The effect of network degradation on speech recognition

Gabriel Joubert and Etienne Barnard

Human Language Technologies Research Group
Meraka Institute/ University of Pretoria, Pretoria, 0001
s2107809@tuks.co.za, ebarnard@csir.co.za

Abstract

We describe a system, based on open-source tools, that was developed in order to study the effect of network degenerations in Voice-over-Internet-Protocol applications on speech-recognition accuracy. Sophisticated play-out algorithms are found to enhance recognition accuracy when network jitter occurs, but do not generally compensate for packet loss successfully. We also confirm that packet loss is a more significant problem for larger loss burst-lengths.

1 Introduction

As packet-switched communications networks become increasingly popular, VoIP (Voice over Internet Protocol) is predicted to become the standard means of spoken telecommunication. As a consequence, a significant amount of research has been undertaken on the effect of various packet-switching phenomena on the perceptual quality of speech (see, for example, [1]), and the effects of degenerating effects such as network delay, packet loss, and packet jitter – that is, variable packet delay – are fairly well understood. Less is known about the effects of such degenerations on automatic speech-processing algorithms, and with the growth of both packet-switched networking and automated speech services, this is certain to become a significant issue.

The current research therefore addresses two issues: on the one hand, we describe a simple test harness, based on open-source software components, which can be used to investigate the relationship between network degradation (the degenerating effect of the network) and speech-recognition accuracy. On the other hand, we also provide initial results for a simple recognition task, to indicate the magnitude of accuracy impairment that occurs under various states of network congestion.

2 Experimental setup

Figure 1 shows setup used to measure the effect of network traffic degeneration during a VoIP transmission, on speech-recognition accuracy. Sentences from the TIMIT database [2] were selected as basis for comparison. The open-source toolkit SOX [3] was used to code the samples in the u-law format, which is widely used for VoIP transmission. In order to emulate network conditions a trace file and a sample file in signed word format are required.

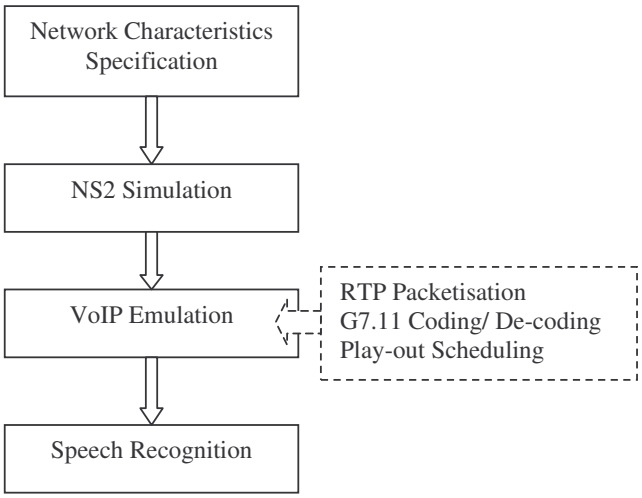

Figure 1.
VoIP performance measuring setup

The trace file was created using the NS2 network simulation tool [4]. Hoene [5] describes an add-on to NS2, which is able to produce packet traces. This add-on was enhanced in order to simulate the specific network degenerations of interest to us. These conditions include a jittering window, a packet loss rate and a discriminating delay in order to manipulate the packet reception order. Hoene further described a play-out scheduler-simulating tool. This tool is able to emulate the effect of network conditions on a sample sound file. The tool uses a combination of the PESQ [6] and ITU E-Model [7] to evaluate the quality of the resulting sample. We adapted the system to measure quality in terms of speech recognition. Since we are interested in the different components, which affect speech recognition, we kept the original tool's ability to accommodate more than one play-out scheduler. The output of the emulating tool is a sound file in the signed word format as described in [8]. The sound file is converted to Sphere wav format [8] for input to the speech-recognition system. The speech-recognition

system was trained using the HTK [9] toolkit. In order to evaluate the effect of different speech recognition algorithms we used tied tri-phone word detection (which will be referred to as hmm37) and monophonic word detection (which will be referred to as hmm09). We also implemented two different play-out algorithms. Ramachandran [10] proposed 4 play-out algorithms. The "exp-avg" algorithm, which estimates the mean delay through an exponentially weighted delay, the "fast-exp" algorithm that adapts quicker to increases in delays, the "min-delay", which uses the minimum delay of the packets received in the current talk spurt and the "spk-delay" which is developed to detect spikes. We used an implementation of the "spk-delay" play-out scheduler (which will be referred to as Ramjee), as described above. The other play-out scheduler we used implemented a "Fixed Delay-Spike" algorithm.

3 VoIP performance measuring tests

A series of tests were performed in an attempt to find relationships between various network conditions and the ability of a speech recognizer to accurately recognize the utterances. Our goal is to ultimately understand these relationships to such extend that speech recognition performance predictions can be made on account of the known network parameters. With "measure" we refer to the result of a speech-recognition algorithm's attempt to recognise an utterance.

