Filtering and Subspace Selection for Spectral Features in Detecting Speech Under Physical Stress

Jouni Pohjalainen, Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
jouni.pohjalainen@aalto.fi, paavo.alku@aalto.fi

Abstract

This paper investigates approaches to modeling the time evolution of short-time spectral features in paralinguistic speech type classification, where we focus on detection of speech influenced by physical exertion. The time series model consists of autoregressive processes of multiple time scales and orders and is trained to describe the long-term dynamics of a given target speech class. The model is applied in two ways in improving long-term modeling in the detection task: 1) to perform predictive filtering of the features and 2) to automatically select instantaneous classification subspaces. The spectrum analysis method underlying the short-time features is also varied between the standard discrete Fourier transform and a time-weighted linear predictive method which yields smooth all-pole spectrum envelope models. Configurations of the proposed methods are evaluated in the Physical Load task of the Interspeech 2014 Computational Paralinguistics Challenge and show improvement over the baseline timbral classifier and the challenge baseline. Also the interrelationships among the methods are discussed.

Index Terms: computational paralinguistics, physical load, modulation filtering, spectrum analysis

1. Introduction

Automatic detection of speaking styles, speaker states and speaker traits is a rapidly emerging field of study in speech technology. For detection of physical and emotional states, potential applications arise in, e.g., context-aware speech interfaces, call center service monitoring and in assisting speech and speaker recognition systems to adapt their acoustic models according to the speaking situation.

Modeling of temporal dynamics of speech is central in many of the above mentioned applications of computational paralinguistics. Often, the temporal dynamics are modeled by using a large set of long-term functionals of short-term acoustic parameters, and a machine learning system capable of tackling high-dimensional feature spaces is applied for the classification. This approach is demonstrated in, e.g., the baseline system of each Interspeech Computational Paralinguistics Challenge up to date [1]. In other studies, systems targeted for a particular application, such as emotion recognition, have used various customized approaches to modeling the temporal information, e.g., [2] [3] [4].

The application under study in this paper is automatic detection of physical stress, or physical load, according to the Physical Load sub-challenge of the Interspeech 2014 Computational Paralinguistics Challenge (ComParE2014) [1]. The goal is to combine robust short-time spectrum modeling with a versatile long-term modeling approach in order to tackle this complex detection task [5] [6] [7] [1], which is a challenge even for human listeners [6]. Recently, autoregressive predictive filtering of features using specifically tailored modulation time scales was proposed for automatic detection of anger in telephone speech [8] and for detecting broad emotional states according to the activation and valence dimensions [9]. The approach proposed was found to improve robustness in the presence of noise mismatch as well as the modeling of clean speech. In other recent studies, spectral envelope features based on time-weighted linear predictive methods have been shown to improve the robustness of classification systems [10] [11] [12]. In this study, these two techniques are combined: multiscale autoregressive modeling of long-term feature dynamics [9] is complemented with automatic selection of an instantaneous classification feature subspace, proposed in this paper, to process spectral short-term feature vectors; to obtain the initial short-term spectral features, we use a recently proposed, robust time-weighted linear predictive method [12].

Each method under study is applied as a straightforward modification of a basic acoustic classifier using short-term timbral features. Modulation filtering time scale parameters are optimized for two types of spectral features: those based on the discrete Fourier transform and those based on the linear predictive all-pole method. The experiments are evaluated by analyzing the interrelationships of long-term autoregressive modeling and the types of spectral features. The best results obtained are evaluated against two baselines: that of the basic classifier, without modifications, and the official baseline of ComParE2014 obtained by a high-dimensional, feature-intensive machine learning approach (support vector machine, SVM). Finally, the implications of the results obtained are discussed.

