Degradation of Speech Recognition Performance over Lossy Data Networks

Dimitris Pratsolis  
Department of Electronic and  
Computer Engineering  
Technical University of Crete  
Chania, Greece  
pratsdim@telecom.tuc.gr

Nikos Tsourakis  
Department of Electronic and  
Computer Engineering  
Technical University of Crete  
Chania, Greece  
tsourak@telecom.tuc.gr

Vassilis Digalakis  
Department of Electronic and  
Computer Engineering  
Technical University of Crete  
Chania, Greece  
vas@telecom.tuc.gr

ABSTRACT
In this work we investigate the effects of lossy data networks on the speech recognition performance, utilizing a stock information corpus. Within the context of the current paper we present the models of the lossy channel that were used in order to perform the specific simulations. The resulted voiceprints, after the transmission through the lossy channels, were fed to an Automatic Recognition System (ASR) in order to calculate the degradation in the performance. The specific procedure can be very helpful in extracting information that can be used for the design and the parameter tuning of the underlying data network.

Categories and Subject Descriptors

General Terms
Algorithms, Measurement, Performance

Keywords
Speech recognition performance; Lossy data networks; Gilbert-Elliot model; Three-state Markov model.

1. INTRODUCTION
The trend of the continuously increasing use of data communication is expanding to the mobile wireless world, as it has already taken place in the world of landline communications. The need for data access is growing and so is the demand for new applications, especially multimedia, that take advantage of the offered medium.

As wireless data networks evolve, the design of new communication protocols increases in size and complexity. The proper evaluation of the current networks provides the basis for the optimization of future protocols. A number of techniques are available for modeling and simulating the channel conditions of the underlying data network. The most common techniques include simulation, analysis of empirical data and analytical models [1].

Gilbert [5] appears to be the first to present a burst error model utilizing a Markov or multi-state model, a work that was later extended by Elliot [6] and Cain and Simpson [4]. Higher state Markov models that represent error distributions were also described by Blank and Trafton [3]. Others followed a different approach with the identification of the statistical distribution of gaps, e.g. hyperbolic distributions [9] and Pareto distributions that model inter-error gaps [2]. Lewis and Cox [7] concluded that there is a strong positive correlation between adjacent gaps in measured error distributions. In IP networks the packet loss, typically caused by congestion, is modeled with similar techniques.

In our work we focus on packet-based IP networks, as they are becoming an attractive alternative especially for wireless voice communications.

The volume of the required data in order to perform simulations and the availability of the network being tested, which may even be under design or deployment, often prohibit the utilization of real traces. One can consequently generate synthetic traces (e.g. from voiceprints) that simulate different conditions of the network and perform the necessary experiments. This is the procedure that we chose to follow and will describe in the following sections.

2. SIMULATION MODELS
In the context of our work we incorporated two error models, namely the Gilbert-Elliot model and a Three-state Markov model.

2.1 Gilbert-Elliot Model
The Gilbert-Elliot (GE) [5][6] channel model is a two-state Markov model, which is widely used to simulate the bursty packet loss behavior. This channel model has been shown to be able to effectively capture the bursty packet loss behavior of the Internet and wireless channels.

Furthermore, the special structure of the Markov model makes it analytically tractable. The two states of the GE model are denoted as G (good) and B (bad), as illustrated in Figure 1. In state G, packets are assumed to be received correctly and timely, whereas in state B, packets are assumed to be lost.
3. EVALUATION PROCESS

The word error rate (WER) is a commonly used metric to evaluate speech recognizers. A different error metric and in some cases more appropriate, is the natural language error rate (NLER). Generally, in speech recognition applications, we usually are interested in the interpretation of a spoken utterance rather than its accurate transcription. The NLER is simply defined as the number of NL errors occurred during the examination of the utterances, divided by the number of the reference utterances expressed with the following equation:

\[
\%\text{NLER} = \left( \frac{\text{NL errors}}{\text{utterances}} \right) \cdot 100\%
\]

During the evaluation process we calculated the NL error rate after processing a test set derived from a stock information application, which consisted of 1000 utterances (8KHz, 8 Bit, PCM). We applied the error models discussed earlier, using different configuration parameters and simulated different strategies of the ASR system for handling transmission errors. The configuration parameters for each model are the transition probabilities presented earlier. The recognizer can incorporate three different strategies for a missing packet:

1. Replacing the missing packet with silence.
2. Ignoring the missing packet.
3. Replacing the missing packet with the previous one.

