INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Say what? A dataset for exploring the error patterns that two ASR engines make

Meredith Moore, Michael Saxon, Hemanth Venkateswara,
Visar Berisha, Sethuraman Panchanathan

Arizona State University, United States
mkmoore7@asu.edu

Abstract

We present a new metadataset which provides insight into where and how two ASR systems make errors on several different speech datasets. By making this data readily available to researchers, we hope to stimulate research in the area of WER estimation models, in order to gain a deeper understanding of how intelligibility is encoded in speech. Using this dataset, we attempt to estimate intelligibility using a state-of-the-art model for speech quality estimation and found that this model did not work to model speech intelligibility. This finding sheds light on the relationship between how speech quality is encoded in acoustic features and how intelligibility is encoded. It shows that we have a lot more to learn in how to effectively model intelligibility. It is our hope that the metadataset we present will stimulate research into creating systems that more effectively model intelligibility.

Index Terms: intelligibility, quality, estimation models, error detection, auditory perception, automatic speech recognition

1. Introduction

Intelligibility refers to the ability for an utterance to be understood. It is a complex concept which involves a speaker and a listener and complicated perceptual systems. Speakers can be human or machine, listeners can be human or machine, and the ability for the speaker to be understood by the listener is described as the intelligibility of the speech signal. Traditional intelligibility problems are associated with characterizing the degradation of speech through a lossy telecommunication system. These metrics rely on the comparison of degraded output speech with the corresponding clean input. While this characterization of quality is crucial in designing end-to-end communication systems, other important notions of intelligibility exist outside of this paradigm. Indeed high-quality, non-degraded speech signals can still be difficult to understand, be it due to accents, speech disorders, or background noise.

An ideal intelligibility estimation system would provide an objective measure that applies across a broad swath of settings such as those mentioned above. It could be used to measure performance/uncertainty of automatic speech recognition (ASR) systems or the intelligibility of speech, providing insight into what kinds of data the ASR system is likely to predict incorrectly. These systems could also be used to measure how intelligible human speech is–something that could be used in a clinical setting (measuring performance in speech therapy), or an educational setting (learning a language).

Existing methods in intelligibility estimation either attempt to characterize degradation, and really only estimate quality, or directly measure machine intelligibility of speech using an ASR system and ground truth transcript. Neural network approaches should be well-suited to estimate intelligibility, however, there are no ground-truth intelligibility-labeled datasets available. Moving from the assumption that ASR intelligibility can serve as a rough proxy for human intelligibility, we create an inclusive metadataset of word error rate (WER) labels for multiple publicly available English speech datasets[1], including healthy native speech, healthy accented speech, and disordered native speech, assessed by two different popular ASR models. We demonstrate the nontriviality of this problem by showing that metrics designed to capture intelligibility in the quality sense are not well-suited for assessing broader senses of intelligibility, in particular, that neural approaches well-adapted to quality assessment fail to learn any useful intelligibility assessment capabilities.

2. Previous Work

2.1. Speech Intelligibility Metrics

The Articulation Index (AI) was the first mentioned measure of intelligibility that aimed to be a short-cut around intelligibility-testing with human subjects in a lab [1]. The main goal of the AI is to account for distortions in the frequency domain (noise, bandpass-limit) when speech is transmitted. The Speech Intelligibility Index takes intelligibility estimation one step further and accounts for nonlinear distortions (peak clipping) as well as for distortions in the time domain (reverberation, echoes, AGC)[2]. The Speech Intelligibility Index is also an updated and expanded version of AI that is driven by per frequency band signal to noise ratios.

Metrics like the AI and SII have served as estimators for speech intelligibility for many years, however, these metrics are what is referred to as 'intrusive' in that they require both the clean and degraded versions of the speech sample [3]. This constraint limits the usefulness of these intelligibility metrics in real-time as the clean speech counterpart is not generally available.

2.2. Speech Quality Metrics:

A closely related concept to intelligibility is speech quality. Speech quality refers to the extent to which a human finds the speech pleasing. Traditionally, being able to estimate the quality of a speech sample was useful in the development of telecommunication systems [3]. Modern telecommunication systems necessitate the use of digital speech coding. Speech coders with a low bit rate may have a negative effect on the quality of the speech as the reconstructed signal becomes a noticeably lossy reconstruction to the recipient.

