USING AUDIO QUALITY TO PREDICT WORD ERROR RATE IN AN AUTOMATIC SPEECH RECOGNITION SYSTEM

Randall Fish, Qian Hu, Stanley Boykin

The MITRE Corporation, 202 Burlington Road, Bedford, MA. 01730-1420
Email:{fishr, qian, sboykin}@mitre.org

ABSTRACT

Faced with a backlog of audio recordings, users of automatic speech recognition (ASR) systems would benefit from the ability to predict which files would result in useful output transcripts in order to prioritize processing resources. ASR systems used in non-research environments typically run in "real time". In other words, one hour of speech requires one hour of processing. These systems produce transcripts with widely varying Word Error Rates (WER) depending upon the actual words spoken and the quality of the recording. Existing correlations between WER and the ability to perform tasks such as information retrieval or machine translation could be leveraged if one could predict WER before processing an audio file. We describe here a method for estimating the quality of the ASR output transcript by predicting the portion of the total WER in a transcript attributable to the quality of the audio recording.

Index Terms— Acoustic noise, Speech Recognition

1. BACKGROUND

The quality of the output transcript of an automatic speech recognition system is typically captured by the word error rate (WER) metric;

$$WER = \frac{S + I + D}{N} * 100 \qquad \text{(Eq. 1)}$$

where **S** is the number of incorrect words substituted, **I** is the number of extra words inserted, **D** is the number of words deleted and **N** is the number of words in the correct transcript. While we will adopt the common practice of referring to WER as a percent, it must be understood that it is possible to have WER exceed 100%.

The WER threshold for acceptable performance is different for different applications. It has been shown that good document retrieval performance is possible even with a 66% WER [1] but that precision begins to fall off rapidly when WER gets above 30% [2]. While there is no accepted standard performance metric for machine translation, studies have been performed showing the correlation between WER and one common machine translation metric the Bleu score [3,4]. Knowing the WER threshold for usable transcripts and an estimate of WER, processing resources can be allocated to only those files predicted to yield usable results. Our goal here is to provide such a WER prediction metric.

Every automatic speech recognition system has a limited vocabulary. If a spoken word is not included in this vocabulary, the transcribed word will be in error and WER will increase. Additionally, ASR systems rely on a language model (LM) which assumes that some sequences of words are more likely than others. If an incorrect language model is used, the WER will increase. If an ASR system intended for office dictation is applied to recordings of radio conversations between pilots and the control tower, both out of vocabulary (OOV) words and a mismatched language model will contribute to a higher WER. The final component contributing to WER is how well the actual audio characteristics match the ASR system's acoustic model.

In our proposed approach, an estimate is made of the expected WER caused by OOV and the system language model for recordings of a particular type; for example the family of intercepted control tower to pilot radio communications. The audio quality tool described here is then used to estimate the delta WER (DWER) added to this estimate based upon the audio quality of individual files under evaluation. This total predicted WER may then be used to estimate the quality of the anticipated ASR transcripts for the ultimate application such as machine translation.

2. EXPERIMENTAL SET-UP

2.1 Estimating OOV & LM Components of WER

To arrive at an estimate of the WER to be expected even with a perfect acoustic model match, the following steps are taken. First, a representative subset of the family of files to be processed is manually transcribed creating reference transcripts. These transcripts are re-recorded using a quality microphone in a quiet environment and the resulting files transcribed by the ASR system. These "clean" recording are considered to have perfect audio quality and their average measured WER becomes the estimate of the WER resulting from a mismatch in the language model and the use of out of vocabulary words for the entire collection of recordings.

2.2 Audio Quality Evaluation

In our work, audio quality was evaluated using the "**SP**eech **Q**uality **A**ssurance" (SPQA) tool [5] from NIST. Discussions of audio quality typically refer to "signal-to-noise ratio". However, this only makes sense if one has access to the signal and noise components separately. The recordings which are the target of this work contain both speech and noise combined and one must estimate the component of the signal attributable to noise. The NIST SPQA tool calculates the RMS power within a sliding 20ms window. Each reading contributes to a histogram used to identify the power during non-speech times attributable to noise and the power when both noise and speech are present. Using these as indicated in equation #2, a "speech-to-noise ratio" (SNR) metric is calculated.

