Using Deep Speech Recognition to Evaluate Speech Enhancement Methods

Shamoon Siddiqui, Ghulam Rasool, Ravi P. Ramachandran, and Nidhal C. Bouaynaya
Department of Electrical and Computer Engineering
Rowan University, New Jersey, USA
{siddiq76, rasool, ravi, bouaynaya}@rowan.edu

***Abstract*—Progress in speech-related tasks is dependent on the quality of the speech signal being processed. While much progress has been made in various aspects of speech processing (including but not limited to, speech recognition, language detection, and speaker diarization), enhancing a noise-corrupted speech signal as it relates to those tasks has not been rigorously evaluated. Speech enhancement aims to improve the signal-to-noise ratio of a noise-corrupted signal to boost the speech elements (signal) and reduce the non-speech ones (noise). Speech enhancement techniques are evaluated using metrics that are either subjective (asking people their opinion of the enhanced signal) or objective (attempt to calculate metrics based on the signal itself). The subjective measures are better indicators of improved quality but do not scale well to large datasets. The objective metrics have mostly been constructed to attempt to model the subjective results. Our goal in this work is to establish a benchmark to assess the improvement of speech enhancement as it relates to the downstream task of automated speech recognition. In doing so, we retain the qualities of subjective measures while ensuring that evaluation can be done at a large scale in an automated fashion. We explore the impact of various noise types, including stationary, non-stationary, and a shift in noise distribution. We found that existing objective metrics are not a strong indicator of performance as it relates to an improvement in a downstream task. As such, we believe that Word Error Rate should be used when the downstream task is automated speech recognition.**

***Index Terms*—speech enhancement, distribution shift, signal-to-noise, benchmark**

I. Introduction

The human brain is a miraculous piece of hardware with robust software that is capable of focusing on a specific signal of interest from a wide spectrum of signals. This process, which is effortless in our biological systems, has proven extremely challenging to replicate in communication systems. *Speech enhancement* (SE) refers to the method of improving the quality of an audio signal so that the speech components are more prevalent. Lin *et al.* [22] state that “the objective of enhancement may perhaps be to improve the overall quality, to increase intelligibility, to reduce listener fatigue, etc.,” and efforts to accomplish this date back to the middle of the 20th century [9] [4]. We expand that definition and make it more general by stating that “the objective speech enhancement is to improve the quality of an audio signal to improve the quality of a downstream task.”

The source code for this paper has been made publicly available.[1]

A. Motivation

Voice as a user interface is already a dominant paradigm in many areas, including (but not limited to) smart assistants (Siri, Alexa, Google, etc.), automated telephone systems, and machine translation software. Systems like those rely on converting a speech signal into natural language to be processed in some manner.

When noise corrupts a speech signal, the quality and intelligibility of the speech can be severely compromised [30]. In the case of additive noise (which is the scenario of this paper), noise reduction is accomplished at the expense of introducing speech distortion [30]. The question of how to measure the relative improvement of speech signal needs to be considered. Evaluating whether or not (and to what degree) an audio signal is “high quality” is not as straightforward as one might initially suspect.

Subjective assessments, such as Mean Opinion Score (MOS), involve asking people their opinion on various aspects of the signal [33]. Three of the most common metrics are signal (SIG), background noise intrusiveness (BAK), and overall quality (OVRL) [14]. These are covered in more detail in **Section III**. However, in practice, these metrics do not scale for the development of automated systems. By their nature, subjective measures are dependent on the population metrics characteristics being polled, which can be difficult to measure and reproduce reliably.

To that end, objective metrics have been established that aim to give a consistent measure that can be automated (these are discussed in greater detail in **Section III**). While objective metrics resolve the issues of population characteristics and scalability, they are not reliable indicators for all classes of downstream processing tasks. Many objective measures have been constructed in an attempt to model subjective measures or some combination of other objective measures; therefore, at their core, they still suffer from many of the limitations of subjective measures.

Our work demonstrates that performance in downstream tasks is a more suitable objective measure by showing how current metrics fall short, specifically focusing on Deep Neural Network (DNN) based Automated Speech Recognition (ASR).

