Effects of Lombard Reflex on Deep-Learning-Based Audio-Visual Speech Enhancement Systems

Daniel Michelsanti[1], Zheng-Hua Tan[1], Sigurdur Sigurdsson[2], Jesper Jensen[1,2]

[1] Aalborg University, Department of Electronic Systems, Denmark
[2] Oticon A/S, Denmark
{danmi,zt,jje}@es.aau.dk {ssig,jesj}@oticon.com

AALBORG UNIVERSITY
DENMARK

CASPR
Centre for Acoustic Signal Processing Research

Motivation

- **Speech enhancement**: task of estimating the clean speech of a speaker immersed in an acoustically noisy environment (Fig. 1).

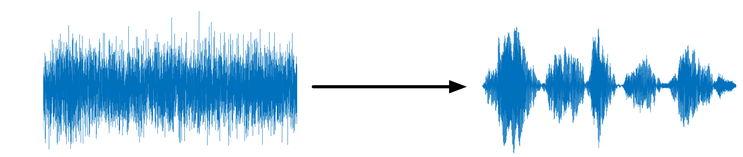

Fig. 1: Speech enhancement.

- Important in several applications:
  - Speech recognition.
  - Speaker verification.
  - Hearing aids.
- **Lombard effect**: Reflex occurring when speakers talk in a noisy environment.
- Current deep-learning-based systems do not take Lombard effect into account: they are trained with neutral (non-Lombard) speech utterances recorded under quiet conditions to which noise is artificially added.
- We study the effects that the Lombard reflex has on deep-learning-based audio-visual speech enhancement systems.

Experiments

- Pipeline shown in Fig. 2:
  - Neural network architecture inspired by [1] and identical to [2].
  - Single modality systems: one of the encoder is removed.
- Systems trained on the utterances from the **Lombard GRID** corpus [3], to which **speech shaped noise** is added at several signal to noise ratios (SNRs).
- Systems tested on speakers observed (**seen speakers**) and not observed (**unseen speakers**) during training.
- Models used in this study shown in Table 1.

Modality	Training Material	
	Non-Lombard Speech	Lombard Speech
Vision	VO-NL	VO-L
Audio	AO-NL	AO-L
Audio-visual	AV-NL / AV-NL*	AV-L / AV-L*

Table 1: Models used for the seen and the unseen (indicated with a *) speaker cases.

Pipeline for Audio-Visual Speech Enhancement

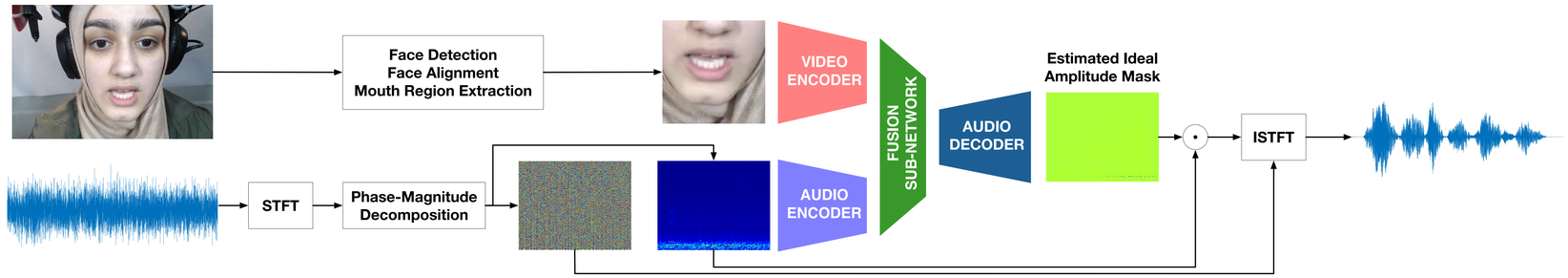

Fig. 2: Pipeline for audio-visual speech enhancement.

Results

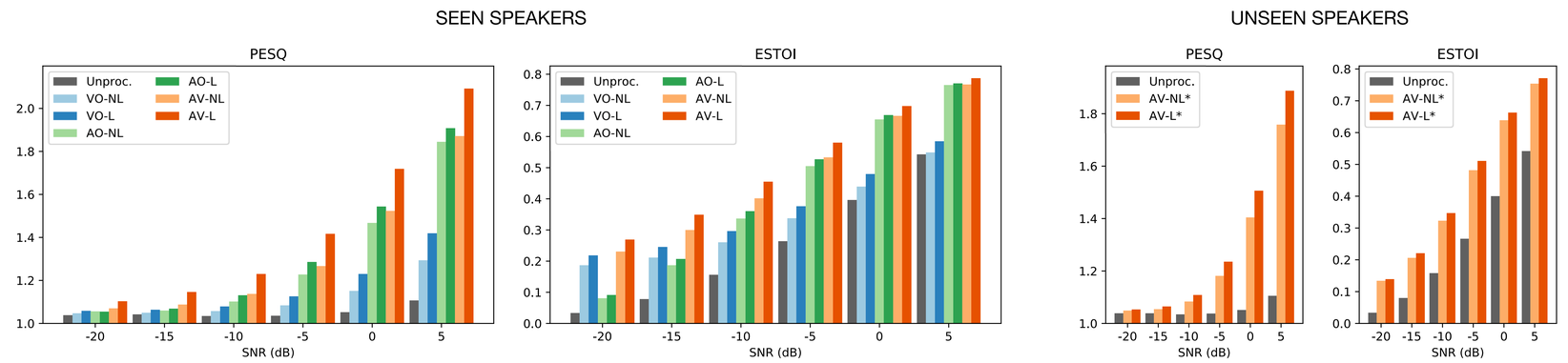

Fig. 3: Results in terms of estimated speech quality (PESQ) and estimated speech intelligibility (ESTOI) for seen (left) and unseen (right) speakers. *Unproc.* indicates the scores for the unprocessed signals.

Conclusions

- The Lombard effect has an impact on audio-only, video-only and audio-visual speech enhancement systems.
- There is a benefit of as much as 5 dB by taking into account the mismatch between neutral and Lombard speech in the design of audio-visual systems.
- Future works include listening tests to validate the findings obtained with objective measures of speech quality and speech intelligibility.

References

[1] A. Gabbay, A. Shamir and S. Peleg, "Visual speech enhancement," *Proc. of Interspeech*, 2018.

[2] D. Michelsanti, Z.-H.Tan, S. Sigurdsson and J. Jensen, "On training targets and objective functions for deep-learning-based audio-visual speech enhancement," *arXiv preprint arXiv:1811.06234*, 2018.

[3] N. Alghamdi, S. Maddock, R. Marxer, J. Barker and G. J. Brown, "A corpus of audio-visual Lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, 2018.