Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 76 (2016) 186–200

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Predicting the intrusiveness of noise through sparse coding with auditory kernels

Raphael Ullmann *, Hervé Bourlard

Idiap Research Institute, Martigny, Switzerland
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Received 10 January 2015; received in revised form 14 June 2015; accepted 21 July 2015
Available online 29 July 2015

Abstract

This paper presents a novel approach to predicting the intrusiveness of background noises in speech signals as it is perceived by human listeners. This problem is of particular interest in telephony, where the recently widened range of transmitted audio frequencies has increased the importance of appropriate background noise reduction strategies. Current approaches predict the average noise intrusiveness score that would be obtained in a subjective listening test by combining different signal features related to physical properties (e.g., signal energy, spectral distribution) or psychoacoustic estimations (e.g., loudness) of noise. The combination and/or implementation of such features requires expert knowledge or the availability of training data. We present a novel approach that is based on a model of efficient sound coding, using a sparse spike coding representation of noise. We show that the sparsity of these representations implicitly models several factors in the perception of noise, and yields predictions of noise intrusiveness scores that compare to or outperform traditional features, without the use of training data. Our evaluation datasets and used performance metrics are based on standardized methods for the evaluation of quality prediction models.

Keywords: Noise intrusiveness; Noise reduction; Objective quality assessment; Sparse spike coding; Perceptual models

1. Introduction

Speech telecommunication is increasingly taking place over mobile terminals, exposing the speech capture process to unpredictable background noise in the speaker's environment. Noise that is transmitted with the speech signal is perceived as a quality degradation by remote interlocutors and reduces the overall quality of experience. This issue has recently received renewed interest with the extension of the traditional telephone band to 50–7000 Hz in many cellular networks (e.g., Orange, 2013).

Noise reduction (NR) processing can be used to attenuate background noise in a speech signal and thereby improve its perceived quality. In order to evaluate the quality benefit of NR, the perceived intrusiveness of background noise needs to be assessed. Subjective listening tests, as standardized in International Telecommunication Union (ITU-T) Recommendation P.835 (2003), can be used to assess the intrusiveness of background noise in speech signals with and without NR processing. However, subjective tests are expensive and time-consuming to conduct. It is therefore desirable to have an *objective measure* (i.e., an algorithm) that can predict the average listener's quality score, called Mean Opinion Score (MOS), for a given test speech signal. Predictions of objective measures are also repeatable, making them especially useful for the comparison and optimization of NR algorithms.

* Corresponding author at: Idiap Research Institute, Martigny, Switzerland.
E-mail addresses: raphael.ullmann@idiap.ch (R. Ullmann), bourlard@idiap.ch (H. Bourlard).

http://dx.doi.org/10.1016/j.specom.2015.07.005
0167-6393/

While significant research efforts have been devoted to the objective assessment of *overall* speech quality (e.g., Rix et al., 2001; Malfait et al., 2006; Beerends et al., 2013a), this score makes no distinction between different causes of quality issues. In improving the quality of noisy speech however, it is useful to distinguish between issues that are due to the presence of noise (i.e., noise intrusiveness) and to other signal distortions (e.g., from speech coding or transmission errors). This can help avoid applying overly strong NR processing that would also degrade the signal. Consequently, the objective assessment of background noise intrusiveness is a focus of an ongoing work item at ITU (ITU-T Study Group 12, 2013).

The present paper seeks to find a signal feature that correlates well with subjective judgments of noise intrusiveness. In the subjective test method (ITU-T Rec. P.835, 2003), listeners are instructed to focus only on the background noise part to rate noise intrusiveness, and report any distortions to the speech part in a different score. This means that noise intrusiveness only pertains to the presence of noise itself, and not to its impact on the speech part (e.g., its impact on speech intelligibility).

Objective measures often use a comparison of the test signal to an undistorted reference to assess quality, an approach known as full-reference assessment and shown in Fig. 1. Such a comparative approach cannot be used for intrusiveness, since the focus is on the noise only and the reference is noise-free. Instead, existing approaches analyze the test signal (e.g., during speech pause sections) and extract multiple features in time and frequency. A challenge is thus to use only few, highly predictive features to estimate a noise intrusiveness MOS, since subjectively scored training data is expensive to collect.

This paper proposes a novel, single feature to predict perceived intrusiveness. Our approach is based on a sparse noise representation as a model of high-level sensory coding. 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 with simple noise types, and find that the number of atoms in the representation models several factors in the perception of noise. We then evaluate our approach on a dataset of noise-corrupted telephone speech recordings, and show that the proposed feature and subjective intrusiveness scores are highly correlated ($\overline{|R|} > 0.95$). The proposed feature uses no training data, and outperforms or compares to a traditional feature based on loudness.

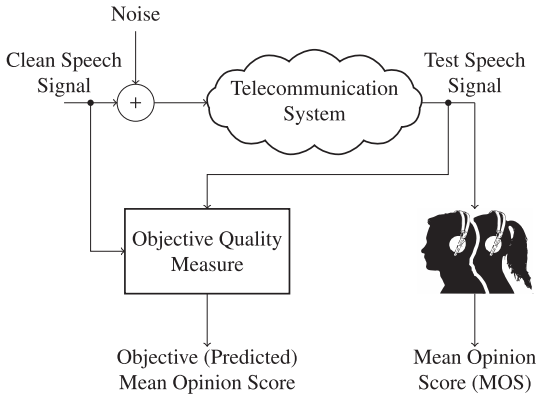

Fig. 1. Signal flow in full-reference objective quality assessment. Listeners rate the perceived quality of test speech signals, whereas the full-reference objective measure also uses the clean speech recording to help compute its quality prediction. The telecommunication system may alter the signal at its input through noise reduction, speech coding and decoding, or other modifications.

1.1. State of the art

What aspects of noise are perceptually relevant has been studied extensively in the context of environmental noise annoyance. Briefly, 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 apply detailed models of the processing at the outer, middle and inner ear to derive an intensity 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 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 follow these principles and combine multiple features (Gautier-Turbin and Le Faucheur, 2005; Reimes et al., 2011; Narwaria et al., 2012).[1] Main features in these measures model perceived intensity using the short-time noise loudness, log energy or filterbank energies, respectively.

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