IDIAP RESEARCH REPORT

SPARSE GAMMATONE SIGNAL MODEL PREDICTS PERCEIVED NOISE INTRUSIVENESS

Raphael Ullmann[a] Hervé Bourlard

Idiap-RR-07-2014

APRIL 2014

[a]Idiap Research Institute

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Sparse Gammatone Signal Model Predicts Perceived Noise Intrusiveness

Raphael Ullmann[1,2] and Hervé Bourlard[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Swiss Federal Institute of Technology, Lausanne, Switzerland
{raphael.ullmann, herve.bourlard}@{idiap.ch, epfl.ch}

Abstract

Is it possible to predict the intrusiveness of background noise in speech signals as perceived by humans? Such a question is important to the automatic evaluation of speech enhancement systems, including those designed for new wideband speech telephony, and the goal of a future ITU quality assessment standard.

In this paper, we show that this is possible by modeling the encoding of the noise signal at the auditory nerve. Indeed, recent research suggests that sparse signal representations may be indicative of the encoding process in the auditory system, making them interesting for modeling human sound perception.

Here, we further explore this hypothesis, and decompose background noise in the speech signal into a sparse combination of gammatone functions, resulting in a sparse, physiologically grounded representation of the noise. We then show that the number of gammatones required to encode the noise is directly correlated with the perception of noise intrusiveness. Furthermore, we show that an established measure of noise intrusiveness based on this new representation outperforms the same measure based on the traditional loudness model.

Index Terms: Psychoacoustics, perceptual quality assessment, noise intrusiveness, noise annoyance, noise reduction, sparsity

1. Introduction

Sparse audio signal representations have attracted recent interest in general audio coding [1–3] and auditory modeling [4, 5]. While sparsity is desirable in audio coding to achieve high compression ratios, the rationale behind sparse signal representations in auditory modeling is the *efficient coding hypothesis*, which states that the human sensory system encodes stimuli in a way that maximizes the amount of information carried to the brain but minimizes the number of neuronal impulses [6, 7].

The use of sparse signal representations for auditory modeling was further substantiated by the work of Smith and Lewicki [8], who showed that a sparse signal model trained on a natural sound ensemble successfully predicted auditory filter shapes.

In this work, we apply a sparse signal modeling approach to the assessment of perceived intrusiveness of environmental noises in speech recordings. More precisely, we focus on the case of noise reduction in telephony, where the assessment of noise intrusiveness is useful to optimize the perceptual quality of the telecommunication service to the end-user.

This paper is structured as follows: We explain the technical background of the speech quality assessment task in Section 2. Section 3 introduces the auditory model, which is based on a sparse signal decomposition using a dictionary of gammatones. We present the datasets of subjectively annotated speech recordings that we use for evaluation in Section 4, and show results in Section 5. We provide a deeper analysis of our results in Section 6 and conclude with some closing remarks in Section 7.

2. Problem statement

Mobile telephony often takes place in noisy environments, but interlocutors generally perceive noise in the transmitted speech signal as a quality impairment. Therefore, modern telecommunication systems apply noise reduction processing in order to attenuate background noises in the speech signal.

Noise reduction usually also partially degrades the foreground speech signal, so a compromise between sufficient noise attenuation and tolerable speech degradation is necessary. The strength of these two effects

can be determined through subjective listening tests as defined by the International Telecommunication Union (ITU) in Recommendation P.835 [9], in which listeners rate the *background noise intrusiveness*, *speech degradation* and *overall quality* of noise-corrupted speech recordings processed by noise reduction on a five-point scale.

An ongoing effort at ITU [10] seeks to predict these quality scores algorithmically from the signal, enabling the evaluation of noise reduction systems without time-consuming subjective tests. In this paper, we focus on predicting the average listener score for noise intrusiveness. In the target application of the ITU effort, the distorted, noise-corrupted speech recording to be evaluated (the *test signal*) is always provided with a matching undistorted, noise-free speech recording (the *reference signal*). Here, we only use the reference signal to help identify speech pause sections in the test signal. Indeed, our model predicts the intrusiveness of noise from the (noisy) speech pause sections.

