Aalto University
School of Electrical
Engineering

Utilization of the Lombard effect for the intelligibility enhancement of telephone speech

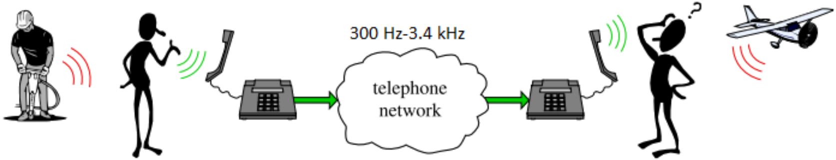
Emma Jokinen, Paavo Alku

Department of Signal Processing and Acoustics
Aalto University, Finland

emma.jokinen@aalto.fi

January 8, 2015

Introduction



- In mobile communications, the quality and intelligibility of the speech signal can be degraded by many factors, e.g.
  - The transmission through the radio channel and the low bit-rate coding used
  - Environmental noise in one or both ends of the communication channel

Figure adapted from [*Sauert et al.* 2006]

Post-processing of telephone speech

- Signal processing methods applied at the receiving side of the communication channel
- Do not require any changes to existing speech codecs
- Used to combat the effect of degradations on quality and intelligibility
- Special requirements for algorithms
    - Real-time processing in short speech frames
    - Low computational complexity

Quality vs. intelligibiility

- Traditionally post-processing methods are intended for quality enhancement, for instance
  - Suppression of quantization noise
  - Reduction of far-end noise in the signal
- In adverse background noise conditions, the intelligibility of speech is severely compromised

$\rightarrow$ Methods especially designed for intelligibility enhancement are needed

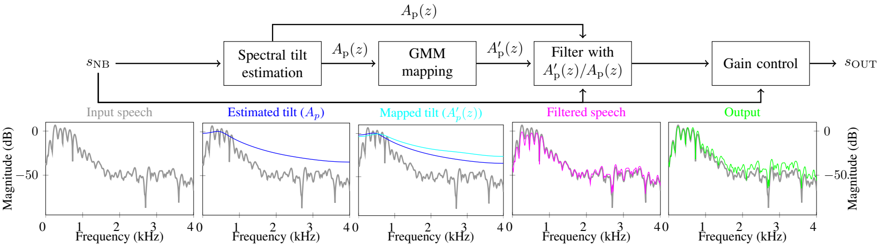
Intelligibility enhancement

- Good results have been achieved with fixed high-pass filtering [*Hall and Flanagan* 2010]
- More advanced techniques are based on modelling how humans hear or understand speech using, e.g.
  - Speech intelligibility index [*Sauert and Vary* 2010; *Taal et al.* 2013]
  - Glimpse proportion [*Tang and Cooke* 2012]
  - Auditory models [*Taal et al.* 2014]
- Few techniques model the Lombard effect, i.e., modifying the production of speech by humans in adverse conditions
- By imitating the Lombard effect, hopefully more natural-sounding modifications can be achieved

Proposed Lombard modelling

- The Lombard effect consists of multiple time and frequency-domain modifications, e.g.
  - Increase in $F0$
  - Decrease in spectral tilt
  - Changes in formant frequencies
- The change in spectral tilt has been shown to be important for the intelligibility increase in Lombard speech [*Lu and Cooke* 2009]
- A statistical, GMM-based mapping of spectral tilt from normal to Lombard speech is proposed [*Jokinen et al.* 2014a;b]

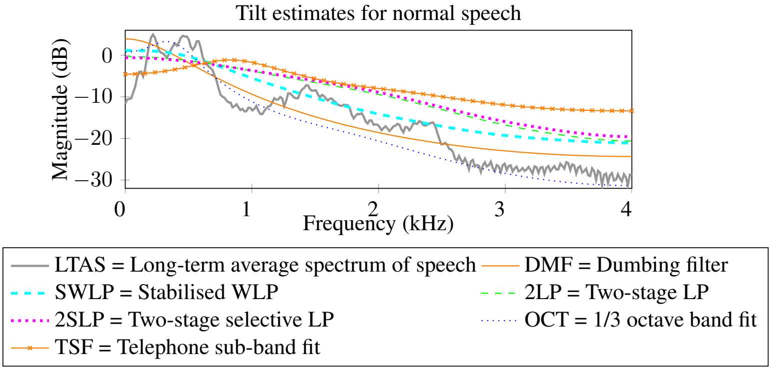
Proposed Lombard modelling



Proposed Lombard modelling
Spectral tilt estimation

1. Dumbing filter (DMF) [*Mizuno and Abe* 1995]
   - $H(z) = 1/(1 - gz^{-1})^2$, where g depends on the autocorrelation coefficients
2. Stabilized weighted linear prediction (SWLP) [*Magi et al.* 2009]
   - All-pole modelling technique where the square of the residual is temporally weighted
3. Two-stage LP (2LP) [*Jokinen et al.* 2012]
   - 20th order LP followed by 6th order LP
4. Two-stage selective LP (2SLP)
   - 2LP where first LP analysis is frequency selective
5. 1/3-octave band energy fit (OCT) [*Lu and Cooke* 2009]
   - All-pole filter fit to 1/3-octave band energies
6. Telephone sub-band magnitude fit (TSF) [*Kontio et al.* 2007]
   - All-pole filter fit to average magnitudes of sub-bands

