staircase procedure. They showed that the consistency of an adaptive measurement is independent of the rate of convergence to the threshold stimulus, i.e., how smoothly the stimulus converged to the anticipated threshold. With no clear definition of consistency, rating the consistency of psychometric data remains a highly subjective task for the experimenter. Consequently, experimenters may be biased in their rating by the visual representation of the track (see example in Figure 3.1). Our simulations indicate that a model-based consistency measure may provide relevant information in a clinical setting.

3.6.4 Test Optimization

Unoptimized speech tests vary in threshold between sentences. This results in a smearing of the esti-

mated psychometric function because of a superposition of various psychometric functions with different thresholds (Kollmeier 1990). We have shown that SiPGOF can be used to distinguish between one and two psychometric functions. The consistency measure may be used to optimize speech tests without the need for human subjects when used together with computational models of speech intelligibility, such as the framework of auditory discrimination experiments (FADE) (Schädler, Warzybok, Hochmuth, et al. 2015). A bootstrapped consistency measure is already included in FADE. However, our empirical data did not include heterogeneous (i.e., unoptimized) speech tests to test the consistency measure explicitly as a predictor of homogeneity.

3.6.5 Consistency Measure for Nonconverging Measurement Procedures

Psychometric measurements do not necessarily aim to converge at a constant response, as there needs to be a trade-off between the number of trials (i.e., responses during a full test) and the time needed to complete the measurement.

This is why modern procedures heavily rely on assumptions about the psychometric function and only query stimuli with maximum expected gain of new information (Thomas Brand and Kollmeier 2002; Doire, Brookes, and Naylor 2017; Schlittenlacher, Turner, and B. C. J. Moore 2018). However, this comes at a cost: inconsistent responses at the beginning of the test can misguide the procedure. Interleaved procedures can monitor the performance of a subject (Leek, Hanna, and Marshall 1991) by concurrent estimates of the threshold, but lead to longer measurements. Additionally, if the procedure does not aim to converge, inconsistent measurements are harder to detect by visual inspection. As a result, inconsistent tracks of these procedures are difficult to recognize without an objective consistency measure.

3.6.6 Future Extensions

The SiPGOF presented here includes the comparison of a single-state to a two-state model. Future extensions of the SiPGOF may be generalized to more states. A Bayesian version might then automatically find the smallest number of states needed for a given track, by its inherent Occam's Razor mechanism (see Occam factor MacKay 2006, Ch. 28).

Declarations

Funding

This work was funded by: Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 415895050 - ” Experiments and models of speech recognition across tonal and non-tonal language systems (EMSATON)” , Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’ s Excellence Strategy –EXC 2177/1 - Project ID 390895286 - ”Hearing4all”

Conflict of Interest

The authors report no conflicts of interest.

Ethics Approval

Does not apply.

Consent to Participate

All participants in the expert rating consented to participate.

Consent for Publication

All authors approved the publication of this manuscript.

Availability of Data and Materials

Empirical tracks of the German matrix sentence test are available under Hülsmeier 2022.

Code Availability

SiPGOF is implemented in Python and supplements this article: M. K. Scharf 2025

Author’s Contributions

MS and BK were responsible for the conception of the method. MS was responsible for the implementation, simulation, analysis, and writing of the manuscript. AW and SH were responsible for the expert rating.

Aknowledgement

The authors thank Arne Leijon for his valuable comments on implementation and the manuscript.

3.7 Appendix to the Article

3.7.1 Guess-rate of the Matrix Sentence Test

For the closed matrix sentence test, the sentence-specific guess rate and the word-specific guess rate are identical, which may lead to confusion between the two. Guessing a word from 10 choices results in a word-specific guess rate of 0.1, while guessing a sentence from a 5-word sentence (trial) yields a sentence-specific guess rate of:

$$\frac{1}{5}\sum_{i=0}^{5}(5-i)\binom{5}{i}\cdot 0.1^{5-i}\cdot 0.9^{i} = 0.1 \tag{3.16}$$

This means that a test subject of the forced-choice matrix sentence test can, on average, never score worse than 1/2 word correct out of a sentence. Although there is no difference in the guess rate, the underlying psychometric functions are not identical, as the slope of the sentence-specific psychometric function cannot be steeper than the slope of the word-specific psychometric function (for implications of this on test development, see Kollmeier, Warzybok, et al. 2015).

3.7.2 Rejection Criterion

The rejection criterion was calculated from the receiver operating characteristic (ROC) of simulated, inconsistent tracks and a histogram of SiPGOF for unlabeled empirical matrix tests from Schädler, Hülsmeier, Warzybok, et al. 2020 and Hülsmeier 2022. The ROC maps false-positive rate (p_{FP}, i.e., inconsistent tracks were falsely identified as consistent) to true-positive rate (p_{TP}, i.e., consistent tracks were correctly identified as consistent).

$$\mathrm{ROC} : p_{\mathrm{FP}} \rightarrow p_{\mathrm{TP}} \tag{3.17}$$

From the distribution of SiPGOF for unlabeled, empirical tracks, we can estimate the cumulative distribution H to classify a track as consistent (i.e., the probability of a track to have a SiPGOF below the threshold above which tracks are considered inconsistent; see Figure 3.5).

$$\mathrm{H} : \mathrm{SiPGOF} \rightarrow p_{\mathrm{consistent}} \tag{3.18}$$

If we assume that the number of inconsistent tracks in the empirical data is small, we can approximate:

$$p_{\mathrm{consistent}} \approx p_{\mathrm{TP}} \tag{3.19}$$

The cumulative probability function H (Eqs. 3.18) is bijective (see Figure 3.5), so we can invert it and substitute into 3.17:

$$p_{\mathrm{FP}} \rightarrow \mathrm{SiPGOF}_{\mathrm{threshold}} \tag{3.20}$$

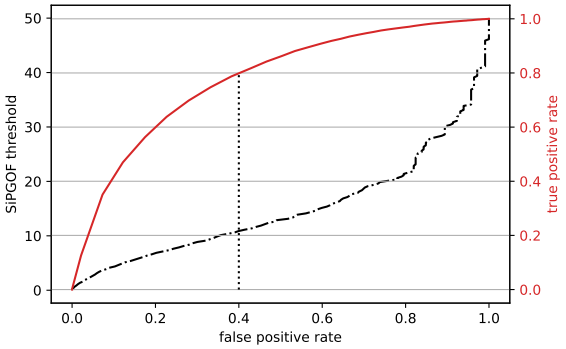


Figure 3.6: SiPGOF threshold and estimated true positive rate:
black: map of false positive rate to threshold SiPGOF according to Eqs. 3.20. **red:** ROC of SiPGOF for simulated inconsistent tracks as in Figure 3.3a. The vertical, dotted line marks the estimated performance of a consistency test with a threshold SiPGOF = 10.

The function 3.20 is shown in Figure 3.6 in black. A threshold of $SiPGOF \approx 10$ yields a true positive rate of $p_{\mathrm{TP}} \approx 0.8$ (8 out of 10 consistent tracks are correctly classified as consistent) and a false positive rate of $p_{\mathrm{FP}} \approx 0.4$ (four out of ten inconsistent tracks are not identified as such).

If SiPGOF is used to test for inconsistency, this test will have a sensitivity (inconsistent classified as inconsistent) of ≈ 60% and a specificity (consistent classified as consistent) of ≈ 80%.

Chapter 4:

On the Limited Generalizability of Specialized Band Importance Functions for Predicting Mandarin Speech Intelligibility

submitted to
Speech Communication

Maximilian Karl Scharf[1], *Anna Warzybok*[1], *Lena L.N. Wong*[2], *Fei Chen*[3], *Birger Kollmeier*[1]

[1] Medical Physics and Cluster of Excellence "Hearing4all",
Carl von Ossietzky Universität Oldenburg, Germany.
[2] Clinical Hearing Sciences (CHearS) Laboratory, Faculty of Education,
The University of Hong Kong, Hong Kong SAR, China.
[3] Department of Electrical and Electronic Engineering,
Southern University of Science and Technology, Shenzhen 518055, China.
maximilian.scharf@uni-oldenburg.de

Abstract

This article provides evidence that language- and gender-specific band importance functions (BIFs) of the speech intelligibility index (SII) in Mandarin Chinese perform comparably to generic BIFs for predicting speech intelligibility (SI) in broadband noise. The speech recognition thresholds for 50 % correctly understood words (SRT_{50}) of six male and five female speakers with two different speaking styles (plain and Lombard speech) in ICRA1-noise were predicted with language- and gender-specific and generic BIF. The predictive power of the SII appears to be independent of the BIF, which raises the question whether BIFs are a crucial parameter for developing generalizable models. We suggest using machine-learning-based SI models, which dynamically choose the most appropriate weighting in the time- and frequency domain.

Keywords: SII, Band-importance, Mandarin Chinese

4.1 Introduction

The efficient prediction of speech intelligibility (SI) is a valuable tool for the fast optimization of equipment for acoustic communication. Historically, the focus was primarily on hardware optimization, but the field has increasingly shifted towards optimizing signal processing and predictive algorithms. One widely used predictor of SI is the speech intelligibility index (SII)ANSIS3.5 1997, which produces a scalar value on the standard interval [0-1]. Its popularity stems from its high correlation with measured SI and its relatively simple calculation. Although the SII was defined initially for English, it has since been extended to other languages through a set of free model parameters: the band importance function (BIF)[1], which is a spectral weighting in the computation of the SII.

Language-specific BIFs are typically derived for a specific speech test and lead to reliable SI predictions for this speech material Wong et al. 2007. This approach is supported by the findings of Yoho et al. 2018, who reported that differences in BIFs across speech materials are larger than across speakers. It is also known that BIF vary depending on both the target stimulus, i.e., speech material ANSIS3.5 1997; DePaolis, Janota, and Frank 1996; E. W. Healy, Yoho, and Apoux 2013, and the masker, i.e., type of noise Bosen et al. 2024; Shen and Langley 2023.

However, it remains unclear whether these differences in BIF across languages, talker gender, and speech materials meaningfully improve SI prediction, or whether they primarily add more tuning parameters that optimize the SI prediction for a specific test without improving generalizability.

4.1.1 Aim of this Study

The generalizability of the SII with highly specific BIF, such as those optimized for a tonal language or particular talker gender, to comparable speech material has not yet been comprehensively evaluated. If BIFs represent truly generalizable parameters, one would expect that the most appropriate, material-specific BIFs should provide the most accurate predictions for related speech materials. This study investigates this assumption in the context of the tonal language Mandarin Chinese by comparing the predictive performance of several different BIFs for the SII. Lombard and plain speech of the same speakers are used to vary only the vocal effort in this comparison.

Importantly, the focus is on forward prediction, evaluating SI without fitting the model to the empirical data. SI data are averaged over listeners to remove listener-specific variance, reflecting how the SII is used in practice: predicting SI for unknown speech materials.

By comparing specialized, language- and gender-specific BIFs to generic and constant BIFs, this study examines whether optimizing BIFs for highly specific conditions translates into improved predictive power, or whether general BIFs are sufficient for practical applications. This approach directly addresses the open question of BIF generalizability and overfitting in SII-based predictions.

4.1.2 The Speech Intelligibility Index (SII)

The SII ANSIS3.5 1997, a successor of the articulation index (AI) Kryter 1962, has a long history as an objective measure of SI Pavlovic 1987 and remains widely used in research and applied acoustics (see, for example, C. F. Hauth et al. 2020; Amlani, Punch, and Ching 2002; Lopez-Poveda et al. 2017). It is based on the concept that different frequency regions contribute independently to the transmission of speech information. This approach is related to the Shannon-Hartley theorem, which connects the maximum rate of information transmission C to the signal to noise ratio (SNR) and bandwidth B:

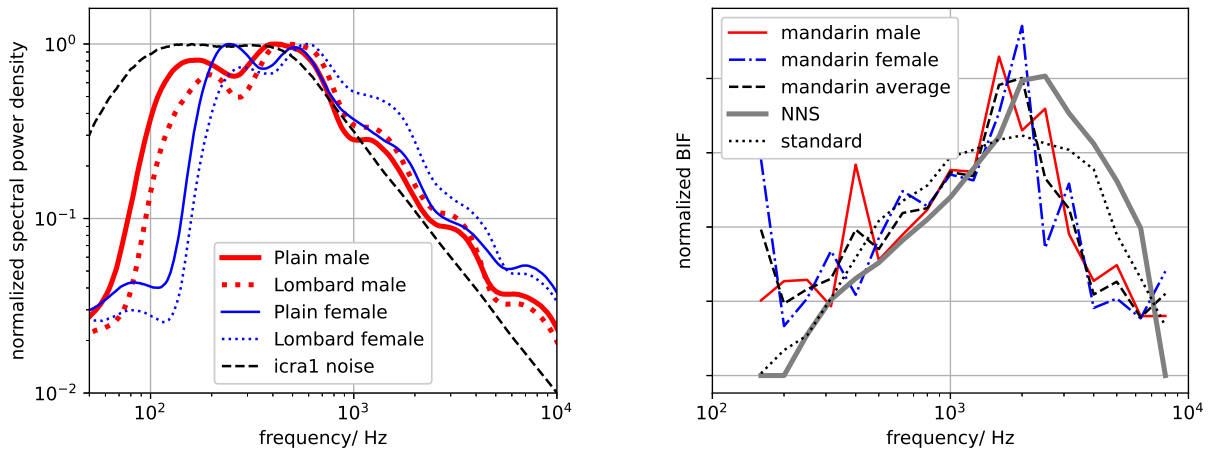
$$C = \int \log_2 \left(1 + \mathrm{SNR}(B)\right) dB \tag{4.1}$$

Although human speech communication does not reach this theoretical limit, the theorem motivates the concept that individual frequency bands provide different amounts of usable information. More importantly, not every band may contribute to information transmission by the same amount, such that an additional weighting is required to tailor the contribution of each band to the encoding of the message. The SII formalizes it by combining band audibility with empirically derived importance weights, the BIF:

$$\mathrm{SII} = \sum_{i=1}^{N} \mathrm{BIF}_i A_i \tag{4.2}$$

Where A_i is the audibility function of band $i = 1, \ldots, N$, which is derived from the long-term spectra of the target speech and masker. The A_i does not depend on the phase of the spectrum, which determines the spectro-temporal structure of the signal. Consequently, the BIF is needed to capture these spectro-temporal structures of the signal. This means that BIFs have to be specific for the language, gender, and speech type of the target, which all influence the spectro-temporal structure of the speech material.

[1] In some publications, the band importance function (BIF) is referred to as the frequency importance function (FIF).



(a) Average spectrum of the four test conditions

(b) Band importance function (BIF) for Mandarin Chinese

Figure 4.1: Comparison of spectrum and band importance function:
(a) Lombard and Plain speech for male and female voices in comparison to the spectrum of ICRA1. (b) from (J. Chen, Huang, and Wu 2016), band importance function various nonsense syllable tests where most of the English phonemes occur equally often (NNS) and standard BIF (ANSIS3.5 1997) for third-octave bands.

4.1.3 Extension of the SII to Tonal Languages

A major challenge for any SI-model is to distinguish between language- and speaker-specific SI-variation. A common approach is to use a BIF that is specific to a combination of language, speech material, and talker's gender. This practice implicitly assumes that these factors capture stable and meaningful differences that improve SI-prediction. However, previous work has shown that inter-speaker variability can exceed cross-language variability and that these effects are not fully independent (Hochmuth et al. 2015). This complicates the interpretation of how much of a "language-specific" BIF is truly attributable to language properties versus characteristics of the particular speakers or materials used during its derivation.

Deriving a BIF requires substantial speech recognition data, making speaker-specific BIFs impractical in conventional procedures. Recent methodological advances, such as the quick-BIF approach by Shen and Kern 2018, substantially reduce the required number of trials, allowing individualized BIF estimation with only 200-300 items, which amounts to 20 minutes of testing (Shen, Yun, and Y. Liu 2020). Although promising, such individualized procedures are not yet standard, and most practical applications rely on non-individualized BIF tables (ANSIS3.5 1997; Beutelmann, Thomas Brand, and Kollmeier 2010).

For Mandarin Chinese, J. Chen, Huang, and Wu 2016 and Du et al. 2019 have published BIFs, which were derived for different speech materials and talker gender. These studies reported systematic differences between Mandarin and English BIFs for comparable speech material, as well as improved SI predictions when using these language and gender-specific BIFs.

Given the availability of the Mandarin BIFs, the SII can, in principle, be used to predict Mandarin SI, provided that the BIFs generalize sufficiently well to related speech materials and speaker characteristics. This study evaluates this assumption.

4.2 Methods

4.2.1 Empirical Data

The analysis relied on the empirical speech recognition threshold (SRT) of the Mandarin Chinese matrix sentence test Hu, Warzybok, et al. 2022, published by F. Chen, Pan, et al. 2025. The dataset contains the SRT for $T = 11$ speakers (six male and five female) measured in stationary, speech-shaped ICRA1 noise. Each speaker produced two speaking styles ($S = 2$): plain speech and Lombard speech[2]. Using plain and Lombard speech enables variation in vocal effort while keeping speaker identity constant. The spectrum of the four conditions (gender x speaking style) and the ICRA1 noise is shown in

[2]Lombard speech (Lombard 1911) is sometimes referred to as speech with raised vocal effort.

Figure 4.1a.

The SRT for each speaker and speaking style was measured with $L = 13$ normal hearing participants, using the adaptive procedure of Thomas Brand and Kollmeier 2002. This procedure adapts the SNR to converge towards the 50% point of the psychometric function. The SRT was estimated by a maximum likelihood fitting. Compared to fixed SNR recognition scores, SRT measurements exhibit smaller measurement uncertainties (with respect to changes in SNR) because they are obtained near the inflection point of the psychometric function, where the slope is maximal (see Kollmeier, Warzybok, et al. 2015).

To minimize listener-specific effects, consistent with the intended use of the SII, the speaker- and speaking style-specific $\mathrm{SRT}_{t,s}$ was calculated from the mean SRT over all listeners l:

$$\mathrm{srt50}_{t,s} = \frac{1}{L}\sum_{l} \mathrm{srt}_{t,l,s} \tag{4.3}$$

The distribution of srt_l of each speaker and style was used to estimate the measurement uncertainty of the speaker-specific SRT from the standard deviation:

$$\Delta\mathrm{srt}_{t,s} \approx \sqrt{\frac{1}{L^2}\sum_{l}\left(\mathrm{srt}_{t,l,s} - \mathrm{srt}_{t,s}\right)^2} \tag{4.4}$$

These uncertainties are marked as error bars in the figures of the supplementary material and capture listener-specific variability and measurement errors from the matrix test. The mean uncertainty of the empirical SRT across speakers and speaking styles was 0.58 decibel (dB) SNR.

4.2.2 Band Importance Functions (BIFs) for this Speech Material

BIF for Mandarin Chinese

BIFs for Mandarin Chinese were taken from J. Chen, Huang, and Wu 2016, who derived gender-specific and average BIFs for random, phonetically balanced, monosyllabic words presented in stationary, speech-shaped noise. The BIF for monosyllabic words in Mandarin has been shown to be similar to the BIF of disyllabic words Du et al. 2019. This supports the use of the BIF of J. Chen, Huang, and Wu 2016 for the Mandarin matrix sentence test in ICRA1 noise, whose speech material consists of syntactically fixed, randomly selected disyllabic words scored at the word level.

Generic BIF

To assess how the Mandarin BIFs compare with more general alternatives, three generic BIFs specified in the ANSIS3.5 1997 were included in the analyses.

- **The NNS BIF** is the band importance function various nonsense syllable tests where most of the English phonemes occur equally often (NNS). It is a commonly used BIF in English and describes the frequency weights for consonant-vowel, vowel-consonant, and consonant-vowel-consonant speech tests, in which most of the English phonemes occur equally often (Fletcher and Galt 1950; ANSIS3.5 1997).
- **the standard SII third-octave BIF** uses the canonical weighting table defined in ANSIS3.5 1997, Table 3.
- **A flat BIF** inspired by Pavlovic 1987 and F. Chen and Loizou 2010, assigns equal importance to all third-octave bands.

These three generic BIFs differ in shape and linguistic basis from the Mandarin BIF, making them suitable comparators for evaluating whether language-, gender- or material-specific weighting offers predictive advantages (see Figure 4.1b).

4.2.3 Speech Recognition Prediction with the SII

SRT predictions were obtained using the SII combined with both the gender- and language-specific BIFs as well as the generic BIFs described above.

Two reference values, $\mathrm{SII}_{\mathrm{ref}}$, were used to map the predicted SII to SRT: the SII from the original Mandarin Matrix test Hu, Xi, et al. 2018, and the average SII at the SRT of each speaker and speaking style.

For the reference from the original Mandarin Matrix test, empirical speech recognition thresholds for 50 % correctly understood words (SRT_{50}) obtained with the same listeners and in the same noise condition (ICRA1) were averaged across listeners, and the mean SRT was converted into an SII value using the Mandarin, female BIF J. Chen, Huang, and Wu 2016 and long-term spectra of the original test material. When predicting the SRT with different BIFs, the same BIF was used in the computation of the corresponding average SII reference to maintain internal consistency.

The SRT was computed as the SNR at which the predicted SII matched $\mathrm{SII}_{\mathrm{ref}}$:

$$\mathrm{srt}_{50}(\mathrm{BIF}) = \underset{\mathrm{SNR}}{\operatorname{argmin}} \left|\mathrm{SII}(\mathrm{SNR}, \mathrm{BIF}) - \mathrm{SII}_{\mathrm{ref}}(\mathrm{BIF})\right| \tag{4.5}$$

Because the SII varies monotonically with SNR, this procedure yields a unique solution.

For each speaker and speaking style, the long-term third-octave band spectra of the speech signals were measured from stationary speech-shaped noise. These signals were generated from the recorded sentences so that the spectrum was equal to the long-term spectrum of the concatenated sentences. The procedure was implemented in MATLAB.

4.2.4 Statistical Measures

For the comparison of predicted SRT with empirical SRT, we report the Pearson correlation coefficient (R_p), root mean square error (RMS), and bias. The RMS is defined as:

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{N}(x - x')^2} \tag{4.6}$$

For a list of N empirical measurements x and predictions x'.

The bias is the offset between a linear interpolation to $x \to x'$ with unity slope and the 45° line, i.e., the line of equality:

$$\mathrm{bias} = \frac{1}{N}\sum_{N}(x - x') \tag{4.7}$$

To test for significant differences between correlations, Steiger´s Z test of the R psych package Revelle 2007 with Bonferroni correction was used.

4.3 Results

Table 4.2 summarizes the comparisons between predicted and empirical SRT separately for plain and Lombard speech and for both speaking styles combined, using five different BIF, and two reference conditions. The comparison between predictions with Mandarin gender-specific and constant BIF is shown in Figure 4.2, where each speaker is marked with a number.