3. Auditory model

Our model is based on a decomposition of the background noise signal following the *spike coding* approach first proposed by Lewicki and Sejnowski [11]. Under this approach, the discrete time-domain signal $x(n)$ is represented as a linear combination of predefined kernel functions $\phi_1(n), \cdots, \phi_M(n)$:

$$x(n) = \sum_{m=1}^{M} \sum_{i=1}^{I_m} \alpha_{m,i} \phi_m(n - r_{m,i}) + \epsilon(n) \tag{1}$$

where $r_{m,i}$ and $\alpha_{m,i}$ denote the (arbitrary) temporal offset and gain of the instance $i \in [1, I_m]$ of kernel ϕ_m, respectively, and $\epsilon(n)$ simply denotes the residual error of the representation. The set of predefined kernel functions $\phi_1(n), \cdots, \phi_M(n)$ is often called a *dictionary*. The number of occurrences I_m of a predefined kernel ϕ_m in the representation is unconstrained.

Compared to models based on the short-time Fourier transform (STFT), this approach does not subdivide the signal into a series of frames, which may obscure transients by spreading their energy over the duration of a frame. Instead, the representation preserves the precise temporal position of individual signal components, a necessary property for essential auditory tasks such as sound source localization.

When the M kernels in the dictionary are modeled after auditory filter shapes, each kernel occurrence in the obtained representation can be thought of as a local population of auditory nerve spikes with average firing rates encoded by the kernel gain α [12]. In an over-simplification of terminology, we will refer to each kernel occurrence in the representation as a *spike*.

Equation (1) admits an infinite number of solutions. *Sparse* solutions to (1) are characterized by a small total number of spikes $K = \sum_m I_m$. Matching Pursuit [13] is an iterative algorithm to derive a (suboptimal) sparse solution to (1), adding one spike to the representation at each iteration. Smith and Lewicki [8] have shown that this modeling approach indeed predicts auditory filter shapes as the most efficient dictionary for encoding an ensemble of natural sounds, lending further credence to sparse spike coding as an auditory model.

3.1. Implementation

In our experiments, we used a dictionary of *gammatone* functions, i.e., mathematical models of auditory filters defined as

$$\phi_m(t) = t^3 e^{-2\pi b_m t} \cos(2\pi f_m t + \varphi), \quad t \geq 0 \tag{2}$$

with b_m a bandwidth parameter and f_m the center frequency of kernel ϕ_m. A good fit to experimental human perceptual data is obtained for $b_m = 1.019\ \mathrm{ERB}(f_m)$ [14], with

$$\mathrm{ERB}(f) = 0.108\ f + 24.7 \tag{3}$$

the equivalent rectangular bandwidth at frequency f in Hz [15].

Figure 1 shows the frequency responses of the $M = 32$ gammatones in our dictionary. We use Slaney's auditory toolbox [16] to generate gammatones with center frequencies between 100 and 7195 Hz, spaced regularly on the ERB scale. We rescale the norm of gammatones to unity, thus the α parameter in (1) describes the energy of each spike in the representation. Sparser solutions could be obtained with a larger dictionary, however previous work has shown the benefit of increasing the dictionary size from 32 to 64 gammatones to be small, and even negligible beyond 64 [4].

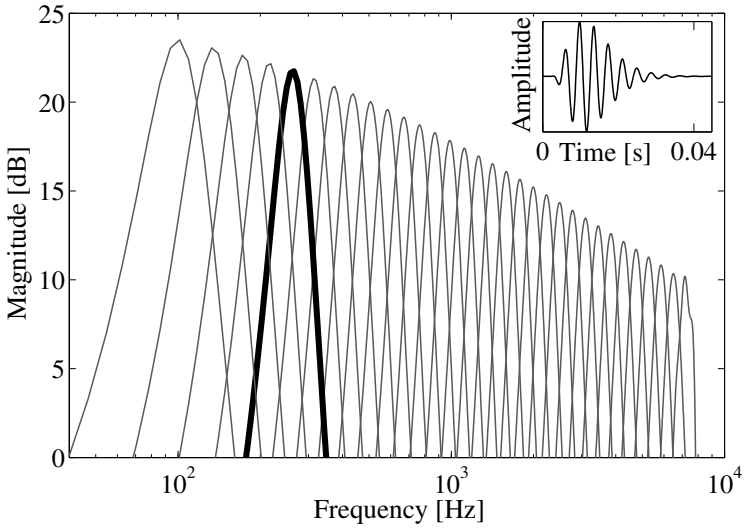

Figure 1: *Frequency responses of gammatones in our dictionary. The time-domain waveform corresponding to the highlighted gammatone is shown in the inset.*

