INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Subjective intelligibility of deep neural network-based speech enhancement

Femke B. Gelderblom[1], *Tron V. Tronstad*[1], *Erlend Magnus Viggen*[1]

[1]Acoustics Research Centre, SINTEF Digital, Trondheim, Norway

femke.gelderblom@sintef.no, tronvedul.tronstad@sintef.no, erlendmagnus.viggen@sintef.no

Abstract

Recent literature indicates increasing interest in deep neural networks for use in speech enhancement systems. Currently, these systems are mostly evaluated through objective measures of speech quality and/or intelligibility. Subjective intelligibility evaluations of these systems have so far not been reported. In this paper we report the results of a speech recognition test with 15 participants, where the participants were asked to pick out words in background noise before and after enhancement using a common deep neural network approach. We found that, although the objective measure STOI predicts that intelligibility should improve or at the very least stay the same, the speech recognition threshold, which is a measure of intelligibility, deteriorated by 4 dB. These results indicate that STOI is not a good predictor for the subjective intelligibility of deep neural network-based speech enhancement systems. We also found that the postprocessing technique of global variance normalisation does not significantly affect subjective intelligibility.

Index Terms: speech enhancement, deep neural network, subjective evaluation, speech intelligibility

1. Introduction

The field of speech enhancement (SE) aims to improve the quality and/or intelligibility of speech that has been degraded [1]. In the past few years, deep neural networks (DNNs) [2, 3] have emerged as a promising approach for SE, outperforming earlier approaches. SE has been proven useful as a preprocessing step for automatic speech recognition systems to decrease their word error rates [4, 5, 6], but the field also aims to make degraded speech easier to understand and/or more comfortable to listen to for humans [5, 7, 8].

The performance of each of these SE approaches with respect to intelligibility improvement is typically evaluated through objective measures. Especially popular measures are STOI [9], PESQ [10], or the word error rates of speech recognition systems. PESQ was originally designed as a measure for speech quality rather than intelligibility, but was then found to also correlate reasonably well with subjective intelligibility [11]. None of today's objective measures of intelligibility can perfectly predict intelligibility to humans, and their correlation depends on the type of speech degradation present [9, 12].

Thus, listening tests are necessary to quantify the benefit of DNN-based SE for human listeners. Listening tests for speech *quality* have previously been reported in the literature with positive results [5, 7, 8]. Quality is however highly subjective, since whether a signal sounds 'good' or 'poor' is based on listeners' preferences. Intelligibility tests are more objective in nature as these allow for quantitative scoring of how much information the listener actually understood. To our knowledge, and despite its popularity, no one has tested the predictive power of STOI for DNN-based SE against subjective listening tests.

In this work we report the results of a series of listening tests for *intelligibility*, where our test subjects attempted to comprehend speech in background noise, before and after DNN-based speech enhancement. Here, we evaluate whether STOI correctly predicts change in subjective intelligibility for a reasonably common DNN setup. Additionally, we analyse the effect of the 'global variance normalisation' postprocessing step (described in sec. 2.1.3) on intelligibility.

2. Methods

2.1. DNN system overview

The speech enhancement system is loosely based on the system Xu et al. proposed in [8], but omits pre-training with restricted Boltzmann machines as their results indicate that the effect of pre-training was negligible. The DNN was implemented using Keras 1.0.5 [13].

2.1.1. Speech and noise preparation

For training, clean speech was combined with noise to obtain noisy speech. The clean speech was obtained from the Norwegian-language library 'Språkbanken' [14], to ensure that the DNN trained on the same language as used during subjective evaluation. The setup of Språkbanken is similar to that of the more widely used TIMIT. The clean speech database was divided into a training set, a validation set, and a test set (not used for this article). Care was taken to ensure that each set was balanced with respect to gender and dialect, and that no specific speakers or sentences occurred in more than one set. The final training set consisted of 1932 sentences from 137 unique speakers, while the validation set contained 816 sentences from 48 speakers.

