Spectral and Cepstral Audio Noise Reduction Techniques in Speech Emotion Recognition

Jouni Pohjalainen, Fabien Ringeval, Zixing Zhang, Björn Schuller

Spectral and Cepstral Audio Noise Reduction Techniques in Speech Emotion Recognition

Jouni Pohjalainen[1], Fabien Ringeval[1,2], Zixing Zhang[1], Björn Schuller[1,3]
[1]Chair of Complex and Intelligent Systems, University of Passau, Germany
[2]Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, France
[3]Department of Computing, Imperial College London, UK
{firstname.lastname}@uni-passau.de

ABSTRACT

Signal noise reduction can improve the performance of machine learning systems dealing with time signals such as audio. Real-life applicability of these recognition technologies requires the system to uphold its performance level in variable, challenging conditions such as noisy environments. In this contribution, we investigate audio signal denoising methods in cepstral and log-spectral domains and compare them with common implementations of standard techniques. The different approaches are first compared generally using averaged acoustic distance metrics. They are then applied to automatic recognition of spontaneous and natural emotions under simulated smartphone-recorded noisy conditions. Emotion recognition is implemented as support vector regression for continuous-valued prediction of arousal and valence on a realistic multimodal database. In the experiments, the proposed methods are found to generally outperform standard noise reduction algorithms.

Keywords

noise reduction; denoising; speech emotion recognition

1. INTRODUCTION

Audio signals are commonly affected by factors other than the signal of interest. Additive noise arises due to other sound sources competing with the target signal and due to the combined effects of the recording equipment and the transmission channel. Convolutive variation appears due to the channel and acoustic reverberation. The performance of machine learning systems is generally negatively affected by these effects, because they may mask target signals, such as speech, and cause mismatch in feature statistics between the training and recognition conditions.

Additive noise is a ubiquitous problem that affects the performance of machine-learning speech systems in many common environments. In the field of automatic speech recognition (ASR), this is addressed on different stages of the system, including feature extraction, feature enhancement, machine learning models and their training [13, 29] and also by using better hardware or multiple microphones.

Noise reduction for speech audio is often evaluated from the viewpoint of listeners in terms of enhancing either the subjective quality or speech intelligibility [16]. Of these, intelligibility, whose improvement is the more difficult problem [16], is potentially relevant for ASR. Computational paralinguistic analysis of speech, whose general aim is to uncover various external attributes of speech not related to its linguistic message (the concern of ASR), has emerged as an active field of research within the past decade [28]. Two of its most central topics are speaker recognition (identification and verification) and, more recently, automatic recognition of emotion in speech. These applications, which are important for multimedia content analysis, retrieval and processing, do not have an obvious connection with the traditional objectives of speech enhancement. Nevertheless, in speaker recognition, traditional single-channel noise reduction systems are frequently found helpful as preprocessing in improving system performance under noisy or mismatched conditions [9, 19, 25, 26]. General-purpose signal-based denoising methods, such as power spectral subtraction and minimum-mean-square-error estimation of log-spectral amplitude, have been found to improve speaker recognition performance quite reliably across different back-end machine learning systems, test conditions (type and level of degradation) and training conditions (noise-matched or mismatched), even more so than auditory-based methods designed to improve subjective quality [9]. In general, the relative performance of different denoising methods is quite strongly affected by the aforementioned factors [9, 25].

In this paper, following on the results of the studies focusing on signal denoising in text-independent speaker recognition [9, 25], a machine learning task that is commonly based on identifying the distribution of the short-time magnitude spectrum, we compare denoising techniques from the viewpoint of audio and multimedia content analysis in speech emotion recognition. It is another central paralinguistic problem with other applications ranging from ASR systems and call centers to intelligent user interfaces [28]. Work on noise robustness in speech emotion recognition has also been started [27, 32], but to our knowledge, signal denoising has not been explicitly studied for this purpose. As emotional or stressed speech often coincides with acoustically noisy environments, it is worthwhile to investigate the performance of general-purpose noise reduction solutions in improving short-time spectral representation of the two pri-

mary emotion dimensions – activation (arousal), i. e., the intensity of emotional expression, and valence, i. e., the positive/negative affective dimension [10].

