Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement

Femke B. Gelderblom, Tron V. Tronstad, Erlend Magnus Viggen

***Abstract*—Speech enhancement systems aim to improve the quality and intelligibility of noisy speech. In this study, we compare two speech enhancement systems based on deep neural networks. The speech intelligibility and quality of both systems was evaluated subjectively, by a Speech Recognition Test based on Hagerman sentences and a translation of the ITU-T P.835 recommendation, respectively. Results were compared with the objective measures STOI and POLQA. Neither STOI nor POLQA reliably predicted subjective results. While STOI anticipated improvement, subjective results for both models showed degradation of speech intelligibility. POLQA results were overall hardly affected, while the subjective results showed significant changes in overall quality, both positive and negative, in many of the tests. One of the systems was trained to remove all noise; a strategy that is common in speech enhancement systems found in the literature. The other system was trained to only reduce the noise such that the signal-to-noise ratio increased with 10 dB. The latter system subjectively outperformed the system that attempted to remove noise completely. From this, we conclude that objective evaluation cannot replace subjective evaluation until a measure that reliably predicts intelligibility and quality for deep neural network based systems has been identified. Results further indicate that it may be beneficial to move away from more aggressive noise removal strategies towards noise reduction strategies that cause less speech distortion.**

***Index Terms*—speech enhancement, artificial neural networks, subjective evaluation, speech intelligibility, speech quality**

I. Introduction

THE field of speech enhancement (SE) deals with improving speech signals that have been degraded by noise [1]. Speech enhancement is commonly applied in automatic speech recognition (ASR) systems as a preprocessing step to improve these systems' accuracy in noisy environments [2], [3], [4]. Recently, research into this application has flourished, resulting in significant performance increases of ASR systems. This success has also lead to a renewed interest in the application of speech enhancement for human listeners, where the goal is to make the speech easier to understand (i.e., increase *speech intelligibility*) and/or more comfortable and less tiring to listen to (i.e., increase *speech quality*) [3], [5], [6]. The latter application is especially important within the fields of telecommunication and hearing assistive technology.

There exists a wide range of SE techniques. As in many other fields, techniques based on deep neural networks (DNNs) [7], [8] are currently receiving a lot of interest due to their potential to outperform earlier techniques. For ASR systems, performance is measured by a SE system's ability to decrease the word error rate. For human listeners, performance is ideally determined through subjective evaluation of speech intelligibility and/or speech quality [1]. These tests generally compare the listeners' evaluations of noisy speech before and after enhancement, to quantify the effect of different SE strategies.

However, since these subjective evaluations are time-consuming to perform, *objective measures* are often calculated instead. These objective measures typically quantify a degraded speech signal in comparison to a clean speech signal. For speech intelligibility, a popular objective measure is STOI, which performs well against competing intelligibility measures [9] and has a reference implementation freely available [10]. For speech quality, popular measures are PESQ [11] and its successor POLQA [12]. Although PESQ also has a downloadable reference implementation [13], licenses must be purchased to use PESQ and POLQA.

When evaluating the change in intelligibility or quality obtained with SE systems, measures based on clean speech and *unenhanced* noisy speech are calculated to establish a reference. Then, the same measures are calculated for clean speech and *enhanced* noisy speech. Comparison of these results then predicts how much the SE system affects speech intelligibility or quality.

However, these objective measures have been designed to predict intelligibility or quality for relatively simple degradations, such as additive noise, and do not necessarily perform well for more complex degradations [9], [14], [15]. DNN-based SE systems perform a complex nonlinear processing of the noisy signal, and multiple authors have found that STOI is not a reliable predictor of whether or not a given DNN-based system actually improves speech intelligibility [16], [17], [18]. Until a specific objective measure has been shown to give reliable predictions for these systems, time-consuming subjective evaluations are required to test DNN-based SE systems.

Femke B. Gelderblom and Tron V. Tronstad and Erlend Magnus Viggen are with the Acoustics Research Centre, Connectivity Technologies and Platforms, SINTEF Digital, Trondheim, Norway. Erlend Magnus Viggen is additionally with the Centre for Innovative Ultrasound Solutions, Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, NTNU – Norwegian University of Science and Technology, Trondheim, Norway, and is supported by grant no. 237887 from the Research Council of Norway.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. The material consists of tables containing detailed results of the statistical analysis. Contact tronvedul.tronstad@sintef.no for further questions about this material.

When training a DNN using supervised learning techniques, we must always specify the format of the input and the target output. In speech enhancement, a common input is the logarithmic half-spectra of several adjacent semi-overlapping frames of the noisy speech signal, and a common target output is the logarithmic half-spectrum of one corresponding frame of the clean speech signal [5], [6], [17], [4], [19], [20], [21], [22], [23], [24]. In this way, the training process leads the DNN towards returning perfectly noise free speech. This approach has shown significant merit for application in ASR systems.

While SE systems often manage to reduce or even remove the presence of noise, the output speech is generally audibly degraded by this process, especially at low signal-to-noise ratios (SNRs), as SE systems cannot perfectly distinguish between speech and noise when attempting to remove only the latter [25], [24]. In fact, our previous study on one possible realization of a DNN-based SE system [17] found that this degradation significantly reduced the speech intelligibility, compared to that of the noisy speech before the enhancement was carried out. We found that the speech recognition threshold (SRT), which is the SNR at which 50 % of words are understood, degraded on median by around 4 dB, in stark contrast with the positive performance predicted by STOI.

However, this is not surprising, when put in the perspective that humans are very sensitive to degradation in speech signals, while capable of scoring 100 % intelligibility despite noisy conditions. This motivates studying training methods that look for a suitable compromise between noise reduction and speech degradation, in addition to methods that focus on finding noise free speech.

Supervised training of a DNN involves optimizing some statistical measure, called the loss function, which is based on the difference between the desired DNN output and the actual DNN output for a given input. A typical loss function in DNN-based SE is the mean-squared-error (MSE) value based on the target clean speech and the DNN's output. By iteratively adjusting the weights of the DNN to obtain a lower MSE, the training process moves the DNN's output towards the target output.

One way of shifting the "focus" of the DNN training towards speech and away from noise, in the hope of indirectly reducing speech degradation, would be to use more speech-aware loss functions. Kumar et al. proposed using a weighted squared error based on absolute thresholds of hearing, but did not report results that allow for direct performance of this loss function to a standard MSE approach [25]. Others investigated using STOI as a training target, but did not obtain improvements of such a degree that it is obvious that they will show in a subjective evaluation [26], [27], [28]. We also investigated a number of other options, such as an MSE loss function weighted according to the SII band importance weights [29] or gammatone weights inspired by the objective intelligibility measure by Dau et al. [30]. However, none of our unpublished pilot studies based on these approaches showed enough promise to warrant continuing with subjective testing on a larger scale.

Another alternative to guide the training process is to go away from using a noise free target. This article investigates using a DNN target output that is not perfectly clean speech; rather, it corresponds to the input signal at a 10 dB higher SNR. This target, which is closer to the input, ensures that noise is still significantly reduced relative to the speech, but in a less aggressive manner. This may reduce the overall speech degradation, and consequently increase the speech quality and intelligibility compared to the more aggressive clean-speech target where the noise reduction likely has a stronger negative impact on the speech [25], [24]. A 10 dB improvement in SNR is clearly perceptible, since it perceptually corresponds to a halving/doubling of the loudness of the noise/speech signal [31]. Even though intelligibility improvement rates have been shown to vary a lot between test situations (from 1 % per dB to 44 % per dB, with a mean value of 7.5 % per dB) [32], a 10 dB improvement of the SNR should always be clearly measurable in subjective testing. The optimal may both be higher (less noise) or lower (less distortion), but finding an optimal value of the target's SNR improvement is out of the scope of this study.

