RECOVERING OF PACKET LOSS FOR DISTRIBUTED SPEECH RECOGNITION

Pedro Mayorga, Richard Lamy and Laurent Besacier
{Pedro.Mayorga-ortiz, Richard.Lamy, Laurent.Besacier}@imag.fr

GEOD Team (Groupe d'Etude sur l'Oral et le Dialogue)-CLIPS Laboratory
Université Joseph Fourier-Institut National Polytechnique de Grenoble
BP 53 –38041 GRENOBLE Cedex 9
(FRANCE)

ABSTRACT

This work deals with the packet loss problem in a *Distributed Speech Recognition* architecture. A packet loss simulation model is first proposed in order to simulate different channel degradation conditions. In these conditions, the performance of our continuous French speech recognition system is evaluated for packets containing different numbers of speech feature vectors. Several reconstruction strategies, to recover lost information, are proposed and evaluated. The results first confirm the intuitive fact that the word error rate obviously increases with the size of the lost packets and with the channel degradation level. However, it is shown that simple reconstruction strategies allow to recover acceptable performance. The most efficient ones are those using interleaving technique to distribute the speech information among packets, combined with interpolation methods to estimate lost acoustic features.

1. INTRODUCTION

Speech recognition technology tends to be more and more embedded in mobile phones or other portable communication terminals. If keyword recognition for basic commands (isolated speech recognition for small vocabulary) can be performed locally on the terminal with DSP processors, it is generally not the case for large vocabulary continuous speech recognition which implies more complex treatments and the use of huge language models (for instance, a 60000 words trigram language model requires about 300 Mbytes of RAM). In this case, a distant speech recognition server is accessed by the terminal which needs to use continuous speech recognition technology. For some very low consumption devices, with few memory available to store acoustic models, this concept of distant or remote recognition is still interesting, even for costless tasks like keyword recognition or speaker verification technologies.

Different strategies can be proposed to access a distant speech recognition server. Recently, some solutions were proposed where speech spectral analysis (the extraction of acoustic features for recognition, which is a low-cost task) is performed on the client terminal and the acoustic features are directly transmitted to the distant speech recognition server. The expression *Distributed Speech Recognition*[1] is generally used to define this type of architecture [1]. This approach allows to avoid speech degradation due to speech coding before transmission from the client to the server, since the acoustic parameters are extracted on the client clean signal before being transmitted to the recognition server. However, this architecture does not solve the transmission errors that can occur on the network. For instance, a data packet transmitted from the client to the server can be lost on GSM wireless network or on the Internet.

This paper is mostly related to the packet loss problem in a Distributed Speech Recognition architecture. Firstly, a realistic packet loss simulation model is proposed in order to evaluate the ASR performance degradation that can occur for different data transmission conditions (*section 2*). Then, different reconstruction strategies are proposed to cope with packet loss problem (*section 3)*. These methods are evaluated on our continuous french speech recognition system (*section 4*). Finally *section 5* gives some conclusions and draws some perspectives.

2. PACKET LOSS SIMULATION: THE GILBERT MODEL

In order to repair the lost packets in the audio data flows, it is necessary to simulate how the packets are lost on the network. If we suppose that the speech feature vectors are transmitted over the Internet, the process of audio packets loss can be characterized with the Gilbert model [2] [3] of two states (*figure 1*). One of the states (state 1) represents a packet loss, the other state (state 0) represents the case where packets are correctly transmitted : p is the probability of going from state 0 to state 1, and q the probability of going from state 1 to the state 0.

This model is then characterized by two parameters, *p* and *q*, which are the transitions probabilities between both states. Different values of p and q define different packet

[1] http://www.icp.inpg.fr/ELRA/aurora.html

loss conditions that can occur on the Internet. The probability that n consecutive packets are lost is $p(1-q)^{n-1}$. If (1-q)>p, the probability of losing a packet is greater after having already lost a packet than after having successfully received a packet [3] ; this is generally the case on Internet data transmission where packet losses occur as bursts. Note that $p+q$ does not necessarily equal *1*. When *p* and *q* parameters are fixed, the mean number of consecutive packets lost can be easily calculated and is equal to p/q^2. The higher is this quantity, the stronger should be the degradation.

For our experiments, the different values of *p* and *q*, representative of what can really occur [2][3], are given in *Table 1*.