This work was supported by the U.S. Department of Education Graduate Assistance in Areas of National Need (GAANN) Grant Number P200A180055.

[1] https://github.com/shamoons/speech-enhancement-asr

B. Contributions

1) Establish a baseline for future work of the downstream metric of Word Error Rate (WER) applied to speech enhancement methods;
2) Demonstrate that current objective metrics are not an adequate indicator for quality improvement with regards to speech enhancement;
3) Evaluate existing DNN-based speech enhancement methods;

II. Speech Enhancement Methods

There are several methods for SE [27] [34] [35]; however, this work evaluates SOTA methods that rely on DNNs with differing mechanisms in their method of action. As a form of control, a *Wiener Filter*, which is not DNN based, was also used.

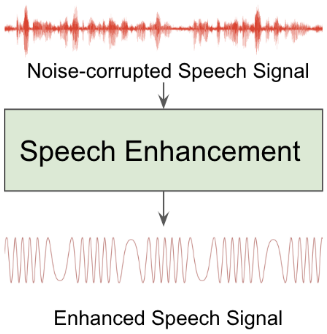

Fig. 1: **Typical Speech Enhancement**: Noise-corrupted signal is passed to some enhancement mechanism, and the output is an aligned "enhanced" signal that is free from noise.

A. Wiener Filter

Wiener Filtering [40] [24] [1] is a commonly used method for speech enhancement tasks. The filter attempts to estimate the clean speech signal from an additive noise-corrupted speech signal. This approach attempts to reduce the error between the desired clean signal $s(n)$ and the estimated signal $s'(n)$. The implementation we use [23] is a local-mean filter with x being the noise-corrupted speech signal. The enhanced signal, y, is defined as:

$$y = \begin{cases} \frac{\sigma^2}{\sigma_x^2} m_x + (1 - \frac{\sigma^2}{\sigma_x^2}) & \sigma_x^2 \geq \sigma^2 \\ m_x & \sigma_x^2 < \sigma^2 \end{cases} \tag{1}$$

where m_x and σ_x^2 are local estimates of the mean and variance, respectively, and σ^2 is a noise threshold noise parameter that is estimated as the average of the local variances.

B. Generative Adversarial Network

Generative models aim to construct data from some distribution; one such model is the *Generative Adversarial Network* (GAN) [13]. GANS learn a model that can sample from $p_{model}(x)$, without having to define the model explicitly (*implicit density estimation*). A GAN is typically comprised of two networks that are competing with each other, a generative model G and a discriminative model D that attempt to "outsmart" each other. Specifically, G learns to emulate the data distribution of the training data and D learns to determine the probability that a sample came from the training data or G. By iteratively training the two competing networks, G learns to create synthetic data that is (ideally) indistinguishable from natural data [13]. GANs have found success in a variety of fields including (but not limited to): hyperrealistic face image generation [17], data augmentation [3], music generation [10] [8], voice/music source separation [37] [11].

The potential use cases of GANs are constantly evolving as their capabilities increase. They initially caught the public imagination by producing photorealistic images of natural data, but have also been used to create realistic speech. The first of these efforts, WaveGAN [8], was able to create intelligible words and other natural sounds.

Since then, Pascual *et al.* [31] demonstrated that a GAN could also be used to enhance speech that has been corrupted with additive noise. Their **Speech Enhancement Generative Adversarial Network** (SEGAN) architecture attempts to generate clean speech conditioned on some noisy speech.

In the SEGAN architecture, the **G network** (generator) performs the actual enhancement, whereas the **D network** (discriminator) attempts to distinguish between clean speech and noise-corrupted enhanced speech. If the D network detects **fake** speech, it learns to evaluate the quality of enhanced speech more accurately. If it detects **true** speech, the discriminator weights are frozen, and the encoder and decoder of the G network are trained via standard backpropagation.