In the study reported in this article, we trained two DNN-based SE systems based on these two targets, as described in Section II-A. We subsequently generated a large number of sound clips where clean sentences were mixed with different background noises at various SNRs and enhanced with either of the SE systems, as described in Section II-B. Our test subjects (Section II-E) were asked to perform subjective evaluations of the speech intelligibility (Section II-C) and speech quality (Section II-D) of these clips. Additionally, we calculated STOI and POLQA scores for comparison (Section II-F). We provide our results in Section III and discuss them further in Section IV, before we conclude in Section V.

II. Method

A. Data and DNN setup

In this work, we used the same general DNN setup as in our previous work [17], which is loosely based on the system by Xu et al. [6] and implemented using Keras [33]. As the details are given in [17], we will only give the essentials here.

The clean speech for training and validation of the DNN was taken from the Norwegian speech audio dataset NB Tale [34]. This forms part of the Norwegian language library Språkbanken, and is set up similarly to the widely used English-language TIMIT dataset. 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 [35]. The clean speech was divided into training, validation, and test sets that did not overlap in either speakers or sentences, with 1932 sentences from 137 speakers in the training set and 816 sentences from 48 speakers in the validation set. We chose to use Norwegian primarily because of our access to Norwegian native speakers as test subjects. However, we expect our results to be transferable to e.g. English, as the two are closely related Germanic languages.

During training and validation, the input was based on noisy speech constructed by combining this clean speech with noises taken from the Aurora database [36], the NOISEX-92

database [37], and Guoning Hu's collection [38]. We chose the same 104 noises for training and 15 unseen noises for validation as Xu et al. Both sets contained both stationary and non-stationary noise sources. For each set, we combined every type of noise with every sentence in that set, giving us a total of 1984 hours of training data and 98 hours of validation data. For the input data, six different SNRs uniformly spaced from −5 dB to 20 dB were used during training. Before they were combined, the speech and noise signals were downsampled to 8 kHz, the lowest sampling rate among the noise types.

The input was constructed from single-sided log-power spectra of frames of this noisy speech. Each frame was found from a 256-sample (32 ms) Hann window of the time signal. Adjacent frames overlapped by 50 % in time. These windowed frames were Fourier transformed and redundant information above the Nyquist limit was discarded, giving a single-sided spectrum. Then, the log-power spectrum was found by taking the base-10 logarithm of the magnitude of each frequency bin. The final input vectors were found by stacking 21 such log-power spectra, based on the adjacent overlapping frames, after each other. The task of the DNN was to enhance only the middle frame, and the stacking thus provided the DNN with 160 ms of past context and 160 ms of future context.

When training the DNN, we used two different training targets, leading to two different DNN models:

- **Model 1:** Here, the training target was the single-sided log-spectrum of a frame of clean speech, unaffected by noise. This is the model we reported earlier [17].
- **Model 2:** Here, the training target was the single-sided log-spectrum of a frame of clean speech mixed with the exact same noise as in the input, but at a 10 dB higher SNR.

The loss function was a standard mean squared error between the DNN output and the training target.

In both models, the DNN was a simple feedforward network with three hidden layers in addition to the input and output layer. Each hidden layer used LeakyReLU activation functions. The models were trained with 50 % dropout in the hidden layers using the Adam optimizer. We trained a number of different candidate networks over the same ranges of hyperparameters for both models. The ranges included hidden layers with 1024, 2048, and 3072 units. The final network for each model was chosen as the best epoch of all the candidate networks, according to the STOI scores that we evaluated for the validation set at 0 dB SNR after every epoch. The resulting Model 1 used 2048 nodes per hidden layer and a learning rate of 10^{-5}, while Model 2 used 3072 nodes and a rate of 10^{-2}. The final epochs for Model 1 and Model 2 were the 8th and the 33rd epochs, respectively.

In order not to change the experimental procedure more than necessary, we picked Model 2 based on STOI scores in the same way as we picked Model 1 in [17]. However, as earlier work indicates that STOI is not a robust predictor of the intelligibility of DNN-based SE systems, as explained in Section I, this approach is hardly ideal as we cannot truly expect the maximum-STOI epochs to perform best in a subjective evaluation. However, given the relatively minor performance changes reported in [39], [26], [27], [28] and the fact that we merely used STOI for selection rather than as a training target (with an expected weaker effect), we do not expect that the approach with STOI as selection criteria will have had a major impact on our results. Until one or more objective measures are identified as a robust predictor of intelligibility and/or quality, determining the best epoch or the best hyperparameters will remain problematic, as subjective evaluations of sufficient precision are generally too time-consuming to be feasible for anything other than a final test of a trained system. While an extensive study into this topic is outside the scope of this article, Section IV does compare STOI and POLQA scores with subjective evaluations of speech intelligibility and speech quality, respectively.

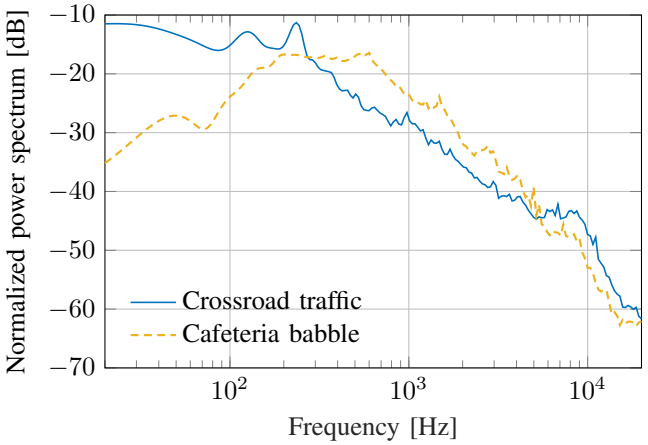

Fig. 1. Long-term average spectra of the two background noises used in the subjective evaluations. These spectra were computed using Welch's method using 2048-sample Hann windows, after the sounds had been normalized to have RMS values of 1.

When the trained network was used to enhance noisy speech, the process of reconstructing a waveform from the DNN output essentially consisted of reversing the steps used to create the input data. As the log-spectrum output does not contain phase information, we used the noisy input phase in this process. Unlike in our previous publication [17], we did not use the global variance normalization preprocessing step, for two reasons: We found then that it did not affect the results of the subjective intelligibility evaluation, and including it as a factor would double the already considerable number of tests to be performed by the test subjects.

B. *Generation of test sounds*

For the subjective evaluations, we generated a variety of single-channel clips of speech in noise at various SNRs. We generated clips both without enhancement and with enhancement by Models 1 and 2.

In all the clips, the base speech was a randomly generated five-word Hagerman sentence in Norwegian, generated as described by Øygarden [40]. Each sentence was built up the same way: [Name], [Verb], [Numeral], [Adjective], [Noun], with 10 possible options for each class of word. As a basis, we generated 500 reference speech clips of unique, noise-free sentences.

We then mixed these files with background noise at various SNRs, as described later in Sections II-D and II-C. Two different types of noise were used. We used one 17-second clip of traffic noise from a crossroad in Trondheim, and one 25-second clip of babble noise recorded in a university cafeteria during a lunch break. Neither type of noise was present in either the training or the validation described in Section II-A. Both noises were originally recorded with a sampling rate of 44.1 kHz. The long-term average spectra of both noises before downsampling is shown in Figure 1.

Finally, these unenhanced noisy clips were run through the two trained DNN models described in Section II-A. Thus, for each SNR, we ended up with 3000 unique degraded clips: The 500 reference clips, times two types of noise, times three types of enhancement (Unenhanced, Model 1, and Model 2). Figure 2 shows spectrogram examples of one speech clip at different points in this process.

C. Speech intelligibility

The speech recognition threshold (SRT), which is a common measure of speech intelligibility [1], was determined using the same method as in our previous work [17]. The five-word test sentences were built up from five word categories as described in Section II-B. The test subjects' task was to select the words they could hear using a graphical user interface with ten possible words per category, a total of 50 words. Guessing was allowed, but the test was not forced choice.

