DEEP-LEARNING-BASED AUDIO-VISUAL SPEECH ENHANCEMENT IN PRESENCE OF LOMBARD EFFECT

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

[1]Dept. of Electronic Systems, Aalborg University, 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).

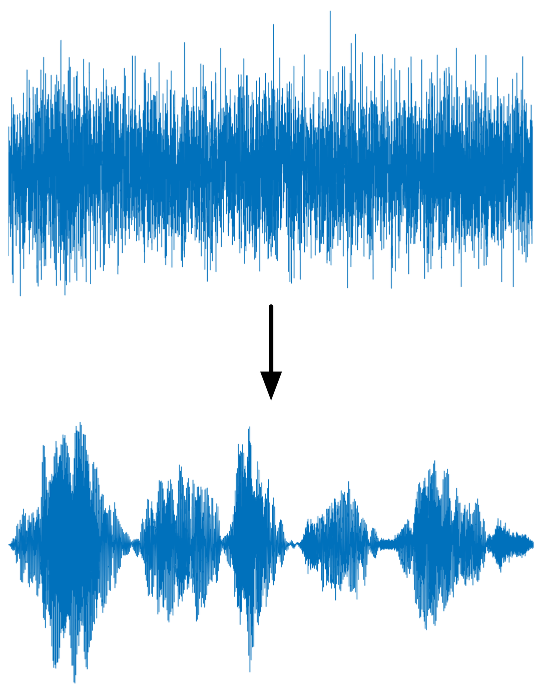

Fig. 1: Speech enhancement.

- 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:
  - Architecture inspired by [1].
  - Single modality systems: one of the encoder is removed.
- Systems trained on the utterances from the **Lombard GRID** corpus [2], to which **speech shaped noise** is added at several signal to noise ratios (SNRs).
- Systems tested on speakers observed (**seen speakers**) during training to isolate the impact of Lombard effect from other factors.
- Models used in this study shown in Table 1.

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

Table 1: Models used in this study.

References

[1] A. Gabbay et al., "Visual speech enhancement," *Proc. of Interspeech*, 2018.

[2] N. Alghamdi et al., "A corpus of audio-visual Lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, 2018.

[3] D. Michelsanti et al., "On training targets and objective functions for deep-learning-based audio-visual speech enhancement," *Proc. of ICASSP*, 2019.

Training Targets and Objective Functions

	Direct Mapping (DM)	Indirect Mapping (IM)	Mask Approximation (MA)
Short Time Spectral Amplitude (STSA)	$J = a\sum_{k,l}(A_{k,l} - \widehat{A}_{k,l})^2$	$J = a\sum_{k,l}(A_{k,l} - \widehat{M}_{k,l}R_{k,l})^2$	$\boxed{J = a\sum_{k,l}(M^{\text{IAM}}_{k,l} - \widehat{M}_{k,l})^2}$
Log Spectral Amplitude (LSA)	$J = a\sum_{k,l}(\log(A_{k,l}) - \log(\widehat{A}_{k,l}))^2$	$J = a\sum_{k,l}(\log(A_{k,l}) - \log(\widehat{M}_{k,l}R_{k,l}))^2$	-
Mel-Scaled Spectral Amplitude (MSA)	$J = b\sum_{q,l}(\overline{A}_{q,l} - \widehat{\overline{A}}_{q,l})^2$	$J = b\sum_{q,l}(\overline{A}_{q,l} - \widehat{\overline{M}}_{q,l}\overline{R}_{q,l})^2$	-
Log Mel-Scaled Spectral Amplitude (LMSA)	$J = b\sum_{q,l}(\log(\overline{A}_{q,l}) - \log(\widehat{\overline{A}}_{q,l}))^2$	$J = b\sum_{q,l}(\log(\overline{A}_{q,l}) - \log(\widehat{\overline{M}}_{q,l}\overline{R}_{q,l}))^2$	-
Phase Sensitive Spectral Amplitude (PSSA)	$J = a\sum_{k,l}(A_{k,l}\cos(\theta_{k,l}) - \widehat{A}_{k,l})^2$	$J = a\sum_{k,l}(A_{k,l}\cos(\theta_{k,l}) - \widehat{M}_{k,l}R_{k,l})^2$	$J = a\sum_{k,l}(M^{\text{PSM}}_{k,l} - \widehat{M}_{k,l})^2$

Table 2: Taxonomy proposed in [3]. Here: $a = \frac{1}{TF}$, $b = \frac{1}{TQ}$, $M^{\text{IAM}}_{k,l} = \frac{A_{k,l}}{R_{k,l}}$ and $M^{\text{PSM}}_{k,l} = \frac{A_{k,l}}{R_{k,l}}\cos(\theta_{k,l})$. In this study, the highlighted objective function is used.