C. Variational Constrained Autoencoder

Autoencoders allow for efficient dimensionality reduction of complex data by training a network, known as the *encoder*, to learn a set of weights θ_{enc} that turns the input (image, audio, real values, etc.) into a lower-dimensional representation, typically defined by a vector. This vector is then fed to a *decoder* network that learns weights θ_{dec} that attempts to reconstruct the initial input (given to the *encoder* network) given the autoencoded representation. Applications of autoencoders include:

- creating a projection onto a lower-dimensional space, which can help remove the impacts of noise;
- compressing input to reduce computational complexity;

As an extension to autoencoders, *Variational Autoencoders* [18] allow for *generative modeling*. Rather than simply learning a deterministic representation, in such a model, some input X' is encoded into a latent space representation, Z. This space is sampled to generate some new data X. In **Figure ??**, we see that the *encoder* learns two embeddings, a mean and a standard deviation, and then samples from that distribution when training the *decoder*. Variational autoencoders allow us to generate data that are close to some initial input but perturbed in some specific way.

While autoencoders have been used for denoising applications previously [25], Braithwaite *et al.* [5] proposed a generative model subject to a constraint on the variance of the distributions, dubbed **Speech Enhancement Variational Constrained Autoencoders** (SEVCAE):

> Let X and $\widetilde{X}$ be random variables representing blocks of clean and noisy speech, respectively. X and $\widetilde{X}$ have distribution $p_D(x)$ and $p_D(\widetilde{x})$, respectively, both defined by the data. The speech enhancement problem can be formulated as learning the distribution $p(x|z)$, where Z is a set of latent features that describe the clean speech being generated. We wish to learn the distribution over latent features given the noisy data, $q(z|\widetilde{x})$.

The SEVCAE architecture is simpler than comparable methods, including SEGAN. As a result, it has reduced computational time during training and inference. In their evaluation, Braithwaite *et al.* rely on *subjective evaluation* to show its performance [5].

III. Existing Metrics

There is no commonly accepted standard for measuring how improved a speech signal is after SE is applied. Broadly, measuring speech quality falls into two categories: subjective and objective. In this section, we outline the methodology of *subjective metrics* as well as the *objective metrics* that we study.

A. Subjective Metrics

Much work has gone into designing, evaluating, and quantifying subjective metrics; Hu *et al.* [14] describe various subjective metrics from a rating of 1 (bad) to 5 (excellent): speech only, which asks for the degree of signal distortion (SIG), background noise, which asks for how intrusive the background noise is (BAK) and the overall quality (OVRL).

However, across various benchmarks and evaluations, the total number of listeners that were contributed ratings is low, under 100 participants [14] [41]. Given that subjects are asked to rate enhanced speech across the three areas (SIG, BAK, OVRL), no correlation is made with regards to predicted performance on any downstream task.

B. Objective Metrics

Several objective metrics are used in the literature, including segmental SNR (SSNR), weighted-slope spectral distance (WSS) [21], various linear predictive coding (LPC) based objective measures and cepstrum distance measures (CEP) [14]. We restrict the evaluation to 2 of the more popular methods: PESQ and STOI.

1) Perceptual Evaluation of Speech Quality (PESQ): This is one of the most common objective metrics and is defined by the International Telecommunication Union (ITU) as [32]:

> PESQ compares an original signal $X(t)$ with a degraded signal $Y(t)$ that is the result of passing $X(t)$ through a communications system. The output of PESQ is a prediction of the perceived quality that would be given to $Y(t)$ by subjects in a subjective listening test.

As such, it attempts to model what listeners might rate the perceived quality of an improved (or degraded) signal. While this certainly shows a strong correlation between what the automated PESQ method will calculate and what people generally say, it does not account for the weaknesses and limitations of MOS [36]. It has been shown that MOS reporting tends to be given as a precise number, whereas in reality it is a statistical measurement. As such, the variance across the population is rarely reported or considered. Also, it has been shown that MOS does not adequately account for bias, mood, a priori estimates, and other factors [19]. Therefore, an automated measure that attempts to model and emulate MOS is subject to the same shortcomings and biases.

The *ITU-T Recommendation P.862* [15] provides raw scores between -0.5 and 4.5, however, the updated *P.862.1 Recommendation* [16] provides a mapping to align with MOS standards to a range of 1 to 5.