The test subject responses were given as input to an adaptive psychometric function estimation procedure called the Ψ-method [41], which continuously estimated the SRT during the test. The final threshold estimate was found after 20 sentences (i.e. 100 words in total). All parameters used in the method were identical to the ones used in our previous study [17], i.e. a guess and lapse rate of 0.01, psychometric function based on a cumulative normal probability density function, and stimulation range of the SNR from -36 dB to 10 dB in 2 dB steps.

The method was implemented in MATLAB [42] and the sentences were presented binaurally for all test subjects. An external sound card (Edirol UA-25) was connected with USB cable to the computer. Headphones (Howard Leight Sync Stereo Headband) with sound attenuating properties were used for the playback. The test was performed in an ordinary single room office with low background noise level. The background noise level was not measured during the test, but considering the headphones' sound attenuating properties and the signal levels involved, the results should not be affected.

Since the results from our previous study [17] did not pass the normality distribution assumption, we decided to use Wilcoxon tests to decide if differences were significant.

D. Speech quality

Speech quality was assessed using the method described in ITU-T P.835 [43]. The ordinal scales presented in the recommendation were translated to Norwegian by comparing and combining the official English and French version, together with a Danish version presented by [44]. The English and Norwegian versions can be seen in Table I and II respectively. Note that the Norwegian noise scale is slightly different than the English version. Instead of using "slightly noticeable", "noticeable but not intrusive" and "somewhat intrusive" as rating 4, 3 and 2, the Norwegian version uses "noticeable but not intrusive", "somewhat intrusive" and "intrusive". The reason for changing the scale was an observation made during a pilot test for the study. Several of the participants noted that it was difficult to distinguish between "slightly noticeable" and "noticeable but not intrusive". To cope with this problem, we adapted the French version [45], which uses a slightly different scale, in the translation.

TABLE I
ENGLISH VERSION OF THE ORDINAL SCALES USED IN ITU-T P.835 [43].

Rating	Speech	Noise	Overall quality
5	Not distorted	Not noticeable	Excellent
4	Slightly distorted	Slightly noticeable	Good
3	Somewhat distorted	Noticeable but not intrusive	Fair
2	Fairly distorted	Somewhat intrusive	Poor
1	Very distorted	Very intrusive	Bad

TABLE II
NORWEGIAN TRANSLATION OF THE ORDINAL SCALES USED IN ITU-T P.835.

Rating	Speech	Noise	Overall quality
5	Ikke forvrengt	Ikke hørbar	Veldig god
4	Litt forvrengt	Hørbar, men ikke påtrengende	God
3	Ganske forvrengt	Litt påtrengende	Middels
2	Betydelig forvrengt	Påtrengende	Dårlig
1	Voldsomt forvrengt	Veldig påtrengende	Veldig dårlig

Three different signal to noise ratios (SNRs) were tested for both noise types; 0 dB, 10 dB, and 20 dB. Each combination of noise type, SNR, and enhancement (including unenhanced clips) was tested twice for different sentences, giving 36 sentences per test subject. As the subjects were asked to rate the speech, noise, and overall quality of each sentence, each subject made a total of 108 evaluations. The sound playback and the test environment was the same as in the speech intelligibility test described in Section II-C.

All participants were given an instruction before starting the test and they were allowed to adjust the sound volume to their preferred level. They were also presented examples of the sounds to be used in the test. These examples were randomly taken from all the available sentences, and they were presented to give the test participants some idea of what to expect during the test.

Since it is not certain that the rating scale used has equal steps size between all ratings (i.e. it is not necessary the case that the size of the quality change going from 5 to 4 is the same as when going from 2 to 1), an ordinal scale analysis was

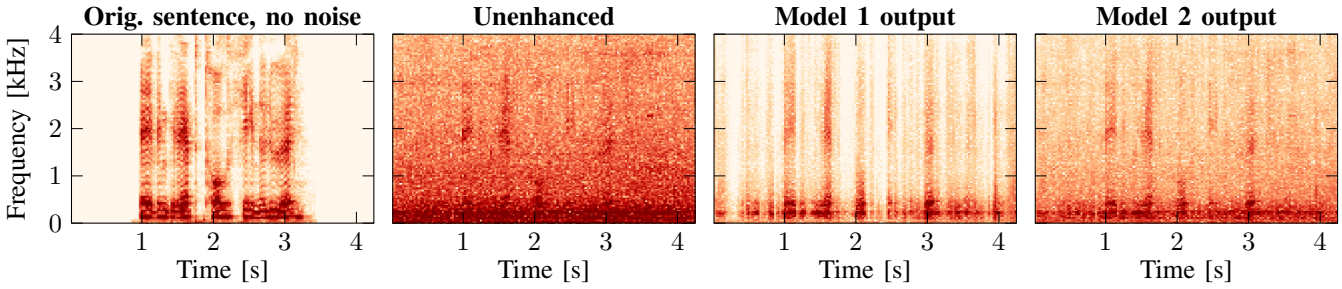

Fig. 2. Spectrograms of the utterance *"Eivind grep tolv fine luer"*, before and after added noise from crossroad traffic at SNR $= -5$ dB, and after enhancement by Model 1 and 2. Each spectrogram is plotted with a dynamic range of 50 dB.

performed to evaluate the results. A cumulative link model (*clm*) from the *ordinal* package [46] in R [47] was used to determine if the models were significantly different from the reference without SE.

E. Test subjects

The speech recognition test was performed by 12 persons, from 40 to 66 years of age (mean value 53.1). These individuals were a subset of the 15 participants from the listening test in our previous study [17]. It is assumed that the learning effect is large for the SRT test, therefore we used the same participants as last time to reduce the time needed for training.

23 persons attended the speech quality test, 8 females and 15 males, from 38 to 74 years of age (mean value 54.7). None of the listeners had performed any subjective listening tests within the last three months.

F. Objective measures

While the subjective evaluations described in Sections II-D–II-E give us the ground truth, it is still interesting to compare these results with those of objective measures. This comparison gives us more information about the reliability of the tested objective measures for DNN-based SE systems. In this work, we calculated the intelligibility measure STOI using the STOI reference code [10] and the quality measure POLQA using the implementation in the software Voice Quality Testing by GL Communications Inc. [48]. Even though PESQ has previously been used as an objective measure for the speech quality of DNN-based SE systems [5], [6], we chose to evaluate its successor POLQA due to licensing rights.

The STOI measures were calculated using the same files as in the speech intelligibility test, with SNRs from -36 dB to 10 dB in 2 dB steps. As a preprocessing step before the STOI calculations, the reference clips and degraded clips were upsampled from 8 kHz to 10 kHz. The POLQA measures were calculated from the same files used in the speech quality test, namely with SNRs of 0 dB, 10 dB, and 20 dB. The POLQA scores were calculated with the High Accuracy and Level Alignment modes activated.

III. Results

A. DNN output

The most basic way to analyze model performance is by investigating the error between the target output and the actual output. Figure 3 shows the mean and the standard deviation of the error (the difference between the target output in dB and the actual output in dB) for both models at various SNRs of the input. These statistics were calculated over each frame of the validation set. Frames where the speech signal was silent are excluded from these statistics. This means that the leftover noise shown in 2 during non-speech periods is not included in the error analyses.

