MULTI-SCALE MODULATION FILTERING IN AUTOMATIC DETECTION OF EMOTIONS IN TELEPHONE SPEECH

Jouni Pohjalainen and Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

ABSTRACT

This study investigates emotion detection from noise-corrupted telephone speech. A generic modulation filtering approach for audio pattern recognition is proposed that utilizes inherent long-term properties of acoustic features in different classes. When applied to binary classification along the activation and valence dimensions, filtering the baseline short-time timbral features in both the training and detection phase leads to significant improvement especially in noise robustness. Automatic selection of training data based on the filter's prediction residual further improves the results.

Index Terms— emotion detection, speech analysis, computational paralinguistics

1. INTRODUCTION

Many potential applications exist for robust automatic recognition of emotions in speech. These include adapting speech recognizers to the speaker's emotional state, providing paralinguistic input to user interfaces and monitoring the quality of call center service. The potential applications are one reason for speech emotion recognition becoming a topic of active research in recent years [1] [2] [3]. Automatic emotion recognition is often studied in the sense of identifying emotions among a fixed set of classes such as anger, joy, sadness or surprise [1]. Some practically oriented studies target the detection of individual emotions such as anger [4] [5] [6] [7]. In addition, binary classifications are studied along various dimensions, such as between different emotions [3] or along two central affective dimensions, namely activation (arousal) and valence. These dimensions, which can be characterized as distinguishing between calm/excited and positive/negative emotions, respectively, are generally viewed in psychology as the two most important ones to represent emotions, although additional dimensions have been studied for more complete low-dimensional representation [8] [9].

Relatively few studies still specifically target system robustness in real-world conditions, but this aspect has gradually gained momentum [1] [10] [11] [12]. These previous studies on the robustness aspect usually investigate the emotion-class identification problem. On the other hand, especially anger detection systems have typically been evaluated with real-world call center data [4] [5] [6]. However, it can be argued that broader characterizations of the caller's emotional state could prove more useful in realistic call center applications, and this approach has indeed been investigated in some studies [13] [14]. Emotion identification with multiple classes may be unnecessarily complex and error-prone, while anger detection systems may tend to focus on the easily detectable hot anger (high vocal effort is relatively easy to detect [15] [16]) but perform poorly with cold anger.

This work was supported by the EC FP7 project Simple4All (287678).

In order to address specifically the issue of recognizing, broadly, the caller's emotional state, this study focuses on activation and valence analysis of telephone speech. In order to build a robust system applicable in real-world conditions, the focus is on *mismatched* acoustic conditions in the recognition phase, with changing types of far-end ambient noise corruption. Concerning the study material, with real telephone speech it would be difficult to relate the results to other studies and to assess the proportional effects of the problem's inherent difficulty, noise, bandwidth and channel on the system performance. In addition, the spectrum of emotions in spontaneous speech would be less diverse than in acted databases. Thus, to focus on the telephone channel and its noise robustness issues in a quantitative manner, this study uses acted speech from the widely used Berlin database of emotional speech [17] after processing it with a simulated telephone transmission channel. However, the system is also evaluated with the original, high-quality speech material in order to more generally validate the proposed approach.

The approach chosen to tackle these issues is a general, customizable method for modeling long-term modulation dynamics in the speech signal on multiple time scales simultaneously. It is based both on the known perceptual relevance of such temporal information in auditory pattern recognition and the specific usefulness of this information in emotion recognition from speech, which has been observed in many previous studies. These considerations are discussed in Section 2 before introducing the method in Section 3.

