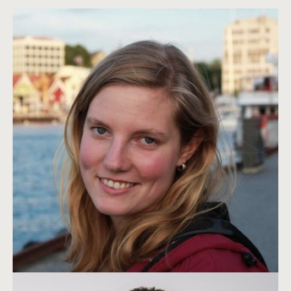
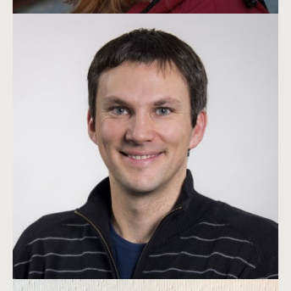
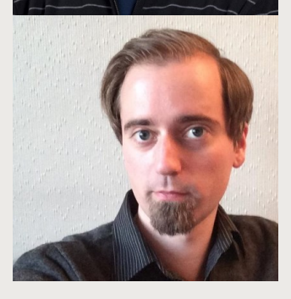
SINTEF

Subjective Intelligibility of Deep Neural Network-Based Speech Enhancement

Femke B. Gelderblom

Tron V. Tronstad

Erlend M. Viggen


SINTEF Digital, Trondheim, Norway

Introduction

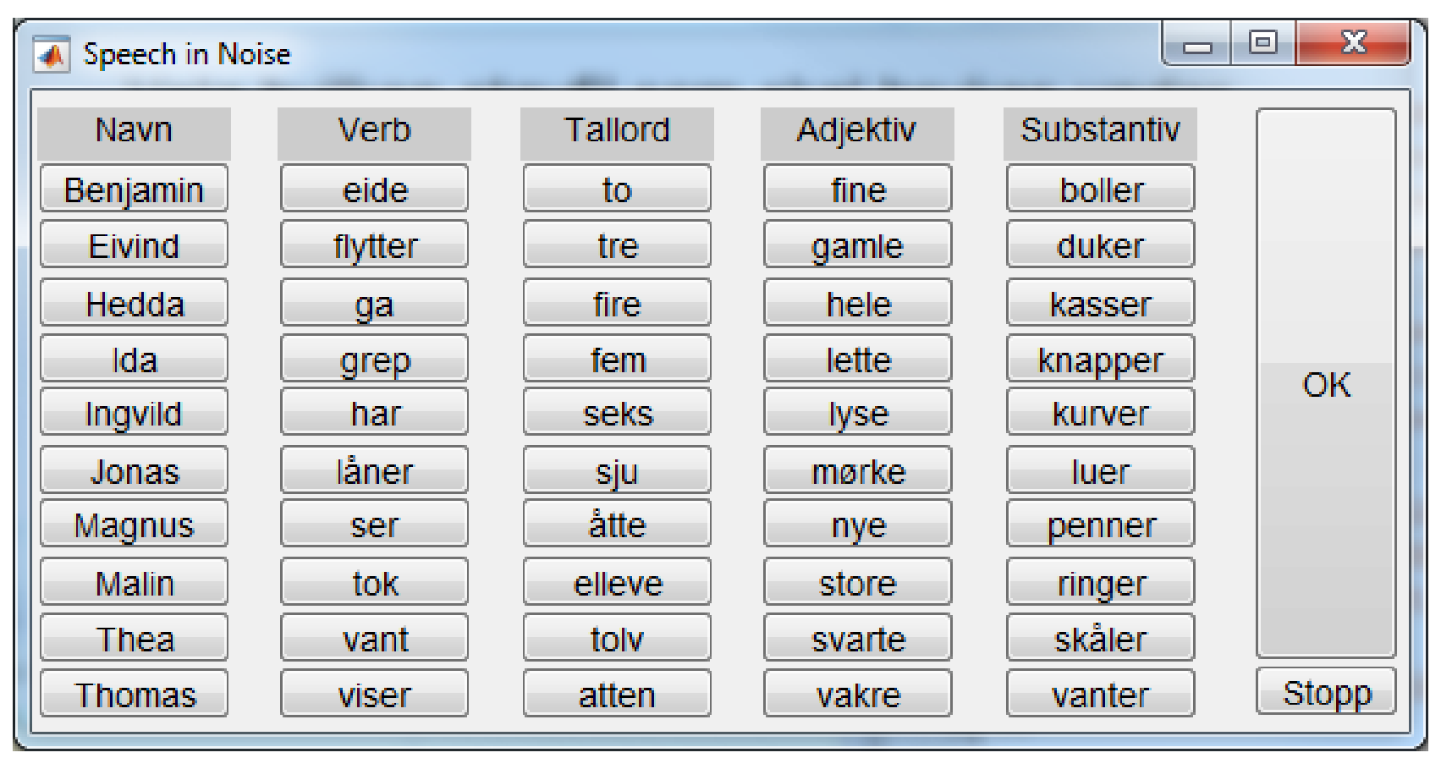
- The intelligibility of DNN-based speech enhancement systems is evaluated through objective measures such as STOI *(Taal et al., 2011)*
- However, STOI does not always correctly predict intelligibility *(Jensen & Taal, 2016)*

Does STOI correctly predict the intelligibility of DNN-based speech enhancement systems? We performed a subjective evaluation test to find out.

