Extended Weighted Linear Prediction Using the Autocorrelation Snapshot - A Robust Speech Analysis Method and its Application to Recognition of Vocal Emotions

Jouni Pohjalainen, Paavo Alku

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

Abstract

Temporally weighted linear predictive methods have recently been successfully used for robust feature extraction in speech and speaker recognition. This paper introduces their general formulation, where various efficient temporal weighting functions can be included in the optimization of the all-pole coefficients of a linear predictive model. Temporal weighting is imposed by multiplying elements of instantaneous autocorrelation "snapshot" matrices computed from speech data. With this novel autocorrelation-snapshot formulation of weighted linear prediction, it is demonstrated that different temporal aspects of speech can be emphasized in order to enhance robustness of feature extraction in speech emotion recognition.

Index Terms: linear prediction, spectrum analysis, speech emotion recognition

1. Introduction

Accurate parametrization of the short-time magnitude spectrum is central in speech processing applications. Frame-level feature extraction in automatic recognizers, including automatic speech recognition (ASR), speaker recognition and paralinguistic analysis systems, is often based on representations of the short-time magnitude spectrum such as the widely used mel-frequency cepstral coefficients (MFCCs).

Degradations such as additive noise or a poor acoustic channel affect the quality of the magnitude spectrum model. While this problem affects all speech processing systems, it can be considered to be particularly relevant for a recognition system: even if the feature extraction module manages, despite the noise, to capture the characteristic information that would otherwise be sufficient for class discrimination, there will still exist a *mismatch* between the training and recognition conditions. Other sources of mismatch are related to speakers, e.g., with respect to variable vocal effort or fundamental frequency.

While additive noise corruption and channel mismatch can be tackled in various stages of the recognition process, including speech enhancement preprocessing (e.g., spectral subtraction, Wiener filtering [1]), feature postprocessing (e.g., cepstral mean normalization [1]) and the recognition models (e.g., discriminative training [2]), a combination of mutually complementary techniques in different stages is often employed. For instance, feature extraction based on robust spectrum models can be successfully combined with speech enhancement preprocessing [3]. Therefore, and also because of being important in other speech processing applications, robust spectrum analysis is an important area of study. However, because the range of potential degradations and mismatch conditions affecting the spectra is wide and varied, it is not sufficient to study it from the perspective of resistance against simple corruptions such as additive white noise. Instead, it would be very beneficial to have a more generic spectrum analysis framework that can be adjusted according to the degradations encountered.

Earlier studies have noted that MFCCs based on linear predictive all-pole models tend to be more robust in the presence of additive noise corruption than conventional MFCCs using the Fourier spectrum analysis [4]-[7]. In recent years, members of the *weighted linear prediction* family of spectrum analysis methods have been studied by the present authors in terms of their noise and channel robustness as part of the MFCC feature extraction procedure in automatic speech recognition [4] [5] and speaker verification [3] [6]. The weighted linear predictive spectrum estimate has also recently been shown to have potential in formant estimation of high-pitched speech [8].

The original *temporally weighted* linear prediction [9] was *stabilized* [10] and then generalized to lag-weighted linear prediction [6], thereby expanding the applicability of the concept. In the present study, thus far the most general formulation of weighted linear prediction is introduced, incorporating all the previously discussed methods and also allowing for many new, potentially useful weighting schemes. Its main properties and the effects of weighting applied to instantaneous *autocorrelation snapshot* matrices are discussed. Weighting schemes for the method are, in this study, derived from additive-noise robustness considerations. The noise robustness of the method is evaluated and compared to conventional methods in a realistic application, the automatic recognition of emotion in speech.

2. Extended weighted linear prediction

2.1. Snapshot Formulation of Weighted Linear Prediction

In linear prediction (LP), the short-time magnitude spectrum of a signal is parametrized as an all-pole filter with z-domain transfer function $H(z) = 1/(1 - \sum_{k=1}^{p} a_k z^{-k}) = 1/A(z)$ [11]. In the time domain, each sample is assumed to be predictable as $\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k}$, i.e., a linear combination of the coefficients a_k, $1 \leq k \leq p$, and p past signal samples, where p is referred to as the prediction order.

