# "Can you hear me now?"
# Automatic assessment of background noise intrusiveness and speech intelligibility in telecommunications

THÈSE N$^{O}$ 7010 (2016)

PRÉSENTÉE LE 14 JUIN 2016

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE L'IDIAP
PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Raphaël Marc ULLMANN

acceptée sur proposition du jury:

Prof. J.-Ph. Thiran, président du jury
Prof. H. Bourlard, directeur de thèse
Prof. M. Cooke, rapporteur
Dr J. Beerends, rapporteur
Dr H. Lissek, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2016

Pour Sarah, qui nous a toujours encouragés à poursuivre nos études.

# Acknowledgments

Writing a thesis is a long and arduous process, and a large number of people contributed to this work through their availability, support and friendship.

First, I wish to thank Dr. Jens Berger of SwissQual AG for encouraging me to pursue a Ph.D., and entering into an industrial collaboration that ensured a direct link with the telecommunications industry, and with the standardization work at the International Telecommunication Union during the 18 months of the CTI project.

I am also indebted to my thesis director, Prof. Hervé Bourlard, for giving me the privilege to join the Idiap speech group, and for his availability, insightful feedback and guidance throughout those four years. Thank you Hervé for your tireless support.

I would like to thank Dr. John Beerends, Dr. Hervé Lissek, Prof. Martin Cooke, and Prof. Jean-Philippe Thiran for kindly accepting to be part of my thesis jury, and for their stimulating feedback.

I have been very fortunate to share an office with three very talented and knowledgeable researchers: Drs. Marc Ferràs, Mathew Magimai.-Doss and Srikanth Madikeri. Thank you for countless fruitful discussions, advice and your friendship through the ups and downs of those last four years. Your interest and comments significantly helped me progress in my work. In particular, Dr. Ferràs shared some insightful comments on the effect of the threshold parameter in Section 4.6.2, and Dr. Magimai.-Doss suggested the experiment of Section 4.3.2.

I wish to thank Dr. Hervé Lissek of the Laboratory of Electromagnetics and Acoustics (LEMA) at EPFL for having generously provided access to the facilities in his lab in order to record speakers and conduct listening experiments. Many thanks go to Dr. Lukas Rohr, as well as Drs. Jens Berger and Anna Llagostera of SwissQual AG for their help in setting up audio equipment and recording speakers. I also thank Gilles Courtois and Baptiste Crettaz for interesting discussions and support while being at LEMA.

I highly appreciated the interesting discussions with the colleagues of Study Group 12 at the International Telecommunication Union, in particular Vincent Barriac, Dr. John Beerends, Dr. Irina Cotanis, Dr. Hans Wilhelm Gierlich, Dr. Ludovic Malfait, Prof. Sebastian Möller, Prof. Alexander Raake and Christian Schmidmer, all of whom I would like to thank here.

My time at Idiap Research Institute was enriched by the company of exceptional colleagues. Thank you Afsaneh, Ajay, Alexandre, Alexandros, André, Branko, Cijo, Cosmin, Dario, David, Dhananjay, Dimitri, Gulcan, Harsha, Hui, Ivana, Jagan, James, Joan Isaac, Joël, John, Kenneth, Lakshmi, Laurent, Leo, Manuel, Marco, Marzieh, Mohammad, Nesli, Nikos, Paco, Paul, Pedro, Phil, Pierre-Edouard, Pranay, Ramya, Rémi, Roy, Rui, Salim, Samira, Thomas and Xingyu, for your wit, kindness and good times spent together. I also met great people when I came to EPFL for classes or study: thank you Anıl, Christina, Christophe, Damien, Didrik, Dimitri, Gabriel, Hamed, Marina, Murat, Pınar, Sasan, Sibylle, Tom, Vijay and Yoav.

I highly enjoyed working with Rubén Braojos Lopez and Elvira Pirondini during my time in the EDEE Student Committee: thanks to your enthusiasm and energy, everything we organized seemed effortless.

Thank you my dear friends in the Lausanne, Zurich and Basel regions, for always being there and encouraging me.

This work was successively funded by two Swiss governmental agencies, the Commission for Technology and Innovation (CTI) and armasuisse, both of which are acknowledged here. My thanks go to Drs. Florent Monay, Yann Rodriguez and Phil Garner for their help with the associated proposal writing and project reporting. I also thank the Idiap IT team, especially Frank, Louis-Marie and Norbert for their help with technical issues big and small.

This thesis benefited from the availability of two comprehensive, high-quality datasets for intelligibility assessment, provided by the Organizers of the Blizzard Challenge and the Public Safety Communications Research program (PSCR), both of which are acknowledged here. My thanks also go to Nicolas Gninenko for his very useful analyses and evaluations of the PSCR dataset during his internship at Idiap.

During those four years, my parents and parents-in-law have shown exemplary patience, understanding and support; thank you so much.

Most importantly, thank you Emilie for your unwavering love, support and patience.

*Lausanne, 30 May 2016* Raphaël

# Abstract

This thesis deals with signal-based methods that predict how listeners perceive speech quality in telecommunications. Such tools, called *objective* quality measures, are of great interest in the telecommunications industry to evaluate how new or deployed systems affect the end-user quality of experience. Two widely used measures, ITU-T Recommendations P.862 "PESQ" and P.863 "POLQA", predict the overall listening quality of a speech signal as it would be rated by an average listener, but do not provide further insight into the composition of that score. This is in contrast to modern telecommunication systems, in which components such as noise reduction or speech coding process speech and non-speech signal parts differently. Therefore, there has been a growing interest for objective measures that assess different quality *features* of speech signals, allowing for a more nuanced analysis of how these components affect quality. In this context, the present thesis addresses the objective assessment of two quality features: *background noise intrusiveness* and *speech intelligibility*.

The perception of background noise is investigated with newly collected datasets, including signals that go beyond the traditional telephone bandwidth, as well as Lombard (effortful) speech. We analyze listener scores for noise intrusiveness, and their relation to scores for perceived speech distortion and overall quality. We then propose a novel objective measure of noise intrusiveness that uses a sparse representation of noise as a model of high-level auditory coding. The proposed approach is shown to yield results that highly correlate with listener scores, without requiring training data.

With respect to speech intelligibility, we focus on the case where the signal is degraded by strong background noises or very low bit-rate coding. Considering that listeners use prior linguistic knowledge in assessing intelligibility, we propose an objective measure that works at the phoneme level and performs a comparison of phoneme class-conditional probability estimations. The proposed approach is evaluated on a large corpus of recordings from public safety communication systems that use low bit-rate coding, and further extended to the assessment of synthetic speech, showing its applicability to a large range of distortion types.

The effectiveness of both measures is evaluated with standardized performance metrics, using corpora that follow established recommendations for subjective listening tests.

*Keywords:* Speech quality, intelligibility, noise intrusiveness, objective assessment, speech perception, sparse coding, posterior features

# Résumé

Cette thèse traite de méthodes automatiques permettant de prédire la perception de la qualité vocale dans les systèmes de télécommunication. Ces outils, appelés *instruments d'évaluation de la qualité*, sont d'une grande utilité pour l'industrie des télécommunications afin d'évaluer comment des systèmes existants ou nouveaux impactent la qualité perçue par les utilisateurs. Deux instruments de mesure largement utilisés, les Recommandations UIT-T P.862 "PESQ" et P.863 "POLQA", prédisent la qualité vocale d'un signal de parole telle qu'elle serait perçue par un utilisateur moyen, mais ne donnent pas d'indication supplémentaire sur la composition de ce score. Cependant, les systèmes de télécommunication modernes comportent souvent des composants de débruitage ou de codage qui ont un effet différent sur les parties de parole et de bruit dans le signal. En conséquence, des instruments permettant de mesurer différentes *caractéristiques* de la qualité des signaux de parole ont récemment gagné en importance. C'est dans ce contexte que cette thèse aborde l'évaluation automatique de *la gêne perçue des bruits de fond*, ainsi que de *l'intelligibilité de la parole*.

La perception des bruits de fond est étudiée par le biais d'une collection récente de parole bruitée de systèmes de télécommunication. Cette collection contient notamment des signaux de systèmes récents à bande audio élargie par rapport à la téléphonie traditionnelle, ainsi que des enregistrements de parole à effet Lombard (i.e., avec effort vocal). L'impact de ces facteurs sur la perception de la gêne des bruits fond, sur la déformation et sur la qualité globale de la parole est évalué. Les relations entre ces trois caractéristiques de qualité sont également établis. Ces analyses ouvrent sur un nouvel instrument de mesure de la gêne de bruits de fond, basé sur une représentation parcimonieuse du bruit comme modèle de la perception auditive à haut niveau. Les résultats obtenus avec ce modèle montrent une forte corrélation avec les évaluations subjectives de la gêne, sans nécessiter de données d'apprentissage.

L'évaluation automatique de *l'intelligibilité* de la parole se concentre sur les scénarios de dégradation par bruits de fond ou par codage à très bas débit. Partant du constat que les utilisateurs appliquent leurs connaissances linguistiques dans l'évaluation de l'intelligibilité, une nouvelle approche basée sur des paramètres phonétiques est proposée. Plus particulièrement, une mesure d'intelligibilité est déterminée en comparant les séquences de probabilités a-posteriori de phonèmes du signal de parole original et du signal dégradé. L'instrument de mesure développé est appliqué à une large collection d'enregistrements de parole codée à bas débit provenant de systèmes de communication d'urgence, et étendu à l'évaluation de la parole synthétique, montrant ainsi son utilité pour des types de distorsions variés.

L'efficacité des méthodes proposées est évaluée à travers des mesures de performances standardisées et sur des bases de données conformes aux recommandations sur l'évaluation subjective de la qualité.

*Mots clefs :* Qualité vocale, intelligibilité, gêne du bruit, évaluation objective, perception de la parole, codage parcimonieux, paramètres postérieurs

# Zusammenfassung

Diese Doktorarbeit befasst sich mit signalbasierten Verfahren zur Vorhersage der wahrgenommenen Sprachqualität in Telekommunikationssystemen. Solche sogenannte instrumentelle Verfahren sind in der Telekommunikationsindustrie von grossem Interesse, da sie erlauben, den Einfluss von neuen oder bereits eingesetzten Systemen auf die Qualitätswahrnehmung des Endbenutzers zu bestimmen. Zwei weit verbreitete Verfahren, ITU-T Empfehlungen P.862 "PESQ" und P.863 "POLQA", schätzen die Gesamtqualität, wie sie von einer durchschnittlichen Versuchsperson beurteilt würde, geben aber keinen Einblick auf die Zusammensetzung diese Urteils. Dies steht im Gegensatz zu modernen Telekommunikationssystemen, in welchen Komponenten wie Geräuschunterdrückung und Sprachkodierung Sprach- und Nichtsprachanteile im Signal unterschiedlich beeinflussen. Aus diesem Grund stehen zunehmend instrumentelle Verfahren, welche verschiedene *Qualitätsmerkmale* von Sprachsignalen bewerten im Vordergrund, da sie Aufschlüsse über die Auswirkung verschiedener Komponenten auf die Gesamtqualität geben. In diesem Kontext befasst sich diese Arbeit mit der instrumentellen Schätzung der *Lästigkeit von Hintergrundgeräuschen* und der *Sprachverständlichkeit*.

Die Wahrnehmung von Hintergrundgeräuschen wird mittels einer eigens erstellter Datenbank an geräuschbehafteten Sprachsignalen aus Telekommunikationssystemen erforscht. Erstmals wurden dabei auch Signale aus neueren Systemen mit erweiterten Audiobandbreiten, sowie Sprachaufnahmen mit dem Lombard-Effekt (Sprechweise im Störgeräusch) berücksichtigt. Der Einfluss dieser Faktoren auf die wahrgenommene Lästigkeit, Sprachverzerrung und Gesamtqualität wird untersucht, und die gegenseitige Abhängigkeit dieser drei Qualitätsmerkmale analysiert. Die gewonnen Erkenntnisse werden zur Entwicklung eines neuen instrumentellen Schätzers der Lästigkeit von Hintergrundgeräuschen angewandt, welcher auf einer spärlichen Darstellung des Geräuschsignals als Modell der auditiven Kodierung beruht. Das entwickelte Verfahren benötigt keine Trainingsdaten und errechnet Schätzwerte, die stark mit der Beurteilung von Probanden korrelieren.

Die instrumentelle Bestimmung der Sprachverständlichkeit wird in Bezug auf die Verständlichkeit im Störgeräusch, sowie auf Verzerrungen durch sehr niedrig bitratige Kodierer angegangen. Davon ausgehend, dass Versuchspersonen sprachliche Kenntnisse zur Beurteilung der Sprachverständlichkeit anwenden, wird ein neues, auf Phonemmerkmalen basierendes Messverfahren entwickelt. Bei diesem Verfahren werden a-posteriori Wahrscheinlichkeiten von Phonemmerkmalen zwischen dem Original und dem verzerrten Sprachsignal verglichen, und die gemessenen Unterschiede als Schätzmass angewandt. Das entwickelte Verfahren wird mit einer umfangreichen Datenbank an niedrig bitratigen Sprachsignalen aus Kommunikationssystemen der öffentlichen Sicherheit evaluiert, und ferner auf die Verständlichkeitsschätzung künstlicher Sprache erweitert.

Die Wirksamkeit der entwickelten Ansätze wird mittels standardisierter Gütemasse überprüft, unter Benutzung von Testdaten, welche nach empfohlenen Methoden gesammelt wurden.

*Stichwörter:* Sprachqualität, Sprachverständlichkeit, Lästigkeit von Hintergrundgeräuschen, instrumentelle Sprachqualitätsschätzung, Sprachwahrnehmung, spärliche Signalkodierung, Posterior-Merkmale

# Contents

# List of Figures

# List of Tables

# List of acronyms

**3GPP / 3GPP2** Third-Generation Partnership Project (2)

**AI** Articulation Index, an objective intelligibility measure based on the signal-to-noise ratio of a speech signal [French and Steinberg, 1947; Kryter, 1962]

**ANN** Artificial Neural Network

**ANSI** American National Standards Institute

**API** Application Programming Interface

**ASR** Automatic Speech Recognition

**codec** Contraction of "coder"-"decoder"

**dB** Decibels

**DRT** Diagnostic Rhyme Test, an intelligibility test method

**DTW** Dynamic Time Warping

**ETSI** European Telecommunications Standards Institute

**HMM** Hidden Markov Model

**HTK** Hidden Markov model Toolkit [Young et al., 2006]

**IEC** International Electrotechnical Commission

**ITU** International Telecommunication Union

**ITU-T** Telecommunication Standardization Sector of ITU

**kbps** kilobits per second

**MFCC** Mel-Frequency Cepstral Coefficients

**MLP** Multilayer Perceptron

**MOS** Mean Opinion Score

**MPTK** Matching Pursuit Toolkit [Krstulovic and Gribonval, 2006]

**MRASTA** Multi-resolution RelAtive SpecTrAl filters [Hermansky and Fousek, 2005]

**MRT** Modified Rhyme Test, an intelligibility test method

**NB** Narrowband, the traditional 195–3700 Hz telephone audio bandwidth

**NR** Noise Reduction

**OIM** Objective Intelligibility Measure

**PESQ** Perceptual Evaluation of Speech Quality, standardized in ITU-T Rec. P.862 [2001]

**PLP** Perceptual Linear Prediction [Hermansky, 1990]

**POLQA** Perceptual Objective Listening Quality Assessment, standardized in ITU-T Rec. P.863 [2011]

**PSCR** Public Safety Communications Research, a joint research program involving several U.S. governmental agencies [http://www.pscr.gov/]

**rmse** root-mean-square error

**SII** Speech Intelligibility Index, an objective intelligibility measure based on the Articulation Index (AI) [ANSI S3.5, 1997]

**SNR** Signal-to-noise ratio

**STI** Speech Transmission Index [Steeneken and Houtgast, 1980; 2002]

**SUS** Semantically Unpredictable Sentence

**SWB** Super-wideband, i.e., 50–14 000 Hz audio bandwidth

**TTS** Text-To-Speech

**VoIP** Voice over IP (Internet Protocol)

**WA** Word Accuracy, a measure of intelligibility

**WB** Wideband, i.e., 50–7000 Hz audio bandwidth

**WER** Word Error Rate

# 1 Introduction

“Ladies and gentlemen, before this presentation begins, would you please silence your phones.” According to the International Telecommunication Union (ITU), there were an estimated seven billion mobile-cellular telephone subscriptions in the world in 2015 [ITU, 2016]. This is in addition to the many other users of fixed-line and voice over IP (VoIP) telecommunication services. The technologies that are used to process, encode and transmit speech in these services have considerably evolved over the last decades, and continue to do so. When developing or deploying new speech technologies, it is desirable to verify their impact on the end-user quality of experience. This is of particular interest to the telecommunications industry, which strives for high speech quality in order to reduce customer churn and promote new services.

The most reliable way to assess perceived speech quality is to conduct a subjective listening test with a representative panel of users. However, such tests are costly and time-consuming, and quickly become prohibitive if they are to be performed repeatedly during the development of a new speech technology. There is thus a strong motivation for automatic methods, called *objective measures*, that can predict speech quality as it would be evaluated in a subjective test. In addition to being a fast and inexpensive alternative to subjective testing, objective measures provide repeatable quality estimations, making them particularly useful for the development and monitoring of telecommunication systems.

This thesis focuses on signal-based objective measures, i.e., methods that are based on an analysis of the speech signal as it would be presented to a user of the telecommunication system under test. This approach has the advantage of being technology independent, and considers all sources of impairments or distortions in a transmission chain. ITU’s telecommunication standardization sector, called ITU-T, has standardized several signal-based speech quality measures, e.g., ITU-T Recommendations P.862 [“PESQ”, Rix, Beerends, Hollier, et al., 2001], P.563 [Malfait, Berger, and Kastner, 2006] and P.863 [“POLQA”, Beerends et al., 2013a; b]. These measures predict the *overall* listening quality of a speech signal as it would be rated by an average listener.

Recently however, there has been a growing interest in objective measures that predict specific quality *features*, allowing for a more fine-grained assessment of speech quality. Quality features provide useful information when the optimization of overall quality involves trade-offs, or for processing steps that only affect certain parts of the signal. Specifically, we address two quality features in this thesis:

- The perception of background noise in speech, and its assessment with the quality feature *background noise intrusiveness*. This is motivated by the growing use of telecommunication services in environments with uncontrolled noise levels, and by the recent extension of the telephone band to so-called wideband or super-wideband audio frequency ranges by mobile network operators. Both factors have increased the importance of noise reduction processing in telecommunications, which can be assessed with the subjective test method defined in [ITU-T Rec. P.835, 2003]. However, there is currently no ITU standard for the objective assessment of noise reduction processing.

- The *intelligibility* of speech in high levels of background noise, and/or after processing with very low bit-rate speech codecs. Speech intelligibility is of a particular interest in mission-critical telecommunication systems, and for the development of recent speech coding approaches based on speech synthesis.

The relevance of both topics is reflected by ongoing or recent work items at ITU-T on the objective assessment of noise reduction processing [ITU-T Study Group 12, 2013] and on the subjective assessment of speech intelligibility [ITU-T Rec. P.807, 2016], respectively.

## 1.1 Objectives

The main goal of this thesis is to propose new objective measures for the assessment of noise intrusiveness and speech intelligibility, focusing on their application in speech telecommunications. This involves finding signal representations that are suitable to predict human perception and — if possible — avoid the use of subjectively scored training data, since such data tends to be scarce.

As purely signal-based approaches, the proposed objective measures should be applicable to a wide range of distortion types as found in current telecommunication systems. In particular, this includes recent very low bit-rate speech coding approaches that are based on speech synthesis principles, and which may change both spectral and temporal properties of transmitted speech signals. Our focus on telecommunications further involves the use of evaluation data and metrics that follow recommended practices for subjective testing and performance evaluation, respectively.

## 1.2 Main contributions

The main contributions of this work are summarized below:

- Design and collection of a new dataset of noise-corrupted speech, with further distortions as found in contemporary telecommunication systems, and corresponding subjective scores. Speech signals and subjective scores are compliant with ITU-T Rec. P.835 [2003], which specifies the subjective test procedure for the evaluation of telecommunication systems that include noise reduction processing.
- Statistical analysis of the effect of signal bandwidth, presence of Lombard (i.e., effortful) speech and presentation level on the subjective quality features speech distortion, background noise intrusiveness, and overall quality [Ullmann, Bourlard, et al., 2013].
- Analysis of the interdependency of the three quality features, and of the possibility of exploiting this dependency and an existing overall quality measure toward their objective assessment [Ullmann, Berger, et al., 2013; Berger and Ullmann, 2013].
- Design of a novel objective measure of noise intrusiveness that is based on a sparse representation of noise in an auditory-inspired basis, and avoids the use of training data [Ullmann and Bourlard, 2016].
- Development of a new objective intelligibility measure that can be used to assess speech signals that have undergone changes to both spectral and temporal structure, as may be the case in low bit-rate telecommunication systems [Ullmann, Magimai.-Doss, et al., 2015].
- Extension of the developed intelligibility measure to the objective assessment of clean and noise-corrupted synthetic speech, and to automatic intelligibility assessment at the word level [Ullmann, Rasipuram, et al., 2015].

As can be seen from the above list, most of these contributions have already been published. The presentation in this thesis is more in-depth, and often includes results from additional experiments. These differences to the initial publications are stated in the appropriate chapters.

## 1.3 Thesis structure

The next thesis chapters are organized as follows:

- Chapter 2, *Background*, defines some of the terminology used in this thesis and provides a general introduction to subjective and objective quality assessment in speech telecommunications. The datasets that we use for our experiments are presented, and the evaluation metrics that are computed to determine the performance of proposed objective measures are explained. Finally, we introduce some of the signal modeling techniques that are exploited in our proposed approaches.

- Chapter 3, *Perception of background noise in speech telecommunications*, presents our first contribution, which are the datasets of subjectively evaluated, noise-corrupted speech that we collected. The subjective scores are analyzed in detail and provide insights on the perception of signal bandwidth, Lombard speech and presentation level in terms of the three quality features speech distortion, noise intrusiveness and overall quality. The results from our analyses highlight the importance of the exact task that is given to listeners in a subjective test, and also show the interdependency of the three quality features.

- Chapter 4, *Objective assessment of background noise intrusiveness*, is motivated by the insights gained in the previous chapter, and investigates a novel, sparse representation of noise in an auditory-inspired basis as an abstract model of auditory coding. The number of kernels or *atoms* in this representation is shown to model several factors in the perception of noise, and is used to derive a feature that highly correlates with subjective noise intrusiveness scores, outperforming or comparing to a traditional loudness-based feature.

- Chapter 5, *Objective intelligibility assessment for speech telecommunications*, deals with the case where very high levels of background noise, or low bit-rate coding may degrade speech to the point of compromising its intelligibility. We propose a novel objective intelligibility measure that is based on a comparison of phoneme posterior probabilities between original and degraded speech recordings. The proposed approach is evaluated on speech recordings degraded by low bit-rate speech codecs, and on a large dataset of noise-degraded speech from public safety communication systems.

- Chapter 6, *Objective intelligibility assessment of synthetic speech*, extends the work of Chapter 5 to the assessment of signals generated with text-to-speech (TTS) systems. These signals are used as examples of very strong changes in spectral and temporal structure that very low bit-rate coding may produce, but where the resulting speech signal may still be intelligible. The approach of Chapter 5 is evaluated on clean and noise-corrupted synthetic speech, and extended to perform intelligibility assessment with a *model* of the expected phonetic content. The model is shown to provide a more flexible reference for synthetic speech assessment, and allows to assess intelligibility at the word level. Finally, the analyses in this chapter also reveal the limitations of standard performance metrics.

- In Chapter 7, we review the key conclusions from our work, and indicate possible avenues for future research.

# 2 Background

In this chapter, we first define the mathematical notation and specific terms used in this thesis. We then provide some context on quality assessment approaches in telecommunications, explain the performance metrics with which these approaches are evaluated, and present the datasets used in our experiments. Finally, we introduce the different types of signal features, as well as two particular machine learning methods — artificial neural networks (ANN) and hidden Markov models (HMM) — that we will exploit in this work.

## 2.1 Notation and terminology

This thesis uses boldface symbols in lower- and uppercase to denote vectors and matrices, respectively. Subscripts are used for vector or time indices, and superscripts for class indices. Unless specified otherwise, vectors are assumed to be column vectors, i.e., $\mathbf{x} = [x_1, \dots, x_N]^\top$, with $(\ )^\top$ the transpose operator. Signal position indices are enclosed in parentheses for the continuous-time case, and in square brackets for discrete time. Finally, $P(\cdot)$ denotes the probability of a discrete random variable.

The terms introduced here are used throughout this thesis:

**Phones** describe the full set of speech sounds independently of the language. Phones are indicated inside square brackets, usually using symbols from the International Phonetic Alphabet (IPA).

**Phonemes** are the set of sounds that distinguish one word from another in a language [Gold et al., 2011, Chap. 23.2]. Depending on the language, some phones are therefore not used, and one phoneme can cover multiple phones if they carry the same meaning (so-called allophones). By convention, phonemes are enclosed in slashes, e.g., /k/.

**Speech intelligibility** as defined in [Kollmeier et al., 2008, Chap. 4.2] is “the proportion of speech items (e.g., syllables, words, or sentences) correctly repeated by (a) listener(s) for a given speech intelligibility test.” In this thesis, we focus on word-level intelligibility.

For the sake of disambiguation and brevity, we also define the following word meanings in the context of speech telecommunications:

**Reference** and **test** speech are the input and output signals, respectively, of a telecommunication system. The signals may be inserted (fed) and captured (recorded) either electrically, i.e., using a wired interface, or acoustically, i.e., at transducers of the system under test.

**Channel** refers to a transmission medium as used in telecommunications, typically a wired or radio connection.

**Condition** is used to describe the technical circumstances affecting the transmission of a speech signal, e.g., bandwidth limitation, addition of noise, speech coding and decoding, or transmission over lossy channels.

**Rating** denotes a single, subjective evaluation of a test speech signal by an individual listener.

**Score** describes a numerical value that is indicative of the quality of a test speech signal, as obtained by averaging ratings from multiple listeners, or with an objective measure.

In this thesis, we evaluate the objective scores of proposed measures by comparing them to ground truth subjective scores, after averaging both scores per condition (i.e., across speakers or sentences). This is because in telecommunications, the interest generally lies in assessing the impact of the system and its parameters on a representative sample of speech recordings, avoiding variability due to a particular speaker or sentence [Rix, Beerends, Kim, et al., 2006, Sec. II-D]. We introduce subjective and objective assessment approaches in the next section, and performance metrics for the evaluation of objective measures in Section 2.3.

## 2.2 Quality assessment of speech signals in telecommunications

We now briefly review quality assessment approaches, focusing on signal-based methods. This leaves out so-called *parametric* approaches, which seek to estimate speech quality from the specifications of the telecommunication system [e.g., ITU-T Rec. G.107, 2015, known as the “E-model”], or from auxiliary measurements such as the occurrence of channel losses [e.g., Raake, 2006]. Signal-based assessment has the advantage of being technology independent, and allows to consider the entire transmission chain (so-called end-to-end assessment).

### 2.2.1 Quality features

Speech at the output of a telecommunication system affects the user experience in different ways, e.g., with regard to voice transmission quality, conversation effectiveness and ease of communication [Möller, 2000, Sec. 2.2]. In practice, user experience is assessed either through listening-only tests, focusing on the perception of the transmitted voice, or in simulated telephone conversations, allowing to also consider interaction factors like echoes or delays.

The results of these tests are quantified in terms of so-called *quality features* such as overall quality, listening effort, or speech intelligibility [ITU-T Rec. P.800, 1996; ANSI S3.2, 2009].[1] Recently, new quality features have been defined for specific signal properties, with the goal of assessing trade-offs in telecommunication system design, or components that only affect certain parts of the signal. Examples are the quality features *speech distortion* and *background noise intrusiveness* studied in this thesis, which were defined to assess the effect of noise reduction processing [ITU-T Rec. P.835, 2003]. Further quality features are currently under study at ITU-T, e.g., to determine perceptually orthogonal signal properties, or to help identify technical issues with expert listeners [ITU-T Study Group 12, 2011a; b].

While intelligible speech remains the fundamental feature of a telecommunication system, these other quality features become relevant once intelligibility is sufficient [Möller and Heusdens, 2013, Sec. IV]. The interest in new quality features can thus be seen as a result of the increased quality and complexity of telecommunication systems.

### 2.2.2 Subjective assessment approaches

Subjective quality assessment involves multiple listeners and generally takes place in a controlled environment (i.e., with regard to ambient noise levels, playback equipment and absence of distractions), although recently crowdsourcing approaches have also been investigated [e.g., Ribeiro et al., 2011]. Listeners are presented with test speech signals containing words or short sentences, pronounced by different speakers and in a random order of presentation. The task given to listeners depends on the quality features of interest, and can be categorized into *judgment* and *functional* tests [Van Heuven and Van Bezooijen, 1995].

Judgment tests collect listeners' *opinion* of a given quality feature, usually by means of an ordinal rating scale (i.e., where scale items can be sorted from lowest to highest). Absolute category rating (ACR) scales [e.g., ITU-T Rec. P.800, 1996, Annex B] are used most often, and require subjects to select the scale item that best describes their opinion of a test speech recording, given their *expectations* of speech telecommunications. A short training session with a variety of quality levels is presented before the actual test to set expectations, and to ensure that listeners use the entire scale range. Nevertheless, test results still depend on personal experience and cultural factors, and will thus vary between listener panels [ITU-T Rec. P.1401, 2012, Sec. 7.3].

By contrast, functional tests evaluate how well speech from the system under test fulfills its *communicative purpose* [Van Heuven and Van Bezooijen, 1995]. Such tests are typically used to assess intelligibility, and can follow a closed- or open-response format. The Diagnostic Rhyme Test (DRT) [Voiers, 1967] and the Modified Rhyme Test (MRT) [House et al., 1965] are two popular closed-response intelligibility tests, and require listeners to recognize one word per recording, given two or more possible responses that differ in one phoneme only.

[1] Note that we cited the most recent versions of ITU-T Rec. P.800 and ANSI S3.2 here. However, the original versions of these standards date from 1976 and 1971, respectively.

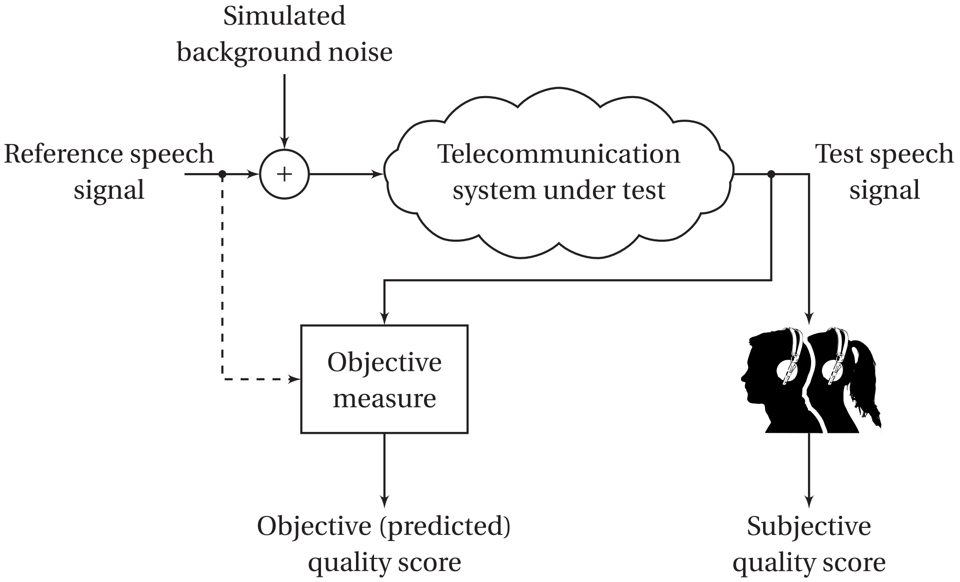

Simulated background noise
Reference speech signal
+
Telecommunication system under test
Test speech signal
Objective measure
Objective (predicted) quality score
Subjective quality score

Figure 2.1 – Signal flow in subjective and objective quality assessment as used in this thesis. Listeners evaluate quality features of the system under test using the test speech signal alone. A signal-based objective measure predicts the listener score from the same signal (*no-reference* model), or by means of a comparison to the reference signal (*full-reference* model).

Conversely, open-response tests allow for unconstrained listener feedback and are typically used with longer stimuli. An example is the SUS (semantically unpredictable sentences) test [Benoît et al., 1996], where subjects transcribe entire sentences. Open-response tests can be more difficult to conduct due to ambiguities in listener responses (e.g., spelling errors), but are less prone to *ceiling effects* in high-intelligibility conditions, where closed-response tests may offer insufficient differentiation [Schmidt-Nielsen, 1992].

The datasets used in this thesis include scores from both judgment and functional tests, and are presented in Section 2.4. Specifically, we use the judgment test defined in ITU-T Rec. P.835 [2003] to assess the impact of background noise and noise reduction processing with the quality features *speech distortion*, *background noise intrusiveness* and *overall listening quality*. For speech intelligibility, we use publicly available datasets with MRT and SUS test results for recordings from public safety communication systems and speech synthesizers, respectively.

### 2.2.3 Objective assessment approaches

The goal of an objective measure is to predict the outcome of a subjective test. Since these outcomes generally consist of quality scores obtained by averaging listener ratings, an objective measure can be thought of as a model of an average listener. Signal-based measures can estimate a quality score from the same test speech signal that was presented in a subjective test, or take the reference signal as additional input.

These two measurement approaches are referred to as *no-reference* and *full-reference* assessment, respectively, and are shown schematically in Figure 2.1. No-reference measures typically use a model of speech production or look for specific degradations to estimate quality features. Such models are also called *non-intrusive*, as they only require access to one end of the transmission chain, making them interesting for monitoring deployed systems. Conversely, full-reference measures require control of both ends when testing a transmission chain, and are thus also known as *intrusive* models. Full-reference measures perform a comparison of the reference and test signal, e.g., using a similarity metric, to derive a quality score. These scores are typically more accurate than the ones obtained with no-reference approaches, since even small degradations can be detected. For a recent overview of standardized objective measures, the reader is referred to [Möller, Chan, et al., 2011].

The objective measures proposed in this thesis are all based on a full-reference approach, and predict quality scores for background noise intrusiveness and speech intelligibility. State-of-the-art measures for either quality feature are reviewed in the respective chapters, i.e., Chapter 4 for noise intrusiveness, and chapters 5 and 6 for intelligibility, respectively.

## 2.3 Performance metrics

In this section, we introduce the metrics that we will use to evaluate the prediction accuracy of proposed objective measures. We use two metrics that compare objective predictions to subjective ground truth scores: Pearson's correlation coefficient, and a modified measure of prediction error. Both metrics are specified in the ITU standard for the evaluation of quality prediction models [ITU-T Rec. P.1401, 2012], and will be used in sections 3.6, 4.6, 5.5 and 6.5.

### 2.3.1 Correlation coefficient

Pearson's correlation coefficient $R$ is the most frequently used metric for objective measures. The correlation between (objective) predictions $\mathbf{o}$ and (subjective) scores $\mathbf{s}$ is defined as

$$R(\mathbf{s},\mathbf{o}) = \frac{\sum_i^N (s_i - \bar{s})(o_i - \bar{o})}{\sqrt{\sum_i^N (s_i - \bar{s})^2}\sqrt{\sum_i^N (o_i - \bar{o})^2}} \,. \tag{2.1}$$

This metric measures the *linear relationship* between $\mathbf{s}$ and $\mathbf{o}$, but has several drawbacks:

- As mentioned in Section 2.2.2, subjective scores can be biased due to cultural effects or imbalanced test designs [ITU-T Rec. P.1401, 2012, Sec. 7.3]. Therefore, the relationship between objective and subjective scores need not be linear, but only *monotonic*, i.e., the objective score should predict the increase or decrease of perceived quality.

- The correlation coefficient does not consider inter-subject variability, i.e., the confidence interval around each score $s_i$ that results from averaging multiple listener ratings.

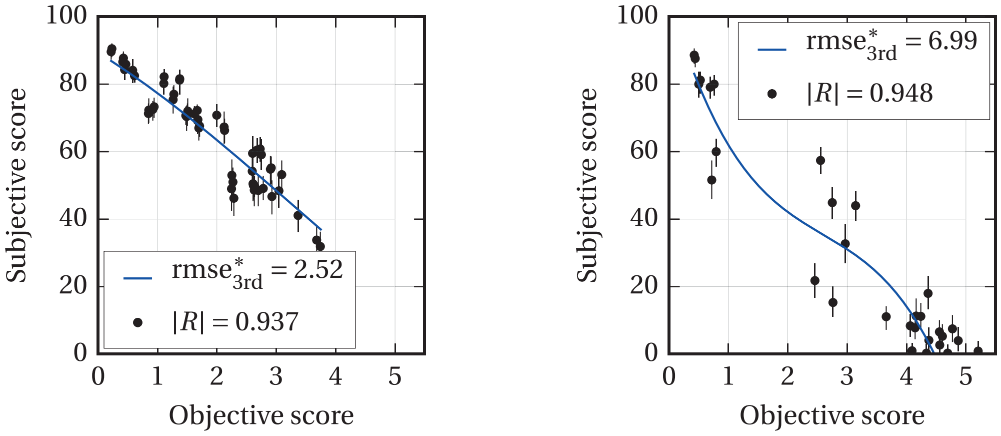


Figure 2.2 – Example scatter plots with corresponding performance metrics $|R|$ and rmse$^*_{3\text{rd}}$. Note the higher correlation coefficient $|R|$ in the right-hand plot, despite the wider spread of data points. The blue line shows the monotonic, third-order polynomial used to compute the metric rmse$^*_{3\text{rd}}$. Error bars indicate 95% confidence intervals of subjective scores.

- Data points at the scale ends have a stronger influence on correlation, making it easy to increase the metric by including extreme conditions (e.g., heavily distorted speech).

### 2.3.2 Modified prediction error

A new performance metric, *modified prediction error*, was introduced in ITU-T Rec. P.1401 [2012] to address some of the issues of the correlation coefficient. The metric is similar to the traditional root-mean-square error (rmse), but uses a monotonic, third-order polynomial mapping to compensate non-linearities between **s** and **o**. Additionally, prediction errors are reduced by the 95% confidence interval around each score $s_i$ to consider inter-subject variability. The resulting metric is denoted rmse$^*_{3\text{rd}}$ and defined as

$$\text{rmse}^*_{3\text{rd}} = \sqrt{\frac{1}{N-4}\sum_{i=1}^{N} \max\left(|s_i - o'_i| - \text{CI95}_i \,,\, 0\right)^2} \tag{2.2}$$

with $o'_i$ the objective score after polynomial mapping and CI95$_i$ the 95% confidence interval of the $i^{\text{th}}$ subjective score $s_i$, respectively [ITU-T Rec. P.1401, 2012]. The polynomial that minimizes (2.2) is computed separately for each dataset through constrained optimization.

As an illustrative example of both metrics, Figure 2.2 shows the results of an objective measure for two datasets. The (absolute) correlation $|R|$ is higher for the right-hand dataset because most data points are located at the scale ends, even though intermediate scores are not well predicted. By contrast, the prediction error rmse$^*_{3\text{rd}}$ reflects how data points are more tightly distributed around the mapping function in the left-hand dataset.

**Confidence interval estimation**

When $s_i$ describes a mean opinion score (MOS), such as perceived noise intrusiveness or speech quality, it can be expected that listeners' individual ratings are approximately normally distributed around the mean. In this case, the corresponding confidence interval $\mathrm{CI95}_i$ can be calculated from the standard deviation $\sigma_i$ of ratings $r$,

$$\mathrm{CI95}_i = t\,(0.05, S)\,\frac{\sigma_i}{\sqrt{S}} \tag{2.3}$$

$$\sigma_i = \sqrt{\frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\left(r_{i,k,l} - \overline{r_{i,k}}\right)^2}{S-1}} \tag{2.4}$$

with $K$ the number of recordings for condition $i$, $L$ the number of listeners, $S$ the number of ratings and $t\,(0.05, S)$ the $5^{\mathrm{th}}$ percentile of the $t$ distribution for $S$ degrees of freedom [ITU-T Rec. P.1401, 2012]. Note that $S = L \cdot K$ if all listeners evaluate all recordings.

With intelligibility scores, the above approach is usually not applicable. This is because individual intelligibility ratings are often not normally distributed, with a long one-sided tail of listeners having high error rates. In this case, confidence intervals can be estimated through the bootstrap [Efron and Tibshirani, 1986], a numerical method that consists in repeatedly sampling the available data, at random and with replacement, to estimate a statistic. In this thesis, we use the adjusted bootstrap percentile interval of means of resampled listener ratings, with 10 000 resampling steps, to estimate confidence intervals of mean intelligibility scores.

### 2.3.3 Statistical significance tests

We use two tests to determine whether differences in subjective scores between conditions, as well as in the prediction performance of objective measures are statistically significant.

Differences in subjective scores are evaluated with the Wilcoxon signed-rank test, a paired difference test that determines whether differences in ratings from the same listeners follow a distribution that is symmetric around zero. The test requires that listener ratings be measured on an ordinal scale (i.e., where scale items can be sorted from lowest to highest), but does not assume equal distance between scale items, nor a normal distribution of listener ratings. It is therefore particularly appropriate for *opinion* ratings, where the perceptual distance between adjacent scale items need not be constant (e.g., "excellent"–"good" vs. "good"–"fair").[2]

We evaluate differences between objective measures with the modified prediction error $\mathrm{rmse}^*_{\mathrm{3rd}}$ described in Section 2.3.2, using the F-test of equality of variances as specified in [ITU-T Rec. P.1401, 2012, Sec. 7.7]. Significant differences between correlation coefficients are not evaluated, due to the weaknesses of this performance metric as discussed above. In particular, the modified prediction error, instead of the correlation coefficient, was also used to select the

[2] Studies on the perceptual distance of opinion scale items have produced inconsistent results, see, e.g., Karaiskos et al. [2008, Sec. 7] (equal distance) vs. Möller [2000, Chap. 8.2] (unequal distance).

winning algorithms in the ITU-T P.OLQA competition [Rapporteur for Question 9/12, 2009].

Testing the same hypothesis in multiple comparisons leads to an increase in type I errors, i.e., the probability of finding a significant difference only by chance (false positive). A common approach to this problem is to perform Bonferroni correction, which involves a division of the significance level by the number of comparisons. We use the slightly less conservative Holm-Bonferroni correction, a stepwise procedure that has higher statistical power than Bonferroni's method [Holm, 1979; Abdi, 2010].

## 2.4 Datasets

We now introduce the different speech datasets that are used in our experiments:

- The *Perceptual Assessment of Noise DAtasets ("PANDA")* are a specific contribution of this thesis and address the perception of noise-corrupted speech in terms of the quality features speech distortion, noise intrusiveness, and overall quality. We will briefly introduce them here, with a more in-depth presentation in Chapter 3, page 23. An objective measure of background noise intrusiveness is presented in Chapter 4, page 41.

- The *PSCR audio library* deals with speech intelligibility in low bit-rate telecommunication systems under adverse acoustic conditions (i.e., high background noise levels and/or distortions due to wearing a breathing mask). Recordings and corresponding subjective scores in the datasets are based on the Modified Rhyme Test (MRT), and are made available by the Public Safety Communications Research (PSCR) program.[3] We evaluate objective intelligibility scores for this dataset in Chapter 5, page 61.

- The *Blizzard Challenge results* datasets contain speech synthesized with different text-to-speech (TTS) systems that were evaluated as part of the yearly Blizzard Challenge.[4] Speech recordings of semantically unpredictable sentences (SUS) and accompanying subjective intelligibility scores are publicly available for academic research. The objective assessment of synthetic speech intelligibility is addressed in Chapter 6, page 77.

### 2.4.1 Perceptual Assessment of Noise DAtasets ("PANDA")

We collected the "PANDA" datasets to investigate the perception of speech degraded by background noise in current telecommunication systems as used by the general public (i.e., commercial cellular and landline as well as Voice over IP systems). These systems generally provide highly intelligible speech transmission, such that listeners' attention turns to other quality features like speech naturalness, continuity, and the presence of noise [Möller and Heusdens, 2013, Sec. IV].

[3] http://www.pscr.gov/

[4] http://www.cstr.ed.ac.uk/projects/blizzard/data.html

We have focused on the quality features speech distortion, noise intrusiveness and overall quality, which are relevant to the assessment of noisy speech processed with noise reduction (as commonly applied in these telecommunication systems). The three quality features are defined in ITU-T Rec. P.835 [2003] and are designed to assess the trade-off between applying strong noise reduction (which may distort foreground speech) and keeping some residual background noise (which listeners may perceive as intrusive).

The collected speech data consists of recordings of short sentences from four speakers, with digitally added noise (ten noise types with SNRs between 3 and 40 dB) and further processing with different noise reduction and speech codec implementations as found in contemporary telecommunication systems. The total duration of test speech recordings is 97 minutes. A complete description of the speech material and collection of subjective scores, with subsequent analysis and discussion, is given in Chapter 3, page 23.

### 2.4.2 Public Safety Communications Research (PSCR) audio library

The Public Safety Communications Research [PSCR, 2013] audio library is a publicly available collection of speech recordings degraded by strong noises, acoustical impairments and different speech coding schemes, with corresponding subjective intelligibility scores. The library originated from an effort to select appropriate communication systems for U.S. fire agencies, and is organized in three datasets collected in 2008, 2010 and 2012, respectively.

