PERFORMANCE ANALYSIS OF DISTRIBUTED SPEECH RECOGNITION OVER IP NETWORKS ON THE AURORA DATABASE

D. Quercia[†], *L. Docio-Fernandez*[††], *C. Garcia-Mateo*[††], *L. Farinetti*[†], *J.C. De Martin*[§]

[†]Dip. Autom. e Informatica/[§]IRITI-CNR
Politecnico di Torino
C.so Duca degli Abruzzi, 24
I-10129 Torino, Italy
quercia|farinetti|demartin@polito.it

[††]E.T.S.I. Telecomunicacion[*]
Dpto. Teoria de la Señal y Comunicaciones
Universidad de Vigo
36200 Vigo-Spain
ldocio|carmen@gts.tsc.uvigo.es

ABSTRACT

We present results on the performance of Distributed Speech Recognition operating over simulated IP networks. ETSI AURORA front-end running at client nodes extracts the speech parameters, packetizes and sends them as real-time IP traffic to a remote recognizer based on Continuous Density Hidden Markov Models. The experimental framework is the ETSI STQ-AURORA Project Database 2.0. The impact of transmission over IP networks is modeled by (1) random losses, (2) losses generated by a Gilbert model and (3) network simulations. Results show that random losses and moderately bursty losses do not significantly affect the recognition performance. Strongly bursty packet losses, as those generated by real-time and Web traffic competing over a network bottleneck, instead, can have a very negative impact on recognition performance, indicating that DSR over the Internet, to be successful, requires high levels of Quality of Service.

1. INTRODUCTION

The increasing use of both the Internet and Automatic Speech Recognition (ASR) systems makes Internet-based Distributed Speech Recognition (DSR) services very attractive. Such services are based on a client-server architecture. Simple, low power, client devices quantize and packetize the speech data (usually in the form of speech feature vectors) and transmit it over the communication channel to a remote ASR server that performs speech recognition. The architecture of the service considered in this paper is shown in Figure 1.

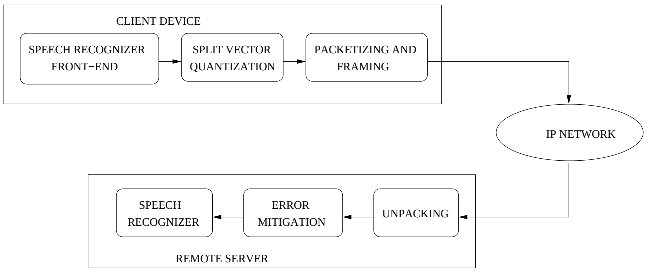

Fig. 1. Block diagram of an IP-based DSR system.

The design of a speech recognition system which operates over IP networks differs from the case of a system operating over the PSTN. When developing Distributed Speech Recognition (DSR) in the current Internet one must consider the effect of packet losses and delay. On congested IP networks, in fact, routers will discard packets if their packet in-flow exceeds their outflow for a given data route. Packet losses can be isolated single losses or successive multiple packet losses. It is, therefore, useful to study how packet losses affect the speech recognition performance and the effectiveness of error concealment techniques. As we will see, bursty losses can negatively affect the performance of IP-based DSR.

We present the results of a performance evalutation analysis of Internet-based DSR using the ETSI AURORA Project database. The IP network is modelled under increasingly more realistic scenarios, culminating in network simulations of a bottleneck network topology.

2. EXPERIMENTAL FRAMEWORK

In this Section, the experimental framework used to carry out the DSR experiments is presented and discussed. Different network scenarios have been considered, but one single task: speaker independent connected digit recognition.

[*]This work has been partially supported by the Spanish project TIC2000-1005-C03-02 “Tecnologías del habla multilingüe”.

2.1. Database

To evaluate and compare the performance of the DSR scenarios proposed in this paper, the ETSI STQ-AURORA Project Database 2.0 experimental framework was adopted [1] [2]. The source speech for this database is the TIdigits clean database, consisting of a connected digits task spoken by American English talkers. This database was downsampled to 8KHz and filtered with the G.712 “standard” frequency characteristic defined by the ITU [3].