Periods of silence lasting longer than 75 ms were trimmed to 75 ms where their levels were 40 dB or more below the peak of the given sentence, to capture the average dynamic range of speech [11]. The 75 ms length was arbitrarily chosen as a compromise between minimising the number of quiet training samples, and maintaining a clear separation between words.

Noisy speech was obtained by combining the clean speech with the same 104 noises Xu et al. used in [8], all obtained from either the Aurora database [15] or Guoning Hu's collection [16]. Six different signal-to-noise ratios (SNRs) ranging from -5 dB to 20 dB, with SNRs applied at sentence level, were used for training. This range was chosen, despite the need for lower SNRs during speech intelligibility testing, as a DNN trained with a more suitable SNR range, but otherwise equal hyperparameters, actually performed worse in terms of STOI values at all SNRs.

The noisy speech, along with clean speech (with 'infinite SNR'), was used as input for the DNN. This lead to a total of 1984 hours of training data. Noisy speech for validation was obtained by combining the clean validation speech with the 15 unseen noises Xu et al. specified in [8], obtained from either the Aurora or NOISEX-92 databases [15, 17]. This resulted in 98 hours of validation data. Both the noisy and clean speech sig-

http://dx.doi.org/10.21437/Interspeech.2017-1041

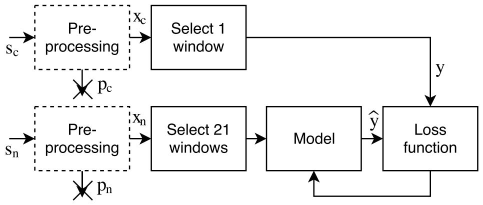


Figure 1: *Diagram of training procedure. The clean and noisy phases output by the preprocessing steps are discarded.*

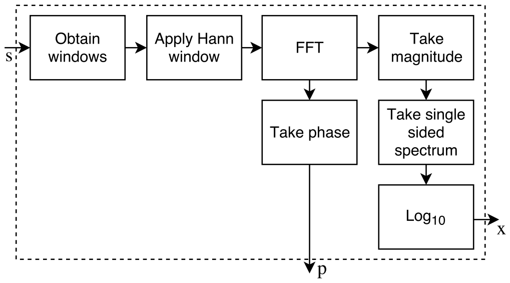


Figure 2: *Diagram of preprocessing steps*

nals were down-sampled to 8 kHz, as this was the lowest sampling rate of any of the original signals.

2.1.2. *Training*

Figure 1 shows a block diagram of the training procedure. The model learns in a supervised manner, with the standard mean squared error (MSE) loss function

$$\text{MSE} = \frac{1}{n}\sum_{k}\left(\hat{y}_k - y_k\right)^2, \tag{1}$$

where $\hat{y}_k$ and y_k represent the kth frequency bins of the enhanced and clean log-power spectral features, respectively. The features were obtained through the preprocessing steps shown in Figure 2. During preprocessing, the signal is first separated into windows that overlap by 50 %. The windows consist of 256 samples, and thus represent a timeframe of 32 ms at 8 kHz. The Hann window function is then applied to each window before the result is Fourier transformed. Redundant information above the Nyquist frequency is discarded from the resulting magnitude spectrum to obtain a single-sided output. Finally, log-power spectrum features are calculated for each window. After preprocessing, the input vector is obtained by stacking 21 sequential 50 % overlapping windows that contain the log-power spectral features. This provides the DNN with 160 ms historic and 160 ms future context. The phase of both clean and noisy speech is ignored during training. No normalisation of input or output was applied.

The DNN model is a multi-layer perceptron, a feedforward neural network with fully connected layers. It has three hidden layers, each with 2048 nodes and LeakyReLU activation functions. The model is trained with 50 % dropout on the hidden layers using the Adam optimiser with a learning rate of 10^{-5}. The activation function of the output layer is linear.

Training continued until the STOI value reached a maximum for the validation set at the 8th epoch. The model's state at this epoch was used for enhancement. We also trained a number of different models with different hyperparameters; the model described here was selected due to its better STOI performance.