Subjective evaluation

Speech in noise test

- Speech: Male voice, **random Hagerman sentences in Norwegian** *(Øygarden, 2009)*
- Noise: Traffic from a crossroads in Trondheim
- Subjects had to pick out words:



Speech in Noise

Navn	Verb	Tallord	Adjektiv	Substantiv
Benjamin	eide	to	fine	boller
Eivind	flytter	tre	gamle	duker
Hedda	ga	fire	hele	kasser
Ida	grep	fem	lette	knapper
Ingvild	har	seks	lyse	kurver
Jonas	låner	sju	mørke	luer
Magnus	ser	åtte	nye	penner
Malin	tok	elleve	store	ringer
Thea	vant	tolv	svarte	skåler
Thomas	viser	atten	vakre	vanter

- Adjusted SNR dynamically using the Ψ method to efficiently determine participants' psychometric functions
- **Goal: Find the speech recognition threshold** (lowest SNR at which 50 % of words are understood)
- Test was run for baseline clips and DNN-enhanced clips

Participants

- 15 native Norwegians, aged 39–65
- All were naive listeners given a training session before the test started

Sound examples

bit.ly/2uhLWcL

DNN-based speech enhancement

Setup

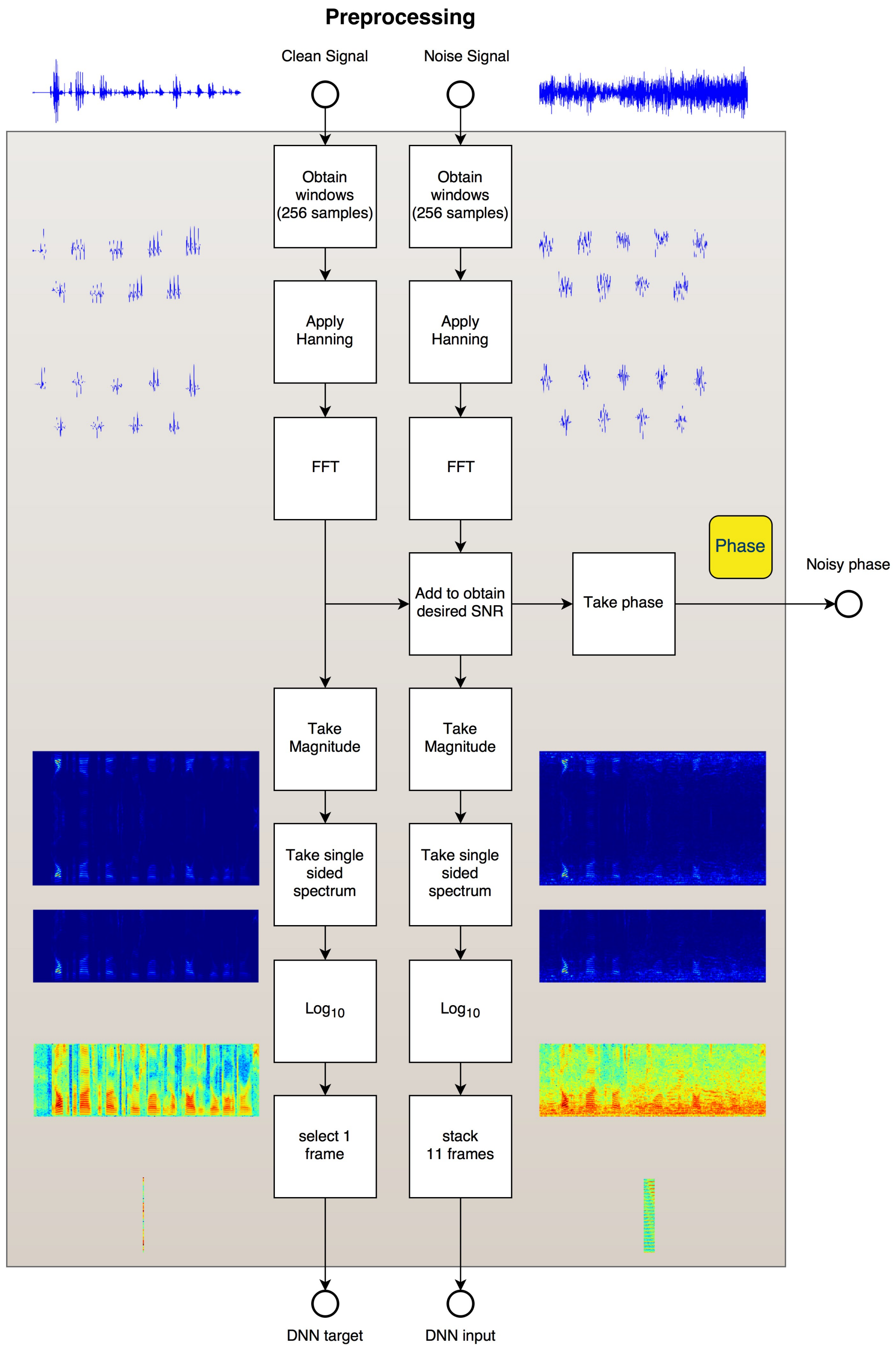
- Closely based on *Xu et al., 2015*
- Multilayer feed-forward network
- Input/output: log-frequency spectra with frames of 256 samples (32 ms at 8 kHz)
- Input: Stacked frames of noisy speech
- Target: One frame of clean speech