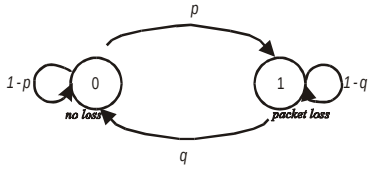

Figure 1: Gilbert Model

condition	1	2	3	4	5
p	0,10	0,05	0,07	0,20	0,25
q	0,70	0,85	0,67	0,50	0,40
p/q²	0,20	0,07	0,15	0,80	1,57

Table 1: Different packet loss conditions

3. RECONSTRUCTION STRATEGIES

An error control mechanism is needed when the number of lost packets is so important that the efficiency of a recognition system is degraded. The most typical mechanisms belong to two classes. Firstly, the mechanisms of Automatic Repeat reQuest (ARQ) are closed-loop mechanisms, which are based on the retransmission of the packets that did not arrive at their destination. The lost control algorithms by automatic retransmission ARQ cannot be used for the audio transmission in real time. Secondly, the Forward Error Correction mechanisms (FEC) are open-loop mechanisms based on the transmission of redundant information [4] [5]. In order to protect against the lost of packets, it is necessary to transmit the redundancy of the last n-1 packets in packet n. Another approach of the audio transmission is the technique of interleaving. This mechanism is based on the distribution of samples, of such form that if a packet is lost, the data can be reconstructed by a repair mechanism [5], for example the calculation of the average (simple interpolation) or the repetition technique. The reconstruction techniques used in our experiments are more precisely described in the following sub-sections.

3.1. Interleaving (EN)

The technique of interleaving distributes the effect of the lost packets; that means that the information of a speech part is distributed in the other packets [5] as shown in *Figure 2*.

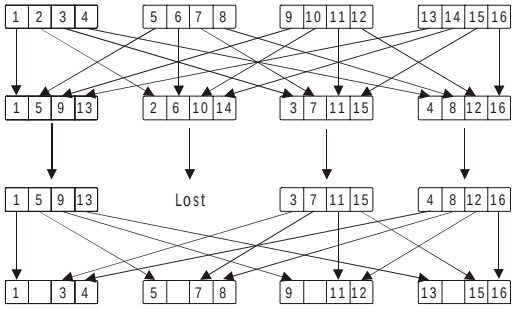

Figure 2: Interleaving Technique

The speech units are regrouped in crossed form before their transmission, in such a way that the units are distributed and separated to supply a distance between the lost ones. In the receiver, they are rearranged into their original form. As we can see, the lost of a packet results in the lost of several speech units distributed in the other packets.

3.2. Repetition (RSEN)

The repetition technique consists in replacing the lost packets by copies of the last received packet. Here we have a tradeoff between the low complexity of calculation and a reasonably good performance [5]. One can observe the process of repetition in *Figure 3*.

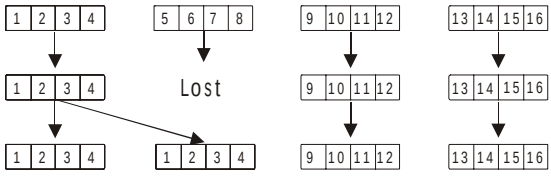

Figure 3: Repetition Technique

3.3. Simple Interpolation by Averaging (ISEN)

Simple interpolation or calculation of the average consists in interpolating by using the packets after and before the lost packet [6], as is shown in *Figure 4*. The advantage of this technique is its simplicity; nevertheless its efficiency decreases as the number of lost packets increases.

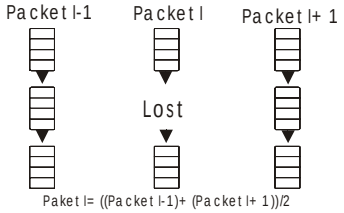

Figure 4: Simple Interpolation Technique

3.4. Interleaving With Repetition (REN, VREN)

In this method the idea is to apply the technique of interleaving before sending the data and the technique of repetition as a repairing technique in the receiver. *Figure 5* shows the process. The repairing can be applied either before rearranging the features (REN) or after (VREN).