3.1 Variation in loss burst length

In our first test we've examined the effect of altering the loss burst length of the emulated network. We kept the jitter window size at zero. The Ramjee play-out scheduler and the hmm37 speech-recognition algorithm were used for this test. Six utterances from the TIMIT speech database were used. We've followed the following steps for this test.

- We've measured the speech recognition performance of the 6 selected utterances.
- The utterances were sent through an emulated network with a loss rate of 10% and a loss burst length of **1 packet**. The resulting utterances were measured.
- The utterances were sent through an emulated network with a loss rate of 10% and a loss burst length of **10 packets**. The resulting utterances were measured.

3.2 Play-out schedulers and the effect of jitter

In this test we've examined the effect of jittering on different play-out algorithms. We've used the Ramjee play-out scheduler, the Fixed Delay-Spike play-out scheduler, the hmm37 speech-recognition algorithm and the same 6 utterances as in the previous test. The idea behind the test was to vary the jitter and study the effect on the system using different play-out schedulers. We've performed the following steps for this test.

- We've measured the speech recognition performance of the 6 selected utterances.
- The utterances were sent through an emulated network with a loss rate of 0% and a **jitter widow size of 0.005**, using the **Ramjee** play-out scheduler. The resulting utterances were measured.
- The utterances were sent through an emulated network with a loss rate of 0% and a **jitter widow size of 0.005**, using the **Fixed Delay-Spike** play-out scheduler. The resulting utterances were measured.
- The previous 2 steps were repeated for a jitter window size of 0.010.

3.3 Play-out schedulers and packet order

A continue delay (a constant delay) for all packets will have no effect on the speech recognition algorithm's ability to recognise a received utterance. It will delay a sound pulse but it will not change the shape of the pulse. According to the RTP protocol used by the emulator, packets are being transmitted at intervals of 0.02s. If a packet is delayed for longer than the interval time, it might arrive out of order. In this test we've studied the effect of out-of-order arrival on different play-out schedulers. The hmm37 speech-recognition algorithm was used for this test. The following steps outline the procedure for this test.

- We've measured the speech recognition performance of the 6 selected utterances.
- The utterances were sent through an emulated network with a loss rate of 0%, a delay rate of 10%, a packet delay magnitude of 0.15s (which corresponds to about 7 packets), a delay burst length of **1 packet** and a jitter widow size of 0s, using the **Ramjee** play-out scheduler. The resulting utterances were measured.
- The utterances were sent through an emulated network with a loss rate of 0%, a delay rate of 10%, a packet delay magnitude of 0.15s, a delay burst length of **1 packet** and a jitter widow size of 0s, using the **Fixed Delay-Spike** play-out scheduler. The resulting utterances were measured.
- The utterances were sent through an emulated network with a loss rate of 0%, a delay rate of 10%, a packet delay magnitude of 0.15s, a delay burst length of **10 packets** and a jitter widow size of 0s, using the **Ramjee** play-out scheduler. The resulting utterances were measured.

- The utterances were sent through an emulated network with a loss rate of 0%, a delay rate of 10%, a packet delay magnitude of 0.15s, a delay burst length of **10 packets** and a jitter widow size of 0s, using the **Fixed Delay-Spike** play-out scheduler. The resulting utterances were measured.

3.4 Performance prediction

In order to attempt VoIP speech recognition, performance prediction a lot of relationships need to be considered. For this initial investigation of performance prediction we've limited the random components of the network. A defined network, with reproducible characteristics was emulated. One hundred utterances from the TIMIT speech database were sent through this network and measured. Measurements for each utterance were taken at loss rate intervals of 4%, from 0% to 40%. The average degeneration at each loss rate was calculated. Emulated accuracy was predicted using the formula:

$$y = (D * x/100 - x) * (-1)$$

Where $x =$ original accuracy, $D =$ average degeneration and $y =$ predicted accuracy. The test was performed using the Ramjee play-out scheduler and the hmm37 speech recognition algorithm.

4 Results

The following results were for a relative small test sample of an urge research area.

4.1 Results: Variation in loss burst length

The "Orig." column refers to the original accuracy of the selected utterances and the "10%" refers to the specified loss rate. The "1" and "10" refers to the selected loss burst-length.

Orig. (%)	10% / 1 (%)	10% / 10 (%)
75.0	62.5	50.0
87.5	87.5	75.0
80.0	90.0	70.0
58.3	58.3	50.0
75.0	75.0	62.5
87.5	100.0	62.5

Table 1.
Variation in loss burst length.

4.2 Results: Play-out schedulers and the effect of jitter

The top row refers to the jitter window size, while the "R" and "S" refers to the Ramjee- and Fixed Delay-Spike play-out algorithms respectively.

0.005		0.010	
R/37 (%)	S/37 (%)	R/37 (%)	S/37 (%)
75.0	50.0	75.0	37.5
87.5	67.5	87.5	37.5
80.0	50.0	80.0	10.0
58.3	58.3	58.3	33.3
75.0	62.5	75.0	25.0
87.5	62.5	87.5	0.0

Table 2.
Play-out schedulers and the effect of jitter.