We use the Matching Pursuit Tool Kit (MPTK) [17] to derive sparse solutions to (1). In computing the decomposition, MPTK will also adapt the gammatone phase φ for each spike via the use of the Hilbert transform.[1] The result of Matching Pursuit decomposition is a set of values

$$\Theta_K = \{\alpha_{m,i},\ r_{m,i},\ \varphi_{m,i}\},\quad m = [1, M],\ i = [1, I_m] \tag{4}$$

describing the gain, temporal offset and phase of each of the $K = \sum_m I_m$ gammatone spikes representing the signal $x(n)$ in the spike coding model, along with a residual error $\epsilon(n)$.

The number of spikes K in the obtained representation depends on how many Matching Pursuit iterations are carried out. We will discuss the choice of stopping criterion in Section 5.

4. Datasets

We evaluate the spike coding model on 2 datasets previously described in [18], which we briefly summarize below:

We recorded short (2–4 seconds) sentences from 4 speakers, to which we added noises from a collection of environmental noise recordings [19], increasing the duration to 6 seconds with temporal structure *leading noise — speech+noise — trailing noise*. The background noise types and the signal-to-noise ratios (SNR) at which they were added are given in Table 1.

We then processed the 6-seconds recordings in accordance with our focus on telephony. This consisted of either applying noise reduction processing [20] and standard speech codecs [21–23] (*simulated conditions*), or of transmitting the recordings with commercial handsets over land-line and cellular networks (in Europe and the USA), and recording the transmitted signals at the far end (*live conditions*). Live conditions represent the full signal processing chain in telephony, including distortions from transmission errors and in-handset noise reduction.

In total, we generated 65 different conditions (repeated across the 4 speakers), which we partitioned into 2 datasets, such that a single dataset could be scored in about 1 hour. Two thirds of generated conditions have audio bandwidths beyond the traditional telephone band, representing wideband or super-wideband (up to 50–14 000 Hz) speech technologies. Furthermore, $\sim$30% of conditions are noise-free for a balanced subjective test design as per ITU requirements [24].

A panel of 27 listeners evaluated the perceptual quality of the test signals in both datasets as defined in ITU Recommendation P.835 [9]. The mean listener score for noise intrusiveness, averaged over the 4 speakers in each condition, is the Mean Opinion Score (MOS) that our model should predict.

[1] Creating a gammatone dictionary in MPTK format is non-trivial. We intend to release MATLAB code that shows how to do this.

Noise type	SNR [dB]
Jackhammer (distant)	3
Pub (unintelligible babble)	4
Road (nearby traffic)	5
Train station (train engine, announcement)	8
Car (inside, cruising at 80 km/h)	8
Train (inside)	12
Office (keyboard clicks, intermittent speech)	15
Cafeteria (cutlery falling on trays, laughter)	17
Crossroad (distant traffic)	10, 20, 30, 40

Table 1: *Used background noise types and SNR. The final SNR after noise reduction processing may be lower.*

4.1. Perception of noise intrusiveness

A statistical analysis of subjective scores (see [18]) revealed that neither sentence content nor speaking style (regular vs. effortful speech) had a significant effect on noise intrusiveness scores. This is likely due to the used ITU test protocol, which instructs listeners to focus only on the signal component of interest (i.e., the background noise when scoring noise intrusiveness) [9]. We also found that noise intrusiveness was more determined by the *absolute playback level* of background noise, rather than the signal-to-noise ratio between speech and noise.

This independence of noise intrusiveness from the speech signal could be a particularity of our datasets, but it simplifies the following experiments, since we will estimate noise intrusiveness solely from the background noise in the test signal.

5. Results

5.1. Applying the proposed model

We apply our model to all background noise conditions, defined as conditions with a SNR $\leq$ 40 dB (22 and 23 conditions in dataset 1 and 2, respectively). To speed up further calculations, and because noise energies above 7 000 Hz tend to be very low, we band-limit all signals to 100–7 000 Hz and operate at 16 kHz.