We propose a family of general-purpose, adaptive signal denoising techniques that can be simply configured with respect to spectral smoothing, temporal smoothing and adaptation rate of the noise model in order to tackle different nonstationary noise conditions and application requirements. They operate in a log-spectral or cepstral domain, which differ in terms of spectral smoothing facilitated by the cepstral representation. These methods are compared against published implementations of baseline methods suggested by the previous studies, firstly by using general, objective signal enhancement quality measures. The most promising variants are then compared as preprocessing for an emotion recognition system performing continuous-valued prediction of arousal and valence. In this study, we focus on relatively similar noise conditions in the training and testing phase, so that denoising is used to recover important information in both training and testing.

2. NOISE REDUCTION

2.1 Previous Work

Classical single-channel noise reduction methods include variants of spectral subtraction [2, 3], where an averaged noise (magnitude or power) spectrum is subtracted from the noisy signal spectrum while keeping the resultant spectral magnitudes positive, and Wiener filtering [16], often implemented in practice using an iterative approach [12].

The performance of the minimum mean square error (MMSE) [7] and log-spectral amplitude MMSE (Log-MMSE) estimators [8] still remains among the best of the published methods [20]. In part, this can be attributed to their decision-directed estimation approach, which bases the spectral estimate of each frame partially on the estimates from previous frames via the *a priori* SNR estimate updated by using a memory coefficient [5]. Spectral subtraction [21] and the decision-directed MMSE [20] methods have also been applied in the spectral modulation domain in order to better handle nonstationary noise.

In recent comprehensive evaluations looking at noise reduction preprocessing techniques to combat mismatch in speaker recognition, MMSE, Log-MMSE and spectral subtraction have performed well [9, 25]. Power spectral subtraction also resulted in large performance gains in [26]. Therefore, as the baseline noise reduction methods in this study, we choose two published implementations of spectral subtraction, one working in the power spectral domain (SS-P [16]) and one in the magnitude spectral domain (SS-M [4]), and two published implementations of Log-MMSE (MMSE-1 [16]; MMSE-2 [4]).

2.2 Methods Under Study

The methods proposed in this study are based on obtaining the noise model by averaging over low-energy frames in the representation domain – log-spectral or cepstral – and, optionally, by adapting the noise model constantly based on the (Gaussian) similarity of the short-term moving average of the representation to the initial noise model. A general overview is presented in Fig. 1.

The audio signal is processed in Hann-windowed frames of 30 ms extracted every 10 ms. The discrete Fourier transform

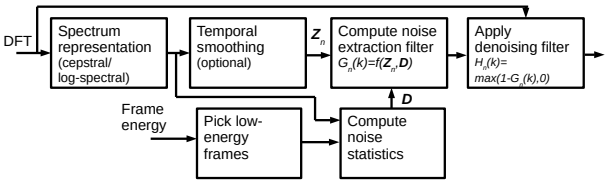

Figure 1: *The noise reduction framework.*

(DFT) of each audio frame is first transformed into either a logarithmic magnitude spectrum or a truncated cepstrum [6, 22]. The real cepstrum is a frequency transform (specifically, the inverse Fourier transform) of the logarithmic spectrum. A truncated cepstrum, where only the lower-order coefficients are retained, thus represents a smoothed version of the original logarithmic spectrum [22]. With the sampling rate of 44100 Hz used in this study, we choose to retain the first $J = 50$ cepstral coefficients in accordance to a rule of thumb of choosing the length of the truncated cepstrum to be less than an expected pitch period [22].

In the lower branch of Fig. 1, low-energy frames within the analysis block of six seconds are automatically found by clustering the frame logarithmic energies into two clusters by using k-means and selecting the frames assigned to the cluster with the lower mean value. The *initial noise model* $D = \{\mu_j, \sigma_j^2\}$, consisting of a mean and variance for each (spectral or cepstral) coefficient, is then obtained by averaging over these frames. In the upper branch, temporal smoothing is applied to the complete sequence of the noisy vectors $\{\boldsymbol{Y}_n\}$ by using the equation

$$\boldsymbol{Z}_n = \beta \boldsymbol{Z}_{n-1} + (1-\beta)\boldsymbol{Y}_n, \tag{1}$$

where n is the frame index, $\beta = (w-1)/w$ and w is an "equivalent rectangular window length" of the moving average integrator in the sense that the contribution of each new observation is weighted by $1/w$. Setting $w = 1$ causes no temporal integration to be performed. In the next stage, the similarity of the noisy representation $\boldsymbol{Z}_n = (Z_{n,1}, \ldots, Z_{n,J})$ with the initial noise model is evaluated as

