Aalto University
School of Electrical
Engineering

Robust Spectral Representation Using Group Delay Function and Stabilized Weighted Linear Prediction for Additive Noise Degradations

Dhananjaya Gowda, Jouni Pohjalainen, Paavo Alku and Mikko Kurimo

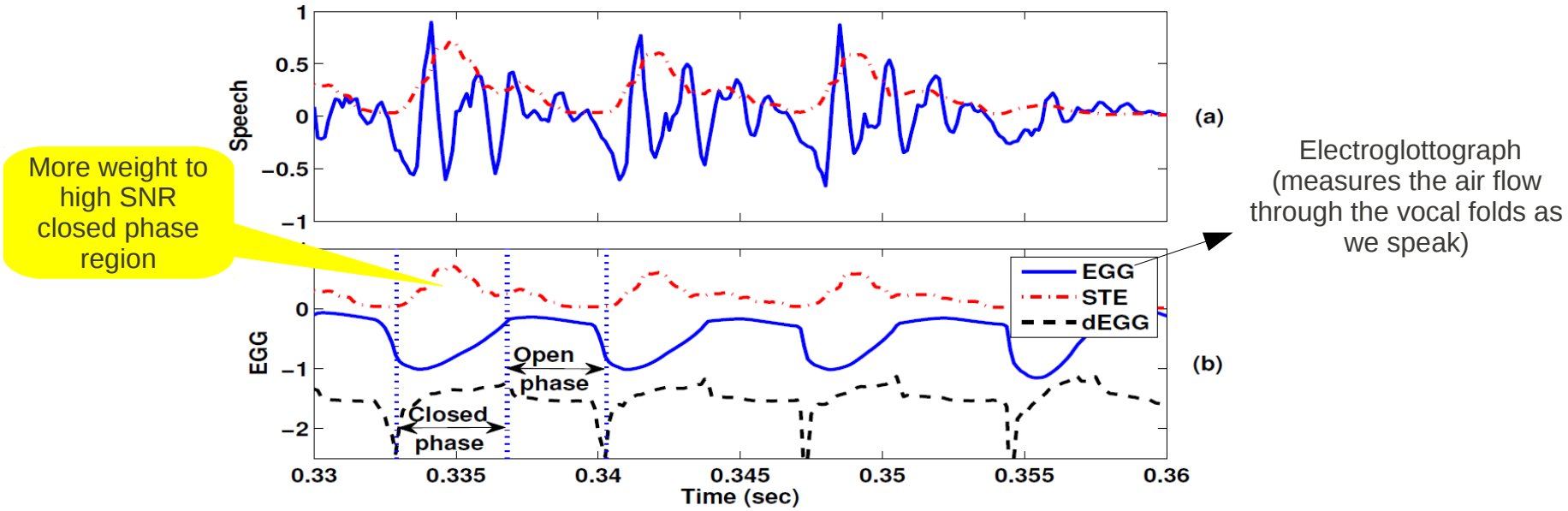
Dept. of Signal Processing and Acoustics
School of Electrical Eng., Aalto University, Finland

Outline

- Weighted linear prediction (WLP)
  - Stabilized weighted linear prediction (SWLP)
- Group delay (GD) of an all-pole model
- SWLP-GD spectrum
- Robustness of SWLP-GD spectrum
- Speaker recognition experiments
- Conclusions

Weighted linear prediction

- Idea: give more importance/weight to reduce prediction errors in the close phase region of the glottal cycle
- Provides better estimates of the vocal tract
- Noise robust as the focus is now on high SNR region
- Short time energy (STE) is one such weight function
  - [*Ma et al., Speech Comm. 1993*]
- Stabilized WLP (SWLP) ensures stability of the estimated poles
  - [*Magi et al., Speech Comm. 2008*]