2. Detection System

2.1. Short-Term Feature Extraction

The speech signal, after pre-emphasis with filter $1 - 0.97z^{-1}$, is divided into Hamming-windowed frames of 25 ms with a 10 ms shift interval. For each such frame, 12 mel-frequency cepstral coefficients (MFCCs), excluding the zeroth one, are obtained by the process chain of 1) magnitude spectrum analysis, 2) computation of mel-filterbank energies, 3) logarithm and 4) discrete cosine transform [15]. A 39-dimensional feature vector is formed by concatenating the MFCCs with logarithmic frame energy, whose value is mean- and variance-normalized over the utterance, and Δ and $\Delta\Delta$ coefficients [15].

Typically, the initial spectrum analysis step in the MFCC computation is performed using the discrete Fourier transform, implemented by FFT (fast Fourier transform). In this study, we also investigate an alternative method which belongs to the family of spectrum analysis algorithms known as extended weighted linear prediction (XLP) [10] [11] [12]. Specifically, we employ a version of this general formulation which was, in an earlier study on speech emotion recognition, observed to result in smooth spectra and showed better noise robustness behavior than standard methods [12].

The most generic XLP formulation [12] solves the weighted normal equations

$$\sum_{k=1}^{p} a_k \sum_{n} Q_{n,j,k} s_{n-k} s_{n-j} = \sum_{n} Q_{n,j,0} s_n s_{n-j}, \qquad (1)$$
$$1 \leq j \leq p,$$

where s_n is a signal within the analysis interval and a_k are the coefficients of an all-pole model $H(z) = 1/(1 - \sum_{k=1}^{p} a_k z^{-k})$. $Q_{n,j,k}$ is a *snapshot weighting function* associated with time instant n and the pair of autocorrelation lags (j, k). The term "snapshot" is used because this weighting is applied to instantaneous products of signal samples, of the form $s_{n-i}s_{n-j}$, which, when summed over the time index n, constitute autocorrelation estimates $R_{i-j} = \sum_n s_{n-i}s_{n-j}$. When the autocorrelation "snapshot" terms are weighted by $Q_{n,j,i}$, the result is weighted autocorrelation $\tilde{R}_{i-j} = \sum_n Q_{n,j,i} s_{n-i} s_{n-j}$. Setting $Q_{n,j,i}$ a constant for all n, j and i results in conventional linear prediction (LP) [14]. In [12], the weighting function was instead obtained recursively as

$$Q_{n,j,k} = \frac{m-1}{m} Q_{n-1,j,k} + \frac{1}{m}\left(s_n^2 + |s_{n-j}||s_{n-k}|\right), \qquad (2)$$

with $Q_{n,j,k}$ initialized as zero outside the analysis interval. This weighting scheme consists of first-order lowpass filtering, using memory coefficient $(m-1)/m$, of the quantity $s_n^2 + |s_{n-j}||s_{n-k}|$. It therefore assigns most weight to the snapshot terms when 1) the absolute value of the snapshot, $|s_{n-j}||s_{n-k}|$, maintains a high value over a sufficiently long period and/or 2) the term s_n^2 maintains a high value, meaning that the short-time energy of the signal is large. This decreases the instantaneous random variation of spectrum estimates and results in smooth spectra whose time evolution follows smooth trajectories, as demonstrated in Fig. 1, in comparison with conventional LP. Increasing m makes the spectra smoother. The reduced spectral variability may contribute to robustness.

In previous studies on the detection of high vocal effort [13], it has been found useful to combine LP and FFT spectra in the following manner. First, the magnitude spectrum envelope is computed using LP. Second, the spectral fine structure, which comprises F0 and its harmonics, is obtained by eliminating the spectrum envelope from the FFT magnitude spectrum using cepstral source-filter separation. Specifically, the signal is transformed into the cepstral domain [15], liftered by suppressing to zero the cepstral coefficients corresponding to lags less than $(F_s/500) + 1$, where F_s is the sampling rate in Hz, and then transformed back to the spectral domain. This way, periodic excitation information up to 500 Hz is retained in the liftered excitation spectrum. Finally, the all-pole envelope and the cepstrally separated fine structure/excitation spectrum are multiplied together. The resulting magnitude spectrum can be denoted as LP-CR (linear prediction with cepstral residual) spectrum, or XLP-CR when based on an XLP spectrum envelope.