Across all conditions, the correlations between predicted and empirical SRT varied only marginally across BIFs. For plain speech, correlation coefficients ranged from 0.81 to 0.90, whereas Lombard speech yielded consistently higher values between 0.93 and 0.96. When both speaking styles were combined, correlations fell within a similarly narrow interval of 0.88 to 0.94. These ranges indicate that none of the tested BIFs, neither the language- and gender-specific Mandarin functions nor the generic alternatives, provided a clear or systematic advantage in predictive accuracy. In particular, the Mandarin gender-specific BIFs performed no better than the Mandarin average BIF or the generic alternatives (NNS, standard, and constant weighting). Also, RMS error and prediction biases showed comparable patterns across BIFs. Using the average SII reference condition, RMS errors ranged from 0.93 to 1.08dB, with biases close to 0dB. With the original test SII reference, RMS errors were slightly larger and varied between 1.17 and 1.32dB. The biases were consistently negative (from −0.80 to −1.01dB). However, for both reference values, these biases were consistent across BIFs, so the relative pattern of results was unaffected.

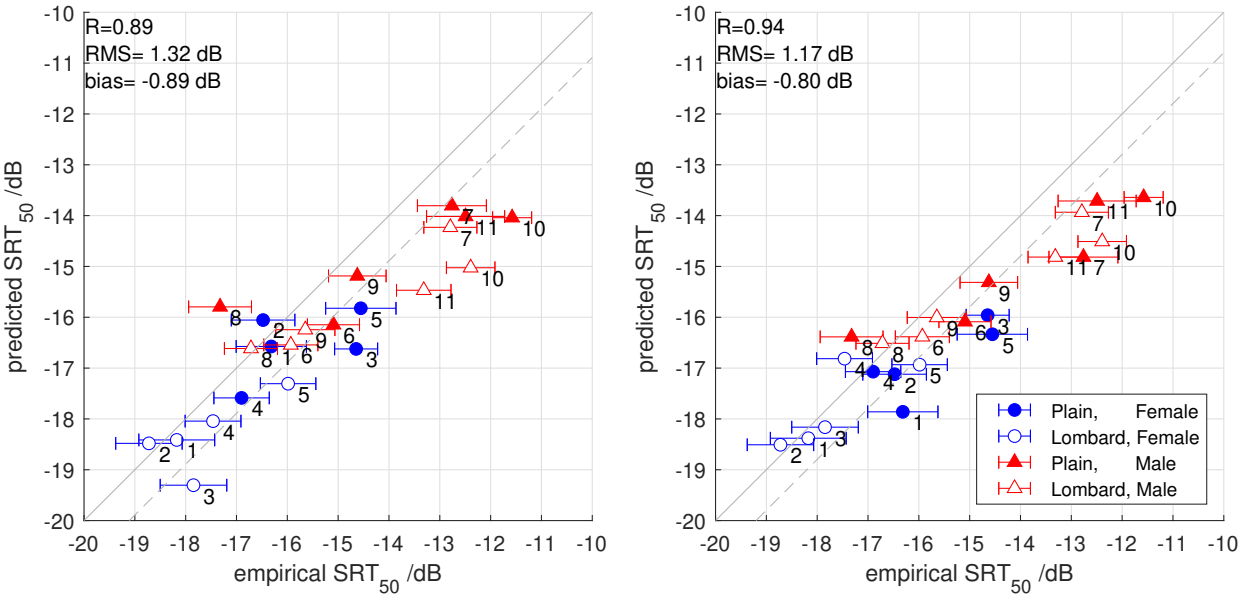
Most importantly, Steiger´s Z tests showed no statistically significant differences between any pair of BIFs with respect to their correlation with empirical SRT (see Table 4.1).

Even the constant weighting function, effectively the SII without a BIF, achieved correlations and RMS errors that were indistinguishable from those obtained with highly specific Mandarin BIFs. This finding suggests that the additional structure embedded in the language- and gender-specific BIFs did not translate into measurable improvements in predictive performance for the Mandarin matrix test.

Table 4.1: Bonferroni corrected p-value of Steiger' s Z test of correlation differences for combined plain and Lombard SRT prediction:
Upper triangle for the mean SII reference. The lower triangle for the reference SII from the original Mandarin speaker Hu, Xi, et al. 2018.

	Mandarin gender-specific [a]	Mandarin average [a]	NNS[b]	standard [b]	constant
Mandarin gender-specific [a]		1.00	0.69	0.59	0.97
Mandarin average [a]	0.96		0.56	1.00	1.00
NNS[b]	0.63	1.00		1.00	1.00
standard [b]	0.38	1.00	1.00		1.00
constant	1.00	1.00	1.00	1.00	

[a] J. Chen, Huang, and Wu 2016,[b] table 3 of ANSIS3.5 1997



(a) Average spectrum of the four test conditions

(b) Band importance function (BIF) for Mandarin Chinese

Figure 4.2: Comparison of empirical SRTs (F. Chen, Pan, et al. 2025) and predictions with the reference SII of the original Mandarin matrix test (Hu, Xi, et al. 2018):
(a) Mandarin gender-specific BIF of J. Chen, Huang, and Wu 2016, (b) constant BIF. Numbers indicate the speaker. Uncertainties (see Equation 4.4) are marked with error bars.

Table 4.2: Pearson correlation coefficient (R_p) of the SII SRT prediction with empirical SRT for Mandarin Chinese matrix sentence tests in ICRA1 with different band importance function (BIF) and two reference conditions:
Summary for 6 male and 5 female speakers and 13 listeners. No significant difference between BIFs exists for the correlations. For the definition of root mean square error (RMS) and bias see methods section 4.2.4.

reference	speaking style / BIF	R_p	R_p	R_p	RMS/dB	bias/dB
		Plain	Lombard	both		
average SII	Mandarin gender-specific[a]	0.81	0.93	0.88	1.08	0.18
	Mandarin average[a]	0.83	0.93	0.88	1.07	0.17
	NNS[b]	0.87	0.93	0.91	0.93	0.11
	standard[b]	0.87	0.94	0.91	0.97	0.14
	constant	0.90	0.96	0.94	0.95	0.16
orig test SII[c]	Mandarin gender-specific [a]	0.83	0.94	0.89	1.32	-0.89
	Mandarin average[a]	0.86	0.94	0.90	1.29	-0.88
	NNS[b]	0.90	0.94	0.93	1.31	-1.01
	standard[b]	0.90	0.95	0.93	1.24	-0.89
	constant	0.90	0.96	0.94	1.17	-0.80

[a] J. Chen, Huang, and Wu 2016,[b] table 3 of ANSIS3.5 1997,[c] Hu, Xi, et al. 2018

4.4 Discussion

The aim of this study was to examine whether published Mandarin BIFs generalize to a closely related but independent Mandarin speech test, namely the Mandarin matrix sentence test. Based on the previous findings that Mandarin BIFs derived from monosyllabic and disyllabic speech materials are comparable Du et al. 2019, and on the assumption that BIFs capture stable, language-dependent weightings of acoustic-phonetic cues, the working hypothesis was that Mandarin-specific BIFs should improve SII-based SRT prediction for the Mandarin matrix sentence test.

The results did not support this hypothesis. None of the tested Mandarin BIFs provided a measurable predictive advantage over generic BIFs or even a constant weighting function. This suggests

that the applied Mandarin BIFs did not generalize to the matrix sentence speech material, despite the linguistic similarity between the derivation and evaluation corpora.

Rather than contradicting the literature, this outcome highlights a structural limitation of BIFs within the SII framework. When BIFs are derived from a specific speech corpus and then evaluated on the same corpus, they can improve prediction because they absorb speech-material-specific properties that the SII cannot explicitly represent. However, such fitted BIFs may encode characteristics tied to the original speech test, its lexical composition, tone distribution, or spectro-temporal structure, rather than a stable perceptual weighting. When transferred to a different but related Mandarin speech test, these BIFs may therefore lose predictive value.

This issue may be particularly pronounced for tonal languages such as Mandarin. Because lexical tone is encoded through pitch contour and fine-grained spectro-temporal cues, a purely spectral weighting has limited capacity to reflect the relative importance of tonal information. As suggested previously (M. K. Scharf et al. 2022), models lacking spectro-temporal features may therefore force tonal cues into the BIF parameters in a manner that is inherently test-specific rather than language-general.

It is important to consider whether the lack of BIF influence could simply be explained by the energy distribution of the present stimuli. From the spectrum of speech and masker (see Figure 4.1a), it is evident that most energy is concentrated below 1 kHz. Together with the fact that the BIFs differ primarily above 1 kHz, it might not be surprising that the choice of BIF did not strongly influence the SRT prediction. At the same time, the SRT is based on a weighted ratio of target and masker, such that bands with relatively low absolute energy but high SNR may have a substantial contribution to the resulting SRT. In fact, as was noted by F. Chen, Pan, et al. 2025, the larger SRT difference between plain and Lombard speech for female speakers is clearly visible as a change in SNR above 1 kHz. Thus, differences at higher frequencies should be relevant and should, in principle, interact with the BIF. The fact that they did not suggests that the BIFs failed to capture a generalizable weighting rather than that the stimuli lacked informative high-frequency differences.

A limitation of this study is the absence of Mandarin BIFs for Lombard speech. Whether Lombard-specific BIFs would improve prediction remains uncertain, but given that language- and gender-specific BIFs did not improve the SRT prediction accuracy, substantial improvements from Lombard-specific BIFs appear unlikely. From the spectral differences between gender and speaking style, one would have expected that gender-specific BIFs would have a significant influence on the prediction, which we did not observe. We do not question that there may exist a BIF that leads to a better prediction accuracy of our data than with the BIFs that were applied here. However, such a BIF would likely miss any generalizability to comparable data, as it would need to capture effects that were not included in the BIF of J. Chen, Huang, and Wu 2016.

A further consideration for future analyses is the inclusion of talker- and listener-specific-proficiency factors, as proposed by Pavlovic 1987. Such factors could, in principle, reduce inter-individual variability by separating acoustic contributions from listener or speaker proficiency. In the present study, speaker-specific SRT values were averaged across listeners, and the resulting standard deviations already capture listener variability. Therefore, proficiency factors were not required to address the main question of this study. Nevertheless, incorporating these factors in future work could provide a more detailed decomposition of variability, isolating the contribution of BIFs from other factors.

Finally, the SII standard itself acknowledges the limited role of highly specialized BIFs: *"For communication systems where the nature of the message and proficiency of the listeners and talkers may vary greatly [...], the SII, calculated with the importance function for average speech used in this standard, is a better descriptor of the quality of the communication system with respect to intelligibility than any SII calculated with a different importance function"* ANSIS3.5 1997, sec. 6. While language-specific BIFs derived for a specific speech test may improve the SRT prediction for that specific speech test, their predictive performance tends to decline when applied to comparable speech material, even within the same language.

This raises a broader question of whether BIFs can serve as generalizable model parameters at all, or whether they must be re-estimated for each new speech test. One promising alternative is the use of automatic speech recognizer (ASR)-based speech intelligibility models such as the framework of auditory discrimination experiments (FADE) (Schädler, Warzybok, Hochmuth, et al. 2015), the phone-based binaural intelligibility model (PHOBI) (Roßbach, Kirsten C. Wagener, and Meyer 2023), the intelligibility model of Saddler and McDermott 2024, or other recent frameworks. These models may overcome the inherent limitations of the standard SII, including its inability to incorporate spectro-temporal cues. They may more effectively identify and weight the most relevant speech features that determine intelligibility. By leveraging these models, it may be possible to achieve more robust predictions across different speech materials

and speaker characteristics.

4.5 Conclusion

This study demonstrated that language- and gender-specific band importance functions (BIFs) of the speech intelligibility index (SII) did not provide an advantage over generic or constant BIFs in the prediction of empirical speech recognition threshold (SRT) of the Mandarin matrix sentence tests in unmodulated speech noise with a long-term spectrum of a male speaker at normal vocal effort (Wouter A. Dreschler et al. 2001) (ICRA1).

These results question the generalizability of specialized BIFs to comparable speech material. These findings support the use of simpler or more generalizable weighting schemes and highlight the potential of approaches that dynamically extract and weight relevant speech features for robust intelligibility prediction.

Author Declarations

Authors Contribution

MS was responsible for the analysis and writing of the manuscript. AW, LW, and FC were responsible for the empirical data. AW and BK contributed to the writing of the manuscript.

Funding

This research was supported by: the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 415895050 - "Experiments and models of speech recognition across tonal and non-tonal language systems (EMSATON)", the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy –EXC 2177/1 - Project ID 390895286 - "Hearing4all", the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 352015383 –"SFB 1330 A5 Model-based diagnostics and hearing aid fitting in complex acoustic scenarios: Perceptive Principles, Algorithms and Applications", the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) –Project ID 465121786 –(GRUSTAD)

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Data Availability

The audio recordings are available under Hu, Warzybok, et al. 2022. The code for the speech intelligibility (SI) predictions is available from the authors upon request.

Chapter 5:

Spectro-temporal vs. Spectral Features to Predict the Lombard Gain in Mandarin Chinese

submitted to
PLOS One

Maximilian Karl Scharf[1], *Anna Warzybok*[1], *Lena L.N. Wong*[2], *Fei Chen*[3], *Birger Kollmeier*[1]

[1] Medical Physics and Cluster of Excellence "Hearing4all",
Carl von Ossietzky Universität Oldenburg, Germany.
[2] Clinical Hearing Sciences (CHearS) Laboratory, Faculty of Education,
The University of Hong Kong, Hong Kong SAR, China.
[3] Department of Electrical and Electronic Engineering,
Southern University of Science and Technology, Shenzhen 518055, China.
maximilian.scharf@uni-oldenburg.de

Abstract

Background: In background noise, speakers adapt their speech production, giving rise to Lombard speech, which often improves speech intelligibility (SI). While intelligibility benefits of Lombard speech have been extensively studied in non-tonal languages, it remains unclear whether spectro-temporal cues, which are critical for tonal contrasts, are necessary to predict both the Lombard gain (LG) (i.e., intelligibility improvement relative to plain speech) and absolute SI in Mandarin Chinese.

Methods: Predictions of two SI-models were compared, namely, an automatic speech recognizer (ASR)-based approach using spectral or spectro-temporal features and the speech intelligibility index (SII)-based model using spectral features. Predicted LG and absolute speech recognition threshold (SRT) values, for five female and six male speakers in stationary speech-shaped noise, were compared with empirical data.

Results: For both models, spectral features alone are sufficient for accurate prediction of the LG with $\chi^2/\nu < 1.4$ for both models. In contrast, predictions of absolute SRT were most accurate when spectro-temporal features were included, capturing substantial inter-speaker variability.

Conclusions: Despite the tonal nature of Mandarin, spectro-temporal features are not required to predict the LG. However, they are essential to predict the absolute SRTs, which vary across speakers.

Keywords: speech intelligibility prediction, tonal language, Lombard effect

5.1 Introduction

Speech intelligibility depends on the acoustic cues that carry linguistic information in a given language. In non-tonal languages such as English or German, these cues are primarily determined by spectral properties of the speech signal. In tonal languages like Mandarin Chinese, however, dynamic F0-contour changes within a single syllable carry lexical meaning and therefore constitute essential segmental cues. These tonal patterns reflect an inherently spectro-temporal structure, raising the possibility that temporal dynamics contribute more strongly to speech intelligibility in tonal languages than in non-tonal ones, particularly for absolute[1] speech intelligibility (SI) prediction.

An additional factor influencing the acoustic realisation of speech is the Lombard effect, which refers to the changes in a speaker's voice when speaking in loud acoustic environments (Lombard 1911).

The acoustic changes associated with the Lombard effect include an upward shift in fundamental frequency (F0), frequency shifts in the first and second formants (F1, F2), and a reduced speaking rate (see, for example, Hanley and Steer 1949; Hansen 1996; Pittman and Wiley 2001; Lu and Cooke 2008; Alghamdi et al. 2018), modifying both spectral and temporal properties of speech. These changes enhance audibility and clarity of speech cues, improving speech recognition (Dreher and O'Neill 1957; Summers et al. 1988; Junqua 1993).

The improvement in speech intelligibility between plain and Lombard speech is commonly referred to as the Lombard gain (LG). Spectral changes, such as changes in F0 and formants, strengthen the contrast of speech cues against masking sounds, while temporal changes, including a slower speaking rate, give listeners more opportunities to process the signal, particularly in fluctuating noises (Lu and Cooke 2009). In contrast to absolute SI, LG reflects a relative benefit and may therefore rely on a different subset of acoustic cues.

Although these mechanisms have been extensively studied in non-tonal languages, it remains unclear whether the same acoustic dimensions account for LG and absolute SI in tonal languages.

5.1.1 Tonal Languages

For tonal languages, the Lombard effect may interact with the dynamic F0 patterns that carry lexical meaning. While spectral cues such as shifts in F0 and formants are likely important, the contribution of temporal information to LG is still questionable.

Recent findings for Mandarin Chinese (F. Chen, Pan, et al. 2025) suggest that spectral measures, such as α-ratio and speech-weighted signal to noise ratio (SNR), correlate with the LG, while F0-related changes are also predictive. However, the relative contributions of spectral versus spectro-temporal cues to absolute SI and LG in tonal languages remain unresolved.

To address this question, we explore computational SI models as a controlled analysis tool that allows us to isolate specific acoustic features. The models are not the primary object of investigation in this study; instead, they allow us to manipulate the acoustic information available for prediction selectively, thereby disentangling spectral and spectro-temporal contributions under otherwise identical conditions.

5.1.2 SI Models of this study

The primary model was the framework of auditory discrimination experiments (FADE) of (Schädler, Warzybok, Hochmuth, et al. 2015). FADE implements a machine learning backend, adapted from automatic speech recognizer (ASR) systems, similar to the "optimum detector" approach (Dau, Kollmeier, and Kohlrausch 1996). Stimuli are first preprocessed by a frontend, which is inspired by auditory-system mechanisms, allowing precise control over the type of information entering the model. This control makes it possible to directly compare speech recognition threshold (SRT) predictions using spectral or spectro-temporal representations of the input signals. Importantly, FADE does not assume any fixed or predefined combination of spectral and temporal cues; instead, it automatically selects the features that lead to the best speech recognition performance.

FADE has previously demonstrated high accuracy across several non-tonal languages in both stationary and fluctuating noise conditions (Schädler, Warzybok, Ewert, et al. 2016a), outperforming the extended speech intelligibility index (SII) (eSII, see Rhebergen, Versfeld, and Wouter. A. Dreschler 2006). It also provides accurate predictions for normal-hearing and hearing-impaired listeners (Schädler, Hülsmeier, Warzybok, et al. 2020; Hülsmeier, Warzybok, et al. 2020), and it still outperforms state-of-the-art commercial ASR systems (Polspoel et al. 2025). In the context of tonal languages, FADE successfully predicted the average LG for Cantonese speakers (M. K. Scharf et al. 2022), which further motivated its use in the present study.

Despite these advantages, the ASR backend of FADE is implemented as a hidden Markov model with restricted transitions (see method section). As

[1] By "absolute" SI prediction we denote a reference-free, ab-initio SI prediction as opposed to a "relative" SI prediction of a deterioration or improvement in relation to a predefined standard.

a result, even a spectral-only input representation may still permit some spectro-temporal inference if its changes across states are decoded. Therefore, the classical SII was used as a second model approach, which decomposes speech and noise into long-term spectral components, providing a strictly spectral model of intelligibility (ANSIS3.5 1997).

The SII has proven to be a simple yet powerful SI model for non-tonal languages (Beutelmann, Thomas Brand, and Kollmeier 2010; Lopez-Poveda et al. 2017). If the LG in Mandarin Chinese is fully explainable by spectral changes in the speech signal only, a strictly spectral model such as the SII is expected to be sufficient to account for the changes in speech intelligibility.

5.1.3 Aim of this Study and Rationale

The presented study investigates whether spectral information alone is sufficient to predict the LG in Mandarin Chinese, or whether additional spectro-temporal cues are required. In addition, we examine whether the same conclusions hold for absolute SI, quantified by the SRT.

Because spectral features are a subset of spectro-temporal features, accurate predictions based solely on spectral inputs would imply that temporal cues introduced by the Lombard effect do not provide additional, independent intelligibility-relevant information. Under the assumption that human listeners make optimal use of the available cues, this would indicate that spectral changes are the main driving factor of the LG in Mandarin Chinese.

To address this, two complementary comparisons were carried out:

(1) FADE with spectro-temporal versus FADE with spectral only features This comparison isolates the contribution of the input representation, as both conditions share the same computational backend. Any differences in predicted LG, therefore, stem solely from the presence or absence of temporal information in the frontend.

(2) FADE with spectral features versus SII This comparison ensures that predictions based on spectral information are evaluated with a strictly spectral model. As the models (i.e., FADE and SII) differ in their computational architectures, this step verifies that any effects observed in comparison (1) are not artefacts of residual temporal sensitivity in FADE's ASR-based backend.

Together, the two comparisons allow us to determine whether spectro-temporal cues provide indispensable information for predicting the LG, while also controlling for potential confounds arising from the underlying model architecture.

5.2 Methods

5.2.1 Empirical Data

The empirical data of SRT of plain and Lombard speech in Mandarin were previously published by F. Chen, Pan, et al. 2025. The speech material consisted of Mandarin matrix test sentences (Hu, Xi, et al. 2018), a syntactically fixed, five-word sentence test with a constrained vocabulary. Recordings were obtained from six male and five female native Mandarin speakers for both plain and Lombard speech. To induce Lombard speech, unmodulated speech noise with a long-term spectrum of a male speaker at normal vocal effort (Wouter A. Dreschler et al. 2001) (ICRA1) noise at 80 decibel (dB) sound pressure level (SPL) was presented over headphones to the speaker. The plain and Lombard speech recordings were subsequently used to measure speech recognition and assess the speaker-specific LG with a group of 13 native normal-hearing listeners. The SRT was measured adaptively with the procedure proposed by Thomas Brand and Kollmeier 2002 in stationary ICRA1 noise. Twenty-two measurements were conducted with each listener so that the SRT was obtained for each speaker and speaking style (11 speakers x two speaking styles).

The average across listeners was used as speaker-specific SRT for each condition, and the uncertainty of the mean was estimated from the standard deviation across listeners. F. Chen, Pan, et al. 2025 reported speaker-specific SRTs in the range of -17.3dB to -11.6dB for plain speech and -18.7dB to -12.4dB for Lombard speech. The speaker-specific LGs varied from 3.2dB to -0.6dB. The data showed a significant mean LG across speakers.

5.2.2 FADE Prediction

Schädler, Warzybok, Hochmuth, et al. 2015 proposed the framework of auditory discrimination experiments (FADE) for SRT modeling. FADE uses a hidden Markov model in conjunction with a Gaussian mixture model (HMM-GMM) backend adapted from ASR systems. This design allows the model to evaluate intelligibility from the acoustic input while retaining the flexibility to exploit spectral and temporal cues without any a priori assumptions about their combination.

FADE Frontend: Feature Representation

FADE supports a wide variety of speech signal feature types for entry into the ASR backend. To test the contributions of spectral versus spectro-temporal cues to the Lombard gain in Mandarin Chinese, feature representations were selected that selectively encode temporal information:

1. Log-mel spectrograms (31 bands, 16 kHz sampling rate, without differential, acceleration, or higher order coefficients ETSI 2003) primarily capture spectral content within each time frame, providing a baseline representation without explicit temporal modulations.