$$\gamma_{n,j} = \exp(-(Z_{n,j} - \mu_j)^2/\sigma_j^2), \qquad 1 \le j \le J, \tag{2}$$

yielding scores in the range $(0, 1)$. The components of the *current* noise model at the nth frame $\mu_{n,j}$ (initially, $\mu_{0,j} = \mu_j$) are then *adapted* towards the direction of the current smoothed noisy representation $\boldsymbol{Z}_n$, in proportion α to the similarity $\gamma_{n,j}$ of the latter to the *initial* noise model:

$$\mu_{n,j} = \mu_{n-1,j} + \alpha\gamma_{n,j}(Z_{n,j} - \mu_{n-1,j}). \tag{3}$$

In both log-spectral and cepstral domains, spectral-domain division corresponds to subtraction. Therefore, the DFT-domain *noise extraction filter* $G_n(k)$ is represented by $\mu_{n,j} - Z_{n,j}$ in the analysis domain, and can be obtained by transforming this quantity back to the DFT domain. It is then used to create the noise suppression filter, with the lower limit of suppression at zero: $H_n(k) = \max(1 - G_n(k), 0)$. The modified frames are combined using an overlap-add method to produce the enhanced signal.

One motivation for processing the signal in one of these domains is to be able to apply cepstral smoothing in generating the noise and noisy signal models by temporal averaging

and integration. The log-spectral representation is both an intermediate stage in cepstrum computation and a special case in the sense that it is equivalent to a non-truncated cepstral representation. Considering distance metrics in the cepstral and log-spectral domains, the cepstral Euclidean distance can be shown to be a lower bound for the root-mean-square (RMS) log-spectral distance; with the inclusion of more cepstral coefficients in the truncated cepstrum, more spectral fine structure information is preserved and the cepstral Euclidean distance approaches the RMS log-spectral distance from below [11]. The logarithmic spectrum representation is frequently used in audio signal analysis due to the large dynamic range of hearing in the amplitude dimension [33], and is also used in denoising by the Log-MMSE method [8]. In this paper, methods using the two representations are distinguished by referring to them as cepstral noise reduction (CNR) and log-spectral noise reduction (LNR). The variants of these main types are further distinguished from each other based on the temporal smoothing parameter w and the adaptation rate α (Eqs. 1-3), which control the focus of the noise suppression filter on different modulation frequencies of the nonstationary noise and the target signal.

3. EXPERIMENTS

3.1 Overview and Material

We evaluate the denoising methods on the RECOLA multimodal affective interaction corpus [24]. It includes 46 multimodal (audio, video and physiological data) recordings of French-speaking participants involved in a dyadic collaborative task. Affective dimensions expressed by the participants were evaluated by six annotators for the first five minutes of each recording. This was done for arousal and valence separately. Obtained labels were then resampled to a constant 40 ms frame rate and averaged over all raters by considering inter-evaluator agreement, to provide a 'gold standard' [23]. In order to ensure speaker-independence in the experiments, the corpus was split into three partitions, by balancing the gender and the age of the subjects: training (16 subjects), validation (15 subjects) and testing (15 subjects).

In adding noise to the audio material, the recordings were convolved with the impulse response of a smartphone [18]. This step was performed to simulate talking over a smartphone in different places. Different types of additive noise were then added, using the CHiME-2013 database [1] for simulating a living room environment (*CHiME*), and data collected from the freesound platform[1] to simulate public transport (*trains*) environments. We collected, in total, 230 minutes of noise, to match the duration of the RECOLA database. In order to provide realistic conditions, we concatenated all the recordings of noise into three independent partitions and added them to the smartphone-simulated recordings of the RECOLA database, with two different signal-to-noise ratios (SNRs): 0 and 6 dB.

In the first part dealing with acoustic distance measures (Section 3.2), we use the first six recordings of the RECOLA training set (30 minutes in total) and study the quality of noise reduction using these distance measures. The original audio is used as reference signal.