In conventional LP, the model parameters are obtained by minimizing the sum of squares of the prediction error $E_{\mathrm{LP}} = \sum_n (s_n - \sum_{k=1}^{p} a_k^{\mathrm{LP}} s_{n-k})^2$ by setting its partial derivatives with respect to each coefficient a_k^{LP} to zero. This results in the LP normal equations [11] $\sum_{k=1}^{p} a_k^{\mathrm{LP}} \sum_n s_{n-k} s_{n-j} = \sum_n s_n s_{n-j}$, $1 \leq j \leq p$, whose solution yields the LP model $\{a_j^{\mathrm{LP}}\}$. The range of summation of n is chosen in this work to correspond to the autocorrelation method, in which the energy is minimized over a theoretically infinite interval, but s_n is

considered to be zero outside the actual analysis window [11].

The underlying motivation of weighted linear predictive methods is temporal emphasis of those parts within the short-time signal frame which are deemed, using some heuristic analysis function, to be the most reliable and least likely to have been severely corrupted. Weighted linear prediction (WLP), proposed by Ma et al. [9], is a generalization of LP in which a time-dependent weighting function is applied to the squared prediction error values so as to emphasize the correct prediction of selected samples and down-weight others. In WLP, the quantity to be minimized is $E_{\mathrm{WLP}} = \sum_n (s_n - \sum_{k=1}^{p} a_k^{\mathrm{WLP}} s_{n-k})^2 W_n$ and the normal equations thus become $\sum_{k=1}^{p} a_k^{\mathrm{WLP}} \sum_n W_n s_{n-k} s_{n-j} = \sum_n W_n s_n s_{n-j}$, $1 \leq j \leq p$. LP is obtained as a special case when $W_n = C$, a constant that cancels out from both sides of the equations.

A further generalization of both LP and WLP, termed extended weighted linear prediction (XLP), was recently proposed [6]. In this formulation, the quantity to be minimized is $E_{\mathrm{XLP}} = \sum_n (s_n Z_{n,0} - \sum_{k=1}^{p} a_k^{XLP} s_{n-k} Z_{n,k})^2$, where the partial weights $Z_{n,j}$ separately weight each lagged signal value at each time instant n, allowing more control over the focus of the all-pole modeling. The resulting normal equations are

$$\sum_{k=1}^{p} a_k^{XLP} \sum_n Z_{n,k} s_{n-k} Z_{n,j} s_{n-j} = \sum_n Z_{n,0} s_n Z_{n,j} s_{n-j}, \quad (1)$$
$$1 \leq j \leq p.$$

WLP is obtained as a special case when $Z_{n,j} = \sqrt{W_n}$ and, similarly to the WLP case, LP is obtained when $Z_{n,j} = C$.

This study proposes a yet more generic formulation of the XLP method which encompasses each of the above mentioned methods for estimating LP all-pole models as special cases but facilitates more versatile weighting schemes. The general normal equations of extended weighted linear prediction can be written as

$$\sum_{k=1}^{p} a_k \sum_n Q_{n,j,k} s_{n-k} s_{n-j} = \sum_n Q_{n,j,0} s_n s_{n-j}, \quad (2)$$
$$1 \leq j \leq p,$$

where $Q_{n,j,k}$ is the weighting function. It can be seen that WLP is obtained when $Q_{n,j,k} = W_n$ and LP is obtained when $Q_{n,j,k} = C$. The version of XLP originally formulated by Pohjalainen et al. [6] is obtained when

$$Q_{n,j,k} = Z_{n,j} Z_{n,k}, \quad (3)$$

i.e., when the weighting function can be factorized as a product of lag-specific partial weights. However, an interesting additional property becomes evident by examining Eq. 2 in matrix form: defining $\boldsymbol{Q}_n = (Q_{n,j,k})$, $\boldsymbol{a} = (a_1 \dots a_p)^T$ and $\boldsymbol{s}_n = (s_{n-1} \dots s_{n-p})^T$, the normal equations can be expressed as

$$\left(\sum_n \boldsymbol{Q}_n \odot \left(\boldsymbol{s}_n \boldsymbol{s}_n^T \right) \right) \boldsymbol{a} = \sum_n Q_{n,j,0} s_n \boldsymbol{s}_n, \quad (4)$$