We use voice activity detection on the (noise-free) reference signals and a temporal alignment (similar to [25]) to easily identify (noisy) speech pause sections in the test signals. For each condition, we find the test signal with the shortest active speech duration, and compute sparse spike coding representations of the concatenated noise sections, as described in Section 3.1.

An important aspect in computing the spike coding representations is the choice of stopping criterion for the Matching Pursuit decomposition. Applications in general audio coding often set a criterion on the energy of the residual signal $\epsilon(n)$, i.e., the decomposition stops when a desired *coding fidelity* is achieved. For perceptual modeling, it makes more sense to set a threshold on the stimulus intensity of auditory nerves, modeled here by the energy of spikes α.

We thus halt the decomposition once the energy of newly extracted spikes falls below a fixed *spiking threshold* α_{min}. We set α_{min} such that the test signal with the lowest noise level in our datasets (a 40 dB SNR condition) produces 0 spikes (i.e., $\epsilon(n) = x(n)$), and an infinitesimally higher noise level will have at least 1 gammatone spike in its representation.

Analyzing the spike coding representations obtained with this stopping criterion, we find that the average *number of spikes* needed to encode 1 second of background noise signal (the "spike rate") strongly correlates with subjective noise intrusiveness scores. Figure 2 shows this relationship. For our 2 datasets, the (absolute) correlation is $\rho = 93.7\%$ and 92.1%, respectively.

Note that the 2 datasets mainly differ in the noise types they contain. The overall test design and processing steps are very similar, and we only partitioned our data to avoid overly long subjective tests.

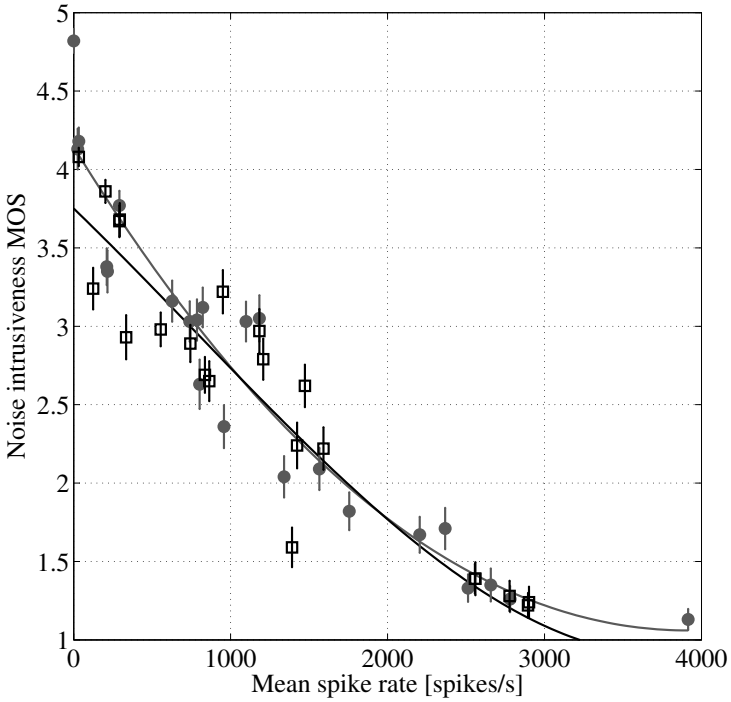

Figure 2: *Number of gammatone spikes per second needed to encode background noise vs. perceived noise intrusiveness, rated on a scale from 1 ("very intrusive") to 5 ("not noticeable"). Conditions from dataset 1 and 2 are shown as black unfilled squares and gray circles, respectively. Error bars represent 95% confidence intervals of subjective scores.*

5.2. Comparison to other measures

We compare our new measure against these baselines:

Log energy, i.e. noise level in dB SPL (sound pressure level).

Weighted log energy, applying a spectral weighting curve to model the ear's frequency-dependent sensitivity. We use the equal loudness level contour from the ISO 226:2003 standard [26]. We flip the contour upside down with a 0 dB gain at 1 kHz to obtain the spectral weighting curve.