2. mel frequency cepstral coefficients (MFCC) without differential, acceleration, or higher order coefficients are derived from a downsampled inverse Fourier-transformed of the log-mel spectrogram. They also represent spectral information only, offering a complementary spectral-only baseline to verify that results are not dependent on a particular spectral representation.

3. separable gabor filter bank (SGBFB) features apply a two-dimensional filter to the log-mel spectrogram, explicitly capturing both spectral and temporal modulations within each time frame. This representation is motivated by simple auditory principles, modeling cortical neuron responses and providing a means to incorporate the dynamic temporal patterns (Schädler and Kollmeier 2015; Schädler, Warzybok, Ewert, et al. 2016b) that may be especially relevant for tonal languages and Lombard speech.

By comparing model predictions across these feature sets, we can systematically evaluate the added value of temporal cues in the prediction of the LG in Mandarin Chinese, while controlling for differences in spectral representations.

FADE Backend: The Generalized Optimum Detector

Central to FADE is the concept of an optimum detector, which extracts the most relevant information from the input feature representation to perform the detection or discrimination task. Unlike the optimum detector approach of Dau, Püschel, and Kohlrausch 1996, FADE generalizes to any psychoacoustic discrimination task, such as speech recognition, using a HMM-GMM as ASR backend. This flexibility makes FADE a suitable tool for testing the importance of cues that contribute to LG in Mandarin Chinese. In contrast, traditional SI-models, such as the SII, make fixed assumptions about the importance of signal features.

The optimum detector operates in two steps: in the first step, the speech signals of the test material (i.e., the sound files of the speaker for the condition of interest) are mixed with random segments of masking noise across a range of SNR. HMM-GMMs are then trained on the labeled mixtures for each SNR. The underlying Markov model assumes eight consecutive states per word (three representing silence and six states representing the beginning or end of a sentence). For the definition of the grammar in the language model, word transitions are restricted according to the grammar of the matrix test. Importantly, this is not an additional modeling assumption, but rather reflects the experimental condition of listeners who are trained before actual measurements to get familiar with the limited vocabulary of the test and with the sentence structure. Including the same grammar in FADE therefore ensures that the model operates under the same structural constraints as the human listeners. Accordingly, FADE models an optimally trained human listener.

In the second step of the optimum detector, a new set of data, constructed in the same manner but using new segments of noise, is used to test the performance of the HMM-GMMs. The resulting model performance as a function of training- and testing-SNR defines the predicted SRT: the lowest test SNR at which word recognition rate reached 50% is returned as the SRT. Uncertainties of the predictions are estimated by bootstrapping the word-recognition performance.

FADE is run independently for each speaker and speaking style (plain, Lombard), ensuring that the optimum detector is always trained on the specific material for which the prediction is made.

The Markov assumption of the FADE backend with its eight states per word warrants further inspection for predictions with spectral features. This is because, although only spectral information may enter the backend, the Markov chain could, in principle, model temporal changes between frames. However, this quickly exhausts all available states for a word; for example, an upwards shift or second tone in Mandarin over 500 ms with a window shift of 10 ms would require in the order of 50 states for log-mel features. As it is central to the conclusion of this study that the model can be made insensitive to temporal features, an additional, purely spectral model (the SII, see Section 5.2.3) was included in this study for comparison.

5.2.3 SII-based Prediction

To complement the FADE predictions, the speech intelligibility index (SII) (ANSIS3.5 1997) was used as a purely spectral model. Unlike FADE, the SII does not incorporate temporal modulations or spectro-temporal interactions and assumes that the filtered long-term spectrum fully determines SI.

In this study, third-octave band-pass filters were implemented according to ANSIS1.11 2004 and used for the calculation of the SII. While these filters differ from the log-mel or MFCC features of FADE, they share the essential property that only long-term spectral information is available to the model.

Definition of the SII

For the SII calculation, speech and noise were decomposed into third-octave bands. The band-specific audibility was computed and weighted using Mandarin language- and gender-specific band importance functions (BIFs) (J. Chen, Huang, and Wu 2016).

Reference Condition for SII-based Predictions

To convert the SII into a predicted SRT, a reference SII (SII_{ref}) was calculated from empirically measured SRTs of the original Mandarin matrix test (Hu, Xi, et al. 2018) in stationary ICRA1 noise. For a new speech in noise condition, the SNR is adjusted such that the SII matches the reference value:

$$SRT_{pred} = \underset{SNR}{\operatorname{argmin}} \left[SII_{ref} - SII(SNR) \right] \tag{5.1}$$

The SNR that corresponds to this match is taken as predicted SRT.

5.2.4 Statistical Measures for a Comparison of Models

When comparing a model prediction with empirical data, it is important to account for measurement uncertainties and to understand how well a "perfect" model would perform.

Here, the "perfect" SI-model is defined as one that predicts the empirical SRT exactly when the measurement uncertainties are zero. In practice, both measurement and model predictions are associated with uncertainties, so even a perfect model deviates from the empirical data.

To quantify the agreement between model predictions and empirical data while considering these uncertainties, the measure of goodness of a model (χ^2) was used:

$$\chi^2 = \sum_{i}^{N} \frac{(a_i - b_i)^2}{\alpha_i^2 + \beta_i^2} \tag{5.2}$$

where a_i and b_i are the empirical and predicted values, respectively, while α_i and β_i are the standard deviations of measurement and prediction uncertainties. Good agreement of the prediction B with a measurement A is reflected by a value of χ^2 close to the mean of the underlying χ^2-distribution. This is equal to the number of independently predicted b_i, also called the degrees of freedom ν.

A ratio χ^2/ν (also called normalized χ^2) close to one indicates that the model predictions deviate from empirical data by an amount consistent with the reported uncertainties. Values above one suggest that the model does not fully explain the data. $\chi^2/\nu \approx 0$ can only happen if measurement and prediction errors are correlated or overestimated.

The likelihood that the observed deviation arises purely from uncertainty can be evaluated using the χ^2-distribution. The probability $p_\nu \left(X > \chi^2\right)$ for the observed χ^2 is the likelihood that the mismatch between observation and prediction is the result of measurement errors and prediction uncertainties.

Commonly used statistical measures such as Pearson correlation coefficient (R_p), bias, and root mean square error (RMS) are included for comparability. In general, a good model should exhibit high R_p, low RMS, and bias and χ^2/ν in the range of unity.

5.3 Results

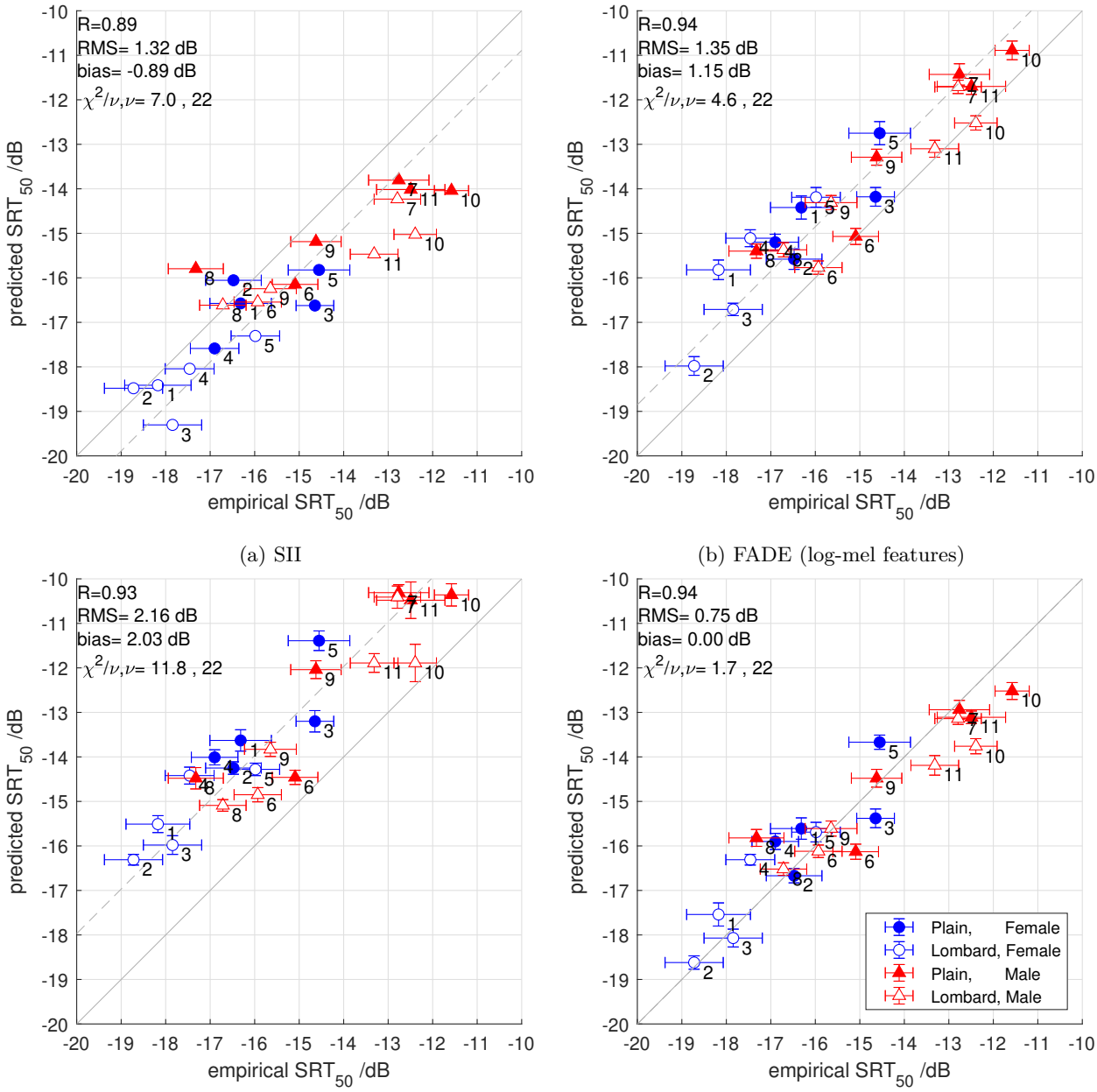
5.3.1 Absolute SRT50 prediction

Statistical comparisons of empirical and predicted SRTs are summarized in Table 5.1.

For absolute SRT predictions with the FADE model, the prediction accuracy strongly depended on the feature representation used. The best overall agreement with the empirical absolute SRTs was observed for FADE using SGBFB features with a high correlation (R_p=0.94), the lowest RMS error (0.75dB), and negligible bias (−0.003dB).

The normalized chi-squared was comparatively small ($\chi^2/\nu = 1.7$, $p_\nu = 0.021$), indicating that, although deviations remain statistically significant, the prediction errors were substantially closer to the expected uncertainty range than for any other model–feature combination. FADE using spectral features (log mel or MFCC) achieved similarly high correlations; however, the RMS error increased by a factor of about two for log-mel features and by almost a factor of three for MFCC features. Similarly, the bias increased markedly: predictions based on log-mel features showed a systematic offset of 1.15dB, whereas the MFCC-based predictions exhibited an even larger bias of 2.03dB. The normalized chi-squared values reflected these deviations, with $\chi^2/\nu = 4.6$ for log-mel and $\chi^2/\nu = 11.8$ for MFCC, indicating that both spectral feature sets failed to reproduce the empirical SRTs within the expected uncertainty range.

The SII performed less accurately than any FADE variant. Correlation was moderate (R_p=0.89), with an RMS error of 1.3dB. The bias (−0.89dB) was also noticeably larger in magnitude than for the FADE model with SGBFB. Most importantly, the normalized chi-squared remained high ($\chi^2/\nu = 7.1$, $p_\nu < 0.001$), indicating that the discrepancies between SII predictions and measured SRTs were larger than expected from the measurement uncertainty.



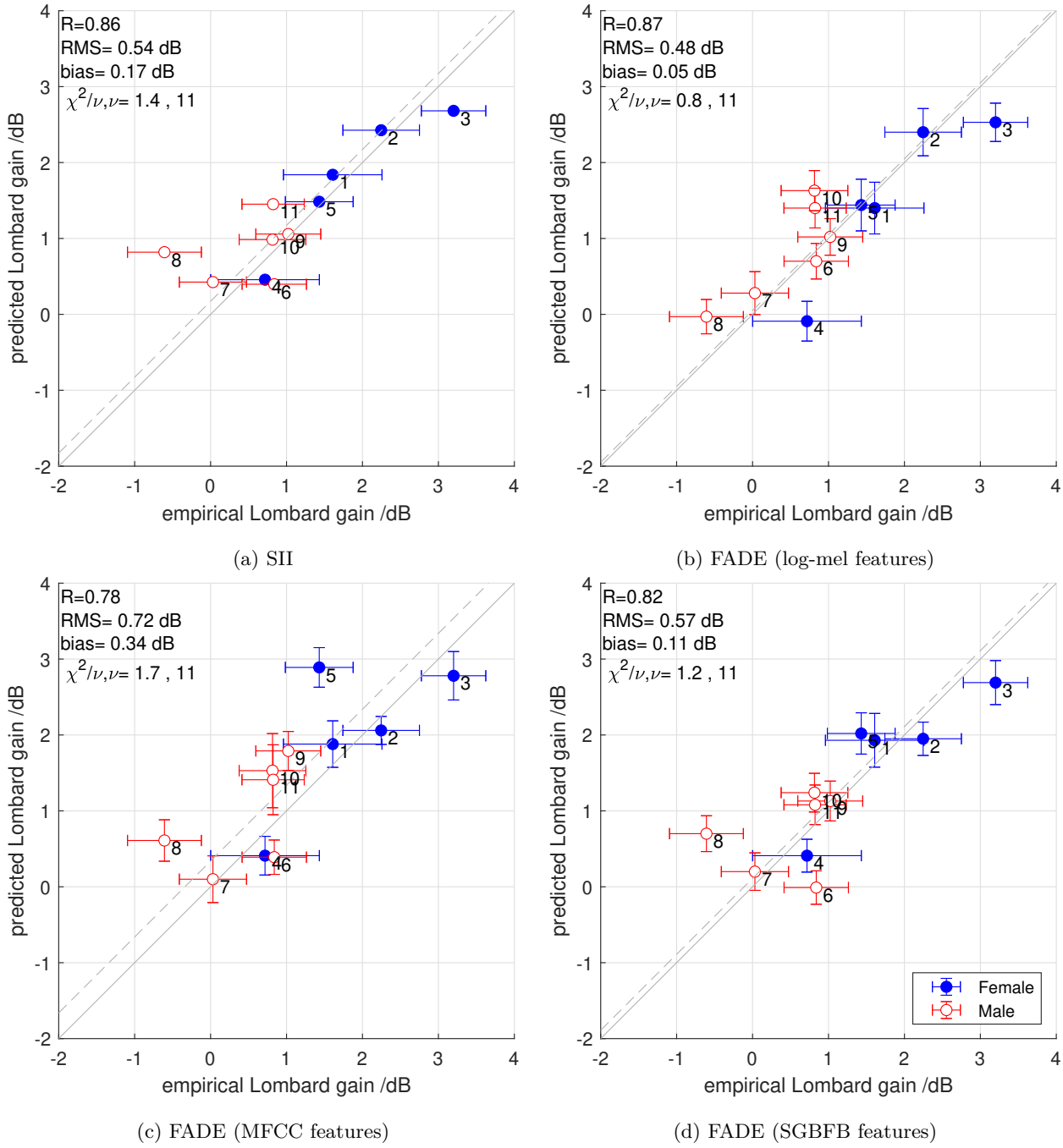
(a) SII

(b) FADE (log-mel features)

(c) FADE (MFCC features)

(d) FADE (SGBFB features)

Figure 5.1: Absolute SRT prediction of Mandarin Chinese in ICRA1:
Each speaker is assigned to a number. The dashed line represents a linear interpolation with a slope of unity. The bias is the offset between the dashed and continuous line. Measurement- as well as model uncertainties (reported as standard deviation) are displayed as error bars. **5.1a:** SII-based prediction with a Mandarin female reference speaker from the original Mandarin Matrix sentence test of Hu, Xi, et al. 2018 with gender-specific BIF of J. Chen, Huang, and Wu 2016.**5.1b** and **5.1c**: FADE prediction based on log-mel and MFCC features, respectively. These capture only spectral information.**5.1d:** FADE prediction based on SGBFB features, which encode spectro-temporal information.



(a) SII

(b) FADE (log-mel features)

(c) FADE (MFCC features)

(d) FADE (SGBFB features)

Figure 5.2: Lombard gain prediction of Mandarin Chinese in ICRA1:
Each speaker is assigned to the same number as in Figure 5.1. The dashed line represents a linear interpolation with a slope of unity. The bias is the offset between the dashed and continuous lines. Measurement as well as prediction uncertainties (reported as standard deviation) are displayed as error bars. **5.2a:** SII-based prediction with a Mandarin female reference speaker from the original Mandarin Matrix sentence test of Hu, Xi, et al. 2018 with gender-specific BIF of J. Chen, Huang, and Wu 2016. **5.2b** and **5.2c**: FADE prediction based on log-mel and MFCC features, respectively. These capture only spectral information.**5.2d:** FADE prediction based on SGBFB features, which encode spectro-temporal information. Overall, all approaches show equally accurate predictions.

Table 5.1: Absolute SRT and LG prediction of empirical SRTs for Mandarin Chinese in ICRA1: Displayed are SII-based prediction and FADE predictions with different frontends. Pearson correlation coefficient (R_p), root mean square error (RMS), and bias are provided. χ^2/ν gives the normalized chi-squared measure. $p_\nu\left(X > \chi^2\right)$ gives the probability for χ^2 above or equal to the reported value for a χ^2-distribution with ν degrees of freedom. (See section 5.2.4)

	model	frontend	R_p	RMS/dB	bias/dB	χ^2/ν	ν	$p_\nu\left(X > \chi^2\right)$
abs. SRT	FADE	log-mel	0.94	1.35	1.15	4.6	22	<0.001
	FADE	MFCC	0.93	2.16	2.03	11.8	22	<0.001
	FADE	SGBFB	0.94	0.75	-0.003	1.7	22	0.021
	SII	-	0.89	1.32	-0.89	7.1	22	<0.001
LG	FADE	log-mel	0.87	0.48	0.05	0.8	11	0.640
	FADE	MFCC	0.78	0.72	0.34	1.7	11	0.067
	FADE	SGBFB	0.82	0.57	0.11	1.2	11	0.280
	SII	-	0.86	0.54	0.17	1.4	11	0.165

The measured and predicted SRTs, separated by speaker and speech style, are presented in Figure 5.1. Although both models produced SRTs that were generally too low for speakers 10 and 11, the remaining deviations appeared unsystematic, with no identifiable dependence on speech style or speaker gender.

5.3.2 Lombard Gain Prediction

Statistical comparisons of measured and predicted LG are summarized in Table 5.1. FADE showed overall good agreement with the measured data. The log-mel features achieved the highest correlation ($R_p = 0.87$) together with the lowest error ($RMS = 0.48$dB) and negligible bias (0.05dB). The normalized chi-square value ($\chi^2/\nu = 0.8$,$p_\nu = 0.64$) indicated no significant difference between prediction mismatch and uncertainties.

FADE with SGBFB features, which was the best-performing configuration for absolute SRT predictions, also performed well for LG estimation. Its correlation ($R_p = 0.82$) and error ($RMS = 0.57$dB) were slightly poorer than those obtained with log-mel features. The fit quality ($\chi^2/\nu = 1.2$, $p_\nu = 0.28$) likewise indicated no statistically significant deviation from the data. Despite the small differences across metrics, SGBFB features maintained a low bias (0.11dB) and stable performance across tasks. FADE with MFCC features showed the weakest performance among FADE variants, reflected in lower correlation ($R_p = 0.78$), slightly increased prediction error ($RMS = 0.72$dB). However, the chi-squared statistic ($\chi^2/\nu = 1.7$, $p_\nu = 0.067$) indicates that the mismatch is only slightly larger than expected from measurement and prediction uncertainties. Its bias was more pronounced (0.34dB) but remained small in practical terms.

For comparison, the SII showed a similar correlation ($R_p = 0.86$) but a slightly higher error ($RMS = 0.54$dB). Its bias remained small (0.17dB), and the normalized chi-square value ($\chi^2/\nu = 1.4$, $p_\nu = 0.165$) indicated an acceptable, though less precise, fit than the best FADE variants.

5.4 Discussion

5.4.1 Absolute SRT50 Prediction

The present results demonstrate that spectro-temporal features are crucial for accurate, absolute SRT prediction. Speaker-specific SRTs can be predicted by both models; however, FADE with SGBFB features clearly outperformed all tested approaches. The SGBFB configuration of FADE showed the highest correlation with empirical data, the smallest RMS error, and no bias (-0.003dB). Nevertheless, the measurement uncertainties were sufficiently small to reveal that neither FADE nor SII is a perfect model, as the remaining variance is higher than the expected variance from measurement uncertainties ($p_\nu \leq 0.021$ for all models; see Table 5.1). This indicates that factors beyond measurement uncertainties contribute to the mismatch and that more accurate models are theoretically possible.

The superior performance of FADE with SGBFB features for absolute SRT prediction suggests that encoding temporal information is essential for a speaker-specific SRT prediction, since only SGBFB capture dynamic spectro-temporal cues. Features such as MFCC, log-mel spectrograms, or the long-term spectrum used in the SII cannot encode these temporal aspects, which limits their predictive accuracy.

These findings align with Schädler, Meyer, and Kollmeier 2012, who demonstrated the advantage of Gabor filter bank (GBFB) over MFCC in ASR tasks, highlighting the importance of spectro-temporal cues. Similar conclusions have been drawn across multiple languages, including German, Polish, Russian, Spanish, English, and Cantonese (Schädler, Warzybok, Ewert, et al. 2016a; M. K. Scharf et al. 2022), especially in fluctuating noise conditions (Schädler, Warzybok, Ewert, et al. 2016b). Together, these results reinforce the role of spectro-temporal encoding as a key determinant of absolute SRT predictions for speech in stationary noise.

5.4.2 Lombard Gain Prediction

In contrast to absolute SRT predictions, the prediction of the LG relied primarily on spectral cues. FADE predicted the LG accurately even when using only spectral features (log-mel or the derived MFCC), with normalized chi-squared close to unity, indicating that the remaining mismatch could be fully explained by prediction- and measurement uncertainties.

FADE with SGBFB features, although superior for absolute SRT predictions, did not provide additional predictive value for LG, suggesting that the temporal information is largely redundant for this measure. The SII showed comparable results, demonstrating that the information in the filtered long-term spectrum is sufficient to capture the relative improvements in SRT due to Lombard speech.

These findings are consistent with the acoustical properties of Lombard speech in Mandarin (F. Chen, Pan, et al. 2025), where spectral properties of Lombard speech have been shown to correlate with the LG.