For training the recognition models a set of 8440 utterances of the above speech are used, containing the recordings a total of 55 male and 55 female adult speakers.

For test 4004 utterances from 52 male and 52 female speakers in the TIDigits test set are split into four subsets with 1001 utterances in each. Recordings of all speakers are presented in each subset. Such test set is the referred in [4] as **clean-test set a**.

2.2. Front-End

The client front-end is based on the proposal by AURORA WI007. It consists of a cepstral analysis scheme where each feature vector has 14 components: 13 Mel frequency cepstral coefficients (MFCCs); the MFCC of order 0; and the logarithmic frame energy. For the cepstral analysis the following operations are applied over the speech signal:

- Signal offset compensation with a notch filter;
- Preemphasis with a factor of 0.97;
- Hamming windowing of 25 ms length;
- FFT based mel filterbank with 23 frequency bands in the range from 64 Hz up to half of the sampling frequency, i.e., 4 KHz;
- Frame shift of 10 ms.

2.3. Network scenarios

In order to measure the influence of missing speech packets on the ASR system performance three different IP network scenarios have been considered. Such scenarios simulate the packet losses produced by the IP channel.

2.3.1. Random losses

We have investigated the recognition performance when the IP network is modelled as a random loss channel, i.e., each packet has the same loss probability.

Various amounts of random packet loss ranging from 10% to 40% have been simulated. For low packet loss ratios (PLR), packet losses are predominantly single packet losses. Approximately 94% of the bursts consists of four packets or less.

2.3.2. Gilbert–model losses

Packet losses are not independent on a frame-by-frame basis, but appear in bursts. Bolot [5] studied the distribution of packet loss in the Internet and concluded that this could be approximated by a Markovian loss model such as the Gilbert or Elliott models. Thus, we have simulated the IP network by using a 2-state Markov model, known also as a Gilbert model. The tests were run under the loss conditions reported in Table 1, where $ulp = p/(p+q)$ is the unconditional loss probability and $clp = 1 - q$ is the loss probability conditioned on the event that the previous packet was lost. Conditions 1 and 2 exhibit predominantly solitary losses and fairly insignificant number of burst losses. Approximately 90% of the bursts at conditions 1 and 2 consist of three packets or less; while at conditions 3 and 4 the 90% of the bursts consist of five packets or less.

condition	1	2	3	4
clp	0.147	0.33	0.50	0.60
ulp	0.006	0.09	0.286	0.385

Table 1. Channel IP gilbert-based loss conditions.

2.3.3. Network simulations

Finally, we have analyzed the recognition performance when the DSR is carried out over a bottleneck network topology, with bottleneck bandwidth in the environs of 64 Kb/s, similar to the bandwidth of an ISDN- or dial-up connection. The protocol used to send and receive speech packets was the Real-time Transport Protocol (RTP). Our analysis was targeted at examining a scenario in which the user is speaking, while at the same time interfering FTP (e.g., Web or email) traffic is going on. Some characteristics of this scenario are the following:

- Competing traffic: on/off TCP sources;
- Playout buffer: 100 ms;
- Duration: 350 s.

To evaluate performance, we used the network simulator ns-2 from UCB/LBNL [6]. All the simulations used a simple bottleneck topology. Three FTP and three RTP sources are placed at one end of the bottleneck link and the six related receivers are placed at the other end. The routers associated with the bottleneck link use a droptail strategy and have a buffering capacity of 2.5kB. In the random and Gilbert loss scenarios we have assumed that the client device transmits speech using the MFCC feature vectors with one frame (10 ms) per packet. In this bottleneck scenario we have considered three different packet sizes, namely: 1, 2 and 5 frames per packet (8, 14 and 32 bytes respectively, with RTP header compression). For each scenario, different values of the bottleneck bandwidth were set so that the speech packets loss ratio was equal to approximately 5%, 10%, 15%, and 20%, for a total of 12 network simulations.