2. LONG-TERM TEMPORAL DYNAMICS IN SPEECH AND AUDIO

2.1. Approaches to Modeling Long-Term Information

In speech and audio processing, the long-term modulation characteristics of short-term acoustic parameters are typically utilized for two purposes. One is to improve the system's robustness by *filtering* out temporal changes which are unlike the usual temporal dynamics of the target signal and are thus more likely to be unpredictable, noisy components. This is the case, for example, in the widely used RASTA modulation filtering [18]. It employs an IIR band-pass filter $H(z) = 0.1z^4 \frac{2+z^{-1}-z^{-3}-2z^{-4}}{1-0.98z^{-1}}$ separately on each spectral (or more generally, timbral) feature across time (speech frames) and captures the typical modulation frequencies of speech (with the typical frame interval of 10 ms). It has been used for improving channel and noise robustness in automatic speech and speaker recognition.

Apart from improved focus on the target signal, another reason for utilizing long-term temporal information is to *distinguish* between broad classes. In speech processing, modeling class-discriminative long-term characteristics has been found to be important in applications such as language identification [19], paralinguistic speaker state and trait classification [20], speaker recognition [21] and speech emotion recognition. These applications get benefit

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

from long-term dynamic information which conventional short-time features and delta features [22] can not accurately capture.

Classification of vocal emotions, the focus of this study, benefits from long-term modeling, as different emotional speech classes generally have quite different temporal dynamics [23]. Some of the approaches proposed to modeling this temporal information in speech emotion recognition are feature extraction explicitly utilizing the modulation spectrum [2] [12], mel-cepstral analysis for low frequencies [5], modeling F0 contours [3], autoregressive features [24] and vector autoregressive classification models [25]. A common, generic approach which is widely used in computational paralinguistics (including emotion recognition) is to compute a large number of long-term statistics and functionals of a comprehensive set of frame-based short-time features and to utilize a machine learning approach capable of handling high dimensionalities [1] [20] [26].

The current study proposes a generic, simple method for modeling long-term temporal dynamics of different classes by *multi-scale* autoregressions. The method is applied to emotion detection. In contrast to the mentioned specialized approaches to capturing this information in computational paralinguistics studies, the proposed method is application-independent and utilizes intermediate, *class-specific* modulation filtering of short-term features across signal frames. It is computationally light and can be straightforwardly plugged in between the initial short-time feature extraction phase and the final classification phase of a generic audio pattern recognition system. To clarify, its place as an optional component in the order of processing steps in such a system would be as follows:

- Training phase:
    1. Feature extraction
    2. *The proposed method*
    3. Training of classification models
- Classification phase:
    1. Feature extraction
    2. *The proposed method*
    3. Classification decisions

In addition to modeling class-discriminative information, the proposed method also acts as a filter removing uncharacteristic and potentially noisy dynamic components from the feature representation, similarly to RASTA apart from being always tuned to a particular class. Consequently, one of the primary goals of this study is to investigate whether the proposed approach can improve the classification system's noise robustness in realistic, adverse conditions such as far-end noise corruption over the telephone channel.

2.2. Perceptual Importance of Modulation in Speech Communication

In the auditory system, short-term acoustic characteristics are analyzed by the auditory periphery, in particular by the cochlea in the inner ear. In general, the nuclei along the neural auditory pathway from the cochlea towards the auditory cortex generate progressively more sophisticated and longer-term representations of the auditory sensation while also acting as relay stations for lower-level representations from the earlier stages [27] [28]. However, the exact functionality of the different stages of the neural pathway is less well understood than that of the inner ear. The first stage after the cochlea, the *cochlear nucleus* contains neurons with different time responses: primary-like, onset, chopper, pauser and buildup. Many different kinds of abstractions of the original auditory stimulus are generated already in specialized regions of the cochlear nucleus. These are passed on through the *superior olivary complex*, the initial site of bilateral representation of the acoustic environment, to the *inferior colliculus*, believed to be specialized in the representation of pitch and in localizing sound sources consisting of complex temporal variations. The cells of the inferior colliculus display *modulation frequency selectivity* and phase-lock to amplitude modulations of the stimulus. In the *medial geniculate nucleus* (MGN), the final waystation on the way to the auditory cortex, the cells also phase-lock to amplitude modulation but with lower temporal resolution, i.e., with lower modulation frequencies being represented. In addition to its role as a relay station for an auditory pathway conveying all the information necessary to characterize acoustic events, it has been suggested that the MGN is also involved in a second pathway that allows the auditory cortex to selectively label stimuli with perceptual qualities. Thus, it would play an essential role in the perception of the acoustic environment.