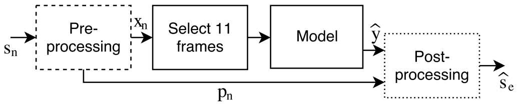


Figure 3: *Diagram of enhancement procedure*

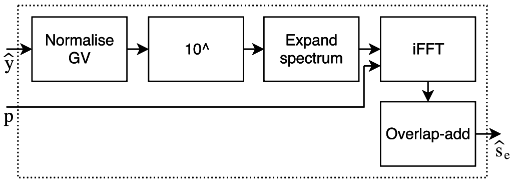


Figure 4: *Diagram of postprocessing steps*

2.1.3. *Enhancement*

After training, the model could be used to enhance noisy speech. Figure 3 shows the enhancement procedure, and Figure 4 shows the postprocessing steps.

Postprocessing mainly consists of reversing the steps that were taken during preprocessing, using the *noisy* phase for waveform reconstruction. The first step, global variance normalisation (GVN), is the exception to this reversal. This step aims to prevent over-smoothing by enforcing the variance of the enhanced speech to be equal to the variance of actual clean speech. During GVN, the DNN's output features are multiplied with a frequency bin independent factor calculated as

$$\beta = \sqrt{\frac{\text{var}_{m,k}[y_k(m)]}{\text{var}_{m,k}[\hat{y}_k(m)]}}, \tag{2}$$

where $\text{var}_{m,k}$ represents the variance over all values of m and k, with m indexing examples in the training set, and k indexing frequency bins. Furthermore, from the law of total variance we can calculate this variance as

$$\underset{m,k}{\text{var}}[a_k(m)] = \frac{1}{K}\sum_{k}\underset{m}{\text{var}}[a_k(m)] + \underset{k}{\text{var}}\left(\underset{m}{\text{var}}[a_k(m)]\right), \tag{3}$$

where K equals the total number of frequency bins and $a_k(m)$ represents either $y_k(m)$ or $\hat{y}_k(m)$. This specific method for the calculation of the global variance combines readily with Welford's online algorithm for variance computation, which is well suited to working with large data sets [18]. Two systems were tested for this work; one with, and one without the GVN step.

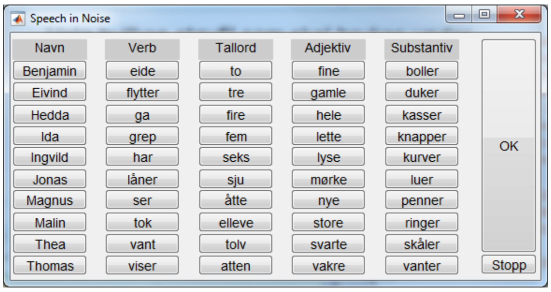
2.2. Objective evaluation

The short-term objective intelligibility (STOI) measure [9] was used to test the model's performance. The advantages of STOI include a documented strong correlation with subjective speech intelligibility [9] and the possibility to compare obtained results with earlier publications [8]. Additionally, unlike with some other popular objective measures like PESQ, use of STOI is not restricted by licencing.

Objective evaluation results were obtained both for the validation set and for the signals used during subjective testing.

2.3. Subjective evaluation

The subjective evaluation of intelligibility was performed using a speech recognition test. Figure 5 shows the user interface



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

Figure 5: *The GUI of the Norwegian-language subjective test*

implemented in MATLAB [19]. Random five-word sentences, all uttered by the same male speaker, were presented at different SNRs to determine the speech recognition threshold (SRT). All sentences were in Norwegian and structured the same way: [Name], [Verb], [Numeral], [Adjective], [Noun], with 10 options for each. The subjects' task was to pick out which word in each category was in the sentence they just heard. The speech material has been taken from Øygarden's hearing in noise test, which is based on Hagerman sentences [20].

To keep the subjective test to a manageable length, only one noise file was used: a road traffic recording from a crossroad in central Trondheim, a common type of background noise in cities. Each sentence was mixed with a random section of this noise file at the desired SNR. The SNR was calculated from the root-mean-square (RMS) value for the sentence without noise and the RMS value for the selected section of the noise signal. The background noise was kept constant at a comfortable level while the speech was varied to achieve the correct SNR. The speaker, utterances, and noise used in this test had not been included during DNN training nor during validation.

