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

Jouni Pohjalainen and Paavo Alku

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

INTERSPEECH 2014

Aalto University
School of Science
and Technology

The focus of this paper

- Recently developed robust feature generation methods:
  - Extended weighted linear prediction (spectrum analysis)
  - Multi-scale autoregressive modeling
    - modulation filtering
    - instantaneous selection of classification subspaces
- Apply them to the present paralinguistic task (Physical Load in ComParE 2014)

Weighted linear prediction (WLP)

- A linear predictive all-pole model

  $$H(z)=1/\left(1-\sum_{k=1}^{p} a_k z^{-k}\right)$$

  has time-domain prediction error

  $$e_n=s_n-\sum_{k=1}^{p} a_k s_{n-k}$$

- To solve a_k, WLP minimizes $E_{WLP}=\sum_n W_n e_n^2$ over the analysis frame

- This way, different temporal locations of the analysis frame can be (de-)emphasized

Extended Weighted Linear Prediction (XLP)

- In WLP, $\partial E_{WLP} / \partial a_j = 0, \quad 1 \leqslant j \leqslant p,$ leads to the normal equations

$$\sum_{k=1}^{p} a_k \sum_n W_n s_{n-k} s_{n-j} = \sum_n W_n s_n s_{n-j}$$

- In XLP, each autocorrelation "snapshot" $s_{n-k} s_{n-j}$ is weighted by a separate value:

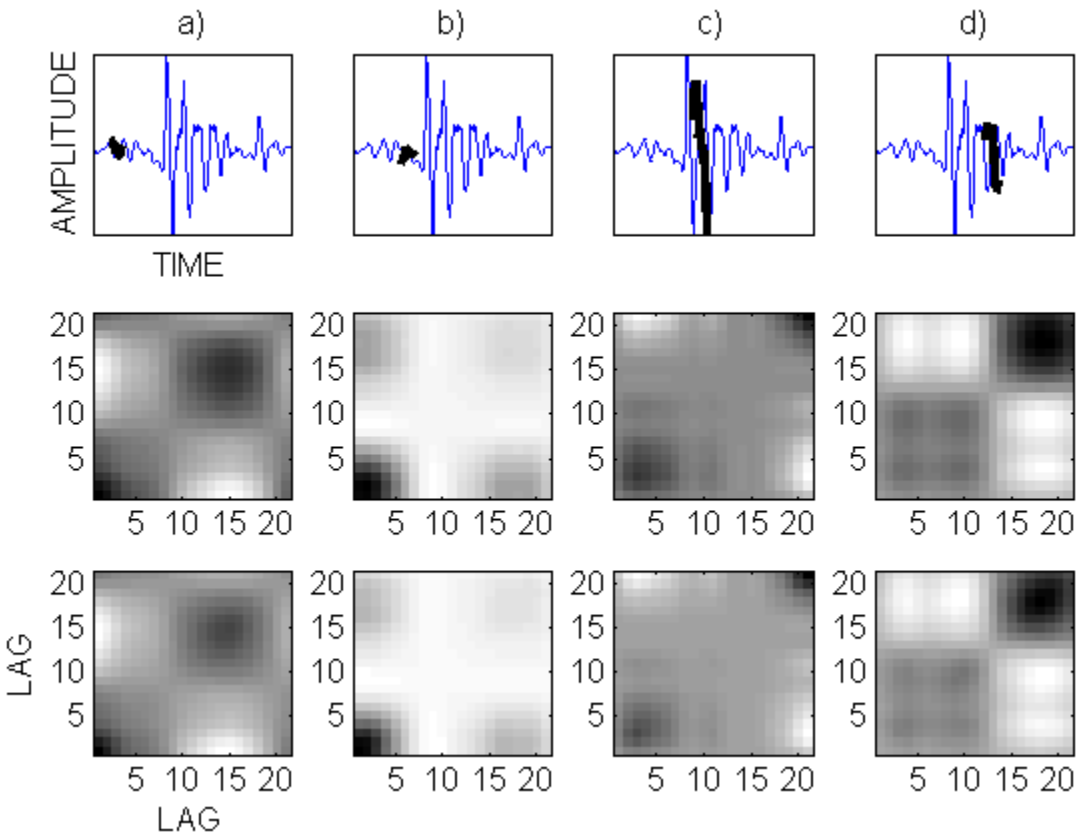
$$\sum_{k=1}^{p} a_k \sum_n Q_{n,j,k} s_{n-k} s_{n-j} = \sum_n Q_{n,j,0} s_n s_{n-j}$$