2) Short-Time Objective Intelligibility (STOI): As the name implies, this is an objective metric that attempts to rate the perceived "intelligibility" of a signal relative to some reference signal [38]. Like PESQ, given that it needs a reference signal, it is said to be *intrusive*. Rather than attempt to create an automated MOS, STOI is designed to measure the intelligibility of a signal that has been processed by a time-frequency (TF) weighting, such as in the case of speech enhancement. To show that STOI is an adequate measure of intelligibility, Taal *et al.* [38] had "15 normal-hearing native Danish speaking subjects," evaluate the intelligibility and compare to the predictions given by STOI. While the correlation was high ($\rho = 0.95$), the interesting point is the lack of correlation to a downstream task (such as ASR) and the limited set of participants. The range of scores is between 0 (low intelligibility) and 1 (high intelligibility).

IV. Experimental Design

The various SE methods that are being evaluated measure themselves differently, so in an attempt to establish a baseline to compare various methods, this work looks at two commonly used and cited "objective" metrics (PESQ, STOI) and our implementation of WER, described above and **Section IV-B** respectively. PESQ and STOI are calculated by comparing the **enhanced** signal to the reference **clean** speech signal, whereas WER is calculated over the **enhanced** signal only. We evaluated six different noise conditions, six distortion levels, and three SE methods, along with a *not enhanced* setting. This resulted in a total of 144 conditions that were evaluated; each condition was run two times, with 250 speech samples per trial. Some sentences are short (1 word), whereas others are 50 or more words, so running multiple trials of each noise condition gave a better representation of the underlying metrics.

A. Data

1) Speech Corpus: Since the pre-trained DeepSpeech model included the LibriSpeech [29] *train* audio sets, we utilized the *test-clean* dataset from that corpus. Each speech

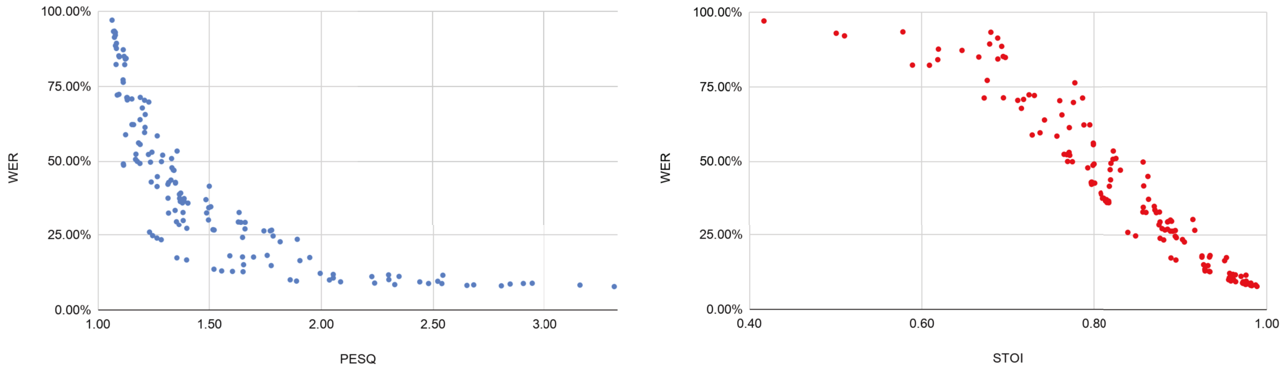

Fig. 2: Across all noise conditions, we see PESQ (blue) and STOI (red) as they relate to WER. PESQ has a range from 1 (low quality) to 5 (high quality); STOI has a range from 0 (low quality) to 1 (high quality).

file contains one sentence from freely available audiobooks sampled at 16KHz, along with aligned transcripts free from punctuation constraints. The audio is without background noise or distortion.

2) Noise: A mix of stationary and non-stationary noise sources were added to clean speech from the NOISEX-92 database [39]:

1) White Noise
2) Voice Babble
3) F16 Fighter Jet
4) M109 Tank
5) Machine gun

To introduce a shift in distribution, another type of noise was created by concatenating three sources (at random, without replacement), which is referred to as *Shift.3*.