Each state in the error models corresponds to a time slot of 20 ms, which is associated with the transmission of a packet of 160 bytes. Thus, a packet transmission is successful only if the error model is in state G for all slots needed for the packet to be transmitted, while it fails otherwise. Each utterance is therefore split into packets of 160 bytes, and the error models determine which one of them will be received by the ASR.

In the Gilbert-Elliot model two parameters must be defined, namely the packet loss probability \( P_b \) and the average burst length \( L_b \). For each parameter, the range of possible values is provided and each pair constitutes the current configuration of the model. The parameter values used in our simulations are summarized in the following table:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Min</th>
<th>Max</th>
<th>Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_b )</td>
<td>1%</td>
<td>15%</td>
<td>1%</td>
</tr>
<tr>
<td>( L_b )</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

After executing 50 repetitions for each configuration, the number of different experiments derived from the specific model simulations is:

\[
(\text{values of } P_b) \cdot (\text{values of } L_b) \cdot (\text{ASR strategies}) \cdot (\text{repetitions}) = 15 \cdot 4 \cdot 3 \cdot 50 = 9000
\]

We used a similar process for the Three-State Markov model, where \( P_{GB}, P_{BG_L}, P_{BG_S}, P_{BG_L} \) and \( k \) are the configuration parameters with a range presented in the table below:
Similarly, the number of experiments obtained from each configuration and for 50 repetitions is 3000.

For each experiment, we simulated the packet loss for all 1000 sentences of the test set and fed the resulting sentences to the ASR system in order to calculate the NL error rate. We should note that the system could recognize among 500 unique stock names uttered in different contexts.

4. EVALUATION RESULTS
In this section we will present the graphs obtained from the simulations with the two models. The specific results are compared with the NL error rate obtained from the original clean test set (baseline), which yields a NLER of 4.64%.

4.1 Gilbert-Elliot Model Evaluation Results
Initially, four graphs that correspond to the Gilbert-Elliot model will be presented. The NL error rate associated with each one of the three ASR strategies discussed earlier is depicted in Figure 3, 4 and 5. The specific error rate, as an average from all repetitions, is calculated with respect of the packet loss probability \( P_{gb} \) in the x-axis and the average burst error length \( L_b \). Each graph contains four plots that correspond to the different values of \( L_b \).

When \( L_b \) equals to 1 it is guaranteed that no consecutive error packets will be encountered in the packet error burst as \( P_{BB} = 0 \).

The strategy of replacing the lost packets with the preceding one yields to the best results, as in most cases there is a strong correlation of adjacent speech samples and consequently of adjacent speech packets. The worst results were obtained using the strategy of replacing the missing packets with silence.

In all configurations we calculated the amount of lost packets, which is more or less the same.

4.2 3-state Markov Model Evaluation Results
For the three-state Markov model the NL error rate associated with each one of the three ASR strategies is depicted in Figure 6, 7 and 8. The specific error rate is calculated with respect to the probability \( P_{GB} \) in the x-axis and the parameter \( k \). Each graph contains four plots that correspond to the different values of \( k \).

We should note that when \( k \) equals 0, the specific three-state Markov model becomes a Gilbert-Elliot model and no transitions to state LB take place.

Again the third ASR strategy yields to the best results and the calculated amount of lost packets is similar in the four configurations, which differ in the way they are distributed within the waveform.

The histogram in Figure 9 corresponds to \( P_{GB}=2\% \) and \( k=0.05 \) and presents the real effects of applying the model on the waveforms. In the x-axis the number of consecutive lost packets is depicted while the y-axis represents the total number of the corresponding error bursts that occurred after the application of the model (the average number from all repetitions). For the specific configuration we encountered 228 lost triplets and 137 lost quadruplets. The total number of packets in the test set is 79190.
Comparing the results from the two models, we can conclude that the Gilbert-Elliot model definitely offers smaller NL error rates for the same packet loss probability. For example for $P_{GB}=P_{GB}=5\%$ and using the strategy of replacing the lost packet with the previous one, we yield to NLER between 5%-10% for the GE model and 14%-16% for the three-state Markov model.

One can utilize specific models that better simulate existent data networks (GPRS, 3G etc). Ideally, the analysis should be performed on real voiceprints collected in the ASR server, after the transmission over the data network. The analysis would therefore provide an objective measurement of the expectations for a new speech recognition deployment.

6. ACKNOWLEDGMENTS

The corpus was provided by Dialogos Speech Communications.

7. REFERENCES