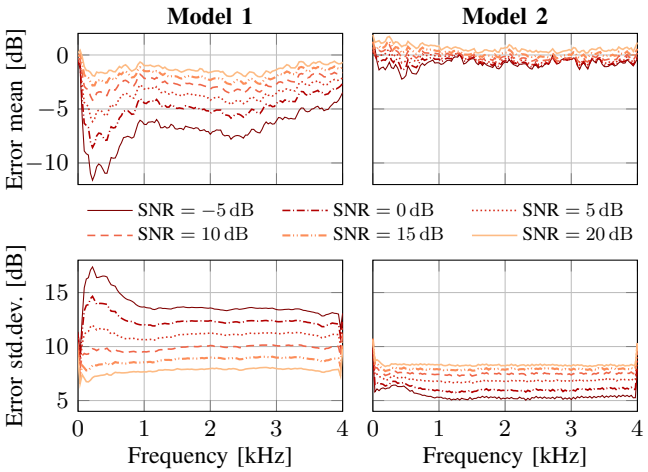

Fig. 3. Statistics for the outputs of the two models, calculated from the difference between the models' target outputs in dB and the actual outputs in dB

We find that Model 2 generally hits its target much better (less biased and with lesser spread) than Model 1 does. This does not necessarily tell us that Model 2 outperforms Model 1 as a SE system, only that it is better at achieving its given task. We also see that Model 1 has a large negative mean error that increases with decreasing SNRs. This shows that its predicted "enhanced" output is higher than the noise-free target output, which indicates that there is still quite a lot of noise left in the output, and that this becomes increasingly true with worsening SNR. For Model 2 the statistics depend less on SNR, indicating that the Model 2 task difficulty is more similar for low and high SNRs than it was for Model 1. Indeed, the standard deviation results show the opposite behaviour with respect to SNR as the Model 1 results did. Model 2 shows less spread (i.e., performs its task with higher accuracy) at lower SNRs. Although this might seem counter-intuitive at first (a SE

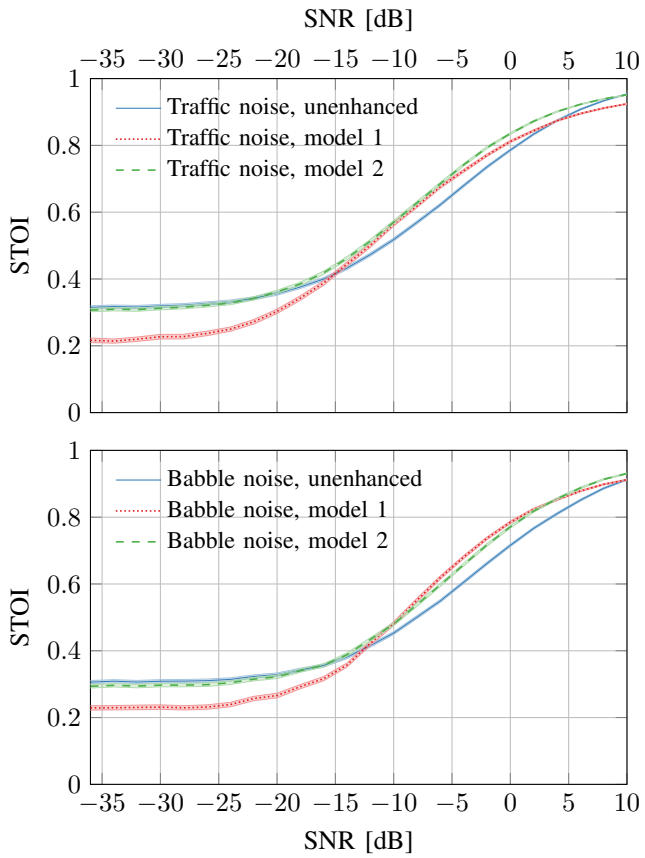

Fig. 4. STOI scores for crossroad traffic noise (upper) and cafeteria babble noise (lower). The lines indicate mean scores, and the shaded areas indicate approximate confidence intervals for the mean scores.

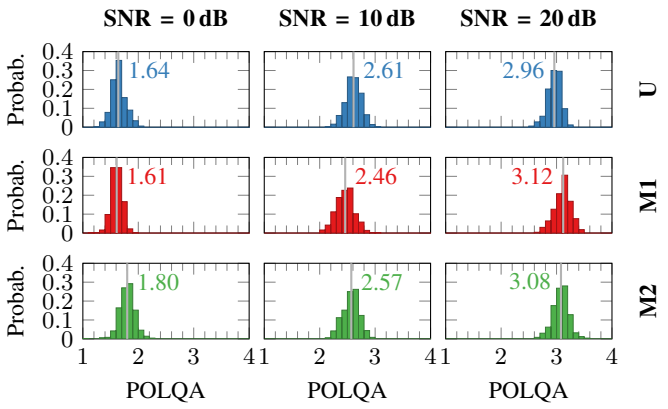

Fig. 5. Histograms of the relative probability of POLQA scores for clips with crossroad traffic noise over three different SNRs. The clips were either unenhanced (U), or enhanced with Model 1 (M1) or Model 2 (M2). The vertical gray lines and numbers represent median values.

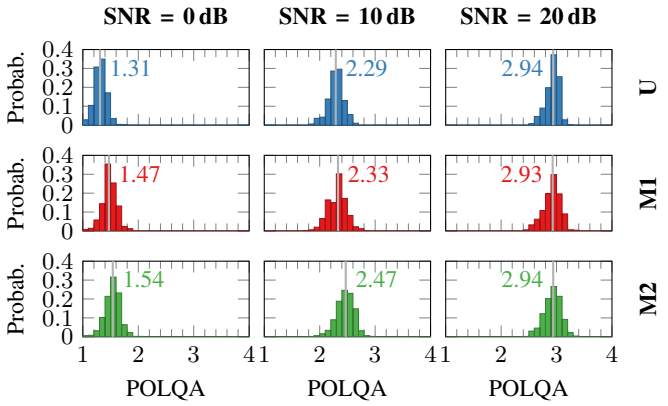

Fig. 6. Histograms of the relative probability of POLQA scores for clips with cafeteria babble noise over three different SNRs. The clips were either unenhanced (U), or enhanced with Model 1 (M1) or Model 2 (M2). The vertical gray lines and numbers represent median values.

model is generally not expected to do better at worse SNR), it makes sense from the perspective that in a situation with a lot of noise, it is easier for this noise to be identified and as such easier to be reduced by 10 dB.

B. Objective measures

As described in Section II-B, 500 clips were available from each combination of SNR, noise, and enhancement, i.e., one clip for each of the 500 original clean speech clips. Thus, we could use these various clips to calculate statistics for STOI and POLQA scores for each of these combinations.

The mean values of the STOI scores are shown as lines for each type of noise and enhancement in Figure 4. Additionally, as the STOI scores of the 500 clips for each combination of SNR, noise, and enhancement were approximately normally distributed, we calculated approximate confidence intervals for these mean values, which are also shown in Figure 4. Due to the high number of clips, the confidence intervals of the various enhancements are quite small and seldom overlap with the means of the other enhancements. Thus, the STOI values unambiguously rank the three enhancements for most SNRs.

The smaller number of SNRs where we calculated POLQA scores allows us to show the scores' distribution in more detail, through the histograms in Fig. 5 and Fig. 6. The median scores are shown as lines together with the median value.

Even if the distribution of the POLQA scores differ between the models with varying skewness and variance, the statistical analysis of the differences was performed using a two sample t-test. The t-test assumes normally distributed data, but it has been shown that for large sample sizes, the t-test might be more robust than the non-parametric tests when the data are a continuous variable [49]. While the mean value might not be the best descriptor for the data, the test does gives a good indication of whether the results differ or not. Note that the median has been used in the illustration if Fig. 5 and Fig. 6 as this is a slightly better descriptor for skewed data. Table III shows the results from the test performed with the function *t.test* in R. An F-test to compare variances was also performed (not shown) and used to decide if pooled variance should be used in the t-test.

C. Subjective speech quality

The results from the speech quality test are illustrated in Fig. 7 and Fig. 8. For more details about the statistical analysis the reader is referred to the supplementary material provided online. The setup for each figure is the same,

TABLE III
RESULTS FROM TWO-SAMPLE T-TEST PERFORMED ON THE POLQA SCORES FOR THE UNENHANCED SIGNAL (U), MODEL 1 (M1) AND MODEL 2 (M2). THE CONFIDENCE INTERVAL (95 % CI) MEANS THE CHANGE IN MEAN POLQA SCORE FOR THE MODELS BEING COMPARED.