B. Word Error Rate

For each trial of 250 speech samples, the WER is calculated as shown below:

$$WER = \frac{\sum_i^n LD(n)}{\sum_i^n len(n)} \tag{2}$$

1) n is a specific sentence
2) LD is the word distance compared to the ground truth (commonly known as the *Levenshtein Distance* [21])
3) len is the number of words in sentence n

As a baseline, after evaluating five trials of 250 samples of clean speech, the mean and standard deviation of the WER is 7.33% and 0.52%, respectively. This is consistent with the reported results of the DeepSpeech baseline.

C. Control

The *no enhancement* baseline was evaluated by adding each noise source to the LibriSpeech *test-clean* data samples at the various SNRs outlined in **Section IV-E**. This was then evaluated without any SE methods.

D. Models

In an attempt to ensure our results accurately reflected the ones of the various SE methods, we utilized the pre-trained models that were made available by Pascual *et al.* [31] and Braithwaite *et al.* [5] for the SEGAN and SEVCAE respectively.

E. Signal-to-Noise Ratio (SNR)

The SNRs that were evaluated were between [0, 25] dB with a 5 dB interval. Since the utterances are short, each speech file had additive noise at each SNR. While some of the SE methods utilized negative SNRs in their evaluation, none perform well enough to warrant further study at negative SNRs. Similarly, an SNR above 25 dB results in a WER very close to that of the clean speech (with no additive noise).

F. Automated Speech Recognition (ASR)

The downstream task to evaluate the various SE methods, is ASR, as implemented by DeepSpeech [6]. The pre-trained model (version 0.6.0) was used to convert the speech to text, which was trained on the Fisher [7], LibriSpeech [29], Switchboard [12] and the Mozilla Common Voice English corpora. The benchmark WER (as stated in **Section IV-B**) for DeepSpeech is $\approx 7.5\%$, which is not SOTA, but it was chosen for the following reasons:

- *active open source community* - as the focus of this work is not ASR, an implementation was used that is widely supported, active and contributed to;
- *poor performance with noise* - DeepSpeech is known to perform poorly ($\approx 67\%$ WER) in noisy environments, which creates for a better testbed to evaluate the improvement of various SE methods;
- *close to human accuracy* - human accuracy on clean speech is $\approx 5.8\%$ [2], which is reasonably close to the baseline WER that we verified with DeepSpeech;

V. Results

While many methods that perform SE offer **subjective** and **objective** measures, we only consider **objective** evaluation as per the metrics described earlier. Specifically, we track PESQ and STOI against WER, given the setup described in **Section IV**.

Taken across all conditions of SNR, enhancement method, and noise, **Figure 2** shows the relationship between PESQ/STOI and WER. To measure how closely good PESQ / STOI are as objective measures, we consider the *coefficient of determination*, R^2, with WER since it measures the proportion of the variance in WER that is predictable from PESQ/STOI. Across all conditions, PESQ and STOI, the R^2 value is 0.583 and 0.906, respectively, which indicates that STOI is more strongly correlated with WER than is PESQ.

A. Subset Analysis

To better understand the conditions under which WER is explained by the objective measures, we look at subsets of noise conditions. **Table I** shows the various R^2 values as they relate to WER for the segments of *enhancement method*, *noise types*, and *SNRs*.

B. ANOVA Analysis

A two-way Analysis of Variance (ANOVA) was performed on the WER that compares the four speech enhancement methods (including *no enhancement*) and the various noise types. The ANOVA was performed separately for each SNR and is based on five trials. The WER of two cases is statistically indistinguishable if the 95% confidence of the achieved WER overlap. Otherwise, they are statistically distinguishable.

A separate ANOVA is done for two cases, namely, (1) only using the white, babble, F16 and Shift.3 noise types, and (2) only using the machine gun and M109 tank noise. The WER for the more non-stationary machine gun and M109 tank noise is significantly different than for the other four types of noise. Doing a separate ANOVA for these two cases gives us proper insight into the relative performance of the four enhancement methods.