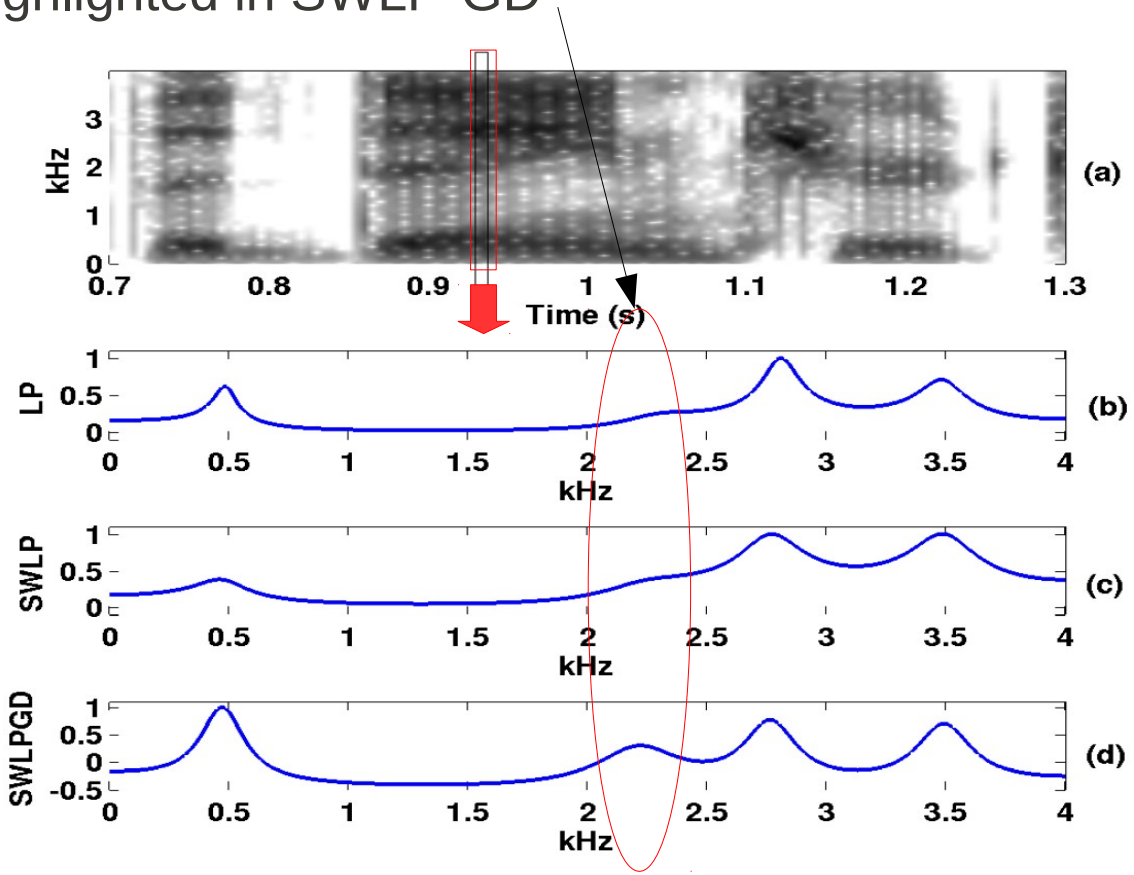


Group delay of an all-pole system

- Group delay (GD) function – negative derivative of phase spectrum
- GD function is additive in nature (w.r.t. individual resonances) as against multiplicative magnitude spectrum
- Formant peaks are better resolved
- Formant peaks are better highlighted even under degradations
- Can be computed from the inverse filter impulse response
- Avoids phase unwrapping

SWLP-GD spectrum

- Computed as the group delay function of the SWLP spectrum
- SWLP tends to smooth the spectrum due to weighting
- SWLP-GD brings back the formant resolution
- Weak formants better highlighted in SWLP-GD



Robustness of SWLP-GD

- Objective measure #1
  - average log spectral distortion (LSD)
  - LSD between normalized spectra from clean and degraded speech
  - Spectra normalized to unit energy
- Data
  - VTR database (192 utterances, 24 speakers, 8 female & 16 male)
  - Degradations from NOISEX database
- Observations
  - STRAIGHT marginally better than LP
  - SWLP better than STRAIGHT
  - SWLP-GD improves upon SWLP and performs the best

$$d_{LS} = \frac{2}{K} \sum_{k=0}^{K/2} |20 \log_{10}(H[k]) - 20 \log_{10}(\hat{H}[k])|$$

SNR (dB)	SPECTRUM TYPE				
	SWLP-GD	SWLP	LP	DFT	STRT
BABBLE					
10	**1.1**	1.7	2.8	3.5	2.9
5	**1.5**	2.5	4.0	4.6	3.9
0	**1.9**	3.3	5.2	5.7	5.0
-5	**2.2**	4.2	6.3	6.7	6.0
FACTORY					
10	**1.4**	2.3	3.8	4.5	3.6
5	**1.7**	3.2	5.2	5.8	4.9
0	**2.0**	4.2	6.5	6.8	6.1
-5	**2.3**	5.0	7.6	7.7	7.1
VOLVO					
10	**0.2**	0.2	0.3	1.1	0.6
5	**0.4**	0.4	0.6	1.5	0.9
0	**0.6**	0.7	1.1	2.1	1.2
-5	**0.9**	1.1	1.8	2.9	1.7
WHITE					
10	**1.7**	3.7	5.6	5.8	5.0
5	**2.0**	4.9	7.0	7.0	6.4
0	**2.3**	6.0	8.3	8.1	7.8
-5	**2.5**	7.0	9.4	8.8	8.8

Robustness of SWLP-GD (contd..)

