Why Speech Recognizers Make Errors?  
A Robustness View

Hong Kook Kim\textsuperscript{1} and Mazin Rahim\textsuperscript{2}

\textsuperscript{1} Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea  
\textsuperscript{2} AT&T Labs-Research, 180 Park Avenue, Florham Park, NJ 07932, USA  
hongkook@gist.ac.kr, mazin@research.att.com

Abstract
The performance of large vocabulary speech recognizers often varies depending on the input speech and the quality of the trained models. The particular attributes that cause recognition errors are a research area that has not been well studied. This paper addresses this issue from a robustness perspective using a large amount of field data collected from natural language dialog services. In particular, we present a method for tracking time-varying or nonstationary extraneous events, such as music, background noise, etc. We show that this measure is a better predictor of recognition errors than a standard measure of stationary signal-to-noise ratio (SNR). Combining the two measures provides a data selection algorithm for detecting problematic speech.

1. Introduction
Although most commercial automatic speech recognition (ASR) systems are trained on huge amounts of speech data, they still suffer from various kinds of robustness problems, resulting in a degradation in the word accuracy. These degradations are known to be attributed, but not limited, to background noise, coarticulation effects, channel distortion, accent and dialects [1]. The philosophy that this problem will go away with more data does not seem to hold. Most commercial recognizers are typically trained on hundreds if not thousands of hours of speech and still suffer from these same effects.

In an effort to improve robustness in ASR, several research labs have reported novel algorithms that have been shown to minimize the acoustic mismatch between the training model and the testing environment. These algorithms which include some form of either enhancement, normalization or adaptation operate either in the feature domain, e.g., [2], or in the model domain, e.g., [3][4]. Although much progress has been made in this area, the majority of the robustness scenarios that have been studied were essentially simulated by adding stationary noise, for example.

In this paper, we present new insights into the robustness problem by analyzing spoken dialog data that have been collected from various telephony-based customer care services. Callers were prompted to speak naturally when interacting with these services. In this paper, we use this data to study how recognition accuracy is affected by the problem of poor robustness. Our goal is to create a diagnostic tool that can provide a better insight of “why recognizers make errors.”

Our study of the effect of noise concludes that local variations in the SNR have more impact on deteriorating the ASR performance than a standard utterance-based SNR measure. The two measures combined provide an even better overall predictor of recognition error rate.

The organization of this paper is as follows. Section 2 describes the stationary SNR measure that we have adopted in our study. Section 3 provides an extension to our SNR measure that allows us to track local nonstationary behavior of noise. Section 4 presents experimental results demonstrating the correlation of stationarity and nonstationarity measurements of noise to the ASR performance. Finally, we summarize our findings in Section 5.

2. Stationary Quantity of Noise: Stationary Signal-to-Noise Ratio
In order to investigate the effect of environmental noise on the ASR performance, we first have to detect and measure background noise. There are essentially three different approaches for detecting background noise. The first approach has been developed in the context of speech coding. This approach, generally known as “voice activity detection (VAD),” has been widely used for classifying speech and nonspeech events [5]. VAD helps to reduce the transmission rate in speech coding applications. It has also been used for speech/silence segmentation to improve decoding speed in ASR [6]. Although VAD can be implemented cost effectively and with low latency, the algorithm is based on a set of thresholds that need careful attention when applied to ASR. The second approach for background noise detection is based on energy detection (e.g., [7]). Speech frames can be clustered based on their energy values as being either speech or silence. An energy histogram is constructed and a threshold is selected such that speech frames are best separated from silence. The third approach for detecting noise is through forced alignment, either by preserving the state segmentations during recognition, or by performing forced alignment.
with recognized transcriptions. An alternative representation of the third approach is to train a binary classifier or a Gaussian mixture model to separate speech from silence.

Let \( I(n) \) be the identifier for the \( n \)-th analysis frame of a given utterance, where \( I(n) = 1 \) indicates that the \( n \)-th frame belongs to the speech interval, and \( I(n) = 0 \) indicates that the \( n \)-th frame belongs to the silent interval. If the number of frames for the utterance is \( L \), the average log energy for the speech intervals, \( SP \), is given by

\[
SP = \frac{1}{\sum_{n=1}^{L} I(n)} \sum_{n=1}^{L} E(n) |_{I(n)=1},
\]

where \( E(n) \) is the log energy of the \( n \)-th frame, which is defined as \( 10 \log_{10} \sum_{i=1}^{N} s^2(n) \) where \( N \) is the number of samples in a frame and \( s(n) \) is a sample value. Similarly, the average log energy for the silent intervals, \( NP \), is computed by

\[
NP = \frac{1}{L - \sum_{n=1}^{L} I(n)} \sum_{n=1}^{L} E(n) |_{I(n)=0}.
\]

As a result, the stationary SNR measurement for the utterance is estimated as the difference between the average log energies of speech and silent intervals:

\[
SNR = SP - NP.
\]

Fig. 1 shows the procedure for estimating SNRs for the three approaches outlined earlier. The particular VAD algorithm used in this study has been standardized for the global system for mobile communications voice services [8]. An example is illustrated in Fig. 2 for the utterance “I need to inquire about a bill that was sent,” spoken by a female. The last interval of this utterance contains significant background events. The estimated SNRs were 30.67 dB, 36.29 dB, and 23.52 dB for VAD, energy clustering, and forced alignment, respectively. Note that both VAD and energy clustering consider the last interval to be speech, whereas forced alignment tends to be more robust to such effects. In general, we have experienced the forced alignment approach to be more robust for speech/non-speech detection.