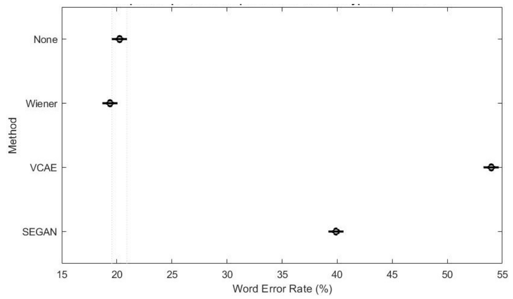

Fig. 3: ANOVA Analysis: Comparison of the enhancement methods using white, babble, F16 and Shift.3 noise at an SNR of 15 dB

The first set of results compare the WER for both cases of noise groupings. The results are consistent for each SNR in that doing no enhancement, and Wiener filtering (1) are statistically indistinguishable, and (2) achieve a lower WER than SEVCAE and SEGAN with statistical significance. **Figure 3** shows a comparison of the 95% confidence intervals for the four methods using white, babble, F16 and Shift.3 noise types at an SNR of 15 dB. **Figure 4** shows a comparison of the 95% confidence intervals for the four methods using machine gun and M109 tank noise at an SNR of 25 dB.

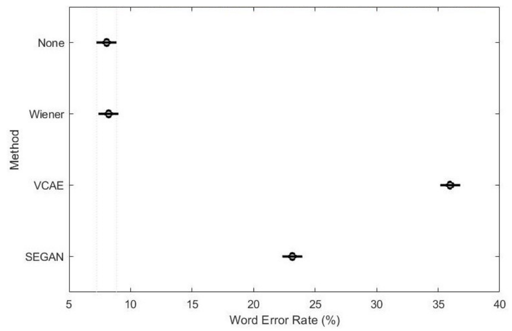

Fig. 4: ANOVA Analysis: Comparison of the enhancement methods using machine gun and M109 tank noise at an SNR of 25 dB

The second set of results compare the WER for various noise types. Again, the results are consistent for each SNR. **Figure 5** shows a comparison of the 95% confidence intervals for the four noise types: white, babble, F16, and Shift.3 noise at an SNR of 20 dB. White noise leads to the highest WER with statistical significance. Babble noise leads to the lowest WER with statistical significance. The WER when comparing machine gun and M109 tank noise is statistically indistinguishable at every SNR.

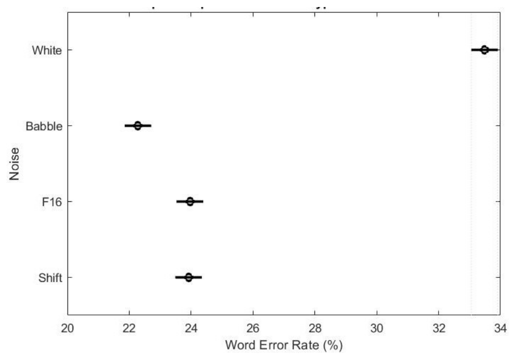

Fig. 5: ANOVA Analysis: Comparison of the 95% confidence intervals for the four noise types: white, babble, F16 and Shift.3 noise at an SNR of 20 dB

VI. Summary and Conclusions

A. Discussion

First, we note that there is a correlation between the existing objective measures (PESQ and STOI) and WER, which indicates that for some applications and conditions, the existing objective measures may be suitable to gauge enhancement

		Enhancement Methods				Noise Types						Signal-to-Noise Ratios					
	Aggregate	*SEGAN*	*Wiener*	*SEVCAE*	*None*	*White*	*Babble*	*F16*	*M109*	*Machine Gun*	*Shift.3*	*0 dB*	*5 dB*	*10 dB*	*15 dB*	*20 dB*	*25 dB*
PESQ	0.583	0.813	0.567	0.877	0.541	0.764	0.651	0.672	0.563	0.534	0.698	0.757	0.788	0.810	0.811	0.857	0.885
STOI	0.906	0.944	0.952	0.909	0.947	0.843	0.961	0.914	0.964	0.947	0.955	0.656	0.751	0.867	0.934	0.952	0.976