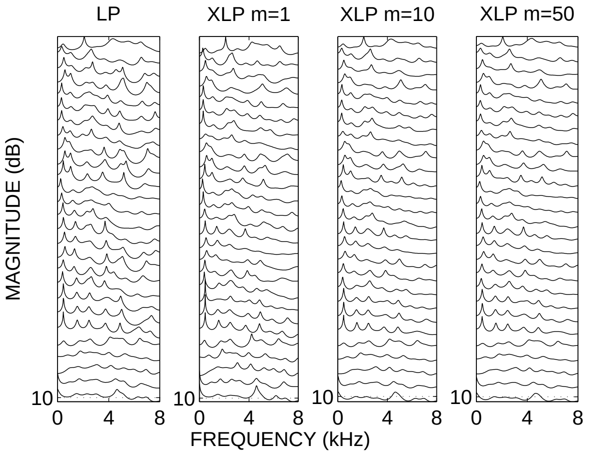

Figure 1: *LP and XLP spectra (with varying m) over one utterance of the Training material resampled at 16 kHz.*

2.2. Modeling of Feature Long-Term Dynamics

Long-term modeling is applied to the feature vector sequence where $y_{i,t}$, $1 \leq i \leq d$, constitute the feature vector describing the tth frame. In [9], a new approach was introduced which extended the autoregressive modeling of feature modulation dynamics presented in [8]. In the new approach, multiple autoregressive filters, parametrized by filter-specific prediction orders $q_{i,j}$, intercept terms $b_{0,i,j}$, autoregressive coefficients $b_{k,i,j}$ and frame skip parameters $S_{i,j}$ are used to generate predictions of the features such that

$$\hat{y}_{i,t,j} = b_{0,i,j} + \sum_{k=1}^{q_{i,j}} b_{k,i,j} y_{i,t-kS_{i,j}} \qquad (3)$$

is the predicted value of $y_{i,t}$ according to the jth filter trained to represent the time evolution of the ith feature. The filter coefficients $b_{k,i,j}$, $0 \leq k \leq q_{i,j}$, are obtained in the training phase by least squares estimation of autoregressive prediction coefficients to model the behavior of the ith feature across frames. In the estimation as well as in prediction, only every $S_{i,j}$th lagged value is considered for the filter $b_{k,i,j}$, $0 \leq k \leq q_{i,j}$. Varying the frame skip parameter $S_{i,j}$ across the filter index j, while the order of the filters stays constant as $q_{i,j} = q$, results in each feature being modeled in multiple different time resolutions, as in [9], while another alternative is to fix the time scale as $S_{i,j} = S$ and to instead vary the prediction order $q_{i,j}$.

For each frame and feature, the most accurate prediction over all filters is used as the final prediction. That is, the predicted value for the ith feature at the tth frame is chosen as

$$\hat{y}_{i,t} = \arg\min_{\hat{y}_{i,t,j}} (y_{i,t} - \hat{y}_{i,t,j})^2, \qquad (4)$$

i.e., as the output of the filter that results in the lowest squared prediction error. These predictions $\hat{y}_{i,t}$ are used either to *replace* the original features $y_{i,t}$ (filtering) in both training and classification phase, as done in [8] [9], or to select an instantaneous classification subspace by examining the prediction error (residual) $e_{i,t} = y_{i,t} - \hat{y}_{i,t}$, as described in Section 2.3.1. Both techniques emphasize time behavior typical to the target signal.

2.3. Classification Rule

The approach chosen is to detect one class of speech as opposed to another class. In practice, this is equivalent to binary classification, which is the goal of the evaluation task. In theory, however, the detector is trying to separate a target class, denoted as

1, from another class, denoted as 0, using a logarithmic likelihood ratio statistic in accordance with the Bayes rule.