5.4.3 General Remarks

This study examined the role of spectral and spectro-temporal acoustic cues in predicting both absolute SI and the LG in Mandarin Chinese. As only a single tonal language was considered here and a comparison to non-tonal languages is beyond the scope of the current study, we cannot resolve whether tonal languages make use of spectro-temporal encoding differently from non-tonal languages. However, we found that the spectro-temporal features of speech are essential for accurate SRT prediction of Mandarin Chinese, similar to non-tonal languages (Schädler, Warzybok, Ewert, et al. 2016a). This suggests that spectro-temporal representations are required to capture speaker-specific intelligibility differences, even when lexical tone is an inherent property of the language.

Previous work with bilingual speakers of English and Cantonese (M. K. Scharf et al. 2022) supports the view that spectro-temporal features are not language-specific but rather encode the physical properties of speech production adequately, such as time-dependent spectral filtering due to articulatory movements in the vocal tract.

In contrast, the prediction of the LG was equally accurate using FADE with either spectro-temporal or purely spectral features, as well as with the SII model. Within the measurement uncertainties of this study, spectro-temporal features did not provide additional predictive information beyond spectral cues for LG. This indicates that Lombard-induced intelligibility benefits in Mandarin are primarily driven by spectral changes, rather than by additional temporal information introduced by speaking rate or other dynamic modifications. At least in continuous ICRA1 noise, the Mandarin Chinese LG can be fully accounted for by spectral changes in the speech signal.

A limitation of this study is the size of the empirical dataset, which included 22 complete matrix tests from 13 listeners each. At the same time, it represents, to the authors' knowledge, the most extensive and most precise Mandarin plain/Lombard SRT dataset published. Nevertheless, this sample size may still be insufficient to detect subtle contributions of temporal cues to the LG, given the measurement uncertainty of approximately 0.5 dB. Accordingly, the absence of a spectro-temporal effect on LG should be interpreted as an upper bound on its potential contribution rather than definitive proof of its absence. Future studies with more listeners may clarify whether spectro-temporal features contain a small, previously undetectable influence on LG.

Similarly, although stationary, speech-shaped noise (ICRA1) was used here, other types of maskers, such as fluctuating or competing speech, may highlight temporal contributions that were not observed in the present study. The use of the Mandarin matrix test with monosyllabic and disyllabic words provides a controlled stimulus set. Still, future studies with everyday sentences or connected speech may uncover additional temporal effects, particularly those related to prosodic or contextual speech dynamics.

5.5 Conclusions

In summary, the main findings are:

1. Spectro-temporal speech features are essential for accurate absolute SRT prediction in Mandarin Chinese, capturing speaker-dependent differences in intelligibility.

2. Within the measurement uncertainty of this study, spectral features alone contain all necessary information to predict the Lombard

gain (LG) in continuous, speech-shaped noise (ICRA1) for Mandarin Chinese.

3. The speech intelligibility index (SII) and the framework of auditory discrimination experiments (FADE) are equally powerful models for predicting the LG. If only the LG, and not the absolute SRT, is of interest, the SII is preferable due to its lower complexity and faster calculations.

4. Future work using larger listener cohorts, additional noise types, and more realistic speech material is needed to determine whether subtle temporal contributions to LG emerge under less controlled conditions and whether these findings generalize across tonal and non-tonal languages.

Acknowledgements

Funding

This work was funded by: the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 415895050 - "Experiments and models of speech recognition across tonal and non-tonal language systems (EMSATON)", the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany' s Excellence Strategy –EXC 2177/1 - Project ID 390895286 - "Hearing4all" and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) –Project ID 352015383 –"SFB 1330 A5 Model-based diagnostics and hearing aid fitting in complex acoustic scenarios: Perceptive Principles, Algorithms and Applications", the Deutsche Forschungsgemeinschaft (DFG) - Project ID 465121786 "Deutsch-russische statistische Audiologie: Datenaufbereitung und audiologische Profil-Analyse für die Diagnostik von Hörstörungen und ihre Kompensation (GRUSTAD)"

Conflict Statement

The authors declare that there is no conflict of interest.

Data Availability

The FADE model is available under Schädler, Hülsmeier, Grimm, et al. 2021. The speech test audio used in this study is available in Hu, Warzybok, et al. 2022. The speech intelligibility (SI) data is available from F. Chen, Pan, et al. 2025.

Author Contributions

MS was responsible for the simulation, analysis, and writing of the manuscript. AW, LW, FC, and BK were responsible for the data and funding. AW and BK contributed to the analysis and writing.

Chapter 6:

Lombard Effect for Bilingual Speakers in Cantonese and English: importance of spectro-temporal features

reprint from
Interspeech 2022 Proceedings

Maximilian Karl Scharf[1] *,Sabine Hochmuth*[1] *, Lena L.N. Wong*[2] *, Anna Warzybok*[1] *, Birger Kollmeier*[1]

[1]Medical physics and cluster of excellence Hearing4all, CvO University Oldenburg, Germany
[2]Clinical Hearing Sciences (CHearS) Laboratory, Faculty of Education, The University of Hong Kong, Hong Kong SAR, China
maximilian.scharf@uni-oldenburg.de, anna.warzybok-oetjen@uni-oldenburg.de

DOI: 10.21437/Interspeech.2022-10235

Abstract

For a better understanding of the mechanisms underlying speech perception and the contribution of different signal features, computational models of speech recognition have a long tradition in hearing research. Due to the diverse range of situations in which speech needs to be recognized, these models need to be generalizable across many acoustic conditions, speakers, and languages.

This contribution examines the importance of different features for speech recognition predictions of plain and Lombard speech for English in comparison to Cantonese in stationary and modulated noise. While Cantonese is a tonal language that encodes information in spectro-temporal features, the Lombard effect is known to be associated with spectral changes in the speech signal. These contrasting properties of tonal languages and the Lombard effect form an interesting basis for the assessment of speech recognition models.

Here, an automatic speech recognition-based (automatic speech recognizer (ASR)) model using spectral or spectro-temporal features is evaluated with empirical data. The results indicate that spectro-temporal features are crucial in order to predict the speaker-specific speech recognition threshold speech recognition threshold (SRT) in both Cantonese and English, as well as to account for the improvement of speech recognition in modulated noise, while effects due to Lombard speech can already be predicted by spectral features.

Keywords: prediction of speech recognition, Lombard effect, tonal and non-tonal languages, speech features

6.1 Introduction

6.1.1 General aim

The aim of this study is to characterize the speech features involved in the Lombard effect for a comparison between a tonal language and a non-tonal language while minimizing effects due to variation between speakers.

6.1.2 Lombard Effect

Discovered in 1911 (Lombard 1911), the Lombard effect describes an increase in amplitude, duration, as well as a change in the spectrum of speech, together with a change in visual cues when the speaker is introduced to loud sound fields. More precisely, the Lombard effect leads to a continuous upward shift of the fundamental F0 frequency and formants F1, F2, as well as an increased average vowel duration (Alghamdi et al. 2018). This change in speech production is also referred to as increased vocal effort. Additionally, it has been reported by Junqua 1993 that the spectral change of increased vocal effort leads to improved speech recognition.

6.1.3 Tonal and non-tonal languages

Tonal languages, such as Cantonese, are spoken by a large fraction of the human world population and differ from non-tonal languages, such as English, in that information is encoded by changes in pitch within syllables. This has implications for the fundamental frequency f0, which will be modulated over time to encode information. Thus, modeling a tonal encoding of information requires a model that incorporates both spectral and temporal features as an internal representation of speech signals. The contrasting property of tonality with spectro-temporal encoding of information and the Lombard effect, which acts on the long-term spectrum of speech, makes it possible to assess the requirements on speech features that are used in speech recognition models. Established and highly successful models for non-tonal languages are commonly used to describe speech recognition by spectral features only (ANSIS3.5 1997). In the past, and with some success, these models have also been applied to tonal languages (J. Chen, Huang, and Wu 2016).

6.1.4 Spectral and spectro-temporal features of speech

The current literature (see above) suggests that an improvement of the speech recognition threshold due to Lombard speech is connected to spectral changes of speech, while tonal languages make use of spectro-temporal coding of information, which remains mostly unchanged during Lombard speech.

Acoustic speech features are commonly derived from a short-time Fourier transform (short time Fourier transform (STFT)) with logarithmically scaled frequency and amplitudes. For the frequency, an approximation of the perceptually motivated mel scale is typically used, which differs from $\log_{10}$ in the base of the logarithm as well as in the vertical and horizontal intercept (ETSI 2003):

$$\mathrm{Mel}(f) = \log_{10^{1/2595}}(1 + \frac{f}{700}) \qquad (6.1)$$

Automatic speech recognition systems can utilize this spectrum of a time-windowed section of a signal as a feature vector, which is updated periodically for a fixed time shift and can be displayed on a 2D-grid as a spectrogram, called a log-mel-spectrogram. A drawback of this data representation is that temporal changes do not enter the feature vector and need to be included by means of derivatives with respect to time.

One method to circumvent this problem is to calculate the Mel Frequency Cepstral Coefficients mel frequency cepstral coefficients (MFCC). MFCC take the log-mel-spectrogram and calculate the real part of its Fourier transform. Cepstra contain information about the temporal structure of a signal and allow to deconvolve, e.g., the glottal excitation signal from the vocal tract filter and the room impulse response that are characteristic for different quefrencies. If more information about the temporal structure beyond the range of the time window of the STFT is required, derivatives of the MFCC with respect to time are sometimes included in the feature vector.

Another method of collecting temporal information of speech for a feature vector is the use of Separable GaborFilter Banks, short separable gabor filter bank (SGBFB) (Schädler and Kollmeier 2015). These work like a set of activation maps for the log-mel-spectrogram under the constraint that the resulting features are uncorrelated. SGBFB can be calculated efficiently, because of their property that the required 2D-filter can be separated into a filter for the time- and frequency-axis. Due to the fact that these activation maps spread over many time samples, it is expected that SGBFB contain more relevant temporal information for speech recognition than MFCC. SGBFB have been shown to be superior to MFCC for accurate simulations of speech recognition in fluctuating noise conditions (Schädler, Warzybok, Ewert, et al. 2016a).

6.1.5 Influence of noise on speech recognition

Besides the characteristics of the tested speech, the spectro-temporal structure of the masking noise is relevant for the resulting speech recognition thresh-

old for which 50% of words are understood correctly (SRT). Generally, continuous noise leads to higher SRT than noise with a spectro-temporal structure, i.e., fluctuating noise, where prospective listeners are able to make extensive use of information that is present in sections of low signal to noise ratio (SNR), which is also referred to as "listening in the gaps". For the purpose of this study, two types of standardized noise were used (Wouter A. Dreschler et al. 2001; Kirsten Carola Wagener, Thomas Brand, and Kollmeier 2006): unmodulated speech noise with a long-term spectrum of a male speaker at normal vocal effort (Wouter A. Dreschler et al. 2001) (ICRA1), which is a stationary signal with speech-like spectrum, and modulated ICRA1 noise with a duration of pauses no longer than 250ms (Kirsten Carola Wagener, Thomas Brand, and Kollmeier 2006) (ICRA5-250), which has a temporal structure with gaps no longer than 250ms, such that the SRT of icra5-250 is expected to be lower than for ICRA1.

6.1.6 Hypothesis

This study aims to model the absolute SRT together with the Lombard Gain, which is the difference between SRT in Plain and Lombard speech, for bilingual speakers of Cantonese and English. Because of the different speech features that are relevant for the prediction of the Lombard gain and an absolute SRT prediction of tonal languages, the following hypotheses are tested:

1. spectro-temporal features are important to describe the speaker specific SRT across different types of masker while
2. spectral features suffice to describe the Lombard gain.

6.2 Methods

6.2.1 Empirical Data

Two female and two male bilingual talkers recorded the matrix sentence test corpus (Kollmeier, Warzybok, et al. 2015) in both Cantonese (Hong Kong variety dialect of Cantonese that is spoken in Hong Kong (YUE)) and English for plain and Lombard speech each. For the purpose of this study, the Matrix sentence test was chosen, as it consists of a fixed grammatical structure with unpredictable content, which does not contain any information encoded in stress patterns. The Lombard effect was induced by ICRA1 noise with 80 decibel (dB)-sound pressure level (SPL). For obtaining the SRT, these matrix sentence recognition tests were conducted with 15 normal hearing listeners, who were monolingual natives of the given language, in two different noise conditions: ICRA1 and ICRA5-250. The adaptive procedure with word-scoring used for this test was according to Thomas Brand and Kollmeier 2002, and each adaptive track was scanned manually for any abnormality prior to further processing. The absolute SRT of a speaker was calculated as the mean across all listeners, while the Lombard gain was calculated as the mean across the individual Lombard gains for each listener. The measurement error of the mean σ_μ was estimated by:

$$\sigma_\mu \approx \sigma/\sqrt{N} \tag{6.2}$$

Where σ is the standard deviation across N listeners.

6.2.2 FADE Modeling approach

The Framework of Auditory Discrimination Experiments' (FADE) approach to speech recognition threshold modeling as proposed by Schädler, Warzybok, Hochmuth, et al. 2015 is based on an automatic speech recognizer (ASR), which acts as a model of a well-trained human listener. The FADE approach supports arbitrary types of features, based on the speech signals, to enter the ASR back end. For the simulation, the publicly available FADE distribution (Schädler, Hülsmeier, Grimm, et al. 2021) was used with default settings.

The ASR-based SRT prediction consisted of the following steps: first, the audio of each sentence of the used test lists was mixed with random sections of the masking noise for N different SNR. Second, N Hidden Markov Models in conjunction with a Gaussian Mixture Model (hidden Markov model in conjunction with a Gaussian mixture model (HMM-GMM)) were trained on the features that were derived from the labeled signals for each SNR. The log-mel-spectrogram, from which the features were calculated, was based on a STFT with a 25ms hamming window and a time shift of 10ms. The underlying Markov model assumed 16 consecutive states per word and four states for silence, beginning or end of a sentence. Additionally, only transitions between words according to the matrix test syntax were possible. As a third step, a new set of K matrix test audio signals, mixed with random sections of noise for a range of K different SNR, was used to test the performance of the N differently trained models. The resulting model performances as a function of N training- and K testing-SNR were used to calculate the SRT in a last step: the lowest value of the K testing SNR for a rate of 50% correctly identified words is returned as SRT. As the simulation is done for discrete values of testing SNR, the final prediction of the SRT is based on an interpolation of the data. Uncertainties in the model prediction originate from the size of the data set used for testing.

The FADE simulations were done for every combination of the two types of input features for the ASR (MFCC, SGBFB), speakers (4), language (ENG, YUE), vocal efforts (plain, Lombard), and masking noises (ICRA1, ICRA5-250), resulting in 64 computations.

6.2.3 Statistical measures

The predictive power of the simulations was quantified by the Pearson correlation coefficient Pearson correlation coefficient (R_p), the root mean square error root mean square error (RMS), the bias of a linear regression, and the χ^2/ν measure with ν degrees of freedom. χ^2/ν respects both measurement- σ_{emp} and simulation- uncertainty σ_{sim}:

$$\chi^2 = \sum_n \frac{(\text{sim}_n - \text{emp}_n)^2}{\sigma^2_{\text{emp},n} + \sigma^2_{\text{sim},n}} \tag{6.3}$$

Where sim_n denotes the result of a single simulation with index n and emp_n stands for the corresponding empirical measurement.

6.3 Results

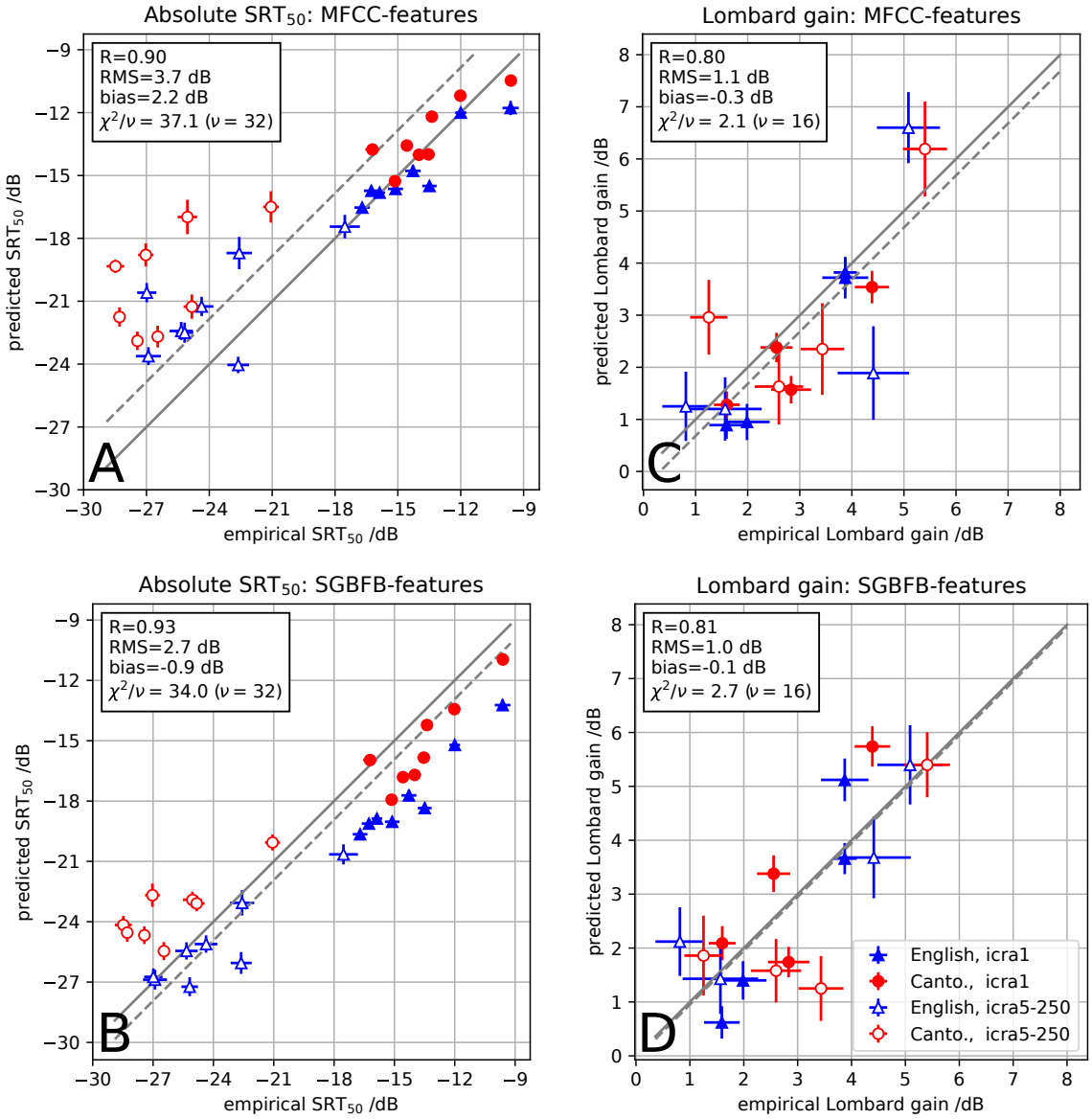


Figure 6.1: SRT and Lombard gain prediction of FADE:
with MFCC (A and C, respectively) and SGBFB features (B and D, respectively) for four bilingual speakers in stationary noise (ICRA1) and fluctuating noise (ICRA5-250). The solid line corresponds to perfect correlation, the dashed line depicts a linear interpolation.

6.3.1 Simulation results

The scatter plots of predicted vs empirically obtained speech recognition thresholds (SRT) are shown in figure 6.1, where the prediction of the absolute SRT as well as Lombard Gain is displayed for MFCC- and SGBFB-based predictions. The predictions and empirical findings of absolute SRT correlate with $R_p \geq 0.9$ for both feature sets.

Panel A and B show that the FADE prediction based on SGBFB features was able to describe the data with the smallest RMS, offset, and χ^2/ν where ν is the degrees of freedom of the modeled data. As the FADE prediction method is only working with the sound files, no information about the empirical SRT has flown into the prediction, such that ν is equal to the number of data points.

The MFCC based prediction in panel A showed a good agreement between simulations and empirical data in stationary noise ICRA1 (R_p=0.85). However, the correlation is weak for the ICRA5-250 noise condition (R_p=0.44), which is in line with previous findings (Schädler, Warzybok, Ewert, et al. 2016a) that FADE can predict the SRT in fluctuating noise better with spectro-temporal feature sets (e.g., SGBFB) than with either temporal or spectral feature sets in isolation. The MFCC are not sufficient to take advantage of the temporal gaps in the masker as efficiently as human listeners do. This observation holds for both languages. On the other hand, the predictions based on SGBFB features of panel B coincide very well with the empirical data for both noise conditions. A negative bias of -2.6dB for ICRA1 remained for SGBFB features.

Panels C and D show the Lombard gain, which is the difference between the SRT in plain and Lombard speech. Accordingly, the remaining degrees of freedom of the data ν are cut in half. Both feature sets show high correlation with vanishing bias together with a value of χ^2/ν close to unity, which means that the deviation of the model from perfect correlation with the measurement is in agreement with the uncertainties. It can be observed that the Lombard gain prediction was independent of the used features.

6.4 Discussion

The discussion will distinguish between the absolute SRT prediction and prediction of the Lombard gain according to the two initial hypotheses that spectro-temporal features are important to describe the speaker-specific SRT while spectral features suffice to describe the Lombard gain.

6.4.1 Absolute SRT prediction

Panel A and B of figure 6.1 show the absolute SRT prediction for the two different speech features MFCC in A and SGBFB in B.

Based on the calculated statistical measures, the SGBFB based prediction was superior to the MFCC. Especially in the fluctuating noise condition ICRA5-250, MFCC based predictions failed to describe the empirical data, which is in agreement with known properties of MFCC (Schädler, Warzybok, Ewert, et al. 2016a). However, SRT in ICRA1 was correctly predicted with MFCC for both languages, which means that the MFCC features provided to the ASR backend of FADE were sufficient to predict SRT for both English and the tonal language Cantonese in stationary ICRA1 noise. On the other hand, the SGBFB based prediction worked for all noise conditions, such that the individual SRT of a speaker was predicted with higher accuracy, which is in line with hypothesis 1, that spectro-temporal features are important to describe the individual SRT. Even so, a bias of -2.6dB is present for the SGBFB based prediction in the stationary ICRA1 noise, that is expected: FADE works as a model of an ideally trained listener, which may lead to a negative offset, if the listeners employed are neither highly trained nor selected to be free from any minor sign of a hearing deficiency. Note that according to Hülsmeier, Buhl, et al. 2021, even normal-hearing listeners have to be modeled with a small processing deficit (denoted as level uncertainty) to obtain precise model predictions with FADE.

Even though SGBFB features resulted in the best prediction of the absolute SRT, as can be seen in figure 6.1 panel B, χ^2/ν is not in the range of unity, which shows that the model, or the applied features, still need improvement to describe the data to full agreement with measurement uncertainties.

6.4.2 Lombard gain prediction

Panel C and D of figure 6.1 show the Lombard gain prediction for the two different speech features MFCC in C and SGBFB in D.