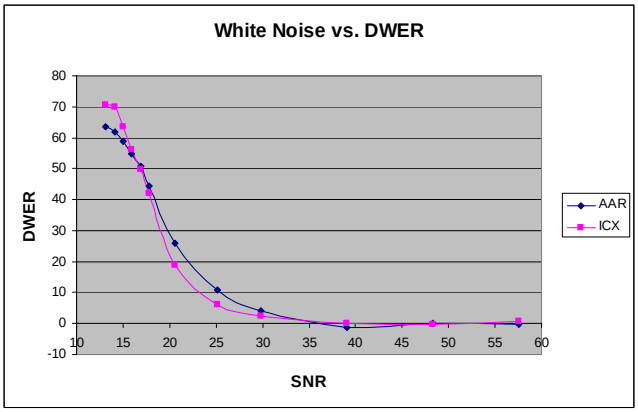

Figure #1 Delta WER resulting from the addition of various levels of white noise.

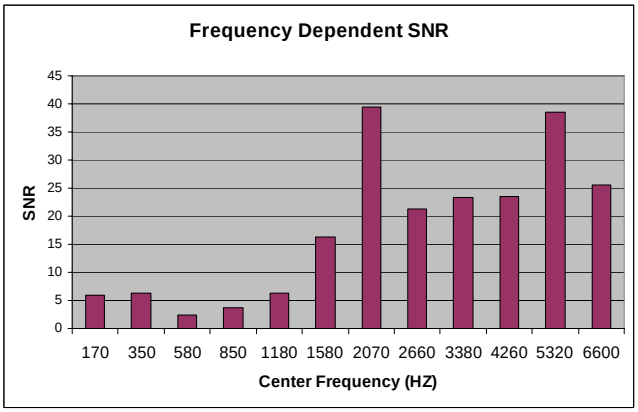

Figure #2: AAR audio file speech-to-noise ratio in mel-warped frequency bins

$$SNR = 10 * Log\left(\frac{peak\_speech\_power}{mean\_noise\_power}\right) \quad \text{(Eq. 2)}$$

Two files were used during our search for correlations between noise and DWER. The "ICX" file was a 4.5 minutes (717 words) reading of fifteen news stories relating to the Iraq constitution. The "AAR" file was a 2.6 minute (475 words) extract from an original recording of two male speakers conducting an interview on a topic outside the anticipated domain of the ASR system. The "clean" recordings of these files were made using GoldWave™ configured for 16kHz sampling rate, PCM (no compression) and a Labtec™ microphone. These clean recordings had WERs of ICX = 16.9, AAR = 28. Since the ICX transcript was taken from printed web news stories, there was no original audio for evaluation. The actual original AAR recording had WER = 82.7 indicating significant degradation due to audio quality.

3. PREDICTION METRICS

In our search for an audio quality metric we investigated the impacts of white and colored noise and the use of single or multiple audio quality features.

3.1 Single Feature White vs. Colored Noise

We first evaluated the impact of "white noise" (noise whose energy is equally distributed across the frequency spectrum). White noise at various amplitudes was added to the original "clean" recordings and the resulting SNR to WER relationship was determined. Since we are interested in the DWER rather than total WER, Figure #1 shows the DWER for our two files at various SNR settings.