TABLE I: **Coefficients of determination (R^2) with WER**: for various subsets of Enhancement Methods, Noise Types and SNRs. Higher values indicate a stronger correlation. Lower values indicates weaker correlation. Degree of determination is strongly dependent on noise conditions.

quality. However, we observe that under certain conditions, the correlation is much less convincing; for example, in the case of *Machine Gun* noise, which is non-stationary and intermittent, PESQ is a poor indicator for WER, whereas STOI is a good indicator. This suggests that the current objective measures are **highly** dependent on conditions. In any real-world scenario, it's unlikely that we would know the distribution of the noise and/or the SNR (although in the case of SNR, many methods to exist to estimate it [28]).

Second, as shown in **Table I** under *Signal-to-Noise Ratios*, the objective measures studied correlate more strongly with WER when the SNR is high (less noise). Lower R^2 values are observed at lower SNRs (more noise). However, as a matter of practicality, this is the exact opposite of what one would want. If the SNR is high, then the speech signal is already quite clear, and therefore, enhancement is not needed. It's only at lower SNRs that we should be concerned about the quality of an enhanced speech signal. **Figure 6** shows how the R^2 increases almost monotonically with SNR.

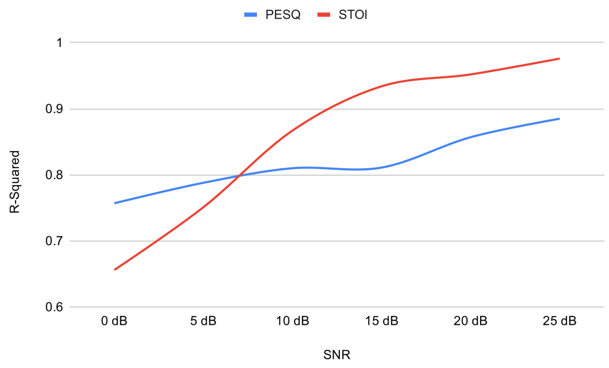

Fig. 6: R^2 values with WER as SNR increases.

B. Conclusion

In attempting to understand why the various DNN-based SE methods (SEVCAE and SEGAN) have a higher WER, in general than simply doing nothing or a simple filter, we consider that DNNs operate under the principle of minimizing some loss function. As a result, they don't have a sense of context, which seems to be important for humans as we can fill in gaps from noisy speech. In this work, we demonstrate the need for a new paradigm of evaluating speech enhancement methods. Our experiments demonstrate that existing objective measures are inadequate and lack any truly consistent predictive capabilities for how an enhanced speech signal would be utilized by a downstream task (ASR in this case). Possible followup research directions include: evaluating other deep neural speech enhancement methods, such as the ones proposed in [20] [26] [34] and others; expanding the evaluation to some of the other objective measures outlined in **Section III**; considering other downstream tasks (such as speaker diarization, language detection, and others); incorporating context into speech enhancement. Many SE methods rely on subjective MOS results as an evaluation metric, but our results demonstrate that by utilizing a downstream task, we can benefit from an objective evaluation that is scalable, reproducible, and perhaps most important: **meaningful**.

References

[1] MA Abd El-Fattah, Moawad Ibrahim Dessouky, Salah M Diab, and Fathi El-Sayed Abd El-Samie. Speech enhancement using an adaptive wiener filtering approach. *Progress in Electromagnetics Research*, 4:167–184, 2008.

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016.

[3] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[4] L. L. Beranek. The design of speech communication systems. *Proceedings of the IRE*, 35(9):880–890, Sep. 1947.

[5] DT Braithwaite and W Bastiaan Kleijn. Speech enhancement with variance constrained autoencoders. *Proc. Interspeech 2019*, pages 1831–1835, 2019.

[6] Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, and Erich Elsen. Deep Speech : Scaling up end-to-end speech recognition. pages 1–12.

[7] Christopher Cieri, David Miller, and Kevin Walker. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, 2004.

[8] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. In *ICLR*, 2019.