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

XLP weighting of autocorrelation

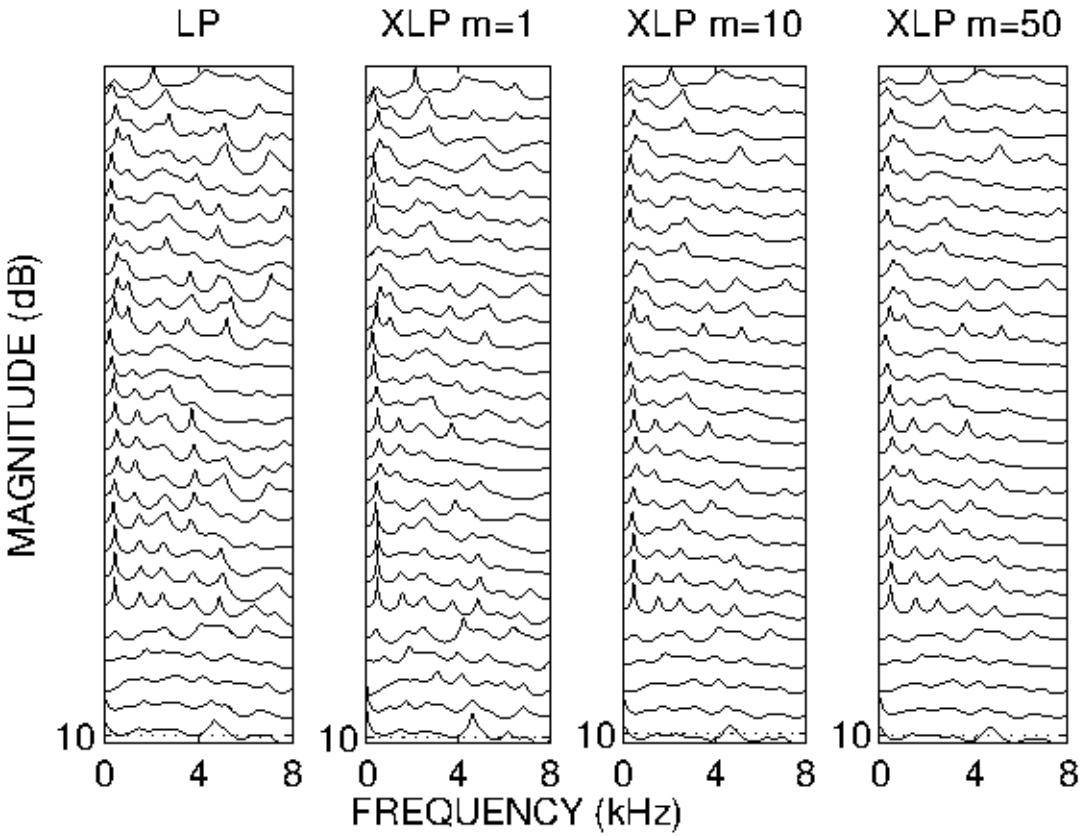
- Weighting example of autocorrelation snapshot matrices $[s_{n-1}\, s_{n-2} \ldots s_{n-p}]^T [s_{n-1}\, s_{n-2} \ldots s_{n-p}]$

  - Unweighted snapshot
  - Weighted snapshot



XLP spectrum analysis example

- Linear prediction (LP) and XLP with different values of the time-averaging parameter m



Multi-scale autoregressive models

- Autoregressive (AR) modeling of the temporal dynamics of the feature trajectories
- A single AR model with order q and sample interval S for the ith feature:

$$\hat{y}_{i,t} = b_{i,0} + \sum_{k=1}^{q} b_{i,k} y_{i,t-kS}$$

  - The time resolution is constrained by choices of q and S

- N AR models with different time scales:

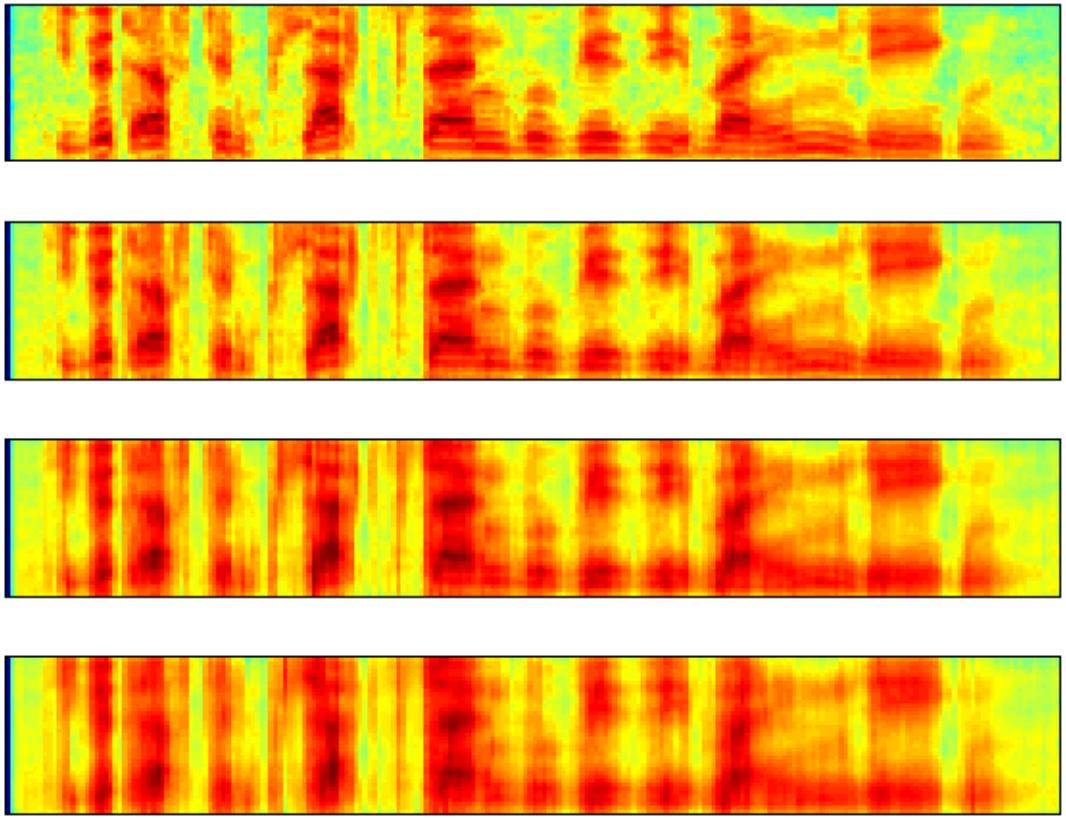
$$\hat{y}_{i,t,j} = b_{0,i,j} + \sum_{k=1}^{q_{i,j}} b_{k,i,j} y_{i,t-kS_{i,j}}, 1 \leqslant j \leqslant N$$

Multi-scale autoregressive models

- Applied for two separate tasks
- Modulation filter for feature trajectories
  - *Replace* each $y_{i,t}$ by its *predicted* value $\hat{y}_{i,t}$ that minimizes $(x_{i,t} - \hat{y}_{i,t,j})^2$
  - $b_{k,i,j}$ have been estimated for target class
- Subspace selection
  - At time t, perform GMM classification using features for which $(y_{i,t} - \hat{y}_{i,t})^2$ is low according to high/low clustering over the utterance

Multi-scale filtering example

- Top: unprocessed auditory spectra (from MFCCs)
- Below: MFCCs filtered by models representing emotions anger, neutral and happiness



The detection system

- 39-dim (12 MFCCs + utterance-normalized log energy) + Δ + ΔΔ
- Spectrum analysis in base MFCC computation is varied (FFT, XLP, XLP-CR)
- Training phase: estimate AR coefficients to represent the target class (low physical load)
- Filtering the feature trajectories across frames
- Classification using 64-component GMMs (optionally, in a subspace consisting of locally selected features)

Physical load detection results

		DEVELOPMENT SET
SPECTRUM ANALYSIS AND LONG-TERM FILTER	LONG-TERM OPERATIONS	UNWEIGHTED AVERAGE RECALL % (MAX/EER)
FFT S=2, q in {4,8,12}	FILTERING SUBSPACE FILT. & SUBSP.	64.0 / 60.4 63.4 / 59.4 63.2 / 59.9
XLP S=2, q in {4,8,12}	FILTERING SUBSPACE FILT. & SUBSP.	65.8 / 64.8 69.7 / 68.0 65.7 / 64.8
XLP-CR S=3, q in {4,8,12}	FILTERING SUBSPACE FILT. & SUBSP.	67.9 / 64.3 64.9 / 62.5 68.2 / 64.3

Physical load detection results

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	
CHALLENGE BASELINE (SVM)	67.2	71.9

Conclusions

- The feature computation methods proposed earlier for robust emotion recognition also showed good robustness performance in recognition of physical stress state
- Spectrum models which include fine detail, such as FFT, could be enhanced by autoregressive long-term filtering
- XLP with the newly proposed classification subspace selection approach gave the best results

What have we learnt?

- Similar approaches can work for different paralinguistic tasks
- Robustness is an important consideration with practical, realistic databases
  - Depending on the task at hand, it may not be realistic to record even training data in a sound-proofed anechoic chamber
- Combinations of different methods may work well together to complement each other
  - But how to find the best configuration?

References and software

- [1] J. Pohjalainen and P. Alku: Extended weighted linear prediction using the autocorrelation snapshot – a robust speech analysis method and its application to recognition of vocal emotions, INTERSPEECH 2013
- [2] J. Pohjalainen and P. Alku: Multi-scale modulation filtering in automatic detection of emotions in telephone speech, ICASSP 2014

- Matlab code: http://www.acoustics.hut.fi/research/robustness