As can be seen from the statistical measures with similar correlation coefficient, RMS of about 1dB and bias below 0.3dB, the two prediction methods are equivalent in their predictive power. χ^2/ν is close to unity for both models, which means that the deviation of the prediction from perfect correlation with the empirical data is in agreement with measurement uncertainties together with the precision of the model. It is evident that the Lombard gain, which is associated with spectral changes of speech, can be adequately modelled by MFCC. MFCC can only contain information from within

the time window of the STFT, which was in this case 25 ms. These findings provide evidence for hypothesis 2 that spectral features suffice to describe the Lombard gain.

6.5 Conclusions

The present study has shown that ASR based models of speech recognition such as FADE are a powerful tool for the prediction of SRT in stationary and fluctuating noise conditions for Cantonese and English. However, the choice of input features is important for a correct prediction of the individual SRT of a speaker. SGBFB features, which are similar to activation maps for the log-mel-spectrogram, were superior to MFCC, which can only contain information limited by the time window of the STFT. On the other hand, it was shown that the Lombard gain can be predicted by both MFCC as well as SGBFB, because changes in SRT due to the Lombard effect appear to be mainly associated with a spectral change of speech. In summary, it was found that:

1. spectro-temporal features such as SGBFB are important to predict the SRT for both Cantonese and English for speakers in stationary and in modulated noise.
2. spectral features such as MFCC suffice to describe the Lombard gain of a speaker, for both stationary and modulated noise.

Although SGBFB have been shown to be superior to MFCC for individual SRT prediction, the model was not able to compute predictions with maximum possible agreement with empirical data, which will require further investigations of the model or the used features.

Furthermore, the aspect of differences between tonal and non-tonal languages has not yet been fully explored in this contribution. Since bilingual speakers were used, further analysis of the data may uncover if significant differences in the Lombard Gain for Cantonese and English exist even for the same speaker. This might be related to the usage of different speech cues for enhancing speech recognition via the Lombard effect across languages.

Funding

This work was supported by the DFG grant "Experiments and models of speech recognition across tonal and non-tonal language systems (EMSATON)" grant number 415895050.

Chapter 7:

Outlook

7.1 General Discussion and Summary

This work showed that the Lombard gain (LG) in the two tonal languages, Mandarin Chinese and Cantonese, as well as the non-tonal language English, can be fully explained by spectral properties of speech. To show this, we used the automatic speech recognizer (ASR)-based framework of auditory discrimination experiments (FADE) and the widely used speech intelligibility index (SII). While there may be spectro-temporal changes between plain and Lombard speech, which may contain additional information, these were shown to be of negligible importance for exact predictions.

Furthermore, it was shown that the band importance function (BIF), which is commonly used to tailor the SII to a specific speech test, does not generalize between comparable conditions in Mandarin Chinese. This observation has potentially far-reaching consequences, as the SII is an ideal objective function for machine learning approaches to audiological problems. For speech intelligibility (SI) prediction of the Mandarin speech test in unmodulated speech noise, any BIF could be used with the same predictive power. For machine learning approaches that specifically target speech processing of Mandarin Chinese, the SII with language-specific BIFs might not be an adequate objective function for optimization.

As we were able to reproduce literature findings that the traditional SII is not a suitable SI model for absolute speech recognition threshold (SRT) predictions, our findings underline the necessity of a feature extraction that captures relevant spectro-temporal cues and does not require an unmixed masker and target for the prediction task. For this purpose, machine learning approaches are promising models. Besides FADE, which was used in this work, other machine learning SI models have been proposed (Roßbach, Kirsten C. Wagener, and Meyer 2023; Chiang et al. 2024; Saddler and McDermott 2024).

The model of inconsistent psychometric tracks that has been developed as part of this book is a novel method to screen experimental data. How to handle inconsistent responses during psychometric measurement procedures has been studied by numerous authors. However, this is the first time that a complete model of the underlying process has been proposed, which may generalize to any measurement procedure. Applications may range from fully automatic hearing tests with humans or animals to the assessment of attention.

Furthermore, this book introduced a calibration procedure for smartphone microphones with a measurement uncertainty of 3.0dB with a reference level of 92.8 ± 1.6dB-SPL. This level uncertainty is small enough for meaningful audiological measurements. As this procedure only requires an empty, standardized glass bottle, it is a powerful, first-order approximation to a microphone calibration for users in low to middle-income countries. Calibrating microphones this way may become part of a headphone calibration protocol for mobile audiological mobile health (mHealth) applications without the need for expensive equipment.

7.2 Future Research Directions

7.2.1 SI Prediction of the Lombard Effect in Tonal Languages

The conclusions of the presented LG predictions depend on the measurement uncertainty and, thereby, the number of listening experiments. The Lombard effect is known to influence temporal cues of speech (section 1.5). However, temporal features were not required to predict the SI correctly. It might be possible to resolve these lower-order temporal effects with more listening experiments of the plain and Lombard matrix tests, which are published under Hu, Warzybok, et al. 2022.

The mean measurement uncertainty of the Mandarin Plain-Lombard data in chapters 5 and 4 is 0.58dB for the average SRT for each speaker. To halve this uncertainty, one would require four times more participants in listening experiments, which is 52 listeners performing 22 different speech tests. Together with an overhead for training and failed adaptive procedures, this would require in the order of 1500 matrix tests. Such a study could be possible with the help of the virtual hearing clinic (Schell-Majoor, Büchner, and Kollmeier 2023) and parallel consistency screening with the single psychometric function goodness of fit measure (SiPGOF) that was proposed in chapter 3.

7.2.2 Non-monotonous Psychometric Function in Tonal Languages

During the optimization phase of the Cantonese matrix test (see section 1.2.2), it was observed that the optimization procedure did not converge on a homogeneous SRT (Hu, Hochmuth, et al. 2022). The optimization procedure of matrix sentence tests makes a linear approximation of the word-specific psychometric function. It adjusts the SRT of each word proportional to the slope of the psychometric function. A similar effect was observed for SRT measurements of plain and Lombard Cantonese speech in reverberation, where the standard adaptive procedure of the matrix test (Thomas Brand and Kollmeier 2002) did not converge. This procedure also assumes a monotonous, sigmoid psychometric function.

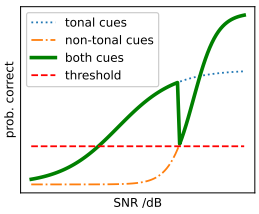


Figure 7.1: Non-monotonous psychometric function from two underlying psychometric functions:
If the probability of correct recognition with the most salient cues falls below a certain threshold, other cues such as tonal features may be used. This might lead to a non-monotonic psychometric function.

One hypothesis is that the word-specific psychometric function is not a monotonous function of the SRT. The cause for non-monotonicity could be two interweaved psychometric functions, similar to the consistency measure in chapter 3. It is known that tonal cues are more robust to band pass filters than consonants and vowels (Fu et al. 1998) and that they carry enough information to reconstruct sentences (see section 1.4).

This suggests that the psychometric function for word recognition might be split up into two separate components: one for tonal and one for spectral features. A listener could, for instance, default to the most salient features and switch to a more robust, but less informative encoding in challenging listening conditions where the information content of the default cues is below a certain threshold (see figure 7.1). A non-monotonic psychometric function could lead to the observations above.

7.2.3 Consistency Measure for Psychometric Experiments

The threshold for a classifier of consistency with the SiPGOF (see chapter 3) was derived for the German male matrix test. Similar thresholds could be derived for other matrix tests with a database of tracks. Ideally, these tracks should be manually screened for inconsistency by experts to achieve a good estimate of the false positive rate (rating consistent tracks as inconsistent).

The SiPGOF could also be extended to audiograms. Such an extension would only require minor changes to the model, as it was built to generalize to other tests. This would require a study in which experts rate a database of tracks from audiogram measurements. These ratings could then be used to verify the predictive power of the SiPGOF for audiograms. A threshold could also be derived from an unrated database of these tracks, as was done for the matrix test.

Further practical applications of the SiPGOF could be an automatic second opinion with a traffic light coding in a measurement software, or an objective rating of the performance of animal models during tests.

The SiPGOF could be extended to a fully Bayesian Model, which, instead of computing the maximum log-likelihood (l_{max}) for a model with one or more states, could be used to predict the most likely number of internal states directly. This would directly lead to a

binary classification into an underlying model with a single or multiple states, corresponding to consistent and inconsistent response behavior, respectively. An advantage of this approach is that a classification would not require a threshold, which, for the current implementation of the SiPGOF, has to be calculated from empirical data.

7.2.4 Headphone Calibration Procedure

Chapter 2 only reported the first step of a headphone calibration: the calibration of a reference microphone. The subsequent step, the calibration of headphones with the help of the microphone, has not been demonstrated so far. From theoretical considerations, we can estimate this additional measurement uncertainty to be below the measurement uncertainty of the reference level. As these uncertainties are uncorrelated, their combined effect should remain below the 5dB uncertainty of clinical audiograms. This should be demonstrated experimentally in a future study.

7.2.5 Database of Reference Levels

The calibration procedure of chapter 2 should be extended to more bottle shapes, as it is currently only available for the 330ml Vichy-shape bottle, which is the most common bottle for small beverages in Germany. A larger database of the average maximum achievable sound pressure level on bottle-flutes would allow for simple calibration in other regions than Germany. A database of these reference levels could help local stakeholders and developers of mHealth products to achieve smaller measurement errors with their measurements in comparison to entirely uncalibrated equipment.

This project requires an active community that can monitor the local market of bottles and contribute reference measurements. Therefore, it might be necessary to kickstart such a community with a sufficiently large database that external contributors can then extend. This project proposal has been nominated for the second revision of the *World Hearing Forum Students Changemakers Award 2025*.

Appendix A:

Statutory Statements

Author's Statement and Use of AI

This text was written in LaTeXwith Overleaf, and a Dropbox extension to enable fast switching between platforms as well as automated backups. The citations in the literature reference were managed with JabRef. Graphics were created with Python, Matlab, PowerPoint, Inkscape, and TikZ. Among the spellcheckers were Grammarly, Languagetools, and Hemingway editor, which include GPT models.

This thesis was written by myself (see Table A.1).

Thanks

I want to thank the Collaborative Research Centre SFB 1330 "Hörakustik" for the scholarship at the beginning of my studies in Oldenburg. Furthermore, I would like to thank the German Research Foundation (DFG) for funding the grant "Experiments and models of speech recognition across tonal and non-tonal language systems (EMSATON)" grant number 415895050 that made this research possible. I also thank the graduate schools of the University of Oldenburg and the Excellence Cluster "Hearing4All" for supporting my studies.

Most of all, I would like to thank Prof. Kollmeier, who made most of this possible and gave me the great freedom to pursue this research. I also thank the working groups of Prof. Wong in Hong Kong and Prof. Chen in Shenzhen, without whom this topic wouldn't have been possible. My thanks go out to my great colleagues and the hearing science community in Oldenburg, who are more than can be named here and have helped me numerous times.

Finally, I want to thank my wife An-Yi for her loving support and thoughtful discussions about my research.

Table A.1: List of own contributions to the published chapters:

Chapter	Own contribution
2 **Microphone Calibration Estimation for Mobile Audiological Tests with Resonating Bottles**	• Development of the method • Data collection of the uncalibrated measurements and analysis • Writing the manuscript
3 **A Consistency Measure for Psychometric Measurements**	• Development of the method • Simulation and analysis • Writing the manuscript
4 **Band Importance Functions For Mandarin Chinese Do Not Generalize Across Similar Speech Material**	• Model predictions and analysis • Writing the manuscript
5 **Spectro-temporal vs. Spectral Features to Predict the Lombard Gain in Mandarin Chinese**	• Model predictions and analysis • Writing the manuscript
6 **Lombard Effect for Bilingual Speakers in Cantonese and English: importance of spectro-temporal features**	• Model predictions and analysis • Writing the manuscript

Appendix B:

Angle-resolved Frequency Response of Smartphones

Since no reliable information on the projected directionality and transfer functions of typical smartphones is published, these were measured with a loudspeaker array of a TASCAR experiment (Grimm and Herzke 2025) to provide a first estimate of the hardware for the study of chapter 2. The experiment included four smartphones. Because these four smartphones are not a representative sample, the results were not published with the article in chapter 2. However, a few general observations can be made, which are discussed in the following.

Methods

The experimental setup is shown in Figure B.1. The smartphone was held above a calibrated microphone inside a loudspeaker array. The calibrated microphone was used to measure and correct the frequency response of each loudspeaker, such that the frequency response of each loudspeaker was flat with a standard deviation (STD) of 1 decibel (dB). 17 Frequency sweeps were played back from each loudspeaker with a TASCAR scene and recorded with the smartphones. The recording was deconvoluted with the frequency sweep (see Müller and Massarani 2001) and the reflections of the room removed with a rectangular time window. The forward direction was defined as pointing upwards from the built-in microphone in Figure B.1. The directions of the loudspeaker array were interpolated to a full sphere with a Voronoi tessellation.

Figure B.1: Test setup of a smartphone calibration:
A smartphone was placed on top of a calibrated microphone (microphone facing upwards). Frequency responses were recorded at several angles of incidence.

Results and Discussion

The frequency responses of the non-representative smartphone sample without protective cover in the 45° forward direction are shown in the top panels of figures B.3,B.4,B.5, and B.6. The average microphone sensitivity[1] in the frequency range (0-300) Hz and (1-1.3) kHz as a function of direction is shown in the lower two panels.

The frequency response of the entire sample in the 45° forward direction is compared in Figure B.2 for a calibration with the calibration estimation procedure of chapter 2.

From the few smartphones that were tested, we found that the resonance frequency of the 330ml bottles ($161.5-176$ Hz, see chapter 2) is just high enough for a reasonable calibration. Larger bottles with lower resonance frequency according to Equation 2.2 may not lead to a consistent calibration because of the high-pass behavior of the microphones (see Figure B.2).

We also found substantial differences between the frequency responses of the smartphones. The smartphone that was built to withstand rough environments (Figure B.6) did not show a flat transfer function, which is a necessary condition to calibrate with a single

[1]Note the difference between microphone sensitivity and the sensitivity of a binary discrimination test (see chapter 3).

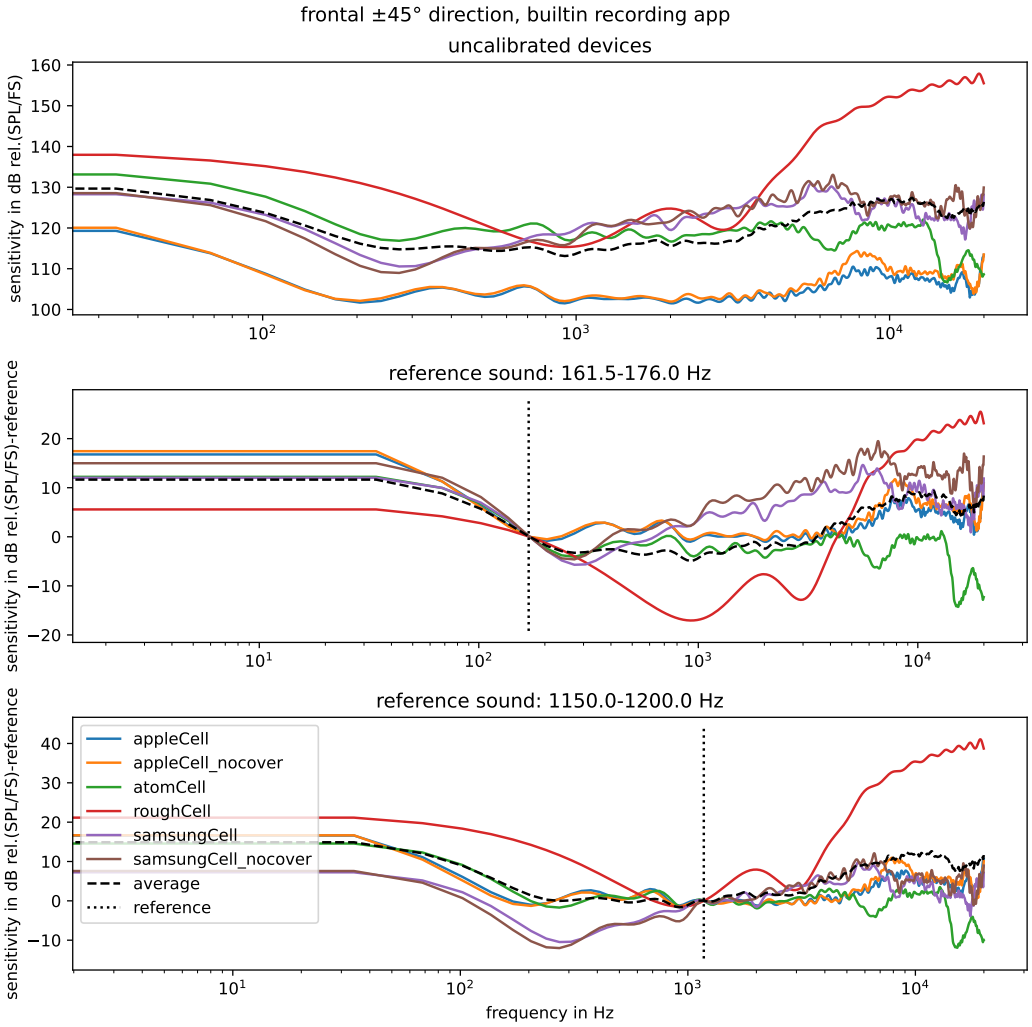


Figure B.2: Mean frequency response of the microphones before and after a calibration at $167 - 177$ and $1150 - 1200$ Hz according to the calibration estimation procedure of chapter 2. Frequency responses are measured for a $\pm 45°$ cone in the forward direction.

reference level and unknown transfer function. In the case of a non-flat transfer function, it might be reasonable to calibrate with more than one resonance frequency of the bottle and interpolate the frequency response. Higher resonance frequencies are produced by blowing with a faster air stream at the opening of the bottle-flute according to Equation 2.1. The second resonance frequency of 330ml bottles is in the range of 1150-1200Hz. Calibration with this resonance frequency is shown in the lower panel of Figure B.2.

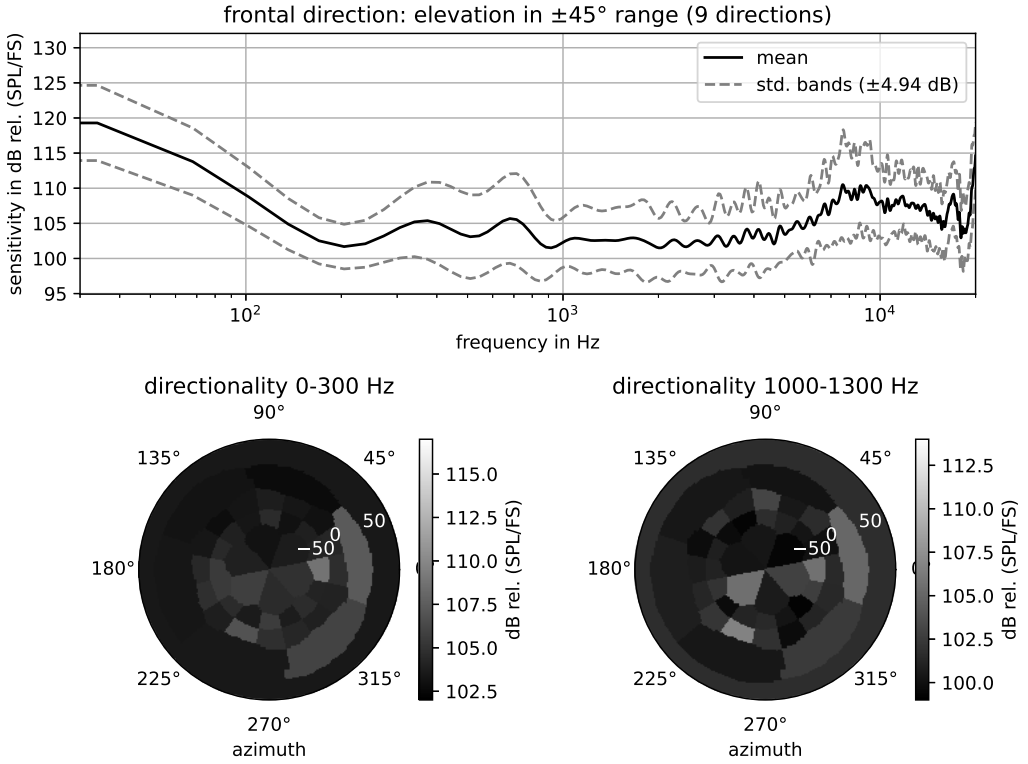


Figure B.3: Frequency response and projected directionality of an Apple iPhone 5 microphone:
Upper panel: average frequency response in forward direction. **Lower panels:** projected directionality for the average sensitivity in the frequency range (0-300) Hz and (1-1.3) kHz.

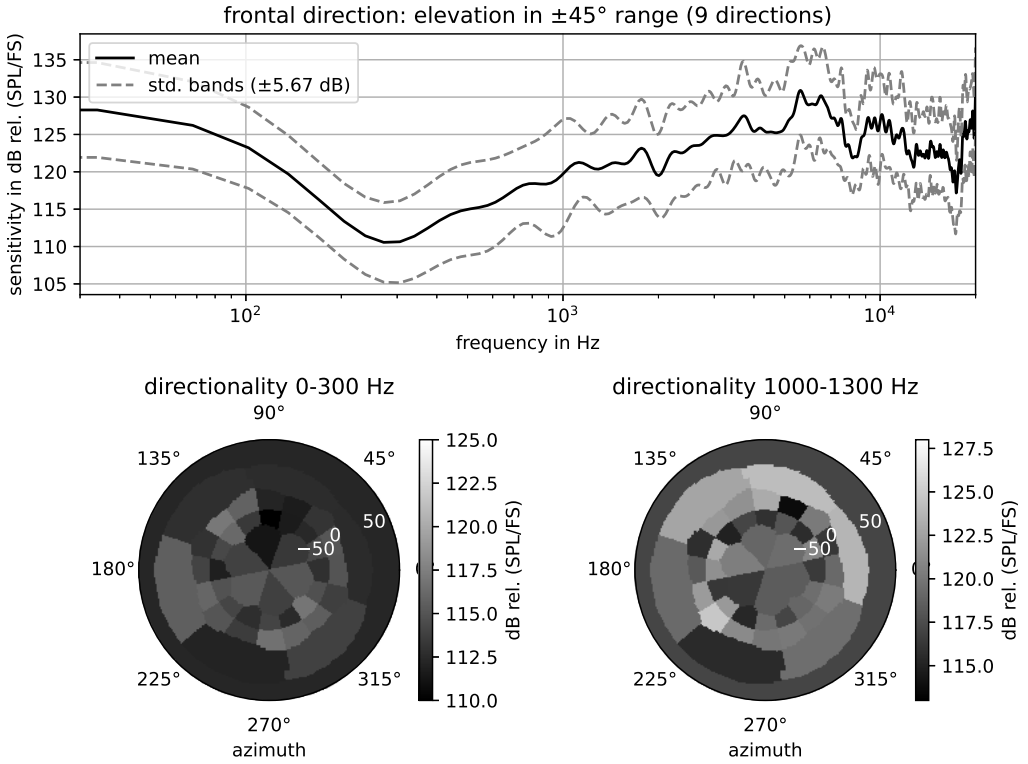


Figure B.4: Frequency response and projected directionality of a Samsung Galaxy microphone:
Upper panel: average frequency response in forward direction. **Lower panels:** projected directionality for the average sensitivity in the frequency range (0-300) Hz and (1-1.3) kHz.