[9] Harris Drucker. Speech processing in a high ambient noise environment. *IEEE Transactions on Audio and Electroacoustics*, 16(2):165–168, 1968.

[10] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. *7th International Conference on Learning Representations, ICLR 2019*, pages 1–17, 2019.

[11] Zhe-Cheng Fan, Yen-Lin Lai, and Jyh-Shing R Jang. Svsgan: Singing voice separation via generative adversarial network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 726–730. IEEE, 2018.

[12] J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1, March 1992.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[14] Yi Hu and Philipos C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, 16(1):229–238, 2008.

[15] Itu-T. Perceptual evaluation of speech quality (PESQ). *Networks*, 862:749–752, 2001.

[16] ITU-T. P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO. *ITU-T Recommendation*, 2003.

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[19] Hendrik Knoche, Hermann G De Meer, and David Kirsh. Utility curves: Mean opinion scores considered biased. In *1999 Seventh International Workshop on Quality of Service. IWQoS'99.(Cat. No. 98EX354)*, pages 12–14. IEEE, 1999.

[20] Simon Leglaive, Umut Simsekli, Antoine Liutkus, Laurent Girin, and Radu Horaud. Speech Enhancement with Variational Autoencoders and Alpha-stable Distributions. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019-May:541–545, 2019.

[21] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[22] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, Dec 1979.

[23] Jae S Lim. Two-dimensional signal and image processing. *Englewood Cliffs, NJ, Prentice Hall, 1990, 710 p.*, 1990.

[24] Philipos C Loizou. *Speech enhancement: theory and practice*. CRC press, 2013.

[25] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In *Interspeech*, pages 436–440, 2013.

[26] Juan Manuel Martin-Donas, Angel Manuel Gomez, Jose A. Gonzalez, and Antonio M. Peinado. A deep learning loss function based on the perceptual evaluation of the speech quality. *IEEE Signal Processing Letters*, 25(11):1680–1684, nov 2018.

[27] Tokunbo Ogunfunmi, Ravi Prakash Ramachandran, Roberto Togneri, Yuanjun Zhao, and Xianjun Xia. A primer on deep learning architectures and applications in speech processing. *Circuits, Systems, and Signal Processing*, pages 1–27, 2019.

[28] Russell Ondusko, Matthew Marbach, Ravi P Ramachandran, and Linda M Head. Blind signal-to-noise ratio estimation of speech based on vector quantizer classifiers and decision level fusion. *Journal of Signal Processing Systems*, 89(2):335–345, 2017.

[29] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[30] Mahdi Panchami, Wei-Ping Zhu, Benoit Champagne, and Eric Plourde. Recent developments in speech enhancement in the short-time fourier transform domain. *IEEE Circuits and Systems Magazine*, 2016.

[31] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

[32] ITU-T Recommendation. Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*, 2001.

[33] ITUT Recommendation. Vocabulary for performance and quality of service, 2006.

[34] Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018-April:5069–5073, 2018.

[35] Björn Schuller, Martin Wllmer, Tobias Moosmayr, and Gerhard Rigoll. Recognition of noisy speech: A comparative survey of robust model architecture and feature enhancement. *Eurasip Journal on Audio, Speech, and Music Processing*, 2009, 2009.

[36] Robert C. Streijl, Stefan Winkler, and David S. Hands. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227, 2016.

[37] Y Cem Subakan and Paris Smaragdis. Generative adversarial source separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 26–30. IEEE, 2018.

[38] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. A SHORT-TIME OBJECTIVE INTELLIGIBILITY MEASURE FOR TIME-FREQUENCY WEIGHTED NOISY SPEECH Signal Information & Processing Lab , 2628 CD Delft , The Netherlands Oticon A / S 2765 Smørum , Denmark. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4214–4217, 2010.

[39] Andrew Varga and Herman JM 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*, 12(3):247–251, 1993.

[40] Norbert Wiener. Extrapolation, interpolation, and smoothing of stationary time series, vol. 2, 1949.

[41] Yi Hu and P.C. Loizou. Subjective Comparison of Speech Enhancement Algorithms. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, volume 1, pages I–153–I–156. IEEE, 2006.