Training

- **Trained and validated on the Norwegian-language speech corpus Språkbanken**
- Loss function: Mean squared error
- Trained for SNR ∈ {-5, 0, 5, 10, 15, 20} dB
- Manually optimised hyperparameters to improve STOI on validation set

Enhancement

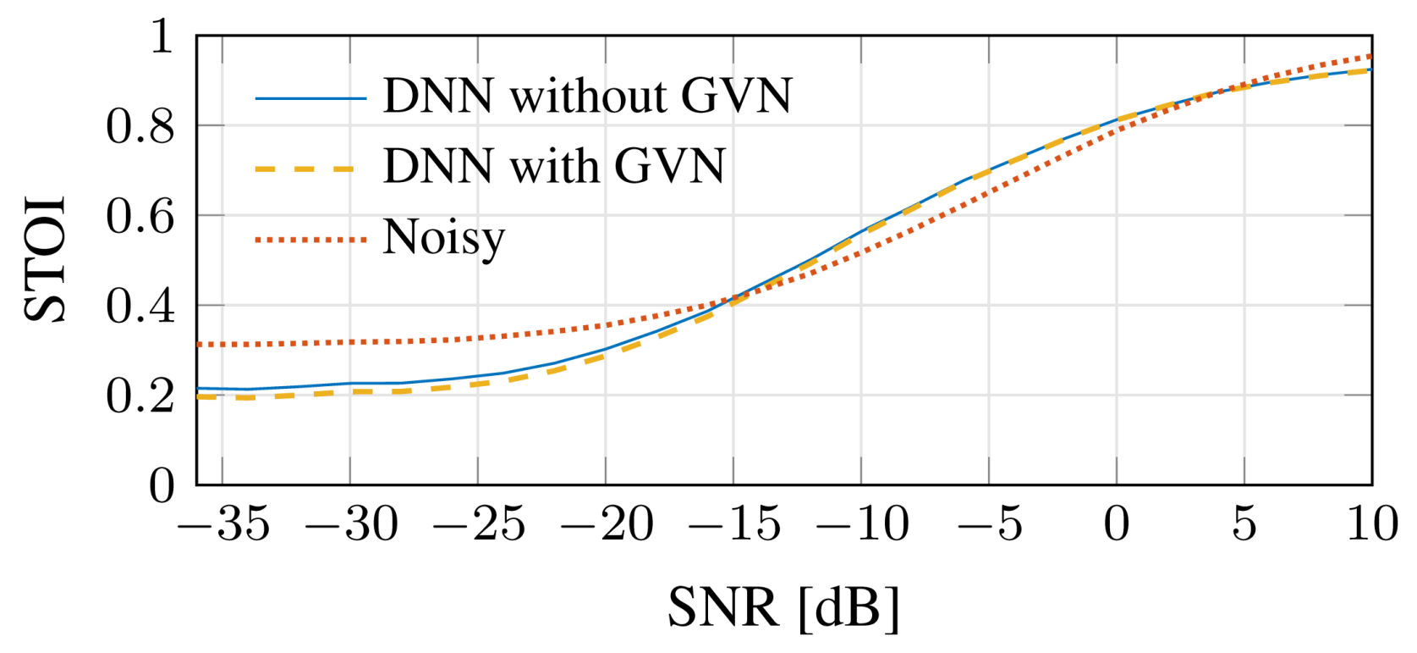
- Converted DNN output into samples using the phase from the noisy DNN input
- Tested with and without global variance normalisation (GVN) post-processing step