Mean loudness, which also considers simultaneous and temporal masking effects in the perception of sound intensity. In [27], the loudness of stationary environmental noises was found to be highly correlated with perceived noise annoyance. We use Zwicker's calculation method [28], and compute the mean over the noise duration.

Percentile loudness, defined as the loudness value that is exceeded during a percentage p of the measurement time. Percentile loudness has become a standard measure of non-stationary noise annoyance, with values for p between 2 and 10% [29–31]. We use $p = 5\%$, as recommended in [29, sec. 16.1.3].

Besides the correlation coefficient, we compute a more sophisticated, recently standardized performance metric called *epsilon-insensitive root mean square error* (unitless, denoted rmse*) [32]. This metric computes a monotonous third-order polynomial to map model results to subjective scores, and evaluates the overall prediction error outside of the confidence intervals of subjective scores. The black and gray curves in Figure 2 show the mappings for the mean spike rate measure.

Table 2 shows the performance metrics for the baseline measures of noise intrusiveness for our 2 datasets. The two measures based on log energy capture the rough trend of perceived noise intrusiveness, but do not provide satisfactory predictions. Measures based on loudness, which are also at the core of models of overall speech quality [e.g. 33, 34] provide better results.

While loudness predicts perceived sound intensity, the *percentile of loudness over time* models how short-time intensities are integrated to an overall perception. A percentile value close to the maximum can be seen as a simple model of the "peak&end rule" from experimental psychology [35], by which subjects

Baseline measures	Dataset 1		Dataset 2	
	ρ [%]	rmse*	ρ [%]	rmse*
Log energy	85.1	0.420	89.4	0.379
Weighted log energy	91.5	0.301	91.7	0.328
Mean loudness	91.1	0.282	95.0	0.216
5%-Perc. loudness	93.5	0.238	95.0	0.235

Table 2: *Prediction performance of baseline measures, evaluated by the correlation with subjective scores (ρ) and prediction error after polynomial mapping (rmse*, lower values are better).*

are disproportionately influenced by peak events (e.g., transients or bursts in non-stationary noises) in their overall judgment.

As a comparison to the loudness measure, we also apply the percentile calculation to our proposed measure of the number of gammatone spikes over time. The resulting prediction performance is shown in the first section of Table 3 (boldface values), and is a clear improvement over the baseline measures, both in terms of correlation and prediction error.

Spike rate measures	Dataset 1		Dataset 2	
	ρ [%]	rmse*	ρ [%]	rmse*
Gammatone dict.				
Mean	93.7	0.237	92.1	0.223
5%-Percentile	**98.3**	**0.062**	**96.1**	**0.166**
ERB-Gabor dict.				
Mean	93.7	0.237	92.0	0.223
5%-Percentile	98.2	0.065	95.4	0.190
Regular Gabor dict.				
Mean	93.2	0.254	92.4	0.229
5%-Percentile	97.5	0.131	94.2	0.251

Table 3: *Prediction performance of spike rate-based measures for different dictionaries. Metrics are the same as in Table 2.*

6. Discussion

All measures compared here depend on the noise energy in different ways. In our proposed measure, the decomposition of a high-energy noise signal produces more spikes with energies above the spiking threshold α_{min} and thus a higher spike rate.

Since all compared measures depend on the noise energy, their different prediction performances are due to how they rank different *noise types* with similar energies relative to one another.

Manual inspection of results by noise type reveals that the worst mis-predictions of the log energy and weighted log energy measures are due to strongly overestimated intrusiveness of babble (speech-like) noises. The loudness measures achieve better predictions, probably by modeling how part of the overall noise floor is masked by voiced segments in the speech babble, resulting in a loudness that is lower than the sum of individual signal components. Finally, in the spike rate measures, it is known that gammatone kernels sparsely encode speech [2, 8], thus the obtained spike rate remains comparatively low.

Besides noise energy, the sparsity of obtained representations also depends on the used dictionary. Components in the noise signal that do not correlate well with any single gammatone kernel from the dictionary will be represented by combinations of multiple gammatones as part of the decomposition [13].

We create two modified dictionaries to evaluate this dependency:

- A dictionary of 32 Gabor kernels with center frequencies and half-magnitude bandwidths identical to the gammatone dictionary ("ERB-Gabor" dictionary)
- 32 Gabor kernels with constant bandwidth that tile the frequency space linearly ("regular Gabor" dictionary).