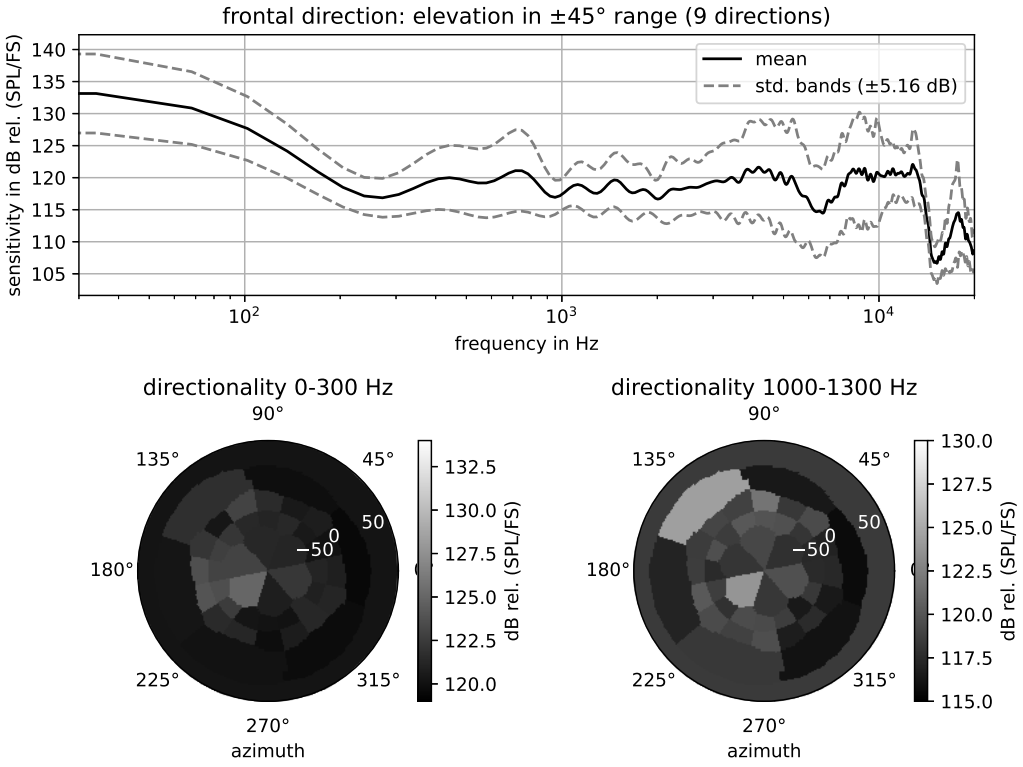


Figure B.5: Frequency response and projected directionality of an Unihertz Atom microphone:
Upper panel: average frequency response in forward direction. **Lower panels:** projected directionality for the average sensitivity in the frequency range (0-300) Hz and (1-1.3) kHz.

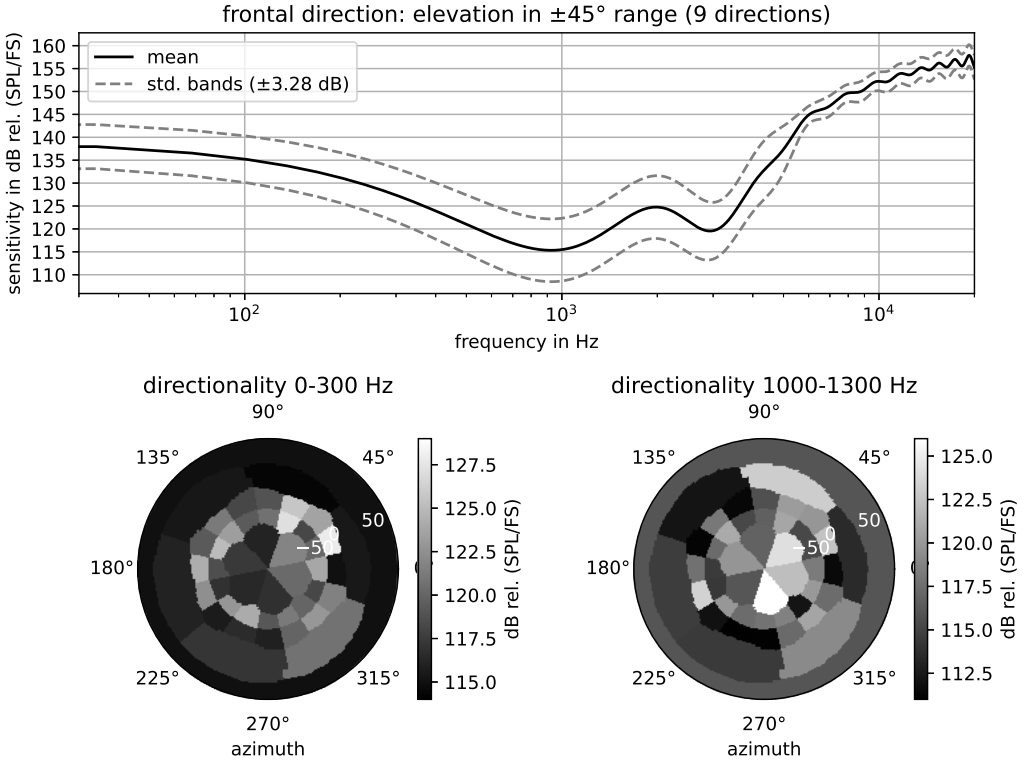


Figure B.6: Frequency response and projected directionality of a rough-phone microphone:
Upper panel: average frequency response in forward direction. **Lower panels:** projected directionality for the average sensitivity in the frequency range (0-300) Hz and (1-1.3) kHz.

Appendix C:

Reference condition for the SII

For low signal to noise ratios (SNRs), the SII is not a linear function (see figure 1.12). In this region of low[1] SNR, the SII exhibits a shallower slope, which makes the mapping of SII to SRT (Equation 1.36) more sensitive to small deviations of the reference SII. In the study of chapter 5 the reference SII_{ref} was 0.0408, which according to Thomas Brand, Roettges, and C. Hauth 2021 might lead to unreliable predictions.

Methods

To mitigate this, we applied the reference shifting method of Thomas Brand, Roettges, and C. Hauth 2021 to predict the SRT of chapter 5 with the SII. For this, the SRT for the SII_{ref} was temporarily increased by 6dB to bring the calculation into the linear range. After computing the predicted SRT with the mapping procedure (described in 1.3.5), the resulting SRT value was shifted back by -6dB.

The empirical data is from chapter 5, and the analysis follows the methods of this chapter.

Results

In Table C.1, the predictions with and without shifting are compared to the prediction with the mean SII of the empirical data.

For absolute SRT prediction, the +6dB shifting method of the reference condition leads to a smaller bias (-0.15dB vs. -0.89dB, while Pearson correlation coefficient (R_p) and root mean square error (RMS) stay comparable to the prediction without +6dB shift. However, the prediction accuracy remains poor, with $\chi^2/\nu = 6.8$. The mean SII reference leads to the best prediction of the absolute SRT

Both reference condition and mean SII lead to a LG prediction that is in agreement with the measurement uncertainty, while the prediction with the +6dB shifting method does not. Of all LG predictions with the SII, it has the largest RMS, absolute bias, χ^2/ν, and smallest R_p.

[1] The SII is approximately linear over the interval (-15,15)dB.

Table C.1: Comparison of the reference condition for absolute SRT and LG prediction of empirical SRTs for Mandarin Chinese in ICRA1 with the SII:
Displayed are SII-based predictions with different reference conditions. Pearson correlation coefficient (R_p), root mean square error (RMS), and bias are provided. χ^2/ν gives the normalized chi-squared measure. $p_\nu\left(X > \chi^2\right)$ gives the probability for χ^2 above or equal to the reported value for a χ^2-distribution with ν degrees of freedom (see section 1.3.6).

	method	R_p	RMS/dB	bias/dB	χ^2/ν	ν	$p_\nu\left(X > \chi^2\right)$
SRT	reference condition	0.89	1.32	-0.89	7.1	22	<0.001
	shifted reference condition	0.86	1.31	-0.15	6.8	22	<0.001
	mean empirical SII	0.88	1.08	0.18	4.2	21	<0.001
LG	reference condition	0.86	0.54	0.17	1.4	11	0.165
	shifted reference condition	0.81	0.70	-0.21	2.4	11	0.006
	mean empirical SII	0.85	0.55	0.16	1.5	10.5	0.128

Discussion

The +6dB shifted reference condition did not significantly improve prediction accuracy. This was not expected, because such a correction is supposed to be more robust against small deviations of the reference condition from an idealized reference that is maximally comparable to the data. Moreover, the +6dB shifted reference condition resulted in a worse LG prediction than for the other SII references. Hence, for a straightforward calculation of SII-based predictions, corrections of nonlinearity are not beneficial in the case of this dataset.

We observed that the SII prediction of the absolute SRT was the least accurate when the reference condition was taken from the original female speaker of the Mandarin matrix test. This is because of the larger bias in comparison to the mean SII with reduced degrees of freedom ν. However, the LG prediction was similar for the two reference conditions and in agreement with the empirical data, which is reflected by χ^2/ν close to unity.

The average SII reference had the highest absolute SRT prediction accuracy because of a smaller bias. Accordingly, the mean SII of the empirical data may be preferred over a reference SII of a different but comparable test. However, depending on the application, this may pose a problem in cases of small samples or even predictions of a single SRT: The mean SII reference will not produce a meaningful prediction as the degrees of freedom ν approach zero.

Conclusion

The absolute SRT prediction with the SII can be improved with an appropriate reference condition. Correcting for nonlinear effects, which are induced by the shape of the reference discrimination function, did not provide a benefit in prediction accuracy.

Appendix D:

Some Interesting Phenomena

The following two short articles may be interesting to musicians and have no immediate methodological connection to the previous chapters.

Harmony from Amplification

As a musician, I find it quite fascinating to think about the implications of the nonlinear amplification of our auditory system. It turns out that this amplification leads to harmonic distortions, which are equal to the first standard intervals of the diatonic scale for just intonation (see table D.1).

An amplification of a signal $s(t)$ can be described with a monotonic function $f : s \to S$. We develop f in a Taylor series to the cubic term (quadratic term does not exist because of monotonicity):

$$f(s) \approx a_0 + a_1 s + a_3 s^3 \tag{D.1}$$

Without limiting generality, we set $a_0 = a_1 = 0$ and consider a signal of two pure tones with (angular[1]) frequencies ω_L, and ω_H (L stands for lower frequency and H stands for higher frequency):

$$s(t) = \cos(\omega_\mathrm{L} \cdot t) + \cos(\omega_\mathrm{H} \cdot t) \tag{D.2}$$

We amplify this signal $s(t)$ with the cubic approximation of f and find:

$$S(t) = \frac{a_3}{4}\Big(3\cos((2\omega_\mathrm{L} - \omega_\mathrm{H}) \cdot t) + 9\cos(\omega_\mathrm{L} \cdot t) + 9\cos(\omega_\mathrm{H} \cdot t) + 3\cos((2\omega_\mathrm{L} - \omega_\mathrm{H}) \cdot t) + 1\cos(3\omega_\mathrm{L} \cdot t) + 3\cos((2\omega_\mathrm{L} + \omega_\mathrm{H}) \cdot t) + 3\cos((\omega_\mathrm{L} + 2\omega_\mathrm{H}) \cdot t) + 1\cos(3\omega_\mathrm{H} \cdot t)\Big) \tag{D.3}$$

Figure D.1 shows the spectrum. In the immediate frequency range around ω_L and ω_H there are two additional pure tones with frequency $\omega'_\mathrm{L} = (2\omega_\mathrm{L} - \omega_\mathrm{H})$ below and $\omega'_\mathrm{H} = (2\omega_\mathrm{H} - \omega_\mathrm{L})$ above the frequency of the undistorted signal. We hear these two additional pure tones as the so-called distortion product[2]. This distortion is hard-coded in our auditory system.

The distortion product ω'_H is directly connected to the intervals of the diatonic scale. This becomes obvious if we rewrite the distortion $d : \omega \to \omega'$ as:

Table D.1: Naming convention of frequency ratios for the diatonic scale for just intonation

ratio	name
2:1	perfect octave
3:2	perfect fifth
4:3	perfect fourth
5:4	major third
6:5	minor third

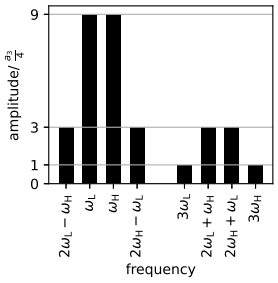


Figure D.1: Spectrum of $(\cos(\omega_\mathrm{L} t) + \cos(\omega_\mathrm{H} t))^3$

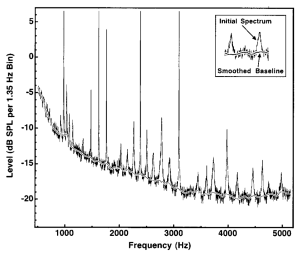


Figure D.2: Spectrum of spontaneous otoacoustic emission: Measured in the outer ear canal. Peaks correspond to resonances because of feedback in the mechanical amplification mechanism of the outer hair cells. Adopted from Pasanen and McFadden 2000.

[1] $\omega = 2\pi f$

[2] We can even record these distortion products from the ear, because of the outer hair cells, which act as an active, nonlinear amplification stage. This is a common screening method for infants to check the function of the ear. In addition to that, there is an interesting effect of otoacoustics that can sometimes be observed with newborn children: the gain of the amplification is not homogeneous along the cochlea, which means that some frequencies are amplified slightly stronger than others. This can lead to a feedback loop and can result in an audible ringing sound from the infant's ears (see figure D.2).

$$\omega_{n+2} = 2\omega_{n+1} - \omega_n \tag{D.4}$$

With the starting interval $\omega_2 = 2\omega_1$ (a pure octave), we yield the next interval of the diatonic scale as the distortion product of the previous interval:

$$\frac{\omega_{n+1}}{\omega_n} = \frac{n+1}{n} \tag{D.5}$$

A perfect fifth is the distortion product of an octave, a perfect fourth is the distortion product of the perfect fifth, and so on. The fundamental building blocks of harmony theory follow from a Taylor expansion of the amplification in our inner ear.

Bad Intonation

If you play a deep bass instrument like a tuba, you may have noticed that it is impossible to play very short, very deep tones with excellent intonation. This is to some degree linked to our perception of pitch, and the same math that is responsible for Heisenberg's uncertainty relation.

The core of the uncertainty principle is that particles are waves of a particle field. These waves follow a similar[3] mathematical description as sound waves of a sound field.

To demonstrate this, we consider a wave with the shape of a normal/Gaussian distribution, which is an eigenfunction of the Fourier transformation[4].

$$\mathscr{F} : \exp\left[-\frac{t^2}{2\sigma_t^2}\right] \to \sqrt{2\pi}\sigma_t \exp\left[-2\pi^2 f^2 \sigma_t^2\right] = \frac{1}{\sqrt{2\pi}\cdot\sigma_f} \exp\left[-\frac{f^2}{2\sigma_f^2}\right] \tag{D.6}$$

A Gaussian with standard deviation (STD) σ_t in the time domain is mapped to a Gaussian with STD $\sigma_f = 1/(2\pi\sigma_t)$ in the frequency domain. So, the shorter the impulse in the time domain, the more smeared out the spectrum. The same analogy holds for a particle, for which the Fourier transform connects position to the de Broglie wavelength, which is proportional to the momentum of a particle and is known as Heisenberg's uncertainty relation.

This means, for a tuba, that if you play a subcontra b-flat[5] (just above the lowest audible pitch with $f \approx 29.135$Hz, see figure 1.4) repeatedly with 240 beats per minute, you will produce sounds with a frequency half width of $\Delta f \approx \pm 1.5$Hz. Accordingly, the largest ratio of the pure and repeating "smeared" tone is:

$$\frac{f}{f - \Delta f} \approx 1.06 \approx \frac{16}{15} \tag{D.7}$$

The frequency spread is as large as a small second (ratio of 16:15) above and below, and audible as bad intonation. This fundamental limit on intonation is not noticeable at higher frequencies because here, Δf is perceived as much smaller.

[3] One major dissimilarity is that sound fields are longitudinal waves of a carrying medium, or "aether", while the fields such as those for photons (i.e., light) are transversal electro-magnetic waves with no "aether".

[4] When appropriately scaled, see Vaidyanathan 2008 for an interesting discussion on the eigenfunctions of $\mathscr{F}$. Here we use the standard (non-unitary) Fourier transform.

[5] German translation: Eng:"B-flat" means Ger:"B" and Eng:"B" means Ger:"H".

Appendix E:

Layman's Introduction

What is it that I am working on?

Every scientist should have a nice answer to this general question. After all, science is not only research but also the communication of knowledge. But there is no communication without understanding. We can measure how understandable spoken words are (understanding as in hearing) through speech tests. When we discuss how well a spoken word or sentence can be understood, we call it intelligibility. In the previous chapters, I referred to it as speech intelligibility (SI).

Tonal Languages:

Tonal languages are special in the way sounds are put together to form a meaningful word. I concentrated on the difference between tonal and non-tonal languages. Tonal languages like Mandarin Chinese have a melodic component called a tone. The tone conveys information about what we say. This is like how we change the melody of a sentence when we turn a statement into a question.[1]

You might wonder, "Is English then also a tonal language?" No! Even though we use melodic cues to add information about questions and especially emotions in English, tones are special. Tones define the melody of a single syllable and can change the meaning of the syllable entirely.

Lombard Effect:

Picture yourself as a young student with normal hearing. You' re in a quiet restaurant with only a few guests. You have dinner with a friend and talk about your life as a scientist. As this turns into a lengthy political discussion about the WissZeitVG[2], the restaurant fills with new guests. After some time, the background noise from other guests increases. You may find yourself speaking louder, but you still understand each other clearly.

We call this speaking style the Lombard effect. It happens when you raise your voice in a loud place. You do this to make sure people can still understand you despite the background noise. We can find Lombard speech in all languages[3], but it' s not clear if it interacts with tonality. We did not know if Lombard speech in tonal languages makes any useful changes to the tones that improve speech intelligibility. My research on two tonal languages has shown that tones do not influence the improved speech intelligibility of Lombard speech (see chapter 5 and 6).

[1] We call the melody of an entire sentence prosody.

[2] Short for "Wissenschaftszeitvertragsgesetz". A German law to partly abolish employee protection for young scientists.

[3] It even occurs with animals, dating back to an old evolutionary invention (Luo, Hage, and Moss 2018).

SI models

In science, once we can measure something with repeatable experiments, we want to know why we see what we see. To achieve this, we develop models with which we can predict our experiments. These are shortcuts to skip the long experiment. We put in a description of our experiment and immediately[4] receive a prediction of the measurement.

In this book, I call these models SI models. They predict how well we can understand speech in background noise, and come in a large variety. I focused on a successful model called framework of auditory discrimination experiments (FADE) and compared it to other popular models. One of these models is the speech intelligibility index (SII). Working with the SII, I found that a common adaptation of the SII to tonal languages does not generalize between speech tests in a way it was thought to generalize (see Chapter 4).

Consistency of Data:

The story might end here if we[5] hadn't had a problem with our empirical data. While developing the speech tests, my colleagues found that a standard method to improve test homogeneity[6] didn' t work. We faced another issue when some recorded speech intelligibility data didn't pass our quality checks. The measurement procedure used did not always produce a meaningful result.

Quality checks of hearing-test data, or psychometric data, are crucial for any quantitative[7] analysis. But there is no general understanding of inconsistent data in the field of audiology. So I set out to define a model of inconsistency and applied it to our data. I found that it is possible to predict expert ratings of consistency with a formula that we showed to be suboptimal. We propose a better way to rate the consistency of audiological tests in chapter 3. We also found a link to the failing optimization strategy of speech tests. One need for the success of our optimization procedure is that we can describe the test with a "monotonous psychometric function". The model that I describe can check if a test fulfills this need.

The Bottle-whistle:

You can now open your "Feierabendbier"[8] because you will need it. More precisely, you will need the empty bottle for this last scientific excursion. What we will attempt is really the first thing that you have to do at the beginning of an experiment. We call this calibration. Calibration is key in quantitative sciences. It helps turn a physical effect into a number you can use for math. Everyone has to agree on this translation of effect to number, for that is what a calibration does. I created a new calibration method. It has enough accuracy for meaningful audiological measurements using smartphones.

The Cluster of Excellence "Hearing4All" in Oldenburg needed a calibration group. Their goal was to find the simplest and most precise calibration procedure for the virtual hearing clinic (VHC). The VHC is a smartphone website. With it, we can compare audiological tests on the smartphone to lab hearing tests. The goal was to establish a scientific comparison of remote hearing tests with clinical tests. These verified remote tests can diagnose hearing impairment in areas with poor infrastructure.

We tested various unusual calibration methods for microphones in mobile devices. These included things like yelling, toilet flushes, and whistling on beer bottles. A key step for the

[4] In practice, the models also need some time to compute the result, but a useful model is always faster than the experiment.

[5] The collaborators of the cooperation between Oldenburg, Shenzhen, and Hong Kong: Prof. Lena Wong, Prof. Felix Chen, Dr. Hongmei Hu, Dr. Sabine Hochmuth, Dr. Ania Warzybok, and Prof. Birger Kollmeier

[6] Homogeneity of speech tests means that every word is equally good to understand as the others. Think of it like sorting dark and white socks for laundry. To make this experiment homogeneous, you' d bleach all the white socks and dye the dark ones black. Now every sock is equally challenging to sort into the white or dark bin, but for what a terrible price!

[7] Quantitative science from the Latin word *quantitas*. It means the science of quantities, which we describe with numbers.

[8] A beer that you might drink after work with fellow PhD students at occasions like the "Promovierendenstammtisch" of the PhD-council.

calibration process was to identify a reference sound source with a fixed sound pressure level (SPL). The SPL is the physical property of sounds that we perceive as loudness.

Chapter 2 introduces bottle-whistles as a candidate for a microphone calibration method. Bottle-whistles are so-called driven oscillators. This driving force is the air that you blow over the bottle. The speed of the air is connected to the excitation frequency, which leads to the fact that you can not whistle as loudly as you want on a bottle. If you whistle on a bottle, chances are high that you will reach the greatest level (SPL) with very little practice. The differences among whistlers using the same bottle are small. This makes bottle-whistles a good reference for SPLs in audiological experiments.