In subjective speech quality assessments, speech quality is measured by a Mean Opinion Score (MOS) which asks listeners

[1]The metadataset is available at https://t2m.io/T3PoCbdh.

http://dx.doi.org/10.21437/Interspeech.2019-3096

to rate the speech on a scale of one (bad) to five (excellent) in terms of the quality of the sound [4]. The MOS is a perceptual score as the listener perceives the speech through their human auditory system. A subset of speech quality metrics focuses on the human perception of speech. One of these perceptual speech quality metrics is the Perceptual Estimation of Speech Quality (PESQ) metric.

In the paper that inspired this project [5], the authors propose a non-intrusive quality metric that estimates the PESQ score. This eliminates the need for a 'gold standard' speech sample to compare a degraded signal to. Quality-Net consists of a bidirectional long short term memory network that converges very quickly–usually after around 1000 samples.

2.3. ASR Error Estimation/Detection

ASR error detection is the problem of detecting when an ASR system is going to make an error. ASR error detection is useful in evaluating the performance of ASR systems in real-time situations where the transcripts are not available. In [6], the authors use recurrent neural networks to predict when an ASR system is going to make an error. In predicting if an ASR is going to make an error, the network is predicting whether or not the ASR system is able to understand the speech signal–if it is intelligible or not. This same thought process can be used to argue that ASR error detection can approximate whether or not a human would understand the utterance.

While ASR WER is not a fully satisfactory proxy for human intelligibility, there has been some recent research that indicates that the difference between human and machine intelligibility is smaller might be expected. In [7], researchers at Microsoft were able to achieve human parity on conversational speech recognition. In [8], we demonstrate that despite a wide variation in the range of intelligibility of the speech that is tested, ASR WER loosely approximates the intelligibility of humans when applied to speech from individuals with voice disorders.

Some metrics attempt to infer a more general sense of intelligibility directly from the performance of an ASR system. Goodness of Pronunciation [9] is a measure of articulatory precision that compares the log probability of a "true" phoneme from a transcript to the most likely phoneme given a frame of speech data, as assessed by the internal probabilities of an ASR system. This and related metrics have been applied in both educational [10] and clinical contexts [11].

Error estimation in ASRs is a very similar task to ASR error detection, but rather than predicting a 1 or 0 for whether or not there was an error, it becomes a regression problem where the end goal is to approximate the WER for the given utterance. To make the problem of ASR error estimation simpler, there are several papers that rather than taking a regression approach, turn the task into a classification task. In [12], the researchers simplify the task of estimating WER into two classification tasks–classifying the numerator of WER (S+D+I), and classifying the denominator (N), and then dividing the two to get the e-WER. In this approach, the authors also use a variety of acoustic features as input into the network.

2.4. Speech Quality/Intelligibility Relationship

Speech quality and intelligibility are inherently related. Traditionally, the main cause of a decrease in intelligibility was the poor quality of the reconstructed speech sample from compression. More recently, however, communication systems like voice over internet protocol (VOIP) are reconstructed with more than sufficient speech quality that the speech is intelligible. In [13] the authors evaluated the relation between speech intelligibility and quality ratings and found a non-linear relationship. In the case where the intelligibility of the speech is high, quality can vary across the whole range, while in the case of low intelligibility, the quality is mainly determined by the intelligibility. If an individual is able to understand 90% of what was said, the range of MOS values ranged from a 2-5, however, if the listener was able to understand less than 90% of what was said, the MOS score was ubiquitously between scores of 1 and 2. Despite the relationship between speech quality and speech intelligibility being non-linear, we hypothesized that re-purposing a network that performed well on quality estimation for WER estimation would result in a model that effectively predicted the WER.