Proposed Lombard modelling
Spectral tilt estimation



Proposed Lombard modelling
GMM mapping

- Gaussian mixtures with $M = \{5, 10, 50, 100\}$ full-covariance components considered
- Both the parameter representation (LP, LSF, RC and LAR) and number of components were varied
- The model parameters were trained with the expectation-maximization algorithm

Proposed Lombard modelling
Speech material

- Two Finnish databases of parallel normal and Lombard recordings
  - Training database: 360 sentences from 6 speakers (3 male)
  - Development database: short recordings from 18 speakers (9 male)
- A subset of the training data was selected utilizing the speech intelligibility index
- All samples were pre-processed to resemble narrowband telephone speech
- Voiced frames of normal and Lombard samples were aligned using dynamic time warping

Proposed Lombard modelling
Selected GMM mapping

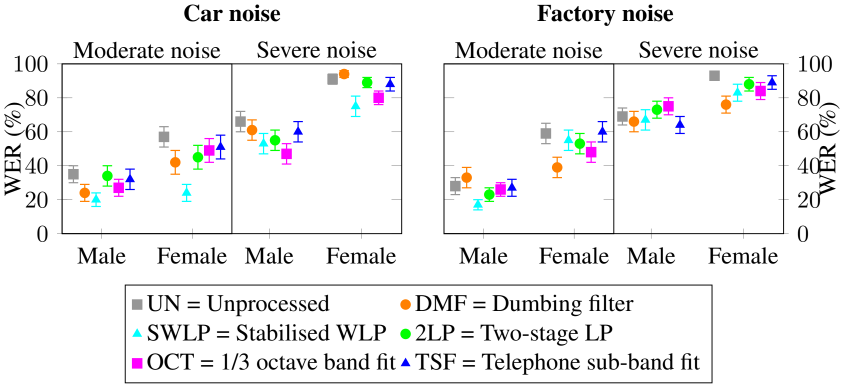
- The best models were selected based on explained variance (R^2) and log-spectral distortion

	DMF	SWLP	2LP	OCT	TSF
Parameter representation	RC	LSF	LSF	LSF	LSF
Number of mixtures	5	10	10	50	50
R^2	0.97	0.99	0.99	0.91	0.95

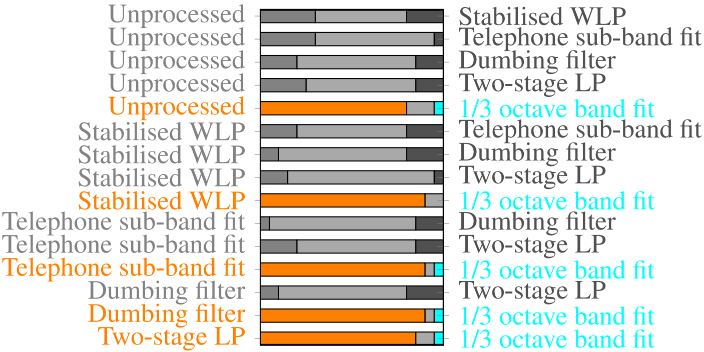
Subjective evaluation

- Finnish sentence material with 4 speakers (2 male)
- Samples were preprocessed to resemble narrowband telephone speech
- 10 listeners
- The evaluation consisted of
  1. a word-error rate (WER) test with two types of noise
     - Car noise (SNR levels: -5 dB, and -10 dB)
     - Factory noise (SNR levels: 0 dB, and -5 dB)
  2. a pair comparison test concerning the overall quality

Results
WER test



Results
Preference test

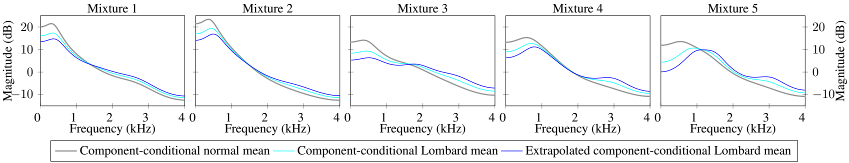


Proposed Lombard modelling
Extrapolation

→ The original GMM models trained with SWLP features were extrapolated

- Linear extrapolation of the component-conditional Lombard vector means
$$\vec{\mu}'_{y|i} = (\vec{\mu}_{y|i} - \vec{\mu}_{x|i})\gamma + \vec{\mu}_{x|i},$$
where γ controls the amount of extrapolation
- The maximum γ was chosen by restricting the number of resonances in the output
- GMMs with 5 and 10 components were considered with LSF parameters