Acronyms

χ^2	measure of goodness of a model
γ	guess rate
λ	lapse rate
$\mathbf{r}$	sequence of responses
$\mathbf{s}$	sequence of stimuli
AI	articulation index
ASR	automatic speech recognizer
AUC	area under the ROC-curve
BIF	band importance function
dB	decibel
dB HL	decibel relative to the hearing threshold of a normal hearing person
EM	expectation maximization algorithm
eSII	extended SII
FADE	framework of auditory discrimination experiments
FAIR	findable, accessible, interoperable, and reusable
FS	full scale
GBFB	Gabor filter bank
GMM	Gaussian mixture model
HMM	hidden Markov model
HMM-GMM	hidden Markov model in conjunction with a Gaussian mixture model
ICRA1	unmodulated speech noise with a long-term spectrum of a male speaker at normal vocal effort (Wouter A. Dreschler et al. 2001)
ICRA5-250	modulated ICRA1 noise with a duration of pauses no longer than 250ms (Kirsten Carola Wagener, Thomas Brand, and Kollmeier 2006)
JND	just noticeable difference
KIC	known inconsistency
$\mathrm{l}_{\max}$	maximum log-likelihood
LG	Lombard gain

LMS	log-mel-spectrum
LTI	linear time invariant system
MFCC	mel frequency cepstral coefficients
mHealth	mobile health
NNS	band importance function various nonsense syllable tests where most of the English phonemes occur equally often
OLSA	Oldenburger matrix sentence test
PF	psychometric function
PHOBI	phone-based binaural intelligibility model
R_p	Pearson correlation coefficient
R_s	Spearman correlation coefficient
RETSPL	reference equivalent threshold sound pressure level
RMS	root mean square error
ROC	receiver operating characteristic
S_r	Strouhal number
SDT	speech detection threshold
SGBFB	separable gabor filter bank
SI	speech intelligibility
SII	speech intelligibility index
SiPGOF	single psychometric function goodness of fit measure
SNR	signal to noise ratio
SPL	sound pressure level
SRT	speech recognition threshold
SRT_{50}	speech recognition threshold for 50 % correctly understood words
STD	standard deviation
STFT	short time Fourier transform
STI	speech transmission index
STOI	short time objective intelligibility measure
VHC	virtual hearing clinic
YUE	dialect of Cantonese that is spoken in Hong Kong

Bibliography

Alghamdi, Najwa et al. (June 2018). "A corpus of audio-visual Lombard speech with frontal and profile views". In: *Journal of the Acoustical Society of America* 143.6, pp. 523–529. DOI: 10.1121/1.5042758 (cit. on pp. 36, 72, 82).

Allahbakhsh, M. et al. (Mar. 2013). "Quality Control in Crowdsourcing Systems: Issues and Directions". In: *IEEE Internet Computing* 17.2, pp. 76–81. DOI: 10.1109/mic.2013.20 (cit. on p. 59).

Alster, M. (Sept. 1972). "Improved calculation of resonant frequencies of Helmholtz resonators". In: *Journal of Sound and Vibration* 24.1, pp. 63–85. ISSN: 0022-460X. DOI: 10.1016/0022-460x(72)90123-x (cit. on p. 43).

Amlani, Amyn M., Jerry L. Punch, and Teresa Y. C. Ching (Sept. 2002). "Methods and Applications of the Audibility Index in Hearing Aid Selection and Fitting". In: *Trends in Amplification* 6.3, pp. 81–129. ISSN: 1084-7138. DOI: 10.1177/108471380200600302 (cit. on p. 64).

Andersen, Asger Heidemann et al. (Mar. 2017). "A non-intrusive Short-Time Objective Intelligibility measure". In: DOI: 10.1109/icassp.2017.7953125 (cit. on p. 24).

ANSIS1.11 (June 2004). *Specification for Octave, Half-Octave, and Third Octave Band Filter Sets*. Acoustical Society of America. URL: https://archive.org/details/gov.law.ansi.s1.11.2004/ (cit. on p. 74).

ANSIS3.5 (June 1997). *Methods for calculating the speech intelligibility index SII*. Acoustical Society of America (cit. on pp. 24, 31, 64–69, 73, 74, 82).

Auvray, Roman et al. (June 2016). "Specific features of a stopped pipe blown by a turbulent jet: Aeroacoustics of the panpipes". In: *The Journal of the Acoustical Society of America* 139.6, pp. 3214–3225. DOI: 10.1121/1.4953066 (cit. on p. 43).

Barker, Jon and Martin Cooke (May 2007). "Modelling speaker intelligibility in noise". In: *Speech Communication* 49.5, pp. 402–417. ISSN: 0167-6393. DOI: 10.1016/j.specom.2006.11.003 (cit. on p. 24).

Barlow, Roger (1989). *Statistics. a guide to the use of statistical methods in the physical sciences*. Ed. by F. Mandel, Ellison R. J., and D. J. Sandiford. Wiley, p. 204. 204 pp. ISBN: 0471922943 (cit. on p. 34).

Baxter, William H (1992). "New rhyme categories for Old Chinese". eng. In: *A Handbook of Old Chinese Phonology*. Vol. 64. Germany: De Gruyter, Inc. ISBN: 311012324X (cit. on p. 34).

Bechhoefer, John (2011). "Kramers–Kronig, Bode, and the meaning of zero". In: *American Journal of Physics* 79.10, pp. 1053–1059. DOI: 10.1119/1.3614039 (cit. on p. 27).

Beutelmann, Rainer, Thomas Brand, and Birger Kollmeier (Apr. 2010). "Revision, extension, and evaluation of a binaural speech intelligibility model". In: *Journal of the Acoustical Society of America* 127.4, pp. 2479–2497. DOI: 10.1121/1.3295575 (cit. on pp. 65, 73).

Bianco, Roberta et al. (Jan. 2021). "Reward Enhances Online Participants' Engagement With a Demanding Auditory Task". In: *Trends in Hearing* 25, p. 233121652110259. ISSN: 2331-2165. DOI: 10.1177/23312165211025941 (cit. on p. 50).

Bishop, Christopher M. (2019). *Pattern recognition and machine learning*. eng. Corrected at 8th printing 2009. Information science and statistics. New York, NY: Springer Science+Business Media, LLC. 758 pp. ISBN: 0387310738 (cit. on pp. 28, 52).

Bogert, Bruce P, MJR Healy, and John W Tukey (1963). "The Quefrency Alanysis of Time Series for Echoes". In: *Proc. Symp. on Time Set. Anal.* Ed. by M. Rosenblatt. Wiley NY (cit. on p. 27).

Bosen, Adam K. et al. (Nov. 2024). "Frequency importance for sentence recognition in co-located noise, co-located speech, and spatially separated speech". In: *The Journal of the Acoustical Society of America* 156.5, pp. 3275–3284. ISSN: 0001-4966. DOI: 10.1121/10.0034412 (cit. on p. 64).

Boujo, E. et al. (Jan. 2020). "Processing time-series of randomly forced self-oscillators: The example of beer bottle whistling". In: *Journal of Sound and Vibration* 464, p. 114981. DOI: 10.1016/j.jsv.2019.114981 (cit. on p. 43).

Bracha, A and SY Tan (Jan. 2015). "Robert Bárány (1876–1936): The Nobel Prize-winning prisoner of war". In: *Singapore Medical Journal* 56.01, pp. 5–6. DOI: 10.11622/smedj.2015002 (cit. on p. 34).

Brand, Thomas and Birger Kollmeier (June 2002). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests". In: *Journal of the Acoustical Society of America* 111.6, pp. 2801–2810. DOI: 10.1121/1.1479152 (cit. on pp. 21, 23, 51, 52, 54, 55, 60, 66, 73, 83, 88).

Brand, Thomas, Saskia Roettges, and Christopher Hauth (2021). "A workaround for problems in predicting very low speech recognition thresholds". In: *procedings of teh Congress of European Federation of Audiology Societies EFAS*. URL: https://vimeo.com/551904093/41776b90c6 (cit. on p. 97).

Bruce, Robert V. (1990). *Bell. Alexander Graham Bell and the conquest of solitude*. [Nachdr. der Ausg.] Boston, Little, Brown, 1973. Ithaca [u.a.]: Cornell Univ. Press. 564 pp. ISBN: 0801424194 (cit. on p. 18).

Buhl, Mareike (Feb. 2022). "Interpretable Clinical Decision Support System for Audiology Based on Predicted Common Audiological Functional Parameters (CAFPAs)". In: *Diagnostics* 12.2, p. 463. DOI: 10.3390/diagnostics12020463 (cit. on p. 24).

Carlyon, Robert P. and Brian C. J. Moore (Feb. 1986). "Continuous versus gated pedestals and the ''severe departure'' from Weber's law". In: *The Journal of the Acoustical Society of America* 79.2, pp. 453–460. ISSN: 1520-8524. DOI: 10.1121/1.393759 (cit. on p. 25).

Castner, T. G. and C. W. Carter (1933). "Developments in the application of articulation testing". eng. In: *Bell System Technical Journal* 12.3, pp. 347–370. ISSN: 0005-8580 (cit. on p. 17).

Chen, Fei and Philipos C. Loizou (Dec. 2010). "Analysis of a simplified normalized covariance measure based on binary weighting functions for predicting the intelligibility of noise-suppressed speech". In: *The Journal of the Acoustical Society of America* 128.6, pp. 3715–3723. ISSN: 1520-8524. DOI: 10.1121/1.3502473 (cit. on p. 66).

Chen, Fei, Changjie Pan, et al. (Mar. 2025). "Understanding the Lombard Effect for Mandarin: Relation Between Speech Recognition Thresholds and Acoustic Parameters". In: *Trends in Hearing* 29. ISSN: 2331-2165. DOI: 10.1177/23312165251324266 (cit. on pp. 23, 35, 65, 68, 69, 72, 73, 79, 80).

Chen, Jing, Qiang Huang, and Xihong Wu (Oct. 2016). "Frequency importance function of the speech intelligibility index for Mandarin Chinese". In: *Speech Communication* 83, pp. 94–103. ISSN: 0167-6393. DOI: 10.1016/j.specom.2016.07.009 (cit. on pp. 31, 65–69, 75–77, 82).

Chiang, Hsin-Tien et al. (Nov. 2024). "Multi-objective non-intrusive hearing-aid speech assessment model". In: *The Journal of the Acoustical Society of America* 156.5, pp. 3574–3587. ISSN: 1520-8524. DOI: 10.1121/10.0034362 (cit. on pp. 31, 87).

Dau, Torsten, Birger Kollmeier, and Armin Kohlrausch (Apr. 1996). "A quantitative prediction of modulation masking with an optimal-detector model." In: *The Journal of the Acoustical Society of America* 99.4, pp. 2565–2574. ISSN: 0001-4966. DOI: 10.1121/1.415040 (cit. on p. 72).

Dau, Torsten, Dirk Püschel, and Armin Kohlrausch (June 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. ISSN: 1520-8524. DOI: 10.1121/1.414959 (cit. on p. 74).

Denk, Florian et al. (Mar. 2021). "The Missing 6 dB Revisited: Influence of Room Acoustics and Binaural Parameters on the Loudness Mismatch Between Headphones and Loudspeakers". In: *Frontiers in Psychology* 12. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.623670 (cit. on p. 46).

DePaolis, Rory A., Claus P. Janota, and Tom Frank (Aug. 1996). "Frequency Importance Functions for Words, Sentences, and Continuous Discourse". In: *Journal of Speech, Language, and Hearing Research* 39.4, pp. 714–723. ISSN: 1558-9102. DOI: 10.1044/jshr.3904.714 (cit. on p. 64).

Dequand, S. et al. (Mar. 2003). "Simplified models of flue instruments: Influence of mouth geometry on the sound source". In: *Journal of the Acoustical Society of America* 113.3, pp. 1724–1735. DOI: 10.1121/1.1543929 (cit. on p. 43).

Dieck, Teena tom et al. (2022). "Wav2vec behind the Scenes: How end2end Models learn Phonetics". In: (cit. on p. 24).

DIN EN 60268-7:2021-08 Elektroakustische Geräte - Teil 7: Kopfhörer und Ohrhörer (IEC 60268-7:2010 + COR1:2012 + A1:2020) (2021). DOI: 10.31030/3271664 (cit. on p. 46).

DIN EN 61305-1:1996-02, Hi-Fi-Geräte und -Anlagen für den Heimgebrauch (1996). Deutsches Institut für Normung e. V. DOI: 10.31030/2832224 (cit. on p. 17).

DIN EN 61672-1:2014-07 Elektroakustik - Schallpegelmesser - Teil 1: Anforderungen (2014). German. Tech. rep. Deutsches Institut für Normung e. V. DOI: 10.31030/2154580 (cit. on pp. 43, 44).

DIN EN ISO 8253-1:2011-04, Akustik- Audiometrische Prüfverfahren- Teil1: Grundlegende Verfahren der Luft- und Knochenleitungs-Schwellenaudiometrie mit reinen Tönen (2011). Normenausschuss Akustik, Lärmminderung und Schwingungstechnik. DOI: 10.31030/1705319 (cit. on pp. 43, 46).

Doire, Clement S. J., Mike Brookes, and Patrick A. Naylor (Apr. 2017). "Robust and efficient Bayesian adaptive psychometric function estimation". In: *Journal of the Acoustical Society of America* 141.4, pp. 2501–2512. DOI: 10.1121/1.4979580 (cit. on pp. 21, 52, 60).

Dreher, John J. and John O'Neill (Dec. 1957). "Effects of Ambient Noise on Speaker Intelligibility for Words and Phrases". In: *The Journal of the Acoustical Society of America* 29.12, pp. 1320–1323. ISSN: 1520-8524. DOI: 10.1121/1.1908780 (cit. on pp. 17, 18, 36, 72).

Dreschler, Wouter A. et al. (Jan. 2001). "ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment: Ruidos ICRA: Señates de ruido artificial con espectro similar al habla y propiedades temporales para pruebas de instrumentos auditivos". In: *International Journal of Audiology* 40.3, pp. 148–157. ISSN: 1708-8186. DOI: 10.3109/00206090109073110 (cit. on pp. 32, 55, 70, 73, 83, 105).

Du, Yufan et al. (July 2019). "The effect of speech material on the band importance function for Mandarin Chinese". In: *Journal of the Acoustical Society of America* 146.1, pp. 445–457. ISSN: 1520-8524. DOI: 10.1121/1.5116691 (cit. on pp. 31, 65, 66, 68).

Edwards, Brent W, Christoph Menzel, and Sunil Puria (Jan. 2005). *System and method for remotely administered, interactive hearing tests.* US Patent 6,840,908 (cit. on p. 42).

ETSI (2003). "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; frontend feature extraction algorithm; compression algorithms". In: *ETSI ES* 201.108, p. V1 (cit. on pp. 26, 74, 82).

Fastl, Hugo and Eberhard Zwicker (2007). *Psychoacoustics.* Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-68888-4 (cit. on p. 26).

Fava, Gaetano et al. (Feb. 2016). "The Use of Sound Level Meter Apps in the Clinical Setting". In: *American Journal of Speech-Language Pathology* 25.1, pp. 14–28. DOI: 10.1044/2015_ajslp-13-0137 (cit. on p. 42).

Feng, Yong and Fei Chen (Jan. 2022). "Nonintrusive objective measurement of speech intelligibility: A review of methodology". In: *Biomedical Signal Processing and Control* 71, p. 103204. ISSN: 1746-8094. DOI: 10.1016/j.bspc.2021.103204 (cit. on p. 31).

Fletcher, Harvey and Rogers H. Galt (Mar. 1950). "The Perception of Speech and Its Relation to Telephony". In: *The Journal of the Acoustical Society of America* 22.2, pp. 89–151. DOI: 10.1121/1.1906605 (cit. on pp. 17, 18, 66).

Fontan, Lionel et al. (Sept. 2017). "Automatic Speech Recognition Predicts Speech Intelligibility and Comprehension for Listeners With Simulated Age-Related Hearing Loss". In: *Journal of Speech, Language, and Hearing Research* 60.9, pp. 2394–2405. ISSN: 1558-9102. DOI: 10.1044/2017_jslhr-s-16-0269 (cit. on p. 24).

Frund, I., N. V. Haenel, and F. A. Wichmann (May 2011). "Inference for psychometric functions in the presence of nonstationary behavior". In: *Journal of Vision* 11.6, pp. 16–16. ISSN: 1534-7362. DOI: 10.1167/11.6.16 (cit. on p. 50).

Fu, Qian-Jie et al. (July 1998). "Importance of tonal envelope cues in Chinese speech recognition". In: *The Journal of the Acoustical Society of America* 104.1, pp. 505–510. ISSN: 1520-8524. DOI: 10.1121/1.423251 (cit. on p. 88).

Gardner, Nele and Stefan Zacharias (Mar. 2016). *3-Punkt-Kalibrierung von Apps zur Schallpegelmesung*. Online tutorial. Video. URL: https://www.youtube.com/watch?v=_5Ste5q4Vv0 (cit. on p. 42).

Gerber, E. A. and R. A. Sykes (1966). "State of the art—Quartz crystal units and oscillators". In: *Proceedings of the IEEE* 54.2, pp. 103–116. ISSN: 0018-9219. DOI: 10.1109/proc.1966.4625 (cit. on p. 39).

Grimm, Giso and Tobias Herzke (2025). *HoerTech-gGmbH/tascar: TASCAR: release 0.234.2*. DOI: 10.5281/ZENODO.14878907 (cit. on p. 93).

Hagerman, B. (Jan. 1982). "Sentences for Testing Speech Intelligibility in Noise". In: *Scandinavian Audiology* 11.2, pp. 79–87. ISSN: 0105-0397. DOI: 10.3109/01050398209076203 (cit. on pp. 21, 23).

Hanley, T. D. and M. D. Steer (Dec. 1949). "Effect of Level of Distracting Noise upon Speaking Rate, Duration and Intensity". In: *Journal of Speech and Hearing Disorders* 14.4, pp. 363–368. ISSN: 2163-6184. DOI: 10.1044/jshd.1404.363 (cit. on pp. 35, 36, 72).

Hansen, John H. L. (Nov. 1996). "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition". In: *Speech Communication* 20.1-2, pp. 151–173. ISSN: 0167-6393. DOI: 10.1016/s0167-6393(96)00050-7 (cit. on p. 72).

Hauth, Christopher F. et al. (Jan. 2020). "Modeling Binaural Unmasking of Speech Using a Blind Binaural Processing Stage". In: *Trends in Hearing* 24, p. 233121652097563. DOI: 10.1177/2331216520975630 (cit. on p. 64).

Healy, Eric W., Sarah E. Yoho, and Frédéric Apoux (Jan. 2013). "Band importance for sentences and words reexamined". In: *The Journal of the Acoustical Society of America* 133.1, pp. 463–473. ISSN: 1520-8524. DOI: 10.1121/1.4770246 (cit. on p. 64).

Heil, Peter and Adam J. Peterson (Apr. 2015). "Basic response properties of auditory nerve fibers: a review". In: *Cell and Tissue Research* 361.1, pp. 129–158. DOI: 10.1007/s00441-015-2177-9 (cit. on p. 25).

Hochmuth, Sabine et al. (May 2015). "Talker- and language-specific effects on speech intelligibility in noise assessed with bilingual talkers: Which language is more robust against noise and reverberation?" In: *International Journal of Audiology* 54.sup2, pp. 23–34. DOI: 10.3109/14992027.2015.1088174 (cit. on p. 65).

Holube, Inga and Birger Kollmeier (Sept. 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. ISSN: 1520-8524. DOI: 10.1121/1.417354 (cit. on p. 24).

Houtgast, Tammo et al. (2002). *Past, present and future of the Speech Transmission Index*. Soesterberg: TNO (cit. on pp. 24, 31).

Hu, Hongmei, Sabine Hochmuth, et al. (Nov. 2022). "Development and evaluation of the Cantonese matrix sentence test". In: *International Journal of Audiology*, pp. 1–13. DOI: 10.1080/14992027.2022.2142683 (cit. on pp. 54, 55, 88).

Hu, Hongmei, Anna Warzybok, et al. (2022). *Mandarin matrix recordings: Lombard and plain speech with different speakers*. DOI: 10.5281/ZENODO.7063030 (cit. on pp. 40, 65, 70, 80, 88).

Hu, Hongmei, Xin Xi, et al. (Sept. 2018). "Construction and evaluation of the Mandarin Chinese matrix (CMNmatrix) sentence test for the assessment of speech recognition in noise". In: *International Journal of Audiology* 57.11, pp. 838–850. ISSN: 1708-8186. DOI: 10.1080/14992027.2018.1483083 (cit. on pp. 22, 32, 66–68, 73, 75–77).

Hughes, Ifan (2010). *Measurements and their uncertainties. A practical guide to modern error analysis*. Ed. by Ifan Hughes and Thomas P. A. Hase. Includes bibliographical references (p. [131]-132) and index. Oxford: Oxford University Press. 136 pp. ISBN: 978-0-19-956632-7 (cit. on p. 34).

Hülsmeier, David (2022). *d-hr/hz-2020-data-drop: Initial Data Drop*. Github repository. DOI: 10.5281/ZENODO.6481185 (cit. on pp. 55, 57, 60, 61).

Hülsmeier, David, Mareike Buhl, et al. (June 2021). "Inference of the distortion component of hearing impairment from speech recognition by predicting the effect of the attenuation component". In: *International Journal of Audiology*, pp. 1–15. DOI: 10.1080/14992027.2021.1929515 (cit. on pp. 25, 85).

Hülsmeier, David, Anna Warzybok, et al. (Sept. 2020). "Simulations with FADE of the effect of impaired hearing on speech recognition performance cast doubt on the role of spectral resolution". In: *Hearing Research* 395, p. 107995. ISSN: 0378-5955. DOI: 10.1016/j.heares.2020.107995 (cit. on p. 72).

Ibelings, Saskia et al. (Jan. 2024). "Development of a Phrase-Based Speech-Recognition Test Using Synthetic Speech". In: *Trends in Hearing* 28. ISSN: 2331-2165. DOI: 10.1177/23312165241261490 (cit. on p. 22).

Junqua, Jean-Claude (Jan. 1993). "The Lombard reflex and its role on human listeners and automatic speech recognizers". In: *Journal of the Acoustical Society of America* 93.1, pp. 510–524. DOI: 10.1121/1.405631 (cit. on pp. 72, 82).

Kardous, Chucri A. and Peter B. Shaw (Mar. 2014). "Evaluation of smartphone sound measurement applications". In: *The Journal of the Acoustical Society of America* 135.4, pp. 186–192. DOI: 10.1121/1.4865269 (cit. on p. 42).

Kates, James M. and Kathryn H. Arehart (Dec. 2022). "An overview of the HASPI and HASQI metrics for predicting speech intelligibility and speech quality for normal hearing, hearing loss, and hearing aids". In: *Hearing Research* 426, p. 108608. ISSN: 0378-5955. DOI: 10.1016/j.heares.2022.108608 (cit. on p. 31).