We adapted the spiking threshold α_{min} for each dictionary as per our criterion in Section 5.1. The performance metrics obtained with the two modified dictionaries are shown in the second and third section of Table 3, respectively.

Interestingly, the prediction performance of the mean spike rate remains almost unchanged across dictionaries, but drops for the percentile spike rate. From the different spike coding representations (not shown here), we see that using Gabor dictionaries results in higher spike rates for almost all noise types, but barely changes the relative ranking of noises by spike rate.

However, the *difference in the spike rate over time* between stationary and non-stationary noise sections is less pronounced in the representations computed with the Gabor dictionaries. Hence, the percentile calculation is less effective at modeling the increased intrusiveness of peak noise events. The performance drop is not as strong with the ERB-Gabor dictionary, since its kernels are very similar to those in the gammatone dictionary.

7. Conclusion

The goal of this work was not to build a complete model of noise intrusiveness, which usually involves learning model parameters from a training dataset (as e.g. in [36, 37]). Instead, we sought to identify the *fundamental measures* that best predict the perception of noise intrusiveness. We have evaluated measures that are widely used in perceptual quality and environmental noise assessment (weighted log energy, mean and percentile loudness), and have proposed a new, physiologically grounded measure that is based on a sparse spike coding representation of noise.

The sparsity of these representations directly correlates with subjective noise intrusiveness scores. Applying the percentile calculation from percentile loudness to the sparsity over time further improved predictions, making our proposed measure the best-performing on our datasets.

No parameters were trained on the data: the dictionary is a standard gammatone filterbank, and the spiking threshold is fixed by the pre-defined SNR (40 dB in our case) at which we consider speech to be "noisy". The results illustrate how models from computational biology may find technical applications.

Our analysis is limited by the amount of available subjectively scored data, and it remains to be seen whether the results hold for other types of noises. As part of the ongoing ITU effort [10], other parties will generate datasets with other conditions, on which our new measure can be further evaluated.

8. Acknowledgments

We wish to thank the Laboratory of Electromagnetics and Acoustics (LEMA) at EPFL for their support in recording speakers and in conducting the subjective listening tests.

We also thank SwissQual AG for their cooperation in generating the test data, in particular the speech recordings transmitted over live telecommunication networks.

This work was funded by the Swiss Commission for Technology and Innovation (CTI) under grant 14255.1 PFES-ES.

9. References

[1] E. Ravelli, G. Richard, and L. Daudet, "Union of MDCT Bases for Audio Coding," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1361–1372, 2008.

[2] S. Strahl and A. Mertins, "Sparse gammatone signal model optimized for English speech does not match the human auditory filters," *Brain Res.*, vol. 1220, pp. 224–33, 2008.

[3] R. Pichevar, H. Najaf-Zadeh, L. Thibault, and H. Lahdili, "Auditory-inspired sparse representation of audio signals," *Speech Commun.*, vol. 53, no. 5, pp. 643–657, 2011.

[4] E. C. Smith and M. S. Lewicki, "Efficient Coding of Time-Relative Structure Using Spikes," *Neural Comput.*, vol. 17, no. 1, pp. 19–45, 2005.

[5] Y. Karklin, C. Ekanadham, and E. P. Simoncelli, "Hierarchical spike coding of sound," in *Adv. NIPS 25*, 2012, pp. 3041–3049.

[6] H. B. Barlow, "Single units and sensation: A neuron doctrine for perceptual psychology?" *Perception*, vol. 1, no. 4, pp. 371–394, 1972.

[7] S. B. Laughlin and T. J. Sejnowski, "Communication in Neuronal Networks," *Science (80-.).*, vol. 301, no. 5641, pp. 1870–1874, 2003.

[8] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

[9] ITU-T Rec. P.835, *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*. International Telecommunication Union, Geneva, Switzerland, 2003.

[10] ITU-T Study Group 12, *Perceptual Objective Noise Reduction Assessment (P.ONRA)*. Geneva, Switzerland: Part of the Work Programme for Study Period 2013–2016, see http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=8942.

[11] M. S. Lewicki and T. J. Sejnowski, "Coding time-varying signals using sparse, shift-invariant representations," in *Adv. NIPS 11*, M. J. Kearns, S. A. Solla, and D. A. Cohn, Eds. Cambridge, MA: MIT Press, 1999, pp. 730–736.