The recordings and subjective test design follow the Modified Rhyme Test (MRT) [House et al., 1965; ANSI S3.2, 2009], where listeners are asked to recognize a word embedded in a carrier phrase. Listeners select the heard word from a choice list of six words differing either in the initial or final consonant. An example is the phrase “Please select the word *sent*.” with choice list *went* — *sent* — *bent* — *dent* — *tent* — *rent*. Intelligibility is computed as the percentage of correctly selected words (after correcting for guessing) for a given condition. Each dataset features 50 choice lists of 6 rhyme words pronounced by 4 speakers, totaling $6 \times 50 \times 4 = 1200$ reference recordings. The complete set of rhyme words is listed in Table A.5, page 108.

Conditions in the dataset focus on scenarios in tactical fire scene communications, including strong background noises, acoustical distortions from a breathing mask, and low bit-rate coding schemes over clean and lossy channels. Due to the presence of very strong degradations, conditions cover intelligibility values between 90% and 0% (after correcting for guessing). Note that all recordings are limited to the narrow audio bandwidth (i.e., below 4000 Hz). Table 2.1 provides an overview of conditions in the three datasets.

Complete descriptions of dataset designs, conditions and subjective scoring are available in [Atkinson et al., 2008; 2013[5]; 2012]. The total size of the datasets is $(24+54+30$ conditions$)$ $\times 1200$ phrases = 129 600 test speech recordings, corresponding to 65.1 hours (entire phrases) or 17.3 hours (rhyme words only) of speech.

[5] The report for the 2010 test was published in 2013.

Table 2.1 – Overview of conditions in the three PSCR datasets. See [Atkinson et al., 2008; 2013; 2012] for a complete description of conditions and processing steps.

| Condition | Included in dataset | | |
|---|---|---|---|
| | 2012 | 2010 | 2008 |
| *Background noises* | | | |
| Clean (no noise) | • | • | • |
| Alarm 1 (−2 dB SNR) | • | • | • |
| Alarm 2 (−2 dB SNR) | | • | |
| Pump hum (4 dB SNR) | | | • |
| Rotary saw cutting metal (4 dB SNR) | | | • |
| Chainsaw cutting wood (5 dB SNR) | • | | • |
| “Club” noise (music+speech; 5 dB SNR) | • | • | |
| Water hose (9 dB SNR) | | | • |
| Low air alarm inside mask (15 dB SNR) | | | • |
| *Acoustical impairments* | | | |
| Transparent (no mask) | • | • | • |
| Breathing mask, speech through diaphragm | • | • | • |
| Breathing mask, microphone inside mask | • | • | |
| *Speech coding schemes* | | | |
| AMR codec [3GPP TS 26.090], 12.2 kbps | • | | |
| AMR codec [3GPP TS 26.090], 7.4 kbps | • | | |
| P25 AMBE+2 codec [TIA-102.BABA], 4.4 kbps | • | • | • |
| P25 AMBE+2 codec [TIA-102.BABA], 2.45 kbps | | • | |
| P25 IMBE codec [TIA-102.BABA], 4.4 kbps | | | • |
| 25 kHz analog FM radio | • | • | • |
| 12.5 kHz analog FM radio | | • | • |
| *Simulated radio channels* | | | |
| Transparent (no loss) | • | • | • |
| Simulated bit errors or analog FM noise | | • | |
| Total number of conditions | 24 | 54 | 30 |
| Number of listeners | 10 | 52 | 30 |

The PSCR audio library distribution includes individual intelligibility ratings from the number of listeners listed at the bottom of Table 2.1. From these, we computed the average word accuracy (WA) with 95% confidence intervals for each condition. Since the distribution of individual intelligibility ratings tends to be non-normal, we have used a bootstrap procedure as described in Section 2.3.2 to estimate confidence intervals.

### 2.4.3 Blizzard Challenge results

The Blizzard Challenge was devised by Black and Tokuda [2005] as a yearly evaluation of speech synthesis approaches that are trained on common data. The evaluation consists in gathering subjective scores for speech naturalness, similarity to a target speaker, and intelligibility. Participation to the Challenge is open to researchers in both academia and industry, with the results being published using anonymized identifiers to protect the reputation and commercial interest of poorly ranked participants.

We use results from the 2010 and 2011 challenges [King and Karaiskos, 2010; 2011] to evaluate the objective assessment of synthetic speech intelligibility in English. We selected these two Challenge editions because they also include recordings from a professional voice talent (male and female speaker in 2010 and 2011, respectively) as a highly intelligible reference. We use results for task "EH1" (clean synthetic speech) from both editions, as well as results for the noisy speech subtask "ES2" in the 2010 Challenge, where participants were requested to build a synthetic voice capable of maintaining intelligibility in high levels of speech-shaped noise [Dreschler et al., 2001]. For both tasks, intelligibility is evaluated by means of semantically unpredictable sentences (SUS), which the organizers found to best differentiate between the intelligibility of different voices [King and Karaiskos, 2011].

Table 2.2 gives an overview of the speech material and subjective scores in the three datasets. Contrary to rhyme tests, intelligibility assessment with SUS requires that no listener hear the same sentence twice to avoid training effects. The Challenge organizers thus used a so-called "Latin square" design, where listeners are presented with different subsets of the data that combine to a complete evaluation across listeners. Moreover, some listeners did not complete the entire test, resulting in a total number of answers slightly below the number of possible listener-stimulus combinations.

An example of a typical SUS is "The glass poured the date that cared." [King and Karaiskos, 2010], with sentence lengths varying between six and eight words. During the test, listeners are requested to type in the sentence as they have understood it from the given test speech recording. The intelligibility of a recording of sentence $k$, produced by voice $i$, is determined by the Word Error Rate (WER),

$$\mathrm{WER}_{i,k,l} = \frac{w^s_{i,k,l} + w^i_{i,k,l} + w^d_{i,k,l}}{M_k} \times 100\% \ , \tag{2.5}$$

Table 2.2 – Overview of speech material and subjective scores for synthetic speech intelligibility with semantically unpredictable sentences (SUS) in the three used Blizzard datasets. “2010c” and “2010n” refer to clean and noisy speech data, respectively, in the 2010 Blizzard Challenge. See [King and Karaiskos, 2010; 2011] for complete dataset descriptions.

| Property | Dataset | | |
|---|---|---|---|
| | 2011 | 2010c | 2010n |
| *Recordings* | | | |
| Task identifier in [King and Karaiskos, 2010; 2011] | EH1 | EH1 | ES2 |
| Added speech-shaped noise, in dB SNR | (clean) | (clean) | 0, −5, −10 |
| Number of voices (synthetic+natural) | 12+1 | 17+1 | 12+1 |
| Number of different SUS in the subjective test | 26 | 18 | 39 |
| Total duration of evaluated SUS recordings, in minutes | 12.2 | 12.6 | 58.4 |
| *Subjective scores* | | | |
| Number of listeners | 231 | 177 | 388 |
| Number of SUS evaluated per voice, per listener | 2 | 1 | 3 |
| Total number of answers for SUS recordings | 5706 | 2857 | 14 166 |

with $w^s_{i,k,l}$, $w^i_{i,k,l}$ and $w^d_{i,k,l}$ the number of words substituted, inserted and deleted, respectively, in the transcription by listener $l$, and $M_k$ the ground truth number of words in sentence $k$. The Challenge organizers took care of correcting spelling mistakes in the answers of listeners before computing WERs. The intelligibility of each voice $i$ is then given by the average of WERs across sentences and listeners,

$$\mathrm{WER}_i = \frac{\sum_{l=1}^{L}\sum_{k=1}^{K}\mathrm{WER}_{i,k,l}}{L\cdot K}\ , \tag{2.6}$$

with $L$ and $K$ the number of listeners and sentences, respectively.

Note that there are two possible issues with the above calculation of intelligibility. First, the averaging of per-sentence scores in (2.6) ignores the fact that sentences have different number of words. However, an evaluation on WER scores for Mandarin in [Karaiskos et al., 2008] showed the impact of these differences to be negligible. Second, the definition of per-sentence WER in (2.5) means that WER values > 100% are possible if many incorrect words are entered, which may occur if a listener attempts to “guess” words in a highly unintelligible recording. In this case, the WER depends on listener behavior, since another listener in the same situation may decide not to enter any words, yielding a WER of 100% instead. While we did find evidence of such discrepancies in the noisy speech data, recalculating the per-voice WER after bounding listener WERs to $[0,100]\%$ resulted in a maximum absolute difference of only 0.67%. We thus decided to use the original data as provided by the Blizzard Challenge organizers.

For the purpose of our experiments, we convert per-voice WERs to Word Accuracy (WA) scores, defined as $\text{WA}_i = 100\% - \text{WER}_i$. As with the PSCR data, we determine 95% confidence intervals of WA scores per voice using the bootstrap procedure described in Section 2.3.2.

## 2.5 Features

A speech recording carries information at different levels, e.g., *what* was said, *how*, *by whom* and *in what environment*. In order to make the signal more amenable to analysis at a particular level, it is usual to extract features that only retain the information of interest. The features presented in this section, and the artificial neural networks and hidden Markov models introduced in Section 2.6 and 2.7, respectively, are standard tools in automatic speech recognition (ASR). They are used in this thesis to analyze the phonetic content of speech, with the goal of assessing its intelligibility, as described in Chapter 5 and 6, respectively.

### 2.5.1 Cepstral features

Cepstral features are obtained by taking the Fourier transform of the log-magnitude short-time spectrum of a signal, and are often used as acoustic features from which the phonetic content of speech is estimated in a later step. Two popular cepstral-based features in speech processing are mel-frequency cepstral coefficients (MFCC) [Davis and Mermelstein, 1980] and PLP (perceptual linear predictive) cepstral coefficients [Hermansky, 1990]. Both features are based on a spectral analysis with non-linear frequency resolution modeled after human hearing, and mainly differ in the applied amplitude compression and smoothing of the short-time spectrum.

In this thesis, we use PLP cepstral coefficients, where the spectral analysis involves a frequency-dependent weighting and cubic-root compression of short-time spectral power that approximate the frequency-dependent sensitivity and loudness perception, respectively, of human hearing. This auditory-like short-time spectrum is then smoothed through approximation with an all-pole model in order to preserve the overall spectrum shape, but remove fine structure that is thought to be speaker dependent [Hermansky, 1990]. This operation has been shown to yield greater noise robustness and speaker independence than smoothing through cepstral truncation of MFCCs [Gold et al., 2011, Chap. 22.2].

We use cepstral coefficients (including the zeroth coefficient, i.e., energy) of the short-time PLP spectrum with model order 12. The coefficients are computed with the HTK software [Young et al., 2006], which uses a mel filterbank for PLP analysis instead of the Bark frequency scale originally proposed by Hermansky [1990]. As is commonly done, we normalize the extracted cepstral coefficients by subtracting their means across the length of the analyzed signal, which compensates possible convolutional distortions from the telecommunication channel.

### 2.5.2 Modulation-based features

Cepstral features of mel-cepstral or PLP short-time analyses are often combined with their derivatives over adjacent frames to include information about the temporal dynamics of the speech signal. These so-called delta and acceleration coefficients represent the first and second temporal derivatives of cepstral coefficients, respectively, and have been shown to improve the performance of speaker-independent speech recognition [Furui, 1986].

Extending on this concept, further methods have been proposed that compute cepstral features from a subset of the temporal modulations contained in speech [Hermansky et al., 1991], or directly use temporal modulations as features themselves [Hermansky and Fousek, 2005]. As a second type of acoustic feature besides PLP, we evaluate the latter method, i.e., so-called MRASTA (Multi-resolution RelAtive SpecTrAl) filters, which capture the temporal trajectories of band energies [Hermansky and Fousek, 2005]. Specifically, the features are obtained by filtering band envelopes with first and second derivatives of Gaussian windows with six different widths between $\sigma = 8$ and 60 ms. Since these filters have zero mean, the features are inherently insensitive to convolutional channel distortions.

A perceptual motivation for MRASTA features is the importance of temporal modulations in speech for intelligibility [Drullman et al., 1994; Greenberg et al., 1998]. We evaluate both PLP-cepstral and MRASTA features for posterior feature estimation in Section 5.5.3, page 73.

### 2.5.3 Posterior features

The posterior probability $P(E \mid O)$ describes the probability of an unobserved or *latent* event $E$ given an observation $O$. In speech processing, we use the term "posterior feature" or *posterior* to refer to the discrete probability vector $\mathbf{z} = \left[P\left(c^1 \mid \mathbf{b}\right), \ldots, P\left(c^k \mid \mathbf{b}\right), \ldots, P\left(c^K \mid \mathbf{b}\right)\right]^\top$ of a set of latent symbols $c^1, \ldots, c^k, \ldots, c^K$, given an observed feature vector $\mathbf{b}$. Posteriors are probability distributions, thus we have $\sum_k P\left(c^k \mid \mathbf{b}\right) = 1$.

The type of posterior features used in this thesis are *phoneme posteriors*, i.e., the probability of $K$ phoneme classes $c^1, \ldots, c^k, \ldots, c^K$, given an acoustic feature vector $\mathbf{b}$ (e.g., PLP-cepstral or MRASTA features) that is extracted from the speech signal. We use artificial neural networks (ANN), introduced in the next section, to estimate these probability distributions.

## 2.6 Artificial neural networks (ANN)

An artificial neural network (ANN) is an adaptive model that can be trained to learn a specific mapping from input to output. The mapping is performed by means of inter-connected nodes organized in layers. An ANN comprises an input and output layer, and possibly one or more intermediate layer(s) called *hidden* layer(s). In a feed-forward ANN, the value of each node is a transformed, linear combination of the values of the nodes in the preceding layer,

$$x^k_{(j)} = w^k_{0(j)} + \sum_{i=1}^{I} w^k_{i(j)} \cdot y^i_{(j-1)} \tag{2.7}$$

$$y^k_{(j)} = f\left(x^k_{(j)}\right) \tag{2.8}$$

with $y^k_{(j)}$ the value of node $k$ in layer $(j)$, $w_i$ and $w_0$ the weights and bias for the node values $y^1, \ldots, y^i, \ldots, y^I$ in the preceding layer $(j-1)$, respectively, and $f(\cdot)$ a so-called *activation function* [Bishop, 2006, Chap. 5.1]. Note the change of node index from $i$ to $k$ between layers in (2.7), reflecting the fact that layers need not have the same number of nodes.

A feed-forward ANN with at least one hidden layer that uses a differentiable, nonlinear activation function $f(\cdot)$ is known as multilayer perceptron (MLP). MLPs can be trained to learn *any* nonlinear input-output mapping, if the number of nodes in the hidden layer is large enough [Bourlard and Morgan, 1994, Chap. 4]. The activation function in the hidden layer(s) is usually a sigmoid that maps values to the range $(0,1)$,

$$f_h\left(x^k\right) = \frac{1}{1+\exp\left(-x^k\right)} \; . \tag{2.9}$$

For MLPs that are trained to perform classification, as used in this thesis, the activation function in the output layer is typically the softmax

$$f_o\left(x^k\right) = \frac{\exp\left(x^k\right)}{\sum_{l=1}^{K} \exp\left(x^l\right)} \; , \tag{2.10}$$

ensuring that all $K$ values of the output layer sum to one. It can be shown that the outputs of an MLP are estimates of posterior probabilities, conditioned on the values of the input layer [Richard and Lippmann, 1991; Bourlard and Morgan, 1994, Chap. 6].

To train an MLP, successive input vectors are presented, and a cost function is evaluated based on the resulting output values. The weight and bias terms $w$ of the network are then adjusted to minimize the cost, using the error backpropagation algorithm [Rumelhart et al., 1986; Bishop, 2006, Chap. 5.3]. In this thesis, we use MLPs to estimate phoneme posterior probabilities $\mathbf{z}$, given an acoustic feature vector $\mathbf{b}$ at the input. We use the QuickNet toolkit [Johnson and contributors, 2011] for MLP training, with utterance-level feature normalization at the input, and the frame-level cross-entropy between MLP outputs and target phoneme labels as the cost function. An early stopping criterion with a cross-validation dataset is used to avoid over-fitting the network on the training data [Morgan and Bourlard, 1989].

## 2.7 Hidden Markov models (HMMs) and KL-based HMMs

Hidden Markov models (HMMs) are statistical generative models that are used in speech recognition and other fields to find the succession of events or *states* that best explains an

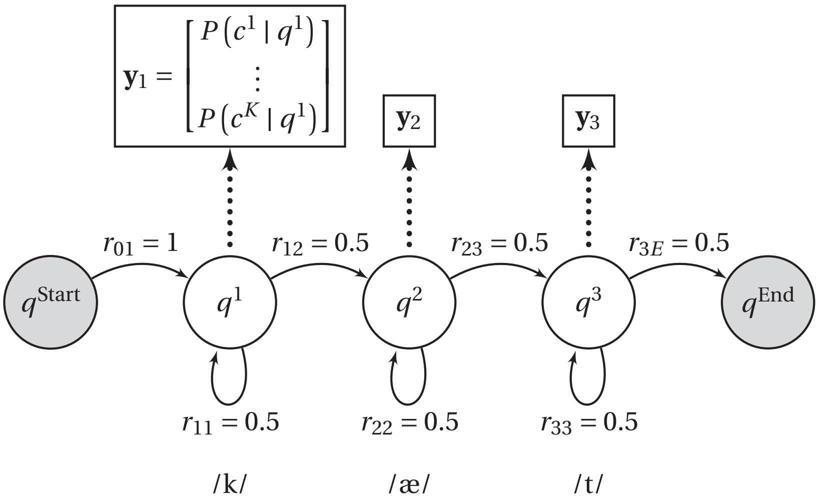


Figure 2.3 – Example of a KL-HMM with three states for the phonemes /k/, /æ/ and /t/. Each state $q^i$ is parameterized by a posterior probability distribution $\mathbf{y}_i$.

observation sequence [see, e.g., Rabiner, 1989]. The model consists of a system that is in one of $I$ possible states $Q = \{q^1, \dots, q^i, \dots, q^I\}$, and produces an observation according to a (known) stochastic process associated with the state. At each time step, the system may either stay in the same state, or transit to a different state. The succession of states is not directly observed (hence the name "hidden"), but assumed to be governed by a second stochastic process, where the probability of transiting to a new state $q^j$ at time $t+1$ only depends on the current state (known as *Markov assumption*),

$$r_{ij} = P\left(q_{t+1} = q^j \mid q_t = q^i, q_{t-1} = q^k, \dots\right) = P\left(q_{t+1} = q^j \mid q_t = q^i\right) , \tag{2.11}$$

with $q_t$ the state at time $t$, $r_{ij}$ the transition probability from state $q^i$ to $q^j$, and $i, j, k \in \{1, \dots, I\}$. Given the specifications for a HMM and a sequence of observed output values, the most probable sequence of states that explains the observations can be determined with the Viterbi algorithm [Viterbi, 1967; Rabiner, 1989].

In this thesis, we use HMMs as acoustic model, i.e., to find the sequence of subword units (e.g., phonemes or phones) that best explains a sequence of observed phoneme posterior probability distributions. The HMM state sequence thus provides a segmentation or labeling of frame-level posteriors. Specifically, we use a type of HMM known as *Kullback-Leibler divergence-based HMM* or KL-HMM [Aradilla et al., 2007], where each state $q^i$ is parameterized by a posterior probability distribution $\mathbf{y}_i = \left[P(c^1 \mid q^i), \dots, P(c^k \mid q^i), \dots, P(c^K \mid q^i)\right]^\top$. Note that the KL-HMM states can represent different subword units (e.g., phones, or context-dependent phonemes) than the ones in the associated distributions $\mathbf{y}_i$.

Figure 2.3 shows an example KL-HMM with three states for the phonemes /k/, /æ/ and /t/, respectively, arranged in a left-to-right topology as usual in speech processing. Given an observation posterior sequence $Z = \{\mathbf{z}_1, \dots, \mathbf{z}_j, \dots, \mathbf{z}_J\}$, the local score for each HMM state is computed with the reverse KL divergence,

$$\mathrm{RKL}(\mathbf{y}_i, \mathbf{z}_j) = \sum_{k=1}^{K} z_j^k \log\left(\frac{z_j^k}{y_i^k}\right) \tag{2.12}$$

as a measure of distance between probability distributions [Cover and Thomas, 1991; Rasipuram, 2014, Chap. 4]. The HMM state sequence $q_1, \ldots, q_j, \ldots, q_J$ that minimizes the global score

$$\min_{\mathcal{Q}(u)} \left( \sum_{j=1}^{J} \mathrm{RKL}(\mathbf{y}_{q_j}, \mathbf{z}_j) - \log\left(r_{q_{j-1}q_j}\right) \right) , \tag{2.13}$$

where $\mathcal{Q}(u)$ denotes the set of all possible state sequences of length $J$ for the utterance $u$ modeled with the KL-HMM, can then be computed with the Viterbi algorithm.

As shown in Figure 2.3, the state transition probabilities $r$ can be set to a constant value, thus KL-HMM training only involves learning the distributions $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_I\}$ associated with the set of states $Q = \{q^1, \ldots, q^i, \ldots, q^I\}$. Training is performed through a Viterbi-EM (expectation maximization) approach, using speech data, phonetic transcriptions from a pronunciation dictionary, and posterior probabilities $\mathbf{z}_j$ estimated with an MLP. Specifically, the E-step consists in segmenting the data with the KL-HMM using the Viterbi algorithm, with a uniform segmentation derived from the phonetic transcription in the initial iteration. During the M-step, the state distributions $\mathbf{y}_i$ for each subword unit $i$ are updated from the posteriors $\mathbf{z}_j$ assigned to the subword. When using $\mathrm{RKL}(\mathbf{y}_i, \mathbf{z}_j)$ as the local score, the optimal update is given by the arithmetic mean of the $N$ posteriors assigned to the subword,

$$y_i^k = \frac{1}{N} \sum_{j \in N} z_j^k \quad \forall \quad k \in \{1, \ldots, K\} , \tag{2.14}$$

with the proof in [Aradilla et al., 2007].

KL-HMMs have been shown to be powerful acoustic models that can be trained on very small amounts (e.g., five minutes) of speech data [Imseng, 2013, Chap. 4; Rasipuram, 2014, Chap. 6]. A state tying approach is used to share data between states for subword units that appear infrequently or not at all in the training data [Imseng, 2013, Chap. 4.6].

## 2.8 Summary

This chapter laid the background for our contributions in the remainder of this thesis. We have given a short introduction to speech quality assessment, explained statistical evaluation methods, and described the datasets used for our experiments. We then presented some state-of-the-art feature extraction and machine learning methods that are exploited in our proposed approaches. The next chapter presents our first contribution, where we investigate the perception of background noise in speech telecommunications.

# 3 Perception of background noise in speech telecommunications

Recently, mobile telephony services have undergone a transition to newer codecs that transmit sound in larger, so-called wideband (WB, 50–7000 Hz), or super-wideband (SWB, 50–14 000 Hz) frequency ranges. The extended bandwidth is claimed to deliver clearer sounding and more life-like speech to end users, dubbed “HD voice” by network operators [Deutsche Telekom, 2011; Swisscom, 2012; Orange, 2013]. Since capturing wideband audio may also include more background noises from the caller’s environment, the industry has put a particular emphasis on noise reduction processing in “HD voice” compatible devices [GSM Association, 2013].

Noise reduction (NR) aims to attenuate background noise without distorting foreground speech, but such perfect separation is hard to achieve. In order to evaluate the effect of NR on perceived quality, the standard subjective test method [ITU-T Rec. P.835, 2003] consists in asking listeners to rate the *foreground speech distortion*, *background noise intrusiveness*, and *overall listening quality* of processed speech recordings. To investigate how noisy speech affects these quality features in wideband telecommunication systems, and with the goal of developing a measure for their objective assessment, we have collected three datasets of speech recordings. A particular focus was put on the following research questions:

1. How do signal bandwidth and bandwidth context (i.e., awareness of band limitations in a subjective test) affect listeners’ quality ratings?
2. Does perceived noise intrusiveness change in the presence of Lombard (i.e., effortful) speech, since it represents a more realistic example of speech in noise?
3. How does presentation level affect the three quality features?

To the best of our knowledge, these questions have not been addressed in previous experiments. In the remainder of this chapter, we first review known effects of noisy speech on quality features in Section 3.1. The collection of test speech recordings and subjective scores for our datasets are described in Section 3.2 and 3.3, respectively. We analyze how different factors affect quality features in Section 3.4, and their mutual dependency in Section 3.5.

Finally, we evaluate in Section 3.6 whether overall quality scores can be predicted with an existing objective measure, and discuss our main findings in Section 3.7.

## 3.1 Related work and contributions

The perceptual impact of noise on speech can be expressed in terms of the P.835 quality features mentioned above, as well as in terms of intelligibility. Each of these quality features depends on properties of the input signal and on further processing that a telecommunication system may apply to it. However, the improvement of one quality feature does not necessarily go hand in hand with the improvement of others. In the following, we briefly review some relevant factors that influence the perception of these quality features.

*Signal bandwidth and bandwidth context* — The wider signal bandwidth afforded by recent telecommunication systems has been shown to increase the perceived absolute quality of clean speech [Wältermann et al., 2010]. However, the wider bandwidth also defines a new context that causes listeners to adjust their expectations. As a result, a call that is reverted to the narrow band after starting in a wideband context will be perceived as degraded by listeners, since they have been made aware of the band limitation [Lewcio and Möller, 2014].

Whereas the long-term average spectrum of *speech* remains quite similar between speakers [Byrne et al., 1994], *noises* can strongly differ in spectral distribution, so their impact on wideband signal capture can be highly variable. Generally, the roughly $1/f$ power distribution of environmental noises [De Coensel et al., 2003] means that the additional noise power tends to be greater in the expanded low- than in the high-frequency range.

Despite its higher quality, wideband speech provides little benefit for intelligibility [Fernández Gallardo and Möller, 2015], since the traditional telephone band already covers most of the frequencies needed for intelligible speech [Allen, 2005, Chap. 4]. However, additional cues at high frequencies may help listeners compensate for speech losses in lower bands [Lippmann, 1996].

*Lombard speech* — The Lombard effect is an adaptation that speakers perform in the presence of noise. It is characterized by an increase in speech level and fundamental frequency, flatter spectral tilt and increased vowel duration [Junqua, 1996]. Several studies have shown that in the presence of noise and at equal SNR, Lombard speech is generally more intelligible than regular (conversational) speech [Van Summers et al., 1988; Junqua, 1993; Lu and Cooke, 2008], although very high vocal effort (i.e., shouted speech) reduces intelligibility [Pickett, 1956].

The impact of these speech adaptations on perceived speech quality and noise intrusiveness has been less considered so far, with noisy conditions in subjective tests consisting of regular speech with digitally added noise instead. This is presumably due to the overhead associated with collecting clean Lombard speech, and the difficulty in controlling the SNR in combined Lombard speech + noise conditions.

*Presentation level* — The relation between presentation level and speech quality was evaluated during the collection of training data for the ITU-T P.OLQA measure. For clean speech, it was found that deviations from the standard average level of 79 dB SPL by −10 and −20 dB result in a reduction of about 1 and 2 MOS, respectively, on an overall quality scale [Malfait, Berger, and Ullmann, 2009].

For noise-corrupted speech, it is known that at equal SNR, intelligibility *decreases* at high presentation levels [Pollack and Pickett, 1958]. At least part of this effect seems to be due to the wider spread of masking as sound levels increase [Fastl and Zwicker, 2007, Chap. 4.1], resulting in a reduced effective signal-to-noise ratio [Dubno et al., 2005].

*Noise reduction* — A key challenge in noise reduction (NR) is to estimate the respective contributions of speech and noise in the input signal. When accurate (i.e., oracle) estimates are available, it is possible to improve the intelligibility of speech by attenuating time-frequency regions of the signal that are dominated by noise [Brungart et al., 2006]. Such improvements are not observed in real-life scenarios with single-track (i.e., mono) speech + noise mixtures, due to under- or over-estimations of noise with state-of-the-art NR algorithms [Hu and Loizou, 2007a; Brons et al., 2012].

Depending on the type of estimation errors, NR processing results in different trade-offs between speech distortion and noise intrusiveness [Brons et al., 2012]. Furthermore, NR algorithms that are most beneficial to overall speech quality are not the same as those that are best for speech intelligibility [Hu and Loizou, 2007a], highlighting again the differences between quality features.

In summary, these studies show that a single quality feature is not sufficient to evaluate and optimize the performance of telecommunication systems for noisy speech. We now present the design, subjective rating and analysis of our datasets, referred to in this thesis as “PANDA” (Perceptual Assessment of Noise DAtasets). These datasets focus on the three quality features defined in ITU-T Rec. P.835 (*speech distortion*, *noise intrusiveness*, and *overall quality*), with conditions from a variety of telecommunication systems. The collected data and insights from its evaluation are the basis for developing an objective measure of background noise intrusiveness in Chapter 4, page 41.

Contributions

- R. Ullmann, H. Bourlard, J. Berger, and A. Llagostera Casanovas [2013]. “Noise Intrusiveness Factors in Speech Telecommunications”. In: *AIA-DAGA Jt. Conf. Acoust.* Merano, Italy, pp. 436–439.
  URL: http://publications.idiap.ch/index.php/publications/show/2619

- R. Ullmann, J. Berger, and A. Llagostera Casanovas [2013]. *Contribution 81 — Derivation of speech degradation scores (S-MOS) from subjective noise intrusiveness and overall quality scores (N- and G-MOS)*. Study Group 12, International Telecommunication Union, Geneva, Switzerland.
  URL: http://idiap.ch/~rullmann/ITU-T_SG12_Q9_Contribution81.pdf
- J. Berger and R. Ullmann [2013]. *Contribution 24 — ITU-T Rec. P.863 as Predictor for P.835 G-MOS in Super-Wideband and Narrowband Experiments*. Study Group 12, International Telecommunication Union, Geneva, Switzerland.
  URL: http://idiap.ch/~rullmann/ITU-T_SG12_Q9_Contribution24.pdf

The latter two documents are ITU-T contributions: they are submitted as input to discussions at Study Group meetings, but are not subject to prior peer review.

## 3.2 Dataset design

### 3.2.1 Reference speech recordings

We recorded 30 short (1.6–2.9 s) sentences in French from eight native speakers (four male). The sentences reflect everyday content as in a phone call, and are organized in groups of three consecutive sentences that are loosely related to the background noises used in the datasets. Recordings were carried out in two sessions, both of which took place in the psychoacoustic chamber of the Laboratory of Electromagnetics and Acoustics (LEMA) at EPFL.

In the first session, speakers pronounced all sentences in a quiet environment and at a normal speech level. In the second session, speakers wore a pair of closed headphones playing background noises at different levels in order to trigger the Lombard effect. The use of closed headphones allows to record clean Lombard speech, but also prevents speakers from hearing their own voice. Therefore, the microphone signal was fed back to the headphones, with speakers adjusting the level beforehand such that it felt equivalent to not wearing headphones. This recording setup is shown schematically in Figure 3.1.

The Lombard effect results in several speech modifications (as discussed in Section 3.1) that also depend on the noise type [Junqua, 1996]. Therefore, we kept the same correspondence of Lombard speech recordings and background noises in the subsequent generation of test speech signals.[1] The list of sentences and matching noise types is given in Section A.1.1 of the appendix, with further technical details on the recording setup in Section A.1.2, page 100.

[1] Of course, the Lombard effect also depends on the noise *level*. However, noise levels in test speech recordings will inevitably differ, due to conditions with noise reduction and other speech activity-dependent processing.

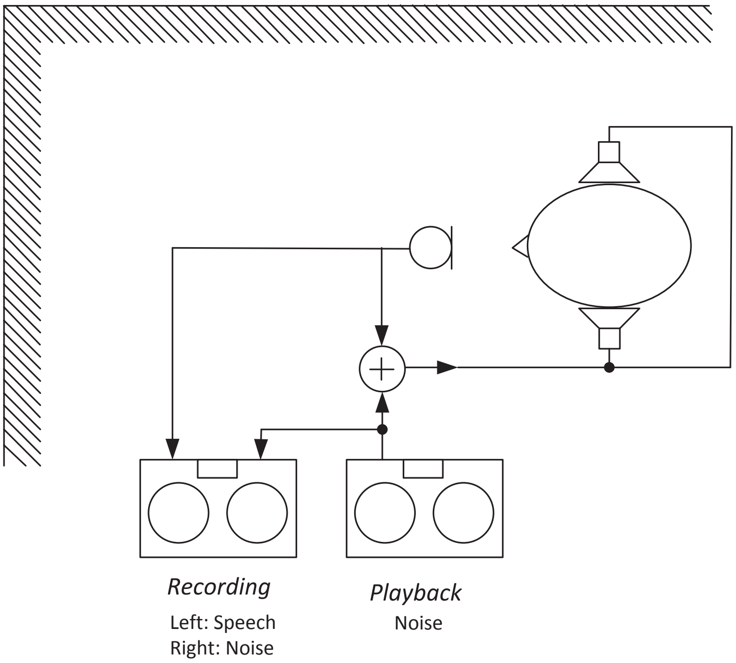


Figure 3.1 – Recording setup for the collection of Lombard speech.

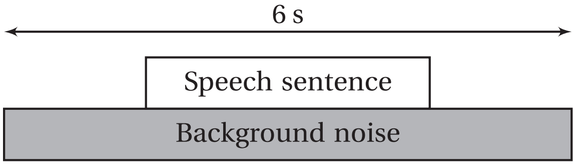


Figure 3.2 – Temporal structure of test speech recordings in the "PANDA" datasets.

### 3.2.2 Generation of test speech recordings

We selected recordings from four speakers (two male) for inclusion in the subjective tests. These recordings were overlaid with six seconds of additive background noise from a collection used for evaluating telecommunication systems [ETSI EG 202 396-1, 2011], resulting in the temporal structure *leading noise — speech+noise — trailing noise*, as shown in Figure 3.2. Consistent with the focus of this thesis, recordings were further processed with speech codecs, noise reduction, and other components as found in telecommunication systems at the time of dataset creation (2012–13). Two types of processing were used:

- *Simulated conditions*, where recordings are processed with software implementations of noise reduction algorithms and/or speech codecs, and
- *Live conditions*, where recordings are transmitted with telephone handsets over live networks and recorded at the receiving handset.[2] Live conditions represent the complete signal processing chain, including distortions from the network.

[2] Live conditions were processed by the industrial partner SwissQual AG, a Rohde & Schwarz company.

The proportion of simulated and live conditions is about 60% and 40%, respectively. In both cases, we processed looped versions of input signals to ensure that any adaptive processing in our conditions (e.g., noise estimation in NR processing) could converge to a stable point. The resulting recordings were partitioned into three sets, named Set 1–3, that could be scored in about 75 minutes each.

Table 3.1 provides an overview of noises and processing conditions in each set. It can be seen that Sets 1 and 2 cover conditions with bandwidths up to SWB, while Set 3 features NB signals only, allowing to compare different bandwidth contexts between subjective tests. Each condition consists of 3 sentences × 4 speakers, i.e., 12 recordings. Some conditions are shared across sets for the purpose of later comparisons, resulting in a total of 81 unique conditions, or 972 test speech recordings (97 minutes).

Moreover, care was taken to ensure balanced test designs by spanning the full quality range of both speech distortion and noise intrusiveness. For this reason, about 30% of conditions in each set are noise free; these recordings contain regular speech. The remaining ~ 70% of conditions are noisy and contain Lombard speech, with the exception of Set 1, where most conditions appear in pairs (once with regular and once with Lombard speech) to study the effect of speech type on subjective scores.

Further details on the processing steps for each condition are given in the test plans listed in the appendix in Section A.1.3, page 101.

## 3.3 Collection of subjective scores

We conducted a subjective test for each set, following the guidelines in ITU-T Rec. P.835 [2003], where listeners provide ratings for foreground speech degradation, background noise intrusiveness and overall quality. These ratings are collected by presenting a triplet of recordings with different sentences, but identical speaker and condition. Listeners rate a different quality feature for each recording in the triplet, as shown in Figure 3.3. Before each recording, listeners are instructed to focus *only on the speech*, *only on the noise*, or on both signal components (for overall quality), respectively.

We invited 90 listeners to participate in the tests (30 per set), which took place — once more — in the psychoacoustic chamber of LEMA. Listeners were native speakers of French, passed a short hearing test and were paid for their participation. Recordings were presented diotically (same signal on both ears) over Grado SR 60 headphones in a random order of speaker and condition, using a software tool from the industrial partner. Screenshots of the rating interface with instructions in French are shown in the appendix in Figure A.1, page 110. After a short training session to familiarize themselves with their task, listeners provided ratings using the scales shown in Table 3.2. The duration of each subjective test was about 75 min. Results from nine listeners who failed to complete the task or to pass the hearing test were removed, leaving ratings from 26, 28 and 27 subjects for the three sets, respectively.

Table 3.1 – Overview of conditions in the “PANDA” datasets.

| Speech processing | Included in | | |
|---|---|---|---|
| | Set 1 | Set 2 | Set 3 |
| *Input filter (for all conditions)* | | | |
| 195–3700 Hz (narrowband) | • | • | • |
| 50–7000 Hz (wideband) | • | • | |
| 50–14 000 Hz (super-wideband) | • | • | |
| *Background noises [ETSI EG 202 396-1]* | | | |
| Jackhammer (distant; 3 dB SNR) | • | • | • |
| Pub (unintelligible babble; 4 dB SNR) | • | • | • |
| Road (nearby traffic; 5 dB SNR) | | • | • |
| Car (inside, cruising at 80 km/h; 8 dB SNR) | • | • | |
| Train station (engine, announcement; 8 dB SNR) | • | | • |
| Train (inside; 12 dB SNR) | • | • | • |
| Office (keyboards, speech fragments; 15 dB SNR) | • | • | • |
| Schoolyard (children cheering; 16 dB SNR) | | | • |
| Cafeteria (cutlery sounds, laughter; 17 dB SNR) | • | • | • |
| Crossroad (distant traffic; 10, 20, 30, 40 dB SNR) | • | • | • |
| *Simulated conditions* | | | |
| No further processing (anchor conditions) | • | • | • |
| NR [Adami et al., 2002], spectral subtractive | • | • | • |
| NR [Adami et al., 2002], Wiener filter | • | • | • |
| GSM half-rate codec [ETSI EN 300 969] | • | | • |
| AMR narrowband codec [3GPP TS 26.090] | | • | • |
| AMR wideband codec [ITU-T Rec. G.722.2] | • | • | |
| EVRC wideband codec [3GPP2 C.S0014-E] | • | | |
| *Live conditions* | | | |
| AMR narrowband codec + in-handset NR | • | • | • |
| AMR narrowband codec, in-handset NR disabled | • | • | • |
| EVRC narrowband codec + in-handset NR | • | | • |
| AMR wideband codec + in-handset NR | | • | |
| Total number of conditions | 33 | 32 | 32 |
| Number of listeners | 26 | 28 | 27 |

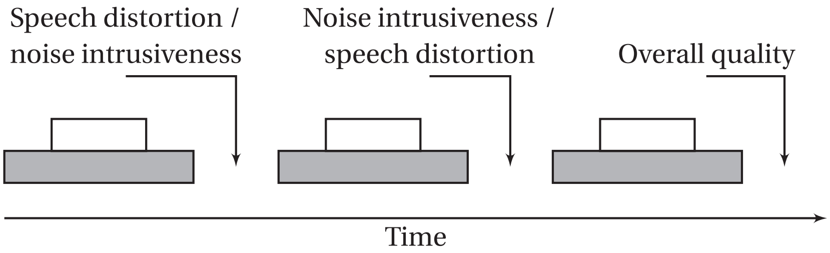

Speech distortion / noise intrusiveness
Noise intrusiveness / speech distortion
Overall quality
Time

Figure 3.3 – Rating order of quality features in P.835 subjective tests. The rating order for speech distortion and noise intrusiveness is inverted for 50% of listeners. Each recording consists of a single sentence embedded in background noise.

Table 3.2 – Subjective rating scales for the quality features defined in [ITU-T Rec. P.835, 2003]. Higher ratings stand for higher quality (i.e., *lower* speech distortion or noise intrusiveness).

| Rating | Speech distortion | Noise intrusiveness | Overall quality |
|---|---|---|---|
| 5 | Not distorted | Not noticeable | Excellent |
| 4 | Slightly distorted | Slightly noticeable | Good |
| 3 | Somewhat distorted | Noticeable but not intrusive | Fair |
| 2 | Fairly distorted | Somewhat intrusive | Poor |
| 1 | Very distorted | Very intrusive | Bad |

The average rating across listeners for a given recording or condition is called Mean Opinion Score (MOS). In the following, we use the abbreviations S-MOS, N-MOS and G-MOS to designate the speech distortion, noise intrusiveness and overall (global) MOS, respectively.

## 3.4 Evaluation of subjective scores

Through a similar design in the proportions and types of conditions, we aimed to obtain comparable listener scores between the three sets. As shown in the top section of Table 3.3, the average MOS across all conditions is very similar for each quality feature. The average G-MOS (overall quality) lies almost exactly in the center of the five-point scale (1.0 to 5.0). We can also observe that the maximum N-MOS is 5.0 in all sets (corresponding to clean conditions), whereas that value is never reached with the two other quality features. This is typical for absolute category rating (ACR) tests, where subjects have no reference signals for comparison, and thus remain unsure about the complete absence of speech distortions.

The bottom section of Table 3.3 gives an overview of per-condition 95% confidence intervals. Our subjects appeared more consistent or confident in scoring noise intrusiveness than speech distortion. This can be explained by the fact that noise is easily assessed during speech pauses, while speech distortions may be hard to perceive in the presence of noise.

Table 3.3 – Overview of subjective scores in the three “PANDA” datasets.

| Dataset | *Minimum* – Average – *Maximum* | | |
|---|---|---|---|
| | S-MOS | N-MOS | G-MOS |
| *MOS per condition* | | | |
| Set 1 (SWB) | *1.5* – 3.4 – *4.9* | *1.2* – 3.3 – *5.0* | *1.3* – 3.0 – *4.9* |
| Set 2 (SWB) | *1.3* – 3.4 – *4.9* | *1.1* – 3.3 – *5.0* | *1.4* – 2.9 – *4.9* |
| Set 3 (NB) | *1.1* – 3.4 – *4.6* | *1.3* – 3.1 – *5.0* | *1.0* – 3.0 – *4.6* |
| *CI95 per condition* | | | |
| Set 1 (SWB) | *0.05* – 0.14 – *0.21* | *0.03* – 0.10 – *0.15* | *0.05* – 0.12 – *0.17* |
| Set 2 (SWB) | *0.06* – 0.14 – *0.21* | *0.03* – 0.10 – *0.16* | *0.06* – 0.12 – *0.20* |
| Set 3 (NB) | *0.05* – 0.16 – *0.21* | *0.02* – 0.11 – *0.15* | *0.02* – 0.12 – *0.19* |

### 3.4.1 Effect of signal bandwidth and bandwidth context

The MOS maxima in Table 3.3 already hinted at a possible difference in scores for different signal bandwidths. In the top section of Figure 3.4, we compare the effect of signal bandwidth for conditions with various levels of “crossroad” noise. These conditions were included in each set to compare the distribution of N-MOS values between sets (so-called *anchors*).

Since there are no distortions other than noise, the S-MOS remains almost constant, but with higher scores for the SWB signals of Sets 1 and 2. Higher absolute quality scores for SWB signals are known from another test method [ITU-T Rec. P.800, 1996] that only focuses on overall quality. The same effect appears here, except for the N-MOS, which is almost identical between sets. The similarity of N-MOS values also indicates that different degrees of intrusiveness are represented similarly within each set, despite having been scored by different listener panels.

Conversely, the conditions in the bottom section of Figure 3.4 were processed with a NB codec, but presented in tests where other conditions had larger bandwidths (SWB context, Set 2) or the same bandwidth (Set 3). The awareness of band limitations in the SWB context results in a compression of S- and G-MOS to lower scores. This is not the case for noise intrusiveness, where scores remain nearly the same and with identical rank order.

Despite the limited number of comparisons that can be drawn here, it appears that listeners rate band limitations in speech and noise differently. Band-limited speech occupies a smaller part of the S- and G-MOS scales. In contrast, listeners do not seem to have any expectations with regard to noise, and use the entire N-MOS scale irrespective of its bandwidth. Finally, in Figure 3.4 the G-MOS reflects the trend of both speech distortion and noise intrusiveness. We will further study the relation between quality features in Section 3.5.