Psychoacoustical studies have examined the temporal properties of hearing by utilizing various approaches and concepts, including temporal integration [29], temporal masking [30] and the ability of listeners to detect sinusoidal amplitude modulation [31]. The latter studies typically examine the temporal modulation transfer function (TMTF), i.e., the sensitivity to amplitude modulation as a function of frequency. Humans have been found to be most sensitive to amplitude modulation at modulation frequencies below roughly 10 Hz (see, e.g., [31]). It has been suggested that spectral and temporal integration characteristics of hearing may be the result of an optimization mechanism (either innate, learned or some combination of these) for detecting patterns in the typical acoustic input, such as speech [32]. In speech, most of the modulation energy is concentrated between 2 and 8 Hz, especially near 4 Hz [33]. Energy in this range is largely affected by phonemic and syllabic variation.

In summary, as the modulation energy in speech concentrates on a certain modulation frequency range, which is also the area hearing is the most sensitive to, perceptually important differences between various speech classes probably also manifest themselves in this region. Systematic, generalizable approaches to modeling this information are thus important topics for study. Moreover, given that the auditory system generates modulation frequency representations on varying temporal scales at different stages of the auditory pathway, being able to separately model these different time scales in automatic recognition of sound classes could potentially offer an advantage and should be specifically investigated. These observations offer perceptual motivation for the particular modeling approach investigated in the present study.

3. THE EMOTION DETECTION SYSTEM

3.1. Feature Extraction

After pre-emphasis with $H_p(z) = 1 - 0.97z^{-1}$, the input signal is arranged into overlapping Hamming-windowed frames of 25 ms with a shift interval of 10 ms. Using the standard processing chain of 1) computation of the the squared magnitude spectrum by fast Fourier transform (FFT), 2) mel-frequency filterbank (here, 40 triangular filters spaced evenly on the mel scale), 3) logarithm and 4) discrete cosine transform, 12 mel-frequency cepstral coefficients (MFCCs) are obtained after exclusion of the zeroth coefficient [22]. These are complemented with the logarithmic frame energy, whose value is locally normalized for the mean and and variance over the utterance, and delta and double-delta coefficients [22] to form a 39-dimensional feature vector.

3.2. The Proposed Filtering Method

In a previous study on detection of angry speech [7], the following method was used to model the long-term dynamics of each MFCC feature. In the training phase, an autoregressive (AR) model is first trained for each feature to represent its long-term behavior *in the target class* according to the training data. These AR models are then used, in both the training and the detection phase of the system, as finite-impulse-response (FIR) filters to generate linear predictions of the features $x_{j,t}$ based on their values over the preceding frames,

$$\hat{x}_{j,t} = c_j + \sum_{k=1}^{r} b_{j,k} x_{j,t-sk}, \quad (1)$$

where j is the index of the feature in the feature vector, t is the frame index, c_j are intercept terms and $b_{j,k}$ are the AR coefficients. Next, the original features $x_{t,j}$ are *replaced* by the predictions $\hat{x}_{t,j}$.

In training the AR models (using least squares estimation) and subsequently using them to generate predictions according to Eq. 1, choosing the integer frame skip parameter as $s = 1$ leads to conventional autoregressions. If s is chosen to be greater than 1, the AR model only sees every sth frame in each prediction. The length of the autoregression history is thereby increased without increasing the number of parameters to be estimated. In the previous study, autoregression order $r = 8$ and frame skip $s = 4$ gave the overall best results in the detection of anger in telephone speech [7]. This choice of parameters corresponds to every prediction being based on frame lags $(4, 8, \ldots, 32)$ and, with a frame shift interval of 10 ms, the total duration of signal history utilized becomes 320 ms.