Pipeline for Audio-Visual Speech Enhancement

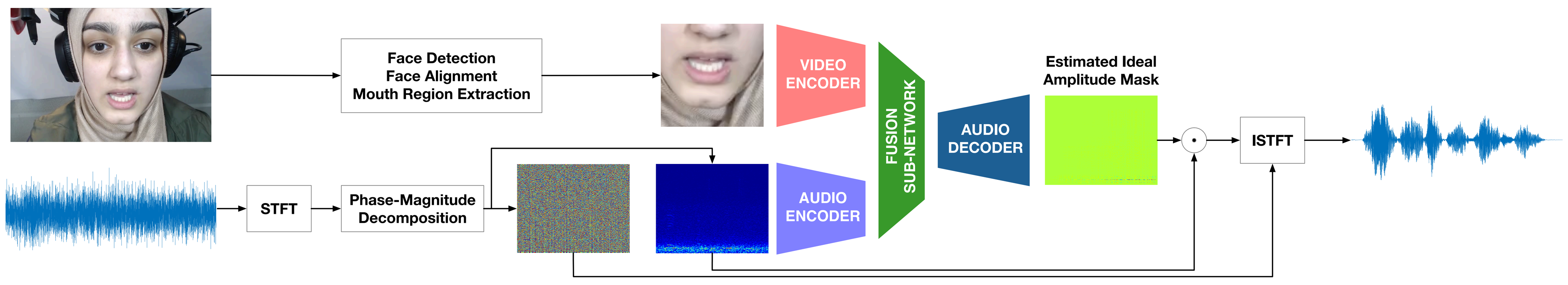

Fig. 2: Pipeline for the audio-visual speech enhancement approach used in this study.