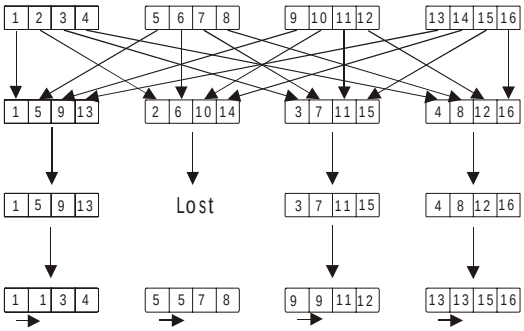

Figure 5: Technique of Interleaving with Repetition

3.5. Interleaving With Average Calculation (IEN, VIEN)

In this method, the idea is to apply the technique of interleaving before sending the data and then, at the time of its arrival to the receiver, the technique of average calculation. *Figure 6* shows the process. The repairing can be applied either before rearranging the features (IEN) or after (VIEN).

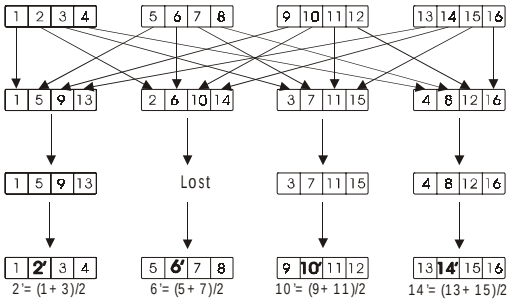

Figure 6: Technique of Interleaving with Average

4. EXPERIMENTS AND RESULTS

4.1. French speech recognition system

Our continuous French speech recognition system RAPHAEL uses Janus-III toolkit [7] from CMU. The context dependent acoustic model (750 CD codebooks, 16 gaussians each) is learned on a corpus which contains 12 hours of continuous speech of 72 speakers extracted from Bref80 [8] database. The system uses 24-dimensional LDA features obtained from 43-dimensional acoustic vectors (13 MFCC, 13 ΔMFCC, 13 ΔΔMFCC, E, ΔE, ΔΔE, zero-crossing parameter) and extracted every 10ms. These 24-dimensional vectors are the feature vectors transmitted through the network to the speech recognition server. The vocabulary (5k words) contains nearly 5500 phonetic variants of 2900 distinct words ; it is specific to reservation and tourist information domain. The trigram language model that we used for our experimentation was computed using Internet documents because it was shown that they give very large amount of training data for spoken language modelling [9].

4.2. Test database

We conducted a series of recognition experiments with 120 sentences focused on reservation and tourist information (CSTAR120 database) used in Cstar project[2].

4.3. Experiments

The first experiment we have made is about the influence of the size of packets and of the degradation conditions on the speech recognition performance. For different packet sizes (packets of 1,2,3 or 6 feature vectors) and the different degradation conditions of the Gilbert model (conditions 1,2,3,4 and 5 described in *Table 1*), we have applied the different reconstruction techniques. *Figure 7* and *Figure 8* present results of this experiment, that is the word error rate (WER) in function of packets size (*Figure 7*), and in function of degradation conditions (*Figure 8*). For both figures, the solid line gives the rate obtained without any reconstruction technique whereas the blocks show the results after applying separately each reconstruction technique of *section 3* and averaging the results. These results first confirm the intuitive fact that the WER obviously increases with the size of the lost packets and with the degradation level. The more interesting result is that the gain of recognition, when reconstruction techniques are applied, increases significantly with the packets size: a gain from 2% for packets of single vector to 8.5% for packets of 6 vectors. The same point can be noticed for the degradation conditions: a gain from 0.2% for the softest condition to 12% for the strongest condition.

The second experiment is performed to show the efficiency of the different reconstruction techniques. We have applied reconstruction techniques on the signals degraded according to the different conditions. *Figure 9* presents the mean performance of each reconstruction strategy averaged over each degradation condition and each packet size. In this graphic, we have two reference lines. The solid one gives the WER without any degradation. The doted one gives the WER obtained for degraded signals without any reconstruction. For comparison purpose, results obtained with a well-known technique, which transmits the sum of consecutive packets

[2] http://www.c-star.org

as redundant information, are also given (RA technique). When regarding these results, we can notice that the 8 techniques have obviously not the same effect. VREN, VIEN and IEN techniques give the best results. Except the RA technique that gives no significant gain (0.7% better than without reconstruction techniques), the 4 remaining techniques bring a gain of about 4%.