The feature distributions of class 1 and class 0 are both modeled with 64-component Gaussian mixture models (GMMs) having a diagonal covariance structure [16]. They are trained using 10 iterations of the expectation-maximization (EM) principle [17] for GMMs [18] after initializing the component mean vectors by utilizing a vector selection algorithm intended for the initialization of iterative cluster-seeking algorithms [19]. The classification system and its training are similar to our previous studies on emotion recognition [8] [12].

Detection is done one utterance at a time. Frame-averaged log likelihoods of the feature vectors having been produced by each GMM are computed and denoted as L_1 and L_0, for classes 1 and 0, respectively. With T denoting the decision threshold, the detection rule for the logarithmic likelihood ratio is

$$L = L_1 - L_0 > T. \quad (5)$$

2.3.1. *Instantaneous Classification Subspace Selection*

In addition to comparing conventional and smoothed spectrum analysis in long-term modeling of spectral feature dynamics, this study also investigates an approach to utilize the long-term models in performing classification decisions. The principle is to use the long-term prediction residual (Section 2.2) $e_{i,t}$ (of the ith feature in the tth frame) in selecting a set of features I_t for the decision for the tth frame. To accomplish this, the likelihoods L_1 and L_0 in Eq. 5 are determined by using frame-specific diagonal-covariance multivariate Gaussian mixture distributions, with dimensionality $|I_t|$, to parametrize the joint distribution of the features $i \in I_t$ by means μ_i and variances σ_i^2. To summarize, only the features that were well predicted by the long-term modulation filter for the tth frame participate in the classification decision of that frame.

The frame-specific feature sets I_t are selected by a simple approach that divides the feature values into two clusters: one with a low prediction error and one with a high prediction error. The squared long-term prediction residuals $e_{i,t}^2$ are clustered using k-means by initializing the mean of one cluster with $\min_{i,t} e_{i,t}^2$ and the mean of the other cluster with $\max_{i,t} e_{i,t}^2$. After 10 iterations, the frame-specific instantaneous feature subsets are chosen as $i \in I_t$ if $e_{i,t}^2$ was assigned to the former, low-value cluster initialized with $\min_{i,t} e_{i,t}^2$. The means in one-dimensional k-means can not cross each other and this approach is guaranteed to divide a set of values into a "high" and a "low" cluster. Other approaches such as thresholding or comparing the squared prediction residual against a distribution learned in the training phase were experimented with, but of these, the proposed approach showed the best results in instantaneous classification subspace selection. This may be due to its ability to exploit any possible natural, bimodal clustering structures that may be present in the squared magnitude of the prediction error of the autoregressive modulation-frequency filters.

3. Experimental Results

The plan of the experiments is to first use cross-validation over the Training set of the ComParE challenge [1] to 1) adjust the time-scale parameters of the long-term modulation filtering and 2) choose the parameters of short-term spectrum analysis (Section 3.1). In this stage, the performance is measured as the unweighted average recall (UAR) at two operating points corresponding to different threshold values T in Eq. 5. Different values of T give rise to different miss rate p_{miss} and false alarm rate p_{fa}. The UAR at equal error rate (UAR-E) corresponds to $p_{\text{miss}} = p_{\text{fa}}$, in which case the UAR-E is given equivalently by $100\% \times (1 - p_{\text{miss}})$ and $100\% \times (1 - p_{\text{fa}})$. The other measure is UAR at minimum half-total error rate (UAR-M) [20], which is the maximum value of $100\% \times (1 - 0.5p_{\text{miss}} - 0.5p_{\text{fa}})$ obtainable by adjusting the threshold.

As the second step, selected modulation filter structures are used for filtering and subspace selection with features obtained using different spectrum analysis methods and these combinations are evaluated on both the Development and Test data of the Challenge (Section 3.2). In this stage, the goal is to set the UAR to its optimum value using the available information.