Noise	SNR	Comparison	p-value	t	df	95 % CI	
Traffic	0 dB	U→M1	< .001	−4.8772	996	[−0.021	−0.050]
		U→M2	< .001	18.641	994.49	[0.139	0.172]
		M1→M2	< .001	25.336	998	[0.177	0.206]
	10 dB	U→M1	< .001	−15.591	998	[−0.168	−0.131]
		U→M2	< .001	−4.824	994.51	[−0.061	−0.026]
		M1→M2	< .001	10.752	998	[0.086	0.125]
	20 dB	U→M1	< .001	17.121	998	[0.129	0.163]
		U→M2	< .001	14.017	998	[0.101	0.134]
		M1→M2	.002	−3.1375	996.94	[−0.047	−0.011]
Babble	0 dB	U→M1	< .001	22.321	996	[0.155	0.184]
		U→M2	< .001	28.238	996	[0.209	0.241]
		M1→M2	< .001	6.5668	991.34	[0.039	0.072]
	10 dB	U→M1	.003	2.9975	998	[0.010	0.048]
		U→M2	< .001	17.891	998	[0.151	0.189]
		M1→M2	< .001	13.84	996.28	[0.121	0.161]
	20 dB	U→M1	.1528	−1.4309	998	[−0.029	0.005]
		U→M2	.8311	−0.213 34	998	[−0.019	0.015]
		M1→M2	.2807	1.0793	998	[−0.009	0.030]

presenting the different quality assessments horizontally, and different SNRs vertically. The bins consist of three groups; the unenhanced reference, Model 1, and Model 2. Each plot also indicates the significance and the direction of the change in score when going from the unenhanced signal (U) to the DNN models (M1: Model 1, M2: Model 2), as well as similarly indicating the change when going from M1 to M2. The changes' significance is indicated by asterisks, and the changes' direction is indicated with arrows. We cannot show the changes' magnitude, as the statistical test we used does not provide this information.

Both models have a negative effect on the quality of the *speech*. All the tested situations have a significant shift in the negative direction, i.e. the speech is more distorted. However, we can see from the M1→M2 comparison that Model 2 does not distort the speech as much as Model 1. This improvement in speech quality from M1 to M2 is significant ($p < .01$).

The *noise* is reduced for both models and all cases except 20 dB SNR have significant differences. For 20 dB SNR, the noise is generally evaluated as "noticeable, but not intrusive".

The *overall* quality results are more mixed. Model 1 does significantly worse for 10 dB and 20 dB SNR for both noise types, and does not have any significant difference for 0 dB SNR. The quality for the latter is not good, however, with score one ("very bad") as the most probable outcome. Model 2, on the other hand, does not have any significant differences in *overall* quality, except for 0 dB SNR with traffic noise, where there is a significant *positive* effect. The overall quality shifts from approximately equal probability for score one and two, to a most probable outcome at score two. Model 2 performs significantly better in all overall quality scores compared to Model 1 ($p < .05$).

D. Speech recognition threshold

The results from the speech recognition test are presented in Fig. 9. Each line represents results from one test subject. We should point out that the "old" reference data from our previous study [17] are similar to the ones in this study. Comparing the two reference results, using an Wilcoxon rank sum test (also known as an independent two-group Mann-Whitney U test), did not show any significant difference (Median $U_{\text{old ref}} = -9.07$ dB ($n_1 = 15$), Median $U_{\text{new ref}} = -9.14$ dB ($n_2 = 12$), $W = 89$, $p = .98$).

All the differences between the reference and the models were tested using a Wilcoxon signed rank test. Table IV shows the test statistics, and also show that all differences are significant ($p < .05$). The median and confidence interval values have been calculated using Hodges-Lehman estimators.

We also compared the two models, and the results can be seen at the bottom of Table IV. The difference between the results for the traffic noise was compared using a Wilcoxon rank sum test since the two data sets had different number of samples. For the babble noise a Wilcoxon signed rank test was used. Again, Model 2 performs significantly better than Model 1 for both noise types. The estimated improvement of the SRT from M1 to M2 is 3.0 dB for traffic noise, and 1.9 dB for babble noise.

IV. DISCUSSION

Model 1 and Model 2 were given different tasks. Where Model 1 was trained to remove noise, Model 2 was trained to only reduce noise such as to improve the SNR by 10 dB.

For both models, we used the noisy phase of the original signal during speech synthetization. Such a noisy phase may be expected to be better suited to the "less noisy" signal (from Model 2) than the "clean" signal (from Model 1) as the former

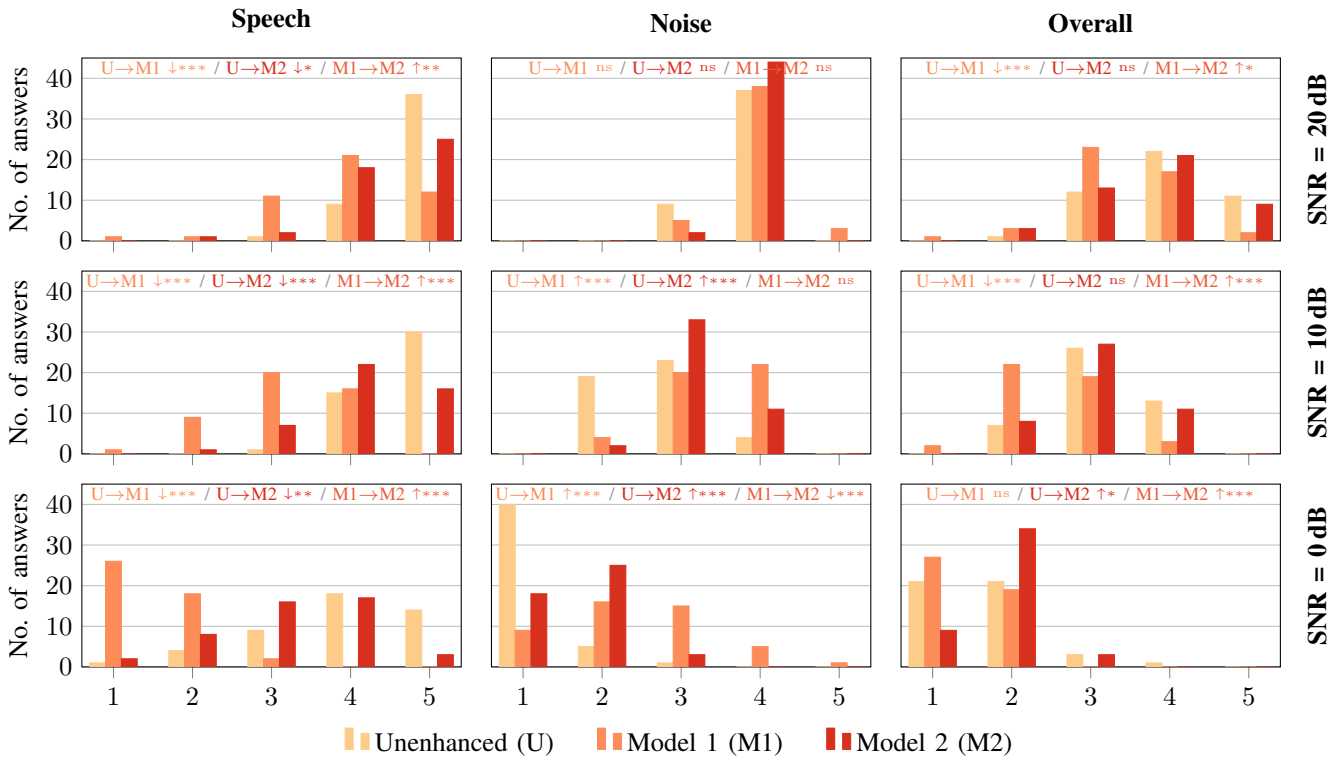

Fig. 7. Speech quality results for the speech (left), noise (middle) and overall (right) evaluation from the ITU-T P.835. The results are from the crossroad traffic noise at three different SNRs; 0 dB (lower), 10 dB (middle), and 20 dB (upper). U, M1, and M2 represent unenhanced, Model 1, and Model 2 respectively in the notation. Three stars (***) indicate $p < .001$, two stars (**) indicate $p < .01$, one star (*) indicates $p < .05$, and *ns* means "not significant". The arrows beside the stars indicate the direction of change.