where $\odot$ denotes element-by-element multiplication. The $(p \times p)$ matrix $\boldsymbol{s}_n \boldsymbol{s}_n^T$ is referred to as the *autocorrelation snapshot matrix* at time n. Substituting $Q_{n,j,k} = C$ and $\boldsymbol{Q}_n = (C)$, the summation on the left-hand side of Eq. 4 yields the (scaled) autocorrelation matrix $\boldsymbol{R} = C \sum_n (\boldsymbol{s}_n \boldsymbol{s}_n^T)$. Similarly, the summation of the *autocorrelation snapshot vectors* $s_n \boldsymbol{s}_n$ on the right-hand side of the equation yields the (scaled) autocorrelation vector $\boldsymbol{r} = C \sum_n s_n \boldsymbol{s}_n$, making Eq. 4 equivalent to the matrix formulations of conventional LP [12]. In computing the autocorrelations in WLP case, i.e., when $Q_{n,j,k} = W_n$ and $\boldsymbol{Q}_n = (W_n)$, the snapshots are weighted differently from each other according to the time instants they are observed at. As a novel idea, this paper proposes weighting the snapshots with an *unconstrained* weighting function $Q_{n,j,k}$, based on the mutual dependencies of the data constituting each snapshot, and thus removing the original XLP constraint given by Eq. 3. The proposed snapshot approach to specifying the weights, which further extends the definition of XLP and is also referred to by that name in the present paper, is able to completely consider the instantaneous mutual dependencies between lagged signal samples involved in each snapshot and should thus allows new, potent data weighting schemes.

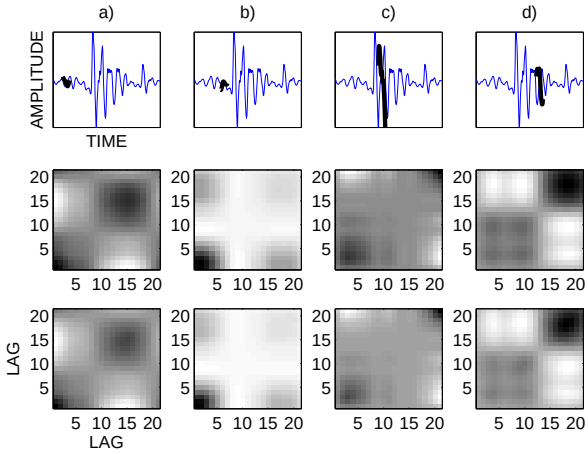

Figure 1: *Short segments of speech, indicated with thick line (top row), and their corresponding unweighted autocorrelation snapshot matrices (middle row; black corresponds to the largest positive value) and the same matrices weighted according to the XLP-S2 scheme (bottom row), with columns a)-d) corresponding to different locations of the snapshot.*

Figure 1 illustrates different snapshot matrices $\boldsymbol{s}_n \boldsymbol{s}_n^T$, as well as their weighted versions, related to different parts of a short-time speech analysis frame.

2.2. Weighting Schemes

The weighting function typically used in WLP is the short-time energy (STE) $W_n = \sum_{i=1}^{p} s_{n-i}^2$, i.e., the energy of past p signal samples. This weighting scheme has good theoretical motivations, as emphasizing the energetic segments focuses on both the glottal closed phase in voiced speech, where the vocal tract resonances are most evident [10], and, in the case of stationary background noise, on different parts of the analysis frame in relation to their local signal-to-noise ratio (SNR). WLP with STE weighting is thus theoretically appropriate for additive-noise robustness and has, indeed, been found to improve the robustness in both large-vocabulary continuous speech recognition [4] and text-independent speaker verification [3]. Incidentally, the STE weighting function is the Frobenius norm, as well as the trace, of the autocorrelation snapshot matrix appearing on the left-

hand side of Eq. 4.

With the original partial-weight formulation of XLP, satisfying Eq. 3, the absolute value sum (AVS) weighting, which can be viewed as an adaptation of the STE scheme to partial weights, has led to improved noise and channel robustness in both speaker verification [6] and speech recognition [5]. This weighting is given by the recursion

$$Z_{n,j} = \frac{p-1}{p} Z_{n-1,j} + \frac{1}{p}(|s_n| + |s_{n-j}|) \quad (5)$$

with $Z_{n,j} = 0$ for all j before the beginning of the frame. In the experimental part of this paper, this method is referred to as XLP-P to denote the property that the weighting is obtained as a product of partial weights according to Eq. 3.