Each subject completed three tests. For each test case, all material was first down-sampled to 8 kHz. One test set was left otherwise untreated ('Noisy'), while for the other cases the speech was enhanced according to the method described in sec. 2.1.3 ('DNN with/without GVN'), where the GVN step was only included for one of these cases. The material of each test set was subsequently up-sampled to 44.1 kHz before being presented to the subject. All sentences were presented binaurally with Sennheiser HDA-200 headphones via an external sound card (Roland Edirol UA-101).

An adaptive procedure called the Ψ method [21] was used to determine the presentation levels during testing. The method uses the entropy of the posterior probability distribution in the determination of the next stimuli level. The Palamedes MATLAB toolbox [22] was used for the realisation of the Ψ method.

The test was not forced choice, but the test subjects were encouraged to guess whenever they thought they (partly) recognised a word. Both the guess and lapse rate were set to 0.01 in the method. The threshold and slope value were allowed to vary in the estimation of the psychometric function. The stimulation range of the SNRs was from -36 dB to 10 dB, in 2 dB steps.

15 persons, with ages from 39 to 65 (Mean $= 54.2$, SD $= 9.5$), participated. The only selection criteria observed was that all participants had to have Norwegian as their first language. All test subjects were given a training session before the three situations (Noisy, DNN with GVN, and DNN without GVN) were tested and the test sequence was randomised between each individual to reduce any further training effect that could occur during the session. The test subjects were also allowed to take a break during the test if they desired.

Table 1: *STOI results for the validation set. Results are averaged over the 15 unseen noise types and stated together with their sample standard deviation.*

SNR	Noisy	DNN without GVN	DNN with GVN
20	0.95 (0.01)	0.92 (0.01)	0.91 (0.01)
15	0.91 (0.02)	0.90 (0.01)	0.89 (0.01)
10	0.85 (0.03)	0.86 (0.02)	0.85 (0.02)
5	0.76 (0.04)	0.80 (0.02)	0.79 (0.02)
0	0.65 (0.04)	0.71 (0.03)	0.71 (0.03)
-5	0.55 (0.04)	0.61 (0.04)	0.60 (0.04)

3. Results

3.1. Objective evaluation

Table 1 shows the STOI results for the validation set. The GVN step shows no significant effect on the STOI results. DNN processing leads to improved scores as compared to the baseline for all SNRs under 10 dB. Looking at our unprocessed 'noisy' baseline, our STOI results at low SNRs are lower by 0.05 than what Xu et al. [8] found using the TIMIT speech library. As we use the same noise types, and we were able to reproduce their 'noisy' STOI scores using TIMIT, this discrepancy shows that STOI predicts different intelligibility for the two libraries under equal noise conditions.

Figure 6 shows a plot of the average STOI scores obtained for the files processed for subjective evaluation. As with the validation set results, the use of GVN did not significantly affect model performance. At higher SNRs, DNN processing performs worse than the noisy baseline. However, for low SNRs STOI scores suggest improvement even outside the training range. According to the objective evaluation, DNN processing ought to be beneficial for all SNRs in between -14 dB and 4 dB.

3.2. Subjective evaluation

Figure 7 shows the results from the subjective tests. Specifically, it shows the differences between the reference and the two DNN models, both for the SRT and the slope of the psychometric function at SRT. All test subjects performed worse on the SRT, while the slope values are more mixed.