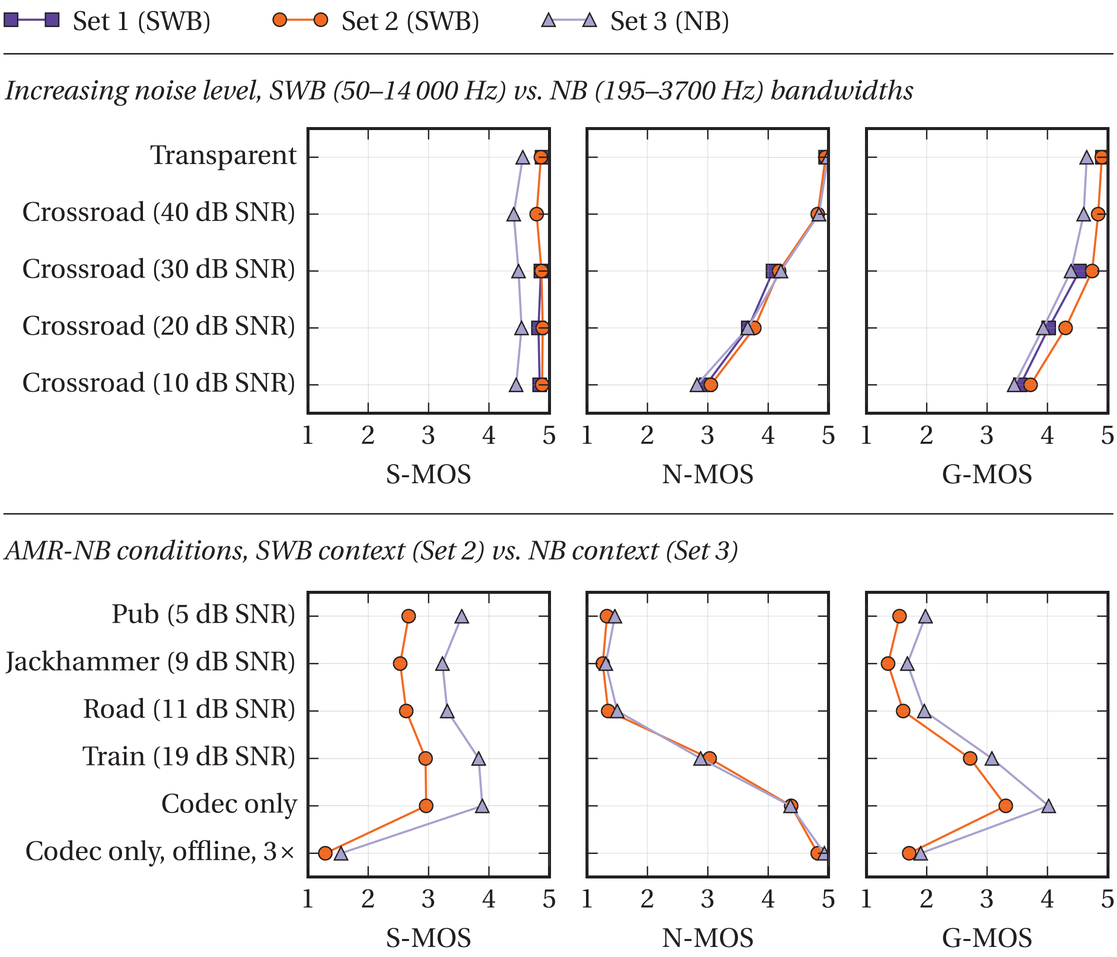


Figure 3.4 – Comparison of subjective scores for different signal bandwidths (top section) and listening test contexts (bottom section).

### 3.4.2 Effect of Lombard speech and presentation level

In designing our tests, we hypothesized that noise-corrupted speech in telecommunications sounds more realistic with the Lombard effect, and may therefore yield lower noise intrusiveness scores. To study this effect, we designed one of the sets (Set 1) to include pairs of conditions differing only in the use of Lombard vs. regular speech, but with the same presentation level and SNR. Additionally, we included conditions with both Lombard speech and higher presentation level, but still with the same SNR.

Since these conditions were all presented to the same listener panel, we can evaluate their effect by means of a paired Wilcoxon signed-rank test on listener ratings. In the following subsections, the terms "significant" (denoted $*$ in figures) and "highly significant" (denoted $**$) refer to differences at the levels $p < 0.05$ and $p < 0.01$, respectively. The increased probability of Type I errors (false positives) with multiple comparisons is compensated through Holm-Bonferroni correction, as discussed in Section 2.3.3 of the Background chapter.

The top section of Figure 3.5 compares subjective scores for both speech types. Noise intrusiveness remains essentially the same for most conditions, but there is a highly significant difference for a condition with strong noise reduction, which only left some noise during speech activity portions. Similarly, speech distortion scores only differ highly significantly for two conditions, both of which are low-SNR conditions with "pub" (babble) noise. These increases in S- and N-MOS remain significant after correcting for multiple comparisons ($p < 0.01$ and $p < 0.05$, respectively), but differences in overall quality are no longer significant.

We speculate that the observed effects are due to the flatter signal envelope of Lombard speech. Specifically, the sustained speech level throughout the sentence length may have better masked the residual noise in the "car" noise condition than regular speech, and helped listeners distinguish between background babble and foreground speech distortions in the "pub" noise conditions. In other words, the few observed quality gains seem to be more related to the signal acoustics than to a better acceptance of noise with Lombard speech.

The effect of presentation level (79 vs. 87 dB SPL) is shown in the bottom section of Figure 3.5. Highly significant differences in N-MOS appear for all background noise conditions, with higher noise levels consistently being rated as more intrusive, despite unchanged SNRs. The comparison in the last row (Lombard speech at standard vs. increased level) supports that the difference is indeed due to presentation level. After correcting for multiple comparisons, all N- and G-MOS differences remain significant ($p < 0.01$ and $p < 0.05$, respectively).

The observed effect of presentation level implies that to some extent, listeners rate noise independently of speech. We will exploit this result for our objective measure presented in Chapter 4, page 41.

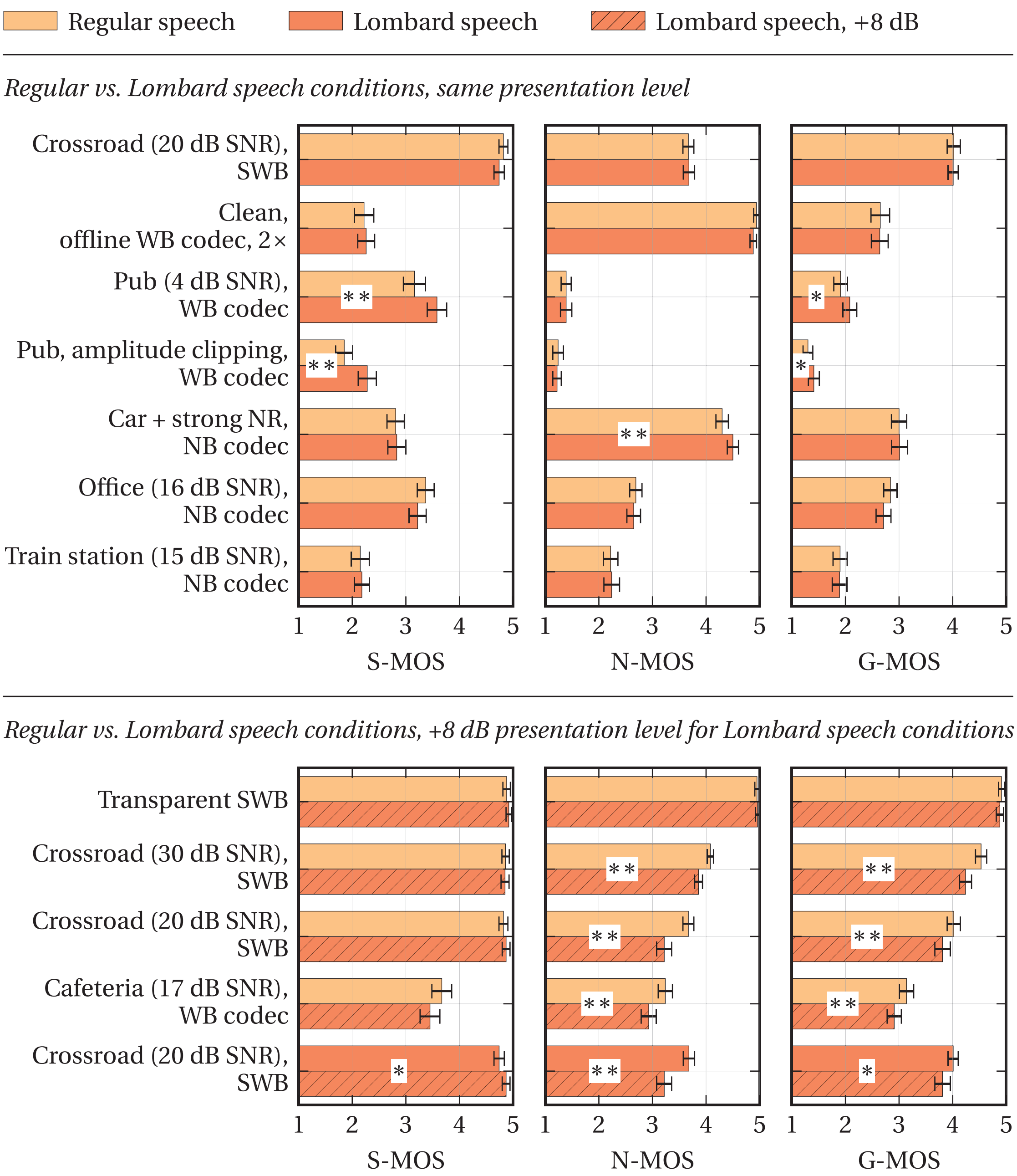


Figure 3.5 – Comparison of subjective scores as a function of speech type (regular vs. Lombard speech, top section) and presentation level (bottom section). Error bars show 95% confidence intervals. Asterisks ($*$ / $**$) indicate significant differences ($p < 0.05$ / $p < 0.01$) between scores. Overall, Lombard speech has little impact on quality scores (top section), whereas presentation level significantly affects perceived noise intrusiveness (bottom section).

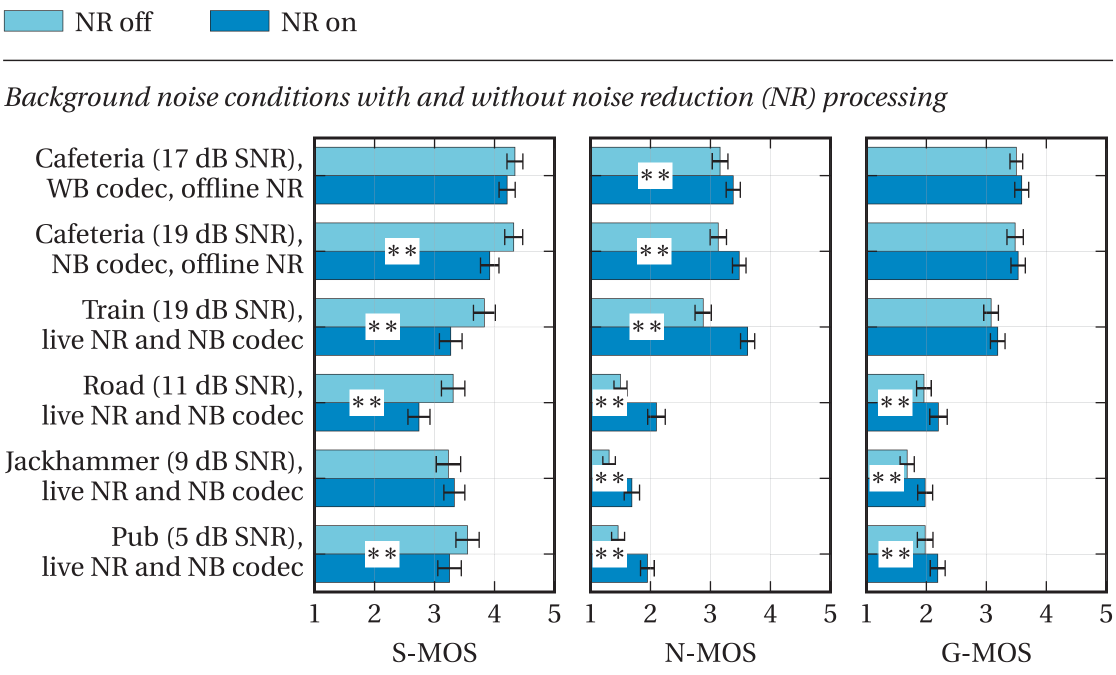


Figure 3.6 – Comparison of subjective scores with and without noise reduction (NR) processing. Error bars show 95% confidence intervals. Asterisks ($*$ / $**$) indicate significant differences ($p < 0.05$ / $p < 0.01$) between scores. NR always reduces noise intrusiveness, but overall quality only improves significantly for low-SNR conditions.

### 3.4.3 Impact of noise reduction processing

The P.835 test method was designed to assess the trade-off between noise intrusiveness and speech distortion that arises with noise reduction (NR). Our datasets include six pairs of conditions with identical noises and processing steps except for the use of NR. Figure 3.6 compares two conditions processed with a commercial NR solution, as well as four conditions from a mobile handset in which NR can be switched on or off.

Unsurprisingly, NR processing always results in a highly significant improvement in terms of noise intrusiveness. However, this improvement is offset by increased speech degradation, which limits the benefit of NR in terms of overall quality to conditions with low SNRs. All differences reported in Figure 3.6 remain significant after correcting for multiple comparisons ($p < 0.01$ for N-MOS, and $p < 0.05$ for S- and G-MOS).

It should be noted that the improvement of overall quality is not the only purpose of NR; removing noise can also help improve the performance of codecs based on a speech production model [e.g., the “EVRC” codec, 3GPP2 C.S0014-E, 2011]. Moreover, cultural preference may change the relative importance of speech distortion and noise intrusiveness in listeners’ overall quality perception. We now analyze this relation between quality features.

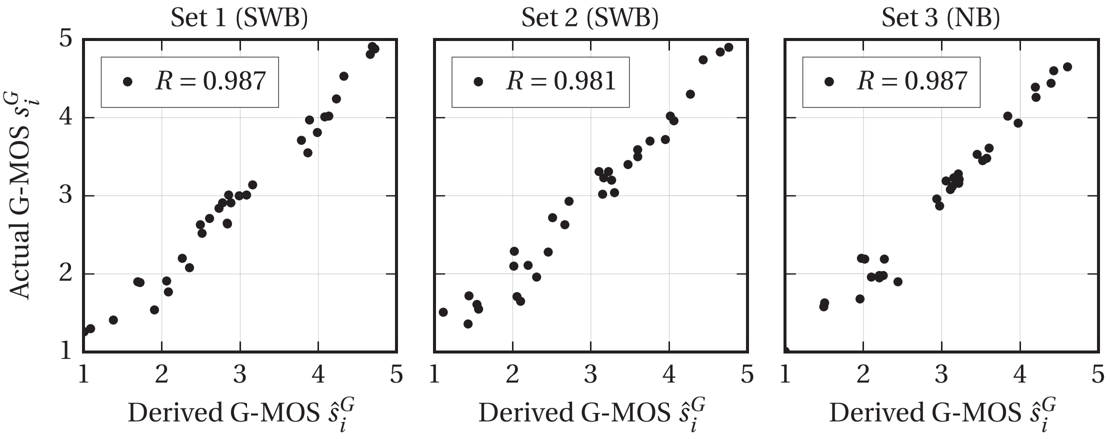


Figure 3.7 – Linear regression of per-condition G-MOS from S- and N-MOS. Values of $\hat{s}_i^G$ outside the range [1, 5] are clipped to the nearest value on the MOS scale (1 condition in Set 3).

Table 3.4 – Comparison of coefficients across datasets for the linear regression in (3.1).

| Dataset | $a$ | $b$ | $c$ |
|---|---|---|---|
| Set 1 (SWB) | 0.696 | 0.402 | −0.697 |
| Set 2 (SWB) | 0.742 | 0.434 | −0.994 |
| Set 3 (NB) | 0.711 | 0.464 | −0.946 |

## 3.5 Relation between the three quality features

Due to the rating order in P.835 tests, where listeners always rate overall quality last, it can be expected that the overall quality rating combines listeners' opinion of speech distortion and noise intrusiveness. In order to analyze this relation in our subjective scores, we have assumed a simple linear relation of the form

$$s_i^G \approx a \cdot s_i^S + b \cdot s_i^N + c = \hat{s}_i^G, \tag{3.1}$$

with $s_i^{G/S/N}$ the G-, S- or N-MOS for the $i^{\text{th}}$ condition, respectively, and $c$ a constant bias term.

Figure 3.7 shows the result of performing a least squares regression for (3.1) in each dataset. There is indeed a very strong linear relation between the G-MOS and the combined S- and N-MOS, as measured by the Pearson correlation coefficient $R$. The respective contributions $a$ and $b$ of speech distortion and noise intrusiveness in the regression are shown in Table 3.4. Speech distortion appears to have exerted a stronger influence on listeners' perception of overall quality. Moreover, the coefficients $a$ and $b$ are very similar across datasets, with the largest difference being the overall bias term $c$.

The observed relation is directly relevant to the design of objective measures for P.835 scores. Instead of designing three separate measures, it may be possible to focus only on the quality

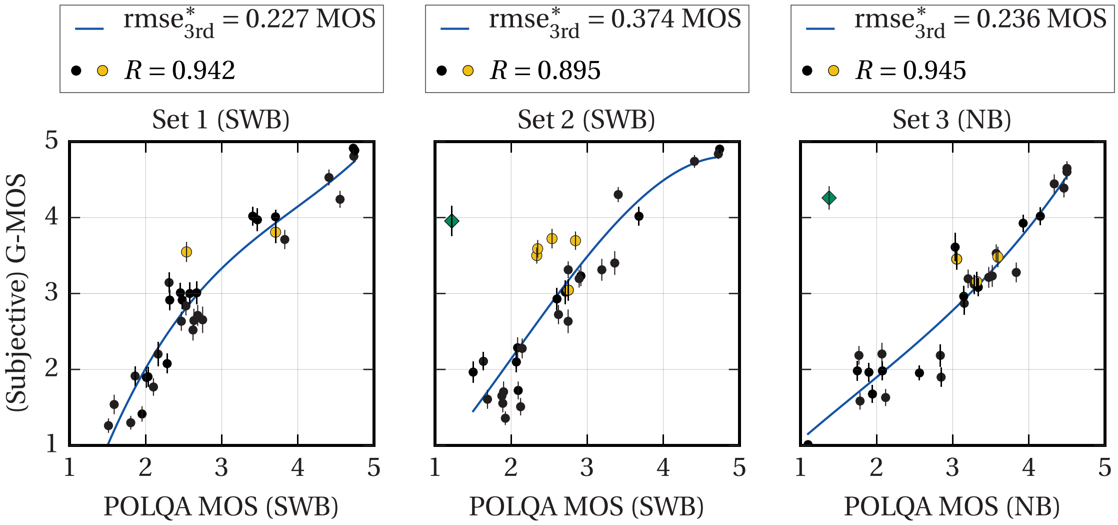


Figure 3.8 – Predictions of the P.835 G-MOS (overall quality) with ITU-T Rec. P.863 “POLQA”. Error bars indicate 95% confidence intervals of subjective scores. Blue lines show the mapping functions used to determine the prediction error $\text{rmse}^*_{3\text{rd}}$. One mis-predicted condition with strong delay jitter (green diamond) was excluded from the evaluation in Sets 2 and 3.

features that are specific to the P.835 test method. In particular, ITU standards for objectively assessing overall quality already exist [ITU-T Recs. P.862 “PESQ”, 2001, and P.863 “POLQA”, 2011], albeit for a different test method [ITU-T Rec. P.800, 1996], which does not first draw listeners’ attention to the trade-offs between speech distortion and noise intrusiveness. We evaluate in the next section how well the P.835 G-MOS can be predicted with P.863 “POLQA”.

Given objective scores for overall quality and for another quality feature, an estimate of the third feature score may be derived by exploiting their interdependency. We provide details of such a derivation in Appendix B, page 111, where we assume the availability of G- and N-MOS to derive the speech distortion MOS (S-MOS).

## 3.6 Prediction of overall quality scores

As seen in the previous sections, the P.835 overall quality score reflects the impact of additive background noise, as well as distortions to the speech component in the signal. Both types of signal degradations were included in the training of the most recent ITU standard for the objective assessment of overall quality, ITU-T Rec. P.863 “POLQA” [2011].[3] This measure was designed to predict subjective scores from a different test method [ITU-T Rec. P.800, 1996], where listeners only provide overall quality ratings. This means that listeners were not necessarily attentive to the trade-off between speech distortion and noise intrusiveness as in a P.835 test, and may thus have scored noisy conditions differently.

[3] In the interest of full disclosure, the author of this thesis is a co-author of the original POLQA algorithm [Beerends et al., 2013a; b]; however, the author has not been affiliated with the POLQA coalition since March 2012.

The POLQA algorithm computes quality predictions through a full-reference approach, i.e., by comparing each test signal to its corresponding reference signal. Figure 3.8 compares the G-MOS of our datasets to POLQA scores, averaged per condition.[4] The algorithm failed to align a Skype condition with strong delay jitter to its reference signals (green diamond in Sets 2 and 3), therefore we exclude this condition from further evaluation.

The overall quality of remaining conditions is rather well predicted, as measured by the two performance metrics introduced in Section 2.3. The prediction error $\text{rmse}^*_{3\text{rd}}$ lies below the worst case of 0.28 MOS indicated in the POLQA standard [ITU-T Rec. P.863, 2011, App. I.3], with the exception of Set 2. Closer analysis reveals an offset for conditions that are mainly degraded by background noise (yellow circles in Figure 3.8); this degradation is judged too severely by POLQA. Due to the large number of background noise conditions in P.835 tests, it is probable that listeners become more accustomed to noise than in general-purpose P.800 tests. The prediction of P.835 G-MOS with POLQA could thus likely be improved by re-training the measure with data that includes a higher proportion of background noise conditions.

## 3.7 Discussion and conclusion

In this chapter, we have presented the design and analysis of our P.835 datasets. The "PANDA" or Perceptual Assessment of Noise DAtasets serve the dual purpose of investigating how background noise is perceived in telecommunications, and providing development data for a new objective measure. Our investigation focused on several factors in noisy telephone speech, which we further discuss here:

*Bandwidth of the signal and bandwidth context* — Our analysis in Section 3.4.1 showed an offset in the maximum scores for speech distortion and overall quality between narrowband and super-wideband tests. Such offsets are well known from other test methods, as discussed in Section 3.1, and can be attributed to used the absolute category rating (ACR) scales, which listeners interpret based on their quality expectations [Möller, 2000, Chap. 6]. In other words, there is always a certain proportion of listeners for whom high-quality speech does not match their expectation of "undistorted" or "excellent" quality, but that proportion becomes smaller with super-wideband signals.

Noise intrusiveness on the other hand did not show such an offset, with average listener scores spanning the full N-MOS scale irrespective of noise bandwidth or bandwidth context.

*Lombard effect and presentation level* — We have used speech recordings with the Lombard effect for noisy conditions, as examples of speech from a caller located in a noisy environment. We hypothesized that such more realistic stimuli would result in a better acceptance of noise by listeners, as measured by the noise intrusiveness rating. However, our analysis in Section 3.4.2

[4] POLQA scores were computed by the industrial partner, since source code for POLQA is not publicly available.

did not reveal a notable difference. The levels of Lombard and regular speech were equalized in our comparison in order to have the same signal-to-noise ratio, leaving only differences in spectral and temporal speech structure.

It is possible that our setup for recording Lombard speech (speakers hearing noise over closed headphones, with microphone feedback) did not realistically simulate communication in a noisy environment. In particular, noise presented over headphones has been found to elicit stronger speech modifications than presentation over loudspeakers [Garnier et al., 2010]. On the other hand, the missing impact on noise intrusiveness may simply be due to the P.835 test method, where listeners are directed to focus on the noise exclusively, without considering the speech component.

This latter notion is supported by the second tested factor of *presentation level*, where higher noise levels were rated as being more intrusive, despite unchanged SNRs. Higher presentation levels also negatively affect intelligibility, as referred to in Section 3.1. However, the lack of an improvement of noise intrusiveness scores with Lombard speech, which has generally higher intelligibility (see Section 3.1), indicates that listeners rate noise intrusiveness independently of speech. Combined with our observations on bandwidth and bandwidth context, these results imply that listeners rate noise without a particular expectation, neither with regard to the noise itself, nor to its relation to the speech signal. We will exploit this finding in the next chapter to objectively assess noise intrusiveness by analyzing the properties of noise only.

*Relation between the three P.835 quality features* — As verified in Section 3.5, listeners combine their opinion of speech distortion and noise intrusiveness in the overall quality rating. This result is consistent with the analysis of Hu and Loizou [2007b, Sec. 5.4], who also found a stronger influence of speech degradation on overall quality scores. We have further evaluated the prediction of overall quality scores with P.863 "POLQA" in Section 3.6, and observed good prediction performance for most conditions.

In conclusion, these results set noise intrusiveness apart as a quality feature that listeners perceive differently. Building on this, we will focus on its objective assessment in Chapter 4.

# 4 Objective assessment of background noise intrusiveness

In Chapter 3, we investigated how listeners rate noise-corrupted speech with respect to the three quality features defined in ITU-T Rec. P.835, i.e., *speech distortion*, *noise intrusiveness* and *overall quality*. Our experiments showed that overall quality scores could be predicted adequately with an existing objective measure (ITU-T Rec. P.863 [2011]), and that the three quality features are interdependent, meaning that they could be predicted with just two objective measures. Moreover, the results showed an important difference between noise intrusiveness and the two other quality features, in that the perception of noise did not seem to be guided by listeners' expectations, nor by its relation to foreground speech.

This latter property is relevant to the design of objective measures, which are often based on a comparison of the test signal to an undistorted reference, an approach known as full-reference assessment (see Section 2.2.3, page 8). Given that subjective noise intrusiveness scores only reflect the presence of noise itself, and that the reference signal is noise-free, such a comparative approach is of limited use here. Instead, existing approaches analyze the test signal (e.g., during speech pauses) and extract multiple noise features in time and frequency. These features are then combined to a predicted noise intrusiveness score, using a regression learned from training data. A challenge is thus to use only few, highly predictive features, since training data with subjective scores is expensive to collect.

In this chapter, we propose a novel, single feature to predict perceived noise intrusiveness. Our approach is based on a sparse noise representation as a model of high-level sensory coding, described in Section 4.2. Our hypothesis is that such a noise representation in an auditory-inspired basis is indicative of its perceived intrusiveness. To validate this hypothesis, we present a study in Section 4.3 with simple noise types, and show that the number of atoms in the representation models several factors in the perception of noise. In Section 4.4, we then propose a method for predicting the intrusiveness of noise in speech recordings and describe the experimental setup for its evaluation in Section 4.5, using the PANDA datasets introduced in Chapter 3. The results in Section 4.6 show a high correlation ($\overline{|R|} > 0.95$) between subjective intrusiveness scores and the proposed feature, and that it outperforms or compares to a traditional loudness-based feature. We conclude with further remarks in Section 4.7.

## 4.1 Related work and contributions

The perception of noise has been extensively studied in the context of environmental noise annoyance. Of the relevant acoustic factors,[1] the perceived intensity of noise emerged as the dominant aspect, with spectral composition and temporal variability as additional factors [Marquis-Favre et al., 2005; Alayrac et al., 2010; Fastl and Zwicker, 2007, Chap. 16.1]. A rough calculation of perceived intensity is the log noise energy in decibels (dB), although better approximations also consider the frequency-dependent sensitivity of human hearing. This can be done by assigning specific weights to the energies within different frequency bands, as given for example in the "A" weighting curve [IEC 61672-1, 2013] and denoted "dB(A)".

More advanced estimations of intensity apply detailed models of the processing at the outer, middle and inner ear to derive an estimate called loudness [see e.g., Fastl and Zwicker, 2007, Chap. 8]. Loudness takes the spectral composition of sound into account to estimate how the relative signal power across frequency bands combines to an overall perceived intensity. These and other features can be calculated either for the long-term average noise spectrum or over short-time intervals.

Calculation over short-time intervals is relevant to the assessment of non-stationary noise. In particular, it is well known that subjective judgments are disproportionately influenced by peak events [Fredrickson and Kahneman, 1993]. Applied to noise perception, this means that simple averaging of short-time features tends to under-estimate the effect of more intrusive segments [Fastl and Zwicker, 2007, Chap. 16.1]. Therefore, the aggregation of short-time features to a predicted score may assign higher weights to some segments, or otherwise consider the variance of feature values over time.

Existing objective measures of noise intrusiveness for telecommunications follow these principles, and analyze multiple features of noise [Gautier-Turbin and Le Faucheur, 2005; Reimes et al., 2011; Narwaria et al., 2012].[2] Main features in these measures model perceived intensity using the short-time noise loudness, log energy or filterbank energies, respectively. Secondary noise features include spectral peakiness, emphasis on time-variant noise structures, or energy variance to account for higher intrusiveness in tonal or non-stationary noises. In each case, a regression is used to combine multiple features to a predicted intrusiveness score.

A disadvantage of these approaches is that the regression parameters to combine different features must be determined either with expert knowledge or using training data, i.e., subjectively scored data that is expensive and scarce. Moreover, the computation of some individual features involves further internal parameters. Examples are the parameters of loudness models that were derived from subjective experiments, or the use of perceptually motivated prepro-

[1] There are also important non-acoustic factors in the perception of noise, e.g., listeners' belief that noise could be avoided [Marquis-Favre et al., 2005]. These factors cannot be assessed through in-laboratory listening tests and are beyond the scope of this thesis.

[2] A modified version of the measure of Reimes et al. [2011] was also adopted in a European standard [ETSI TS 103 106, 2014].

cessing tuned with expert knowledge [Reimes et al., 2011]. The feature proposed in this chapter seeks to assess perceived noise intrusiveness while avoiding these drawbacks. Specifically, the proposed feature directly achieves high correlation with subjective scores, allowing to avoid combination with secondary features. It also makes very few assumptions of the parameters with which intrusiveness is predicted, thus avoiding the need for training data.

Our approach is based on the *sparse coding* theory, which postulates that sensory systems have evolved to encode stimuli efficiently by using only a small subset of a large population of neurons at a time [Barlow, 1972]. A possible mathematical abstraction of sparse coding is the sparse representation of a signal in an overcomplete basis [Olshausen and Field, 1997]. This signal model was substantiated for sound by the work of Smith and Lewicki [2006], who showed that a sparse coding model could predict cochlear filter shapes as the most efficient basis set for a wide range of sound classes. The present work takes this approach further and applies it to the prediction of noise perception.

Contribution

R. Ullmann and H. Bourlard [2016]. "Predicting the intrusiveness of noise through sparse coding with auditory kernels". In: *Speech Commun.* 76, pp. 186–200.
URL: http://dx.doi.org/10.1016/j.specom.2015.07.005

The present chapter is a slightly revised version of the above paper.

## 4.2 Sparse signal representation

We now provide the background on the sparse coding model on which our proposed approach is based. Sparse coding follows the idea that sensory systems have adapted to encode stimuli in a way that maximizes the amount of information carried to the brain, but minimizes the number of neuronal impulses [Barlow, 1972; see also Laughlin and Sejnowski, 2003, "Saving on Traffic"]. A further motivation for a sparse representation of incoming stimuli is to present information succinctly to later processing stages [Olshausen and Field, 1997].

Sparse coding is thus a more abstract, high-level approach to auditory modeling, compared to the features presented in the previous section that modeled acoustical properties in hearing. Although the approach was originally developed as a computational model of sensory coding, our notion is that it may be possible to derive a perceptually meaningful measure of noise from the sparse coding representation.

### 4.2.1 Signal model

We use the signal model proposed by Lewicki and Sejnowski [1999], which defines an approximation $\hat{x}(t)$ of the waveform $x(t)$ through a linear combination of shiftable kernel functions from a set of $M$ kernels $\phi^1(t),\dots,\phi^m(t),\dots,\phi^M(t)$,

$$\hat{x}(t) = \sum_{m=1}^{M}\sum_{i=1}^{I^m} \alpha_i^m \phi^m\left(t-\tau_i^m\right) \tag{4.1}$$

where $\alpha_i^m$ and $\tau_i^m$ denote the gain and time shift of the $i^{\text{th}}$ occurrence of kernel function $\phi^m(t)$, respectively. The temporally localized occurrences of kernels $\phi$ in the approximation yield a spike-like, shift-invariant representation of the signal $x(t)$. This is in contrast to frame-based approaches like the short-time Fourier transform (STFT), where delaying the signal can move signal components between analysis frames and change the estimated spectral magnitude [Lewicki and Sejnowski, 1999]. It also retains the precise temporal location of signal components, which can be a desirable property for essential auditory tasks like sound source localization.

The error of the approximation is $e(t) = x(t) - \hat{x}(t)$ and is called the *residual signal*. The ability to place kernels from the set $\phi^1(t),\dots,\phi^M(t)$ at arbitrary temporal locations $\tau$ means that the basis set is highly overcomplete. Therefore, for a given target residual signal energy $\|e\|^2$, there exist an infinite number of solutions to (4.1). Sparse solutions are characterized by a low total number of kernel occurrences $\sum_m I^m$.

This signal model was used successfully in [Smith and Lewicki, 2005; 2006] to investigate what type of kernel functions $\phi$ allow for sparse representations of natural sounds.[3] The key result from these studies is that the optimal kernels strongly resemble cochlear filter shapes, as obtained from measurements at the auditory nerve. For our objective measure, we follow a reverse approach: starting with a set of auditory kernel functions, we compute a sparse representation of a noise signal $x(t)$ to predict its perception.

### 4.2.2 Choice of kernel functions

Auditory kernels are attuned to different frequencies over the audible frequency range. We use analytical approximations, where kernels are formulated as gamma-modulated sinusoids known as *gammatones*, defined as

$$\phi^m(t) = (t)^{n-1}\exp\left(-2\pi b^m t\right)\cos\left(2\pi f^m t + \varphi\right), \quad t \geq 0 \tag{4.2}$$

with $b^m$ a bandwidth parameter and $f^m$ the center frequency. Gammatones have been used to characterize the impulse response of cochlear nerve fibers in cats; specifically, the output of a gammatone filter was shown to be a good predictor of the corresponding fiber's firing probability [De Boer and De Jongh, 1978]. Therefore, each occurrence of auditory kernels $\phi^m$ in (4.1) may be interpreted as a population of auditory nerve spikes, with average firing rate encoded by the kernel gain $\alpha$ [Lewicki, 2002].

[3] Recordings of 5–40 seconds, featuring natural ambient noises, transients and mammalian vocalizations.

We use the parameters $n = 4$ and $b^m = 1.019\ \mathrm{ERB}(f^m)$ at which gammatones provide a good fit to human psychoacoustic data [Patterson et al., 1992]. $\mathrm{ERB}(f^m)$ denotes the auditory Equivalent Rectangular Bandwidth scale [Glasberg and Moore, 1990] and is given by

$$\mathrm{ERB}(f^m) = 0.108 f^m + 24.7 \quad [\mathrm{Hz}]\,. \tag{4.3}$$

Since a goal of our approach is to reduce the use of parameters obtained through subjective experiments, we also evaluate more analytical gammatone parameters in Section 4.6.2, page 56.

### 4.2.3 Computing a sparse solution

Several methods can be used to sparsely approximate a discrete noise signal $x[n]$ with a linear combination of gammatones $\phi^m[n]$. Here we use a modified version of Matching Pursuit (MP) [Mallat and Zhang, 1993], an iterative method that does not necessarily yield the sparsest possible approximation, but is conceptually simple and computationally tractable.

Briefly, MP projects the input signal $x[n]$ onto the set of unit-normed kernels $\{\boldsymbol{\phi}^m\}$, shifted at all possible time offsets within the length of $x[n]$. The kernel that has the highest correlation with the input signal is added to the initial approximation $\hat{x}[n]_{(1)}$, producing the residual signal $e[n]_{(1)} = x[n] - \hat{x}[n]_{(1)}$. At the following iterations $k \geq 2$, the method uses the previous residual $e[n]_{(k-1)}$ as input signal, and again selects the most correlated kernel to produce an updated approximation $\hat{x}[n]_{(k)}$ and residual $e[n]_{(k)}$. These iterations continue until a given stopping criterion, discussed below, is met.

Selecting the most correlated kernel in each iteration amounts to maximizing the objective function $\left\| e[n]_{(k-1)} \right\|^2 - \left\| e[n]_{(k)} \right\|^2$, i.e., the decrease of residual energy [Goodwin and Vetterli, 1999]. Moreover, the time offset of the most correlated kernel can be determined through convolution, so it is not necessary to actually store time-shifted copies of the set of kernels.

The used modification of MP [Gribonval, 1999] consists in projecting onto analytic kernels

$$\boldsymbol{\phi}_a^m = \boldsymbol{\phi}^m + i \cdot \mathscr{H}\big(\boldsymbol{\phi}^m\big)\,, \tag{4.4}$$

allowing to also adjust the phase $\varphi$ in (4.2) of each gammatone, with $\mathscr{H}(\cdot)$ the Hilbert transform operator. The complete MP calculation steps are formally presented in Appendix C, page 115.

Each MP iteration adds a kernel occurrence with a specific gain, temporal offset, center frequency and phase to the approximation. We perform iterations until the energy $\|\alpha\|^2$ of newly added kernels falls below a set threshold. This stopping criterion can be thought of a lower limit on the firing rate of auditory nerves in the model. We adopt the terminology of earlier work in [Smith and Lewicki, 2005; 2006] and refer to each kernel occurrence in the approximation simply as a *spike*. Given a noise waveform with sound pressure in Pa, it follows that the “spike” energy threshold $\|\alpha_{\min}\|^2$ has units $\mathrm{Pa}^2$.

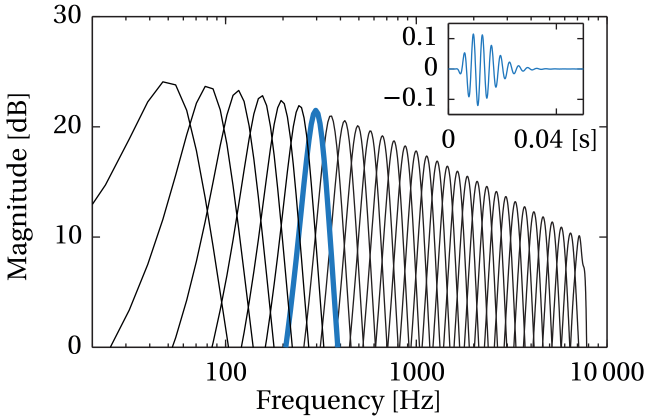


Figure 4.1 – Frequency responses of gammatone dictionary kernels. The waveform for the highlighted (thick) frequency response is shown in the inset. Peak magnitudes differ because kernel waveforms are normalized to unit norm.

## 4.3 Properties of the sparse representation

We perform a series of experiments to study how the sparse coding representation of noise relates to its perception. The following experiments use a set of kernels or *dictionary* of $M = 32$ gammatones generated with the parameters of Section 4.2.2 and shown in Figure 4.1. The gammatones are sampled at 16 kHz and have center frequencies $f^m$ between 50 and 7150 Hz, distributed evenly on the ERB scale. The support of kernels (measured as the length over which amplitudes are $\geq 5 \cdot 10^{-5}$ the gammatone peak) ranges from ~4 to 128 ms. Dictionaries of other sizes and with other kernel types will also be compared in Section 4.6.2, page 56.

### 4.3.1 Sparse representation and effect of signal level

We first study the placement of kernel occurrences (“spikes”) over time in the representation, as well as its dependency on signal level as one key factor in the perception of noise. Figure 4.2 visualizes the sparse approximation of a speech signal (we use speech instead of noise in this example because it nicely combines different signal structures). The waveform of the word “punch” is sparsely approximated using MP until the energy of new spikes falls below 0.01 Pa$^2$. The resulting ~500 spikes are shown in a so-called *spikegram* as dots of different sizes, proportional to the log energy of the spike. Each spike is plotted at the Hilbert envelope peak position and center frequency, respectively, of the corresponding kernel.

The obtained representation exhibits a high degree of localization in both time and frequency, showing the succession of harmonic, transient and unvoiced structure in the signal. The number of spikes (i.e., the sparsity of the obtained representation) depends on how many gammatone kernels with energy above the threshold are extracted from the signal. This number scales *logarithmically* with the average signal energy, due to the exponential decay of successive spike energies [Mallat and Zhang, 1993], as shown by the solid line in the bottom

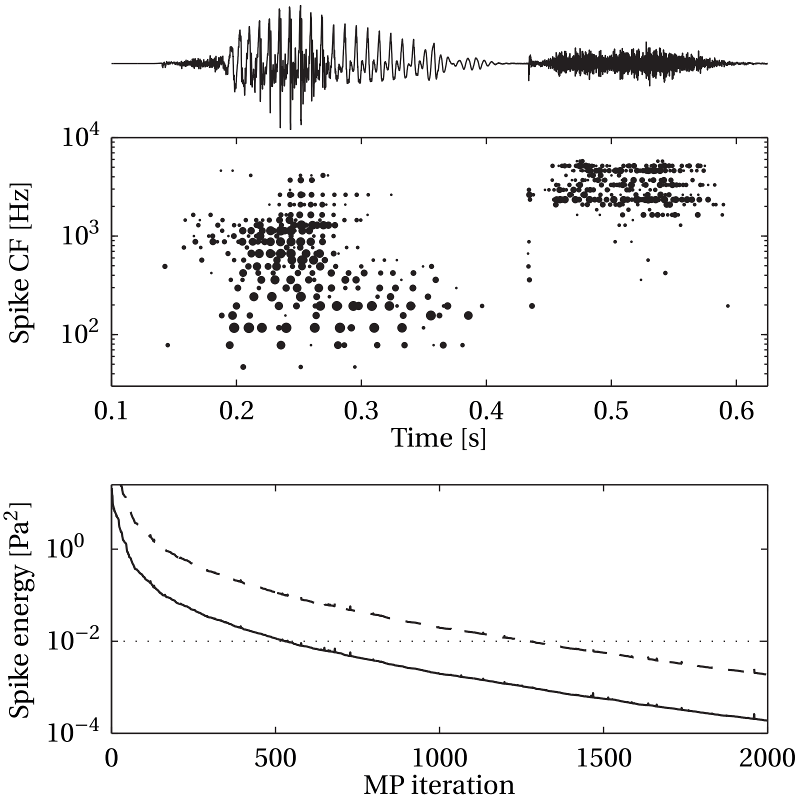


Figure 4.2 – Spikegram of the word “punch”. Dots in the middle panel represent spikes obtained after approximating the top waveform up to energies ≥ $10^{-2}$ Pa$^2$. The bottom panel shows the decrease of successive spike energies extracted from the signal at its original level (solid line) and at +10 dB (dashed line).

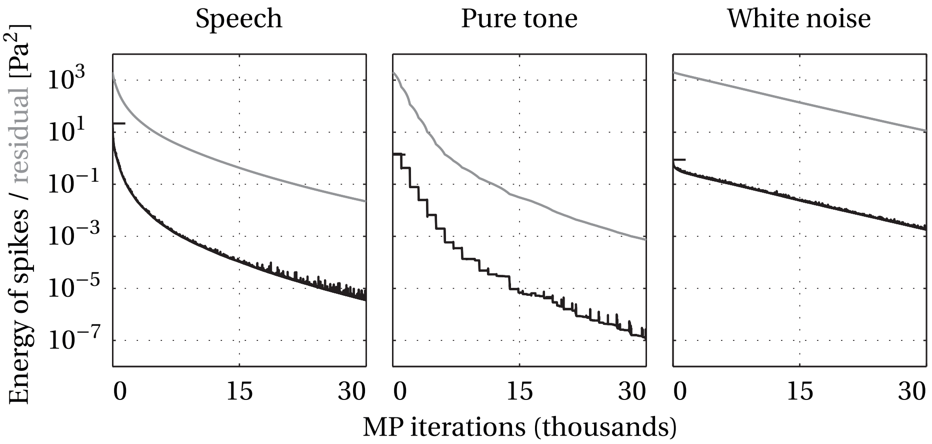


Figure 4.3 – Decay of spike energies for different signal types. We compare three signals with the same duration and average sound level (i.e., the same initial residual error energy $\|\mathbf{e}\|^2$, gray lines). A thick tick mark indicates the maximum spike energy $\|\alpha_{j(1)}\|^2 = \|\mathbf{e}_{(0)}\|^2 - \|\mathbf{e}_{(1)}\|^2$ in each panel. The values and decay rate of spike energies (black lines) depend on how succinctly the dictionary kernels encode structures in the different signals.

panel of Figure 4.2. A 10 dB increase in signal level results in an upward shift of this line (dashed line), and thus in a *linear* increase in the number of MP iterations until the same threshold is reached. A feature based on the number of spikes therefore inherently accounts for the roughly logarithmic human perception of sound intensity.

### 4.3.2 Effect of signal type

In order for the sparse coding representation to be indicative of noise perception, we expect it to differ for different types of noise signals. We study this property with three signal types: speech, a sinusoidal tone at 2 kHz and white Gaussian noise. The signals have the same duration and average level to allow better comparison. Figure 4.3 shows the decay of both spike energies and residual signal energy, which are linked through the relationship $\|\alpha_{j(k)}\|^2 = \|\mathbf{e}_{(k-1)}\|^2 - \|\mathbf{e}_{(k)}\|^2$, with $j(k)$ the kernel selected at the $k^{\text{th}}$ MP iteration (see Appendix C, page 115).

Speech and the sinusoidal tone can both be efficiently approximated with gammatones, so the residual error (gray line) decreases quickly. The decrease is steepest at initial iterations, where high-energy spikes are used to approximate harmonic structure in the speech signal, after which the remaining structure is decomposed into many lower-energy gammatones. Similarly, the pure tone shows a steep initial step-like pattern of spike energies (black line), corresponding to the approximation of repeated tonal structure across the length of the signal. As more and more gammatones are subtracted, the residual signal becomes less tonal, and the decay rate levels off. White noise finally is not well approximated with any gammatone in the dictionary and shows a slow, near-constant decay rate.