In the present study, we first attempted to apply this approach to speech classification along the activation and valence dimensions. However, determination of a suitable combination of order r and frame skip s proved to be difficult in this more general case, even though the combination found in the previous study works well for the special case of detecting anger in speech.

These preliminary results motivated the idea of using *multiple* autoregressive filters, with different time scales (different frame skip parameters in Eq. 1), to accomplish variable focus on different modulation frequencies in different contexts. For each feature j at each time instant t, the filter resulting in the most accurate prediction would be used to generate the final prediction. That is, the prediction of the jth feature in the tth frame according to the nth filter is, analogously to Eq. 1, given by

$$\hat{x}_{j,t,n} = c_{j,n} + \sum_{k=1}^{r} b_{j,k,n} x_{j,t-s_n k}, \quad (2)$$

where $b_{j,k,n}$ are the autoregressive parameters of the nth filter, s_n is the nth filter's frame skip parameter and $c_{j,n}$ its intercept. The predicted value for the jth feature in the tth frame is then chosen as

$$\hat{x}_{j,t} = \arg\min_{\hat{x}_{j,t,n}} (x_{j,t} - \hat{x}_{j,t,n})^2, \quad (3)$$

i.e., as the output of the filter that results in the lowest squared prediction error. This prediction $\hat{x}_{j,t}$ then replaces $x_{j,t}$.

3.3. Decision Rule

For each emotion class, after learning the filter described by Eqs. 2 and 3 to represent that emotion's typical temporal characteristics, the filter is applied to all the training data before training the detector for that emotion. Each such detector uses one diagonal-covariance, 64-component Gaussian mixture model (GMM) to represent the target emotion and another such GMM to represent all other emotions.

In the detection phase, each class detector first again applies the modulation filter optimized for its target class (emotion) X and computes a detection statistic L_X that is the difference of frame-average log likelihoods of the two GMMs over an utterance [7].

The emotions are subsequently combined into larger classes defined by high/low activation and positive/negative valence. The decision statistic for high activation, for instance, is then the maximum of detection statistics L_X among high-activation emotion detectors minus the maximum value of L_X among low-activation detectors.

3.4. Selection of Training Data Using the Filters

Because classifications are made based on the outputs of filters optimized for each class, for congruency it makes sense to consider selecting "typical" training vectors that are well predicted by those same filters. That is, after extracting features in the training phase, estimating the filter parameters to represent a particular class and using them to filter the training data, we can choose to keep only those feature vectors which were predicted well, according to some criterion. Only these vectors will be used in training the GMMs.

In the present work, we investigate an unsupervised criterion for selecting training vectors according to the filtering result. Let $e_{j,t} = x_{j,t} - \hat{x}_{j,t}$ be the prediction residual of the jth feature in the tth frame ($\hat{x}_{j,t}$ given by Eq. 3). We cluster the values $\sum_j e_{j,t}^2$ within a block of 10 s (1000 frames) using k-means into two clusters initialized with $\min_t(\sum_j e_{j,t}^2)$ and $\max_t(\sum_j e_{j,t}^2)$ and keep only the former cluster, i.e., the feature vectors with low prediction error.

4. EXPERIMENTAL EVALUATION

4.1. Speech Material

The test material in this study is the Berlin database of emotional speech [17]. The database is used both as is and by simulating speaking on a telephone in a noisy environment. There are 535 German sentences in the database, spoken by five male and five female actors. The emotion categories are (hot) anger, boredom, disgust, fear, happiness, sadness and neutral. We consider the emotions anger, fear and happiness to represent high activation (arousal) and the other categories to be low activation [8] [9]. This gives 267 and 268 instances of high and low activation, respectively. Similarly, we consider the emotions anger, disgust, fear and sadness to represent negative valence and boredom, happiness and neutral to represent positive or approximately neutral valence [9]. This gives 304 and 231 instances of negative and positive/neutral valence, respectively.