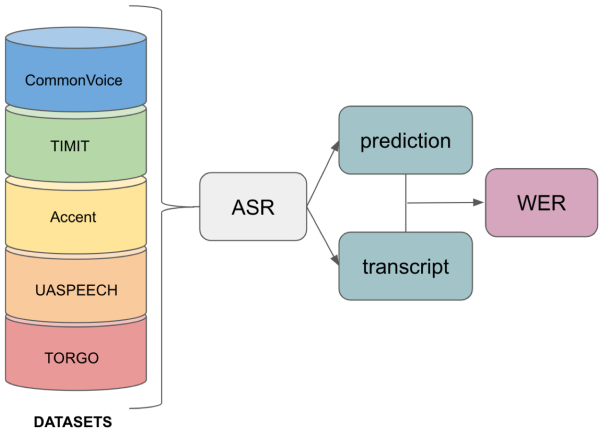

Figure 1: *The metadataset collection process*

3. WER Metadataset

We have annotated several datasets with the output hypotheses of two different ASR systems, and an analysis of where an error was made. This metadataset includes the reference text, the hypothesis of two different ASR systems, the number of substitutions, insertions, deletions, total words, the WER, and the time that it took to get the prediction from the ASR system. The WER is defined by the sum of the number of substitutions (S), insertions (I), and deletions (D) divided by the total number of words in the utterance (N) as shown in Equation 1. It is worth mentioning that because there is a potential for an infinite number of insertions, there is no strict upper bound for WER.

$$WER = \frac{S + I + D}{N} \tag{1}$$

By providing researchers with the WER annotations for these datasets, we hope to stimulate research in the area of intelligibility estimation models. One thing that makes this metadataset unique is the wide range of intelligibility and the diversity of intelligibility in the data that was analyzed. We have collected intelligibility data from the following healthy native, healthy accented, and disordered native speech datasets:

3.1. Dataset Descriptions

3.1.1. Mozilla's Common Voice

Common Voice is an open source speech dataset collected and maintained by Mozilla. Common Voice is collected in a dis-

Table 1: *Metadetaset Metadata: a breakdown of the type of speech included in each dataset and the number of speakers in each dataset.*

Dataset	**Type**	**Speakers**
TORGO/UASPEECH	path	27
Accent	accented	2140
CommonVoice	average	33,541
TIMIT	average	630
Total		36338

tributed fashion through their website. As Common Voice is collected wherever the speaker is, it provides a realistic amount of noise in the data. In Common Voice v1, there are 33,541 speakers included in Common Voice, and 582 hours of validated speech (validated in that someone has listened to the speech sample and has checked that it matches the prompt).

3.1.2. UASPEECH

UASpeech is a dysarthric speech database for universal access [14]. It consists of data from 19 speakers with cerebral palsy speaking 795 isolated words each, as well as speech from age and gender-matched control subjects. UASpeech includes video data of some of the subjects, as well as a percentage intelligibility rating based on human orthographic transcriptions.

3.1.3. TORGO

The TORGO Database is another dataset of dysarthric speech [15]. It includes recordings from eight individuals with dysarthrias, and age and gender-matched control subjects. TORGO is unique in that it also includes articulatory data from individuals with dysarthria.

3.1.4. Speech Accent Archive

The Speech Accent Archive is a dataset from George Mason University that includes speech from 2140 individuals with varying language backgrounds. Both native and non-native English speaking speakers read the same paragraph [16].

3.1.5. TIMIT

TIMIT is an acoustic-phonetic continuous speech corpus that includes speech from different dialects of English [17]. TIMIT includes 630 speakers from different English dialects saying sentences that are phonetically diverse.

3.2. Metadatset Collection

Two different ASR systems were used to obtain this metadata: Google Speech Recognition and CMUSphinx Open Source Speech Recognition. Google Speech Recognition uses deep neural networks while CMUSphinx uses hidden Markov models to achieve its speech recognition. Figure 1 shows the relatively straightforward process of obtaining the data. The speech files were fed into the two ASR systems to obtain the prediction of what was said. Then, using the transcript, the WER was calculated and recorded, along with the number of substitutions, insertions, deletions, and the time taken to obtain the results. The time is included to provide a comparison for alternative techniques such as estimation models that may operate faster.

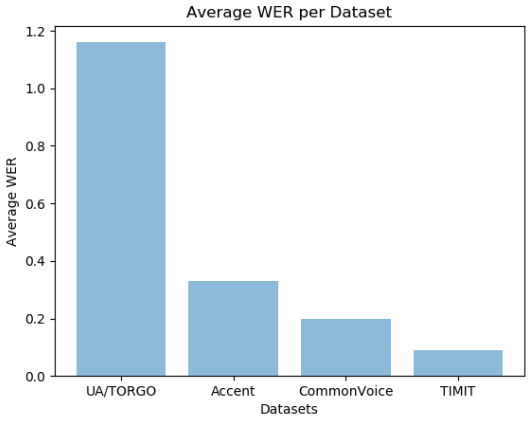

Figure 2: *The average WER per dataset included in the metadatatset.*

Table 2: *Performance of Intellinet in comparison to Quality-Net*

	MSE	**LCC**	**SRCC**
Quality-Net	0.1225	0.9054	0.9065
Intellinet	0.023	0.023	0.007