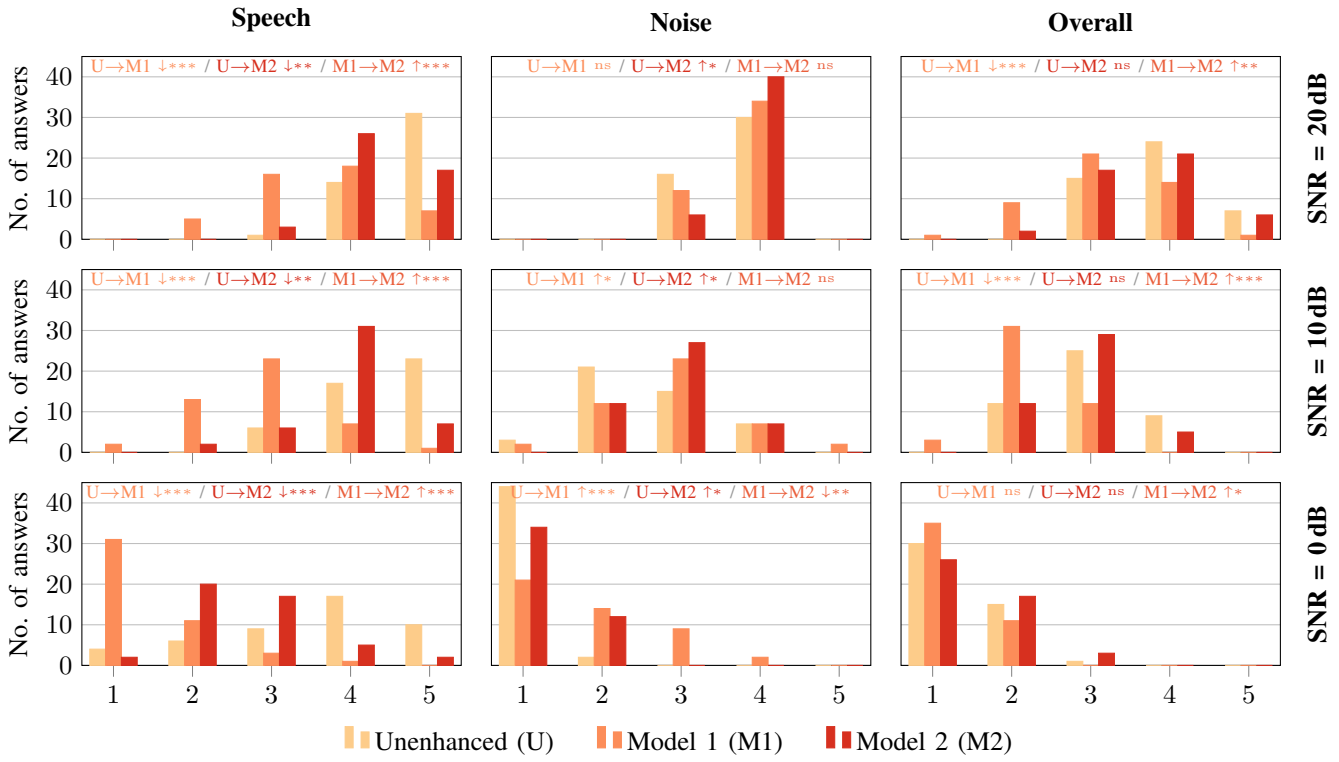

Fig. 8. Speech quality results for the speech (left), noise (middle) and overall (right) evaluation from the ITU-T P.835. The results are from the cafeteria babble noise at three different SNRs; 0 dB (lower), 10 dB (middle), and 20 dB (upper). U, M1, and M2 represent unenhanced, Model 1, and Model 2 respectively in the notation. Three stars (***) indicate $p < .001$, two stars (**) indicate $p < .01$, one star (*) indicates $p < .05$, and *ns* means "not significant". The arrows beside the stars indicate the direction of change.

TABLE IV
SPEECH RECOGNITION THRESHOLD STATISTICS FROM THE ANALYSIS OF THE RESULTS.

Noise	Comparison	n	p-value	V	Median	95 % CI
Traffic	$U_{old} \rightarrow$ M1	15	$< .001$	120	3.9 dB	[3.2, 4.8]
Traffic	U→M2	12	.002	75	1.4 dB	[0.7, 2.0]
Babble	U→M1	12	$< .001$	78	2.4 dB	[1.7, 3.2]
Babble	U→M2	12	.01	70	0.6 dB	[0.2, 1.1]
Traffic	M1→M2	15(12)	$< .001$	161^+	−3.0 dB	[−3.9, −1.6]
Babble	M1→M2	12	.002	75	−1.9 dB	[−2.7, −1.1]

$^+$: Comparison was done with Wilcoxon rank sum test since the data sets had different number of samples. The number is the observed rank sum W.

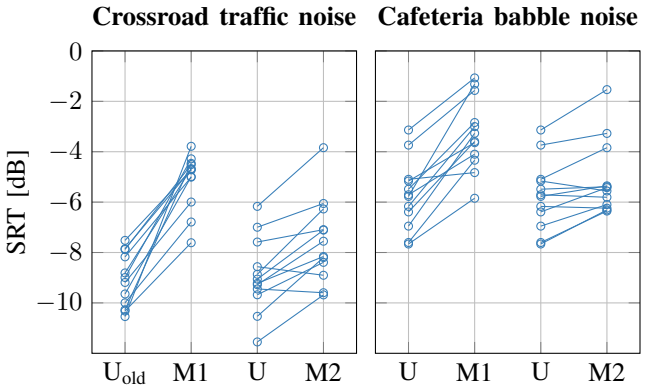

Fig. 9. Speech recognition threshold results for traffic noise and babble. The connected lines represent results from each of the test subjects from unenhanced clips (U) to clips enhanced Model 1 (M1) or Model 2 (M2). For crossroad traffic the results (U_{old} and M1) are taken from our previous study [17].

is closer to the original input from which the noisy phase was taken. Thus, different performance could possibly be the result of better/worse suitability with respect to the speech synthetization process. This in itself would be an advantage of the approach taken in Model 2: After all, the noisy phase is always readily available, whereas a clean phase would have to be approximated. However, the mean and standard deviation results presented in Figure 3 show that there is more going on. First of all, from the rather large standard deviations obtained for Model 1, one can easily argue that the resulting signal is far from "clean", and as such a clean phase won't be optimal either. Also, Model 2 performs better at its given task than Model 1: The fact that the standard deviation of the difference between targeted and obtained output is smaller, shows that the model is more accurate at reducing noise rather than Model 1 is at removing it. There is also a marked lower dependence on SNR, and the model is actually more accurate at noise reduction when the SNR gets worse. This indicates that a DNN-based SE system does indeed have less trouble with reducing noise than with removing it, making the approach worthy of investigation so long as systems aiming to remove noise entirely do not achieve ideal results.

Both models were trained with an equal variety of hyperparameters, and in each case the model with the best STOI score was selected for further subjective testing. This selection method resulted in Model 2 having 3072 nodes per hidden layer, where Model 1 only had 2048 nodes per hidden layer. As such, Model 2 has a larger capacity than Model 1, and one may argue that any differences in the results may be (partly) due to this difference, rather than the difference in noise removal/reduction strategy. However, the statistical results (not reported in this article) akin to those presented in Figure 3 of a model equal to Model 1 but with 3072 nodes per hidden layer, show the same behaviour as the chosen Model 1. During hyperparameter optimization, we also noticed that the lowest MSE obtained for models with a noisy target was generally much lower than for models with a clean target. Given this, we are confident that any performance differences obtained are not due to the different capacities of the model, but due to the different noise cleaning strategies.