It is not immediately clear how to extend the idea of the previously successful AVS weighting scheme (Eq. 5) to snapshot XLP. Therefore, we evaluate two different options which involve recursive low-pass filtering of sums and products of absolute values of lagged samples:

$$Q_{n,j,k} = \frac{p-1}{p} Q_{n-1,j,k} + \frac{1}{p}\left(|s_n| + |s_{n-j}| + |s_{n-k}|\right), \quad (6)$$

which will be referred to as XLP-S1 in this paper, and

$$Q_{n,j,k} = \frac{p-1}{p} Q_{n-1,j,k} + \frac{1}{p}\left(s_n^2 + |s_{n-j}||s_{n-k}|\right), \quad (7)$$

referred to as XLP-S2.

2.3. Spectral smoothing

All-pole filter stability is especially important in applications where they will be used to synthesize signals. However, the stabilized versions of WLP and XLP-P, which typically produce smoother spectra than the original weighted linear predictive methods they are based on, have shown improved robustness also in feature extraction, especially for speaker verification [3] [6].

As originally proven by Magi et al. [10], the all-pole filter produced by WLP becomes stable if, according to the partial-weight XLP notation in Eq. 1 with $Z_{n,j} = \sqrt{W_n}$ [4], the condition

$$Z_{n,j} \leq Z_{n-1,j-1} \quad (8)$$

is satisfied for all n and j. The same condition was subsequently adopted for stabilizing general partial-weight XLP (Eq. 1) using arbitrary $Z_{n,j}$. Model stabilization is thereby performed by replacing these weights with $Z'_{n,j} = \max(Z_{n,j}, Z_{n-1,j-1})$ with $Z_{n,j} = 0$ for $j < 0$ [6]. With snapshot XLP, assuming that the weight matrix $\boldsymbol{Q}_n = (Q_{n,j,k})$ can be factorized as a product of two partial-weight vectors according to Eq. 3, Eq. 8 yields

$$Q_{n,j,k} \leq Q_{n-1,j-1,k-1}. \quad (9)$$

While this condition for stability apparently does not generalize to weights that can not be factorized according to Eq. 3, it has been found that enforcing it nevertheless typically leads to generally smoother spectral models as well as a reduced number of unstable filters. Therefore, we also propose an optional spectral smoothing operation for snapshot XLP which is performed by replacing the original weights $Q_{n,j,k}$ with $Q'_{n,j,k} = \max(Q_{n,j,k}, Q_{n-1,j-1,k-1})$, where $Q_{n,j,k} = 0$ for $j < 0$ or $k < 0$. In our evaluations, we apply the smoothing operation to XLP-S1 because it was observed to improve the quality of the spectral estimate and the classification performance. The operation is not applied to XLP-S2, which already results in smooth spectra, or XLP-P, which has been observed to produce few unstable filters.

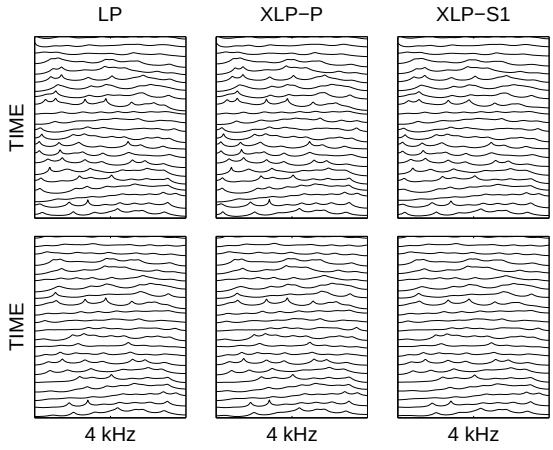

Figure 2: *Example spectra of LP, XLP-P and XLP-S1 over one utterance of the anger emotion category. Upper panels: clean speech. Lower panels: the same utterance with noise corruption by factory noise at SNR 0 dB.*

Figure 2 shows examples of LP and XLP spectra in clean and noisy conditions. XLP-S1 arguably shows less degradation between the clean and noisy case than LP. While XLP-S1 is smoother than LP or XLP-P in the clean case, it does not show a noticeable loss in modeling of the spectral detail under either noise condition.