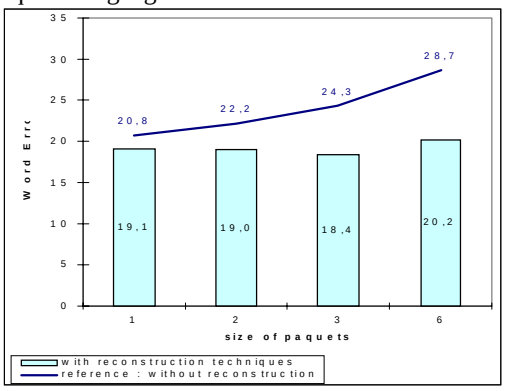

Figure 7: Size of packets vs degradation (WER%)

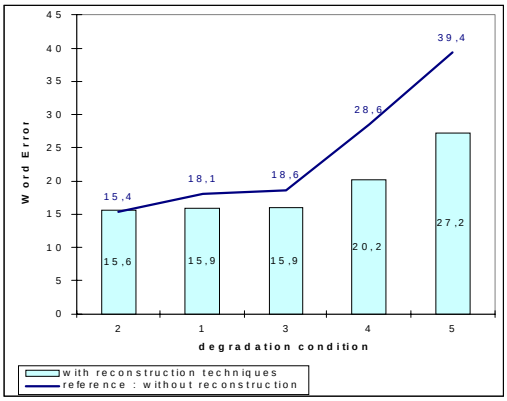

Figure 8: Network Condition vs Degradation (WER%)

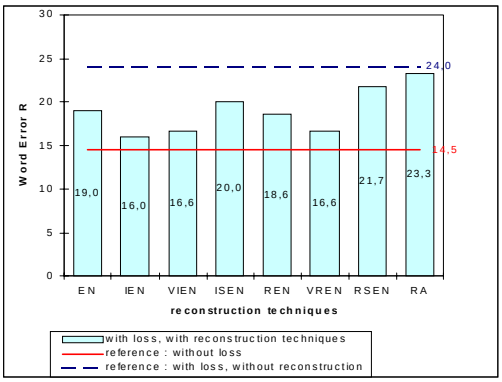

Figure 9: Average performance of the techniques (WER%)

5. CONCLUSION

This work was dedicated to the packet loss problem in distributed speech recognition. A packet loss simulation model was proposed and the performance of our continuous French speech recognition system was evaluated for different degradation conditions and for different packet sizes. Several reconstruction strategies were proposed and evaluated. The results confirm the intuitive fact that the word error rate obviously increases with the size of the lost packets and with the degradation level. It is also shown that simple reconstruction strategies allow to recover acceptable performance. The most efficient ones are those using interleaving technique to distribute the speech information among packets, combined with interpolation methods to estimate lost acoustic features. In further work, we intend to apply other techniques as addition *exclusive-XOR* or/and *Reed-Salomon codes* and to evaluate them on the AURORA[3] database.

6. REFERENCES

[1] W. Zhang, L. He, Y. Chow, R. Yang, and Y Su, "The Study on Distributed Speech Recognition System", *ICASSP 2000 International Conference en Acoustic Speech & Signal Processing*. June 5-9 2000, Istambul, Turkey.

[2] J-C. Bolot, S. Fosse-Parisis, Adaptive FEC-Based Error Control for Internet Telephony, Proc. IEEE Infocom'99, New York, NY, March 1999.

[3] M. Yajnik, S. Moon, J. Kurose and D. Towsley, Measurment and Modelling of the temporal Dependence in Packet Loss, IEEE Infocom'99, New York, March 1999.

[4] J-C. Bolot, A. Vega-Garcia, The Case for FEC-Based Error Control for Packet Audio in the Internet, ACM/Springer Multimedia Systems, 1999.

[5] C. Perkins, O. Hodson, V. Hardman, A Survey of Packet-Loss Recovery Techniques for Streaming Audio, IEEE Network Magazine, pp. 40-48, September/October 1998.

[6] B. Milner and S. Semnani, Robust Speech Recognition Over IP Networks, ICASSP2000, Istambul (Turkey).

[7] Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki-Waibel, N., Waibel, A., and Ward, W. "Recent Advances in JANUS : A Speech Translation System". *Eurospeech*, 1993, volume 2, pages 1295-1298.

[8] L. F. Lamel, J. L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Coprus for French", *Eurospeech*, Gênes, Italy, Vol 2, pp. 505-508, 24-26 september 1991.

[9] Vaufreydaz, D., Akbar, M., Rouillard, J., "Internet documents : a rich source for spoken language modeling", *ASRU'99 Workshop*, Keystone Colorado (USA), pp. 277-280.

[3] http://www.icp.inpg.fr/ELRA/aurora.html