As in our previous study [17] the SE did not improve the speech intelligibility. Even if STOI predicted a slight improvement for both models in the SNR range of interest, our subjective evaluation showed that both models did significantly worse than the unenhanced signal. However, Model 2 performed significantly better with respect to speech intelligibility than Model 1 for both noise types, by 3.0 dB and 1.9 dB for traffic and babble noise respectively. Compared to the unenhanced signal, however, it still has an elevated speech recognition threshold.

For the DNN models used in this study, calculated STOI scores were used to select a final model from model candidates over different sets of hyperparameters and different training epochs. The results show that this approach might not be justified as STOI does not seem to be a good predictor in our case. This means that we may have trained other models that could have performed better in our subjective evaluations, but how to identify these models is as of yet an unsolved problem.

Even though we have shown that our selected DNN-based SE systems did not end up actually improving speech intelligibility, we should point out that other authors have trained DNN-based systems that improve intelligibility to human listeners [16], [50], [51]. Our results do by no means provide evidence that DNN-based SE is not a generally promising approach worthy to be further investigated.

In addition to the speech intelligibility test, this study also evaluated the quality of the signal using the ITU-T P.835 recommendation. The results show that the models did not give a general improvement of the *overall* quality of the signal. No significant change to overall quality was found in 7 out of 12 comparisons of unenhanced and enhanced signals, and

Model 1 did actually significantly ($p < .001$) reduce the overall quality in four of the six tests performed. The only exception was for traffic noise at 0 dB SNR, where Model 2 did significantly ($p < .01$) better than the unenhanced signal.

For the evaluation of the quality of the *noise* separately, the results were as expected. The models were trained to reduce the background noise, and the results verify that they achieve this in 9 out of 12 comparisons. Only the situation with the highest SNR (20 dB), where the noise is already rated as "noticeable but not intrusive", does not show significant improvement by both models. (This comes as no surprise, as it is difficult and arguably unnecessary to improve upon a situation that already does not bother listeners.) Note that the highest noise score, "not noticeable", is almost never used. This may indicate that the step size from score four to five on the noise scale is large, and that it is difficult to show minor improvements of low-noise signals on this scale.

Another observation is that Model 2 performs similarly to Model 1 with respect to noise reduction, except at the lowest SNR (0 dB). This is surprising, since Model 2 does not try to remove the noise, only reduce it. It is, however, supported by the fact that a SNR of 30 dB often is referred to as "effective clean speech", and that people have little benefit of improving the SNR beyond 20 dB. This suggests that it might be beneficial to use variable training targets, with little noise reduction for the signals with high SNR, and progressively more reduction as the SNR gets worse. A common training target at 20 dB SNR could be a possible solution.

The evaluation of the *speech* also comply with the results from previous studies on noise reduction. Reducing noise will, in most cases, also add distortion to the speech signal. While Model 2 does perform better than Model 1 in all cases, it still does add distortion to the speech.

Objective POLQA scores were compared to the overall quality results from the subjective test to see if similar traits could be found. The general impression is, however, that POLQA does not predict the overall quality results from the ITU-T P.835 test. Even if we found significant degradation in quality for Model 1 compared to the unenhanced signal, POLQA did not show a consistent correlation. The POLQA scores were, in general, very similar within each SNR, and the largest difference found was below 0.25. Even if this is more than the theoretical accuracy for POLQA [52], such a small difference would be very difficult to detect in a subjective test. The subjective results does, however, show a significant degradation of the overall quality for Model 1, while POLQA actually shows a minor improvement in half of these situations.

Since the ITU-T P.835 recommendation was not available in a Norwegian version, the quality assessment scales were translated for this study. During the pilot test it was revealed that the initial translation was confusing for the test subjects. Several participants found it hard to differentiate between the noise being "slightly noticeable" and "noticeable but not intrusive". To solve this, we used a slightly different wording, closer to the French version of the recommendation [45]. Hence it might be difficult to compare the noise scores in this paper with other results performed with the English scale. The translation of the overall score labels might also affect the (lack of) correlation with POLQA, but this minor textual change to the scale cannot explain why the POLQA scores and the subjective results are opposite for many of the tested situations.

Another limitation of the study is the spoken material used in the test. All the sentences used, both for the intelligibility and quality test, were uttered by the same male speaker. Strictly speaking, this means that the validity of the results are limited to this speaker, and it might be possible that the models could perform better for other speakers.

Similarly to our previous study [17] the sampling frequency used was 8 kHz. This might affect the results since much high-frequency information that might be important both for speech intelligibility and quality assessment are lost. It is, however, not obvious that an increased sampling frequency would have affected the comparisons in this study since they were all done using the same sampling frequency.

In this study two different background noises were used, traffic noise from a busy crossroad and cafeteria babble. The results showed similar improvements for the two noise types, but it is possible that other types of noise could have given different results. The SRT results are otherwise in accordance with what we expect; it is more difficult to understand speech in babble noise than in traffic noise.

Another possible bias is the effect of hearing loss. The average age of the test subjects was relatively high, hence it is expected that age-related hearing loss could be a problem. None of the participants reported any problems with their hearing, or wore hearing aids, but this does not mean that they do not have an elevated hearing threshold. Such elevation could have affected the results, especially the speech intelligibility, which is known to deteriorate with increasing hearing loss. Since all the comparisons were done within each subject, it is expected that an improvement (or deterioration) of the signal would affect both those with normal hearing and those with hearing loss. It is, however, possible that a speech enhancement is perceived differently for individuals with or without hearing loss.

V. Conclusion

In this study, we compared two similar speech enhancement systems based on deep neural networks. The first system, Model 1, was trained with the target of removing all noise from a noisy speech signal, as was done in previous studies [5], [6], [17]. The second system, Model 2, was trained with the target of improving the noisy signal's signal-to-noise ratio by 10 dB.

A subjective evaluation of speech quality in terms of speech degradation, noise intrusiveness, and overall quality showed some interesting similarities and differences between the two models. From the evaluation of overall quality, Model 2 represents a significant improvement to Model 1 in all six situations tested. Both models significantly reduced the noise intrusiveness except at the highest SNR of 20 dB, with Model 1 outperforming Model 2 only at the lowest SNR of 0 dB. While both models significantly distort the speech at all SNRs, Model 2, with its less agressive training target, distorts speech

to a significantly smaller degree than Model 1 at all SNRs. This reduction in distortion may be the reason why Model 2 outperforms Model 1 by 2–3 dB in a subjective evaluation of speech intelligibility in terms of the speech recognition threshold.

For these reasons, we believe that using less aggressive training targets in DNN-based SE systems, along the lines of our Model 2, is a promising approach that warrants further investigation. However, we must point out that if we compare our subjective evaluation results for the noisy speech enhanced by Model 2 and the unenhanced noisy speech, we find that Model 2 does not perform a general improvement to the signal. Model 2 actually degrades the speech intelligibility slightly, raising the speech recognition threshold by around 1 dB. It however did make a significant improvement to the overall quality in one of the six situations tested, while not affecting performance in a statistically significant manner in the other five situations.

In order to train better DNN-based SE systems than the ones presented here, it is absolutely essential to be able to distinguish between a good system and a bad one without having to run a complete subjective evaluation, as these are prohibitively time-consuming. However, our results comparing the subjective evaluations with the objective measures STOI and POLQA indicate that these measures are not appropriate for this purpose. We found that the STOI results predicted significant improvements in intelligibility for our DNN-based SE systems while the subjective evaluations found significant reductions. We also found that the weak changes in POLQA scores failed to predict the significant changes in speech quality found by the subjective evaluations. Therefore, we must advise against solely using STOI and/or POLQA to evaluate DNN-based SE systems, either for the purpose of choosing which trained model candidate to proceed with, or for the purpose of evaluating the final system in the place of a subjective evaluation.