3. Experiments

3.1. Test material

Speech emotion recognition evaluation was conducted using the Berlin database of emotional speech as evaluation material [13]. The database has been widely adopted in emotion classification studies, also including earlier studies on robustness, e.g., [14] [15] [16]. In total, it consists of 535 utterances of German sentences spoken in seven different emotional styles by five male and five female actors. The emotion categories contained in this database are *anger*, *disgust*, *fear*, *joy*, *sadness*, *boredom* and *neutral*. The whole database was used for the evaluation, which was carried out in a speaker-independent manner as leave-one-speaker-out cross validation [14] [17], i.e., one speaker at at a time was chosen as the test speaker and the material from the other nine was used for training. Unfortunately, a vast array of different ways of using the Berlin database for evaluation can be encountered in the literature: sometimes only subsets of the data are used, for example, by leaving out one of the emotion categories; cross validation is sometimes performed without regard to speaker identity and sometimes with it; different evaluation metrics are used including the total recognition rate, cross-validation-averaged recognition rate, etc. Therefore, a direct large-scale comparison of the results with other approaches in the literature will not be possible in the context of the present study.

In order to study the noise performance, the speech material

was analyzed in the clean form and also artificially corrupted by two types of noise from the NOISEX-92 database with three segmental (frame-averaged) signal-to-noise ratios (SNRs): 20 dB, 0 dB and -20 dB, resulting in seven different noise conditions. The noise types used were *factory1*, which is mechanical noise recorded inside a factory and *babble*, which consists of the simultaneous speech of many speakers. The clean case and each of the six noise cases were used for test material, while classifier training was always conducted using clean speech data only.

3.2. Speech emotion recognition

3.2.1. Overview

In order to evaluate the performance of different features in a realistic application, an emotion recognition system was constructed. The MFCC-based feature extraction and the GMM-based classification are briefly described.

3.2.2. Feature extraction

After pre-emphasis with $H_p(z) = 1-0.97z^{-1}$, the signal is arranged into overlapping Hamming-windowed frames of 25 ms with a shift interval of 10 ms. For each frame, mel-frequency cepstral coefficients (MFCCs) are obtained using the processing chain: 1) compute squared magnitude spectrum by either FFT (the conventional method), LP, XLP-P, XLP-S1 or XLP-S2, 2) apply a mel filterbank to the squared magnitude spectrum, 3) take the logarithm of filtered band energies and 4) perform discrete cosine transform [1]. The mel filterbank has 40 triangular filters with center frequencies spaced evenly on the mel scale. The zeroth MFCC is excluded and the subsequent 12 MFCCs are complemented with logarithmic frame energy which has been mean- and variance-normalized over the complete utterance. Finally, delta and double-delta coefficients of the MFCCs and log energy are appended, resulting in a 39-dimensional feature vector.

3.2.3. Classification

For the purpose of classification according to the Bayes rule [18], a Gaussian mixture model (GMM) with 64 components and a diagonal covariance structure is trained separately for each of the seven emotion classes in the Berlin database using 10 iterations of the expectation-maximization (EM) algorithm for GMM training [19]. Before training each GMM, the mean vectors of the components are initialized by applying a selection approach intended for the initialization of the cluster means in EM-style cluster-seeking algorithms [20]. This procedure is applied to the 12 base MFCCs while the other 27 elements of the feature vector are initialized by averaging within the initial clusters. The component weights are initialized with uniform distributions and the variances with 0.1 times the global variances. In the classification phase, the class decision for each utterance is determined by the identity of the GMM giving the highest average likelihood, averaged over each frame of the utterance.

3.3. Results

Tables 1 and 2 show the classification accuracy for the utterances in the Berlin database. For the clean condition, it can be observed that while FFT shows the best performance (which is comparable to other published approaches using leave-one-speaker-out cross validation on the Berlin database [14] [17]), the proposed XLP methods are already competitive with conventional LP. As noise is increased, the robustness of the XLP methods becomes evident. In the case of factory noise and SNR 0 dB, the performance advantage of newly formulated XLP methods over the conventional FFT is particularly clear.

Table 1: Emotion recognition performance (correct %) under factory noise with MFCC features obtained using five spectrum analysis methods. The best score for each classifier and noise level is shown in boldface. The results pertain to utterances of seven emotion categories in the Berlin database and were obtained using leave-one-speaker-out cross validation.

Signal-to-noise ratio (dB)	Spectrum analysis method				
	FFT	LP	XLP-P	XLP-S1	XLP-S2
clean	**74.0**	70.3	71.8	71.2	71.2
20	55.9	54.0	54.4	**57.0**	54.6
0	27.5	35.3	37.6	39.4	**43.4**
-20	16.5	16.3	15.9	17.8	**20.0**

Table 2: Emotion recognition performance (correct %) under babble noise. See caption of Table 1 for additional details.