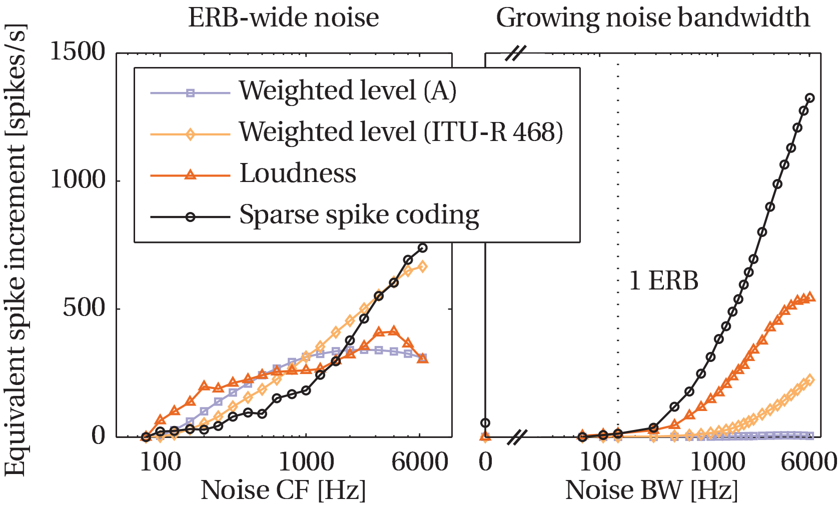


Figure 4.4 – Effect of noise center frequency (CF) and noise bandwidth (BW). A narrowband noise signal is kept at a constant level, but shifted across frequency (left panel) or increased in bandwidth (right panel). In the resulting sparse representation, the number of spikes follows the trend of traditional frequency weighting curves (left panel), and also models the perceived increase in intensity when the bandwidth grows beyond 1 ERB (right panel).

The number of spikes with energies above a fixed threshold thus depends on the type of signal. Depending on the threshold value $\|\alpha_{\min}\|^2$ that is used as stopping criterion, the representation will capture only the higher-energy (i.e., dominant) and tonal structures in the signal, or also include softer and more noise-like structures with lower spike energies. This behavior is consistent with the higher intrusiveness of tonal noises [Marquis-Favre et al., 2005]. Moreover, the dominance of higher-energy structure in complex signals bears a resemblance to the *masking* effect that some objective measures estimate through loudness models.

### 4.3.3 Effect of spectral distribution

The last experiment compares the representation for different types of band-limited noise. Specifically, white Gaussian noise is bandpass filtered to produce noises in different frequency bands or with different bandwidths. The such filtered noises are all rescaled to 59 dB SPL (sound pressure level) and sparsely approximated up to spike energies of $\|\alpha_{\min}\|^2 = 1.16 \cdot 10^{-3}$ Pa$^2$ (this threshold is derived in Section 4.4, page 51).

Figure 4.4 shows how the noise types differ in the number of spikes above threshold. The number of spikes is expressed independently of signal length, in units of spikes/s. In the left panel of Figure 4.4, all noises are 1 ERB (equivalent rectangular bandwidth) wide, but centered at different frequencies. The number of spikes above threshold is higher for noises located in higher bands. As a comparison, a frequency weighting curve used for environmental noise assessment ["A" curve, IEC 61672-1, 2013] and another for noise in audio circuits [ITU-R Rec. BS.468, 1986] also predict narrowband noise in higher bands to be more objectionable.

Their spike count equivalent is obtained by scaling a reference ERB-wide noise to the level calculated with these features, and counting the number of spikes above threshold in it.[4]

The right panel of Figure 4.4 shows results for noises with constant center frequency but different bandwidths, centered on the ERB scale around 1.1 kHz. The number of spikes in the sparse representation increases with larger bandwidths, especially beyond 1 ERB. The same trend is obtained with a model of loudness [Zwicker, 1977], which predicts the increase in perceived intensity when noise fills more than one auditory band [Fastl and Zwicker, 2007, Chap. 8.3]. Frequency weighting curves on the other hand do not predict this effect. Finally, note that a pure tone (points on the ordinate in the right panel of Figure 4.4) also results in a higher spike count with the sparse coding approach.

The observed behavior of the sparse representation can be explained with the used gammatone dictionary. Since kernels in the dictionary are normalized to unit norm, the large-bandwidth gammatones at higher center frequencies have lower peak magnitudes (see Figure 4.1). This means that higher spike gains $\|\alpha\|^2$ are needed to represent noise in high frequency bands, increasing the number of spikes above threshold. The spike count also increases with wider noise, as it enters more and more gammatone bands.

In summary, these experiments support our hypothesis that spikes in an auditory-inspired sparse noise representation are indicative of its intrusiveness. Specifically, the number of spikes above a fixed threshold $\|\alpha_{\min}\|^2$ scales logarithmically with sound pressure; differs by noise type, with a higher sensitivity to high-energy and tonal components in complex noises; follows the trend of frequency weighting curves for narrowband noise; and considers the higher perceived intensity of wider noises. This feature therefore combines properties of multiple features that can be used to predict noise intrusiveness. The spike count reproduces none of the those features exactly, but this is also not necessary, since those features are only intermediate *correlates* of the actual score that we seek to predict.

[4] Minor irregularities in the sparse coding curve can be seen for narrowband noise centered at 250, 500 and 1000 Hz — these are frequencies that coincide with the center frequency of a gammatone in the dictionary. For the same reason, we used a reference ERB-wide noise centered at 1.1 kHz to determine the spike count equivalent of other features, instead of the more conventional 1 kHz.

## 4.4 Noise intrusiveness measure

Given the properties of the sparse noise representation demonstrated in Section 4.3, we propose the following approach to predict the intrusiveness of noise in a test speech recording:

*Noise extraction* — The background noise needs to be extracted from the recording. This can be done e.g., by trying to separate speech and background noise in the recording, or as done here, by analyzing noise in speech pause sections. This is motivated by the P.835 test method introduced in Section 3.3, page 28, where recordings have a specific temporal structure with leading and trailing pause sections. Listeners successively rate speech distortion, noise intrusiveness and overall quality in the recordings. For intrusiveness, listeners are instructed to focus on the background only, meaning that their rating is likely determined by the noise in speech pauses. Objective assessment may thus focus on this part of the signal only, as done with some existing approaches [Gautier-Turbin and Le Faucheur, 2005; Reimes et al., 2011].

*Sparse approximation* — The extracted noise signal is sparsely approximated using a dictionary of auditory kernels. This can be the gammatone dictionary used in Section 4.3, or other dictionary types as compared in Section 4.6.2, page 56. The threshold $\|\alpha_{\min}\|^2$ for the approximation can be selected without training data, based on the standard speech presentation level of 79 dB SPL in listening tests. Since $\|\alpha_{\min}\|^2$ sets a limit on the level of noise structures to include in the approximation, it may be selected such as to consider structures up to a certain dynamic range below the speech level. This can be motivated by the analogy of "spikes" in the model to the firing rate in actual auditory nerve fibers, which varies for changes in sound pressure within a dynamic range of about 40 dB, beyond which the rate response saturates [Sachs and Abbas, 1974]. A value for $\|\alpha_{\min}\|^2$ is derived along this line in Section 4.4.2.

*Spike counting* — The approximation yields a sparse representation of noise with temporally localized kernel occurrences ("spikes"). These spikes are counted, either for the entire noise duration or within short-time intervals. Counts for short-time intervals can be aggregated non-uniformly to model the disproportionate perceptual effect of more intrusive intervals in non-stationary noises. Here we use the fifth percentile, i.e., the short-time count that is exceeded during 5% of the signal duration. Fastl and Zwicker [2007, Chap. 16.1] found that the 5th percentile of instantaneous loudness values best summarized the perceived loudness of non-stationary noise. Percentile aggregation has since become an established method in the assessment of non-stationary sound [e.g., Axelsson et al., 2010; Huber and Kollmeier, 2006]. While other aggregation types could be imagined, the percentile adds minimal complexity. We evaluate both mean and percentile aggregation in Section 4.6.1, page 55.

The aggregated spike count, time-normalized to units of spikes/s, is indicative of the perceived intrusiveness of noise in the recording. In the just described "spike density" feature, none of the parameters are optimized for a specific dataset.

### 4.4.1 Advantages and limitations

An advantage of the proposed approach is that it uses the idea of sparse coding to model higher-level noise perception. This differs from traditional features, which model lower-level *correlates* (e.g., perceived intensity) of noise intrusiveness. Moreover, traditional features such as weighted noise level or loudness use coefficients or models of the inner ear that were obtained from extensive subjective experiments. If multiple such features are used, their respective contributions to intrusiveness must be determined from further training data. In contrast, the proposed approach uses two main parameters (the dictionary of auditory kernels and the spike energy threshold) that are largely derived from physiological measurements in mammals. Our evaluation in Section 4.6.2, page 56, also shows that the exact parameter values are not critical for prediction performance.

A limitation is the higher computational complexity for sparse approximation, even with an efficient implementation of Matching Pursuit [MPTK, Krstulovic and Gribonval, 2006]. This can be an issue for an objective measure targeted at telecommunications, where quality assessment algorithms are sometimes used on low-power mobile platforms. The experiments of Section 4.3 as well as the following evaluations also only use sparse representations with a limited frequency range (50–7150 Hz). This ignores possible high-frequency noise components in super-wideband telecommunication systems, but these components tend to be very weak in environmental noises as considered here [De Coensel et al., 2003]. Finally, the analysis of noise in speech pause sections and assumption of a standard speech level can be seen as further limitations, although they are not specific to the proposed approach.

### 4.4.2 Threshold selection

The value of $\|\alpha_{\min}\|^2$ that is used as stopping criterion for the sparse approximation can be selected through a simple experiment with a speech signal. Here we use a 27-second recording of concatenated sentences from 4 speakers, scaled to the standard average level of 79 dB SPL. The peak spike energy for speech at this level is estimated, and the threshold value is set 40 dB below, as previously motivated by the analogy to the dynamic range of auditory nerve fibers. Specifically, we compute a sparse approximation of this speech signal where the number of spikes multiplied by the support of the longest kernel (128 ms) equals the signal length. Since MP selects the highest-energy spikes first, this approximation contains the peak spikes energies for the signal, which have a median of 11.6 Pa$^2$. The spike energy threshold is thus set to $\|\alpha_{\min}\|^2 = 1.16 \cdot 10^{-3}$ Pa$^2$.

Note that this simple approach does not account for the actual dynamic behavior of auditory neurons [Moore, 2003], and other ways of selecting $\|\alpha_{\min}\|^2$ could be imagined. However, we will see in Section 4.6.2, page 56, that the performance of the proposed feature remains quite stable with other threshold values.

## 4.5 Experimental setup

We evaluate the proposed feature on the PANDA datasets that were introduced in Chapter 3. Other studies on objective noise intrusiveness assessment used unavailable (proprietary) data [Gautier-Turbin and Le Faucheur, 2005; Reimes et al., 2011], while Narwaria et al. [2012] used the NOIZEUS dataset developed in [Hu and Loizou, 2007b]. However, that dataset focused on speech enhancement instead of noise perception. Consequently, the recordings in it have very short leading and trailing pauses (~0.15 s), making it unsuitable for the features compared here, which analyze noise in speech pause sections.

The temporal structure and rating order of recordings in the PANDA datasets were shown in Figures 3.2 and 3.3, respectively. In particular, listeners used triplets of test speech recordings with different sentences, but identical speaker and condition, to provide the three P.835 quality ratings. Since noise intrusiveness is rated either after the first or the second recording, we evaluate the objective measure on the first two sentences in each triplet.

An overview of conditions was given in Table 3.1, page 29, with complete test plans in the appendix, Section A.1.3, page 101. For the evaluation of noise intrusiveness measures, we focus on the background noise conditions in each set, leaving 24, 23 and 24 conditions, respectively. Each condition includes 6-second recordings from four speakers, resulting in a total of $(24+23+24)$ conditions ×4 speakers ×2 sentences = 568 test speech recordings.

### 4.5.1 Intrusiveness features

Given the noise-corrupted test speech recordings and ground truth subjective intrusiveness scores from the dataset, we compare the prediction performance of the spike density feature proposed in Section 4.4 to the traditional, main features discussed in Section 4.1.

As a common preprocessing step to all features, we extract the background noise signal from speech pause sections in each recording, sampled at 48 kHz. Pause sections can be detected through voice activity detection (VAD), but for this experiment we take advantage of the availability of reference signals. Specifically, we first identify speech pauses in the clean reference through VAD, and then find the corresponding locations in the noisy test speech recording with a temporal alignment step [Beerends et al., 2013a, Sec. 4.3.1]. This procedure allows for a more reliable detection of pause sections, especially in case of high noise levels. The detected leading and trailing pause sections are concatenated to a single noise signal $x[n]$, using overlapping 64 ms half-Hann windows at the joint.

Given the noise signal $x[n]$, we evaluate the following features:

*Mean noise level [dB SPL]* — The average sound pressure level of noise in Pa is given by

$$10\log_{10}\left(\frac{1}{N}\sum_{n=1}^{N}\left(\frac{x[n]}{p_0}\right)^2\right) \tag{4.5}$$

with $p_0 = 20$ µPa the reference sound pressure in air.

*Mean noise level [dB(A) SPL]* — This feature applies a pre-filtering with the "A" weighting curve defined in [IEC 61672-1, 2013] before computing the noise level in (4.5).

*Loudness [sone]* — We use a loudness model for temporally variable sounds [Zwicker, 1977] as implemented in the Loudness Toolbox [GENESIS S.A., 2012]. The model estimates the instantaneous loudness of noise over time, which can be averaged to obtain its mean loudness. The short-time estimates may also be aggregated by computing the fifth percentile, as described in Section 4.4. This method was proposed by Fastl and Zwicker [2007, Chap. 16.1] to better model the perceived overall loudness of non-stationary noise. Both types of aggregation are compared in the following section as two separate features.

*Spike density [spikes/s]* — The proposed feature as described in Section 4.4. Due to the restricted frequency range covered by the dictionary, the noise signal is resampled to 16 kHz and band-limited to 7150 Hz for this feature. The spike density in units of spikes/s can either represent an average for the signal or a percentile of short-time spike counts. As with loudness, we evaluate both aggregation types, computing the percentile in the same manner, using 25 ms intervals with 50% overlap.

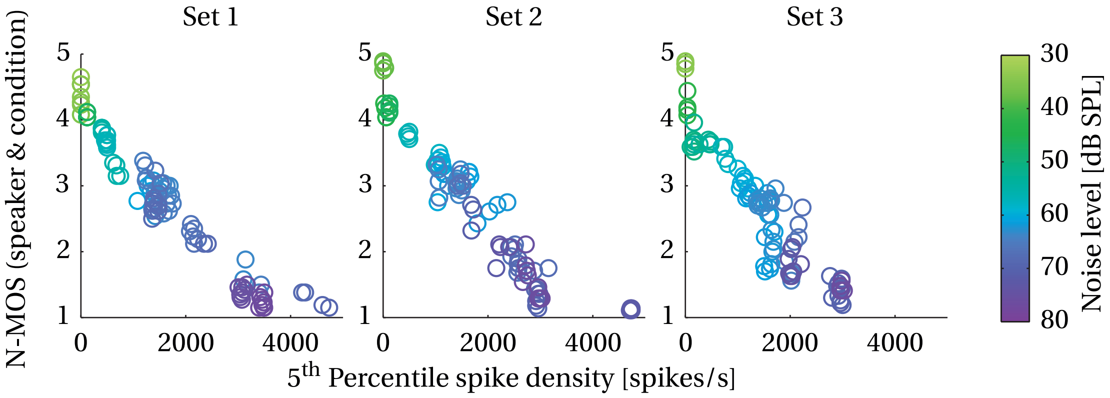


Figure 4.5 – Relation between the proposed feature (5th percentile spike density) and perceived noise intrusiveness. Data points show results per speaker (four speakers per condition).

Table 4.1 – Prediction performance of per-condition noise intrusiveness scores on the three "PANDA" datasets. Best performances are highlighted in boldface. Double asterisks ($**$) denote prediction errors that are highly significantly lower than those of *all other features* in a given set ($p < 0.01$ with Holm-Bonferroni correction).

| Feature | Correlation $\|R\|$ | | | Prediction error $\text{rmse}^*_{3\text{rd}}$ [MOS] | | |
|---|---|---|---|---|---|---|
| | Set 1 | Set 2 | Set 3 | Set 1 | Set 2 | Set 3 |
| Mean noise level [dB SPL] | 0.881 | 0.900 | 0.919 | 0.339 | 0.359 | 0.279 |
| Mean noise level [dB(A) SPL] | 0.930 | 0.926 | **0.942** | 0.230 | 0.277 | 0.234 |
| Mean loudness [sone] | 0.932 | 0.951 | 0.925 | 0.257 | 0.206 | **0.197** |
| Mean spike density [spikes/s] | 0.930 | 0.916 | 0.887 | 0.250 | 0.242 | 0.270 |
| 5th Percentile loudness [sone] | 0.959 | **0.953** | 0.911 | 0.191 | 0.234 | 0.270 |
| 5th Percentile spike density [spikes/s] | **0.970** | **0.953** | 0.939 | **0.087**** | **0.117**** | 0.231 |

## 4.6 Results

Figure 4.5 shows the relation between subjective scores and the values of the 5th percentile spike density feature. Each data point represents the noise intrusiveness mean opinion score (N-MOS) of listeners and the feature value, respectively, for a given speaker and condition. The proposed feature has a strong linear relation to the N-MOS in all three sets. Additionally, data points in the figure are shaded by the mean noise level. It can be seen that the noise level can predict the low intrusiveness of very weak noises (bright shading), but not the subjective rank-order of stronger noises (irregular progression to darker shading).

### 4.6.1 Prediction performance

Table 4.1 shows performance metrics for the different features described in Section 4.5.1. For Sets 1 and 2, the proposed 5th percentile spike density feature (bottom row) achieves the highest prediction performance in terms of both performance metrics introduced in Section 2.3. For Set 3, its performance is the second-highest. For Sets 1 and 2, the prediction error of the proposed feature is highly significantly lower than that of all other compared features ($p < 0.01$ with Holm-Bonferroni correction). For Set 3, there is no significant difference between prediction errors ($p > 0.05$).

#### Effect of percentile aggregation

Without 5th percentile temporal aggregation, loudness and spike density perform comparably, i.e., the prediction errors of mean loudness and mean spike density do not significantly differ ($p > 0.05$). Using 5th percentile aggregation does not significantly change the prediction error of loudness, whereas the spike density feature improves highly significantly for two of the three sets. The sparse representation preserves the temporal localization of peak noise intensities, while loudness may smear these peaks over time to model temporal masking [Zwicker, 1977]. This may explain why percentile aggregation is more beneficial with the spike density feature, even though it was conceived for loudness [Fastl and Zwicker, 2007, Chap. 16.1].

We also tested percentile values $k$ in the range $k = 2, \dots, 10$ as sometimes found in the literature [e.g., Axelsson et al., 2010]. Our tests showed that at all these values, prediction errors of the percentile spike density feature remain either lower or comparable ($p > 0.05$) to those of percentile loudness. With the exception of Set 2 at $k = 2$ and $k = 3$, its error also remains lower or comparable ($p > 0.05$) to that of *all* other features.

In summary, the proposed feature can be used to assess noise intrusiveness with a prediction error that is comparable, and in most cases lower, than that of traditionally used features. This result, in addition to the experiments of Section 4.3, further confirms our hypothesis that a high-level model of sensory coding can be used to assess noise perception.

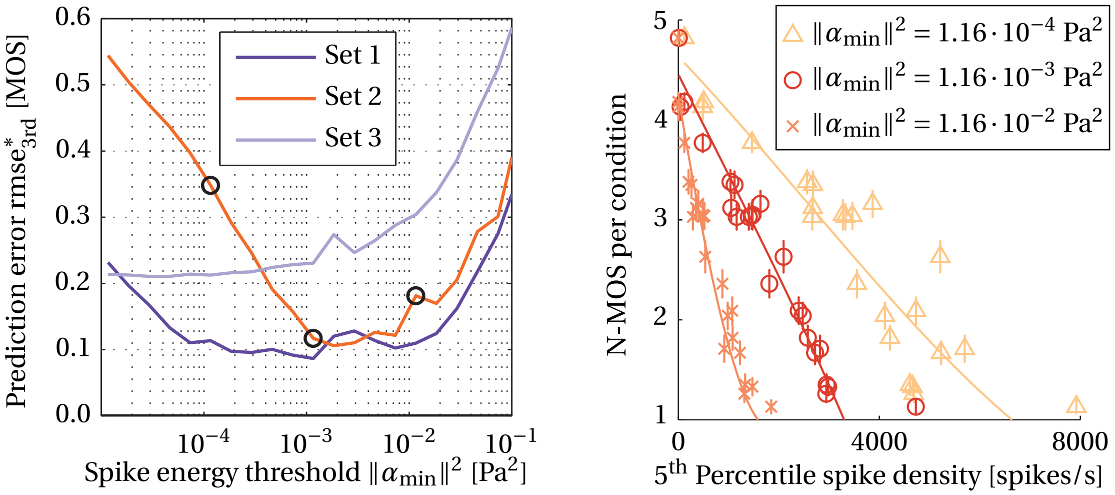
### 4.6.2 Effect of hyperparameters

A key motivation for the proposed approach is that it only uses two main parameters — the spike energy threshold $\|\alpha_{\min}\|^2$ and the dictionary of auditory kernels — that can be largely derived without subjective experiments. We now evaluate the effect of both hyperparameters:

*Spike energy threshold* $\|\alpha_{\min}\|^2$ — This parameter determines the level of noise structures to include in the sparse representation. A way of selecting $\|\alpha_{\min}\|^2$ was described in Section 4.4.2. The effect of changing its value on prediction performance is shown in Figure 4.6a for the three evaluation datasets. The error remains fairly stable within two orders of magnitude around the original threshold value of $1.16 \cdot 10^{-3}$ Pa$^2$. Higher thresholds slowly increase the error in all sets, whereas lower values have less effect, with the exception of Set 2.

Figure 4.6b provides a closer look at results per condition in Set 2. A high threshold excludes low-intensity noise components from the sparse approximation. The intrusiveness of weaker noises is therefore no longer differentiated (orange crosses clustered near the ordinate of Figure 4.6b). Conversely, low thresholds include even very low-level noise components into the approximation. These components increase the percentile spike density, but the increase is no longer proportional to subjective intrusiveness (yellow triangles in Figure 4.6b).

This latter effect may be due to the greedy nature of the Matching Pursuit algorithm. Once many kernels have been extracted from the signal, further iterations will mainly correct distortions in the residual that are due to earlier iterations, and not capture true signal structure [Sturm et al., 2009]. Another explanation is that very low-energy noise components are not perceived by listeners, i.e., that they are *masked* by higher-energy components in the signal.



(a) Performance for different values of $\|\alpha_{\min}\|^2$.

(b) Visualization of the effect of $\|\alpha_{\min}\|^2$ on Set 2.

Figure 4.6 – Influence of the spike energy threshold $\|\alpha_{\min}\|^2$. **(a)** Prediction errors for different threshold values. The threshold for the results in Table 4.1 is $\|\alpha_{\min}\|^2 = 1.16\cdot 10^{-3}$ Pa$^2$ and was derived in Section 4.4.2. **(b)** Results for Set 2 for the threshold values circled in Panel (a). Data points show results per condition. Error bars indicate 95% confidence intervals of subjective scores. Lines show the mapping function used to calculate the prediction error rmse$^*_{3\text{rd}}$.

*Dictionary of auditory kernels* — We used a dictionary of auditorily motivated gammatone kernels with center frequencies as described in Section 4.3 to compute the sparse approximation. While the gammatone shape can be derived from physiological measurements, the distribution of center frequencies and bandwidths followed the data of subjective experiments, i.e., the Equivalent Rectangular Bandwidth (ERB) scale [Glasberg and Moore, 1990]. We therefore evaluate two alternative dictionaries with different parameters.

The first alternative dictionary contains gammatone kernels with *logarithmically* distributed center frequencies within the same range, with bandwidths $B^m$ proportional to center frequency (i.e., as in constant-Q filters),

$$B^m = \max\left(f^m/10\,,\,25\right) \quad \text{[Hz]} \tag{4.6}$$

where the minimum bandwidth of 25 Hz was chosen to avoid overly long kernels at low center frequencies.[5]

The second alternative dictionary uses the same logarithmic center frequency distribution and bandwidths $B^m$, and further replaces gammatone kernels by even Gabor kernels $\phi^m_g(t)$,

$$\phi^m_g(t) = \frac{1}{\sqrt{2\pi}\sigma^m}\exp\left(-\frac{t^2}{2\,(\sigma^m)^2}\right)\cos\left(2\pi f^m t + \varphi\right), \quad t \lesseqgtr 0 \tag{4.7}$$

[5] The calculation of the parameter $b^m$ in the gammatone equation (4.2) for a target bandwidth $B^m$ is given in [Holdsworth et al., 1988].

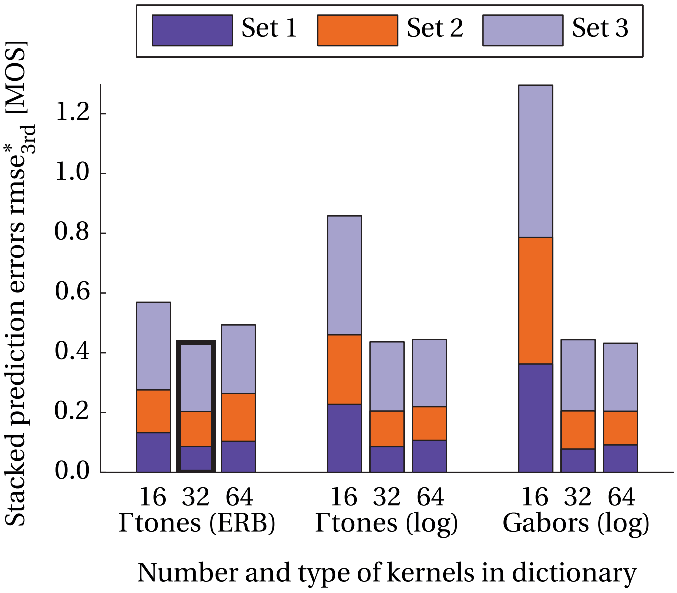


Figure 4.7 – Influence of the dictionary $\boldsymbol{\Phi}$ on prediction errors. "Γtones (ERB)" are the gammatone kernels of Section 4.2.2. "Γtones (Log)" and "Gabors (Log)" have logarithmically distributed center frequencies with fixed center frequency-to-bandwidth ratio, using gammatone and Gabor kernels, respectively. The thick bar represents the original dictionary of Section 4.3.

with $\sigma^m = \sqrt{\ln 2 / \pi B^m}$ the bandwidth parameter. For each dictionary, we also evaluate the effect of the number of kernels $M$. The threshold value $\|\alpha_{\min}\|^2$ is re-determined for each dictionary as per the method described in Section 4.4.2.

Figure 4.7 compares the prediction errors obtained with different dictionary types and sizes. For dictionaries with $M = 32$ or 64 different kernels, there is no significant difference in prediction errors ($p > 0.05$). However, when the dictionary only consists of $M = 16$ kernels, significant differences ($p < 0.05$ with Holm-Bonferroni correction) appear between the original and alternative dictionaries for at least two of the sets.

With $M = 16$ kernels, the frequency space is tiled less densely, so noise approximations up to the threshold $\|\alpha_{\min}\|^2$ are considerably less sparse. Further analysis shows that the decrease in sparsity is most pronounced for the "Pub" (babble) noise. The intrusiveness of this noise type is therefore over-estimated with the alternative dictionaries, but not with the "ERB" gammatone dictionary that matches cochlear filter shapes more closely. This result is in line with that of Smith and Lewicki [2006], who found that the center frequency and bandwidth characteristics of cochlear filters are highly adapted to the acoustic composition of speech.

Overall, these results show that the performance of the proposed feature does not critically depend on the original parameter values. Specifically, prediction errors only increase slowly around the original threshold $\|\alpha_{\min}\|^2$, and dictionaries that avoid knowledge from subjective experiments yield similar error, as long as they allow for a sparse representation of noise.

## 4.7 Discussion and conclusion

In this chapter, we have studied the objective assessment of perceived background noise intrusiveness in telecommunications. A challenge in assessing noise intrusiveness is that the degradation of interest is not the distortion of the speech signal, but the presence of noise. The degradation is thus an added signal element that cannot be assessed through a comparison to a reference signal. Instead, one or more features of noise that are relevant to its perception need to be found. However, predicting intrusiveness from these features should require as little subjective training data as possible, due to the expense in collecting such data.

We took insight from recent work on modeling sensory coding to propose a novel approach, where background noise is analyzed with a sparse coding signal model. We have tested the hypothesis that such a noise representation, computed in a basis of auditory kernels, is indicative of its perception, and verified this hypothesis in Section 4.3. Specifically, the number of kernels in the representation scales logarithmically with sound pressure, differs by noise type, follows the trend of frequency weighting curves for noise, and increases for wider bandwidth noises. The number of kernels over time in the representation can then be analyzed to yield a feature that is highly correlated with noise intrusiveness scores.

Our evaluation on the PANDA datasets introduced in Chapter 3 shows that the proposed feature predicts subjective scores with an error 14–54% below that of percentile loudness. Moreover, the prediction error is either significantly lower or comparable to that of all other evaluated features. Note that the PANDA datasets were collected before, and independently of the proposed feature.

An advantage of the proposed approach is that it uses few hyperparameters, which can be determined without training data or subjective experiments. The obtained prediction error remains low even when these parameters deviate from the proposed original values. This is in contrast to conventional methods, which perform a regression with multiple features, or use features that rely on parameters derived from subjective experiments. These latter features represent acoustic *correlates* of intrusiveness, whereas our approach seeks to model perception at the higher level of sensory coding.

The sparsity of noise in a basis of auditory kernels has no direct physical or perceptual interpretation. When the kernels resemble cochlear filter shapes, each kernel occurrence in the representation can be thought of a local population of auditory nerve spikes [Lewicki, 2002]. The spike rate is commonly thought to be related to loudness, although the precise relationship appears to be more complex [Moore, 2003]. This may provide one indication of how the proposed feature is linked to human perception. On the other hand, our experiments in Section 4.6.2 have shown that prediction performance remains high even with kernels that deviate from cochlear filter shapes, as long as they allow for a sparse representation of the signal. This result is more consistent with the notion of sparsity as a general abstraction of sensory coding, which need not be tied to the cochlea specifically.

# 5 Objective intelligibility assessment for speech telecommunications

The last two chapters discussed the perception of background noise in speech telecommunications. An objective measure was proposed to predict the intrusiveness of noise during non-speech parts. An analysis of subjective scores further showed the mutual dependency between overall listening quality, speech distortion and noise intrusiveness.

An important quality feature that is not addressed with these scores is the *intelligibility* of degraded speech. For example, Hu and Loizou [2007a] evaluated different noise reduction algorithms and found that none of them significantly improved the intelligibility of noisy speech. More importantly, algorithms that resulted in the best overall quality were not the ones that best preserved intelligibility. The three quality features in P.835 may thus be an insufficient characterization of speech quality when intelligibility is key, such as in telecommunication systems for first responders (i.e., firefighters, medical or law enforcement personnel).

Speech intelligibility can be affected by interfering background noise, but also by distortions due to noise reduction, speech coding, channel loss or acoustical impairments. Objective intelligibility measures (OIMs) have often been designed to assess the impact of specific degradations. In the following, we will concentrate on OIMs that are applicable to degradations found in telecommunications. This leaves out some measures that rather focus on the intelligibility of assistive listening devices [see, e.g., Falk, Parsa, et al., 2015, for a review].

In this chapter, we propose a novel objective intelligibility measure for degradations found in telecommunications. Unlike noise intrusiveness, where listeners have no a-priori knowledge or expectation of the noise signal, our approach is based on the notion that listeners apply linguistic knowledge in understanding speech. The proposed approach thus makes use of a language-specific representation to analyze speech beyond the signal level. The remainder of this chapter is structured as follows: We briefly review existing approaches to objective intelligibility assessment in Section 5.1 and highlight their strengths and weaknesses. Section 5.2 explains the motivation for our proposed approach, which is described in Section 5.3. Implementation details and the experimental setup are reported in Section 5.4. We discuss results in Section 5.5, and conclude with further remarks in Section 5.6.

## 5.1 Related work and contributions

One of the first published OIMs for telecommunications was the articulation index (AI) [French and Steinberg, 1947; simplified by Kryter, 1962]. The (simplified) AI is obtained by computing the average signal-to-noise ratio across critical bands $k$:

$$\mathrm{AI} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{AI}_k \qquad \mathrm{AI}_k = \frac{1}{30}\left(6 + 10\log_{10}\left(\frac{|X(k)|^2}{|N(k)|^2}\right)\right), \quad \mathrm{AI}_k \in [0,1] \tag{5.1}$$

with $|X(k)|^2$ and $|N(k)|^2$ the speech and noise spectral power in band $k$, respectively. The resulting score has a monotonic relation to the intelligibility of speech degraded by additive stationary noise and bandpass filtering. A revised version of the AI, which adds corrections for presentation level and the auditory threshold (among others), is standardized as speech intelligibility index [SII; ANSI S3.5, 1997].

The AI is a *macroscopic* measure, in that it is usually computed for the average long-term spectrum of speech from multiple speakers and sentences. If the short-time spectral powers of speech and noise are known, intelligibility may also be predicted at the *microscopic* level by measuring the occurrence of so-called "glimpses", time-frequency regions in which the speech energy exceeds that of noise by a given threshold [Cooke, 2006].

In test signals from telecommunication systems, separate estimates of speech and noise powers are usually not available. Instead, a full-reference method like the speech transmission index [STI; Steeneken and Houtgast, 1980] may be used to apply a specific input signal (the reference) to the system, and compare it to the resulting test signal. STI uses a modulated noise with spectral distribution similar to that of speech as reference, and analyzes the modulation transfer function to the test signal to derive an intelligibility score. This approach has been shown to be suitable to assess the impact of additive noise, reverberation, bandpass filtering and waveform coding on speech intelligibility [Steeneken and Houtgast, 2002]. Like the AI, STI is a macroscopic measure, although an extension to evaluate the STI over short-time segments has recently been proposed [Schwerin and Paliwal, 2014].

Modern telecommunication systems often include signal-dependent processing, such as digital speech codecs that are based on a source-filter model of speech [see e.g., Rabiner and Schafer, 2007, chap. 7.2], or noise reduction that applies a time- and frequency-varying signal gain. Testing these components with the modulated noise signal of the STI may not reflect their effect on speech signals. Consequently, other approaches have been devised to assess intelligibility with actual speech signals processed by the system under test. In that line, Beerends, Van Buuren, et al. [2009] proposed a modification of the PESQ *overall quality* measure [Beerends, Hekstra, et al., 2002], where an auditory spectral representation of the test signal is compared to that of the transmitted reference. Full-reference approaches based on a comparison of spectral features (e.g., short-time auditory spectra or band envelopes) have also been proposed by several other authors [Elhilali et al., 2003; Christiansen et al., 2010; Taal et al., 2011; Voran, 2013].

A limitation of using spectral features for full-reference assessment is that such features may be sensitive to differences in speech tempo and timbre between the reference and test signal. Changes to speech tempo can occur in packet-switched telecommunication systems such as VoIP, which may stretch or compress parts of the signal to compensate for delayed packets. Similarly, it has been proposed to use speech synthesis principles in codecs to achieve lower bit rates [e.g., Lee and Cox, 2001], meaning that the voice in the decoded signal may bear no resemblance to the original voice. While differences in speech tempo can be compensated through a prior alignment step [e.g., Sakoe and Chiba, 1978; Rix, Hollier, et al., 2002], a different speech timbre may change the spectral feature values themselves.

Therefore, recent approaches have been proposed that go beyond the spectral level and seek to assess intelligibility at *phone*, *phoneme* or *word* level. For instance, Teng et al. [2007] compared occurrences of phone bigrams, determined with an ASR (automatic speech recognition) system in reference and test speech, to assess the impact of low bit rate codecs and bit error conditions on intelligibility. Phonetic features were also used to assess pathological speech, by estimating confidence scores over phone segments [Middag et al., 2008]. Meyer and Kollmeier [2010] used *phoneme-level* features in an ASR system to compare word error rates to intelligibility scores for noisy speech. Finally, Maier et al. [2007] compared pathological speech to a reference *word-level* transcription, using a complete ASR system.

In summary, most OIMs that were designed for degradations found in telecommunications use either acoustic or spectral features. On the other hand, phone- or phoneme-level features have been found to be suitable to assess speech with strongly distorted or unnatural timbre, as may be the case with low bit-rate coded, pathological or even synthetic speech. In this chapter, we expand on phoneme-level feature-based approaches to intelligibility assessment. Specifically, we show how phoneme-level features may be used to compute a distance measure between a human speech reference and the output of a system under test, which may degrade the speech signal through additive noise, noise reduction and low bit-rate speech coding. In Chapter 6, page 77, we extend these results to the assessment of synthetic speech, which can differ in both tempo and timbre from the reference.

Contribution

R. Ullmann, M. Magimai.-Doss, and H. Bourlard [2015]. “Objective Speech Intelligibility Assessment through Comparison of Phoneme Class Conditional Probability Sequences”. In: *Proc. ICASSP*. Brisbane, Australia, pp. 4924–4928.
URL: http://dx.doi.org/10.1109/ICASSP.2015.7178907

The present chapter expands on the above paper, adding results for the PSCR dataset (see Section 2.4.2, page 13), a comparison to results obtained with the STOI measure, and further analysis of model hyperparameters. Results from the above paper for synthetic speech are included in Chapter 6, page 77.

## 5.2 Motivation

The core idea in our work is that intelligibility impairments can be seen as cases of mismatch between the test signal acoustics and listener's phonetic and lexical knowledge. In this line of thought, listeners apply lexical knowledge to match phoneme sequences against known words, or to resolve word confusions through context. Subjective tests are often designed to limit the application of lexical knowledge, e.g., by focusing on the recognition of individual words (as used in rhyme tests) or of semantically unpredictable sentences to reduce context. The closed-response format of rhyme tests, where listeners select a target word from several alternatives (see Section 2.2.2, page 7), further reduces the use of lexical knowledge.

Intelligibility impairments that are observed in such tests can thus be largely attributed to a mismatch with phonetic knowledge, i.e., listener's (in-)ability to recognize speech sounds in the test signal. Like lexical knowledge, phonetic knowledge is language-dependent in the sense that listeners learn to differentiate sounds that carry different meaning in their language [Kuhl et al., 1992]. This set of sounds is described by phonemes [see, e.g., Gold et al., 2011, Chap. 23.2.3]. Therefore, it may be possible to objectively assess intelligibility by analyzing the test signal at the phoneme level and measuring the mismatch to a reference phoneme-level representation. Such a reference may be obtained from a highly intelligible recording of the same words, or from a model of the expected phoneme-level content for those words.

The approach proposed in this chapter uses a reference recording and is inspired from recent results in template-based ASR. In this type of ASR, a speech utterance is recognized by comparing it to several example recordings or *templates* of possible target utterances. Soldo, Magimai.-Doss, and Bourlard [2012] recently studied the use of synthetic speech templates, i.e., reference recordings generated with a text-to-speech (TTS) system, with phoneme posterior probabilities as features. They observed that such templates could yield recognition performance comparable to templates of natural speech, indicating that the features were insensitive to speaker characteristics and naturalness. They also found that recognition performance correlated with the *intelligibility* of the synthetic voices used for template generation.

Motivated by these results, we investigate an approach to intelligibility assessment based on the comparison of phoneme posterior probability sequences of the test signal to those of a reference signal. Note that in this approach, the speaker in the reference signal need not be the same as in the test signal. This is particularly relevant to the assessment of very low bit-rate speech codecs, which may use synthetic speech principles and yield test signals with changed speaker characteristics.

## 5.3 Proposed approach

Given a reference speech signal and a test speech signal, the approach performs the steps outlined in Figure 5.1, which consist in:

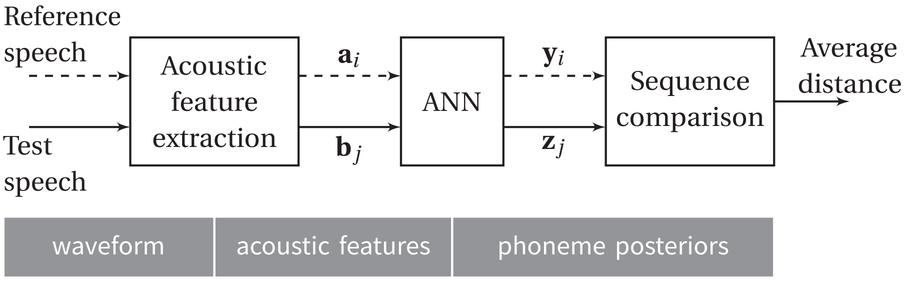


Figure 5.1 – Diagram of the proposed objective intelligibility measure. Phoneme posterior probabilities are estimated with an artificial neural network (ANN). The type of signal feature in each stage of the proposed approach is highlighted in gray at the bottom.

*Acoustic feature extraction* — Two acoustic feature sequences $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_i, \ldots, \mathbf{a}_I\}$ and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_j, \ldots, \mathbf{b}_J\}$ are extracted from the reference and test signal, respectively. The features can be, e.g., cepstral coefficients of a short-time spectrum. Note that the two sequences need not be of the same length, i.e., $I \lesseqgtr J$.

*Posterior probability estimation* — Estimation of the reference phoneme posterior probability sequence $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_I\}$ and test phoneme posterior probability sequence $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_j, \ldots, \mathbf{z}_J\}$, where

$$\mathbf{y}_i = \left[P\left(c^1 \mid \mathbf{a}_i\right), \ldots, P\left(c^K \mid \mathbf{a}_i\right)\right]^\top = \left[y_i^1, \ldots, y_i^K\right]^\top, \tag{5.2}$$

$$\mathbf{z}_j = \left[P\left(c^1 \mid \mathbf{b}_j\right), \ldots, P\left(c^K \mid \mathbf{b}_j\right)\right]^\top = \left[z_j^1, \ldots, z_j^K\right]^\top, \tag{5.3}$$

with $\sum_k y_i^k = \sum_k z_j^k = 1$, and $c^k$ the $k^{\text{th}}$ phoneme class out of $k \in [1, K]$ phoneme classes.

The distributions $\mathbf{y}_i$ and $\mathbf{z}_j$ are obtained through an artificial neural network (ANN) trained to estimate phoneme posterior probabilities. The inputs to the ANN include acoustic features from multiple frames around the current frame $i$ or $j$ in order to consider temporal context for the estimation of phoneme content. ANNs trained with spectral-based features have been shown to learn properties of the spectral envelope [Pinto et al., 2009], i.e., speech properties that are relevant to phoneme discrimination, but show little variability across speakers. In this chapter, we will use multilayer perceptrons (MLPs), a type of feedforward ANNs with three or more layers of nodes. We will evaluate both three- and five-layer architectures in Section 5.5.3.

*Distance calculation* — The sequences $Y$ and $Z$ are compared to calculate a distance score. There are several local distance measures that can be used to compare two phoneme posterior distributions $\mathbf{y}_i$ and $\mathbf{z}_j$ in the sequences [see, e.g., Soldo, Magimai.-Doss, Pinto, et al., 2011].

Here we use the symmetric Kullback-Leibler (SKL) divergence as local distance,

$$\mathrm{SKL}\left(\mathbf{y}_i,\mathbf{z}_j\right)=\frac{1}{2}\sum_{k=1}^{K}y_i^k\log_2\frac{y_i^k}{z_j^k}+\frac{1}{2}\sum_{k=1}^{K}z_j^k\log_2\frac{z_j^k}{y_i^k}\quad. \tag{5.4}$$

A measure based on Kullback-Leibler divergence allows for an information theoretic interpretation of the distance. Specifically, a delta posterior distribution $\mathbf{y}=\delta_{lk}$ (i.e., $y^k=1$ and $y^l=0\,\forall\, l\neq k$) has no uncertainty regarding the actual phoneme class $c^k$. Conversely, a uniform distribution $y^k=\frac{1}{K}\;\forall\; k$ provides no reduction in uncertainty. The Kullback-Leibler (KL) divergence is an asymmetric measure of the *increase in uncertainty* that comes from assuming a different distribution than the actual distribution [Cover and Thomas, 1991]. In the proposed approach, the symmetric KL divergence is used as local distance in (5.4) because the reference and test recordings can be different realizations of the same words, i.e., there is no guarantee of which recording is more intelligible or represents the "actual" distribution.