Extrapolated Lombard modelling

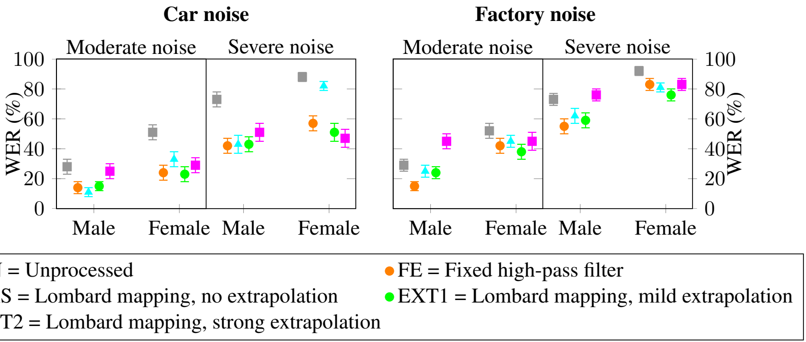


Subjective evaluation

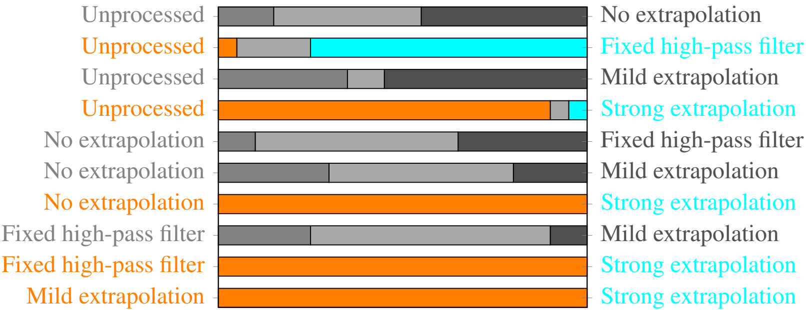
- Finnish sentence material with 4 speakers (2 male)
- Samples were preprocessed to resemble narrowband telephone speech
- 10 listeners
- The evaluation consisted of
  1. a word-error rate (WER) test with two types of noise
     - Car noise (SNR levels: -5 dB, and -10 dB)
     - Factory noise (SNR levels: 0 dB, and -5 dB)
  2. a pair comparison test concerning the overall quality

Results

WER test



Results
Preference test



Conclusion

- GMM-based post-processing method was proposed for intelligibility enhancement of telephone speech
- The maximal intelligibility gain of spectral tilt modification was evaluated by extrapolating the mapping
- Mild extrapolation provided similar improvement as high-pass filtering
- A production-based statistical mapping can follow natural speaker behavior in different noise conditions

References

J.L. Hall and J.L. Flanagan. Intelligibility and listener preference of telephone speech in the presence of babble noise. *J. Acoust. Soc. Amer.*, 127(1): 280–285, 2010.

E. Jokinen, P. Alku, and M. Vainio. Lombard-motivated post-filtering method for the intelligibility enhancement of telephone speech. In *Proc. Interspeech*, 2012.

E. Jokinen, U. Remes, M. Takanen, K. Palomäki, M. Kurimo, and P. Alku. Spectral tilt modelling with GMMs for intelligibility enhancement of narrowband telephone speech. In *Proc. Interspeech*, pages 2036–2040, 2014a.

E. Jokinen, U. Remes, M. Takanen, K. Palomäki, M. Kurimo, and P. Alku. Spectral tilt modelling with extrapolated GMMs for intelligibility enhancement of narrowband telephone speech. In *Proc. IWAENC*, 2014b.

J. Kontio, L. Laaksonen, and P. Alku. Neural network-based artificial bandwidth expansion of speech. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(3): 873–881, 2007.

Y. Lu and M. Cooke. The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Commun.*, 51(12): 1253–1262, 2009.

C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku. Stabilised weighted linear prediction. *Speech Commun.*, 51(5):401–411, 2009.

H. Mizuno and M. Abe. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. *Speech Commun.*, 16(2):153–164, 1995.

B. Sauert and P. Vary. Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement. In *ITG-Fachtagung Sprachkommunikation*, 2010.

B. Sauert, G. Enzner, and P. Vary. Near end listening enhancement with strict loudspeaker output power constraining. In *Proc. IWAENC*, 2006.

C.H. Taal, J. Jensen, and A. Leijon. On optimal linear filtering of speech for near-end listening enhancement. *IEEE Signal Process. Lett.*, 20(3):225–228, 2013.

C.H. Taal, R.C. Hendriks, and R. Heusdens. Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure. *Comput., speech, lang.*, 28(4):858 – 872, 2014.

Y. Tang and M. Cooke. Optimised spectral weightings for noise-dependent speech intelligibility enhancement. In *Proc. Interspeech*, 2012.