4.3 Results: Play-out schedulers and packet order

The "1" and "10" refers to the respective delay burst-lengths.

R/1 (%)	S/1 (%)	R/10 (%)	S/10 (%)
50.0	62.5	75.0	50.0
37.5	87.5	87.5	75.0
80.0	90.0	80.0	70.0
41.7	58.3	58.3	50.0
37.5	75.0	62.5	62.5
25.0	100.0	87.5	62.5

Table 3.
Results: Play-out schedulers and packet order.

4.4 Results: Performance prediction

The results of 20 test utterances:

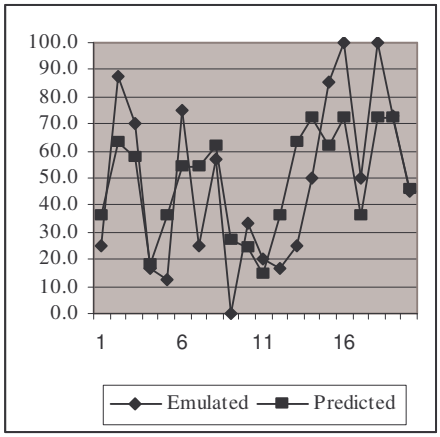

Figure 2.
Results: Performance prediction.

5 Discussion

Table 1 shows that longer loss burst-lengths causes larger degenerations. This may be either due to the G7.11's loss concealment algorithm or it may be that

that the speech recognition algorithm is still able to recognise some words, if only one packet is lost during a loss interval. From table 2 it is clear that adaptive play-out algorithms like Ramjee are successful in handling a moderate amount of jitter. Table 3 surprised us with the suggestion that the Ramjee play-out scheduler is able to adapt to bursts of out-of-order packets but not to individual delays. The predicted performance results as shown in figure 2, correlated with the emulated accuracies with a factor of 0.76. This correlation coefficient indicates that there is still plenty of room for improvement.

6 Conclusion and future work

We have shown that a number of open-source tools can be combined to create an effective tool for studying the effects of packet-switching degenerations on speech-processing algorithms. Although our focus has been speech recognition, it would be simple to extend this protocol to, for example, speaker verification or language identification.

Using this testing tool, we have seen that significant reductions in recognition accuracy occur at high packet losses, especially for large burst-lengths. It is also interesting that sophisticated play-out algorithms, which were designed for perceptual acceptability, do not necessarily improve recognition accuracy (although improved performance under packet jitter was found for the Ramjee algorithm). This work can be extended in a number of interesting ways. We have focused on a specific recognition task; these tests should be repeated for a range of grammars and vocabularies, in order to understand how factors such as perplexity interact with the network degenerations we have studied. It would be very useful if an algorithm could be developed to predict the level of degeneration based on factors such as perplexity [11], jitter, and packet loss. Finally, it seems as if play-out algorithms optimized for recognition accuracy rather than perceptual quality are a worthwhile topic for investigation.

7 Acknowledgments

We are grateful to Marelie Davel for various forms of assistance, including help with HTK.

8 References

[1] Sun, L. Speech Quality Prediction for Voice over Internet Protocol Networks. (A thesis submitted to the University of Plymouth in partial fulfilment for the degree of Doctor Of Philosophy). Plymouth: School of Computing, Communications and Electronics. University of Plymouth, 2004.

[2] The DARPA TIMIT speech database. URL: http:// www.mpi.nl/world/tg/corpora/timit/timit.html [Accessed: 10 November 2005].

[3] Sound eXcange sound converter (SOX). URL: http://sox/sourceforge.net [Accessed: 10 November 2005].

[4] The network simulator NS2. URL: http:// www.isi.edu/nsnam [Accessed: 10 November 2005].

[5] Hoene. C. Wietholter. S. Simulating Playout Schedulers for VoIP – Software Manual. TKN, Technical University of Berlin, Germany, September 2004.

[6] Itu-t. recommendation p.862. perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, 2 2001.

[7] Itu-t. recommendation g.107. the e-model, a computational model for use in transmission planning, 5 2000.

[8] Wavesoftware. URL: http:// www. Waveform-software.com/ SoX_Wrap/soxhelp.htm [Accessed: 10 November 2005].

[9] The Hidden Markov Model Toolkit (HTK). URL: http://htk.eng.cam.ac.uk [Accessed: 10 November 2005].

[10] Ramjee. R. Kurose. J. Towsly. D., Department of Computer Science, University of Massachusetts, Schulsrinne. H., GMD-Focus, Berlin, Germany, Adaptive Playout Mechanisms for Packetized Audio Applications in Wide Area Networks, Not dated.

[11] HLTsurvey. URL: http://cslu.cse.ogi.ed/ HLTsurvey/ch1node4.html [Accessed: 10 November 2005].