4.2. Simulation of Noisy Telephone Speech

Additive noise from the NOISEX-92 database was first added to the signal in order to simulate noise at the location of a mobile station. Three noise types were used: *volvo* (recorded inside a moving car), *factory1* (mechanical factory noise with frequent transient/impulsive sounds) and *babble* (a large number of people talking simultaneously). The noise corruption was done at 16 kHz sampling rate with a controlled segmental signal-to-noise ratio (SNR), i.e., the average over 25-ms frames. As in [15], noise-corrupted speech signals sampled at 16 kHz were then high-pass filtered with the mobile station input (MSIN) filter, which approximates the input characteristics of a mobile terminal [34], and decimated to the sampling rate of 8 kHz. The speech level was normalized to 26 dB below overload point. Finally, the signals were processed with the adaptive multi-rate (AMR) codec [35], which is commonly used for speech coding in the GSM

cellular system, at a bit rate of 12.2 kbps. Fig. 1 shows an example of processed speech and results of applying the filtering method.

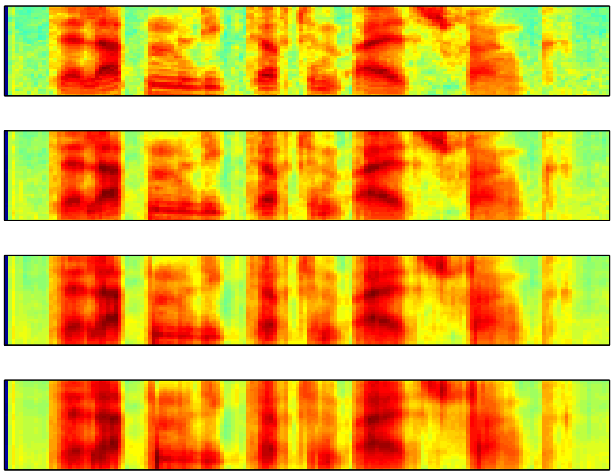

Fig. 1. Top panel: mel-scale spectrogram, with 40 bins, transformed back from MFCCs for a neutral telephone utterance (original label 03a01Nc) corrupted by car interior noise (SNR 0 dB). Lower panels: mel-scale spectrograms for the same utterance after filtering the original MFCCs with multi-scale autoregressive predictors for classes "anger", "neutral" and "happiness".

4.3. Evaluation Method

The evaluation is performed as leave-one-speaker-out cross validation, where one of the ten speakers in turn becomes the test speaker and the speech material of the other nine speakers is used for training. Detection statistics are obtained separately for each utterance for both tasks (activation and valence). They are used to compute equal error rates (EER), a widely used performance measure in evaluating detection and binary classification systems. The EER is the misclassification rate when the detection threshold is set in such a way that it becomes equal in both classes. The differences between the filtering methods and the baseline are statistically analyzed using a significance test appropriate for detection systems [36]. As all the detections use the same analysis block division and original speech material, the "dependent-case" version of this test is employed.

4.4. Results

Tables 1 and 2 show EER scores in various conditions for the baseline (standard MFCCs with energy and deltas) and two class-specific temporal filtering approaches applied to the MFCC features: basic autoregressive (AR) filtering (similar to [7]) and the proposed multi-scale AR filtering. For analyzing the original, clean speech material, the system has been trained with original data. For testing in the telephone conditions with far-end noise corruption, the system has been trained using telephone speech with high-SNR (30 dB) car noise.

We encounter the known phenomenon that valence is more difficult to detect than activation [37]. The class-specific multi-scale filtering outperforms class-specific simple AR filtering (which in [7] outperformed baseline MFCCs with and without RASTA in detecting angry speech) as well as the baseline in both tasks, indicating that it succeeds in capturing class characteristics that are helpful for robustness. In particular, the method improves the robustness of activation detection. When combined with training data selection (Section 3.4), however, it notably also achieves statistically significant improvement in valence detection in the clean speech case.