The speech material of the experiments comes from the Munich Bio-voice Corpus [21] according to the specifications of the ComParE2014 challenge [1]. The speech material was observed by listening to contain somewhat variable audible noise, motivating the use of noise-robust processing methods from [12] and [9]. The original speech material was downsampled to 16 kHz. The speech clips of the corpus are labeled as representing either 'low' or 'high' physical load. The detection system of Section 2.3 is set to detect the 'low' class as its target. The choice of the target class is irrelevant according to Eq. 5, but it affects the long-term dynamic modeling component as the modulation filter is trained to represent the target class.

3.1. Parameter Optimization

In the first part of the experiments, leave-one-speaker-out cross validation over the Training data set is used to examine the spectrum estimation methods for short-time acoustic feature extraction and the time scales for autoregressive modeling of these features. First, the autoregression (AR) order q and frame skip S in the single-AR filtering method from [8] (i.e., j is constrained to a single value in Eqs. 3 and 4) are optimized. Tables 1 and 2 show the results for 39-dimensional MFCC feature vectors obtained with FFT and LP ($p = 20$) spectrum analysis, respectively. Major differences in the optimal time scales between FFT- and LP-based short-time features are not observed, but the LP-based features appear to favor somewhat lower S values, especially $S = 1$ and $S = 2$, according to the UAR-E criterion. In the remainder of this study, we focus on frame skip range $S = 1, 2, 3$ and prediction orders $q = 4, 8, 12$.

Table 1: *UAR-M (%; UAR-E in parentheses) for single-autoregression filtering of FFT-based MFCCs.*

S	$q = 4$	$q = 8$	$q = 12$
1	69.2 (63.4)	65.6 (61.6)	67.4 (62.3)
2	66.8 (64.9)	65.3 (62.9)	64.9 (63.9)
3	67.3 (63.4)	64.7 (61.8)	66.2 (63.9)
4	64.4 (61.0)	63.4 (62.3)	65.8 (60.5)
5	63.7 (59.5)	62.8 (57.4)	64.1 (60.0)

Table 3 compares FFT, LP and LP-CR spectrum analyses without any long-term processing. As the "CR" features outperform FFT and LP at the equal error rate operating point, they are included in further evaluations. The performance boost obtained by the CR spectra can be predicted by previous studies. It is known that high vocal effort impacts F0 and associated harmonics, and CR-spectral features have shown improved robustness performance in detection of high vocal effort, e.g., [13]. Because, e.g., the periodic glottal waveform, which is re-

Table 2: *UAR-M (%; UAR-E in parentheses) for single-autoregression filtering of LP-based MFCCs.*

S	$q=4$	$q=8$	$q=12$
1	66.1 (63.4)	67.9 (61.8)	65.5 (62.3)
2	66.1 (60.5)	66.4 (62.3)	65.3 (62.9)
3	66.1 (62.9)	62.8 (59.5)	64.3 (59.5)
4	65.2 (62.9)	64.9 (59.5)	62.8 (61.8)
5	63.3 (60.5)	60.2 (58.4)	61.0 (56.4)

flected in the spectral fine structure, is known to be affected by physical stress [5] [6], CR features can potentially improve the performance also in this detection task.

Table 3: *Effect of applying an all-pole model as the short-time spectrum analysis (cross validation over Training set).*

Short-time spectrum	UAR-M (UAR-E), %
FFT	67.3 (65.5)
LP	64.8 (62.3)
LP-CR	67.2 (66.5)

3.2. Classification Results

After initial parameter optimization by the cross-validated Training set, the best detection system configurations are chosen using the Development set while using the Training set for training. Table 4 shows predictive, multiple-time-scale modulation filtering applied to three types of short-time features, each based on a different spectrum analysis method.