Results

Objective evaluation

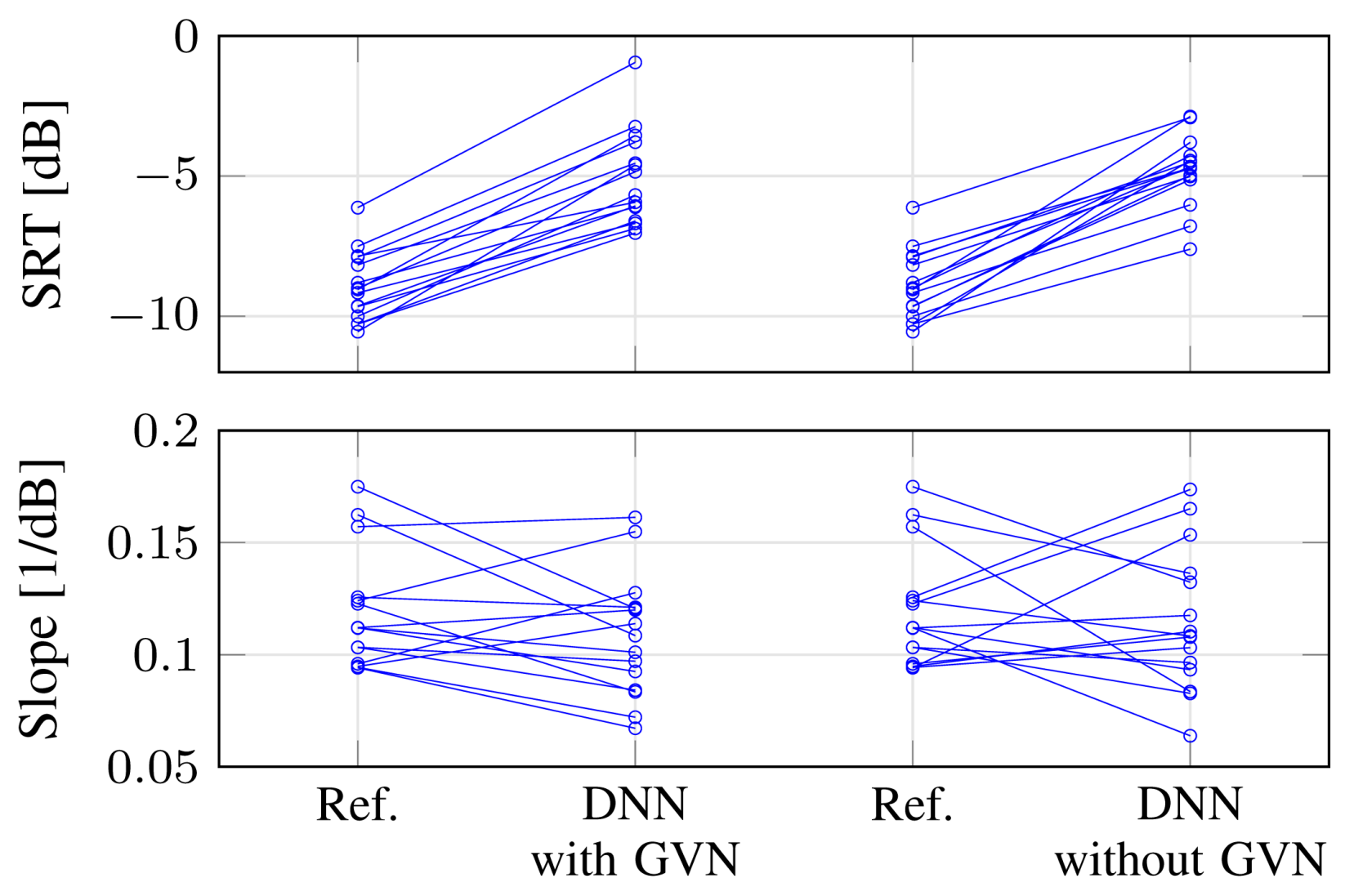
- STOI on the subjective evaluation set:



- **Predicts that this DNN *improves* intelligibility** for SNR ∈ [-14, 4] dB

Subjective evaluation

- Speech recognition threshold (SRT) significantly degrades (4 dB median)
- Slope of psychometric function does not show significant differences
- GVN makes no significant difference



- **Shows that this DNN *reduces* intelligibility**

Main references

- Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 23, 2015.
- C. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE Trans. Audio Speech Lang. Proc., vol. 19, 2011.
- J. Jensen, C. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 24, 2016.
- J. Øygarden, *Norwegian speech audiometry*, Ph.D. thesis, Norwegian University of Science and Technology, 2009.

Conclusion

Our results show a significant degradation in intelligibility, even though STOI scores predicted otherwise. Therefore, we advise against solely relying on STOI when designing DNN-based speech enhancement systems for human listeners.