Signal-to-noise ratio (dB)	Spectrum analysis method				
	FFT	LP	XLP-P	XLP-S1	XLP-S2
clean	**74.0**	70.5	71.8	71.2	71.2
20	**64.1**	60.4	61.9	61.7	62.4
0	38.9	39.8	39.4	41.3	**42.1**
-20	16.5	17.9	16.3	16.8	**18.7**

4. Conclusions

In this study, a new and thus far the most generic formulation of (temporally) weighted linear predictive methods was presented. This framework generalizes both original weighted linear prediction (WLP) [9], which only allows one-dimensional, temporal weighting functions, and the previously proposed version of extended weighted linear prediction (XLP) [6]. At each time instant, the new model allows for unconstrained element-by-element weighting of the autocorrelation snapshot matrix and vector, new concepts introduced in this study. Two new XLP methods, facilitated by the new formulation and with weighting schemes designed to improve robustness against degradations like additive noise, were evaluated as the basis of the MFCC feature representation in the recognition of vocal emotions using the Berlin database of emotional speech. The new speech spectrum analysis methods outperformed the standard FFT and also linear prediction (LP), which itself is known to frequently have a robustness advantage over FFT. Thus, the methods show promise to be used as a basis for different types of robust features in speech emotion recognition. Code can be found at [21].

For the future, the newly proposed formulation of temporally weighted linear prediction allows for many more information weighting strategies in speech spectrum analysis than its predecessors. These weighting schemes can be designed according to different considerations in various applications and targeted on different aspects of the speech signal.

5. Acknowledgements

This work was supported by Academy of Finland (256961) and the EC FP7 project Simple4All (287678).

6. References

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

[2] Pylkkönen, J. and Kurimo, M., “Improving discriminative training for robust acoustic models in large vocabulary continuous speech recognition”, in Proc. Interspeech, Portland, Oregon, USA, September 9–13 2012.

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

[4] Pohjalainen, J., Kallasjoki, H., Palomäki, K. J., Kurimo, M. and Alku, P., “Weighted linear prediction for speech analysis in noisy conditions”, in Proc. Interspeech, Brighton, UK, September 6–10 2009.

[5] 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.

[6] 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.

[7] de Wet, F., Cranen, B., de Veth, J. and Boves, L., “Comparing acoustic features for robust ASR in fixed and cellular network applications”, in Proc. ICASSP, Istanbul, Turkey, June 5–9 2000.

[8] Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M. and Story, B., “Improved formant frequency estimation from high-pitched vowels by downgrading the contribution of the glottal source with weighted linear prediction”, in Proc. Interspeech, Portland, Oregon, USA, September 9–13 2012.

[9] Ma, C., Kamp, Y. and Willems, L. F., “Robust signal selection for linear prediction analysis of voiced speech”, Speech Communication 12(2):69–81, 1993.

[10] Magi, C., Pohjalainen, J., Bäckström, T. and Alku, P.,“Stabilised weighted linear prediction”,Speech Communication 51(5):401–411, 2009.

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

[12] Rabiner, L. R. and Schafer, R. W., “Digital Processing of Speech Signals”, Prentice-Hall, 1978.

[13] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. and Weiss, B., “A database of German emotional speech”, in Proc. Interspeech, Lisbon, Portugal, September 4–8 2005.

[14] Schuller, B., Wimmer, M., Arsic, D., Moosmayr, T. and Rigoll, G., “Detection of security related affect and behaviour in passenger transport”, in Proc. Interspeech, Brisbane, Australia, September 22–26 2008.

[15] Tawari, A. and Trivedi, M., “Speech emotion analysis in noisy real-world environment”, in Proc. Int. Conf. on Pattern Recognition, Istanbul, Turkey, August 23–26 2010.

[16] Chi, T.-S., Yeh, L.-Y. and Hsu, C.-C., “Robust emotion recognition by spectro-temporal modulation statistic features”, J. Ambient Intell. Human Comput. 3:47–60, 2012.

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

[18] Theodoridis, S. and Koutroumbas, K., “Pattern Recognition”, 2nd ed., Academic Press, 2003.

[19] Xu, L. and Jordan, M. I., “On convergence properties of the EM algorithm for Gaussian mixtures”, Neural Computation, 8(1):129–151, 1996.

[20] 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, 1994.

[21] Matlab implementations of different time-weighted linear predictive methods:
http://www.acoustics.hut.fi/∼jpohjala/xlp/, May 29, 2013.