3.3. Metadataset Analysis

While many utterances are understood by the ASR systems, there is enough variability in the metadataset to find some interesting patterns. In Table 1, the type of data, and number of speakers per dataset are shown. As one would expect, the pathological datasets TORGO and UASpeech have the fewest speakers (27) and the highest average WER (1.16). In Figure 2 a bar chart shows the average WER per dataset. The dataset with the smallest average WER is TIMIT (0.09), while the Accent database (0.33) and Common Voice (0.20) sit somewhere in the middle. This distribution of WER is what we would expect. Obtaining the WER from these datasets was a slow process, it took on average 1.87 seconds per utterance, and 0.33 seconds per word for a total of 513533.93 seconds of continuous computation to obtain the WER data that is included in this dataset–that's almost 6 full days of continuous computation.

4. Experiments

In an effort to estimate intelligibility of an utterance, we decided to use a state-of-the-art model for estimating speech quality–Quality-Net [5]. We created Intellinet, our own implementation of Quality-Net to test its ability to predict intelligibility data in the form of estimating the WER. Quality-Net consists of a bidirectional LSTM layer with 100 nodes followed by two fully connected linear layers with 50 nodes each, one linear layer, and then a global averaging layer. The output of the last layer is the global average of the frame-level predictions which is the utterance-level WER prediction. As input into the network, we used spectrograms extracted from TIMIT in order to replicate Quality-Net. We selected a random subset of TIMIT (4200 utterances) to use as the training set and split the rest into a validation set (1049 utterances) and test set (1049 utterances).

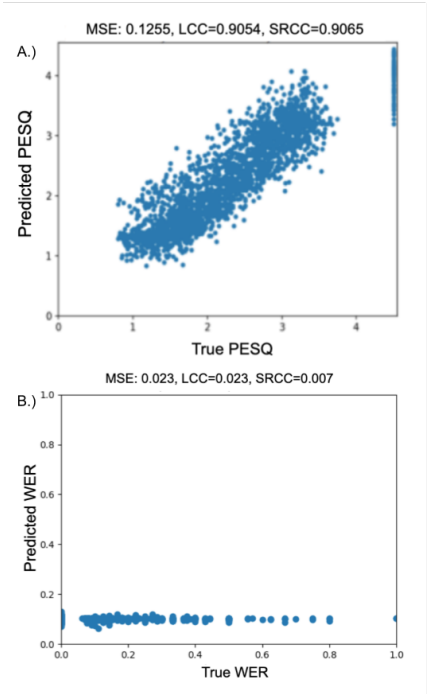

Figure 3: *In A, the correlation between Quality-Net's predictions and labels is demonstrated. B shows the correlation between the true WER and predicted WER when using Quality-Net to model intelligibility*

5. Results

Much like the results seen in Quality-Net, our model quickly converges after around 500 iterations, however, unlike Quality-Net, our model converges to the global average of the labels rather than learning the mapping between the utterances and the WER. This difference in performance is demonstrated in figure 3. In Figure 3A, the results from Quality-Net show the predicted PESQ scores compared to the true PESQ scores, and Figure 3B shows the predicted WER compared to the true WER. The PESQ predictions line up relatively along the line y=x, while the predicted WER is only predicting the global average of the data, and the Linear Correlation Coefficient (LCC) for Quality-Net's predictions is 0.9054, indicating a very strong correlation between the true PESQ value and the predicted PESQ. The Spearman's Rank Correlation Coefficient (SRCC) is 0.9065 for Quality-Net reinforcing the idea that there is a strong correlation between the true and predicted values. However, when we plot the results from the intelligibility data, we see roughly a horizontal line right around where the global average of the WER is, an LCC of 0.023 and an SRCC of 0.007. Both the LCC and SRCC for the intelligibility data are very close to zero, which indicates that there is no correlation between the true WER and the predicted WER. The mean squared error of the intelligibility network is lower than Quality-Net's but this doesn't really say much.

6. Discussion

Using the same network as Quality-Net and the same features from TIMIT, we were unable to replicate the success of Quality-Net with the task of predicting the WER instead of PESQ. While intelligibility and quality may seem like superficially similar tasks, a state of the art model for estimating quality is ill-suited for assessing intelligibility. There are a variety of reasons this might be the case.