The final objective intelligibility score is derived from the global distance between the sequences $Y$ and $Z$. This distance may be calculated in two ways depending on the test signal:

- If test and reference signals have the same temporal structure, the sequences have equal lengths $I=J$ and the average of local distances can be used as intelligibility measure,

$$D(Y,Z)=\frac{1}{J}\sum_{j=1}^{J}\mathrm{SKL}\left(\mathbf{y}_j,\mathbf{z}_j\right)\ . \tag{5.5}$$

- If both signals have different temporal structure, the sequences $Y$ and $Z$ can be aligned through dynamic time warping (DTW) [Sakoe and Chiba, 1978] with path constraints

$$C(i,j)=\mathrm{SKL}\left(\mathbf{y}_i,\mathbf{z}_j\right)+\min\left(C(i,j-1),C(i-1,j-1),C(i-2,j-1)\right) \tag{5.6}$$

  where $C(i,j)$ is the accumulated distance at reference and test time frames $i$ and $j$, respectively. However, no overall constraints are applied. The global distance score is then given by the average distance over the DTW path, i.e.,

$$D(Y,Z)_{\mathrm{DTW}}=\frac{1}{J}C\left(I,J\right)\ . \tag{5.7}$$

## 5.4 Experimental setup

We first test the proposed approach with various low bit-rate coding and bit error/frame loss conditions. The purpose of this initial experiment is to check for any systematic offsets between different speech coding schemes. We then perform a more formal evaluation on the PSCR library (see Section 2.4.2, page 13), which includes recordings degraded by low bit-rate coding, strong background noises and acoustic impairments.

### 5.4.1 Low bit-rate coding and frame loss conditions

As an initial experiment, the proposed approach is tested on the following conditions:

- AMR cellular telecommunications codec [ETSI TS 126 090, 2012], running at the codec's eight different constant bit rates (4.75–12.2 kbps),
- EVRC-B cellular telecommunications codec [3GPP2 C.S0014-E, 2011] running at the codec's standard average bit rates (4.8–9.6 kbps),
- MELP US DoD codec [Supplee et al., 1997] in simple, double and triple cascaded setups (2.4 kbps),
- codec2 free open-source codec [Rowe, 1997; Rowe and contributors, 2013] operating at 2.4 kbps, with bit error rates of 0.0, 0.2, 0.5, 1 and 5%, and
- simulated frame loss (5, 10, 20 and 40%), by silencing randomly selected 20 ms frames of active speech segments.

Each condition is applied to 12 recordings of phonetically balanced, English sentences from 12 speakers (six male) provided in ITU-T Rec. P.501 [2012]. Recordings are 2–3 seconds long, and were pre-filtered with the IRS-send telephone bandpass [ITU-T Rec. P.48, 1988] prior to processing.

### 5.4.2 Public safety communications conditions

The Public Safety Communications Research [PSCR, 2013] audio library consists of three datasets (2008, 2010 and 2012) of speech recordings and was presented in Section 2.4.2, page 13. Each recording contains a carrier phrase with a consonant-vowel-consonant (CVC) word at its end, to be recognized by listeners from a choice list of six rhyming words.

Given that listeners' scores only reflect the intelligibility of the final word, the proposed approach is applied to that part of the sentence only. Specifically, the phoneme posterior sequence is initially estimated for the entire recording, allowing to consider temporal context at the ANN input. The sequence is then truncated to only keep frames for the final word, using the sample-accurate start and end word positions that are provided with the distribution.[1] The distribution provides these positions for the reference recordings only; thus we used a temporal alignment step to determine corresponding positions in the test speech recordings.[2] The total number of test speech recordings is 129 600, corresponding to 17.3 hours of speech.

[1] The PSCR distribution provides time-frequency templates of extracted rhyme words. The determination of exact sample positions from the templates was carried out by Mr. Nicolas Gninenko during his internship at Idiap.

[2] Signal delays are near-constant within groups of 150 recordings per condition [Voran and Catellier, personal communication, 25 Aug., 2015], allowing reliable delay estimation even in case of strong signal degradations.

### 5.4.3 Implementation

We extract Perceptual Linear Predictive (PLP) acoustic features [Hermansky, 1990] from the reference and test signals, i.e., cepstral coefficients derived from a short-time spectrum with frequency resolution and power weighting modeled after human hearing characteristics. PLP calculation involves an approximation of the short-time spectrum with an all-pole model that is designed to remove speaker-dependent fine spectral structure (see Section 2.5.1, page 17).

We compute 39-dimensional PLP cepstral coefficients (the first 12 cepstral coefficients plus energy, with appended delta and acceleration coefficients [Furui, 1986]), using a 25 ms frame size with 10 ms frame shift. The PLP analysis is based on 24 mel band energies in the 125–3800 Hz range.

The resulting acoustic features $\mathbf{a}_i$ or $\mathbf{b}_j$ are fed at the input of an ANN with a 9-frame temporal context (i.e., 4 preceding and 4 following frames), resulting in $9 \times 39 = 351$ input units. The ANN used here is the same multilayer perceptron (MLP) used in the studies in [Soldo et al., 2011, 2012], with a single hidden layer of 5000 units and 45 output units. The MLP is trained on 232 hours of conversational telephone speech (CTS)[3] from over 4000 speakers, with additional 36.3 hours for cross-validation, to estimate posterior probability vectors $\mathbf{y}_i$ or $\mathbf{z}_j$ for 44 English phonemes and silence. MLP training was carried out with the QuickNet toolkit [Johnson and contributors, 2011] by minimizing frame-level cross entropy. General details of MLP training and estimation of posteriors are given in Section 2.5.3, page 18.

Since all evaluated conditions preserve the temporal structure of speech, we use the global distance $D(Y, Z)$ as defined in (5.5) for intelligibility assessment, i.e., without DTW alignment.

### 5.4.4 Comparison to a spectral feature-based measure

We use the short-time objective intelligibility (STOI) measure of Taal et al. [2011] to compare our proposed approach to an objective measure based on spectral features. STOI analyzes the short-time magnitudes of signals within one-third octave bands, and computes the correlations between reference and test signal band envelopes over segments of 384 ms. The individual correlation values are then averaged over bands and segments to yield an objective score in the range $[0, 1]$ that is expected to have a monotonic relation to intelligibility. This approach was found to provide good predictions of subjective intelligibility scores for speech degraded by additive noise, noise reduction, bandpass filtering, and speech coding as used in cellular networks [Taal et al., 2011; Jørgensen et al., 2015]. We used the publicly available MATLAB implementation of [Taal et al., 2010] for our experiments.

[3] The CTS data consists of recordings from the Switchboard-1 [Godfrey and Holliman, 1993], Callhome [Canavan et al., 1997] and Switchboard Cellular [Graff et al., 2001] corpora.

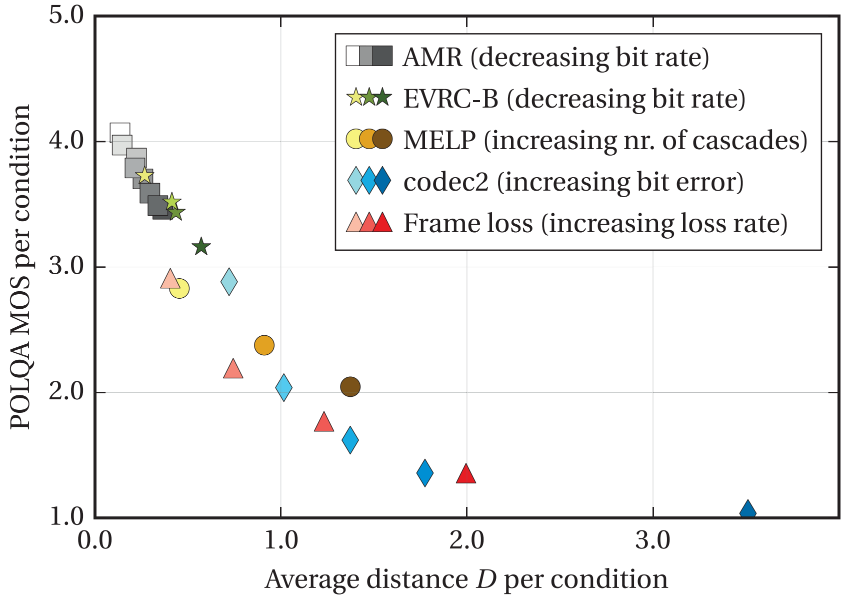


Figure 5.2 – Objective scores for speech *intelligibility* (proposed approach) and *overall quality* (POLQA), for the conditions described in Section 5.4.1. Darker data point shading follows the trends given in the legend parentheses.

## 5.5 Results

### 5.5.1 Low bit-rate coding and frame loss conditions

We calculate average distances $D$ between the original and processed recordings listed in Section 5.4.1, sampled at 8 kHz. Even without subjective intelligibility scores, we can expect codecs with higher bit rates to have higher intelligibility than, e.g., conditions with bit errors of frame losses. We can also expect a trend where lower bit rates of the same codec, or increasing error or loss rates result in lower predicted intelligibility.

Additionally, we compare the *overall quality* of conditions, predicted with ITU-T Rec. P.863 “POLQA” [2011][4], the technological update to ITU-T Rec. P.862 “PESQ” [2001]. Comparing objective scores for intelligibility and quality is interesting, because a reduction in overall quality need not translate to lower intelligibility (e.g., robotic-sounding speech may have low overall quality but still be highly intelligible). On the other hand, good intelligibility is a prerequisite for high overall quality [Côté and Berger, 2014, Sec. 12.1]. This means that we should observe either high or low quality values for high-intelligibility conditions, but only low quality values when intelligibility is low.

Figure 5.2 shows results for both types of objective scores, averaged per condition. Both the AMR and EVRC-B codecs, which operate at comparatively high bit rates, show a range of different quality values as a function of bit rate, but little variation in average distance

[4] The POLQA scores were kindly computed by SwissQual AG, a Rohde & Schwarz company.

Table 5.1 – Prediction performance of per-condition intelligibility scores on the three PSCR datasets (2012, 2010, 2008). Best performances are highlighted in boldface for both the standard case (top section) and the case with added pink noise to model noise at the listener side (bottom section). Connections with asterisks (—∗— / —∗∗—) indicate significant differences between prediction errors ($p < 0.05$ / $p < 0.01$ with Holm-Bonferroni correction).

| Measure | Correlation $\|R\|$ | | | Prediction error rmse$^*_{\text{3rd}}$ [% WA] | | |
|---|---|---|---|---|---|---|
| | 2012 | 2010 | 2008 | 2012 | 2010 | 2008 |
| *No added noise* | | | | | | |
| Average distance $D$ | **0.955** | **0.937** | **0.948** | **1.17** | **2.52** | **6.99** |
| STOI [Taal et al., 2011] | 0.948 | 0.858 | 0.898 | 2.68 | 4.84 | 11.16 |
| *Added stationary pink noise* | | | | | | |
| Average distance $D$ | 0.901 | 0.642 | **0.970** | 3.33 | 8.13 | **3.88** |
| STOI [Taal et al., 2011] | **0.964** | **0.838** | 0.950 | **1.72** | **5.56** | 6.29 |

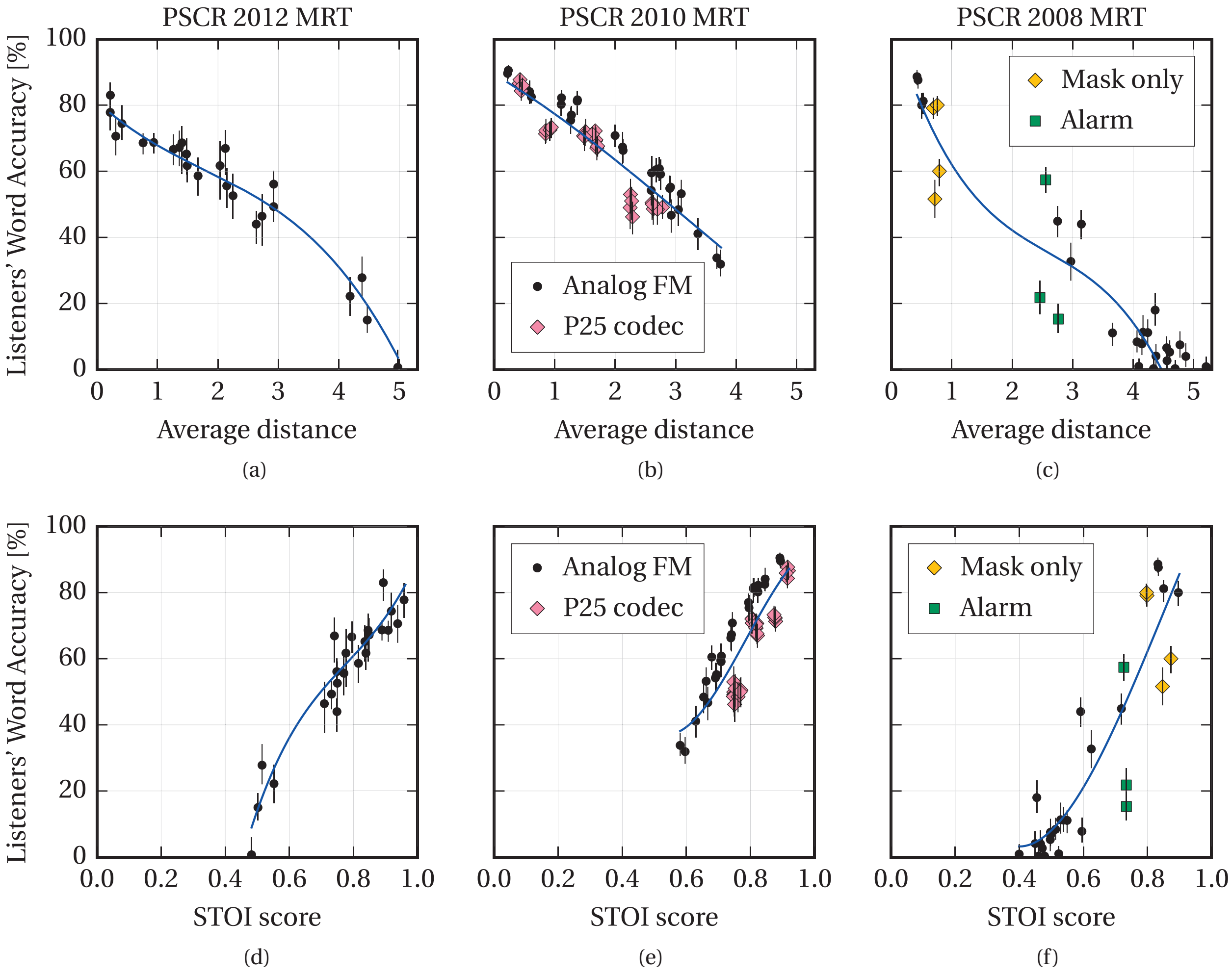
(i.e., high predicted intelligibility). The MELP codec at 2.4 kbps (single encoding, bright circle in Figure 5.2) reaches a lower quality value, but a predicted intelligibility similar to that of the two cellular telecommunication codecs. This seems plausible, given that MELP is a low bit-rate codec designed for mission-critical telecommunications, where intelligibility is key.

On the other hand, conditions with high average distance (low predicted intelligibility) are only found at low objective overall quality scores, as expected. Increases in the number of MELP encoding cascades, codec2 bit errors or frame losses all show the expected trend. Informal listening indicates that speech remains partly intelligible at 40% frame loss, but not in the codec2 condition with maximum bit error rate (dark triangle and dark diamond in Figure 5.2, respectively).

### 5.5.2 Public safety communications conditions

We estimate the intelligibility of rhyme words in all test signals and compute the average distance $D$ per condition. As a comparison to the proposed approach, the STOI score is computed on truncated signals to consider only the final (rhyme) word, with an additional 192 ms of context before the word to avoid signal durations below 384 ms (the measure's minimum).

The top section of Table 5.1 compares the prediction performance of the proposed average distance measure to that of the STOI measure on the three PSCR datasets. The proposed measure consistently achieves a higher correlation to subjective intelligibility scores, as well as a lower prediction error rmse$^*_{\text{3rd}}$. With both measures, the prediction error is largest for the 2008 PSCR dataset.



(a) (b) (c)

(d) (e) (f)

Figure 5.3 – Per-condition intelligibility scores for the PSCR datasets. Top row: proposed average distance measure. Bottom row: STOI measure [Taal et al., 2011]. Error bars indicate 95% confidence intervals of subjective scores. Blue lines show the mapping function used to calculate the prediction error $\text{rmse}^*_{3\text{rd}}$. Highlighted outliers are discussed in the text.

Figure 5.3 provides a closer look at predictions vs. subjective scores for each dataset. Both measures yield good predictions for conditions in the 2012 dataset (panels 5.3a and 5.3d). The 2010 dataset reveals a systematic offset of the STOI measure between two speech coding schemes (panel 5.3e), which does not occur with the proposed measure (panel 5.3b). Finally, panels 5.3c and 5.3f show that both measures failed to predict the effect of some combined impairments in the 2008 dataset. Specifically, for conditions where speech is already distorted by the diaphragm of a breathing mask or by an alarm tone at −2 dB SNR (highlighted outliers), additional differences in speech codec and -level had almost no effect on objective scores.

This latter discrepancy could be due to the fact that listeners were seated in a room with ambient pink noise while evaluating intelligibility (so-called *noise at the listener side*). The 2012 and 2010 tests used a noise SNR of 19 dB relative to the average clean speech level,

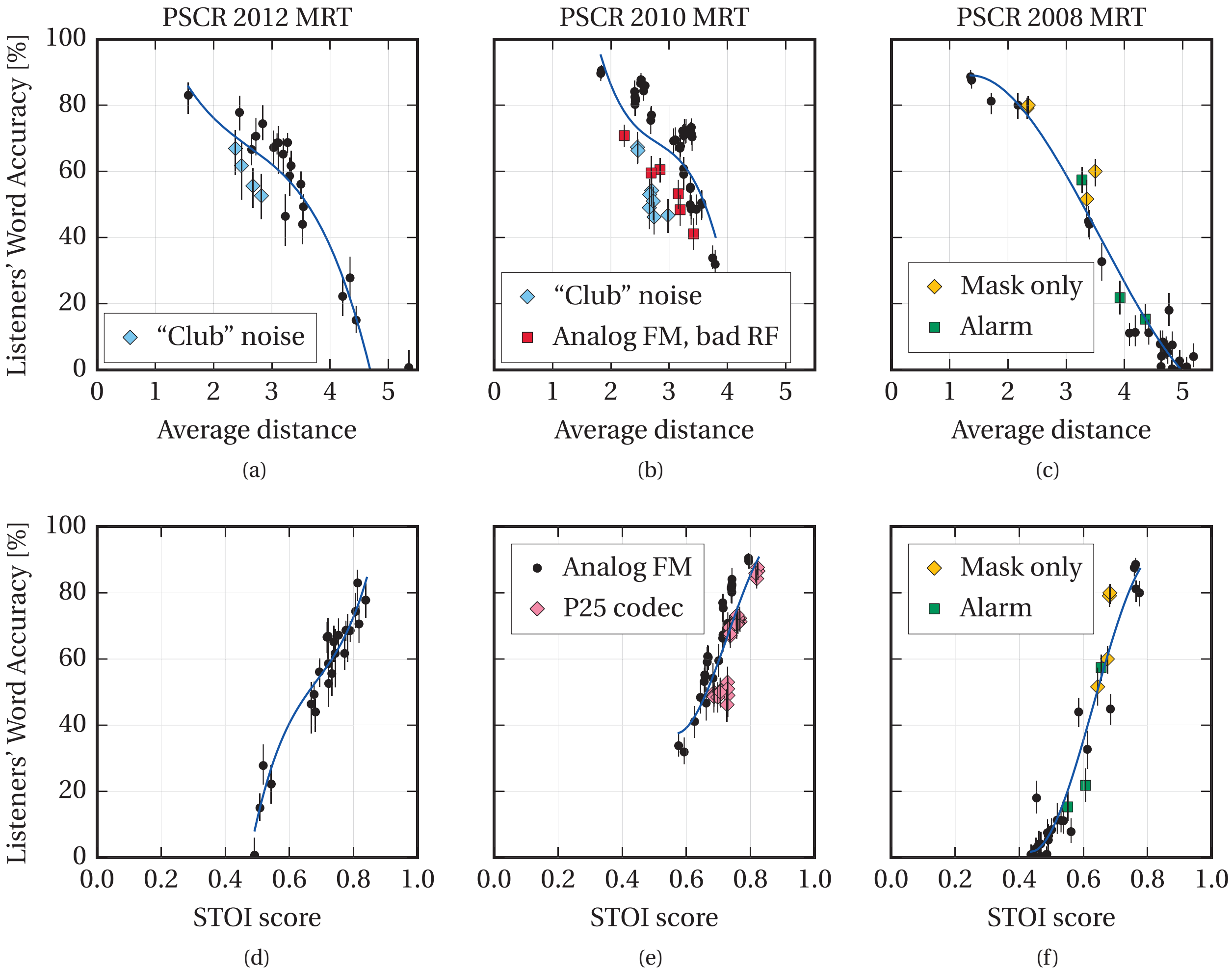


Figure 5.4 – PSCR intelligibility scores, with added pink noise to simulate noise at the listener side. Top row: proposed measure. Bottom row: STOI measure [Taal et al., 2011]. Error bars indicate 95% confidence intervals of subjective scores. Blue lines show the mapping function used to calculate the prediction error $\mathrm{rmse}^*_{3\mathrm{rd}}$. Highlighted outliers are discussed in the text.

whereas the 2008 test used a 12 dB SNR [Atkinson et al., 2008–2013]. In other words, the higher ambient noise level in the 2008 test may have exacerbated some intelligibility impairments for listeners. To test this assumption, we repeated the evaluation after adding pink noise as described in [Atkinson et al., 2008–2013] to all test speech files. The resulting predictions are displayed in Figure 5.4. As can be seen in panels 5.4c and 5.4f, both objective measures now predict additional impairments due to different speech coding systems, suggesting that the high ambient noise level in the 2008 test indeed increased perceived intelligibility impairments.

The lower section of Table 5.1 shows performance metrics after the addition of pink noise. Simulating the presence of ambient noise results in a highly significant ($p < 0.01$ with Holm-Bonferroni correction) reduction of prediction errors in the 2008 test for both measures. However, errors highly significantly increase in the 2012 and 2010 tests for the proposed measure. Here, the addition of noise creates an offset to conditions that already contain broadband noise ("night club" noise and analog FM radio static, highlighted in panels 5.4a and 5.4b, respectively).

Comparing individual predictions in the 2010 and 2012 tests, it appears that adding soft pink noise excessively increased the distance scores of other conditions (i.e., those not containing broadband noise). Since the MLP in our experiments was trained on clean speech only, we speculate that the estimation of phoneme posteriors is too sensitive to low-level noises that have little actual impact on intelligibility. As a result, STOI outperforms the proposed measure in the 2012 and 2010 tests when listener-side noise is added to recordings.

### 5.5.3 Effect of ANN depth and input feature type

The approach proposed in this chapter used a comparison of phoneme-level features to assess intelligibility. The same type of features have been used in ASR (automatic speech recognition) for acoustic modeling, with extensive research on ANN topologies and input feature types. A particular recent trend in ASR is to train ANNs with multiple hidden layers to achieve more accurate estimations of phoneme-level content. Furthermore, the sensitivity of the proposed approach to low-level noises, as observed in the previous section, could be addressed by using more noise-robust acoustic features. In this section, we investigate the potential benefits of both changes to our objective intelligibility measure.

So-called "deep" ANNs have recently been shown to achieve higher accuracy than single hidden layer ANNs trained on the same data [see e.g., Hinton et al., 2012]. In a recent study, Imseng et al. [2013] compared 3-layer and 5-layer MLPs (i.e., with 1 and 3 hidden layer(s), respectively) that had been trained discriminatively for phoneme posterior estimation. They found that the 5-layer MLP consistently achieved higher frame-level phoneme accuracies and lower ASR word error rates for different corpora and with different ASR systems.

Some earlier studies also evaluated alternatives to the cepstral acoustic features used in

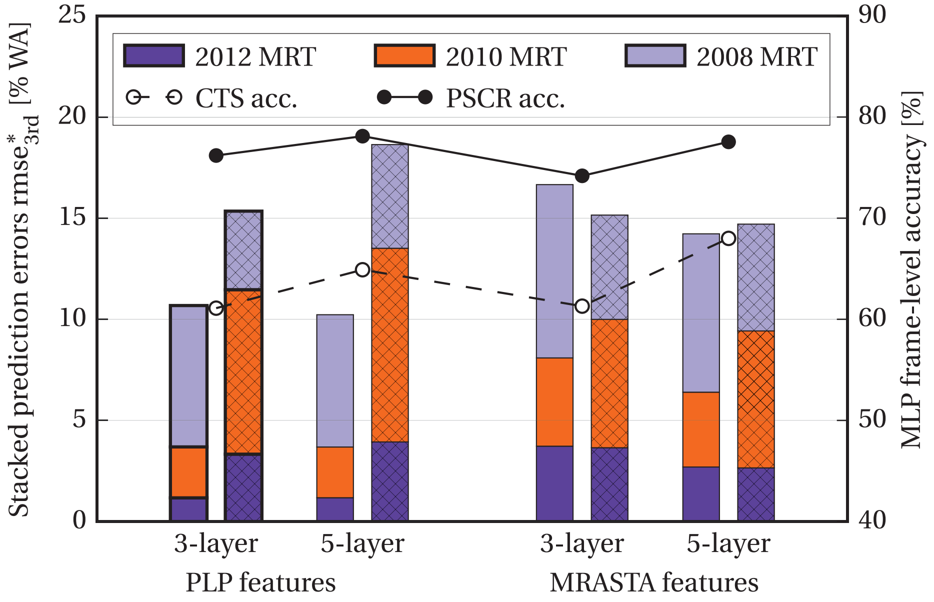


Figure 5.5 – Effect of MLP depth and input feature type on MLP frame-level accuracy (lines) and on prediction errors (bars). Hatched bars show prediction errors for the case with added pink noise to simulate noise at the listener side. The two highlighted (thick) bars correspond to the original hyperparameters and to the results in Table 5.1.

our approach. In particular, several experiments highlighted the importance of temporal *modulations* in the speech envelope for intelligibility [e.g., Drullman et al., 1994; Greenberg et al., 1998]. The same modulation frequencies were shown to be critical for the automatic recognition of clean and noisy speech [Kanedera et al., 1998], and have inspired modulation-based features for ASR [Hermansky et al., 1992; Hermansky and Fousek, 2005].

Therefore, we evaluate MRASTA modulation-based features [Hermansky and Fousek, 2005] as alternative to PLP features, as well as using 3- vs. 5-layer MLPs. The MRASTA features are extracted from the same 24 mel band energies and frame size as for PLP feature calculation, and describe first and second temporal derivatives of band energies over 6 different lengths, plus appended band energies, yielding $(2 \times 6 + 1) \times 24 = 312$-dimensional features. Further details on MRASTA calculation are given in Section 2.5.2, page 18. We evaluate 3- and 5-layer MLPs for both acoustic feature types, trained on the same conversational telephone speech (CTS) data as described earlier in Section 5.4.3. The number of hidden units is the same for both feature types, i.e., 5000 units and 5000–1000–5000 units for the 3- and 5-layer case, respectively.

The dashed line in Figure 5.5 shows the frame-level accuracy obtained with these different MLPs on the CTS cross-validation data. Using 5 layers increases the accuracy of posteriors, with a maximum of 68.0% for the 5-layer MLP with MRASTA input features. However, these gains do not translate to improved intelligibility predictions. Stacked bars in Figure 5.5 show prediction errors for the three PSCR datasets. There appears to be no consistent trend between MLP accuracy and prediction performance for intelligibility.

The lack of improvement could be due to a bias of the MLP to the CTS training material (e.g., conversational speech vs. read speech in PSCR) that appears with the higher number of parameters in the 5-layer case. In order to verify this possibility, we also evaluated MLP accuracies for the (clean) PSCR reference recordings (36 minutes). Accuracy was computed through forced Viterbi alignment of posteriors to the phonetic transcription of references, using the UNISYN pronunciation dictionary [Fitt, 2000]. The results (solid line in Figure 5.5) follow a similar trend to those for CTS material, suggesting that 5-layer MLPs still generalized well to the out-of-domain PSCR data. Moreover, accuracies for PSCR references are on average 20% higher, as could be expected for high-quality read speech.

Regarding the effect of feature type, the proposed approach indeed appears to be less sensitive to the addition of pink noise when MRASTA features are used (plain vs. hatched bars in Figure 5.5). Nevertheless, the increase in prediction errors for the 2010 dataset remains highly significant with all hyperparameters tested here ($p < 0.01$ with Holm-Bonferroni correction). This result supports our notion that the absence of noisy speech during MLP training causes the measure to be overly sensitive to noises that have little impact on intelligibility.

More generally, these results demonstrate a fundamental difference between the objectives in ASR and intelligibility assessment. While ASR seeks to push recognition performance to the highest possible levels, assessment is concerned with finding measures that mimic human behavior.

## 5.6 Discussion and conclusion

We have proposed a novel objective measure of speech intelligibility, based on a distance score between phoneme posterior probability sequences. Our experiments in Section 5.5 show that this approach yields realistic results for low bit-rate coding conditions, and achieves a very good agreement with subjective scores for the PSCR dataset, which includes background noise, speech coding and acoustic impairment conditions. This result is consistent with the study in [Soldo, Magimai.-Doss, and Bourlard, 2012], where such a distance score was found to also correlate with the intelligibility of synthetic speech templates.

The approach proposed here used perceptually motivated features and training data with high pronunciation variability to estimate phoneme posterior probability sequences. The analysis in Section 5.5.3 illustrated that other estimators do not necessarily achieve better prediction performance if they are optimized to predict true as opposed to *perceived* phonetic content. Training an MLP on perceived phonetic content (using annotations for noisy speech) was proposed by Meyer and Kollmeier [2010] to predict intelligibility scores. However, their study was limited to a single noise type and SNR. Moreover, most available datasets of phonetic misperceptions focus on speech in noise [e.g., Cooke and Scharenborg, 2008; Tóth et al., 2015]. It is thus unclear whether a model trained on such data would generalize to other degradation types as found in telecommunications.

The objective measure proposed in this chapter could be further developed along several lines:

- An advantage of the proposed approach is that reference and test speech may be different realizations of the words from the same speaker, or even originate from different speakers. This is relevant to the assessment of very low bit-rate speech codecs that may modify speaker characteristics, but has not been formally evaluated yet.
- Our experiments on the PSCR dataset used known signal positions to assess the intelligibility of rhyme words. In the case that such positions are not known, the approach could be extended to word-level assessment without performing ASR, using an utterance verification approach.
- We used a single reference recording to assess the mismatch to listener's modeled phonetic knowledge. However, it may be argued that listeners possess more than one "internal reference" for each word. The approach could thus benefit from using multiple reference speech recordings, or from replacing reference speech by a statistical model such as a Kullback-Leibler divergence-based HMM (KL-HMM, introduced in Section 2.7, page 19), which models lexical and phonetic content [Aradilla et al., 2007].

We address these points in the following chapter. Specifically, we apply and extend the proposed approach to the assessment of *synthetic speech* intelligibility. Synthetic speech may have completely different speech timbre and tempo from natural speech, making it an interesting evaluation of the approach's insensitivity to speaker characteristics.

# 6 Objective intelligibility assessment of synthetic speech

In the previous chapter, we presented a novel approach to speech intelligibility assessment, based on a comparison of phoneme posterior sequences of reference and test speech signals. A key motivation for this approach was the assessment of low bit-rate codec conditions, where a spectral feature-based comparison of signals may be sensitive to differences in speech timbre or tempo that do not affect intelligibility. In this chapter, we evaluate our approach on *synthetic* speech, as examples of test signals that have very different timbre or tempo than natural (human) speech. Synthetic speech intelligibility assessment is also directly relevant to telecommunications, where recent very low bit-rate codecs (≤ 1 kbps) built on speech synthesis principles have been proposed [e.g., Lee and Cox, 2001; Cerňak, Potard, et al., 2015].

In order to evaluate its potential for synthetic speech, we first apply our approach with natural speech recordings as reference, i.e., we compare phoneme posterior sequences of natural and synthetic speech recordings of the same words to compute an average distance score. In a second step, we address intelligibility assessment at the *word level*, in order to obtain a score that is more comparable to that of human listeners. Specifically, we replace the natural reference recording by a model of the expected phoneme posterior sequence, generated from the (known) textual transcription of the sentence. Aligning the model with the phoneme posterior sequence of the test speech recording provides the word-level segmentation. The intelligibility of each word can then be assessed individually. In addition, the model can be used to represent a more general reference than a single natural speech recording.

This chapter is structured as follows: We briefly review existing approaches to intelligibility assessment of synthetic speech in Section 6.1, and present the overall experimental setup for this chapter in Section 6.2. The approach of Chapter 5, which consists in comparing synthetic speech to a natural speech recording, is evaluated in Section 6.3. Based on the results of this first experiment, we present a modified approach where the reference recording is replaced by a model, as described in Section 6.4. The results of this second approach are analyzed and compared in Section 6.5. We conclude with further comments in Section 6.6.

## 6.1 Related work and contributions

Text-to-speech (TTS) synthesis can be understood as a two-stage process that consists in 1) converting text to linguistic and phonetic information, e.g., by placing pauses and stress and using a pronunciation dictionary, and 2) converting linguistic and phonetic information to an acoustic signal, i.e., by modeling pitch and duration, and generating the speech waveform [Van Bezooijen and Pols, 1990]. Errors or artifacts in either stage may introduce audible distortions in the synthesized signal. Typical examples are discontinuities between adjacent speech units in the output of unit-selection TTS systems, or muffled speech due to over-smoothing of parameters in statistical parametric synthesizers [see, e.g., Zen et al., 2009, for a review].

Due to their different root cause, these distortions do not resemble the ones encountered with natural speech in traditional telecommunication systems, meaning that synthetic speech should not be regarded as a degraded version of natural speech, but as a different class of speech altogether [Norrenbrock et al., 2015]. In addition to overall quality and intelligibility, assessment of synthetic speech has therefore focused on further quality features like perceived *naturalness*, *similarity to a target speaker*, *pleasantness*, or *pronunciation anomalies* [Clark et al., 2007; ITU-T Rec. P.85, 1994].

The disparity between natural and synthetic speech also means that objective measures designed for natural speech may not be readily applied, as verified by the very limited success of studies attempting to assess overall synthetic speech quality with traditional measures. For full-reference measures (i.e., where a test signal is compared to a high-quality reference signal), it was found that traditional measures could not time-align synthetic speech to a natural speech reference [Cerňak and Rusko, 2005; Hinterleitner, Zabel, et al., 2011]. No-reference measures, where a test signal is evaluated against a general model of speech production [e.g., Malfait, Berger, and Kastner, 2006; Kim and Tarraf, 2007], were also shown to provide unsatisfactory prediction performance for synthetic speech [Möller, Kim, et al., 2008]. The authors assumed a failure of these measures to detect distortions that are specific to synthetic speech as one of the main reasons for this result.

Developers of TTS systems have therefore relied on measures for specific artifacts or targets in speech synthesis. Two prominent examples are measures of perceived discontinuities to assess the output of unit-selection TTS systems [e.g., Stylianou and Syrdal, 2001; Vepa et al., 2002] and the mel-cepstral distance [Gray, Jr. and Markel, 1976] between synthetic and natural speech as a measure of speaker similarity [Mashimo et al., 2001; Remes et al., 2013].

More recently however, approaches have been proposed that seek to assess the overall quality or intelligibility of synthetic speech in an integral fashion, i.e., without restriction to specific artifacts. Falk and Möller [2008] trained a reference HMM on natural speech recordings and found that the log-likelihood to synthetic speech yielded promising results for the prediction of overall quality, naturalness and continuity. Other quality features such as "comprehension"[1]

[1] "Comprehension" as defined in ITU-T Rec. P.85 [1994] describes listeners' *opinion* of whether speech is hard to understand, without formally verifying the correct recognition of words as in dedicated intelligibility tests.

were not well predicted, which may be due to the signal features used in their approach (cepstral coefficients of a short-time spectrum). Wang et al. [2012] used the decoder of an ASR system to analyze the phone graph in synthetic speech, and compared it to multiple templates of individual phones with multiple context to derive an intelligibility score. Their approach requires a reference phonetic transcription for the comparison, and achieved high correlation with subjective intelligibility scores for clean synthetic voices in the Blizzard 2010 dataset. An entirely ASR-based approach was proposed by Hinterleitner, Zander, et al. [2015], who used the Google ASR API to compute word error rates (WER) for synthetic speech recordings in the Blizzard 2011 dataset, and obtained a high correlation to subjective intelligibility scores.

Finally, recent work also evaluated the intelligibility of *degraded* synthetic speech, e.g., due to background noise or simulated telephone channels. It was shown that in this case, objective measures that were developed for natural speech *can* be applied to assess the impact of these degradations, e.g., by comparing the original and degraded synthetic signal [Počta and Beerends, 2015]. Traditional objective measures were further shown to predict intelligibility gains obtained by modifying synthetic speech in noise, e.g., through spectral or temporal changes inspired by the Lombard effect [Valentini-Botinhao et al., 2011; Tang et al., 2016].

In summary, objective intelligibility measures (OIMs) developed for natural speech appear to be suitable to assess the impact of *degradations* on synthetic speech, but not the intelligibility of synthetic voices themselves. On the other hand, OIMs specifically developed for synthetic speech have only been evaluated on signals without further degradations. In the following sections, we apply the approach from the previous chapter to both problems, i.e., clean and noise-corrupted synthetic speech assessment. We then present an extension that seeks to improve predictions, and also addresses the assessment of intelligibility at the word level.

Contribution

R. Ullmann, R. Rasipuram, M. Magimai.-Doss, and H. Bourlard [2015]. “Objective Intelligibility Assessment of Text-to-Speech Systems Through Utterance Verification”. In: *Proc. Interspeech*. Dresden, Germany, pp. 3501–3505.
URL: http://isca-speech.org/archive/interspeech_2015/i15_3501.html

This chapter expands on the above paper, adding results for tasks EH1 (clean speech) and ES2 (noisy speech) of the Blizzard 2010 dataset, and a more detailed evaluation. Experiments with the template-based approach include results from our previous contribution [Ullmann, Magimai.-Doss, et al., 2015].

## 6.2 Experimental setup

We conduct all experiments in this chapter on recordings of semantically unpredictable sentences (SUS) from the 2010 and 2011 Blizzard challenges [King and Karaiskos, 2010; 2011],

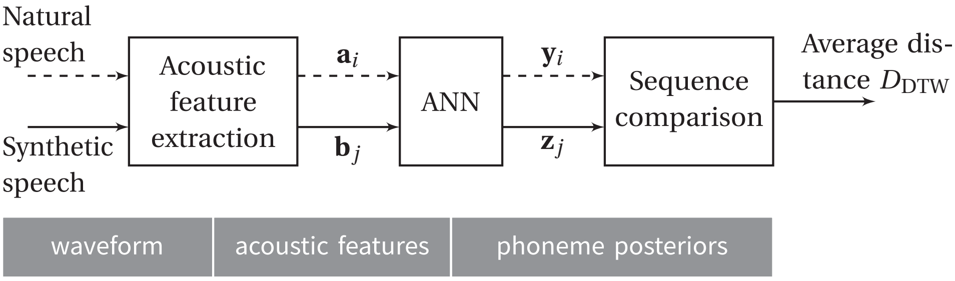


Figure 6.1 – Synthetic speech intelligibility assessment through comparison to a natural speech template. The approach is the same as in Figure 5.1, page 65, except that reference and test speech at the input are replaced by natural and synthetic speech recordings, respectively. Symbols are analogous to Section 5.3, with $\mathbf{a}_i$ and $\mathbf{b}_j$ acoustic feature vectors and $\mathbf{y}_i$ and $\mathbf{z}_j$ phoneme posterior probability distributions at natural and synthetic speech time index $i$ and $j$, respectively.

which were synthesized with 17 and 12 different text-to-speech (TTS) systems, respectively. Both the 2010 and the 2011 challenges also include natural speech recordings from a professional voice talent for all sentences. Additionally, the 2010 Challenge analyzed the intelligibility of natural speech and 12 TTS systems in different levels of speech-shaped noise, resulting in a total of three synthetic speech datasets (two clean and one noisy dataset) for our experiments. Further information about the Blizzard datasets is given in Section 2.4.3, page 15.

The experiments in this chapter use the same setup as in Section 5.4.3 to analyze the phonetic content of speech signals, i.e., 39-dimensional PLP acoustic features and a 3-layer MLP to estimate posterior probabilities for 44 English phonemes and silence. In particular, the acoustic feature extraction, as well as the Conversational Telephone Speech (CTS) data used for MLP training, are limited to the narrow telephone band (< 4 kHz bandwidth), whereas speech recordings in the Blizzard datasets are provided with a bandwidth of up to 24 kHz. Since bandwidths beyond the traditional telephone band provide almost no intelligibility benefit [Côté and Berger, 2014; Fernández Gallardo and Möller, 2015], we do not consider the information in this extra frequency range for our experiments.

## 6.3 Template-based intelligibility assessment

As a first experiment, we apply the method presented in Chapter 5 without further modifications, i.e., synthetic speech is compared to the natural speech recording of the same words. This amounts to treating synthetic speech as a distorted version of the reference recording, even though both recordings are different realizations of the same words. In this sense, the natural reference recording is just an example or a *template* of a pronunciation of those words.

The steps for template-based intelligibility assessment are outlined in Figure 6.1. Since natural and synthetic speech have different temporal structure, the duration of individual words as well as the overall lengths $I$ and $J$ of posterior sequences need not be the same, i.e., $I \lesseqgtr J$.

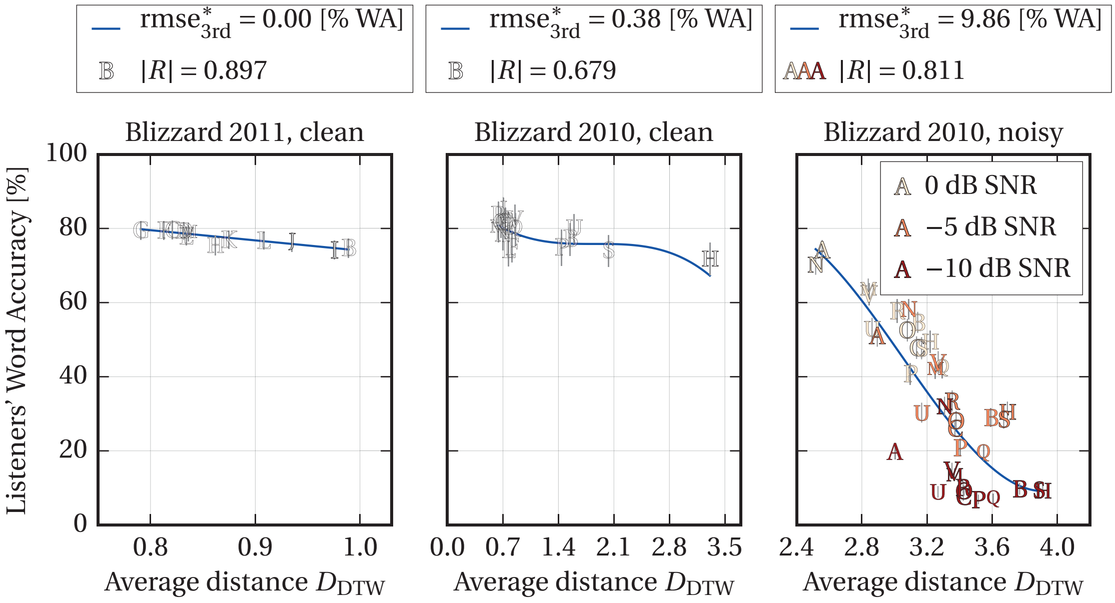


Figure 6.2 – Prediction of per-voice intelligibility through comparison to a reference template. Letters identify different voices, with “A” for natural speech and “B”–“V” referring to *different TTS systems* in the 2011 and 2010 Challenge. Error bars indicate 95% confidence intervals of subjective scores. Blue lines show the mapping function used to calculate the metric rmse$^*_{3rd}$.

Therefore, we apply dynamic time warping (DTW)[2] to align both sequences, and use the global distance score $D_{\mathrm{DTW}}$ as defined in (5.7), page 66, for intelligibility assessment. Due to the DTW path constraints in (5.6), the alignment may be erroneous if segments in the reference signal are more than twice as long as in the test speech recording. In particular, differences in leading and trailing silence lengths can cause the alignment to fail. Thus we adjusted leading and trailing silences in all recordings to the same lengths before applying our approach.[3]

Figure 6.2 compares the per-voice means of average DTW distances to subjective intelligibility scores. Letters represent different voices as referred to in [King and Karaiskos, 2010; 2011], with “A” denoting natural and “B”–“V” synthetic speech, respectively. Since natural speech recordings are used as references, the intelligibility of system “A” recordings is only assessed for the noise-corrupted conditions in the right panel of Figure 6.2. The resulting performance metrics are satisfactory overall, indicating that the method can generalize to synthetic speech.

Looking first at the noisy speech conditions in the right panel of Figure 6.2, the overall trend of natural and synthetic speech intelligibility in noise is well predicted. However, some outliers can be observed at negative SNRs, with the proposed approach failing to predict the higher intelligibility of TTS system “N” compared to natural speech (system “A”). Since the natural system “A” recordings are used as references, there are no pronunciation variabilities between reference and test speech, leading to an offset to synthetic voices.

[2] The DTW implementation was carried out by Mr. Guillem Quer during his internship at Idiap.
[3] One incompletely synthesized sentence from system “V” in Blizzard 2010 was excluded due to its short length.