Kisić, Dominik et al. (July 2022). "The Potential of Speech as the Calibration Sound for Level Calibration of Non-Laboratory Listening Test Setups". In: *Applied Sciences* 12.14, p. 7202. DOI: 10.3390/app12147202 (cit. on pp. 42, 46).

Kollmeier, Birger (1990). "Meßmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache". PhD thesis. Göttingen University (cit. on pp. 22, 31, 60).

Kollmeier, Birger, Anna Warzybok, et al. (May 2015). "The multilingual matrix test: Principles, applications, and comparison across languages: A review". In: *International Journal of Audiology* 54.sup2, pp. 3–16. DOI: 10.3109/14992027.2015.1020971 (cit. on pp. 21, 23, 50, 51, 54, 55, 61, 66, 83).

Kollmeier, Birger and Matthias Wesselkamp (Oct. 1997). "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment". In: *The Journal of the Acoustical Society of America* 102.4, pp. 2412–2421. ISSN: 1520-8524. DOI: 10.1121/1.419624 (cit. on p. 21).

Kryter, Karl D. (Nov. 1962). "Methods for the Calculation and Use of the Articulation Index". In: *Journal of the Acoustical Society of America* 34.11, pp. 1689–1697. DOI: 10.1121/1.1909094 (cit. on p. 64).

Lane, Harlan and Bernard Tranel (Dec. 1971). "The Lombard Sign and the Role of Hearing in Speech". In: *Journal of Speech and Hearing Research* 14.4, pp. 677–709. DOI: 10.1044/jshr.1404.677 (cit. on p. 35).

Leek, Marjorie R. (Nov. 2001). "Adaptive procedures in psychophysical research". In: *Perception & Psychophysics* 63.8, pp. 1279–1292. DOI: 10.3758/bf03194543 (cit. on pp. 24, 50).

Leek, Marjorie R., Thomas E. Hanna, and Lynne Marshall (Sept. 1991). "An interleaved tracking procedure to monitor unstable psychometric functions". In: *The Journal of the Acoustical Society of America* 90.3, pp. 1385–1397. DOI: 10.1121/1.401930 (cit. on p. 60).

Lermoyez, Marcel (Jan. 1921). "Étienne Lombard. (Paris, 1869-1920)". In: *The Journal of Laryngology & Otology* 36.1, pp. 47–48. ISSN: 1748-5460. DOI: 10.1017/s0022215100021575 (cit. on pp. 18, 35).

Levine, Harold and Julian Schwinger (Feb. 1948). "On the Radiation of Sound from an Unflanged Circular Pipe". In: *Physical Review* 73.4, pp. 383–406. ISSN: 0031-899X. DOI: 10.1103/physrev.73.383 (cit. on p. 43).

Lombard, Etienne (1911). "Le signe de l'elevation de la voix". In: *Ann. Mal. de L'Oreille et du Larynx*, pp. 101–119 (cit. on pp. 17, 35, 65, 72, 82).

Lopez-Poveda, Enrique A. et al. (2017). "Predictors of Hearing-Aid Outcomes". In: *Trends in Hearing* 21, p. 233121651773052. ISSN: 2331-2165. DOI: 10.1177/2331216517730526 (cit. on pp. 64, 73).

Love, Elliot K. and Mark A. Bee (Sept. 2010). "An experimental test of noise-dependent voice amplitude regulation in Cope's grey treefrog, Hyla chrysoscelis". In: *Animal Behaviour* 80.3, pp. 509–515. ISSN: 0003-3472. DOI: 10.1016/j.anbehav.2010.05.031 (cit. on p. 35).

Lu, Youyi and Martin Cooke (Nov. 2008). "Speech production modifications produced by competing talkers, babble, and stationary noise". In: *The Journal of the Acoustical Society of America* 124.5, pp. 3261–3275. ISSN: 1520-8524. DOI: 10.1121/1.2990705 (cit. on pp. 36, 72).

— (Dec. 2009). "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise". In: *Speech Communication* 51.12, pp. 1253–1262. ISSN: 0167-6393. DOI: 10.1016/j.specom.2009.07.002 (cit. on pp. 35, 72).

Ludvigsen, Carl et al. (Jan. 1990). "Prediction of Intelligibility of Non-linearly Processed Speech". In: *Acta Oto-Laryngologica* 109.sup469, pp. 190–195. ISSN: 1651-2251. DOI: 10.1080/00016489.1990.12088428 (cit. on p. 31).

Luo, Jinhong, Steffen R. Hage, and Cynthia F. Moss (Dec. 2018). "The Lombard Effect: From Acoustics to Neural Mechanisms". In: *Trends in Neurosciences* 41.12, pp. 938–949. DOI: 10.1016/j.tins.2018.07.011 (cit. on pp. 35, 101).

MacKay, David J. C. (2006). *Information theory, inference and learning algorithms*. Cambridge university press (cit. on p. 60).

Marxer, Ricard et al. (June 2018). "The impact of the Lombard effect on audio and visual speech recognition systems". In: *Speech Communication* 100, pp. 58–68. ISSN: 0167-6393. DOI: 10.1016/j.specom.2018.04.006 (cit. on p. 36).

McLennon, Travis et al. (July 2019). "Evaluation of smartphone sound level meter applications as a reliable tool for noise monitoring". In: *Journal of Occupational and Environmental Hygiene* 16.9, pp. 620–627. DOI: 10.1080/15459624.2019.1639718 (cit. on pp. 42, 45).

Mishap Investigation board (Nov. 1999). *Mars Climate Orbiter Mishap*. Tech. rep. Report 10. Jet Propulsion Laboratory (cit. on p. 36).

Morat, Daniel and Hansjakob Ziemer (2018). *Handbuch Sound: Geschichte - Begriffe - Ansätze*. J.B. Metzler. ISBN: 9783476054210. DOI: 10.1007/978-3-476-05421-0 (cit. on p. 20).

Müller, Swen and Paulo Massarani (2001). "Transfer-function measurement with sweeps". In: *Journal of the Audio Engineering Society* 49.6, pp. 443–471 (cit. on pp. 37, 93).

NASA (1999). *Vortex-street-1*. Public domain. URL: https://commons.wikimedia.org/wiki/File:Vortex-street-1.jpg (cit. on p. 37).

Ooster, Jasper et al. (Jan. 2020). "Speech Audiometry at Home: Automated Listening Tests via Smart Speakers With Normal-Hearing and Hearing-Impaired Listeners". In: *Trends in Hearing* 24, p. 233121652097001. DOI: 10.1177/2331216520970011 (cit. on p. 50).

Oppenheim, A.V. and R.W. Schafer (Sept. 2004). "Dsp history - From frequency to quefrency: a history of the cepstrum". In: *IEEE Signal Processing Magazine* 21.5, pp. 95–106. ISSN: 1053-5888. DOI: 10.1109/msp.2004.1328092 (cit. on p. 27).

Pasanen, Edward G. and Dennis McFadden (Sept. 2000). "An automated procedure for identifying spontaneous otoacoustic emissions". In: *The Journal of the Acoustical Society of America* 108.3, pp. 1105–1116. ISSN: 1520-8524. DOI: 10.1121/1.1287026 (cit. on p. 99).

Pavlovic, Chaslav V. (Aug. 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. ISSN: 1520-8524. DOI: 10.1121/1.395442 (cit. on pp. 31, 64, 66, 69).

Pearson, Karl (July 1900). "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302, pp. 157–175. ISSN: 1941-5990. DOI: 10.1080/14786440009463897 (cit. on p. 34).

Pednekar, Shourav et al. (Mar. 2023). "Weber's Law of perception is a consequence of resolving the intensity of natural scintillating light and sound with the least possible

error". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 479.2271. ISSN: 1471-2946. DOI: 10.1098/rspa.2022.0626 (cit. on p. 25).

Pick, Herbert L. et al. (Feb. 1989). "Inhibiting the Lombard effect". In: *The Journal of the Acoustical Society of America* 85.2, pp. 894–900. ISSN: 1520-8524. DOI: 10.1121/1.397561 (cit. on p. 35).

Pierce, Allan D. (2019). *Acoustics: An Introduction to Its Physical Principles and Applications*. Springer International Publishing. ISBN: 9783030112141. DOI: 10.1007/978-3-030-11214-1 (cit. on pp. 19, 43).

Pittman, Andrea L. and Terry L. Wiley (June 2001). "Recognition of Speech Produced in Noise". In: *Journal of Speech, Language, and Hearing Research* 44.3, pp. 487–496. ISSN: 1558-9102. DOI: 10.1044/1092-4388(2001/038) (cit. on pp. 35, 36, 72).

Plomp, Reinier (Feb. 1978). "Auditory handicap of hearing impairment and the limited benefit of hearing aids". In: *The Journal of the Acoustical Society of America* 63.2, pp. 533–549. ISSN: 1520-8524. DOI: 10.1121/1.381753 (cit. on p. 25).

Polspoel, Sigrid et al. (Apr. 2025). "Automatic development of speech-in-noise hearing tests using machine learning". In: *Scientific Reports* 15.1. ISSN: 2045-2322. DOI: 10.1038/s41598-025-96312-z (cit. on p. 72).

Potash, L. M. (May 1972). "Noise-induced changes in calls of the Japanese quail". In: *Psychonomic Science* 26.5, pp. 252–254. ISSN: 0033-3131. DOI: 10.3758/bf03328608 (cit. on p. 35).

Povey, Daniel et al. (Dec. 2011). "The Kaldi Speech Recognition Toolkit". In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Catalog No.: CFP11SRW-USB. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society (cit. on p. 40).

Rabiner, L.R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2, pp. 257–286. ISSN: 0018-9219. DOI: 10.1109/5.18626 (cit. on pp. 28, 29, 52).

Revelle, William (May 2007). *psych: Procedures for Psychological, Psychometric, and Personality Research*. DOI: 10.32614/cran.package.psych (cit. on p. 67).

Rhebergen, Koenraad S., Niek J. Versfeld, and Wouter. A. Dreschler (Dec. 2006). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise". In: *Journal of the Acoustical Society of America* 120.6, pp. 3988–3997. DOI: 10.1121/1.2358008 (cit. on pp. 31, 72).

Roßbach, Jana, Kirsten C. Wagener, and Bernd T. Meyer (2023). "Multilingual Non-intrusive Binaural Intelligibility Prediction based on Phone Classification". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.4608134 (cit. on pp. 24, 31, 69, 87).

Saak, Samira et al. (Sept. 2022). "A flexible data-driven audiological patient stratification method for deriving auditory profiles". In: *Frontiers in Neurology* 13. DOI: 10.3389/fneur.2022.959582 (cit. on pp. 24, 59).

Saba, Juliana N. and John H. L. Hansen (Feb. 2022). "The effects of Lombard perturbation on speech intelligibility in noise for normal hearing and cochlear implant listeners". In: *The Journal of the Acoustical Society of America* 151.2, pp. 1007–1021. ISSN: 1520-8524. DOI: 10.1121/10.0009377 (cit. on p. 36).

Saddler, Mark R. and Josh H. McDermott (Dec. 2024). "Models optimized for real-world tasks reveal the task-dependent necessity of precise temporal coding in hearing". In: *Nature Communications* 15.1. ISSN: 2041-1723. DOI: 10.1038/s41467-024-54700-5 (cit. on pp. 24, 31, 69, 87).

Schädler, Marc René, David Hülsmeier, Giso Grimm, et al. (2021). *m-r-s/fade: FADE version 2.4.1*. DOI: 10.5281/ZENODO.3734164. URL: https://zenodo.org/record/4730596 (cit. on pp. 80, 83).

Schädler, Marc René, David Hülsmeier, Anna Warzybok, et al. (Jan. 2020). "Individual Aided Speech-Recognition Performance and Predictions of Benefit for Listeners With Impaired Hearing Employing FADE". In: *Trends in Hearing* 24, p. 233121652093892. DOI: 10.1177/2331216520938929 (cit. on pp. 55, 57, 61, 72).

Schädler, Marc René and Birger Kollmeier (Apr. 2015). "Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition". In: *Journal of the Acoustical Society of America* 137.4, pp. 2047–2059. ISSN: 1520-8524. DOI: 10.1121/1.4916618 (cit. on pp. 28, 74, 82).

Schädler, Marc René, Bernd T. Meyer, and Birger Kollmeier (May 2012). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition". In: *The Journal of the Acoustical Society of America* 131.5, pp. 4134–4151. ISSN: 1520-8524. DOI: 10.1121/1.3699200 (cit. on p. 79).

Schädler, Marc René, Anna Warzybok, Stephan D. Ewert, et al. (May 2016a). "A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception". In: *Journal of the Acoustical Society of America* 139.5, pp. 2708–2722. DOI: 10.1121/1.4948772 (cit. on pp. 72, 79, 82, 85).

— (2016b). "Robust automatic speech recognition and modeling of auditory discrimination experiments with auditory spectro-temporal features". PhD thesis. Carl von Ossietzky Universität Oldenburg. URL: http://oops.uni-oldenburg.de/2844/ (cit. on pp. 74, 79).

Schädler, Marc René, Anna Warzybok, Sabine Hochmuth, et al. (May 2015). "Matrix sentence intelligibility prediction using an automatic speech recognition system". In: *International Journal of Audiology* 54.sup2, pp. 100–107. ISSN: 1708-8186. DOI: 10.3109/14992027.2015.1061708 (cit. on pp. 17, 24, 27, 28, 60, 69, 72, 73, 83).

Scharf, M. K. (2023). *maxamazing/MicrophoneCalibrationEstimator: Microphone Calibration Estimation with Resonating Bottles*. DOI: 10.5281/ZENODO.10214506 (cit. on p. 40).

— (2025). *maxamazing/sipgof: Single Psychometric Function Goodness Of Fit Measure*. DOI: 10.5281/ZENODO.10277694 (cit. on pp. 30, 40, 52, 60).

Scharf, M. K. et al. (Sept. 2022). "Lombard Effect for Bilingual Speakers in Cantonese and English: importance of spectro-temporal features". In: *Proc. Interspeech 2022*. Ed. by Hanseok Ko and John H. L. Hansen. Incheon, South Korea: ISCA. DOI: 10.21437/Interspeech.2022-10235 (cit. on pp. 42, 55, 69, 72, 79).

Scharf, Maximilian Karl (Mar. 2023). *MicrophoneCalibrationEstimator*. Online repository. URL: https://github.com/maxamazing/MicrophoneCalibrationEstimator (cit. on pp. 43, 47).

Schell-Majoor, Lena, Andreas Büchner, and Birger Kollmeier (2023). "Auf dem Weg zur virtuellen Hörklinik: Grundlagen und Lösungswege aus dem Exzellenzcluster Hearing4All". de. In: DOI: 10.3205/23DGA009 (cit. on pp. 23, 88).

Schlittenlacher, Josef, Richard E. Turner, and Brian C. J. Moore (July 2018). "Audiogram estimation using Bayesian active learning". In: *The Journal of the Acoustical Society of America* 144.1, pp. 421–430. DOI: 10.1121/1.5047436 (cit. on p. 60).

Schneider, Steffen et al. (2019). "wav2vec: Unsupervised Pre-training for Speech Recognition". In: DOI: 10.48550/ARXIV.1904.05862 (cit. on p. 31).

Serpanos, Yula C. et al. (Nov. 2018). "The Accuracy of Smartphone Sound Level Meter Applications With and Without Calibration". In: *American Journal of Speech-Language Pathology* 27.4, pp. 1319–1328. DOI: 10.1044/2018_ajslp-17-0171 (cit. on p. 42).

Shen, Yi and Allison B. Kern (Jan. 2018). "An Analysis of Individual Differences in Recognizing Monosyllabic Words Under the Speech Intelligibility Index Framework". In: *Trends in Hearing* 22, p. 233121651876177. DOI: 10.1177/2331216518761773 (cit. on pp. 31, 65).

Shen, Yi and Lauren Langley (May 2023). "Spectral weighting for sentence recognition in steady-state and amplitude-modulated noise". In: *JASA Express Letters* 3.5. ISSN: 2691-1191. DOI: 10.1121/10.0017934 (cit. on p. 64).

Shen, Yi, Donghyeon Yun, and Yi Liu (Sept. 2020). "Individualized estimation of the Speech Intelligibility Index for short sentences: Test-retest reliability". In: *The Journal of the Acoustical Society of America* 148.3, pp. 1647–1661. DOI: 10.1121/10.0001994 (cit. on p. 65).

Skowronski, Mark D. and John G. Harris (May 2006). "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments". In: *Speech Communication* 48.5, pp. 549–558. ISSN: 0167-6393. DOI: 10.1016/j.specom.2005.09.003 (cit. on p. 36).

Smits, Cas et al. (Oct. 2022). "The one-up one-down adaptive (staircase) procedure in speech-in-noise testing: Standard error of measurement and fluctuations in the track". In: *The Journal of the Acoustical Society of America* 152.4, pp. 2357–2368. DOI: 10.1121/10.0014898 (cit. on pp. 58, 59).

Stowe, Lauren M. and Edward J. Golob (July 2013). "Evidence that the Lombard effect is frequency-specific in humans". In: *The Journal of the Acoustical Society of America* 134.1, pp. 640–647. DOI: 10.1121/1.4807645 (cit. on p. 35).

Summers, W. Van et al. (Sept. 1988). "Effects of noise on speech production: Acoustic and perceptual analyses". In: *The Journal of the Acoustical Society of America* 84.3, pp. 917–928. ISSN: 1520-8524. DOI: 10.1121/1.396660 (cit. on pp. 35, 36, 72).

Swanepoel, De Wet, Jackie L. Clark, et al. (Jan. 2010). "Telehealth in audiology: The need and potential to reach underserved communities". In: *International Journal of Audiology* 49.3, pp. 195–202. ISSN: 1708-8186. DOI: 10.3109/14992020903470783 (cit. on pp. 42, 50).

Swanepoel, De Wet and James W. Hall (Mar. 2010). "A Systematic Review of Telehealth Applications in Audiology". In: *Telemedicine and e-Health* 16.2, pp. 181–200. ISSN: 1556-3669. DOI: 10.1089/tmj.2009.0111 (cit. on pp. 42, 50).

Taal, Cees H. et al. (Sept. 2011). "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7, pp. 2125–2136. ISSN: 1558-7924. DOI: 10.1109/TASL.2011.2114881 (cit. on p. 31).

Treutwein, Bernhard (Sept. 1995). "Adaptive psychophysical procedures". In: *Vision Research* 35.17, pp. 2503–2522. ISSN: 0042-6989. DOI: 10.1016/0042-6989(95)00016-x (cit. on pp. 24, 50).

Trujillo, James et al. (Aug. 2021). "Speakers exhibit a multimodal Lombard effect in noise". In: *Scientific Reports* 11.1. DOI: 10.1038/s41598-021-95791-0 (cit. on pp. 35, 36).

Uncertainty of measurement -Part 3: Guide to the expression of uncertainty in measurement(GUM:1995) (1995). International Organization for Standardization (cit. on p. 36).

Vaidyanathan, P. P. (May 2008). "Eigenfunctions of the Fourier Transform". In: *IETE Journal of Education* 49.2, pp. 51–58. ISSN: 0974-7338. DOI: 10.1080/09747338.2008.11673800 (cit. on p. 100).

Verpackungswesen, DIN-Normenausschuss (2021a). *DIN 6075-1:2021-06, Packmittel Flaschen Teil 1: Vichyform 1*. DOI: 10.31030/3257687 (cit. on pp. 44, 45).

— (2021b). *DIN 6075-2:2021-06, Packmittel Flaschen Teil 2: Vichyform 2*. DOI: 10.31030/3257693 (cit. on pp. 44, 45).

— (2021c). *DIN EN ISO 12821:2020-02, Verpackungen aus Glas: Kronenmundstück 26 H 180 - Maße (ISO12821:2019)- Deutsche Fassung EN ISO 12821:2019*. DOI: 10.31030/3060769 (cit. on pp. 44, 45).

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2 (cit. on p. 52).

Vogten, L. L. M. (1974). "Pure-Tone Masking: A New Result from a New Method". In: *Facts and Models in Hearing*. Springer Berlin Heidelberg, pp. 142–155. ISBN: 9783642659027. DOI: 10.1007/978-3-642-65902-7_20 (cit. on p. 25).

Völk, Florian and Hugo Fastl (Oct. 2011). "Locating the Missing 6 dB by Loudness Calibration of Binaural Synthesis". In: URL: https://aes2.org/publications/elibrary-page/?id=16014 (cit. on p. 46).

Wagener, K, T Brand, and Birger Kollmeier (1999). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache. I-III: Design, Optimierung und Evaluation des Oldenburger Satztests (Development and evaluation of a sentence test for the German language. I-III: Design, optimization and evaluation of the Oldenburg sentence test)". In: *Zeitschrift für Audiologie (Audiological Acoustics)* 38, pp. 4–15 (cit. on pp. 21–23).

Wagener, Kirsten Carola, Thomas Brand, and Birger Kollmeier (Jan. 2006). "The role of silent intervals for sentence intelligibility in fluctuating noise in hearing-impaired listeners". In: *International Journal of Audiology* 45.1, pp. 26–33. ISSN: 1708-8186. DOI: 10.1080/14992020500243851 (cit. on pp. 55, 83, 105).

Weisser, Adam and Jörg M. Buchholz (Jan. 2019). "Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions". In: *The Journal of the Acoustical Society of America* 145.1, pp. 349–360. DOI: 10.1121/1.5087567 (cit. on p. 42).

Wier, Craig C., Walt Jesteadt, and David M. Green (Jan. 1977). "Frequency discrimination as a function of frequency and sensation level". In: *The Journal of the Acoustical Society of America* 61.1, pp. 178–184. ISSN: 1520-8524. DOI: 10.1121/1.381251 (cit. on p. 39).

Wong, Lena L. N. et al. (Apr. 2007). "Development of the Cantonese speech intelligibility index". In: *The Journal of the Acoustical Society of America* 121.4, pp. 2350–2361. ISSN: 1520-8524. DOI: 10.1121/1.2431338 (cit. on pp. 31, 64).

Xu, Chen et al. (Jan. 2024). "How Does Inattention Influence the Robustness and Efficiency of Adaptive Procedures in the Context of Psychoacoustic Assessments via Smartphone?" In: *Trends in Hearing* 28. ISSN: 2331-2165. DOI: 10.1177/23312165241288051 (cit. on pp. 24, 50).

Yip, Moira (Aug. 2002). *Tone*. Cambridge University Press. DOI: 10.1017/cbo9781139164559 (cit. on p. 34).

Yoho, Sarah E. et al. (Mar. 2018). "Speech-material and talker effects in speech band importance". In: *The Journal of the Acoustical Society of America* 143.3, pp. 1417–1426. ISSN: 1520-8524. DOI: 10.1121/1.5026787 (cit. on pp. 31, 64).

Young, Steve et al. (2002). *The HTK book*. Vol. 3. 175. Cambridge university engineering department, p. 12 (cit. on pp. 28–30).

Zwicker, E (1952). "Die Grenzen der Hörbarkeit der Amplitudenmodulation und der Frequenzmodulation eines Tones". In: *Acta Acustica United With Acustica* 2.3, pp. 125–133 (cit. on p. 46).