Since TCP packets are much larger than speech ones, when speech packets find a TCP packet in front of them, they get delayed and may reach the receiver too late for playback, resulting, as we will see, in long bursts of packet losses. Examination of each simulation reveals that they exhibit predominantly bursty losses. Figure 2 shows the distribution of the burst lengths when the number of frames per packet is one (condition 1). The examination of this figure reveals extremely long error bursts: the mean burst length is approximately 45 packets.

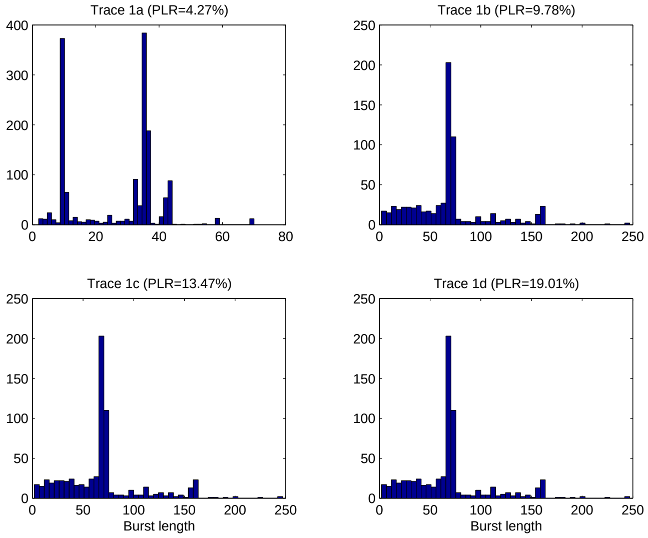

Fig. 2. Burst-length distribution, network conditions 1a-d.

2.4. Recognizer

The reference recognizer is based on Continuous Density Hidden Markov Models (CDHMM). The digits are modeled as whole word HMMs with the following parameters:

- 16 states per word;
- left-to-right models without skips over states;
- mixture of 3 Gaussians per state;
- diagonal covariance matrix.

Two pause models are defined. The first one called “sil” consists of 3 states with a mixture of 6 Gaussians each one. It should model the pauses before and after the utterance. The second pause model called “sp” is used to model pauses between words. It consists of a single state which is tied with the middle state of the “sil” model.

During recognition an utterance can be modeled by any sequence of digits with the possibility of a “sil” model at the beginning and at the end, and a “sp” model between two digits.

For recognition the MFCC of order 0 is not considered. Only the remaining 13 components as well as the corresponding delta and acceleration coefficients are considered.

3. RECOGNITION RESULTS

The recognition results are presented in this Section when applying both the front-end and the recognition scheme as described above.

The experiments performed aimed to show recognition performance for a variety of different network scenarios; and to measure the influence of missing speech packets on the ASR system performance. The packet loss rate and the length of bursts were taken into account.

The total number of complete sentences which were recognized correctly as well as the correct and accuracy percentages are

$$97.10/99.40/99.02$$

for the baseline system without losses. Tables 2, 3, and 4 refer to results when applying the network scenarios described in the previous Section. The results show that the worst performance is obtained when the speech is sent through the bottleneck scenario.

If the packet losses are predominantly solitary losses or the length of burst losses is small, as in the proposed *random loss* and *gilbert conditions* scenarios, there exists a range of traditional error concealment methods which can easily conceal the errors due to the packet loss, resulting in an optimal or at least improved speech recognition performance [7]. Over these scenarios we have examined the effectiveness of the *repetition* error concealment approach to improve the performance.

Table 2 shows the performance of the ASR system when the random loss network is considered. The results show that as packet loss increases performance deteriorates. Good performance recovery is shown with the repetition error concealment technique. With 40% packet loss the baseline performance is down to 83% and is restored to 99%.