The addition of small amounts of white noise, SNR > 40dB, does not effect performance. Similarly, when the added noise is too great, SNR < 10dB, WER becomes large enough to render the system unusable for most if not all applications. We focus our attention on SNR in the range of 15 – 25dB. In this range, a first approximation of the DWER to SNR relationship is:

$$DWER = 165 - \left(6.56 * SNR_{white}\right) \quad \text{(Eq. 3)}$$

To check the repeatability of this SNR to DWER prediction, we divided the single "clean" ICX file into its 15 individual news stories and added white noise resulting in SNR = 15 – 25 dB. While the average DWER for the 15 passages does indeed decrease as expected as the input SNR increases, there is a non-trivial average error (avg. = 8.5) in the predicted DWER at all levels of input noise. However, we applied the single feature white noise relationship (eq. 3) to the original AAR recording. The SNR calculation for the original AAR file was SNR = 26; which our equation predicts will have little or no degradation in the total WER for the file. Since the actual DWER = 54.7, something was obviously wrong. We assumed that treating the noise as white noise was too simplistic. For our remaining experiments we used the AAR recording allowing us to focus on colored rather than white noise.

We began by evaluating the noise in the original AAR file using a bank of twelve 51-tap overlapping bandpass filters whose center frequencies spanned the 0-8khz bandwidth using the mel warping relationship employed by ASR systems to mimic the logarithmic response of the human auditory system (eq 4).

$$Mel(f) = 2595 * \log\left(1 + \frac{f}{700}\right) \quad \text{(Eq. 4)}$$

Looking at figure #2, it is clear that the overall SNR=26 is influenced by the higher frequency bands but does not reflect the poor signal quality in the lower frequencies known to be critical for ASR performance. If we use the frequency specific information from our filters and replace our overall SNR measurement with a simple average of the first eleven frequency bins (Avg11 = 17), our white noise DWER/SNR predictor developed using the ICX file is only off by 2% when applied to the original AAR recording.

To see if this "avg11" metric is robust, we used GoldWave™ to create two types of colored noise. We filtered the white noise file used in earlier experiments to create band limited noise files with bandwidth = 1kHz ranging from 0-8kHz. The bandwidth of these noise files intentionally did not agree with the mel-warped frequency analysis tool since one can't depend upon noise to fall neatly into preferred bandwidths. Noise in only one of these eight frequency bands was added to files at randomly selected amplitudes. We refer to this as "narrow noise". Contributions, scaled to randomly selected relative amplitudes, from all eight

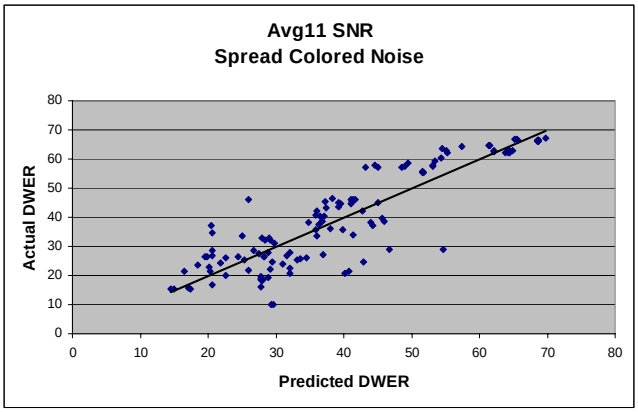

Figure #3 DWER resulting from eight band limited (1kz) noise components added at random amplitudes.

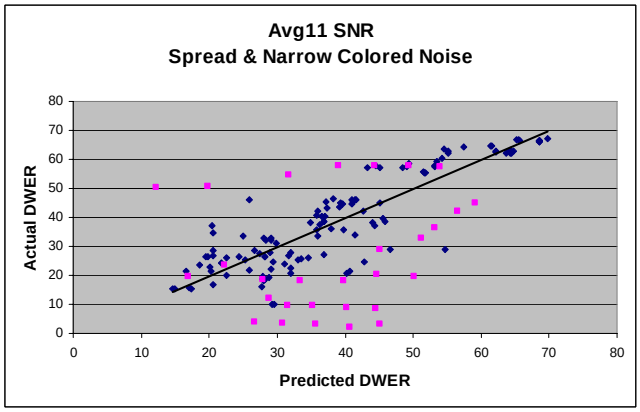

Figure #4 DWER resulting from colored noise files limited to a single band limited noise component and noise spanning the entire frequency band

bandwidth limited noise files were used to create "spread noise". These two types of colored noise files were added to the "clean" recording of the AAR file. The relationship between DWER and the AVG11 SNR metric for the 124 files corrupted with spread colored noise with average SNR in our identified range of 15 – 25dB is shown in Figure #3.