A further offset can be seen for system “H” recordings in both clean and noisy conditions in Blizzard 2010.[4] Speech synthesized with this system is only slightly faster overall (1.2×) than the system “A” reference, but informal listening shows that the speech tempo varies strongly within sentences, with abrupt transitions between phonemes. The DTW path constraints and the speech tempo in the reference template impose minimum durations on phonemes in each sentence, penalizing fast voices. As a result, clean recordings of system “H” are evaluated with an average distance $D_{\mathrm{DTW}}$ greater than that of natural speech at −10 dB SNR, even though the latter is much less intelligible.

Finally, in the 2011 data all subjective confidence intervals intersect with the mapping function for the objective measure, yielding a prediction error $\mathrm{rmse}^*_{3\mathrm{rd}} = 0$ (left panel of Figure 6.2). In this case, the intelligibility of clean synthetic speech is confined to a narrow range that approaches the variability in listener scores. The $\mathrm{rmse}^*_{3\mathrm{rd}}$ also remained zero with confidence intervals computed for the subset of native, paid listeners, thus we decided to keep using the complete set of listener data, as done in [King and Karaiskos, 2010; 2011].

### 6.3.1 Analysis of significant differences

The narrow range of intelligibility differences, combined with the length of 95% confidence intervals of subjective scores, means that the $\mathrm{rmse}^*_{3\mathrm{rd}}$ is not an insightful performance metric for objective measures of clean synthetic speech intelligibility. However, this does not imply that all TTS systems are comparable, with small but significant differences being revealed through paired comparisons of *individual* listener scores.

Table 6.1 shows the results of paired Wilcoxon signed-rank tests in [King and Karaiskos, 2011], with squares indicating significant subjective differences between pairs of systems. For each pair with a significant difference, we evaluate whether the rank-order of objective scores per voice agrees with that of listeners. As shown by green squares in Table 6.1, the rank-order is predicted correctly for all 14 significant differences. Further significant differences with system “A” are grayed out, since objective scores for clean natural speech are not available with our template-based approach.

Repeating the same evaluation for the 2010 Blizzard data, we find that the proposed approach predicts correct rank-orders for 24 out of 27 and 50/53, 45/54, 33/37 significant differences in the clean and the three noisy conditions, respectively. On average, the template-based approach predicts rank-orders correctly for 91% of significant pairwise differences across the three datasets. In determining significant differences, the organizers of the Blizzard Challenge used a significance level of $p < 0.01$ with Bonferroni correction, which is more conservative (i.e., with a lower probability of false positives, at the risk of more false negatives) than the Holm-Bonferroni method used for other evaluations in this thesis. Nevertheless, we chose to use these values so that other authors may perform comparable evaluations on their results.

[4] TTS identifiers refer to *different systems* in the 2010 and 2011 Blizzard challenges and are thus not comparable.

Table 6.1 – Significant intelligibility differences between voice pairs in the 2011 Blizzard data. 🟩 indicate differences where the rank-order of average subjective and objective scores agree (14 out of 14). Results for system "A" are not available with the template-based approach. Significant differences reproduced with permission from [King and Karaiskos, 2011].

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | ╲ | ■ | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| B | ■ | ╲ | 🟩 | 🟩 | | 🟩 | 🟩 | | | | | | 🟩 |
| C | | 🟩 | ╲ | | | | | 🟩 | 🟩 | 🟩 | | | |
| D | ■ | 🟩 | | ╲ | | | | | 🟩 | | | | |
| E | ■ | | | | ╲ | | | | 🟩 | | | | |
| F | ■ | 🟩 | | | | ╲ | | | 🟩 | | | | |
| G | ■ | 🟩 | | | | | ╲ | | 🟩 | 🟩 | | | |
| H | ■ | | 🟩 | | | | | ╲ | | | | | |
| I | ■ | | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | ╲ | | | | 🟩 |
| J | ■ | | 🟩 | | | | 🟩 | | | ╲ | | | |
| K | ■ | | | | | | | | | | ╲ | | |
| L | ■ | | | | | | | | | | | ╲ | |
| M | ■ | 🟩 | | | | | | | 🟩 | | | | ╲ |

In summary, these results show that a comparison of phoneme posterior sequences can be used to assess the intelligibility of synthetic speech. However, the previous analyses also indicated possible issues with the use of a natural speech template as reference, particularly with respect to differences in pronunciation and tempo to synthetic speech.

A further issue that we have not discussed so far is the type of speech material used for assessing intelligibility. The Blizzard Challenge used recordings of semantically unpredictable sentences (SUS), where listeners transcribe any words that they understand in the signal. This is in contrast to the rhyme tests previously used in Chapter 5, where the focus was on a single word. As a result, distortions that are concentrated on a single word and those that are spread throughout the sentence (e.g., an isolated noise burst vs. stationary noise) could result in the same average distance score $D_{\mathrm{DTW}}$, even though they would not affect word-level intelligibility in the same way.

In the next section, we present an extension to our approach that seeks to address these issues.

## 6.4 Model-based intelligibility assessment

### 6.4.1 Motivation

As argued in Chapter 5, intelligibility impairments can be seen as cases of mismatch between the test signal acoustics and listeners' phonetic and lexical knowledge. The rhyme tests used in the previous chapter mostly focused on phonetic knowledge, in that they required listeners to recognize single words given a list of alternatives that differed in one consonant. In contrast, the recognition of semantically unpredictable sentences (SUS) requires both phonetic and lexical knowledge to identify speech sounds and match their sequence to words from a listener's vocabulary. Lexical knowledge is also relevant in the context of synthetic speech assessment, where TTS systems may produce varying pronunciations for a given input text. These pronunciations may differ both temporally and phonetically from a given natural speech template, yet still be valid and intelligible.

One way of accounting for pronunciation variability would be to use multiple templates of different natural speech realizations as references. The method we propose in this section is to replace reference templates by a model that is trained on multiple natural speech recordings. The model represents the expected sequence of phoneme posteriors for a given input text, and can include several alternative pronunciations when those exist. Training the model on multiple speech recordings can also help produce a more general or average representation of reference phoneme-level content.

Finally, since words are clearly separated in text, generating the model from input text also provides knowledge of word boundaries in the posterior sequence, allowing to assess intelligibility separately for each word. Specifically, a distance or uncertainty measure between the synthetic speech signal and the model can be computed for each word, and thresholded to determine which words can be recalled successfully. Our notion is that the computed word recall in percent can be directly related to the word accuracy score of listeners.

### 6.4.2 Proposed approach

Our approach uses a Kullback-Leibler divergence-based Hidden Markov Model (KL-HMM, introduced in Section 2.7, page 19) to generate the sequence of expected phoneme posteriors for a given TTS input text. This reference sequence is aligned with the sequence from the synthetic speech signal, and the mismatch between the two is evaluated for each word.

The architecture of the proposed objective intelligibility measure is shown in Figure 6.3 and consists of the following parts:

*Synthetic speech* — A TTS system takes as input a sequence of words $W = \{w_1, \ldots, w_m, \ldots, w_M\}$ and converts them to speech, with $M$ the total number of words in the input text.

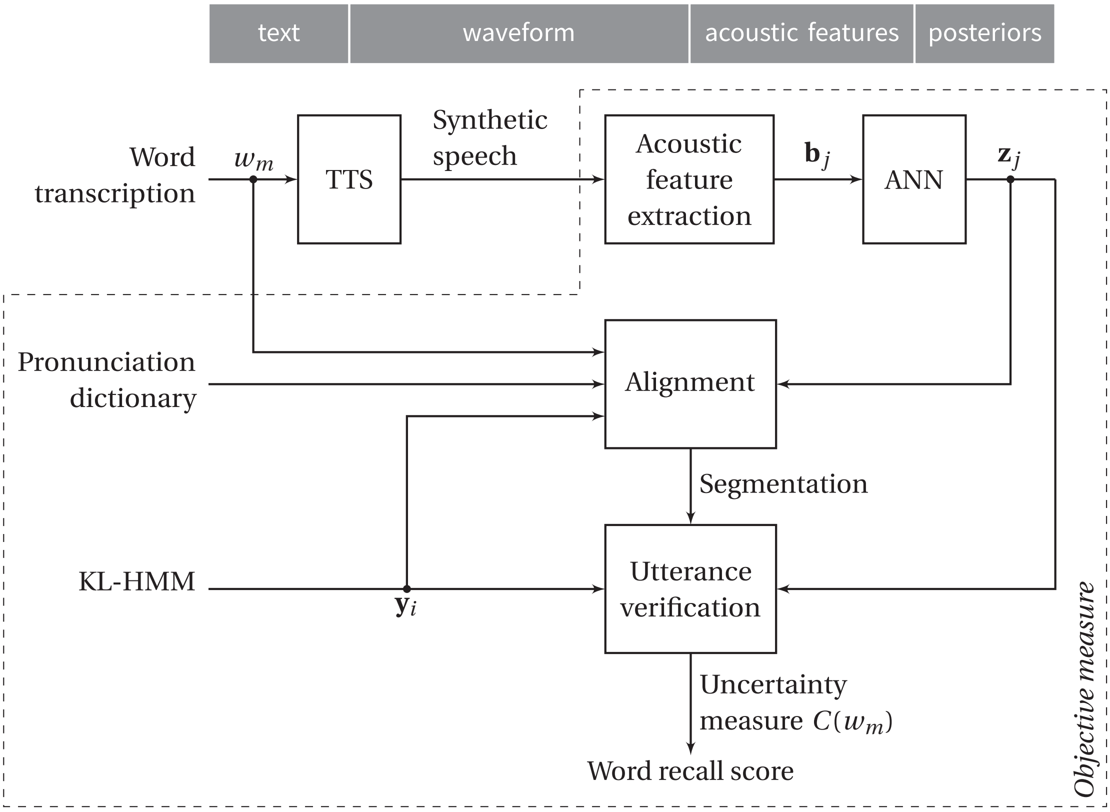


Figure 6.3 – Architecture of the proposed objective TTS intelligibility assessment system. Gray boxes indicate the type of feature across the top row of the flowchart, with "posteriors" for phoneme posterior probabilities.

*Acoustic feature extraction* — An acoustic feature sequence $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_j, \ldots, \mathbf{b}_J\}$ is extracted from the synthetic speech signal produced by the TTS system. The features can be, e.g., cepstral coefficients of a short-time spectrum.

*Test posterior sequence estimation* — The acoustic feature sequence $B$ is converted into a test sequence of estimated phoneme posterior probabilities $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_j, \ldots, \mathbf{z}_J\}$ using an Artificial Neural Network (ANN), where

$$\mathbf{z}_j = \left[P\left(c^1 \mid \mathbf{b}_j\right), \ldots, P\left(c^K \mid \mathbf{b}_j\right)\right]^\top = \left[z_j^1, \ldots, z_j^K\right]^\top, \tag{6.1}$$

with $\sum_k z_j^k = 1$ and $c_k$ the $k^{\text{th}}$ phoneme class out of $k \in [1, K]$ phoneme classes.

*Reference posterior sequence estimation* — The expected sequence of phoneme posterior probabilities for the words $W$ is modeled as a sequence of KL-HMM states. In a KL-HMM, each state $i$ represents a subword unit (e.g., a phone or a context-dependent phoneme), and is parameterized by a categorical distribution $\mathbf{y}_i = \left[y_i^1, \ldots, y_i^k, \ldots, y_i^K\right]^\top$. The state distribution for each subword unit is learned during a prior training step, using phoneme posterior probabilities estimated by an ANN as feature observations, and a pronunciation dictionary. The trained KL-HMM system and the pronunciation dictionary are then used to generate the sequence of subword unit states for a given word.

*Alignment* — The test sequence $Z$ is aligned to the KL-HMM through Viterbi alignment, with the distance

$$\mathrm{RKL}\left(\mathbf{y}_i, \mathbf{z}_j\right) = \sum_{k=1}^{K} z_j^k \log\left(\frac{z_j^k}{y_i^k}\right) \tag{6.2}$$

between the test posterior feature $\mathbf{z}_j$ and the HMM state distribution $\mathbf{y}_i$ as the local score. The alignment provides the segmentation of the test sequence $Z$ at the subword level.

*Utterance verification* — An uncertainty measure $C(w_m)$ is computed for each word $w_m$. The uncertainty measure is based on the local scores calculated in the alignment step, normalized by the number of frames in each subword state and by the number of subword states in each word, similar to the double normalization approach for hybrid HMM/ANN systems [Bernardis and Bourlard, 1998],

$$C(w_m) = \frac{1}{V_m} \sum_{v=1}^{V_m} \frac{1}{e_{vm} - b_{vm} + 1} \sum_{j=b_{vm}}^{e_{vm}} \mathrm{RKL}(\mathbf{y}_{i_{vm}}, \mathbf{z}_j) \tag{6.3}$$

with $i_{vm}$ the $v^{\text{th}}$ subword state in word $w_m$, $b_{vm}$ and $e_{vm}$ the begin and end indices of the test frames aligned with subword state $i_{vm}$, and $V_m$ the number of subword states for word $w_m$, respectively.

*Calculation of word recall* — The word recall is calculated by comparing the uncertainty measure $C(w_m)$ of each word to a decision threshold $\tau$. The value of $\tau$ may be chosen such that the calculated word recall correlates best with intelligibility scores, e.g., using a small development set of subjectively scored synthetic speech recordings.

Alternatively, the threshold $\tau$ may be selected without subjectively scored data, by choosing the value that best separates two distributions $H_0$ and $H_1$ of uncertainty scores. The $H_0$ hypothesis means that the expected word is present in the signal, whereas $H_1$ means that the TTS system synthesized a signal that does not agree with the phonetic and lexical knowledge for the word, as modeled by the KL-HMM.

We obtain uncertainty scores for the $H_0$ distribution from speech signals with known high intelligibility, e.g., undistorted natural speech. Uncertainty scores for the $H_1$ distribution can be obtained by distorting the signals, or — more simply — by using *wrong transcriptions* for utterance verification. The wrong transcription can be a word that sounds similar to the true word or a completely different word, depending on the type of intelligibility test for which we are designing the objective measure. We will compare both approaches to selecting $\tau$ in Section 6.5.1.

### 6.4.3 KL-HMM training

The KL-HMM system is trained on the system "A" natural speech recordings and models crossword context-dependent phonemes. Specifically, each crossword context-dependent phoneme is modeled with three HMM states, resulting in a minimum duration constraint of three frames for each phoneme. The 2010 and 2011 Blizzard Challenge used different semantically unpredictable sentences (SUS), meaning that the corresponding natural speech recordings also cover different phoneme contexts within and across words. We therefore trained two different KL-HMM systems (one for each Challenge year), using 100 SUS recordings of natural speech in the respective Challenge data (4.6 and 4.8 minutes, respectively). Given the very low amount of training data, some phoneme contexts will be observed very rarely or not at all. Therefore, a state tying approach is used to share data between similar states [Imseng, 2013, Chap 4.6].

The word pronunciations needed for training the KL-HMM, as well as for generating the reference sequence for a given input text, are looked up in the CMU pronunciation dictionary [Carnegie Mellon Speech Group, 2014]. Further details on KL-HMM training are given in Section 2.7 of the Background chapter, page 19.

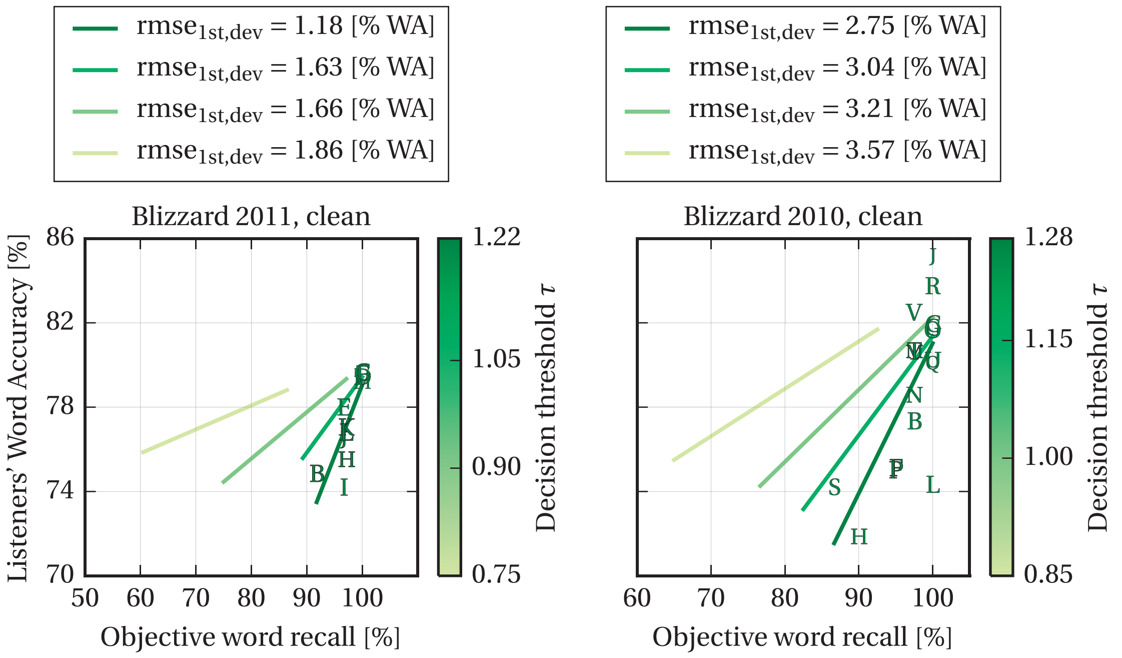


Figure 6.4 – Relation between word recall and word accuracy for different decision thresholds, using a development set of five SUS recordings per TTS system. Lines show the linear fit between both measures at the threshold values $\tau$ indicated on colorbar ticks. For clarity, data points for TTS systems are shown for the best threshold value only.

## 6.5 Results

### 6.5.1 Threshold selection

We compute uncertainty scores $C(w_m)$ for the words in the SUS recordings of each voice in the Blizzard data, using the steps described in Section 6.4.2. The word recall per recording is obtained by comparing these scores to a decision threshold $\tau$. Our notion is that the word recall is directly related to listener's word accuracy (WA). However, word recall describes the verification of *expected* words with the model, whereas listeners' WA scores describe the recognition of *unknown* words. Therefore, we expect word recall to be in a higher numerical range than WA. Since the KL-HMMs for the assessment of Blizzard 2011 and 2010 conditions are trained on different speech recordings, we derive separate thresholds for each model.

#### Selection with a development set

The relation between objective recall and subjective word accuracy (WA) is shown in Figure 6.4. We use the first five SUS recordings that listeners rated for each TTS system in the 2011 and 2010 Blizzard Challenge, respectively, as development set to find a threshold value $\tau$.

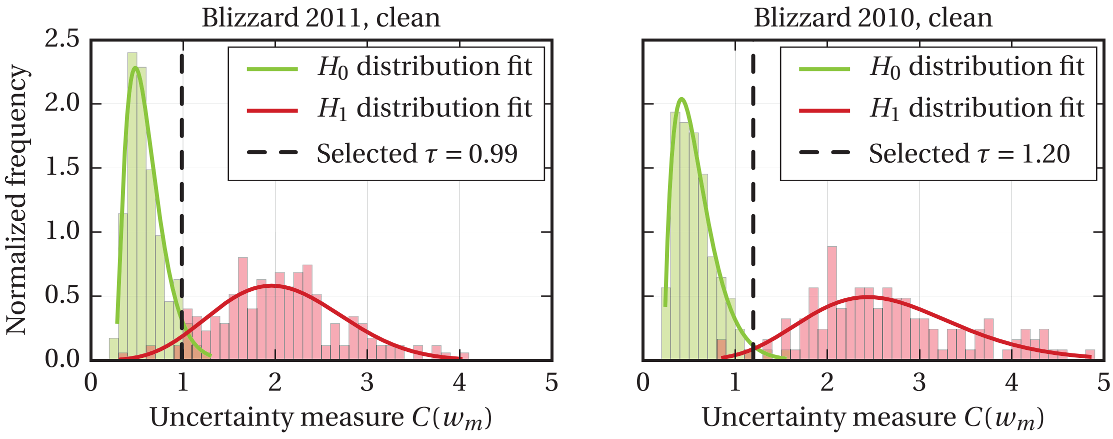


Figure 6.5 – Selection of the decision threshold without development data, using two distributions of uncertainty scores. The distributions are obtained from uncertainty scores $C(w_m)$ for words in natural speech recordings with correct ($H_0$) and wrong ($H_1$) transcriptions, respectively. The threshold is selected such as to minimize the overlap of both distributions.

Specifically, we select the value of $\tau$ that results in the lowest prediction error $\text{rmse}_{\text{1st}}$,

$$\text{rmse}_{\text{1st}} = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}\left(s_i - o'_i\right)^2} \tag{6.4}$$

with $o'_i$ the objective word recall after linear mapping and $s_i$ the subjective WA score for the $i^{\text{th}}$ TTS system, respectively. For simplicity, the prediction error in (6.4) is computed without consideration of subjective confidence intervals.

Lines in Figure 6.4 show the linear mapping from recall to WA for different values of $\tau$. At low threshold values (bright lines), the word recall is in a lower numerical range than subjective WA scores, and the prediction error is also larger. As the threshold is increased to yield recall values in the higher ranges that we expect, the prediction error improves too (dark lines). The smallest prediction error on the two development sets is obtained with decision thresholds $\tau^{2011}_{\text{dev}} = 1.22$ and $\tau^{2010}_{\text{dev}} = 1.28$, respectively, yielding word recall values between 87 and 100%.

### Selection with two distributions of uncertainty scores

As an alternative to using subjectively scored data, we derive decision thresholds $\tau$ by comparing two distributions of uncertainty scores $C(w_m)$ as shown in Figure 6.5. The $H_0$ distributions for either Challenge year represent uncertainty scores for words in 100 system "A" recordings, which were pronounced by a professional voice talent and are highly intelligible. $H_1$ distributions are obtained with the same recordings, but using transcriptions for different words to produce an intentional mismatch with the KL-HMM reference. Specifically, each word in the transcription is substituted at random with a different word from the 100 sentences.

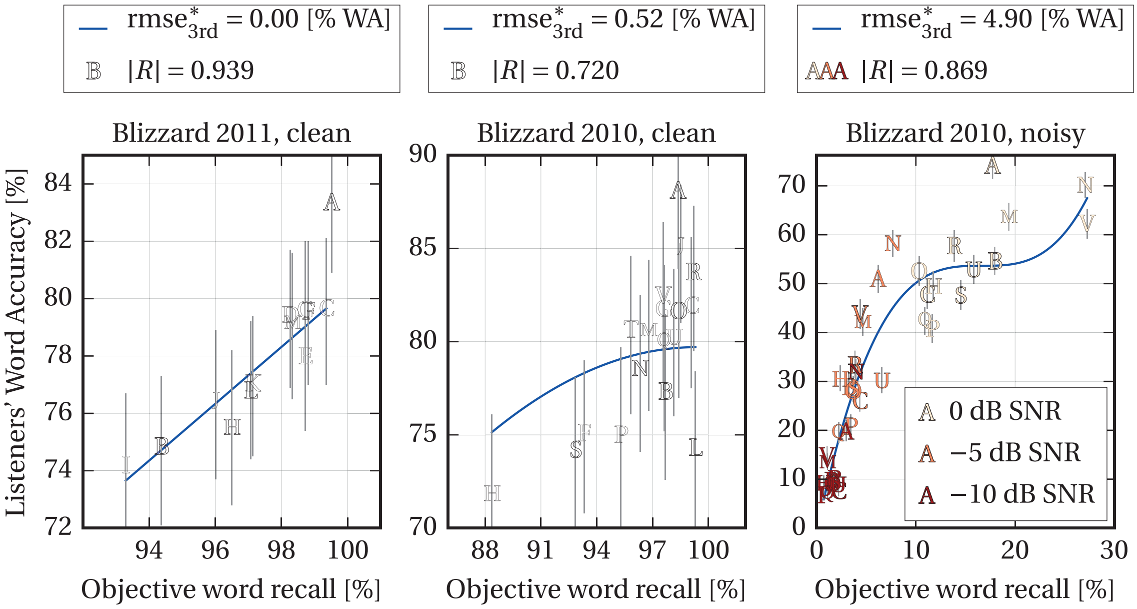


Figure 6.6 – Prediction of per-voice intelligibility through utterance verification, using decision thresholds $\tau_{\text{dev}}$ from development data. Letters identify different voices, with "A" for natural speech and "B"–"V" referring to *different TTS systems* in the 2011 and 2010 Challenge. Error bars indicate 95% confidence intervals of subjective scores. Blue lines show the mapping function used to calculate the $\text{rmse}^*_{\text{3rd}}$. Note the change in axes scales between panels.

We select the threshold value that minimizes the overlap of the fitted Beta distributions for each hypothesis. The resulting decision thresholds $\tau^{2011}_{\text{hypo}} = 0.99$ and $\tau^{2010}_{\text{hypo}} = 1.20$ are close to the values obtained with the development sets, but required no subjectively scored data.

### 6.5.2 Prediction performance

Figure 6.6 shows the objective word recall computed with the proposed approach, using the thresholds $\tau_{\text{dev}}$ derived from the development sets. The utterance verification approach allows to also assess the intelligibility of clean natural speech recordings, denoted system "A" in the left and center panel of Figure 6.6. However, we still computed the performance metrics on top of these two panels without considering system "A", in order to allow a comparison to the earlier results of the template-based approach. As observed in Section 6.3, the performance metric $\text{rmse}^*_{\text{3rd}}$ shows its limitation when evaluating predictions for clean synthetic speech. However, the metric reveals a significant reduction in prediction error for the noisy synthetic speech conditions (an overview of all results will be given on page 92).

Looking first at results for the clean and noisy Blizzard 2010 conditions (center and right panel, respectively), we can observe that the faster speech of system "H" is no longer penalized compared to other voices. This improvement over the template-based approach indicates that the KL-HMM reference provides more flexibility in aligning speech at different tempos.

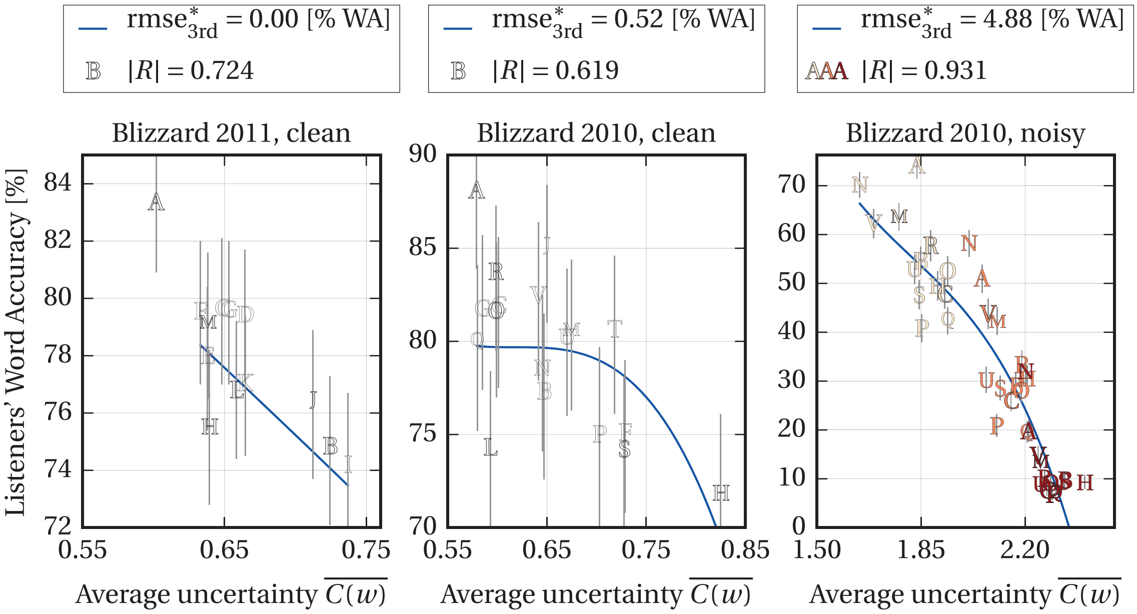


Figure 6.7 – Prediction of per-voice intelligibility through utterance verification, without thresholding. Letters identify different voices, with "A" for natural speech and "B"–"V" referring to *different TTS systems* in the 2011 and 2010 Challenge. Error bars indicate 95% confidence intervals of subjective scores. Blue lines show the mapping function used to calculate the $\text{rmse}^*_{\text{3rd}}$.

The right panel of Figure 6.6 also shows that the natural system "A" recordings are no longer favored over synthetic speech, now correctly predicting the higher intelligibility of system "N" recordings at negative SNRs. Both results can be attributed to the better ability of the KL-HMM system to accommodate pronunciation variability.

Contrary to our earlier observations on the relation between word recall and subjective word accuracy, the recall for the noisy speech conditions in the right panel of Figure 6.6 is in a *lower* numerical range than subjective scores. This may be due to the used threshold $\tau$, which was selected with clean development data, or to a more systematic issue with the proposed approach. In order to clarify this point, we computed results without thresholding, using the average uncertainty of words per sentence $\overline{C(w)} = \frac{1}{M}\sum_{m=1}^{M} C(w_m)$ as the objective score.

The resulting predictions are shown in Figure 6.7. Comparing clean vs. noisy conditions (left and center vs. right panel in Figure 6.7, respectively), we can see that clean conditions with subjective WA scores of ~ 70% have almost half the average uncertainty score $\overline{C(w)}$ of comparable noisy conditions. This points to a general discrepancy in the way that the proposed approach assesses different distortion types. Specifically, intelligibility impairments due to irregularities in pronunciation or prosody appear to be underestimated compared to those that are due to noise. This can also be seen with the 0 dB SNR conditions, where the spread of data points around the mapping function becomes larger (bright letters in the right panels of figures 6.6 and 6.7). At this SNR, intelligibility impairments inherent to the voices become

Table 6.2 – Performance of proposed approaches on three subsets of the Blizzard Challenge. "2010c" and "2010n" refer to clean and noisy speech data in the 2010 Challenge, respectively. Best performances are highlighted in boldface for both the set of TTS systems (top section) and all voices (bottom section). Connections with asterisks (—∗— / —∗∗—) indicate significant differences between prediction errors ($p < 0.05$ / $p < 0.01$ with Holm-Bonferroni correction).

| Approach | Correlation $\|R\|$ | | | Error $\text{rmse}^*_{\text{3rd}}$ [% WA] | | | Correct rank-orders | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2011 | 2010c | 2010n | 2011 | 2010c | 2010n | 2011 | 2010c | 2010n |
| *TTS systems only* | | | | | | | | | |
| Template-based distance $D_{\text{DTW}}$ | 0.897 | 0.679 | 0.814 | **0.00** | **0.38** | 9.11 | **100**% | **89**% | 86% |
| Utt. verification, uncertainty $\overline{C(w)}$ | 0.724 | 0.619 | **0.931** | **0.00** | 0.52 | **4.88** | 93% | 78% | **87**% |
| Utt. verification, recall with $\tau_{\text{dev}}$ | **0.939** | **0.720** | 0.869 | **0.00** | 0.52 | 4.90 | **100**% | 85% | 78% |
| Utt. verification, recall with $\tau_{\text{hypo}}$ | 0.821 | **0.720** | 0.857 | **0.00** | 0.51 | 6.41 | **100**% | 85% | 73% |
| *TTS systems + natural speech* | | | | | | | | | |
| Template-based distance $D_{\text{DTW}}$ | | | 0.811 | | | 9.86 | | | **89**% |
| Utt. verification, uncertainty $\overline{C(w)}$ | 0.790 | 0.654 | **0.921** | **0.00** | **1.02** | 5.56 | 96% | 86% | 88% |
| Utt. verification, recall with $\tau_{\text{dev}}$ | **0.891** | 0.693 | 0.863 | **0.00** | 1.08 | **5.52** | **100**% | 83% | 81% |
| Utt. verification, recall with $\tau_{\text{hypo}}$ | 0.753 | **0.715** | 0.851 | 0.13 | 1.05 | 6.78 | 92% | **88**% | 78% |

more apparent, but the proposed approach underestimates their importance relative to noise.

Since the KL-HMM does not enforce durational constraints other than the minimum of three frames per phoneme, we hypothesize that temporal artifacts in synthetic speech, e.g., disfluencies or concatenation errors, are insufficiently penalized with this type of reference. Informal listening also suggests this to be a reason for the clear offset between natural speech, i.e., system "A", and synthetic voices in clean conditions (left and center panel in figures 6.6 and 6.7), where the prosody of the professional voice talent may have helped listeners understand certain sentence structures or words.

The performance of all proposed approaches is summarized in Table 6.2. Performance metrics are shown both for synthetic speech conditions only (top section), allowing a comparison to the template-based approach of Section 6.3, and for all voices including system "A" (bottom section). Given the frequent saturation of the metric $\text{rmse}^*_{3\text{rd}}$ to zero, the last column header in Table 6.2 also evaluates the proportion of correctly predicted rank-orders, relative to the number of significant differences in subjective scores, as described in Section 6.3.1.

For both clean speech datasets (denoted “2011” and “2010c”), the template-based distance score $D_{\text{DTW}}$ provides the best performance in terms of prediction error and correct rank-orders (top section of Table 6.2). This result can be explained by the relevance of temporal structure at high intelligibility levels, which is implicitly modeled with the natural speech reference template and the path constraints for DTW alignment. Nevertheless, the proposed utterance verification approach achieves a performance that is very close to that of the template-based distance measure.

For the noisy speech data (denoted “2010n”), the utterance verification approach yields a highly significant reduction in prediction errors over the template-based measure (top and bottom sections of Table 6.2). As discussed earlier, we attribute this gain to the more flexible alignment of the KL-HMM reference to speech with different pronunciations or tempo. The impact of these pronunciation variabilities on intelligibility presumably becomes secondary at SNRs ≤ 0 dB, and is thus overestimated with the template-based approach.

Regarding the effect of thresholding in the utterance verification approach (i.e., $\overline{C(w)}$ vs. the recall with $\tau_{\text{dev}}$), it appears that thresholding was beneficial for assessing clean speech, but not for noisy conditions. Given the observed discrepancy between uncertainty scores $C(w_m)$ for clean and noisy conditions, thresholds $\tau$ derived from clean development data may not be appropriate for assessing noise-corrupted speech. Moreover, the benefit for clean speech is in line with the localized nature of speech synthesis distortions, which only affect the recall of the relevant word when thresholding, instead of a distance score for the entire sentence.

Finally, results obtained with either threshold $\tau_{\text{dev}}$ and $\tau_{\text{hypo}}$ are quite similar, although (unsurprisingly) performance with $\tau_{\text{dev}}$ is higher overall.[5] This implies that the proposed approach could indeed be used to assess data without subjective scores for threshold calibration.

## 6.6 Discussion and conclusion

We have shown that a comparison of phoneme posterior probability sequences can be used to assess the intelligibility of both clean and degraded synthetic speech. This result complements our findings in Chapter 5, where the same approach was shown to predict the intelligibility of natural speech in low bit-rate coding and background noise conditions. The proposed approach thus provides a highly versatile method for assessing the intelligibility of speech, based on a measure of mismatch to listeners’ phonetic and lexical knowledge, as modeled by a reference.

[5] The highly significant differences for the 2011 dataset in the bottom section of Table 6.2 can be seen as artifacts of the F-test of equal variances, where any difference to a prediction error of zero will be evaluated as significant.

Our experiments in this chapter evaluated two different types of reference, one consisting of a highly intelligible natural speech recording of the same words, and one using a model that was trained on multiple such recordings. A motivation for the latter reference type is that it is more flexible with regard to pronunciation variabilities at the phonetic and temporal level, as may be expected with different realizations of (synthetic) speech. Given the availability of a sentence transcription, the use of a KL-HMM reference also allows to assess intelligibility at the word level. The resulting word recall provides a measure that is consistent with the subjective evaluation of sentence intelligibility, and is considerably simpler to implement than a full-fledged automatic speech recognition (ASR) system. While the subjective assessment of intelligibility requires specially designed speech material to reduce context (e.g., semantically unpredictable sentences or rhyme words), our objective measure is not limited to particular sentence structures.

On the Blizzard Challenge data, our evaluations showed that either reference type can provide satisfactory predictions of clean and noisy synthetic speech intelligibility. However, we also observed a discrepancy or offset when comparing objective scores between datasets, with the template- and the model-based approach respectively over- and underestimating the impact of speech synthesis distortions relative to that of noise. It is thus likely that either approach would yield worse prediction performance for synthetic speech in moderate background noise (i.e., SNRs > 0 dB), where intelligibility impairments would not be dominated by a single distortion type.

Much of the analysis in this chapter was more qualitative than quantitative, due to the limitations of standard performance metrics. In particular, the modified prediction error $\mathrm{rmse}^{*}_{\mathrm{3rd}}$ cannot reveal performance improvements when differences between conditions become as small as the variability in listener scores. An evaluation at the file level (i.e., predicting the intelligibility of individual TTS *recordings*) was not feasible due to the “Latin square” design used by the Blizzard Challenge organizers, where each recording is rated by a very small subset of listeners, resulting in per-file 95% confidence intervals of up to 30% WA. However, the availability of scores from multiple listeners still allows to reveal small but significant differences between TTS systems. These subjective scores thus hold higher statistical power than an objective score modeling a single average listener.

In the next chapter, we review the overall conclusions of this thesis, and discuss possible ways of addressing the limitations we have found, as well as other avenues of research.

# 7 Conclusions and directions for future research

## 7.1 Conclusions

We have studied the objective assessment of two quality features in speech telecommunications: perceived background noise intrusiveness, and speech intelligibility. With regard to background noise perception, our collected data revealed that listeners assess noise intrusiveness by rating the noise signal alone, without considering its relation to foreground speech, or particular expectations to the noise itself. Specifically, signal bandwidth and the presence of Lombard speech had no significant effect on noise intrusiveness scores, whereas higher presentation levels resulted in significantly higher perceived intrusiveness, despite unchanged SNRs to speech. These results imply that listeners were capable of assessing noise independently of speech, as instructed in the ITU-T Rec. P.835 test method, and thus highlight the importance of considering the exact task that is given to listeners before conducting a subjective test.

Building on these results, we have proposed to objectively assess noise intrusiveness by analyzing the noise signal alone. We have shown that a sparse representation of noise, computed in a basis of cochlear filter shapes, models several effects of noise perception, such as the effects of sound pressure, noise type and spectral distribution. This representation only exploits knowledge of physiologically measured filter shapes, but does not require training data from subjective tests. When aggregated over the length of the analyzed noise signal, the number of filters or atoms in this representation provides a measure of noise intrusiveness that shows very good agreement with subjective scores. Moreover, the obtained prediction performance remains stable in case of deviations from the original hyperparameter values. Together with the analyzed interdependency of speech distortion, noise intrusiveness and overall quality scores, these results provide a comprehensive foundation for the objective assessment of noise reduction in telecommunications.

With respect to speech intelligibility, we have shown that a comparison of phoneme posterior probability sequences between reference and test speech can serve as a basis for the assessment of a wide range of distortion types. Specifically, we have demonstrated good agreement with subjective intelligibility scores for speech distorted by background noise, acoustic impairments and low bit-rate coding, as well as with synthetic speech. The proposed approach is consistent with the subjective assessment of speech intelligibility, in that it evaluates the mismatch to listeners' modeled phonetic and lexical knowledge, but seeks to ignore differences at the spectral or temporal level that are more related to speaker characteristics. This latter insensitivity was partly achieved using an artificial neural network (ANN) that was trained on conversational telephone speech from several thousand speakers. However, it was also shown that training ANNs to provide more accurate estimations of phonetic content does not necessarily yield more accurate intelligibility predictions. Furthermore, the absence of noise during ANN training also appears to result in an underestimation of intelligibility with low noise levels.

In the case of synthetic speech, switching from a reference speech template to a model appears to provide even greater insensitivity to speaker characteristics especially at the temporal level, while also allowing for an automatic segmentation of individual words in a test speech signal. However, our analysis indicates that the allowable degree of temporal variability can be hard to model, and that the proposed approach may not properly predict the relative importance of different impairments (i.e., background noise vs. speech synthesis distortions).

## 7.2 Directions for future research

The work in this thesis could be further developed along several lines:

- The objective assessment of noise intrusiveness in Chapter 4 only considered noise during speech pauses. This was motivated by the fact that listeners exclusively focused on the noise signal for their noise intrusiveness rating, and that foreground speech may mask background noise. However, strong fluctuations in noise level may still be perceptible during speech activity, and thus impact listener scores. Such fluctuations could be detected by exploiting the availability of a reference speech signal to perform a spectral subtraction of foreground speech, and recover the noise signal even during active speech segments.

- Our evaluation in Chapter 4 showed that the proposed spike density feature significantly outperforms or compares to a traditional loudness-based feature. On the other hand, loudness estimations are based on low-level models of the inner ear that have been extensively studied and validated for a wide range of sounds. This feature can therefore be applied confidently for new applications and unseen conditions. By contrast, our novel feature has only been validated for environmental noises in the very specific context of noise intrusiveness in telecommunications. In particular, this context implied a restricted bandwidth and dynamic range of noises. There is thus a need to further link low-level acoustic phenomena in hearing with the higher-level perceptual model developed here.

- The ANNs used for intelligibility prediction in chapters 5 and 6 were trained to classify English phonemes, making the proposed approaches language dependent. While this limitation could be addressed by training ANNs to classify multilingual phones, the inclusion of finer phonetic categories may also result in an objective measure that is sensitive to minor acoustic variations that are not relevant to intelligibility.

- Our experiments on synthetic speech in Chapter 6 highlighted the importance of minimum duration constraints for intelligibility assessment. Such constraints could be taken into account by combining template- and model-based references. The use of a reference template could also help detect pronunciation errors in synthetic speech that are due to errors or omissions in pronunciation dictionaries, and which could not be detected with the KL-HMM reference.

- An advantage of the KL-HMM reference proposed in Chapter 6 is that it can be used to generate a reference sequence of phoneme posterior probabilities for an arbitrary text input, including words that were not part of the training data. This has not been evaluated in this work, but may be particularly useful for synthetic speech assessment, where the correct synthesis of a very large number of sentences could be verified.

- Finally, our evaluation of objective scores in Chapter 6 also revealed limits of standardized performance metrics, especially for datasets with small differences in subjective scores between conditions. While the availability of subjective ratings from multiple listeners still allowed to determine significant differences between conditions, a single objective prediction does not offer comparable statistical power. This could be addressed by modeling multiple listeners, using a distribution of different decision thresholds $\tau$ in our proposed approach, e.g., to model listeners with different levels of linguistic knowledge or listening environments.

# A Dataset descriptions

This section provides technical details on the design, recording, processing and subjective evaluation of speech recordings in the "PANDA" datasets presented in Chapter 3, page 23.

## A.1 PANDA dataset designs

### A.1.1 Sentence material and corresponding background noise signals

Table A.1 lists the 30 sentences recorded for the three datasets, with corresponding background noises used for eliciting the Lombard effect. All noises were drawn from ETSI EG 202 396-1 [2011], with the exception of (self-generated) pink noise.

Table A.1 – Sentences and noises used while recording speech for the three PANDA datasets. The same noise was used for consecutive groups of three sentences.

| Sentence (in French) | Noise type |
|---|---|
| Est-ce qu'il reste des places de parc dans ta rue? | Crossroad |
| Guillaume est avec moi, Virginie vient en bus. | |
| Nous allons prendre des boissons au supermarché. | |
| Hélas, notre train a pris un important retard. | Train |
| Nous roulons actuellement à vitesse réduite. | |
| Le trajet se prolongera d'une quinzaine de minutes. | |
| Dépêche-toi de venir à l'apéro de Christophe. | Schoolyard |
| Il va bientôt recevoir son cadeau d'adieu. | |
| Magalie lui a spécialement préparé une tarte. | |

Continued on next page

Table A.1 – continued from previous page

| Sentence (in French) | Noise type |
|---|---|
| Nous aurons du retard à la fête de mariage. | Car |
| Notre GPS n’a pas indiqué le bon chemin. | |
| Nous avons pris la route du lac dans le mauvais sens. | |
| Je suis allé dîner avec Olivier et Chantale. | Cafeteria |
| Nous sommes partis en avance pour éviter la queue. | |
| Nous continuerons le travail cet après-midi. | |
| Je viens de sortir à l’instant de mon bureau. | Office |
| Sa formation est encore loin d’être achevée. | |
| Rappelez-moi demain à huit heures s’il-vous-plaît. | |
| Ils font du sport sur le gazon, près du grand chêne. | Road |
| Céline et Marc devraient nous rejoindre d’ici une heure. | |
| Ils ont dit qu’ils apporteraient un nouveau jeu de cartes. | |
| Est-ce qu’il nous reste du fromage râpé à la maison? | Pink |
| Les pizzas fraîches sont en action, j’en ai pris cinq. | |
| Il n’y a plus de feuilles de basilic en vente là-bas. | |
| Les gens profitent de l’entracte pour venir discuter. | Pub |
| Les acteurs ont bien joué cette ancienne comédie. | |
| Les billets sur internet peuvent être achetés plus tôt. | |
| Je ne suis pas encore arrivé à la maison. | Jackhammer |
| La conférence a dépassé l’horaire prévu. | |
| Les organisateurs n’espéraient pas ce succès. | |

### A.1.2 Technical details of the recording setup

All speech recordings took place in the psychoacoustic chamber of the Laboratory of Electromagnetics and Acoustics (LEMA) at EPFL. The following equipment was used:

**Microphone** Schoeps MK2 microphone capsule (omnidirectional, free sound field, flat frequency response) with CMC 5-U microphone amplifier and foam pop filter.