Table 3 shows the recognition performance when the network is simulated by the Gilbert model. We can see the impact of packet loss on recognition performance for condition 4, where the loss are predominantly bursty and the mean burst length is 4 packets. As in random loss scenario the repetition error concealment method results in only slight degradation in performance as compared to baseline system without losses.

Table 4 shows the recognition results obtained when the IP network is implemented as the bottleneck topology described in the previous Section. Examination of the recog-

	PLR (%)			
	10	20	30	40
(a)	96.00	91.68	81.97	64.59
	98.93/98.67	97.29/97.15	93.13/93.08	83.56/83.53
(b)	97.10	96.95	96.78	96.45
	99.41/99.03	99.34/98.98	99.24/98.88	99.24/98.81

Table 2. Recognition performance, random packet losses: (a) Without error concealment; (b) With error concealment. First row: %Correct sentences; Second row: %Word correct/%Word accuracy.

	Loss condition			
	1	2	3	4
(a)	96.57	95.38	77.50	57.92
	99.33/98.99	98.75/98.44	91.09/91.01	80.00/79.98
(b)	96.85	96.93	96.18	95.18
	99.42/99.06	99.35/98.96	99.07/98.65	98.73/98.31

Table 3. Recognition performance, Gilbert-model losses: (a) Without error concealment; (b) With error concealment. First row: %Correct sentences; Second row: %Word correct/%Word accuracy.

nition results reveals that the number of deletion errors is very high. This is to expected given the characteristic of the packet losses, bursty with long bursts. The repetition receiver-based recovery methods do not work when the packet losses consist of long bursts: when a significant part of speech signal is lost, nothing can be done, from the acoustic point of view, to improve the recognition performance.

4. CONCLUSIONS

We have analyzed the impact of packet losses on the performance of an Internet-based DSR system using the ETSI AURORA database. Packet losses were modelled by (1) random losses, (2) losses generated by a 2-state Gilbert model and (3) network simulations of a bottleneck topology with interfering FTP traffic.

The results indicate that even relatively high levels single packet losses can be tolerated; however, if packet losses are strongly bursty, as it may happen in the Internet, the consequences can be very negative.

The results also show that for single packet losses and short bursts the repetition-based error concealment technique provides good performance.

Further work will be devoted to examine the packet loss profiles in more detail. The main focus of attention will be on techniques to combat bursty packet losses.

Trace	Without losses	With losses
1a	96.74	47.45
	99.47/99.08	83.97/83.70
1b	97.14	30.37
	99.60/99.15	65.66/65.58
1c	96.68	29.99
	99.52/98.95	63.71/63.53
1d	96.59	30.67
	99.38/99.01	60.71/60.54
2a	96.89	55.49
	99.49/99.13	88.24/87.94
2b	96.66	35.73
	99.45/99.04	67.01/66.74
2c	97.22	27.25
	99.52/99.14	49.57/49.47
2d	97.18	28.76
	99.46/99.23	54.72/54.67
3a	96.00	47.08
	99.42/98.85	78.31/77.98
3b	97.58	33.84
	99.70/99.40	57.94/57.86
3c	96.72	30.85
	99.68/99.08	51.48/51.38
3d	95.19	29.37
	99.37/98.61	54.33/54.08

Table 4. Recognition performance, network simulations.

5. REFERENCES

[1] ETSI ES 201 108 V1.2.2, "Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," Tech. Rep., ETSI, 2000.

[2] H.G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, Sept. 2000.

[3] ITU recomendation G.712, "Transmission performance characteristics of pulse code modulation channels," *ITU-T*, Nov. 1996.

[4] "Aurora project database 2&3," http://www.icp.inpg.fr/ELRA/aurora2.html.

[5] J.C. Bolot, "End-to-end frame delay and loss behavior in the Internet," in *Proc. ACM SIGCOMM*, Sept. 1993, pp. 289–298.

[6] LBL, http://www.isi.edu/nsnam, *Network simulator*.

[7] B. Milner and S. Semnani, "Robust speech recognition over IP networks," in *Proc. of IEEE ICASSP*, June 2000, pp. 1791–1794.