An emotion recognition system is then equipped with different types of noise reduction preprocessing, in both the training and evaluation phase, using the most promising variants of the proposed methods and the relevant baseline methods. The different systems are trained on the training set of RECOLA to predict continuous-valued arousal and valence using support vector regression (SVR) on the mean and variance of 13 mel-frequency cepstral coefficients (MFCCs). In the training phase, we adjust the windowing for mean and variance computation to be optimal according to the validation set, as in [31]. For SVR, we use a linear kernel and also tune the complexity on the validation partition [32]. Post-processing of the predictions is applied using the same methodology as described in [30]. Performance in emotion recognition is evaluated with the concordance correlation coefficient (CCC) [15, 31].

3.2 Acoustic Quality Measures

Several quality measures exist for automatically evaluating denoising and enhancement results for speech signals. These include simple energy-based measures such as SNR and segmental SNR, perceptual measures such as PESQ and distance measures in different representation domains of the magnitude spectrum [16].

The short-time magnitude spectrum is the basis of most audio feature extraction techniques. Audio pattern recognition and machine learning systems generally aim to distinguish sound classes based on the distributions of the *shape* of the short-time magnitude spectrum, and typically are not concerned with the true sound level or its psychoacoustical counterpart, loudness [33]. Therefore, in this study, we choose not to use measures that depend on the signal level in any manner, and whose results would depend on proper gain adjustment applied to the enhanced signal. These include the SNR measures as well as the RMS log-spectral distance [11]. We also do not use perceptual measures that explicitly aim to predict subjective quality or intelligibility. Instead, we focus on the preservation of the information about the shape of the short-time magnitude spectrum. These considerations resulted in the choice of two measures. Firstly, we apply the Itakura distance [6, 14], given by $d_{\mathrm{ITA}} = \log(\boldsymbol{a}\boldsymbol{R}_0\boldsymbol{a}'/\boldsymbol{a}_0\boldsymbol{R}_0\boldsymbol{a}_0')$, where $\boldsymbol{a}_=(1, a_1, \ldots, a_p)$ is the all-pole model under test for its similarity to the reference model $\boldsymbol{a}_0 = (1, a_{0,1}, \ldots, a_{0,p})$, estimated using the Toeplitz autocorrelation matrix $\boldsymbol{R}_0$ [17]. The numerator represents the total squared prediction error of the reference signal using the test model and the denominator represents the error using the reference model. In this study, $\boldsymbol{R}_0$ and the reference model $\boldsymbol{a}_0$ are obtained from the original, i. e., not noisified signal. This is a purely signal-based, non-perceptual distance measure that focuses on accuracy in all-pole modeling, which in turn is known to focus on spectral peaks and formants [17, 22]. As another measure, we apply the Euclidean distance between two mel-frequency cepstral vectors, the noisy test vector and the reference vector obtained from the original signal, consisting of 12 MFCCs while excluding the zeroth coefficient [13]; we denote this measure by d_{MFCC}. This measure has the perceptual aspect of an auditory-based warped frequency scale and it is also connected to the feature extraction procedure of many machine learning systems, including the one used in the present study.

Fig. 2 shows the Itakura and MFCC distances, averaged over a large number of speech frames, computed such that denoised frames are compared against corresponding reference frames in the original signal. Three conditions are con-

[1] https://www.freesound.org/

sidered: smartphone speech, and smartphone with *CHiME* and *trains* noise, each added at 6 dB SNR.

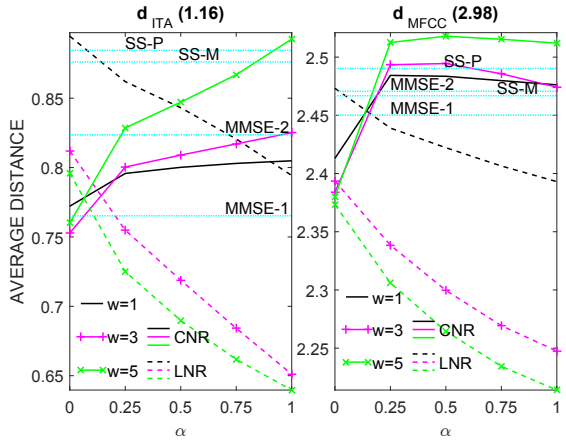

Figure 2: Signal degradation (left: Itakura distance, right: MFCC distance) evaluated for denoised signals against original, frame by frame, averaged over three noise conditions and the 25 % frames with most target signal energy. The performance of the proposed methods is shown with respect to the temporal smoothing parameter w and the noise model adaptation rate α. Corresponding averages for unprocessed noisy signals are shown in parentheses.