While the best fit line results in a prediction for the original AAR file with only a 7% error, there is more variance across the development files than we would like. When the 29 narrow noise files are added (Fig. #4), the fit becomes clearly unacceptable. Most of the files corrupted with narrow noise have better performance than the average SNR would predict. A possible explanation for this is that the acoustic models for most sub-word units rely on information across the full frequency spectrum. When noise is isolated to a single band of frequencies, the acoustic models are robust enough to still make the correct decision. When the noise is everywhere, the models have no good information and make more mistakes.

3.2 SNR Below Threshold Metric

Our simple SNR average assumes that noise in all frequency bands is equally important to DWER. As already noted, this is not true. Using the known frequency dependence and the previous observation that SNR above some maxSNR threshold has no impact on DWER, we developed a new set of metrics. One consideration in the creation of these new metrics was our desire to limit the number of frequency selective filters needed for our DWER prediction. The SNR tool processes the output of each filter in any specified filter bank independently. Therefore each added frequency band increases the processing time which we are trying to minimize. Starting with the same 12 overlapping frequency bins, we experimented with using all 12 filter bank outputs or only the 9 whose regression coefficients contributed at least 1% of the total predicted DWER. Both of these metrics had an average error of 4.7 on the development files. Choosing a different MaxSNR threshold for each filter bank output rather than a single common threshold allowed us to use only 6 outputs from the filter bank and further reduced the average error to 4.0. We next incorporated the observation that once noise is bad enough it causes no further degradation in WER. We therefore set a cap on the maximum difference between the measured SNR and the MaxSNR threshold. This reduced average error to 3.75. Attempts to reduce the number of filters in our filter bank below 6 resulted in an unacceptable reduction in performance. For example, our best 3 band metric doubled the average error.

3.3 Six Non-Overlapping Filters

The six bins with the best performance just described spanned the frequency range from 150-6500Hz. However, it left a gap which ignored noise between 2.6-4.1Khz. We developed a final metric spanning the same 150-6500Hz frequency range with a six bin filter bank of non-overlapping mel-warped filters. While frequencies above 6500Hz were ignored, there was no gap in our identified range of interest. The correlation between predicted and actual DWER for this final metric is shown in Figure #5. This metric resulted in an increase in the average error from 3.75 to 4.7 on the development data but we suspected that it would generalize better to unseen data.

The steps in generating our DWER prediction which incorporates both a maximum and a minimum level of noise within a specified filter bank is as follows. Process the input audio through a bank of filters. Use the SPQA tool to calculate an SNR measurement for each of the filtered outputs. Calculate the difference between this SNR measurement and a filter specific maximum SNR threshold. Set any negative difference to zero and truncate any difference to a global maximum used for all filters. Use the resulting difference values in a linear regression. The particular regression coefficients, thresholds and maximum delta used for our ASR system are shown in Table #1.

	F6	F5	F4	F3	F2	F1	Int
Coefficients	0.94	0.31	0.80	1.96	3.76	-3.56	0.83
Threshold	29	30	35	23	25	21	
Max Delta	17						

Table 1: Regression coefficients used with the 6 filter bank prediction equation. F# is the filter bank (F1 lowest freq.), Threshold is the maximum SNR value (dB) used to calculate a difference and Max Delta is the maximum allowed SNR difference for any filter bank SNR.

Processing the outputs of all six filters in our filter bank requires 30% of the audio file duration when run on a SunBlade 1500. Further savings could be realized by extracting sub-segments of long files and performing the evaluation on only these representative samples.