**Sound card** MOTU 896 mk3 FireWire sound card with 48 V phantom power. A separate 48 V phantom power source was used for the Lombard speech recordings. The microphone signal was digitized with a 48 kHz sampling rate and 24-bit resolution.

**Headphones** Two pairs of Beyerdynamic DT 770 Pro closed headphones were used, one for monitoring, and one for playing back noises to speakers while collecting Lombard speech recordings.

**Computer** PC running Windows 7 and Adobe Audition 3, connected to the sound card via a FireWire cable. Recordings were stored as uncompressed, mono WAV files with a 48 kHz sampling rate and 32-bit resolution.

The first recording session (regular speech) took place on 2012-06-13. Speakers kindly accepted to be “paid” by being offered a slice of fruit pie. The second session (Lombard speech) took place between 2012-06-19 and 2012-06-21 with the same speakers. Speakers were paid CHF 40 for their participation in this second session.

For Lombard speech recordings, a TEAC analog mixer was used to combine the speech and background noise signals. More specifically, the microphone signal was fed back to the headphones worn by the speakers, such as to avoid an “earplug” effect due to the closed shape of the headphone earpiece. The background noises used to provoke the Lombard effect were then mixed over this signal. In order to get accustomed to wearing closed headphones, speakers were first asked to speak all sentences with the headphones on. During this first round, no noises were played back on the headphones. Speakers were asked to position themselves at a distance to the microphone (typically 10–20 cm) such as to achieve a sound level of the feedback signal equivalent to hearing themselves when not wearing headphones.

In a second round, the speakers pronounced all sentences while different background noises were being played back through their headphones (and mixed with the microphone feedback signal). Speakers were given a visual cue to start speaking only after an adaptation period of 8 s into the noise signal, and then read the 3 sentences of the corresponding group. The noise signals were processed with a fade-in effect to avoid discomforting speakers with a sudden onset of loud noises.

### A.1.3 Test plans

Tables A.2, A.3 and A.4 list all conditions in the three PANDA datasets. Italicized condition numbers indicate conditions that are shared between multiple sets. Signal-to-noise ratios refer to the noise level after input filtering, but before further processing (e.g., noise reduction). The presentation level of recordings in Set 1 is shown in the last column of Table A.2. For the two other sets, the presentation level was fixed at 79 dB SPL (diotic) for all conditions. Each condition consists of 12 distinct recordings (3 sentences × 4 speakers).

Table A.2 – Test plan for PANDA Set 1.

| Cond. # | Speech type | Input filter | Noise | SNR [dB] | NR | Codec | Simul. / live | Comment | Level [dB SPL] |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Regular | SWB | – | – | – | – | Simul. | – | 79 |
| 2 | Regular | SWB | Crossroad | 30 | – | – | Simul. | – | 79 |
| 3 | Regular | SWB | Crossroad | 20 | – | – | Simul. | – | 79 |
| 4 | Regular | SWB | Crossroad | 10 | – | – | Simul. | – | 79 |
| 5 | Lombard | SWB | – | – | – | – | Simul. | – | 79 |
| 6 | Lombard | SWB | Crossroad | 20 | – | – | Simul. | – | 79 |
| 7 | Lombard | SWB | – | – | – | – | Simul. | – | 87 |
| 8 | Lombard | SWB | Crossroad | 20 | – | – | Simul. | – | 87 |
| 9 | Lombard | SWB | Crossroad | 30 | – | – | Simul. | – | 87 |
| 10 | Regular | NB | Cafeteria | 19 | – | AMR-NB | Live | – | 79 |
| 11 | Lombard | NB | Cafeteria | 19 | – | AMR-NB | Live | – | 84 |
| 12 | Regular | NB | Train | 19 | – | AMR-NB | Live | – | 79 |
| 13 | Lombard | NB | Train | 19 | – | AMR-NB | Live | – | 82 |
| 14 | Regular | WB | Pub | 4 | – | AMR-WB, 12.65 kbps | Simul. | – | 79 |
| 15 | Lombard | WB | Pub | 4 | – | AMR-WB, 12.65 kbps | Simul. | – | 79 |
| 16 | Regular | WB | – | – | – | 2× AMR-WB, 6.6 kbps | Simul. | – | 79 |
| 17 | Lombard | WB | – | – | – | 2× AMR-WB, 6.6 kbps | Simul. | – | 79 |

Continued on next page

Table A.2 (Test plan for PANDA Set 1) – continued from previous page

| Cond. # | Speech type | Input filter | Noise | SNR [dB] | NR | Codec | Simul. / live | Comment | Level [dB SPL] |
|---|---|---|---|---|---|---|---|---|---|
| 18 | Regular | WB | Pub | 4 | – | AMR-WB, 12.65 kbps | Simul. | Amplitude clipping before codec | 79 |
| 19 | Lombard | WB | Pub | 4 | – | AMR-WB, 12.65 kbps | Simul. | Amplitude clipping before codec | 79 |
| 20 | Regular | NB | Car | 18 | In handset | EVRC-B | Live | – | 79 |
| 21 | Lombard | NB | Car | 18 | In handset | EVRC-B | Live | – | 79 |
| 22 | Regular | WB | Cafeteria | 17 | Inside codec | EVRC-WB, COP 0 | Simul. | – | 79 |
| 23 | Lombard | WB | Cafeteria | 17 | Inside codec | EVRC-WB, COP 0 | Simul. | – | 87 |
| 24 | Regular | WB | – | – | Inside codec | EVRC-WB, COP 0 | Simul. | – | 79 |
| 25 | Regular | NB | – | – | – | AMR-NB | Live | Channel errors | 79 |
| 26 | Lombard | NB | – | – | – | AMR-NB | Live | Channel errors | 79 |
| 27 | Regular | NB | Office | 16 | In device (conf. phone) | AMR-NB | Live | Acoustical insertion | 84 |
| 28 | Lombard | NB | Office | 16 | In device (conf. phone) | AMR-NB | Live | Acoustical insertion | 84 |
| 29 | Lombard | WB | – | – | Spectral subtractive | – | Simul. | – | 79 |
| 30 | Regular | WB | Jackhammer | 7 | Spectral subtractive | – | Simul. | – | 79 |
| 31 | Regular | WB | Jackhammer | 7 | Parametric Wiener filter | – | Simul. | – | 79 |
| 32 | Regular | NB | Train station | 15 | – | GSM-HR | Simul. | – | 79 |
| 33 | Lombard | NB | Train station | 15 | – | GSM-HR | Simul. | – | 79 |

Table A.3 – Test plan for PANDA Set 2.

| | Speech | Input | | SNR | | | Simul. | |
|---|---|---|---|---|---|---|---|---|
| Cond. # | type | filter | Noise | [dB] | NR | Codec | / live | Comment |
| 34 | Regular | SWB | – | – | – | – | Simul. | Different sentences than in Set 1 |
| 35 | Regular | SWB | Crossroad | 40 | – | – | Simul. | – |
| *2* | Regular | SWB | Crossroad | 30 | – | – | Simul. | – |
| *3* | Regular | SWB | Crossroad | 20 | – | – | Simul. | – |
| *4* | Regular | SWB | Crossroad | 10 | – | – | Simul. | – |
| 39 | Lombard | NB | Road | 12 | – | AMR-NB, 10.2 kbps | Simul. | Noise modulated at 2–3 Hz |
| 40 | Lombard | NB | Road | 12 | – | AMR-NB, 10.2 kbps | Simul. | Noise modulated at 3–5 Hz |
| 41 | Lombard | WB | Cafeteria | 17 | Parametric Wiener filter | – | Simul. | Noise raised back to level before NR |
| 42 | Lombard | WB | Cafeteria | 17 | – | – | Simul. | – |
| 43 | Lombard | WB | Cafeteria | 17 | Third-party | – | Simul. | – |
| 44 | Lombard | WB | Cafeteria | 17 | – | – | Simul. | Time-reversed noise |
| 45 | Regular | NB | – | – | – | AMR-NB | Live | – |
| 46 | Lombard | NB | Car | 18 | In handset | AMR-NB | Live | – |
| 47 | Regular | WB | – | – | – | AMR-WB | Live | – |
| 48 | Regular | WB | – | – | – | AMR-WB | Live | Channel errors |
| 49 | Lombard | WB | Pub | 4 | In handset | AMR-WB | Live | – |
| 50 | Lombard | WB | Pub | 4 | In handset | AMR-WB | Live | Channel errors |

Continued on next page

Table A.3 (Test plan for PANDA Set 2) – continued from previous page

| Cond. # | Speech type | Input filter | Noise | SNR [dB] | NR | Codec | Simul. / live | Comment |
|---|---|---|---|---|---|---|---|---|
| 51 | Lombard | WB | Pub | 4 | – | AMR-WB, 12.65 kbps | Simul. | Frequency response of cond. 47 |
| 52 | Lombard | WB | Road | 10 | – | AMR-WB, 12.65 kbps | Simul. | Frequency response of cond. 47 |
| 53 | Lombard | WB | Train | 17 | – | AMR-WB, 12.65 kbps | Simul. | Frequency response of cond. 47 |
| 54 | Lombard | WB | Jackhammer | 7 | – | AMR-WB, 12.65 kbps | Simul. | Frequency response of cond. 47 |
| 55 | Regular | WB | – | – | Spectral subtractive | – | Simul. | Noise estimation from “Pub” noise |
| 56 | Regular | WB | – | – | Parametric Wiener filter | – | Simul. | Noise estimation from “Pub” noise |
| 57 | Lombard | WB | Office | 15 | – | AMR-WB, 12.65 kbps | Simul. | – |
| 58 | Lombard | WB | Office | 15 | – | AMR-WB, 12.65 kbps | Simul. | Different sentences |
| 59 | Lombard | NB | Pub | 5 | – | AMR-NB | Live | In-handset NR disabled |
| 60 | Lombard | NB | Road | 11 | – | AMR-NB | Live | In-handset NR disabled |
| 61 | Lombard | NB | Train | 19 | – | AMR-NB | Live | In-handset NR disabled |
| 62 | Lombard | NB | Jackhammer | 9 | – | AMR-NB | Live | In-handset NR disabled |
| 63 | Regular | SWB | – | – | – | Skype | Live | Strong delay jitter |
| *16* | Regular | WB | – | – | – | 2× AMR-WB, 6.6 kbps | Simul. | – |
| 65 | Regular | NB | – | – | – | 3× AMR-NB, 4.75 kbps | Simul. | – |

Table A.4 – Test plan for PANDA Set 3.

| Cond. # | Speech type | Input filter | Noise | SNR [dB] | NR | Codec | Simul. / live | Comment |
|---|---|---|---|---|---|---|---|---|
| 66 | Regular | NB | – | – | – | – | Simul. | – |
| 67 | Regular | NB | Crossroad | 40 | – | – | Simul. | – |
| 68 | Regular | NB | Crossroad | 30 | – | – | Simul. | – |
| 69 | Regular | NB | Crossroad | 20 | – | – | Simul. | – |
| 70 | Regular | NB | Crossroad | 10 | – | – | Simul. | – |
| 71 | Lombard | NB | Cafeteria | 19 | – | AMR-NB, 12.2 kbps | Simul. | – |
| 72 | Lombard | NB | Cafeteria | 19 | Third-party | AMR-NB, 12.2 kbps | Simul. | – |
| 73 | Lombard | NB | Cafeteria | 19 | – | AMR-NB, 12.2 kbps | Simul. | Time-reversed noise |
| 74 | Regular | NB | – | – | Wiener filter | – | Simul. | Noise estimation from cond. 76 |
| 75 | Regular | NB | – | – | Parametric Wiener filter | – | Simul. | Noise estimation from cond. 77 |
| 76 | Lombard | NB | Schoolyard | 16 | Wiener filter | – | Simul. | – |
| 77 | Lombard | NB | Train station | 15 | Parametric Wiener filter | – | Simul. | – |
| 78 | Regular | NB | – | – | – | Skype | Live | Strong delay jitter |
| 79 | Lombard | NB | Office | 16 | – | AMR-NB, 10.2 kbps | Simul. | – |
| 80 | Lombard | NB | Office | 16 | – | AMR-NB, 10.2 kbps | Simul. | Different sentences |
| 81 | Regular | NB | – | – | – | EVRC-NB | Live | – |
| *45* | Regular | NB | – | – | – | AMR-NB | Live | – |
| 83 | Lombard | NB | Pub | 5 | In handset | EVRC-NB | Live | – |

Continued on next page

Table A.4 (Test plan for PANDA Set 3) – continued from previous page

| | Speech | Input | | SNR | | | Simul. | |
|---|---|---|---|---|---|---|---|---|
| Cond. # | type | filter | Noise | [dB] | NR | Codec | / live | Comment |
| 84 | Lombard | NB | Train | 19 | In handset | EVRC-NB | Live | – |
| 85 | Lombard | NB | Jackhammer | 9 | In handset | EVRC-NB | Live | – |
| 86 | Lombard | NB | Pub | 5 | In handset | AMR-NB | Live | – |
| 87 | Lombard | NB | Road | 11 | In handset | AMR-NB | Live | – |
| 88 | Lombard | NB | Train | 19 | In handset | AMR-NB | Live | – |
| 89 | Lombard | NB | Jackhammer | 9 | In handset | AMR-NB | Live | – |
| *59* | Lombard | NB | Pub | 5 | – | AMR-NB | Live | In-handset NR disabled |
| *60* | Lombard | NB | Road | 11 | – | AMR-NB | Live | In-handset NR disabled |
| *61* | Lombard | NB | Train | 19 | – | AMR-NB | Live | In-handset NR disabled |
| *62* | Lombard | NB | Jackhammer | 9 | – | AMR-NB | Live | In-handset NR disabled |
| 94 | Lombard | NB | Road | 11 | – | AMR-NB | Live | In-handset NR disabled; Amplitude clipping before insertion |
| *25* | Regular | NB | – | – | – | AMR-NB | Live | Channel errors |
| *33* | Lombard | NB | Train station | 15 | – | GSM-HR | Simul. | – |
| *65* | Regular | NB | – | – | – | 3× AMR-NB, 4.75 kbps | Simul. | – |

### A.1.4 On-screen rating interface

Figure A.1 on page 110 shows the on-screen rating interface for the three quality features *speech distortion*, *noise intrusiveness* and *overall quality*. The interface is designed such that the previously given rating is no longer visible after moving to the next quality feature.

## A.2 PSCR word lists

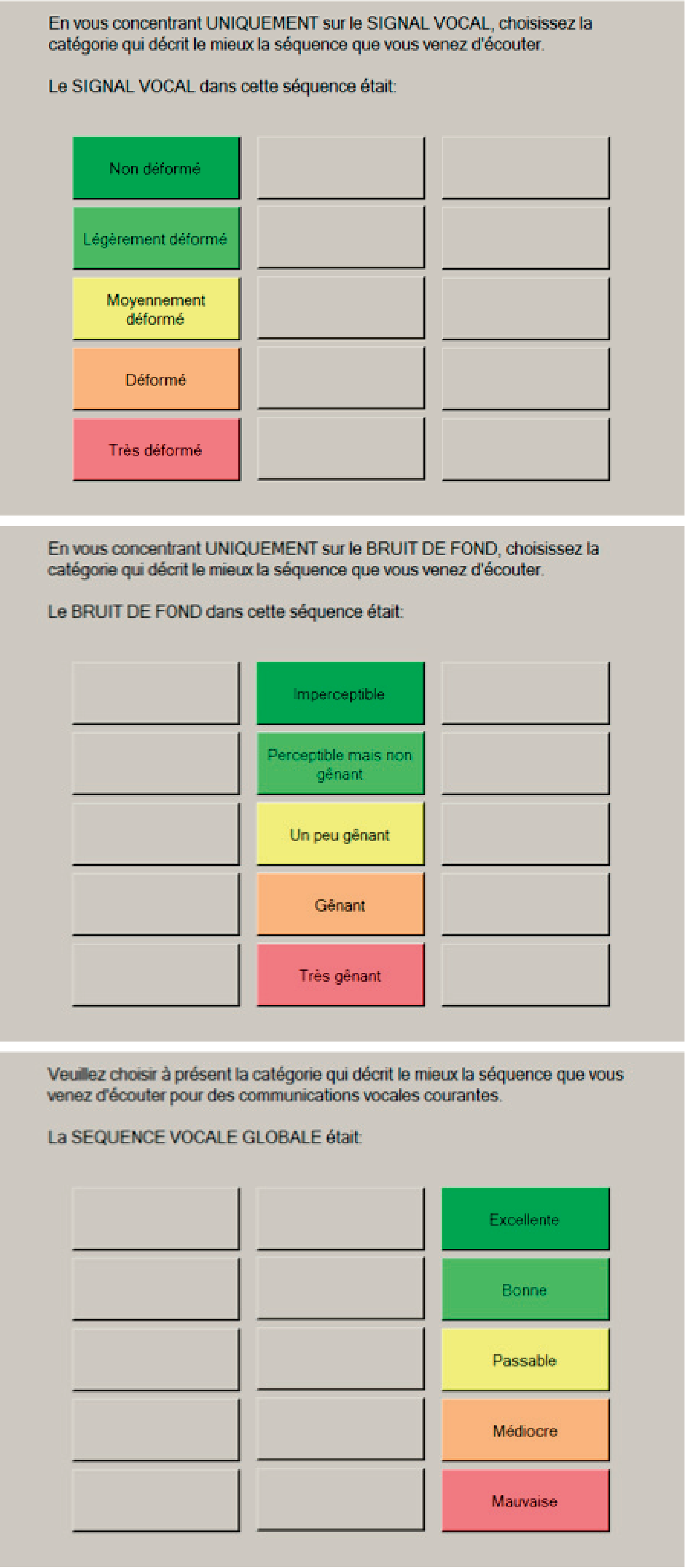
Table A.5 – Lists of rhyming words used in the PSCR [2013] datasets.

| list # | | | | | | |
|---|---|---|---|---|---|---|
| 1 | went | sent | bent | dent | tent | rent |
| 2 | hold | cold | told | fold | sold | gold |
| 3 | pat | pad | pan | path | pack | pass |
| 4 | lane | lay | late | lake | lace | lame |
| 5 | kit | bit | fit | hit | wit | sit |
| 6 | must | bust | gust | rust | dust | just |
| 7 | teak | team | teal | teach | tear | tease |
| 8 | din | dill | dim | dig | dip | did |
| 9 | bed | led | fed | red | wed | shed |
| 10 | pin | sin | tin | fin | din | win |
| 11 | dug | dung | duck | dud | dub | dun |
| 12 | sum | sun | sung | sup | sub | sud |
| 13 | seep | seen | seethe | seek | seem | seed |
| 14 | not | tot | got | pot | hot | lot |
| 15 | vest | test | rest | best | west | nest |
| 16 | pig | pill | pin | pip | pit | pick |
| 17 | back | bath | bad | bass | bat | ban |
| 18 | way | may | say | pay | day | gay |
| 19 | pig | big | dig | wig | rig | fig |
| 20 | pale | pace | page | pane | pay | pave |
| 21 | cane | case | cape | cake | came | cave |
| 22 | shop | mop | cop | top | hop | pop |
| 23 | coil | oil | soil | toil | boil | foil |
| 24 | tan | tang | tap | tack | tam | tab |

Continued on next page

Table A.5 – continued from previous page

| list # | | | | | | |
|---|---|---|---|---|---|---|
| 25 | fit | fib | fizz | fill | fig | fin |
| 26 | same | name | game | tame | came | fame |
| 27 | peel | reel | feel | eel | keel | heel |
| 28 | hark | dark | mark | bark | park | lark |
| 29 | heave | hear | heat | heal | heap | heath |
| 30 | cup | cut | cud | cuff | cuss | cub |
| 31 | thaw | law | raw | paw | jaw | saw |
| 32 | pen | hen | men | then | den | ten |
| 33 | puff | puck | pub | pus | pup | pun |
| 34 | bean | beach | beat | beak | bead | beam |
| 35 | heat | neat | feat | seat | meat | beat |
| 36 | dip | sip | hip | tip | lip | rip |
| 37 | kill | kin | kit | kick | king | kid |
| 38 | hang | sang | bang | rang | fang | gang |
| 39 | took | cook | look | hook | shook | book |
| 40 | mass | math | map | mat | man | mad |
| 41 | ray | raze | rate | rave | rake | race |
| 42 | save | same | sale | sane | sake | safe |
| 43 | fill | kill | will | hill | till | bill |
| 44 | sill | sick | sip | sing | sit | sin |
| 45 | bale | gale | sale | tale | pale | male |
| 46 | wick | sick | kick | lick | pick | tick |
| 47 | peace | peas | peak | peach | peat | peal |
| 48 | bun | bus | but | bug | buck | buff |
| 49 | sag | sat | sass | sack | sad | sap |
| 50 | fun | sun | bun | gun | run | nun |



En vous concentrant UNIQUEMENT sur le SIGNAL VOCAL, choisissez la catégorie qui décrit le mieux la séquence que vous venez d'écouter.

Le SIGNAL VOCAL dans cette séquence était:

En vous concentrant UNIQUEMENT sur le BRUIT DE FOND, choisissez la catégorie qui décrit le mieux la séquence que vous venez d'écouter.

Le BRUIT DE FOND dans cette séquence était:

Veuillez choisir à présent la catégorie qui décrit le mieux la séquence que vous venez d'écouter pour des communications vocales courantes.

La SEQUENCE VOCALE GLOBALE était:

Figure A.1 – Rating interface for the collection of subjective P.835 scores.

# B Derivation of speech distortion from noise intrusiveness and overall quality

In Section 3.5, page 36, it was shown how speech distortion (S-MOS) and noise intrusiveness (N-MOS) could be combined to derive the overall quality score (G-MOS) in P.835 tests. The combination consisted of a simple linear model, shown here again for convenience:

$$s_i^G \approx a \cdot s_i^S + b \cdot s_i^N + c = \hat{s}_i^G, \tag{B.1}$$

with $s_i^{G/S/N}$ the G-, S- or N-MOS for the $i^{\text{th}}$ condition, respectively, and $a$, $b$ and $c$ the regression coefficients.

Assuming the availability of objective estimates $o_i^G$ and $o_i^N$ for G- and N-MOS, respectively, we can exploit the relation between scores to derive an objective estimate of S-MOS $o_i^S$. Rearranging (B.1) and omitting the subscript $i$ for clarity, we obtain:

$$o^S = \left(o^G - b \cdot o^N - c\right) / a. \tag{B.2}$$

In order to verify the feasibility of this approach, we have computed a common set of coefficients $a$, $b$ and $c$ through least squares regression for (B.1), by pooling subjective scores from the three PANDA datasets. We have then applied (B.2) to derive the S-MOS from the ground truth scores $s^G$ and $s^N$. Figure B.1 compares the resulting estimates to the actual S-MOS. The relation is highly linear, but there is a visible bias for conditions with N-MOS $< 2.5$. Due to the high noise intrusiveness in these conditions, the G-MOS remains in a low range, meaning that changes in speech distortion are underestimated.

We can account for this effect by switching to a quadratic model of the form

$$s^G \approx a \cdot s^S + b \cdot s^N + c + d\left(s^S\right)^2 + e\left(s^N\right)^2 + f\left(s^S \cdot s^N\right) = \hat{s}^G, \tag{B.3}$$

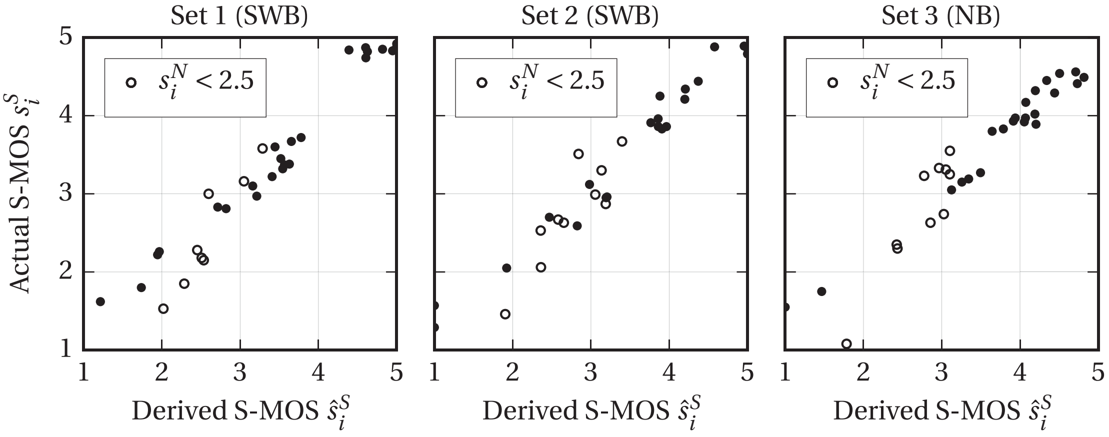


Figure B.1 – Linear regression of per-condition S-MOS from G- and N-MOS, using a common set of coefficients $a$, $b$ and $c$ for the three datasets. Circles show conditions with N-MOS < 2.5.

with additional constraints

$$\frac{\partial \hat{s}^G}{\partial s^S} = a + 2d \cdot s^S + f \cdot s^N \geq 0 \tag{B.4}$$

$$\frac{\partial \hat{s}^G}{\partial s^N} = b + 2e \cdot s^N + f \cdot s^S \geq 0 \tag{B.5}$$

to ensure that $\hat{s}^G$ in (B.3) remains monotonically increasing with respect to both $s^S$ and $s^N$. Indeed, it is reasonable to assume that an increase in either S- or N-MOS should be reflected by a higher G-MOS. Solving (B.3) for $s^S$, we obtain

$$\hat{s}_{1,2}^S = \frac{-\left(a + f \cdot s^N\right) \pm \sqrt{\left(a + f \cdot s^N\right)^2 - 4d \cdot \left(e\left(s^N\right)^2 + b \cdot s^N + c - s^G\right)}}{2d} \tag{B.6}$$

with the additional constraint that the determinant in (B.6) be positive, i.e.,

$$\left(a + f \cdot s^N\right)^2 - 4d \cdot \left(e\left(s^N\right)^2 + b \cdot s^N + c - s^G\right) \geq 0. \tag{B.7}$$

The derivation of speech distortion from noise intrusiveness and overall quality scores can thus be formulated as solving (B.3) with constraints (B.4), (B.5) and (B.7).

As before, we pool subjective scores from the three PANDA datasets to find a set of regression coefficients $a$ to $f$ for the quadratic model. The coefficients can be computed through constrained optimization, with (B.4), (B.5) and (B.7) as constraints, and the error rmse* between derived and actual speech distortion scores $\hat{s}_i^S$ and $s_i^S$, respectively, as objective function:

$$\text{rmse}^* = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \max\left(\left|s_i^S - \hat{s}_i^S\right| - \text{CI95}_i^S\,,\ 0\right)^2}, \tag{B.8}$$

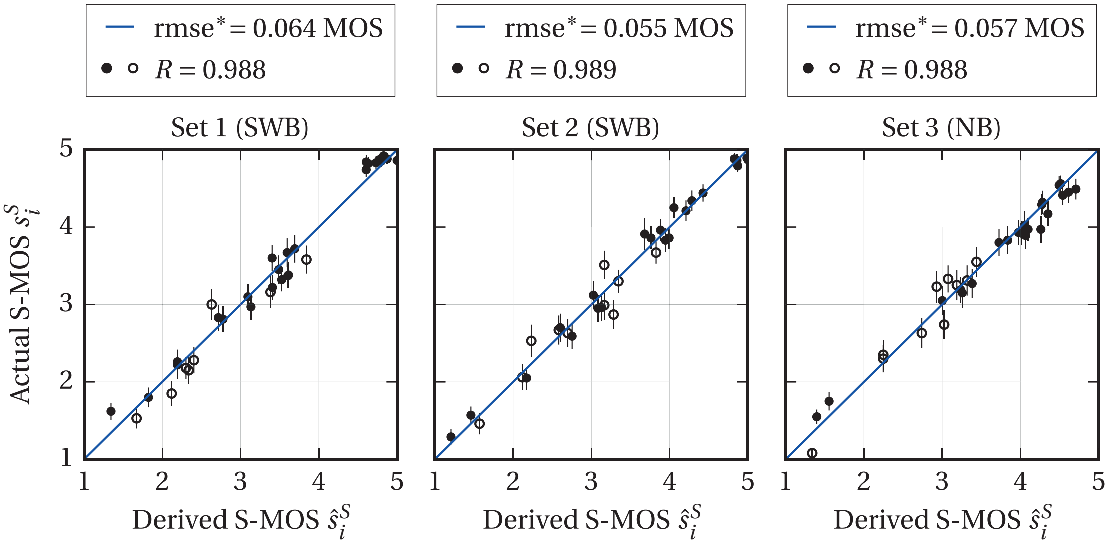


Figure B.2 – Quadratic regression of per-condition S-MOS from G- and N-MOS, using the same regression coefficients for each dataset. Circles show conditions with N-MOS < 2.5; the earlier bias has disappeared. Error bars show 95% confidence intervals. Values of $\hat{s}_i^S$ outside the range [1, 5] are clipped to the nearest value on the MOS scale (1 condition in Sets 1 and 2).

where $\text{CI95}_i^S$ is the 95% confidence interval of the $i^{\text{th}}$ subjective speech distortion score. The error metric rmse* is similar to the $\text{rmse}^*_{\text{3rd}}$ that was introduced in the Background chapter in Section 2.3.2, page 10, except that no further mapping of derived scores $\hat{s}_i^S$ is applied. The initial values for the coefficients $a$ to $f$ before optimization are determined through ordinary least-squares fitting, without consideration of constraints or confidence intervals.

In optimizing (B.8), it turns out that only the first solution $\hat{s}_1^S$ in (B.6) (positive sign in front of the determinant) provides results in the correct numerical range. The obtained coefficients $a$ to $f$ are given in Table B.1. The S-MOS retains a stronger influence on the G-MOS, i.e., $d > e$ in (B.3), as was the case with the linear model in Section 3.5. The S-MOS derived with the quadratic model is shown in Figure B.2. The earlier systematic bias for conditions with N-MOS < 2.5 is no longer visible. Moreover, the difference in the maximum S-MOS that we observed between super-wideband and narrowband signals (Sets 1&2 and Set 3, respectively) is preserved in the derived S-MOS.

Together, these results suggest that an objective assessment of speech distortion is possible, given objective scores for overall quality and noise intrusiveness.

Table B.1 – Coefficients for the quadratic regression in (B.3), using the pooled PANDA datasets.

| Datasets | $a$ | $b$ | $c$ | $d$ | $e$ | $f$ |
|---|---|---|---|---|---|---|
| PANDA Sets 1–3 | 0.397 | 0.390 | −0.160 | −0.029 | −0.052 | 0.131 |

# C Sparse approximation with matching pursuit

We gave a high-level introduction to the Matching Pursuit (MP) algorithm in Section 4.2.3, page 45, where it was used to compute a sparse approximation of a given input noise signal. Formally, the inputs to MP are the discrete-time signal $\mathbf{x}$ and a *dictionary* $\mathbf{\Phi}$ consisting of the set of unit-norm kernels $\{\boldsymbol{\phi}^m\}$, shifted at all possible time offsets within the length of $\mathbf{x}$, i.e.,

$$\mathbf{\Phi} = \left[\boldsymbol{\phi}_\tau^m\right], \quad m = 1, \ldots, M \quad \text{and} \quad \tau = 0, \ldots, N-L \tag{C.1}$$

with $\mathbf{x} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{\phi}^m \in \mathbb{R}^{L\times 1}$ and $\mathbf{\Phi} \in \mathbb{R}^{N\times M(N-L+1)}$. There exist efficient implementations that avoid explicitly storing time-shifted versions of the $M$ kernels [Krstulovic and Gribonval, 2006].

At each iteration $k$, MP projects the discrete-time residual $\mathbf{e} \in \mathbb{R}^{N\times 1}$ onto the columns of the dictionary $\mathbf{\Phi}$ and subtracts the projection with the highest correlation. The modification of MP by Gribonval [1999] used in this thesis allows to also adjust the phase $\varphi$ of gammatones in (4.2) by projecting onto analytic kernels $\boldsymbol{\phi}_a$ instead, resulting in the following calculation steps:

$$\boldsymbol{\phi}_{a,\tau}^m = \boldsymbol{\phi}_\tau^m + i \cdot \mathscr{H}\left(\boldsymbol{\phi}_\tau^m\right) \tag{C.2}$$

$$j(k) = \underset{m,\tau}{\operatorname{argmax}} \left\| \left\langle \boldsymbol{\phi}_{a,\tau}^m \, , \mathbf{e}_{(k-1)} \right\rangle \right\|^2 \tag{C.3}$$

$$\alpha_{j(k)} = \left\langle \boldsymbol{\phi}_{a,j(k)} \, , \mathbf{e}_{(k-1)} \right\rangle \tag{C.4}$$

$$\mathbf{e}_{(k)} = \mathbf{e}_{(k-1)} - \mathfrak{Re}\left(\alpha_{j(k)} \boldsymbol{\phi}_{a,j(k)}\right) \tag{C.5}$$

with $\mathscr{H}(\cdot)$ the Hilbert transform operator, and $j(k)$ the index of the kernel selected at the $k^{\text{th}}$ iteration in the matrix $\mathbf{\Phi}$. The projected signal at the first iteration is $\mathbf{e}_{(0)} = \mathbf{x}$. The signal approximation after $K$ iterations is then

$$\hat{\mathbf{x}}_{(K)} = \mathfrak{Re}\left(\mathbf{\Phi}_{a,J} \boldsymbol{\alpha}_J\right), \quad J = \left\{j(1), \ldots, j(k), \ldots, j(K)\right\} \tag{C.6}$$

with $\boldsymbol{\alpha} \in \mathbb{C}^{M(N-L+1)\times 1}$ the sparse vector containing the gains for each kernel occurrence or *"spike"*.

By the energy conservation property of MP [Mallat and Zhang, 1993], the energy of the original signal is preserved between the approximation and the residual, i.e.,

$$\|\mathbf{x}\|^2 = \left\|\mathfrak{Re}\left(\boldsymbol{\Phi}_{a,J}\boldsymbol{\alpha}_J\right)\right\|^2 + \left\|\mathbf{e}_{(K)}\right\|^2 \tag{C.7}$$

$$= \sum_{k=1}^{K} \left\|\alpha_{j(k)}\right\|^2 + \left\|\mathbf{e}_{(K)}\right\|^2 \tag{C.8}$$

$$\Rightarrow \left\|\alpha_{j(k)}\right\|^2 = \left\|\mathbf{e}_{(k-1)}\right\|^2 - \left\|\mathbf{e}_{(k)}\right\|^2 \tag{C.9}$$

where (C.8) follows from the fact that kernels $\boldsymbol{\phi}$ have unit norm. The energy of spikes is thus equal to the *decrease in residual error energy per iteration* [Goodwin and Vetterli, 1999]. It can be shown that the kernel selection criterion in (C.3) achieves the highest possible decrease in error energy per iteration [Goodwin and Vetterli, 1999, Sec. 3].

The objective of minimizing residual error energy with a dictionary of auditory kernels is only a coarse model of human perception. Modifications to MP that use a psychoacoustic model for the objective function have been proposed [e.g., Pichevar et al., 2011], but our aim is precisely to avoid the numerous experimental parameters associated with such models. Moreover, the successful prediction of cochlear filter shapes in [Smith and Lewicki, 2006] with standard MP suggests that the error energy minimization objective may be sufficient.

# Bibliography

3GPP TS 26.090 (1999). *AMR speech codec; Transcoding functions.* Third Generation Partnership Project. URL: http://www.3gpp.org/DynaReport/26071.htm.

3GPP2 C.S0014-E (2011). *Enhanced Variable Rate Codec, Speech Service Options 3, 68, 70, 73 and 77 for Wideband Spread Spectrum Digital Systems.* Third Generation Partnership Project 2. URL: http://www.3gpp2.org/Public_html/specs/C.S0014-E_v1.0_EVRC_20111231.pdf.

Abdi, H. (2010). "Holm's Sequential Bonferroni Procedure". In: *Encycl. Res. Des.* Ed. by N. Salkind. Thousand Oaks, CA, USA: SAGE Publications, Inc., pp. 1–8. URL: http://dx.doi.org/10.4135/9781412961288.n178.

Adami, A., L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas (2002). "Qualcomm-ICSI-OGI Features for ASR". In: *Proc. ICSLP*, pp. 21–24. URL: http://www.isca-speech.org/archive/icslp_2002/i02_0021.html.

Alayrac, M., C. Marquis-Favre, S. Viollon, J. Morel, and G. Le Nost (2010). "Annoyance from industrial noise: indicators for a wide variety of industrial sources." In: *J. Acoust. Soc. Am.* 128 (3), pp. 1128–39. URL: http://dx.doi.org/10.1121/1.3466855.

Allen, J. B. (2005). *Articulation and Intelligibility.* Morgan & Claypool, pp. 1–124. URL: http://dx.doi.org/10.2200/S00004ED1V01Y200508SAP001.

ANSI S3.2 (2009). *Method for Measuring the Intelligibility of Speech over Communication Systems.* American National Standards Institute. URL: http://webstore.ansi.org/RecordDetail.aspx?sku=ANSI/ASA+S3.2-2009+(R2014).

ANSI S3.5 (1997). *American National Standard Methods for Calculation of the Speech Intelligibility Index.* American National Standards Institute. URL: http://webstore.ansi.org/RecordDetail.aspx?sku=ANSI/ASA+S3.5-1997+(R2012).

Aradilla, G., J. Vepa, and H. Bourlard (2007). “An acoustic model based on Kullback-Leibler divergence for posterior features”. In: *Proc. ICASSP*. Vol. 4. Honolulu, HI, USA, pp. 657–660. URL: http://dx.doi.org/10.1109/ICASSP.2007.366998.

Atkinson, D. J. and A. A. Catellier (2008). *Intelligibility of Selected Radio Systems in the Presence of Fireground Noise: Test Plan and Results*. Tech. rep. 08-453. National Telecommunications and Information Administration. URL: http://www.its.bldrdoc.gov/publications/2490.aspx.

– (2013). *Intelligibility of Selected Radio Systems in the Presence of Fireground Noise: Test Plan and Results*. Tech. rep. 13-495. National Telecommunications and Information Administration. URL: http://www.its.bldrdoc.gov/publications/2720.aspx.

Atkinson, D. J., S. D. Voran, and A. A. Catellier (2012). *Intelligibility of the Adaptive Multi-Rate Speech Coder in Emergency-Response Environments*. Tech. rep. 13-493. National Telecommunications and Information Administration.
URL: http://www.its.bldrdoc.gov/publications/2693.aspx.

Axelsson, Ö., M. E. Nilsson, and B. Berglund (2010). “A principal components model of soundscape perception”. In: *J. Acoust. Soc. Am.* 128 (5), pp. 2836–46.
URL: http://dx.doi.org/10.1121/1.3493436.

Barlow, H. B. (1972). “Single units and sensation: A neuron doctrine for perceptual psychology?” In: *Perception* 38 (4), pp. 371–394. URL: http://dx.doi.org/10.1068/p010371.

Beerends, J. G., R. A. van Buuren, J. van Vugt, and J. A. Verhave (2009). “Objective Speech Intelligibility Measurement on the Basis of Natural Speech in Combination with Perceptual Modeling”. In: *J. Audio Eng. Soc.* 57 (5), pp. 299–308.
URL: http://www.aes.org/e-lib/browse.cfm?elib=14818.

Beerends, J. G., A. P. Hekstra, A. W. Rix, and M. P. Hollier (2002). “Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II — Psychoacoustic model”. In: *J. Audio Eng. Soc.* 50 (10), pp. 765–778.
URL: http://www.aes.org/e-lib/browse.cfm?elib=11062.

Beerends, J. G., C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl (2013a). “Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I — Temporal Alignment”. In: *J. Audio Eng. Soc.* 61 (6), pp. 366–384.
URL: http://www.aes.org/e-lib/browse.cfm?elib=16829.

– (2013b). “Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II — Perceptual Model”. In: *J. Audio Eng. Soc.* 61 (6), pp. 385–402.
URL: http://www.aes.org/e-lib/browse.cfm?elib=16830.

Benoît, C., M. Grice, and V. Hazan (1996). "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences". In: *Speech Commun.* 18 (4), pp. 381–392. URL: http://dx.doi.org/10.1016/0167-6393(96)00026-X.

Berger, J. and R. Ullmann (2013). *Contribution 24 — ITU-T Rec. P.863 as Predictor for P.835 G-MOS in Super-Wideband and Narrowband Experiments.* Study Group 12, International Telecommunication Union, Geneva, Switzerland.
URL: http://idiap.ch/~rullmann/ITU-T_SG12_Q9_Contribution24.pdf.

Bernardis, G. and H. Bourlard (1998). "Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems". In: *Proc. ICSLP*. Sydney, Australia, pp. 775–778. URL: http://isca-speech.org/archive/icslp_1998/i98_0318.html.

van Bezooijen, R. and L. C. W. Pols (1990). "Evaluating text-to-speech systems: Some methodological aspects". In: *Speech Commun.* 9 (4), pp. 263–270.
URL: http://dx.doi.org/10.1016/0167-6393(90)90002-Q.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Black, A. W. and K. Tokuda (2005). "The Blizzard Challenge — 2005: Evaluating corpus-based speech synthesis on common datasets". In: *Proc. Interspeech.* Lisbon, Portugal, pp. 77–80.
URL: http://isca-speech.org/archive/interspeech_2005/i05_0077.html.

de Boer, E. and H. R. de Jongh (1978). "On cochlear encoding: Potentialities and limitations of the reverse-correlation technique". In: *J. Acoust. Soc. Am.* 63 (1), pp. 115–35.
URL: http://dx.doi.org/10.1121/1.381704.

Bourlard, H. and N. Morgan (1994). *Connectionist Speech Recognition — A Hybrid Approach.* Boston, MA: Springer US. URL: http://dx.doi.org/10.1007/978-1-4615-3210-1.

Brons, I., R. Houben, and W. A. Dreschler (2012). "Perceptual effects of noise reduction by time-frequency masking of noisy speech". In: *J. Acoust. Soc. Am.* 132 (4), pp. 2690–2699.
URL: http://dx.doi.org/10.1121/1.4747006.

Brungart, D. S., P. S. Chang, B. D. Simpson, and D. Wang (2006). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation". In: *J. Acoust. Soc. Am.* 120 (6), pp. 4007–4018. URL: http://dx.doi.org/10.1121/1.2363929.

Byrne, D. et al. (1994). "An international comparison of long-term average speech spectra". In: *J. Acoust. Soc. Am.* 96 (4), pp. 2108–2120. URL: http://dx.doi.org/10.1121/1.410152.

Canavan, A., D. Graff, and G. Zipperlen (1997). *CALLHOME American English Speech LDC97S42.* Philadelphia, PA, USA: Linguistic Data Consortium. URL: https://catalog.ldc.upenn.edu/LDC97S42.

Carnegie Mellon Speech Group (2014). *Carnegie Mellon Pronouncing Dictionary Version 0.7b.* Pittsburgh, PA, USA: Carnegie Mellon University. URL: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Cerňak, M., B. Potard, and P. N. Garner (2015). "Phonological Vocoding Using Artificial Neural Networks". In: *Proc. ICASSP.* Brisbane, Australia. URL: http://dx.doi.org/10.1109/ICASSP.2015.7178891.

Cerňak, M. and M. Rusko (2005). "An Evaluation of Synthetic Speech Using the PESQ Measure". In: *Proc. Eur. Congr. Acoust.* Budapest, Hungary, pp. 2725–2728. URL: http://webistem.com/acoustics2008/acoustics2008/cd1/data/fa2005-budapest/paper/334-0.pdf

Christiansen, C., M. S. Pedersen, and T. Dau (2010). "Prediction of speech intelligibility based on an auditory preprocessing model". In: *Speech Commun.* 52 (7-8), pp. 678–692. URL: http://dx.doi.org/10.1016/j.specom.2010.03.004.

Clark, R. A. J., M. Podsiadło, M. Fraser, C. Mayo, and S. King (2007). "Statistical analysis of the Blizzard Challenge 2007 listening test results". In: *Proc. Blizzard Chall. Work.* Bonn, Germany, pp. 1–6. URL: http://www.isca-speech.org/archive_open/ssw6/blizzard_2007/blz3_003.html.

de Coensel, B., D. Botteldooren, and T. de Muer (2003). "1/f Noise in Rural and Urban Soundscapes". In: *Acta Acust. united with Acust.* 89 (2), pp. 287–295. URL: http://www.ingentaconnect.com/content/dav/aaua/2003/00000089/00000002/art00012

Cooke, M. (2006). "A glimpsing model of speech perception in noise". In: *J. Acoust. Soc. Am.* 119 (3), pp. 1562–1573. URL: http://dx.doi.org/10.1121/1.2166600.

Cooke, M. and O. Scharenborg (2008). "The Interspeech 2008 Consonant Challenge". In: *Proc. Interspeech.* Brisbane, Australia, pp. 1765–1768. URL: http://isca-speech.org/archive/interspeech_2008/i08_1765.html.

Côté, N. and J. Berger (2014). "Speech Communication". In: *Quality of Experience – Advanced Concepts, Applications and Methods.* Ed. by S. Möller and A. Raake. Springer International Publishing. Chap. 12, pp. 165–177. URL: http://dx.doi.org/10.1007/978-3-319-02681-7_12.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory.* John Wiley & Sons, Inc.