The studied systems are relatively simple implementations of DNN-based SE. As such, their speech enhancing ability is limited, even as indicated by objective measures. However, there is no reason to assume that there will not also be a mismatch between objective and subjective results in better and/or more complicated DNN-based SE systems. Indeed, similar mismatches have also been found elsewhere [16], [18].

Thus, we believe that we have pointed out an important issue that impedes progress for DNN-based SE systems for direct human applications like in telecommunication and hearing assistive devices. To resolve this issue, we believe that it is essential to identify or develop an objective measure that correlates well with intelligibility and/or quality even for channels with the complex nonlinear degradations that processing with a DNN-based SE system can cause. A dedicated study on this topic should be carried out.

REFERENCES

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

[2] Z.-Q. Wang and D. Wang, "A Joint Training Framework for Robust Automatic Speech Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, pp. 1–11, 2016.

[3] 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 J. Adv. Signal Process.*, vol. 2016, no. 1, 2016.

[4] 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 2014*, Singapore, 2014, pp. 616–620.

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

[6] ——, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 1, pp. 7–19, 2015.

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

[8] F. Chollet, *Deep Learning with Python*. Shelter Island, New York: Manning Publications Co, 2017.

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

[10] C. H. Taal. STOI – Short-Time Objective Intelligibility Measure. [Online]. Available: http://www.ceestaal.nl/code/

[11] "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.

[12] "Perceptual objective listening quality assessment," International Telecommunication Union, ITU-T Recommendation P.863, 2014.

[13] "Recommendation P.862," https://www.itu.int/rec/T-REC-P.862.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "An evaluation of objective quality measures for speech intelligibility prediction." in *Interspeech 2009*, 2009, pp. 1947–1950.

[15] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[16] E. W. Healy, S. E. Yoho, J. Chen, Y. Wang, and D. Wang, "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," *J. Acoust. Soc. Am.*, vol. 138, no. 3, pp. 1660–1669, 2015.

[17] F. B. Gelderblom, T. V. Tronstad, and E. M. Viggen, "Subjective Intelligibility of Deep Neural Network-Based Speech Enhancement," in *Interspeech 2017*. ISCA, 2017, pp. 1968–1972.

[18] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 1, pp. 153–167, 2017.

[19] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *2014 12th Int. Conf. on Signal Process. (ICSP)*. IEEE, 2014, pp. 473–477.

[20] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Cross-language transfer learning for deep neural network based speech enhancement," in *9th Int. Sym. on Chinese Spoken Lang. Process., 2014. ISCSLP-14*. Singapore, Singapore: IEEE, 2014, pp. 336–340.

[21] X. Xiao, S. Zhao, D. H. Ha Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, 2016.

[22] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *9th Int. Symp. on Chinese Spoken Lang. Process. (ISCSLP)*. IEEE, 2014, pp. 250–254.

[23] ——, "Deep neural network based speech separation for robust speech recognition," in *2014 12th Int. Conf. on Signal Processing (ICSP)*. IEEE, 2014, pp. 532–536.

[24] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments," *Speech Commun.*, vol. 95, pp. 28–39, 2017.

[25] A. Kumar and D. Florencio, "Speech Enhancement in Multiple-Noise Conditions Using Deep Neural Networks," in *Interspeech 2016*, San Francisco, USA, 2016, pp. 3738–3742.

[26] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Monaural Speech Enhancement using Deep Neural Networks by Maximizing a Short-Time Objective Intelligibility Measure," *arXiv:1802.00604 [cs, eess]*, 2018.

[27] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually Guided Speech Enhancement Using Deep Neural Networks," in *ICASSP 2018*. Calgary: IEEE, 2018, pp. 5074–5078.

[28] H. Zhang, X. Zhang, and G. Gao, "Training Supervised Speech Separation System To Improve STOI and PESQ Directly," in *ICASSP*, 2018, pp. 5374–5378.

[29] "Methods for Calculation of the Speech Intelligibility Index," American National Standards Institute, Tech. Rep. ANSI S3.5-1997, 1997.

[30] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective"signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[31] S. S. Stevens, "The Measurement of Loudness," *J. Acoust. Soc. Am.*, vol. 27, no. 5, pp. 815–829, 1955.

[32] A. MacPherson and M. A. Akeroyd, "Variations in the Slope of the Psychometric Functions for Speech Intelligibility: A Systematic Survey," *Trends Hear.*, vol. 18, p. 233121651453772, 2014.

[33] F. Chollet, "Keras," GitHub, 2015.

[34] Nasjonalbiblioteket, "NB Tale - a basic acoustic phonetic speech database for Norwegian," 2015.

[35] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, p. 3387, 2009.

[36] 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*, 2000, pp. 29–32.

[37] 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 Commun.*, vol. 12, no. 3, pp. 247 – 251, 1993.

[38] Guoning Hu, "100 Nonspeech Sounds," http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

[39] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, 2018.

[40] J. Øygarden, "Norwegian speech audiometry," Ph.D. dissertation, Norwegian University of Science and Technology, 2009.

[41] N. Prins and F. Kingdom, "Palamedes: Matlab routines for analyzing psychophysical data." http://www.palamedestoolbox.org, 2009.

[42] The MathWorks, Inc., *MATLAB R2017a*. Massachusetts, United States: Natick, 2017, MATLAB version 9.2.0.556344. [Online]. Available: http://www.mathworks.com

[43] ITU-T P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," 2003.

[44] J. Ramsgaard and S. V. Legarth, "Listening test on headset recordings applying the ITU-T P.835 with trained listeners – results from main systems under test," SenseLab, Tech. Rep. SenseLab 006-14(2), 2014.

[45] ITU-T P.835, "Recommendation P.835 (2003) Erratum 1 (05/08)," 2008.

[46] R. H. B. Christensen. (2015) ordinal–Regression models for ordinal data. R package version 2015.6-28. [Online]. Available: http://www.cran.r-project.org/package=ordinal/

[47] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017, R version 3.4.2 (2017-09-28). [Online]. Available: https://www.R-project.org/

[48] G. C. Inc. Voice Quality Testing (VQT) Software (POLQA, PESQ). [Online]. Available: https://www.gl.com/voice-quality-testing-pesq-polqa.html

[49] M. W. Fagerland, "T-tests, non-parametric tests, and large studies—a paradox of statistical practice?" *BMC Med Res Methodol*, vol. 12, no. 1, 2012.

[50] E. W. Healy, M. Delfarah, J. L. Vasko, B. L. Carter, and D. Wang, "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4230–4239, 2017.

[51] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2604–2612, 2016.

[52] "Application guide POLQA," HEAD acoustics, Tech. Rep. Rev0 (3/2012), 2012.

Femke B. Gelderblom received a BSc degree in Applied Physics, and a MSc degree in Biomedical Engineering from Delft University of Technology, the Netherlands, in 2012. Since then, she has been working as a research scientist at the Acoustics group of SINTEF Digital in Trondheim, Norway. She is currently also working towards a PhD degree at the Signal Processing group of the Norwegian University of Science and Technology (NTNU), under supervision of Tor Andre Myrvoll and Torbjørn Svendsen. Her main research interests include speech enhancement, deep learning, and microphone arrays.

Tron Vedul Tronstad is a research scientist at the Acoustics group at SINTEF Digital in Trondheim, Norway. He received a MSc degree in acoustics from the Department of Electronics and Telecommunications at the Norwegian University of Science and Technology (NTNU) in 2007. He received a PhD from the Department of Electronic Systems at the same university in 2018. His main research topics revolves around hearing and hearing damage.

Erlend Magnus Viggen received an MSc degree in applied physics in 2009 and a PhD in acoustics in 2014, both at the Norwegian University of Science and Technology (NTNU). While making most of his contributions to this work, he worked as a research scientist at the Acoustics group of SINTEF Digital in Trondheim, Norway. Currently, Erlend is a post-doc at NTNU, working at the Centre for Innovative Ultrasound Solutions on the topic of ultrasonic logging in petroleum wells. His current research interests include physical acoustics, computational acoustics, and machine learning.