Linguistic context is very important for understanding intelligibility. Realization difficulties become exhibited in varying phonemes for various reasons when considering disordered speech and accented speech. Word errors in ASR systems often involve the mischaracterization of one word to another as a result of the complicated interaction of internal language models and acoustic models. Out of domain acoustic patterns produced by disordered or accented speakers lead to incorrect classification calls made by the ASR model. This means that Quality-net's focus on frame-level details of the speech acoustics will miss broad patterns on a word-level scale – the sort of short-term spectral artifacts in low-quality compressed speech are in no way similar to the broader, more complicated patterns of difficult-to-understand natural speech.

Integrating linguistic data acquired solely from a speech signal and learning broad patterns or acoustics across words and the propensity for an ASR system to make errors from only WER counts is a very tall order for a simple neural network. Much more work in this area is necessary.

7. Conclusions

In this paper, we introduced a new metadataset of WER labels for several popular speech datasets spanning a wide intelligibility range. By adding the resulting transcripts from two different ASR systems, and an analysis of the number of substitutions, insertions, deletions, and total words in the predicted transcript, we hope to stimulate research into modeling intelligibility. We showed that modeling intelligibility proves to be a nontrivial task that requires novel approaches unrelated to quality assessment methods, using an existing state-of-the-art model capable of predicting PESQ scores from speech utterances. This model trains and converges but learns no useful WER prediction capabilities.

Intelligibility and quality are encoded significantly different in speech. The complexity of intelligibility and the diversity of reasons intelligibility difficulties can arise mean that significantly more complicated models integrating local acoustic, global acoustic, and linguistic data are probably necessary to model it adequately. Once such a model is created many clinical and educational applications will be available. Our hope is that a direct speech intelligibility estimation system requiring no transcripts and no 'gold' examples will drive new applications in clinical, educational, and research settings.

8. Acknowledgements

We wish to acknowledge the National Science Foundation (NSF) and their generous support through the NSF Graduate Research Fellowship program, as well as Arizona State University's Center for Cognitive Ubiquitous Computing.

9. References

[1] K. D. Kryter, "Methods for the calculation and use of the articulation index," *The Journal of the Acoustical Society of America*, vol. 34, no. 11, pp. 1689–1697, 1962. [Online]. Available: https://doi.org/10.1121/1.1909094

[2] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speechtransmission quality," *The Journal of the*

Acoustical Society of America, vol. 67, no. 1, pp. 318–326, 1980. [Online]. Available: https://doi.org/10.1121/1.384464

[3] S. Voran, *Estimation of Speech Intelligibility and Quality*. New York, NY: Springer New York, 2008, pp. 483–520. [Online]. Available: https://doi.org/10.1007/978-0-387-30441-0_28

[4] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[5] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," *Interspeech*, 2018.

[6] Y.-C. Tam, Y. Lei, J. Zheng, and W. Wang, "ASR error detection using recurrent neural network language model and complementary ASR," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2312–2316.

[7] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[8] M. Moore, H. Demakethepalli Venkateswara, and S. Panchanathan, "Whistle-blowing ASRs: Evaluating the need for more inclusive automatic speech recognition systems," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-September, pp. 466–470, 1 2018.

[9] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, no. 2-3, pp. 95–108, Feb. 2000. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(99)00044-8

[10] M. Tu, A. Grabek, J. Liss, and V. Berisha, "Investigating the role of L1 in automatic pronunciation evaluation of L2 speech," in *Proc. Interspeech 2018*, 2018, pp. 1636–1640. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1350

[11] M. Saxon, J. Liss, and V. Berisha, "Objective measures of plosive nasalization in hypernasal speech," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[12] A. Ali and S. Renals, "Word error rate estimation for speech recognition: e-WER," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2018, pp. 20–24.

[13] F. Schiffner, J. Skowronek, and A. Raake, "On the impact of speech intelligibility on speech quality in the context of voice over IP telephony," in *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, Sep. 2014, pp. 59–60.

[14] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research." in *Interspeech*, vol. 2008, 2008, pp. 1741–1744.

[15] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, Dec 2012.

[16] S. Weinberger, "Speech accent archive. George Mason University." 2013.

[17] "An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 380–394, July 1994.