Results

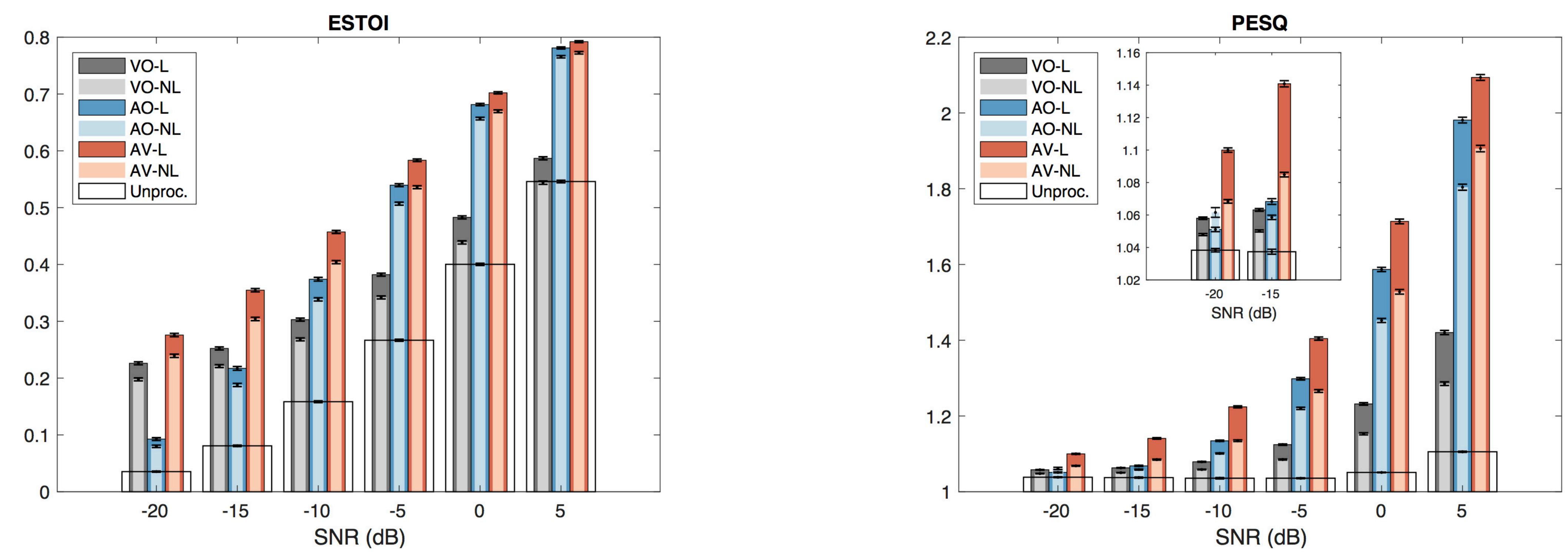

Fig. 3: ESTOI and PESQ results for the various approaches.

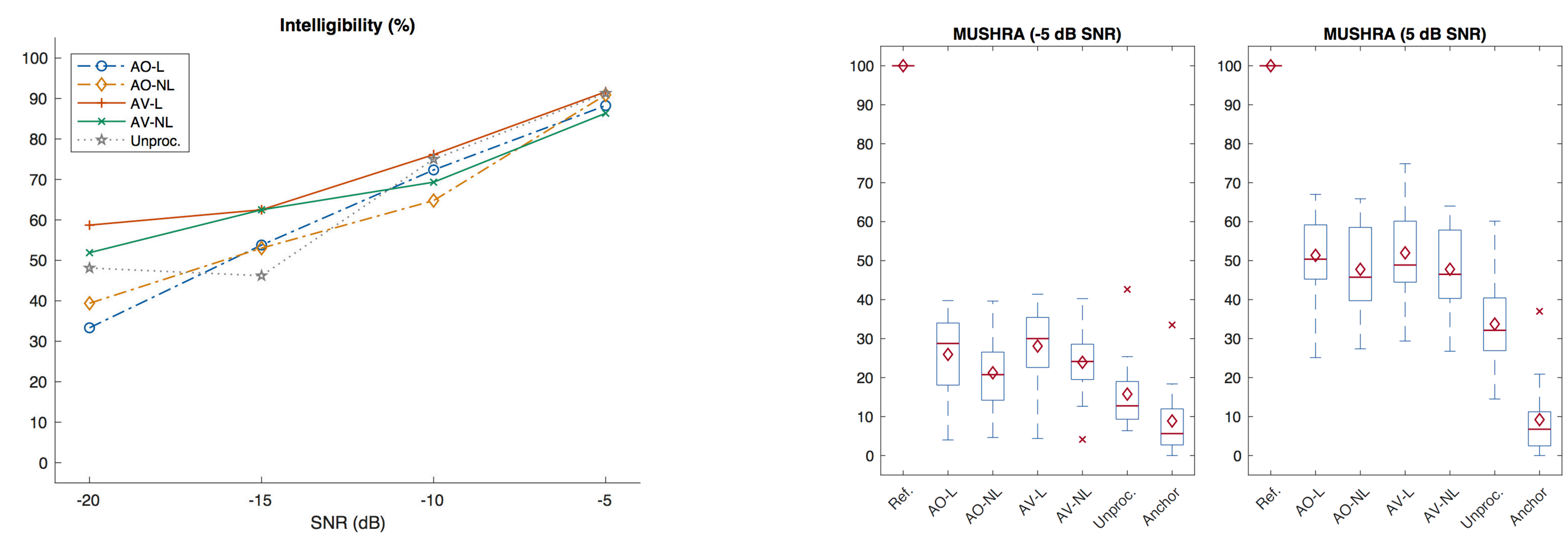

Fig. 4: Listening tests results using audio-visual stimuli to evaluate speech intelligibility and speech quality.

Conclusion

- The Lombard effect affects the performance of speech enhancement systems.
- The impact of visual differences between Lombard and non-Lombard speech on estimated speech intelligibility is higher than the impact of acoustic differences.
- A 5 dB benefit can be observed for the estimated speech quality at low SNRs when the mismatch between neutral and Lombard speech is taken into account in the design of audio-visual systems.
- Listening tests using audio-visual stimuli show that:
  - Signals processed with L systems tend to have higher intelligibility if compared to the other processing conditions.
  - The speech quality of the L systems is statistically significantly better than the one of the NL systems at -5 dB SNR.