To assess the normality of the data, we performed an Anderson-Darling test on all the differences. The SRT differences for the DNN without GVN failed the normality test. The non-normality is presumably a consequence of the small sample size. To cope with this, we performed a Wilcoxon signed rank test to compare the models with the reference. The tests

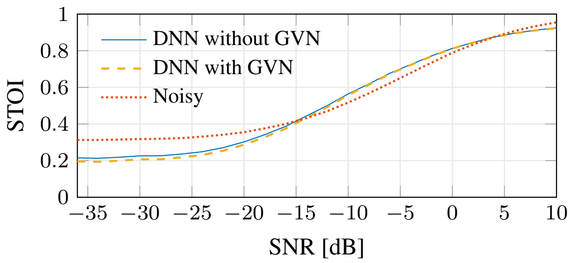


Figure 6: *STOI results for the subjective evaluation set*

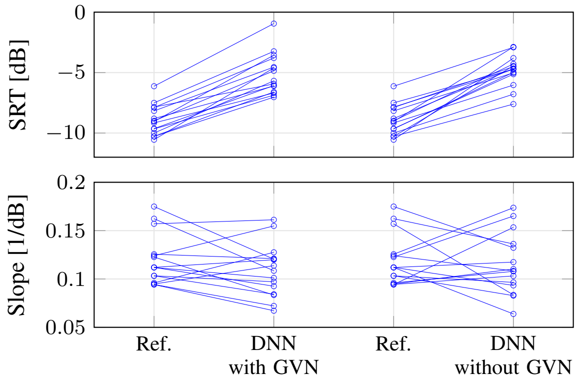


Figure 7: *Comparison between unenhanced reference data and DNN data. Upper: Speech recognition thresholds (SRT). Lower: Slope of the psychometric function at SRT.*

showed a significant difference ($W = 120, p = .0001$ for both) between the models and the reference; not surprisingly, since all the test subjects performed worse on the DNN models (see Figure 7). The differences in median SRT values were (using Hodges-Lehman estimators) 3.8 [3.2, 4.4] and 3.9 [3.2, 4.8] for DNN with GVN and without GVN, respectively. The numbers in brackets are the 95 % confidence intervals.

The slope of psychometric functions were compared using a two sample F-test. Neither DNN with GVN ($F_{14,14} = .91$, NS), nor DNN without GVN ($F_{14,14} = .69$, NS) showed any significant difference from the reference.

4. Discussion

The STOI results for unprocessed noisy validation files from the Norwegian database (Table 1) differ from those obtained for the TIMIT database by Xu et al. [8]. This complicates comparing model performance directly. However, the results are similar to those of Xu et al. in the sense that STOI improvement is arguably insignificant for SNRs of 10 dB and above. For lower SNRs, STOI predicts our system will achieve improvements of up to 6 percent on the subjective scale. This is less than Xu et al. achieved, but significant enough to predict that subjective SRTs ought to decrease, or at the very least, stay the same.

The DNN model was not trained at SNRs below -5 dB, but surprisingly, the STOI results shown in Figure 6 indicate that the model enhances noisy speech with SNRs up to 9 dB below its training range. This means that during subjective testing, 93.8 % of sentences presented to the listener had an SNR that fell in the functional range of the model (from -14 dB to 4 dB). All test subjects also achieved SRT values within this range. Nonetheless, the results from the subjective testing showed that the DNN models performed significantly worse (SRTs increased with approx. 4 dB) than the unprocessed sentences. Even from a conservative perspective where we could say that the changes the model attains in STOI are insignificant, the SRTs should not have increased this much. Thus, STOI significantly overestimates the speech intelligibility of our DNN-based speech enhancement system.

On the other hand, STOI correctly predicts that GVN has no significant effect on speech intelligibility. According to Xu et al. [8], PESQ results are, in contrast, significantly affected when GVN is used during postprocessing of a DNN-based speech enhancement system. This may indicate that GVN matters more to speech quality, but we did not investigate this further.

Our DNN model was selected because it obtained better STOI scores than similar networks trained for a larger range of SNRs or with different hyperparameters. Our results however indicate that STOI fails to predict the intelligibility of a DNN-based speech enhancement system. This directly undermines our model selection criterion. It is therefore possible that one of our other models would have lead to better subjective scores.