3.3 Emotion Recognition

Table 1 shows the emotion recognition results in various noise conditions with approximately matched training. When training is performed with the original audio material, the system is evaluated with the original and smartphone-processed material. For *CHiME* and *trains* noise, the training is performed with 6 dB SNR and the system is evaluated with 0 dB and 6 dB SNR of the same type of noise corruption. This setup requires the noise reduction preprocessing methods to preserve emotion-related information to a sufficient precision in both the training and recognition phases.

The two leftmost results columns indicate that denoising clean speech generally degrades recognition performance. A similar observation has been made in a recent study on feature enhancement with the same data [32]. An exception here, however, is the cepstral approach with moderate adaptation and no temporal smoothing (CNR/1,0.5), which actually improves prediction of the arousal dimension. None of the evaluated methods improve the prediction of arousal under the nonstationary *CHiME* noise; this might be explained by highly variable residual noise corrupting spectral energy trajectories. Otherwise, the proposed methods (CNR, LNR) outperform both the not denoised approach and the standard baseline noise reduction methods.

4. CONCLUSIONS

Noise reduction methods were evaluated with a focus on paralinguistic machine learning applications (two emotion recognition tasks). Baseline methods were chosen based on earlier, related studies and compared with a proposed new approach, which can be configured according to the specific learning task and the noise conditions using a small number of parameters. The experiments involved objective signal degradation measures and recognition of emotions on a multimodal interaction corpus. In the former evaluation, the proposed approach is shown to perform competitively to the standard noise reduction methods (spectral subtraction and MMSE, recommended as robust baseline methods in the field of speaker recognition [9]) over most of the range of the parameter values. By increasing the temporal smoothing and adaptation rate in noise modeling, the log-spectral approach clearly outperforms the standard techniques.

In emotion recognition evaluations, arousal and valence are noticed to place somewhat different requirements on the denoising scheme. Denoising improves the results in a majority of the evaluated training/test setups and in these cases, the proposed methods outperform the standard baselines. The cepstral approach without temporal smoothing of the noise leads to noticeable improvement in the high-SNR conditions. In noisy training and test conditions, temporal smoothing combined with adaptation outperforms all other evaluated approaches. Therefore, it is noted that the proposed methods show promise in audio noise reduction and audio content analysis. In future work, the relationships of different learning tasks, noise conditions and noise reduction approaches can be further investigated.

Table 1: Correlation (CCC) between predicted and ground-truth arousal and valence over the test set for closely matched training while using a given denoising method in both training and testing. The cases in which denoising improves upon the not denoised case are indicated in bold.

Training	Original		*CHiME* 6 dB		*trains* 6 dB	
Noise reduction method (w,α)	Test: Orig. –	Test: phone –	Test: *CHiME* 6 dB	Test: *CHiME* 0 dB	Test: *trains* 6 dB	Test: *trains* 0 dB
	AROUSAL					
none	0.735	0.728	0.670	0.557	0.483	0.379
SS-P	0.704	0.697	0.634	0.553	**0.538**	**0.441**
SS-M	0.708	0.701	0.637	0.535	**0.564**	**0.488**
MMSE-1	0.672	0.680	0.601	0.490	**0.566**	**0.488**
MMSE-2	0.679	0.698	0.620	0.510	**0.514**	**0.430**
CNR (1,0.5)	**0.753**	**0.755**	0.599	0.502	**0.535**	**0.438**
CNR (20,1.0)	0.676	0.650	0.642	0.539	**0.547**	**0.440**
LNR (1,0.5)	0.706	0.664	0.635	0.538	**0.532**	**0.435**
LNR (20,1.0)	0.676	0.650	0.641	0.535	**0.701**	**0.637**
	VALENCE					
none	0.400	0.342	0.173	0.120	0.154	0.108
SS-P	0.328	0.262	0.112	0.107	**0.195**	**0.228**
SS-M	0.386	0.303	0.128	0.114	**0.190**	**0.177**
MMSE-1	0.318	0.294	0.098	0.066	**0.218**	**0.190**
MMSE-2	0.340	0.292	0.087	0.071	**0.219**	**0.160**
CNR (1,0.5)	0.308	0.308	**0.180**	**0.127**	**0.243**	**0.156**
CNR (20,1.0)	0.207	0.179	**0.261**	**0.227**	**0.252**	**0.198**
LNR (1,0.5)	0.359	0.288	**0.174**	**0.124**	0.147	0.107
LNR (20,1.0)	0.278	0.250	0.137	0.109	**0.241**	**0.218**

5. ACKNOWLEDGMENTS

This work has been supported by the European Community's Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu).