Table 4: *UAR-M (%; UAR-E in parentheses) for Development set showing the effect of multi-scale filtering on MFCC features based on FFT, XLP ($p = 20$, $m = 50$) and XLP-CR ($m = 40$).*

Filter	FFT	XLP	XLP-CR
none	63.7 (60.4)	68.5 (68.0)	65.2 (62.0)
$S = 1$, $p \in \{4, 8, 12\}$	61.6 (59.9)	60.8 (58.3)	67.4 (62.5)
$S = 2$, $q \in \{4, 8, 12\}$	64.0 (60.4)	65.8 (64.8)	67.2 (65.9)
$S = 3$, $q \in \{4, 8, 12\}$	59.8 (56.5)	67.3 (64.8)	67.9 (64.3)
$q = 4$, $S \in \{1, 2, 3\}$	62.9 (57.8)	64.0 (62.0)	64.8 (60.9)
$q = 8$, $S \in \{1, 2, 3\}$	61.4 (58.3)	64.5 (62.5)	65.0 (64.1)
$q = 12$, $S \in \{1, 2, 3\}$	63.1 (59.4)	66.9 (66.4)	64.4 (63.6)

Based mainly on Table 4, a modulation filter structure is chosen for each spectrum analysis method. The filter is applied in three modes: predictive filtering as in [8] [9], classification subspace selection based on the prediction residual (Section 2.3.1) and both the filtering and the subspace selection combined. These results are shown in Table 5. The most promising configurations are evaluated also on the Test set with limited classification trials. Table 6 shows the results obtained for the Development and Test sets together with the baselines.

Table 5: *Comparison of long-term processing for three spectrum estimation methods on the Development set. For each spectrum analysis method, the MFCCs are modeled with a multi-scale autoregressive filter chosen according to Table 4. The filter is applied in three different ways: for predictive filtering, for subspace selection or for combined filtering and subspace selection. The Challenge baseline UAR is 67.2 %.*

Spectrum analysis and long-term filter	Long-term operations	UAR-M (UAR-E), %
FFT $S = 2, q \in \{4, 8, 12\}$	filtering	64.0 (60.4)
	subspace	63.4 (59.4)
	filt.&subsp.	63.2 (59.9)
XLP $S = 2, q \in \{4, 8, 12\}$	filtering	65.8 (64.8)
	subspace	69.7 (68.0)
	filt.&subsp.	65.7 (64.8)
XLP-CR $S = 3$,$q \in \{4, 8, 12\}$	filtering	67.9 (64.3)
	subspace	64.9 (62.5)
	filt.&subsp.	68.2 (64.3)

Table 6: *Main results: UAR (%) for the Development (Dev) and Test sets using the best methods found in this study. The scores that exceed the ComParE baseline [1] are marked in boldface.*

Method	Dev	Test
FFT $S = 2$,$q \in \{4, 8, 12\}$ filtering	64.0	69.9
FFT $S = 2$,$q \in \{4, 8, 12\}$ filtering & subspace	63.2	69.9
XLP $S = 2$,$q \in \{4, 8, 12\}$ subspace	**69.7**	68.6
XLP-CR $S = 3$,$q \in \{4, 8, 12\}$ filtering	**67.9**	68.6
FFT-MFCC/GMM baseline	63.7	n/a
SVM (official) baseline [1]	67.2	71.9

4. Conclusions

Automatic detection of physical stress in speech was studied. In initial short-time feature extraction, conventional FFT spectrum analysis was substituted with a form of extended weighted linear prediction (XLP) which produces smooth spectra by focusing on the most salient spectral cues. Short-time MFCC features obtained by both FFT and XLP were processed with a multi-scale long-term autoregressive filter in order to model the temporal characteristics of the target speech class. The short-time XLP approach, which was recently proposed as a robust method to focus on relevant spectral cues in emotion detection [12], and the multi-scale autoregressive approach, which has also recently been successfully applied to robust emotion detection [8] [9], were combined in order to investigate how well their performance in the detection of mental states will carry over to physical states. The results show that XLP spectrum analysis is effective in distinguishing physical stress from speech and gets further benefit from long-term modeling by applying the modeling residual to dynamic GMM classification subspace selection, a new method proposed in this study. In agreement with previous studies, Fourier-based features were enhanced by predictive filtering/smoothing with the same long-term model. The proposed configurations outperformed the basic timbral classifier and also reached the performance level of the Challenge baselines. The methods studied thus hold potential for future use in similar applications.