Table 1. *EER scores (%) for the detection of high-activation emotions anger, fear and happiness. The scores that are statistically significantly better than the baseline in the corresponding noise and channel conditions are indicated in boldface. The maximum prediction lag (determined by maximum s_n) is varied between 400 and 600 ms in order to investigate the effect of low modulation frequencies.*

	Channel and noise condition			
	Original	Telephone (SNR 0 dB)		
	clean	car	factory	babble
baseline MFCCs	7.1	12.7	34.0	22.1
AR: $r = 50$, $s_n = 1$	9.0	12.7	32.5	23.9
$r = 10$, $1 \leq s_n \leq 4$	7.5	12.3	**21.7**	**17.9**
$r = 10$, $1 \leq s_n \leq 6$	6.7	**10.5**	**24.3**	**17.9**
$r = 10$, $1 \leq s_n \leq 5$	7.1	**10.5**	**22.1**	**17.9**
+ training data selection	7.1	**8.2**	**20.2**	**16.5**

Table 2. *EER scores (%) for the detection of negative-valence emotions anger, disgust, fear and sadness. The scores that are statistically significantly better than the baseline in the corresponding noise and channel conditions are indicated in boldface.*

	Channel and noise condition			
	Original	Telephone (SNR 0 dB)		
	clean	car	factory	babble
baseline MFCCs	22.2	27.8	45.2	36.4
AR: $r = 50$, $s_n = 1$	23.5	29.5	**41.7**	34.8
$r = 10$, $1 \leq s_n \leq 5$	21.3	27.8	**40.0**	**32.1**
+ training data selection	**20.0**	**25.2**	**39.4**	**34.2**

5. CONCLUSIONS

A simple, generalizable, customizable and domain-independent modulation filtering method was introduced in this study and applied to detecting emotions in adverse conditions, i.e., noise-corrupted telephone-channel speech. The method significantly improved upon the baseline MFCC features in terms of robustness by filtering them in a way that emphasizes class-specific long-term temporal dynamics. Automatic selection of training data for training the GMM-based detection system led to further improvement.

In light of the results, the filtering appears to both remove uncharacteristic, likely noisy temporal components and may also offer improvement in clean speech by providing more accurate modeling of vocal emotional classes. The results obtained hold promises for future studies in related applications in the field of computational paralinguistics and, for example, in automatic detection of the speaker's mental or physical state from telephone speech.

Subsequently, the filtering method has also been applied to noise reduction in speech enhancement. An implementation of the filter algorithm and speech enhancement examples are available at `http://www.acoustics.hut.fi/research/robustness/`.

6. REFERENCES

[1] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proc. ICASSP*, Honolulu, Hawaii, April 2007.

[2] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, pp. 768–785, 2011.

[3] J. P. Arias, C. Busso, and N. B. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Comput. Speech Lang.*, vol. 28, pp. 278–294, 2014.

[4] F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann, and W. Burleson, "Detecting anger in automated voice portal dialogs," in *Proc. Interspeech*, Pittsburgh, USA, Sept. 2006.

[5] D. Neiberg and K. Elenius, "Automatic recognition of anger in spontaneous speech," in *Proc. Interspeech*, Brisbane, Australia, Sept. 2008.

[6] M. Erden and L. M. Arslan, "Automatic detection of anger in human-human call center dialogs," in *Proc. Interspeech*, Florence, Italy, Aug. 2011.

[7] J. Pohjalainen and P. Alku, "Automatic detection of anger in telephone speech with robust autoregressive modulation filtering," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[8] J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, pp. 1050–1057, 2007.

[9] M. Goudbeek and K. Scherer, "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," *J. Acoust. Soc. Am.*, vol. 128, no. 3, pp. 1322–1336, 2010.