- Objective measure #2
  - Frequency weighted segmental SNR
  - Gives more weight to spectral peaks as against valleys
  - Correlates well with the industry standard PESQ (a measure of speech quality)
  - SWLP-GD performs better than other representations
  - Most affected: white noise followed by factory noise

Frequency weighted segmental SNR

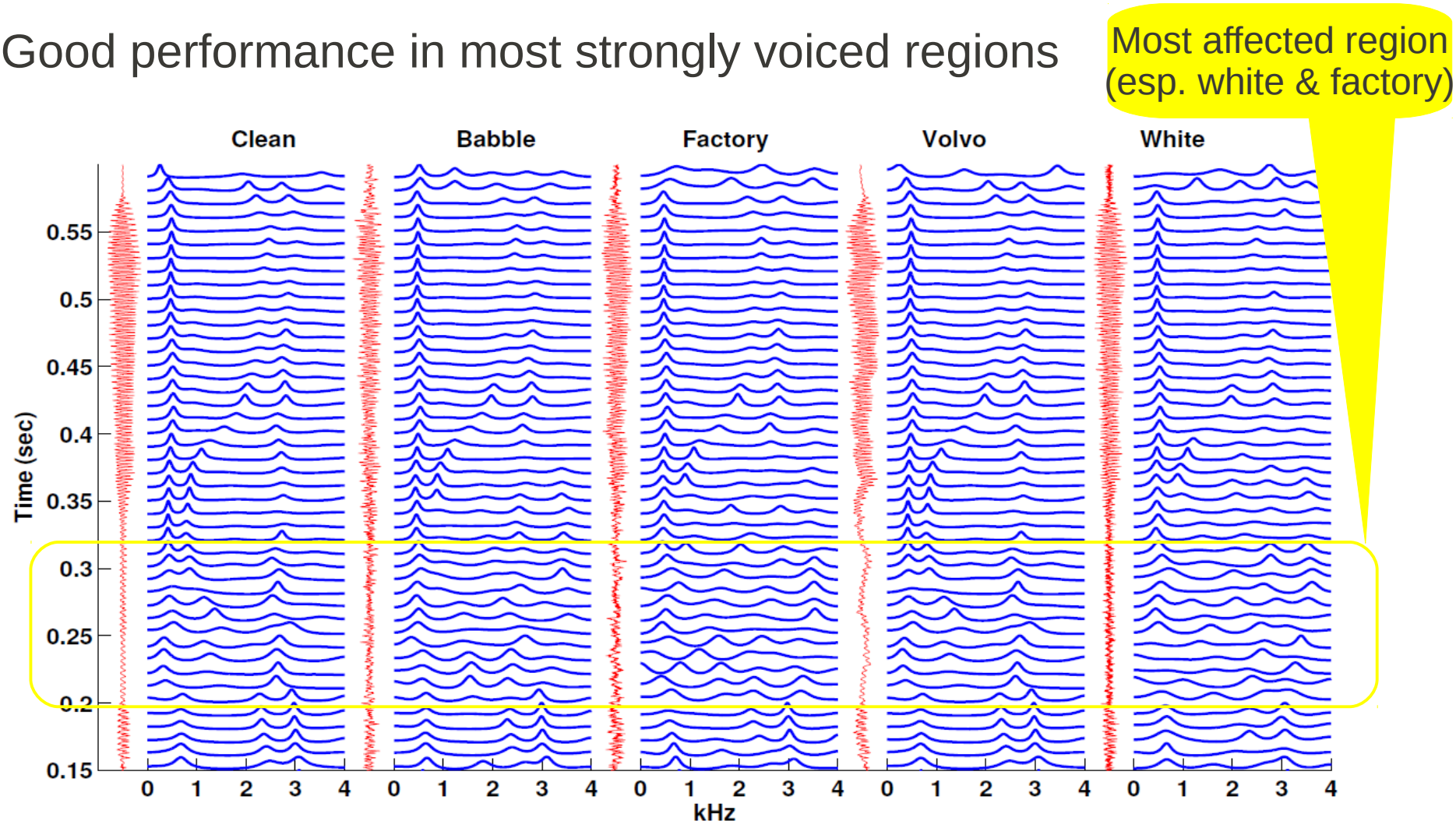
$$S_{fw} = \sum_{k=0}^{K/2} W[k].20\log_{10}\frac{H[k]}{H[k]-\hat{H}[k]}$$