5. References

[1] Schuller, B., Steidl, S., Batliner, A., Epps., J., Eyben, F., Ringeval, F., Marchi, E. and Zhang, Y., “The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load”, in Proc. Interspeech, Singapore, Singapore, September 14–18 2014.

[2] Wu, S., Falk, T. H. and Chan, W.-Y., “Automatic speech emotion recognition using modulation spectral features”, Speech Communication 53(5):768–785, May-June 2011.

[3] Arias, J. P., Busso, C. and Yoma, N. B., “Shape-based modeling of the fundamental frequency contour for emotion detection in speech”, Computer Speech and Language 28(1):278–294, January 2014.

[4] El Ayadi, M. M. H., Kamel, M. S. and Karray, F., “Speech emotion recognition using Gaussian mixture vector autoregressive models”, in Proc. ICASSP, Honolulu, Hawaii, April 15–20 2007.

[5] Hansen, J. H. L. and Patil, S. A., “Speech under stress: analysis, modeling and recognition”, in Speaker Classification I, Lecture Notes in Computer Science 4343:108–137, 2007.

[6] Godin, K. W. and Hansen, J. H. L., “Analysis and perception of speech under physical task stress”, in Proc. Interspeech, Brisbane, Australia, September 22–26 2008.

[7] Patil, S. A. and Hansen, J. H. L., “Detection of speech under physical stress: model development, sensor selection, and feature fusion”, in Proc. Interspeech, Brisbane, Australia, September 22–26 2008.

[8] Pohjalainen, J. and Alku, P., “Automatic detection of anger in telephone speech with robust autoregressive modulation filtering”, in Proc. ICASSP, Vancouver, Canada, May 26–31 2013.

[9] Pohjalainen, J. and Alku, P., “Multi-scale modulation filtering in automatic detection of emotions in telephone speech”, in Proc. ICASSP, Florence, Italy, May 4–9 2014.

[10] Pohjalainen, J., Saeidi, R., Kinnunen, T. and Alku, P., “Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions”, in Proc. Interspeech, Makuhari, Japan, September 26–30 2010.

[11] Keronen, S., Pohjalainen, J., Alku, P. and Kurimo, M., “Noise robust feature extraction based on extended weighted linear prediction in LVCSR”, in Proc. Interspeech, Florence, Italy, August 27–31 2011.

[12] Pohjalainen, J. and Alku, P., “Extended weighted linear prediction using the autocorrelation snapshot - a robust speech analysis method and its application to recognition of vocal emotions”, in Proc. Interspeech, Lyon, France, August 25–29 2013.

[13] Pohjalainen, J., Raitio, T., Yrttiaho, S. and Alku, P., “Detection of shouted speech in noise: human and machine”, Journal of the Acoustical Society of America 133(4):2377–2389, April 2013.

[14] Makhoul, J., “Linear prediction: a tutorial review”, Proceedings of the IEEE 63(4):561–580, April 1975.

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

[16] Reynolds, D. A. and Rose, R. C., “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, IEEE Trans. Speech and Audio Proc., 3(1):72–83, January 1995.

[17] Dempster, A. P, Laird, N. M. and Rubin, D. B., “Maximum Likelihood from Incomplete Data via the EM algorithm”, Journal of the Royal Statistical Society. Series B, 39:1–38, 1977.

[18] Bilmes, J. A., “A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models”, University of California at Berkeley, Technical Report, 1998.

[19] Katsavounidis, I., Kuo, C.-C. J. and Zhang, Z., “A new initialization technique for generalized Lloyd iteration”, IEEE Signal Processing Letters 1(10):144–146, October 1994.

[20] Bengio, S. and Mariéthoz, J., “A Statistical Significance Test for Person Authentication”, in Proc. ODYSSEY04, Toledo, Spain, May 31–June 3 2004.

[21] Schuller, B., Friedmann, F. and Eyben, F., “The Munich BioVoice Corpus: effects of physical exercising, heart rate, and skin conductance on human speech production”, in Proc. 9th Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, May 26–31 2014.