[12] M. S. Lewicki, "Efficient Coding of Time-Varying Signals Using a Spiking Population Code," in *Probabilistic Model. Brain*, R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki, Eds. Cambridge, MA: MIT Press, 2002, pp. 243–255.

[13] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[14] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Audit. Physiol. Percept.*, Y. Cazals, L. Demany, and K. Horner, Eds. Oxford: Pergamon, 1992, pp. 429–446.

[15] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1-2, pp. 103–138, 1990.

[16] M. Slaney, "Auditory Toolbox — Version 2," Interval Research Corporation, Palo Alto, CA, Tech. Rep., 1998.

[17] S. Krstulovic and R. Gribonval, "MPTK: Matching Pursuit Made Tractable," in *Proc. ICASSP*, vol. 3, 2006, pp. 496–499.

[18] R. Ullmann, H. Bourlard, J. Berger, and A. Llagostera Casanovas, "Noise Intrusiveness Factors in Speech Telecommunications," in *AIA-DAGA Jt. Conf. Acoust.*, 2013, pp. 436–439.

[19] ETSI EG 202 396-1, "Background noise database — Binaural Signals," 2011.

[20] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI Features for ASR," in *Proc. ICSLP*, 2002, pp. 21–24.

[21] ETSI TS 126 090, *Adaptive Multi-Rate (AMR) speech codec; Transcoding functions*. European Telecommunications Standards Institute, 2012.

[22] 3GPP2 C.S0014-E, *Enhanced Variable Rate Codec, Speech Service Options 3, 68, 70, 73 and 77 for Wideband Spread Spectrum Digital Systems*. Third Generation Partnership Project 2, 2011.

[23] ITU-T Rec. G.722.2, *Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband*. International Telecommunication Union, Geneva, Switzerland, 2003.

[24] ITU-T Study Group 12, "Agreed Sections Requirement Specification for P.ONRA," in *Quest. 9 Interim Meet. Rep.*, (internal ITU-T meeting document, not publicly available), 2013.

[25] J. G. 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 Part I — Temporal Alignment," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 366–384, 2013.

[26] ISO 226, *Acoustics — Normal equal-loudness-level contours*. International Organization for Standardization, 2003.

[27] M. Alayrac, "Indicateurs de gêne sonore pour l'étude d'impact du bruit d'un site industriel : caractérisation physique et perceptive," Ph.D. dissertation, INSA Lyon, 2009.

[28] E. Zwicker, "Procedure for calculating loudness of temporally variable sounds," *J. Acoust. Soc. Am.*, vol. 62, no. 3, pp. 675–682, 1977.

[29] H. Fastl and E. Zwicker, *Psychoacoustics. Facts and Models.* Springer Berlin Heidelberg, 2007.

[30] B. Kollmeier, *Vorlesung über physikalische, technische und medizinische Akustik [lecture notes]*, Oldenburg, Germany, 2009.

[31] O. Axelsson, M. E. Nilsson, and B. Berglund, "A principal components model of soundscape perception," *J. Acoust. Soc. Am.*, vol. 128, no. 5, pp. 2836–46, 2010.

[32] ITU-T Rec. P.1401, *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*. International Telecommunication Union, Geneva, Switzerland, 2012.

[33] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II — Psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 765–778, 2002.

[34] J. G. 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 Part II — Perceptual Model," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 385–402, 2013.

[35] B. L. Fredrickson and D. Kahneman, "Duration Neglect in Retrospective Evaluations of Affective Episodes," *J. Pers. Soc. Psychol.*, vol. 65, no. 1, pp. 45–55, 1993.

[36] V. Gautier-Turbin and N. Le Faucheur, "A perceptual objective measure for noise reduction systems," in *Proc. Meas. Speech Qual. Net.*, 2005, pp. 81–84.

[37] J. Reimes, H. W. Gierlich, F. Kettler, S. Poschen, and M. Lepage, "The Relative Approach Algorithm and its Applications in New Perceptual Models for Noisy Speech and Echo Performance," *Acta Acust. united Ac.*, vol. 97, no. 2, pp. 325–341, 2011.