3. Time-Varying Quantity of Noise: Nonstationary SNR

The stationary SNR measurement described in the previous section does not reflect the local characteristics of environmental noise and in many cases such measurements can be misleading. For example, although the utterance in Fig. 2 is of high SNR, the average SNR measurement would be low due to the highly nonstationary noise signal in the last part of that utterance. Tracking such nonstationarities is essential for providing a more accurate SNR measurement.

In this paper, we propose a nonstationary measure to quantify the variation of SNR along an utterance, which is referred to as nonstationary SNR (NSNR). NSNR is defined as the standard deviation of noise power normalized by the average signal power, and computed by

\[
NSNR = \left( \frac{1}{L - \sum_{n=1}^{L} I(n)} \sum_{n=1}^{L} (SP - E(n))^2 |_{I(n)=0} - SNR^2 \right)^{\frac{1}{2}},
\]

where \( SP \) and \( SNR \) are the average log energy of speech intervals and the stationary SNR computed from (1) and (3), respectively. From (4), NSNR, which is measured in dB, becomes smaller when the average of the frame-dependent SNR, defined by \( (SP - E(n)) \), approaches the SNR measurement. This implies that smaller variations in the noise characteristics among different frames would result in low measurement of NSNR.
4. Effect of Stationary and Nonstationary SNRs on ASR Performance

In this section, we investigate the impact of SNR and NSNR on the ASR word accuracy across a wide range of customer care speech data.

Our baseline ASR system uses an extended phoneme set that is suitable for multiple task-domains. The acoustic model was built from telephone speech collected from over twenty different datasets [9]. For testing, we used 5171 utterances (54658 words) from a deployed AT&T customer care service. This dataset contains utterances with various types of distortion and background noise, each spoken by a different user. All experiments were performed using AT&T Watson speech recognition system. The language model used in this experiment was a trigram statistical language model. The word accuracy for this dataset was measured at 73.9%.

Fig. 3 shows the word accuracy as a function of estimated SNR when applying VAD, energy clustering, and forced alignment methods. These measurements are computed over the test data and are drawn at 5 dB intervals. Over half of the utterances belonged to the SNR range of 25 ~ 30 dB. The number of utterances whose estimated SNRs were less than 15 dB or greater than 35 dB was very small. The average word accuracy of 73.9% is considered as the “Reference.” As expected, the figure shows a strong correlation between word accuracy and estimated SNR for all three methods. Utterances that scored over 30 dB SNR seem to have a higher word accuracy than the Reference point. On the other hand, the word accuracy seems to be severely degraded for utterances below 20 dB. As a result, we conclude that the estimated SNR is a reasonable predictor of word accuracy and increasing it would translate to improved ASR performance.

In order to study the effect of variations in the noise characteristics on the ASR performance, we measured the standard deviation of noise powers of each utterance and classified all the utterances based on this measurement into five categories, namely [0, 2.5], [2.5, 5], [5, 7.5], [7.5, 10], and [> 10]. In particular, we took only the utterances whose estimated SNR was in a range of 20 ~ 25 dB. Fig. 4 shows the word accuracy according to the degree of nonstationarity when the estimated SNR is in the range of 20 ~ 25 dB. The average word accuracy in this interval is 72%. The figure clearly shows there is a strong correlation between the standard deviation of the noise power and the word accuracy. In particular, this is an inverse relationship. It is interesting to note that lower deviations seem to result in much higher word accuracy than the average measurement for that interval of SNR. At less than 5 dB measurement, the word accuracy is 6 ~ 13% higher than the reference.

As stated earlier in the paper that one goal of our study is to be able to understand why recognizers make errors. Therefore we decided to combine both SNR and NSNR measurements to study their prediction power of the word accuracy. The combined effect of SNR and the nonstationarity measure of noise on the ASR performance is shown in Fig. 5.

This figure was obtained by computing the NSNR measurement for each SNR interval. The intensity bar represents the word accuracy. This provides a 3-D diagram. The size of each interval in this figure represents the amount of utterances available for computation. The diagram shows that higher word accuracies are obtained at high SNR and low NSNR. At a given SNR interval, the word accuracy seems to vary significantly as a function of SNR. In summary, one can conclude that the two measurements are complementary and highly correlated to word accuracy.

One approach to combining those two measurements was to apply a linear regression model. The estimated word accuracy can be represented by

$$\hat{\text{word accuracy}} = \alpha \cdot \text{SNR} + \beta \cdot \text{NSNR} + \delta,$$

where $\text{SNR}_i$ and $\text{NSNR}_j$ are the stationary SNR and nonstationary SNR corresponding to the x-axis and y-axis in Fig. 5, respectively.

The parameters, $\alpha$, $\beta$, and $\delta$, are computed by minimizing the least square of the word accuracy $(1/N) \sum_{i,j=1}^{N} e_{i,j}^2$, where $e_{i,j}$ is the error between the estimated and actual word accuracy for each utterance.
current extending our approach to be able to investigate the impact of other speech quality features such as gender, accent, and loudness on ASR performance. Our findings will help us to understand why recognizers make errors and enable us to explore new algorithms for alleviating these effects.

6. References