6. REFERENCES

[1] J. Barker, E. Vincent, N. Ma, C. Christensen, and P. Green. The PASCAL CHiME speech separation and recognition challenge. *Computer Speech and Language*, 27(3):621–633, Nov. 2013.

[2] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. ICASSP*, Washington, D.C., USA, Apr. 1979.

[3] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 27(2):113–120, Apr. 1979.

[4] M. Brookes. VOICEBOX: Speech processing toolbox for MATLAB, speech enhancement functions. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html#enhance. Accessed: 2016-04-26.

[5] O. Cappé. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. Speech and Audio Processing*, 2(2):345–349, Apr. 1994.

[6] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, 1993.

[7] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, Dec. 1984.

[8] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 23(2):443–445, Apr. 1985.

[9] K. W. Godin, S. O. Sadjadi, and J. H. L. Hansen. Impact of noise reduction and spectrum estimation on noise robust speaker identification. In *Proc. Interspeech*, Lyon, France, Aug. 2013.

[10] M. Goudbeek and K. Scherer. Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *Journal of the Acoustical Society of America*, 128(3):1322–1336, Sept. 2010.

[11] A. H. Gray and J. D. Markel. Distance measures for speech processing. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 24(5):380–391, Oct. 1976.

[12] J. H. L. Hansen and M. A. Clements. Constrained iterative speech enhancement with application to speech recognition. *IEEE Trans. Signal Processing*, 39(4):795–805, Apr. 1991.

[13] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing*. Prentice Hall PTR, 2001.

[14] F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 23(1):67–72, Feb. 1975.

[15] L. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, Mar. 1989.

[16] P. C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[17] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, Apr. 1975.

[18] M. Mauch and S. Ewert. The audio degradation toolbox and its application to robustness evaluation. In *Proc. ISMIR*, Curitiba, Brazil, Nov. 2013.

[19] J. Ortega-Garcia and J. Gonzalez-Rodriguez. Overview of speech enhancement techniques for automatic speaker recognition. In *Proc. ICSLP*, Philadelphia, PA, Oct. 1996.

[20] K. Paliwal, B. Schwerin, and K. Wójcicki. Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. *Speech Communication*, 54(2):282–305, Feb. 2012.

[21] K. Paliwal, K. Wójcicki, and B. Schwerin. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech Communication*, 52(5):450–475, May 2010.

[22] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[23] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30, Nov. 2015.

[24] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proc. FG*, Shanghai, China, Apr. 2013.

[25] S. O. Sadjadi and J. H. L. Hansen. Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions. In *Proc. Interspeech*, Makuhari, Japan, Sept. 2010.

[26] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku. Temporally weighted linear prediction features for tackling additive noise in speaker verification. *IEEE Signal Processing Letters*, 17(6):599–602, June 2010.

[27] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll. Emotion recognition in the noise applying large acoustic feature sets. In *Proc. Speech Prosody*, Dresden, Germany, May 2006.

[28] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. Paralinguistics in speech and language - state-of-the-art and the challenge. *Computer Speech and Language*, 27:4–39, 2013.

[29] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll. Recognition of noisy speech: A comparative survey of robust model architecture and feature enhancement. *EURASIP J. on Audio, Speech, and Music Processing*, 2009:17 pages, 2009.

[30] G. Trigeorgis, F. Ringeval, R. Bruckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? End-to-end speech emotion recognition using a Deep Convolutional Recurrent Network. In *Proc. ICASSP*, Shanghai, China, Mar. 2016.

[31] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016 – depression, mood, and emotion recognition workshop and challenge. In *Proc. AVEC*, Amsterdam, Oct. 2016.

[32] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller. Facing realism in spontaneous emotion recognition from speech: Feature enhancement by LSTM neural autoencoders. In *Proc. Interspeech*, San Francisco, CA, Sept. 2016.

[33] E. Zwicker and H. Fastl. *Psychoacoustics, Facts and Models*. Springer-Verlag, Berlin, 1990.