$$W[k] = H^{\gamma}[k] / \sum_{k=0}^{K/2} H^{\gamma}[k]$$

SNR (dB)	SPECTRUM TYPE				
	SWLP-GD	SWLP	LP	DFT	STRT
BABBLE					
10	**24.8**	22.5	18.4	14.8	16.8
5	**21.2**	17.9	13.7	11.2	13.1
0	**18.3**	13.8	9.7	8.3	9.9
-5	**16.1**	10.5	6.6	6.0	7.2
FACTORY					
10	**22.6**	19.6	15.2	12.0	14.5
5	**19.5**	15.0	10.7	8.6	10.7
0	**17.1**	11.3	7.0	6.0	7.5
-5	**15.4**	8.6	4.5	4.2	5.2
VOLVO					
10	**43.2**	42.0	40.6	28.8	34.1
5	**38.0**	37.0	34.4	25.6	29.9
0	**32.5**	31.6	27.8	22.2	25.8
-5	**27.4**	26.4	21.9	18.8	22.0
WHITE					
10	**19.8**	14.4	10.8	9.1	11.3
5	**17.4**	10.2	6.8	6.1	7.4
0	**15.7**	6.9	3.8	3.9	4.5
-5	**14.5**	4.7	1.8	2.5	2.6

Robustness of SWLP-GD (contd..)

- SWLP-GD spectra for different noise degradations at 0 dB SNR
- Good performance in most strongly voiced regions



Speaker recognition experiments

- Small-scale closed-set speaker recognition experiments
  - Matched and mismatched conditions
- Data - VTR database
  - 24 speakers; 8 female, 16 male
  - Train: 6 utts ; Test: 2 utts
  - Degradations: NOISEX database
- Models and features
  - 32 mixture GMMs
  - 12 MFCCs (c1-c12)
- Results:
  - Overall

    48.8% (DFT), **62.7%** (SWLP-GD)
  - Mismatched

    36.5% (DFT), **54.2%** (SWLP-GD)

C – clean case; B10, F10, V10 and W10 – noisy speech at 10 dB SNR (babble, factory, vehicle and white noise respectively)

DFT		Train				
	Degradation	C	B_{10}	F_{10}	V_{10}	W_{10}
Test	C	24	15	6	10	11
	B_{10}	15	24	20	11	10
	F_{10}	9	15	22	4	11
	V_{10}	11	4	10	24	2
	W_{10}	7	2	1	1	24

SWLP-GD		Train				
	Degradation	C	B_{10}	F_{10}	V_{10}	W_{10}
Test	C	24	15	14	19	10
	B_{10}	13	23	18	14	3
	F_{10}	12	23	24	14	14
	V_{10}	22	16	17	24	2
	W_{10}	6	11	14	3	21

Matched conditions

Mismatched conditions with large improvements

Conclusions

- SWLP-GD – key features
  - Provides robust spectral representation for feature extraction
  - Temporal weighting provides robustness in time domain
  - Group delay function provides robustness in frequency domain
- SWLP-GD – *vs* traditional spectral representations
  - lower log-spectral distortion and higher frequency weighted SNR compared to the traditional DFT, LP or STRAIGHT spectra.
  - performs better than the traditional MFCCs in a small-scale closed-set speaker recognition experiments for mismatched conditions of degradation

References

[1] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, “Stabilized weighted linear prediction,” Speech Communication, vol. 51, no. 5, pp. 401 – 411, 2009.

[2] B. Yegnanarayana, “Formant extraction from linear prediction phase spectra,” J. Acoust. Soc. Am., vol. 63, no. 5, pp. 1638–1640, May 1978.

[3] H. Murthy and B. Yegnanarayana, “Group delay functions and its applications in speech technology,” Sadhana, vol. 36, pp. 745–782, 2011.

[4] C. Magi, T. Bäckström, and P. Alku, “Objective and subjective evaluation of seven selected all-pole modeling methods in processing of noise corrupted speech,” in Proc. 7th Nordic Signal Processing Symposium (NORSIG 2006), Reykjavik, Iceland, June 2006.

[5] C. Ma, Y. Kamp, and L. F. Willems, “Robust signal selection for linear prediction analysis of voiced speech,” Speech Communication, vol. 12, no. 1, pp. 69 – 81, 1993.

[6] L. Deng, X. Cui, R. Pruvenok, J. Huang, and S. Momen, “A database of vocal tract resonance trajectories for research in speech processing,” in Proc. Int. Conf. Acoustics Speech and Signal Processing, Toulouse, France, 2006, pp. I–369–I–372.

[7] D. Gowda, J. Pohjalainen, M. Kurimo, and P. Alku, “Robust formant detection using group delay function and stabilized weighted linear prediction,” in Proc. Interspeech, Lyon, France, August 2013.

Questions?

Thank You!