[10] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Emotion recognition from noisy speech," in *Proc. ICASSP*, Toulouse, France, May 2006.

[11] A. Tawari and M. Trivedi, "Speech emotion analysis in noisy real-world environment," in *Proc. Int. Conf. Pattern Recognition*, Istanbul, Turkey, Aug. 23–26 2010.

[12] T.-S. Chi, L.-Y. Yeh, and C.-C. Hsu, "Robust emotion recognition by spectro-temporal modulation statistic features," *J. Ambient Intell. Human Comput.*, vol. 3, pp. 47–60, 2012.

[13] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers," in *Proc. Interspeech*, Lisbon, Portugal, Sept. 2005.

[14] L. Devillers, C. Vaudable, and C. Chastagnol, "Real-life emotion-related states detection in call centers: A cross-corpora study," in *Proc. Interspeech*, Makuhari, Japan, Sept. 2010, pp. 2350–2353.

[15] J. Pohjalainen, T. Raitio, H. Pulakka, and P. Alku, "Automatic detection of high vocal effort in telephone speech," in *Proc. Interspeech*, Portland, Oregon, USA, Sept. 2012.

[16] J. Pohjalainen, T. Raitio, S. Yrttiaho, and P. Alku, "Detection of shouted speech in noise: human and machine," *J. Acoust. Soc. Am.*, vol. 133, no. 4, pp. 2377–2389, April 2013.

[17] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Lisbon, Portugal, Sept. 2005.

[18] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Speech Audio Process.*, vol. 2, pp. 578–589, 1994.

[19] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht, "Rhythmic unit extraction and modelling for automatic language identification," *Speech Communication*, vol. 47, no. 4, pp. 436–456, 2005.

[20] 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," *Comput. Speech Lang.*, vol. 27, pp. 4–39, 2013.

[21] S. H. Mallidi, S. Ganapathy, and H. Hermansky, "Robust speaker recognition using spectro-temporal autoregressive models," in *Proc. Interspeech*, Lyon, France, Aug. 2013.

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

[23] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech," in *Proc. Interspeech*, Jeju Island, Korea, Oct. 2004, pp. 2193–2196.

[24] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *IEEE Trans. Affective Computing*, vol. 3, pp. 116–125, 2012.

[25] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in *Proc. ICASSP*, Honolulu, Hawaii, April 2007.

[26] J. Pohjalainen, O. Räsänen, and S. Kadioglu, "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits," *Comput. Speech Lang.*, in press; available online 28 November 2013.

[27] B. C. J. Moore, Ed., *Hearing*, Academic Press, 1995.

[28] D. P. Morgan and C. L. Scofield, *Neural Networks and Speech Processing*, Kluwer Academic Publishers, 1991.

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

[30] W. Jesteadt, S. P. Bacon, and J. R. Lehman, "Forward masking as a function of frequency, masker level, and signal delay," *J. Acoust. Soc. Am.*, vol. 71, no. 4, pp. 950–962, April 1982.

[31] S. P. Bacon and N. F. Viemeister, "Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners," *Audiology*, vol. 24, pp. 117–134, 1985.

[32] O. Räsänen and U. K. Laine, "Time-frequency integration characteristics of hearing are optimized for perception of speech-like acoustic patterns," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 407–419, July 2013.

[33] S. Greenberg, "On the origins of speech intelligibility in the real world," in *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, 1997.

[34] Int. Telecommun. Union, *ITU-T G.191, Software tools for speech and audio coding standardization*, 2010.

[35] 3rd Generation Partnership Project, *3GPP TS 26.090, Adaptive multi-rate (AMR) speech codec, transcoding functions*, 2011, version 10.1.0.

[36] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proc. ODYSSEY04, The Speaker and Language Recognition Workshop*, Toledo, Spain, June 2004.

[37] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Proc. Interspeech*, Portland, Oregon, USA, Sept. 2012.