All test sentences were uttered by the same male speaker; it is likely that the DNN model will perform differently for different speakers. Similarly, the results are presumably affected by the choice of background noise. We expect that the traffic noise used here performs better than for example noise that consists mainly of human speech (babble), since the DNN models might try to enhance some of the speakers in the noise as well. Similarly, other types of noise may again be easier for the system to handle. A more comprehensive study of the suitability of STOI as an objective evaluation measure for DNN-based speech enhancement would need to include a variety of speakers and noises. Such a comprehensive study will be time-consuming and the material for the speech-in-noise tests will need to be carefully constructed for unbiased results.

The choice of sampling frequency (8 kHz) might also have affected the results. Increasing the sampling frequency to 16 kHz, or higher, would probably have improved the speech recognition for all the tests [23], but it is not clear if this would have changed the results of this study.

Another possible bias in this study is the effect of hearing loss. As the analysis of the subjective testing looked at the difference between a reference and the DNN models, we assumed that a hearing loss would not alter the results. Only one test subject had a hearing aid, but this was not used during the subjective test. Since the test subjects' ages were relatively high (mean $= 54.2$) it can be assumed that several of the test subjects were affected by presbycusis. Even if the intra-subject change in SRTs should be independent of hearing impairment, this may have affected results.

Our analysis is limited to speech intelligibility, and does not consider the effect of DNNs on speech quality. The relationship between these two parameters is not fully understood. For many communication systems, intelligibility may be approaching 100 %, while user satisfaction is still limited. Here, listening effort tests, where a speech intelligibility test is combined with another task, may provide a good compromise between providing objective results for the more quality related question of how comfortable or easy it is to listen to the enhanced speech.

5. Conclusion

We have tested a DNN-based speech enhancement system with listening tests to determine the subjective intelligibility of processed noisy speech. 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. Our results further show that the postprocessing technique of global variance normalisation does not significantly affect subjective intelligibility.

6. Acknowledgements

We thank Tor Andre Myrvoll for his guidance in setting up the speech enhancement system and valuable insights during discussions. We also thank our volunteer test subjects.

7. References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, 2013.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[3] M. Nielsen, *Neural Networks and Deep Learning*. Determination Press, 2015.

[4] Z.-Q. Wang and D. Wang, "A Joint Training Framework for Robust Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1–11, 2016.

[5] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, 2016.

[6] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *INTERSPEECH*, Singapore, 2014, pp. 616–620.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[8] ——, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[10] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, ITU-T Recommendation P.862, 2001.

[11] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, p. 3387, 2009.

[12] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7539284/

[13] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015, last accessed on 2017-06-01.

[14] Nasjonalbiblioteket, "NB Tale - a basic acoustic phonetic speech database for Norwegian," http://www.nb.no/sprakbanken/show?serial=sbr-31, 2015, last accessed on 2017-06-01.

[15] D. Pearce, H.-G. Hirsch, and others, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions." in *Interspeech*, 2000, pp. 29–32.

[16] Guoning Hu, "100 Nonspeech Sounds," http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html, last accessed on 2017-03-14.

[17] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.

[18] T. F. Chan, G. H. Golub, and R. J. LeVeque, "Algorithms for computing the sample variance: Analysis and recommendations," *The American Statistician*, vol. 37, no. 3, pp. 242–247, 1983. [Online]. Available: http://www.jstor.org/stable/2683386?origin=crossref

[19] The MathWorks, Inc., *MATLAB R2016a*. Massachusetts, United States: Natick, 2016.

[20] J. Øygarden, "Norwegian speech audiometry," Ph.D. dissertation, Norwegian University of Science and Technology (NTNU), Faculty of Art, Department of Language and Communication Studies, 2009.

[21] L. L. Kontsevich and C. W. Tyler, "Bayesian adaptive estimation of psychometric slope and threshold," *Vision research*, vol. 39, no. 16, pp. 2729–2737, 1999.

[22] N. Prins and F. Kingdom, "Palamedes: Matlab routines for analyzing psychophysical data." http://www.palamedestoolbox.org, 2009, last accessed on 2017-03-14.

[23] A. B. Silberer, "Importance of high frequency audibility on speech recognition with and without visual cues in listeners with normal hearing," Ph.D. dissertation, University of Iowa, 2014.