Davis, S. B. and P. Mermelstein (1980). “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”. In: *IEEE Trans. Acoust.* 28 (4), pp. 357–366. URL: http://dx.doi.org/10.1109/TASSP.1980.1163420.

Deutsche Telekom (2011). *Telekom Kunden telefonieren in HD Qualität (press release).* URL: http://www.telekom.com/medien/produkte-fuer-privatkunden/30588.

Dreschler, W. A., H. Verschuure, C. Ludvigsen, and S. Westermann (2001). “ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment”. In: *Audiology* 40 (3), pp. 148–157. URL: http://dx.doi.org/10.3109/00206090109073110.

Drullman, R., J. M. Festen, and R. Plomp (1994). “Effect of temporal envelope smearing on speech reception”. In: *J. Acoust. Soc. Am.* 95 (2), pp. 1053–1064. URL: http://dx.doi.org/10.1121/1.408467.

Dubno, J. R., A. R. Horwitz, and J. B. Ahlstrom (2005). “Recognition of filtered words in noise at higher-than-normal levels: decreases in scores with and without increases in masking”. In: *J. Acoust. Soc. Am.* 118 (2), pp. 923–933. URL: http://dx.doi.org/10.1121/1.1953107.

Efron, B. and R. Tibshirani (1986). “Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy”. In: *Stat. Sci.* 1 (1), pp. 54–75. URL: http://dx.doi.org/10.1214/ss/1177013817.

Elhilali, M., T. Chi, and S. A. Shamma (2003). “A spectro-temporal modulation index (STMI) for assessment of speech intelligibility”. In: *Speech Commun.* 41 (2-3), pp. 331–348. URL: http://dx.doi.org/10.1016/S0167-6393(02)00134-6.

ETSI EG 202 396-1 (2011). *Background noise database — Binaural Signals.* URL: http://docbox.etsi.org/STQ/Open/EG 202 396-1 Background noise database/Binaural_Signals/

ETSI EN 300 969 (2000). *Half rate speech; Half rate speech transcoding.* European Telecommunications Standards Institute. URL: http://webapp.etsi.org/workprogram/Report_WorkItem.asp?WKI_ID=10142.

ETSI TS 103 106 (2014). *Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals — objective test methods.* European Telecommunications Standards Institute. URL: http://webapp.etsi.org/workprogram/Report_WorkItem.asp?WKI_ID=43343.

ETSI TS 126 090 (2012). *Adaptive Multi-Rate (AMR) speech codec; Transcoding functions.* European Telecommunications Standards Institute.
URL: http://webapp.etsi.org/workprogram/Report_WorkItem.asp?WKI_ID=40556.

Falk, T. H. and S. Möller (2008). "Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems". In: *IEEE Signal Process. Lett.* 15, pp. 781–784.
URL: http://dx.doi.org/10.1109/LSP.2008.2006709.

Falk, T. H., V. Parsa, J. F. Santos, K. H. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie (2015). "Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools". In: *IEEE Signal Process. Mag.* 32 (2), pp. 114–124. URL: http://dx.doi.org/10.1109/MSP.2014.2358871.

Fastl, H. and E. Zwicker (2007). *Psychoacoustics. Facts and Models.* Springer Berlin Heidelberg.
URL: http://dx.doi.org/10.1007/978-3-540-68888-4.

Fernández Gallardo, L. and S. Möller (2015). "Phoneme Intelligibility in Narrowband and in Wideband Channels". In: *Proc. DAGA.* Nuremberg, Germany.

Fitt, S. (2000). *Documentation and user guide to UNISYN lexicon and post-lexical rules.* Tech. rep. CSTR, University of Edinburgh. URL: http://www.cstr.ed.ac.uk/projects/unisyn/.

Fredrickson, B. L. and D. Kahneman (1993). "Duration Neglect in Retrospective Evaluations of Affective Episodes". In: *J. Pers. Soc. Psychol.* 65 (1), pp. 45–55.
URL: http://dx.doi.org/10.1037/0022-3514.65.1.45.

French, N. R. and J. C. Steinberg (1947). "Factors Governing the Intelligibility of Speech Sounds". In: *J. Acoust. Soc. Am.* 19 (1), pp. 90–119.
URL: http://dx.doi.org/10.1121/1.1916407.

Furui, S. (1986). "Speaker independent isolated word recognition using dynamic features of speech spectrum". In: *IEEE Trans. Acoust. Speech Signal Process.* 34 (1), pp. 52–59.
URL: http://dx.doi.org/10.1109/TASSP.1986.1164788.

Garnier, M., N. Henrich, and D. Dubois (2010). "Influence of Sound Immersion and Communicative Interaction on the Lombard Effect". In: *J. Speech, Lang. Hear. Res.* 53 (3), pp. 588–608.
URL: http://dx.doi.org/10.1044/1092-4388(2009/08-0138).

Gautier-Turbin, V. and N. Le Faucheur (2005). "A perceptual objective measure for noise reduction systems". In: *Proc. Meas. Speech Qual. Net.* Prague, Czech Republic, pp. 81–84.
URL: http://wireless.feld.cvut.cz/mesaqin2005/papers/7_MESAQIN2005_POMNR_VGT-NLF_France-Telecom_RD.pdf

GENESIS S.A. (2012). *Loudness Toolbox.*
URL: http://genesis-acoustics.com/en/loudness_online-32.html.

Glasberg, B. R. and B. C. J. Moore (1990). "Derivation of auditory filter shapes from notched-noise data". In: *Hear. Res.* 47 (1-2), pp. 103–138.
URL: http://dx.doi.org/10.1016/0378-5955(90)90170-T.

Godfrey, J. and E. Holliman (1993). *Switchboard-1 Release 2 LDC97S62.* Philadelphia, PA, USA: Linguistic Data Consortium. URL: https://catalog.ldc.upenn.edu/LDC97S62.

Gold, B., N. Morgan, and D. P. W. Ellis (2011). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music.* Wiley-Interscience.
URL: http://dx.doi.org/10.1002/9781118142882.

Goodwin, M. M. and M. Vetterli (1999). "Matching Pursuit and Atomic Signal Models Based on Recursive Filter Banks". In: *IEEE Trans. Signal Process.* 47 (7), pp. 1890–1902.
URL: http://dx.doi.org/10.1109/78.771038.

Graff, D., K. Walker, and D. Miller (2001). *Switchboard Cellular Part 1 Transcribed Audio LDC2001S15.* Philadelphia, PA, USA: Linguistic Data Consortium.
URL: https://catalog.ldc.upenn.edu/LDC2001S15.

Gray, Jr., A. H. and J. D. Markel (1976). "Distance measures for speech processing". In: *IEEE Trans. Acoust. Speech Signal Process.* 24 (5), pp. 380–391.
URL: http://dx.doi.org/10.1109/TASSP.1976.1162849.

Greenberg, S., T. Arai, and R. Silipo (1998). "Speech intelligibility derived from exceedingly sparse spectral information". In: *Proc. ICSLP.* Sydney, Australia.
URL: http://isca-speech.org/archive/icslp_1998/i98_0074.html.

Gribonval, R. (1999). "Approximations non-linéaires pour l'analyse de signaux sonores". French. Ph.D. thesis. Université Paris Dauphine - Paris IX.
URL: https://tel.archives-ouvertes.fr/tel-00583662/.

GSM Association (2013). *Minimum Technical Requirements for use of the HD Voice Logo with GSM/UMTS issued by GSMA (Annex C),* pp. 1–16.
URL: http://www.gsma.com/network2020/hd-voice-minimum-requirements/.

Hermansky, H. (1990). "Perceptual linear predictive (PLP) analysis of speech". In: *J. Acoust. Soc. Am.* 87 (4), pp. 1738–1752. URL: http://dx.doi.org/10.1121/1.399423.

Hermansky, H. and P. Fousek (2005). "Multi-resolution RASTA filtering for TANDEM-based ASR". In: *Proc. Interspeech.* Lisbon, Portugal, pp. 361–364.
URL: http://www.isca-speech.org/archive/interspeech_2005/i05_0361.html.

Hermansky, H., N. Morgan, A. Bayya, and P. Kohn (1991). "Compensation for the Effect of the Communication Channel in Auditory-Like Analysis of Speech (RASTA-PLP)". In: *Proc. Eurospeech.* Genoa, Italy, pp. 1367–1370.
URL: http://www.isca-speech.org/archive/eurospeech_1991/e91_1367.html.

– (1992). "RASTA-PLP speech analysis technique". In: *Proc. ICASSP.* Vol. 1. San Francisco, CA, USA, pp. 121–124. URL: http://dx.doi.org/10.1109/ICASSP.1992.225957.

van Heuven, V. J. and R. van Bezooijen (1995). "Quality evaluation of synthesized speech". In: *Speech Coding Synth.* Ed. by W. B. Kleijn and K. K. Paliwal. Elsevier Science. Chap. 21, pp. 707–738. URL: http://hdl.handle.net/1887/1065.

Hinterleitner, F., S. Zabel, S. Möller, L. Leutelt, and C. R. Norrenbrock (2011). "Predicting the Quality of Synthesized Speech Using Reference-Based Prediction Measures". In: *Proc. Elektron. Sprachsignalverarbeitung.* Aachen, Germany, pp. 99–106.

Hinterleitner, F., S. Zander, K.-P. Engelbrecht, and S. Möller (2015). "On the Use of Automatic Speech Recognizers for the Quality and Intelligibility Prediction of Synthetic Speech". In: *Proc. Elektron. Sprachsignalverarbeitung.* Eichstätt, Germany, pp. 105–111.

Hinton, G., L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury (2012). "Deep Neural Networks for Acoustic Modeling in Speech Recognition — The shared views of four research groups". In: *IEEE Signal Process. Mag.* 29 (6), pp. 82–97. URL: http://dx.doi.org/10.1109/MSP.2012.2205597.

Holdsworth, J., I. Nimmo-Smith, R. D. Patterson, and P. Rice (1988). "Implementing a Gamma-Tone Filter Bank". In: *Annex C SVOS Final Rep. (Part A Audit. Filter Bank)*, pp. 1–5. URL: http://www.pdn.cam.ac.uk/groups/cnbh/research/publications/pdfs/SVOS Annex C 1988.pdf

Holm, S. (1979). "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scand. J. Stat.* 6 (2), pp. 65–70. URL: http://www.jstor.org/stable/4615733.

House, A. S., C. E. Williams, M. H. L. Hecker, and K. D. Kryter (1965). "Articulation-Testing Methods: Consonantal Differentiation With a Closed-Response Set." In: *J. Acoust. Soc. Am.* 37 (144), pp. 158–166. URL: http://dx.doi.org/10.1121/1.1909295.

Hu, Y. and P. C. Loizou (2007a). "A comparative intelligibility study of single-microphone noise reduction algorithms." In: *J. Acoust. Soc. Am.* 122 (3), pp. 1777–1786.
URL: http://dx.doi.org/10.1121/1.2766778.

– (2007b). “Subjective comparison and evaluation of speech enhancement algorithms”. In: *Speech Commun.* 49 (7), pp. 588–601. URL: http://dx.doi.org/10.1016/j.specom.2006.12.006.

Huber, R. and B. Kollmeier (2006). “PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception”. In: *IEEE Trans. Audio, Speech Lang. Process.* 14 (6), pp. 1902–1911. URL: http://dx.doi.org/10.1109/TASL.2006.883259.

IEC 61672-1 (2013). *Electroacoustics — Sound level meters — Part 1: Specifications.* International Electrotechnical Commission, Geneva, Switzerland.

Imseng, D. (2013). “Multilingual speech recognition — A posterior based approach”. Ph.D. thesis. École Polytechnique Fédérale de Lausanne (EPFL).
URL: http://infoscience.epfl.ch/record/192460.

Imseng, D., P. Motlicek, P. N. Garner, and H. Bourlard (2013). “Impact of deep MLP architecture on different acoustic modeling techniques for under-resourced speech recognition”. In: *Proc. ASRU.* Olomouc, Czech Republic, pp. 332–337.
URL: http://dx.doi.org/10.1109/ASRU.2013.6707752.

ITU (2016). *Key 2005–2015 ICT data.*
URL: http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx.

ITU-R Rec. BS.468 (1986). *Measurement of audio-frequency noise voltage level in sound broadcasting.* International Telecommunication Union, Geneva, Switzerland.
URL: http://www.itu.int/rec/R-REC-BS.468/en.

ITU-T Rec. G.107 (2015). *The E-model: a computational model for use in transmission planning.* International Telecommunication Union, Geneva, Switzerland.
URL: http://handle.itu.int/11.1002/1000/12505.

ITU-T Rec. G.722.2 (2003). *Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband.* International Telecommunication Union, Geneva, Switzerland.
URL: http://handle.itu.int/11.1002/1000/6506.

ITU-T Rec. P.1401 (2012). *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models.* International Telecommunication Union, Geneva, Switzerland. URL: http://handle.itu.int/11.1002/1000/11688.

ITU-T Rec. P.48 (1988). *Specification for an Intermediate Reference System.* International Telecommunication Union, Geneva, Switzerland.
URL: http://handle.itu.int/11.1002/1000/1735.

ITU-T Rec. P.501 (2012). *Test signals for use in telephonometry.* International Telecommunication Union, Geneva, Switzerland. URL: http://handle.itu.int/11.1002/1000/11459.

ITU-T Rec. P.800 (1996). *Methods for subjective determination of transmission quality.* International Telecommunication Union, Geneva, Switzerland.
URL: http://handle.itu.int/11.1002/1000/3638.

ITU-T Rec. P.807 (2016). *Subjective Test Methodology for Assessing Speech Intelligibility.* International Telecommunication Union, Geneva, Switzerland.

ITU-T Rec. P.835 (2003). *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm.* International Telecommunication Union, Geneva, Switzerland. URL: http://handle.itu.int/11.1002/1000/7041.

ITU-T Rec. P.85 (1994). *A method for subjective performance assessment of the quality of speech voice output devices.* International Telecommunication Union, Geneva, Switzerland.
URL: http://handle.itu.int/11.1002/1000/1774.

ITU-T Rec. P.862 (2001). *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.* International Telecommunication Union, Geneva, Switzerland.
URL: http://handle.itu.int/11.1002/1000/5374.

ITU-T Rec. P.863 (2011). *Perceptual objective listening quality assessment (POLQA).* International Telecommunication Union, Geneva, Switzerland.
URL: http://handle.itu.int/11.1002/1000/11009.

ITU-T Study Group 12 (2011a). *Work item "Perceptual approaches for multi-dimensional analysis" (P.AMD).* International Telecommunication Union, Geneva, Switzerland: Part of the Work Programme for Study Period 2013–2016.
URL: http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=8891.

– (2011b). *Work item "Technical cause analysis" (P.TCA).* International Telecommunication Union, Geneva, Switzerland: Part of the Work Programme for Study Period 2013–2016.
URL: http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=8892.

– (2013). *Work item "Perceptual Objective Noise Reduction Assessment" (P.ONRA).* International Telecommunication Union, Geneva, Switzerland: Part of the Work Programme for Study Period 2013–2016. URL: http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=8942.

Johnson, D. and contributors (2011). *QuickNet.*
URL: http://www1.icsi.berkeley.edu/Speech/qn.html.

Jørgensen, S., J. Cubick, and T. Dau (2015). “Speech Intelligibility Evaluation for Mobile Phones”. In: *Acta Acust. united with Acust.* 101 (5), pp. 1016–1025. URL: http://dx.doi.org/10.3813/AAA.918896.

Junqua, J.-C. (1993). “The Lombard reflex and its role on human listeners and automatic speech recognizers”. In: *J. Acoust. Soc. Am.* 93 (1), pp. 510–524. URL: http://dx.doi.org/10.1121/1.405631.

– (1996). “The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex”. In: *Speech Commun.* 20 (1-2), pp. 13–22. URL: http://dx.doi.org/10.1016/S0167-6393(96)00041-6.

Kanedera, N., H. Hermansky, and T. Arai (1998). “On properties of modulation spectrum for robust automatic speech recognition”. In: *Proc. ICASSP.* Vol. 2. Seattle, WA, USA, pp. 613–616. URL: http://dx.doi.org/10.1109/ICASSP.1998.675339.

Karaiskos, V., S. King, R. A. J. Clark, and C. Mayo (2008). “The Blizzard Challenge 2008”. In: *Proc. Blizzard Chall. Work.* Brisbane, Australia. URL: http://festvox.org/blizzard/blizzard2008.html.

Kim, D.-S. and A. Tarraf (2007). “ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality”. In: *Bell Labs Tech. J.* 12 (1), pp. 221–236. URL: http://dx.doi.org/10.1002/bltj.20228.

King, S. and V. Karaiskos (2010). “The Blizzard Challenge 2010”. In: *Proc. Blizzard Chall. Work.* Kansai Science City, Japan. URL: http://festvox.org/blizzard/blizzard2010.html.

– (2011). “The Blizzard Challenge 2011”. In: *Proc. Blizzard Chall. Work.* Turin, Italy. URL: http://festvox.org/blizzard/blizzard2011.html.

Kollmeier, B., T. Brand, and B. T. Meyer (2008). “Perception of Speech and Sound”. In: *Springer Handb. Speech Process.* Ed. by J. Benesty, M. M. Sondhi, and Y. Huang. Springer Berlin Heidelberg. Chap. 4, pp. 61–82. URL: http://dx.doi.org/10.1007/978-3-540-49127-9_4.

Krstulovic, S. and R. Gribonval (2006). “MPTK: Matching Pursuit Made Tractable”. In: *Proc. ICASSP.* Vol. 3. Toulouse, France, pp. 496–499. URL: http://dx.doi.org/10.1109/ICASSP.2006.1660699.

Kryter, K. D. (1962). “Methods for the Calculation and Use of the Articulation Index”. In: *J. Acoust. Soc. Am.* 34 (11), pp. 1689–1697. URL: http://dx.doi.org/10.1121/1.1909094.

Kuhl, P. K., K. A. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom (1992). “Linguistic experience alters phonetic perception in infants by 6 months of age.” In: *Science* 255 (5044), pp. 606–608. URL: http://dx.doi.org/10.1126/science.1736364.

Laughlin, S. B. and T. J. Sejnowski (2003). “Communication in Neuronal Networks”. In: *Science* 301 (5641), pp. 1870–1874. URL: http://dx.doi.org/10.1126/science.1089662.

Lee, K.-S. and R. V. Cox (2001). “A Very Low Bit Rate Speech Coder Based on a Recognition/Synthesis Paradigm”. In: *IEEE Trans. Speech Audio Process.* 9 (5), pp. 482–491. URL: http://dx.doi.org/10.1109/89.928913.

Lewcio, B. and S. Möller (2014). “Perception of Quality Changes in Wireless Networks”. In: *Quality of Experience – Advanced Concepts, Applications and Methods.* Ed. by S. Möller and A. Raake. Springer International Publishing. Chap. 27, pp. 395–410. URL: http://dx.doi.org/10.1007/978-3-319-02681-7_27.

Lewicki, M. S. (2002). “Efficient Coding of Time-Varying Signals Using a Spiking Population Code”. In: *Probabilistic Model. Brain.* Ed. by R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki. Cambridge, MA: MIT Press. Chap. II-12, pp. 243–255.

Lewicki, M. S. and T. J. Sejnowski (1999). “Coding time-varying signals using sparse, shift-invariant representations”. In: *Adv. NIPS 11.* Ed. by M. J. Kearns, S. A. Solla, and D. A. Cohn. Cambridge, MA: MIT Press, pp. 730–736. URL: https://papers.nips.cc/paper/1514-coding-time-varying-signals-using-sparse-shift-invariant-representations.pdf

Lippmann, R. P. (1996). “Accurate consonant perception without mid-frequency speech energy”. In: *IEEE Trans. Speech Audio Process.* 4 (1), p. 66. URL: http://dx.doi.org/10.1109/TSA.1996.481454.

Lu, Y. and M. Cooke (2008). “Speech production modifications produced by competing talkers, babble, and stationary noise”. In: *J. Acoust. Soc. Am.* 124 (5), pp. 3261–3275. URL: http://dx.doi.org/10.1121/1.2990705.

Maier, A., M. Schuster, A. Batliner, E. Nöth, and E. Nkenke (2007). “Automatic Scoring of the Intelligibility in Patients with Cancer of the Oral Cavity”. In: *Proc. Interspeech.* Antwerp, Belgium, pp. 1206–1209. URL: http://www.isca-speech.org/archive/interspeech_2007/i07_1206.html.

Malfait, L., J. Berger, and M. Kastner (2006). “P.563—The ITU-T Standard for Single-Ended Speech Quality Assessment”. In: *IEEE Trans. Audio, Speech Lang. Process.* 14 (6), pp. 1924–1934. URL: http://dx.doi.org/10.1109/TASL.2006.883177.

Malfait, L., J. Berger, and R. Ullmann (2009). *Contribution 96 — Analysis of ten P.OLQA full-scale super-wideband experiments.* Study Group 12, International Telecommunication Union, Geneva, Switzerland. URL: http://www.itu.int/md/T09-SG12-C-0096/en.

Mallat, S. G. and Z. Zhang (1993). "Matching pursuits with time-frequency dictionaries". In: *IEEE Trans. Signal Process.* 41 (12), pp. 3397–3415. URL: http://dx.doi.org/10.1109/78.258082.

Marquis-Favre, C., E. Premat, and D. Aubrée (2005). "Noise and its Effects — A Review on Qualitative Aspects of Sound. Part II: Noise and Annoyance". In: *Acta Acust. united with Acust.* 91 (4), pp. 626–642. URL: http://www.ingentaconnect.com/content/dav/aaua/2005/00000091/00000004/art00002

Mashimo, M., T. Toda, K. Shikano, and N. Campbell (2001). "Evaluation of Cross-Language Voice Conversion". In: *Proc. Eurospeech.* Aalborg, Denmark, pp. 361–364. URL: http://isca-speech.org/archive/eurospeech_2001/e01_0361.html.

Meyer, B. T. and B. Kollmeier (2010). "Learning from human errors: Prediction of phoneme confusions based on modified ASR training". In: *Proc. Interspeech.* Makuhari, Chiba, Japan, pp. 1209–1212. URL: http://isca-speech.org/archive/interspeech_2010/i10_1209.html.

Middag, C., G. van Nuffelen, J.-P. Martens, and M. de Bodt (2008). "Objective intelligibility assessment of pathological speakers". In: *Proc. Interspeech.* Brisbane, Australia, pp. 1745–1748. URL: http://isca-speech.org/archive/interspeech_2008/i08_1745.html.

Möller, S. (2000). *Assessment and Prediction of Speech Quality in Telecommunications.* Boston, MA: Springer US. URL: http://dx.doi.org/10.1007/978-1-4757-3117-0.

Möller, S., W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann (2011). "Speech Quality Estimation: Models and Trends". In: *IEEE Signal Process. Mag.* 28 (6), pp. 18–28. URL: http://dx.doi.org/10.1109/MSP.2011.942469.

Möller, S. and R. Heusdens (2013). "Objective Estimation of Speech Quality for Communication Systems". In: *Proc. IEEE* 101 (9), pp. 1955–1967. URL: http://dx.doi.org/10.1109/JPROC.2013.2241374.

Möller, S., D.-S. Kim, and L. Malfait (2008). "Estimating the Quality of Synthesized and Natural Speech Transmitted Through Telephone Networks Using Single-ended Prediction Models". In: *Acta Acust. united with Acust.* 94 (1), pp. 21–31. URL: http://dx.doi.org/10.3813/AAA.918004.

Moore, B. C. J. (2003). "Coding of Sounds in the Auditory System and Its Relevance to Signal Processing and Coding in Cochlear Implants". In: *Otol. Neurotol.* 24 (2), pp. 243–254. URL: http://dx.doi.org/10.1097/00129492-200303000-00019.

Morgan, N. and H. Bourlard (1989). “Generalization and Parameter Estimation in Feedforward Nets: Some Experiments”. In: *Adv. NIPS 2.* Ed. by D. S. Touretzky, pp. 630–637. URL: http://papers.nips.cc/paper/275-generalization-and-parameter-estimation-in-feedforward-nets-some-experiments

Narwaria, M., W. Lin, I. V. McLoughlin, S. Emmanuel, and L.-T. Chia (2012). “Nonintrusive Quality Assessment of Noise Suppressed Speech With Mel-Filtered Energies and Support Vector Regression”. In: *IEEE Trans. Audio. Speech. Lang. Processing* 20 (4), pp. 1217–1232. URL: http://dx.doi.org/10.1109/TASL.2011.2174223.

Norrenbrock, C. R., F. Hinterleitner, U. Heute, and S. Möller (2015). “Quality prediction of synthesized speech based on perceptual quality dimensions”. In: *Speech Commun.* 66, pp. 17–35. URL: http://dx.doi.org/10.1016/j.specom.2014.06.003.

Olshausen, B. A. and D. J. Field (1997). “Sparse coding with an overcomplete basis set: A strategy employed by V1?” In: *Vision Res.* 37 (23), pp. 3311–3325. URL: http://dx.doi.org/10.1016/S0042-6989(97)00169-7.

Orange (2013). *Orange launches international high-definition voice service for operators and service providers (press release).* Paris, France. URL: http://www.orange.com/en/Press-and-medias/press-releases-2016/press-releases-2013/Orange-launches-international-high-definition-voice-service-for-operators-and-service-providers

Patterson, R. D., K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand (1992). “Complex sounds and auditory images”. In: *Audit. Physiol. Perception, Proc. 9th Int. Symp. Hear.* Ed. by Y. Cazals, L. Demany, and K. Horner. Oxford: Pergamon, pp. 429–446. URL: http://w3.pdn.cam.ac.uk/groups/cnbh/research/publications/pdfs/Petal92ish.pdf

Pichevar, R., H. Najaf-Zadeh, L. Thibault, and H. Lahdili (2011). “Auditory-inspired sparse representation of audio signals”. In: *Speech Commun.* 53 (5), pp. 643–657. URL: http://dx.doi.org/10.1016/j.specom.2010.09.008.

Pickett, J. M. (1956). “Effects of Vocal Force on the Intelligibility of Speech Sounds”. In: *J. Acoust. Soc. Am.* 28 (5), pp. 902–905. URL: http://dx.doi.org/10.1121/1.1908510.

Pinto, J., G. S. V. S. Sivaram, H. Hermansky, and M. Magimai.-Doss (2009). “Volterra series for analyzing MLP based phoneme posterior estimator”. In: *Proc. ICASSP.* Taipei, Taiwan, pp. 1813–1816. URL: http://dx.doi.org/10.1109/ICASSP.2009.4959958.

Počta, P. and J. G. Beerends (2015). “Subjective and objective measurement of synthesized speech intelligibility in modern telephone conditions”. In: *Speech Commun.* 71, pp. 1–9. URL: http://dx.doi.org/10.1016/j.specom.2015.04.001.

Pollack, I. and J. M. Pickett (1958). “Masking of Speech by Noise at High Sound Levels”. In: *J. Acoust. Soc. Am.* 30 (2), pp. 127–130. URL: http://dx.doi.org/10.1121/1.1909503.

PSCR (2013). *Modified Rhyme Test Audio Library.*
URL: http://www.pscr.gov/projects/audio_quality/mrt_library/overview/.

Raake, A. (2006). “Short- and long-term packet loss behavior: Towards speech quality prediction for arbitrary loss distributions”. In: *IEEE Trans. Audio, Speech Lang. Process.* 14 (6), pp. 1957–1968. URL: http://dx.doi.org/10.1109/TASL.2006.883231.

Rabiner, L. R. (1989). “A tutorial on hidden Markov models and selected applications in speech recognition”. In: *Proc. IEEE* 77 (2), pp. 257–286. URL: http://dx.doi.org/10.1109/5.18626.

Rabiner, L. R. and R. W. Schafer (2007). “Introduction to digital speech processing”. In: *Found. Trends Signal Process.* 1 (1), pp. 1–194. URL: http://dx.doi.org/10.1561/2000000001.

Rapporteur for Question 9/12 (2009). *Temporary Document 12-WP2: Statistical Evaluation Procedure for P.OLQA v.1.0.* International Telecommunication Union, Geneva, Switzerland. URL: http://www.itu.int/md/T09-SG12-090310-TD-WP2-0012/en.

Rasipuram, R. (2014). “Grapheme-based Automatic Speech Recognition using Probabilistic Lexical Modeling”. Ph.D. thesis. École Polytechnique Fédérale de Lausanne (EPFL).
URL: http://dx.doi.org/10.5075/epfl-thesis-6280.

Reimes, J., H. W. Gierlich, F. Kettler, S. Poschen, and M. Lepage (2011). “The Relative Approach Algorithm and its Applications in New Perceptual Models for Noisy Speech and Echo Performance”. In: *Acta Acust. united Ac.* 97 (2), pp. 325–341.
URL: http://dx.doi.org/10.3813/AAA.918412.

Remes, U., R. Karhila, and M. Kurimo (2013). “Objective evaluation measures for speaker-adaptive HMM-TTS systems”. In: *Proc. Speech Synth. Work.* Barcelona, Spain, pp. 177–181. URL: http://ssw8.talp.cat/papers/ssw8_PS2-6_Remes.pdf.

Ribeiro, F., D. Florêncio, C. Zhang, and M. L. Seltzer (2011). “crowdMOS: An Approach for Crowdsourcing Mean Opinion Score Studies”. In: *Proc. ICASSP.* Prague, Czech Republic, pp. 2416–2419. URL: http://dx.doi.org/10.1109/ICASSP.2011.5946971.

Richard, M. D. and R. P. Lippmann (1991). “Neural Network Classifiers Estimate Bayesian a posteriori Probabilities”. In: *Neural Comput.* 3 (4), pp. 461–483.
URL: http://dx.doi.org/10.1162/neco.1991.3.4.461.

Rix, A. W., J. G. Beerends, M. P. Hollier, and A. P. Hekstra (2001). “Perceptual evaluation of speech quality (PESQ) — a new method for speech quality assessment of telephone networks and codecs”. In: *Proc. ICASSP*. Vol. 2. Salt Lake City, UT, USA, pp. 749–752. URL: http://dx.doi.org/10.1109/ICASSP.2001.941023.

Rix, A. W., J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza (2006). “Objective Assessment of Speech and Audio Quality — Technology and Applications”. In: *IEEE Trans. Audio, Speech Lang. Process.* 14 (6), pp. 1890–1901. URL: http://dx.doi.org/10.1109/TASL.2006.883260.

Rix, A. W., M. P. Hollier, A. P. Hekstra, and J. G. Beerends (2002). “Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part I — Time alignment”. In: *J. Audio Eng. Soc.* 50 (10), pp. 755–764. URL: http://www.aes.org/e-lib/browse.cfm?elib=11063.

Rowe, D. G. (1997). “Techniques for Harmonic Sinusoidal Coding”. Ph.D. thesis. University of South Australia. URL: http://www.itr.unisa.edu.au/~steven/thesis/dgr.pdf.

Rowe, D. G. and contributors (2013). *Codec2*. URL: http://rowetel.com/codec2.html.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). “Learning representations by back-propagating errors”. In: *Nature* 323 (6088), pp. 533–536. URL: http://dx.doi.org/10.1038/323533a0.

Sachs, M. B. and P. J. Abbas (1974). “Rate versus level functions for auditory-nerve fibers in cats: tone-burst stimuli”. In: *J. Acoust. Soc. Am.* 56 (6), pp. 1835–1847. URL: http://dx.doi.org/10.1121/1.1903521.

Sakoe, H. and S. Chiba (1978). “Dynamic Programming Algorithm Optimization for Spoken Word Recognition”. In: *IEEE Trans. Acoust.* ASSP-26 (1), pp. 43–49. URL: http://dx.doi.org/10.1109/TASSP.1978.1163055.

Schmidt-Nielsen, A. (1992). *Intelligibility and Acceptability Testing for Speech Technology.* Tech. rep. Washington, D.C., USA: Naval Research Laboratory. URL: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA252015

Schwerin, B. and K. Paliwal (2014). “An improved speech transmission index for intelligibility prediction”. In: *Speech Commun.* 65, pp. 9–19. URL: http://dx.doi.org/10.1016/j.specom.2014.05.003.

Smith, E. C. and M. S. Lewicki (2005). “Efficient Coding of Time-Relative Structure Using Spikes”. In: *Neural Comput.* 17 (1), pp. 19–45. URL: http://dx.doi.org/10.1162/0899766052530839.

– (2006). “Efficient auditory coding”. In: *Nature* 439 (7079), pp. 978–982. URL: http://dx.doi.org/10.1038/nature04485.

Soldo, S., M. Magimai.-Doss, and H. Bourlard (2012). “Synthetic References for Template-based ASR using Posterior Features”. In: *Proc. Interspeech.* Portland, OR, USA. URL: http://www.isca-speech.org/archive/interspeech_2012/i12_2146.html.

Soldo, S., M. Magimai.-Doss, J. Pinto, and H. Bourlard (2011). “Posterior features for template-based ASR”. In: *Proc. ICASSP.* Prague, Czech Republic, pp. 4864–4867. URL: http://dx.doi.org/10.1109/ICASSP.2011.5947445.

Steeneken, H. and T. Houtgast (1980). “A physical method for measuring speech-transmission quality”. In: *J. Acoust. Soc. Am.* 67 (1), pp. 318–326. URL: http://dx.doi.org/10.1121/1.384464.

– (2002). “Validation of the revised STIr method”. In: *Speech Commun.* 38 (3-4), pp. 413–425. URL: http://dx.doi.org/10.1016/S0167-6393(02)00010-9.

Sturm, B. L., C. Roads, A. McLeran, and J. J. Shynk (2009). “Analysis, Visualization, and Transformation of Audio Signals Using Dictionary-based Methods”. In: *J. New Music Res.* 38 (4), pp. 325–341. URL: http://dx.doi.org/10.1080/09298210903171178.

Stylianou, Y. and A. K. Syrdal (2001). “Perceptual and objective detection of discontinuities in concatenative speech synthesis”. In: *Proc. ICASSP.* Salt Lake City, UT, USA, pp. 837–840. URL: http://dx.doi.org/10.1109/ICASSP.2001.941045.

van Summers, W., D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes (1988). “Effects of noise on speech production: Acoustic and perceptual analyses”. In: *J. Acoust. Soc. Am.* 84 (3), pp. 917–928. URL: http://dx.doi.org/10.1121/1.396660.

Supplee, L. M., R. P. Cohn, J. S. Collura, and A. V. McCree (1997). “MELP: the new Federal Standard at 2400 bps”. In: *Proc. ICASSP.* Munich, Germany, pp. 1591–1594. URL: http://dx.doi.org/10.1109/ICASSP.1997.596257.

Swisscom (2012). *Crystal-clear voice quality on your mobile: Swisscom launches HD Voice (press release).* Berne, Switzerland. URL: https://www.swisscom.ch/en/about/medien/press-releases/2012/01/20120131_MM_HD_Voice.html

Taal, C. H., R. C. Hendriks, R. Heusdens, and J. Jensen (2010). *Matlab implementation of the Short-Time Objective Intelligibility (STOI) measure.* URL: http://insy.ewi.tudelft.nl/content/short-time-objective-intelligibility-measure.

– (2011). “An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech”. In: *IEEE Trans. Audio. Speech. Lang. Processing* 19 (7), pp. 2125–2136. URL: http://dx.doi.org/10.1109/TASL.2011.2114881.

Tang, Y., M. Cooke, and C. Valentini-Botinhao (2016). “Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech”. In: *Comput. Speech Lang.* 35, pp. 73–92. URL: http://dx.doi.org/10.1016/j.csl.2015.06.002.

Teng, Y., R. Kubichek, R. Anderson-Sprecher, and J. E. Schroeder (2007). “Objective Speech Intelligibility Measure for Low Bit-Rate Speech Codecs Operating in Noisy Channels”. In: *IEEE Mil. Commun. Conf.* Orlando, FL, USA, pp. 1–7. URL: http://dx.doi.org/10.1109/MILCOM.2007.4455330.

TIA-102.BABA (2009). *Project 25 Vocoder Description.* Telecommunications Industry Association. URL: http://www.tiaonline.org/standards/.

Tóth, M. A., M. L. García Lecumberri, Y. Tang, and M. Cooke (2015). “A corpus of noise-induced word misperceptions for Spanish”. In: *J. Acoust. Soc. Am.* 137 (2), EL184–EL189. URL: http://dx.doi.org/10.1121/1.4905877.

Ullmann, R., J. Berger, and A. Llagostera Casanovas (2013). *Contribution 81 — Derivation of speech degradation scores (S-MOS) from subjective noise intrusiveness and overall quality scores (N- and G-MOS).* Study Group 12, International Telecommunication Union, Geneva, Switzerland. URL: http://idiap.ch/~rullmann/ITU-T_SG12_Q9_Contribution81.pdf.

Ullmann, R. and H. Bourlard (2016). “Predicting the intrusiveness of noise through sparse coding with auditory kernels”. In: *Speech Commun.* 76, pp. 186–200. URL: http://dx.doi.org/10.1016/j.specom.2015.07.005.

Ullmann, R., H. Bourlard, J. Berger, and A. Llagostera Casanovas (2013). “Noise Intrusiveness Factors in Speech Telecommunications”. In: *AIA-DAGA Jt. Conf. Acoust.* Merano, Italy, pp. 436–439. URL: http://publications.idiap.ch/index.php/publications/show/2619.

Ullmann, R., M. Magimai.-Doss, and H. Bourlard (2015). “Objective Speech Intelligibility Assessment through Comparison of Phoneme Class Conditional Probability Sequences”. In: *Proc. ICASSP.* Brisbane, Australia, pp. 4924–4928. URL: http://dx.doi.org/10.1109/ICASSP.2015.7178907.

Ullmann, R., R. Rasipuram, M. Magimai.-Doss, and H. Bourlard (2015). “Objective Intelligibility Assessment of Text-to-Speech Systems Through Utterance Verification”. In: *Proc. Interspeech.* Dresden, Germany, pp. 3501–3505. URL: http://isca-speech.org/archive/interspeech_2015/i15_3501.html.

Valentini-Botinhao, C., J. Yamagishi, and S. King (2011). “Can Objective Measures Predict the Intelligibility of Modified HMM-based Synthetic Speech in Noise?” In: *Proc. Interspeech.*

Florence, Italy, pp. 1837–1840.
URL: http://isca-speech.org/archive/interspeech_2011/i11_1837.html.

Vepa, J., S. King, and P. Taylor (2002). "New objective distance measures for spectral discontinuities in concatenative speech synthesis". In: *Proc. IEEE Work. Speech Synth.* Vol. 4. Santa Monica, CA, USA, pp. 223–226. URL: http://dx.doi.org/10.1109/WSS.2002.1224414.

Viterbi, A. J. (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE Trans. Inf. Theory* 13 (2), pp. 260–269.
URL: http://dx.doi.org/10.1109/TIT.1967.1054010.

Voiers, W. D. (1967). *Performance evaluation of speech processing devices — III. Diagnostic evaluation of speech intelligibility.* Tech. rep. Bedford, MA, USA: Air Force Cambridge Research Laboratories. URL: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=AD0650158

Voran, S. D. (2013). "Using articulation index band correlations to objectively estimate speech intelligibility consistent with the modified rhyme test". In: *IEEE Work. Appl. Signal Process. to Audio Acoust.* New Paltz, NY, USA, pp. 1–4.
URL: http://dx.doi.org/10.1109/WASPAA.2013.6701826.

Wältermann, M., I. Tucker, A. Raake, and S. Möller (2010). "Extension of the E-model towards super-wideband speech transmission". In: *Proc. ICASSP.* Dallas, TX, USA, pp. 4654–4657.
URL: http://dx.doi.org/10.1109/ICASSP.2010.5495199.

Wang, L., L. Wang, Y. Teng, Z. Geng, and F. K. Soong (2012). "Objective Intelligibility Assessment of Text-to-Speech System using Template Constrained Generalized Posterior Probability". In: *Proc. Interspeech.* Portland, OR, USA, pp. 627–630.
URL: http://www.isca-speech.org/archive/interspeech_2012/i12_0627.html.

Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland (2006). *The HTK Book (for HTK Version 3.4).* Cambridge University Engineering Department.

Zen, H., K. Tokuda, and A. W. Black (2009). "Statistical parametric speech synthesis". In: *Speech Commun.* 51 (11), pp. 1039–1064. URL: http://dx.doi.org/10.1016/j.specom.2009.04.004.

Zwicker, E. (1977). "Procedure for calculating loudness of temporally variable sounds". In: *J. Acoust. Soc. Am.* 62 (3), pp. 675–682. URL: http://dx.doi.org/10.1121/1.381580.

**Raphael ULLMANN**
Av. du Mont-d'Or 39, CH-1007 Lausanne
raphael.ullmann@gmail.com
+41 76 488 72 60

Curriculum Vitae as of May 2016
Swiss and French citizen
Born 1982–10–13 (33 years)

Ph.D. Graduate, 5 Years of Industry Experience
Speech & Audio Processing, Telecommunications, Perception

**Profile:** Signal processing engineer with a focus on industrial research. I enjoy building solutions and seeing them work in the real world. I am a self-motivated, reliable and communicative person.

## EMPLOYMENT

2012 – 2016 **Research Assistant at the Speech and Audio Processing Group** —
Idiap / École Polytechnique Fédérale de Lausanne (EPFL), Martigny VS / Lausanne VD

- Designed a psychoacoustic model to predict background noise perception in new "HD Voice" telephony services. Collaboration with the industry. C++ and MATLAB.
  **Result:** Significantly lower prediction error than loudness-based models.
- Developed an algorithm to measure speech intelligibility in emergency communication systems. Linux shell scripting and Python.
  **Result:** Solution works for both telephone and synthetic speech signal analysis.
- Designed subjective listening tests. Analyzed the statistical significance of results.
- Surveyed and monitored the state of the art. Conferred with peer researchers.

2010 – 2012 **Senior Research Engineer (Auditory Perception in Telecommunications)** —
SwissQual AG, Zuchwil SO (now a Rohde & Schwarz company)

- Coauthor of the "POLQA" algorithm for speech quality assessment (polqa.info), winner of a competition at the International Telecommunication Union (ITU). C++.
- Advised mobile phone operators on speech quality matters.
  Contributed to the activities of the Quality of Experience group at ITU.
- Integrated new algorithms within the company's software release cycle.
  Collaborated with the software engineering, hardware and verification teams.

2006 – 2010 **Applied Research Engineer** — SwissQual AG, Zuchwil SO (now Rohde & Schwarz)

- Developed algorithms for speech and audio signal analysis and classification. C++.
  Optimized execution speed on Windows/Intel and Android/ARM platforms.
- Designed and conducted subjective listening tests for various signal impairments.

## PROJECTS

- **Medical Speech Analysis** — Civil Service (2014). Extracted features from patients' recorded speech to predict intubation difficulty. Analyzed links between results and doctor's ratings. MATLAB.
- **Audio Compression** — M.Sc. Thesis (2006). Designed and implemented the psychoacoustic and quantization modules for an AAC encoder (Advanced Audio Coding, successor format to MP3) in C.

## EDUCATION

2012 – 2016 **Ph.D. in Electrical Engineering** —
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne VD
Relevant Courses: Machine Learning | Computational Perception | Seminar on Patents

2001 – 2006 **M.Sc. in Microengineering** — EPFL, Lausanne VD
Relevant Courses: Image Processing | Speech Processing. Overall GPA: 5.48 / 6.0

2003 – 2004 **Exchange Student** (Bachelor Year) — Carnegie Mellon University, Pittsburgh, PA, USA
Honors: Placed on the “Dean’s List”. GPA: 4.0 / 4.0

## ADDITIONAL EXPERIENCE

2014 – 2015 **Doctoral Student Committee Member** — EPFL. In a team of three:

- Organized scientific events (e.g., invited speakers, lab visits, industry lunch).
- Collaborated on improvements to the doctoral program with its director.
- Organized and recruited a coaching team for new students.

## COMPUTER SKILLS

| | |
|---|---|
| **Languages** | C/C++ \| MATLAB \| Python \| Bash |
| **Tools** | Adobe Audition \| MS Visual Studio \| PyCharm \| git \| LaTeX |

## LANGUAGE SKILLS

**German/Swiss-German** and **French** (mother tongues) | **English** (full professional proficiency)

## SELECTED PUBLICATIONS AND PATENT (full list available on Google Scholar)

- R. Ullmann, R. Rasipuram, M. Magimai-Doss and H. Bourlard. Objective **Intelligibility Assessment** of Text-to-Speech Systems Through Utterance Verification. **Winner of a best student paper award** at Interspeech 2015.
- R. Ullmann and H. Bourlard. **Predicting the Intrusiveness of Noise** Through Sparse Coding with Auditory Kernels. Speech Communication, 2016.
- R. Ullmann, H. Bourlard, J. Berger and A. Llagostera Casanovas. **Noise Intrusiveness Factors** in Speech Telecommunications. In Proc. AIA-DAGA International Conference on Acoustics, 2013.
- J. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy and M. Keyhl. Perceptual Objective Listening Quality Assessment (**POLQA**), **The Third Generation ITU-T Standard** for End-to-End Speech Quality Measurement. Journal of the Audio Engineering Society, 2013.
- **European Patent** 2 474 975 “Method for estimating speech quality” for SwissQual AG.

## ACTIVITIES AND INTERESTS

Photography, badminton (weekly practice), discovering other countries.