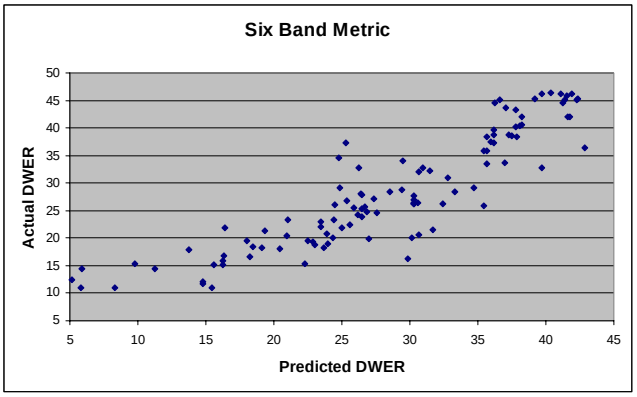

Figure #5 The predicted and actual DWER using the difference between the SNR measurement in six non-overlapping frequency bands and a maximum SNR threshold.

4. VALIDATION

To evaluate our two six band metrics on unseen data, we again used GoldWave™ to create 65 test files with a different type of colored noise than used to this point in developing SNR to DWER relationships. We filtered the white noise file used in earlier experiments to create narrower band limited noise files with bandwidth = 500Hz. Noise files with random relative amplitudes across the entire spectrum from 0-8khz were created as well as files using 3 or 4 of these narrow bins sprinkled across the full frequency range. The average error using the overlapping six bins with the range of uncovered frequencies grew to 26.1 while the non-overlapping contiguous filter bank only increased to 11.3.

Our final evaluation applied our best six band prediction relationship to two files collected from different domains than those used in the metric development. Eval1 was a five minute (711 words) recording of a conversation between three civilians and security personnel during a simulated exercise. Eval2 was an eight minute (463 words) recording of an individual providing situation reports during a simulated exercise. Both recordings were made with an Olympus™ DS-2 digital voice recorder. The performance for these evaluation files is shown in table #2.

File	WER	Predicted WER	Error
Eval 1	84.8	88.4	4%
Eval 2	85.5	83.7	2%

Table #2 Prediction performance of the six band metric applied to field data.

5. CONCLUSION

In this paper we have shown that it is possible to use an audio quality metric to estimate the word error rate in the output transcript of an ASR system. We have shown that the relative importance of noise at different frequencies makes it impossible to predict WER with a single speech to noise estimate across the entire frequency range. Additionally we have seen both that noise below a minimum threshold does not impact performance and that increasing noise beyond a maximum threshold does not further degrade performance. We have shown that an approach which identifies and appropriately weights the noise in critical frequency bands between these two thresholds can make a usable approximation of actual ASR system performance.

Future work should investigate the robustness of the described approach to a wider range of audio files and apply it to multiple ASR systems. While we anticipate that the actual regression coefficients found during our work will differ when the ASR system changes, we believe that the approach is independent of the ASR system.

6. REFERENCES

[1] S. Johnson, P. Jourlin, K. S. Jones, and P.Woodland. “Spoken document retrieval for TREC-8 at Cambridge University.” In *Proc. of the 8th Text REtrieval Conference*, Gaithersburg, MD, 1999.

[2] Johnson, S. E., P. Jourlin, G. L. Moore, K. Sp¨arck Jones, and P. C.Woodland. “The Cambridge University Spoken Document Retrieval System.” In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 49–52. 1999

[3] K. Papineni, et al, “Bleu: A Method for Automatic Evaluation of Machine Translation”, *Research Report RC22176*, IBM, Sept. 2001.

[4] Fu-Hua Liu, Yuqing Gao, Liang Gu, and Michael Picheny “Noise Robustness In In Speech To Speech Translation”, *Research Report* RC22874, IBM, 2003.

[5] http://www.nist.gov/speech/tools/index.htm, “spqa_2.3+sphere_2.5.tar.Z”