EXPERIMENTS ON A CHANNEL ERROR PROTECTION SCHEME FOR DISTRIBUTED SPEECH RECOGNITION

Zheng-Hua Tan, Børge Lindberg and Paul Dalsgaard

Center for PersonKommunikation, Aalborg University,
Fredrik Bajers Vej 7, DK-9220 Aalborg Ø, Denmark
{zt, bli, pd}@cpk.auc.dk

ABSTRACT

Low recognition rates are often experienced in speech recognition over wireless networks where propagation environments may cause high transmission error rates. One way to reduce the sensitivity towards transmission errors is to use a distributed speech recognition architecture (DSR) which eliminates the speech channel and instead uses an error protected data channel to transmit a parameterised representation - suitable for speech recognition - of the speech.

Within the ETSI-DSR standard, two quantised melcepstral frames – each of 10 ms duration - are grouped together and protected with a four-bit Cyclic Redundancy Checking (CRC) forming a frame-pair. However, this scheme increases the Frame Error Rate (FER) over an error-prone transmission channel.

To overcome this, the paper presents a one-frame architecture in which a four-bit CRC is calculated to protect each frame independently. This scheme results in a lower overall probability of one frame in error at the cost of only a marginal increase in data rate from 4800 bits/s to 5000 bits/s.

A number of recognition experiments have been conducted on two different recognition tasks, digits and city names, to verify the introduction of the one-frame CRC protection scheme for a number of simulated transmission channel bit-error rates (BER) ranging from 0 (no transmission channel involved) to 2 %.

Experimental results verify that the one-frame error protection scheme is more robust against channel errors.

1. INTRODUCTION

Spoken language human-computer interaction tuned for flexible and user-friendly communication scenarios is forecast to play an important role in accessing and retrieving information at any time, from anywhere and on a variety of devices [1,2].

Over the past years one important research topic within this area has focused on the problem of distributed speech recognition (DSR) in mobile, wireless and IP networks.

Adopting the client-server architecture, see Figure 1, the modules of a DSR system are split between the terminal (client) and the server. The recogniser front-end is located in the terminal to which it is 'connected' via the transmission network to a remote back-end server in which the speech recogniser is executing. The transmission between the client and the server may be over either a wireless or a wire-line channel network or a combination of the two types.

Non-perfect networks definitely induce a number of constraints to currently used methodologies. Especially, the performance of speech recognition will degrade seriously when used in environments that are influenced by transmission packet loss and channel errors.

It is thus of paramount importance to conduct research on techniques for channel error protection against transmission errors. The speech recognition community has just recently begun to investigate the issue of packet loss and transmission errors [3]. In [4], we proposed a robust channel error protection scheme for DSR that has shown a good performance in a simple recognition task.

In this paper the work is extended with further experiments in order to cover additional recognition tasks.

2. CHANNEL ERROR PROTECTION SCHEME

The first DSR standard published by ETSI in February 2000 aimed at dealing with the degradations of speech recognition over mobile channels, caused by both low bit rate speech coding and channel transmission errors [5, 6].
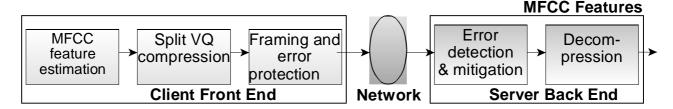

Figure 1: Block Diagram of ETSI Standard

Sync Sequence	Header Field	Frame 1	Frame 2	CRC 1-2	• • •	Frame 23	Frame 24	CRC 23-24
		44 bits	44 bits	4 bits	• • •	44 bits	44 bits	4 bits
		92-bit frame-pair			• • • •	92-bit frame-pair		
16 bits	32 bits	1104 bits						
1152 bits / 144 octets (i.e. 1152 / 240 · 1000 = 4800 bits)								

Table I: Aurora CRC protection scheme and multi-frame format

A DSR system handles these problems by eliminating the speech channel and instead using an error protected data channel to transmit a parameterised representation - suitable for speech recognition - of the speech.

One key point of the ETSI-DSR standard, Aurora, is that the transmission channel is claimed not to affect the recognition system performance and channel invariability is achieved.

The Aurora document [7] shows that no major degradation is observed for strong and medium GSM signal strength. However, for a poor channel, e.g. 4 dB carrier-to-interference (C/I), the recognition performance relatively degrades by from 10.0% to 16.2% for different tasks in comparison to the case of transmission without errors.

The ETSI-DSR standard defines a feature estimation front-end and an encoding scheme for speech input to be transmitted to the speech recognition system in the server. The encoding algorithm is a standard mel-cepstral technique commonly used in many speech recognition systems. The mel-cepstral calculation is a frame-based scheme that produces an output vector every 10 ms.

The frame-based feature estimation algorithm generates a 14-element vector consisting of 13 cepstral coefficients and log Energy. Each feature vector is further compressed to 44 bits via a split-vector quantization to reduce the data rate of the encoded stream. Each frame with the length of 44 bits represents 10 ms of speech. Two of the quantized 10 ms mel-cepstral frames are grouped together as a pair. A four-bit CRC is calculated on the frame-pair and is appended to it, resulting in a 92-bit long frame-pair packet. Twelve of these frame-pairs are combined to fill an 1104 bits feature stream packet. The feature stream is combined with the overhead of the synchronization sequence and the header, resulting in a multi-frame packet with a fixed length of 1152 bits representing 240 ms of speech. The multi-frame packets are concatenated into a bit-stream for transmission via a GSM channel with an overall data rate of 4.800 bits/s, see Table I.

Two types of data transmission can be supported, circuit-switched data and packet data. The Aurora working group defined the DSR standard for circuit switched channels. For packet data networks the DSR draft of the Internet Engineering Task Force (IETF) adopts the same frame-pair architecture and a different multi-frame format [4]. The bit-stream is transformed using the Real Time Protocol (RTP). Both data transmission channels are error prone. Therefore, it is essential to have robust error protection.

Over an error-prone transmission channel – often occurring in mobile communication - this format will cause severe problems.

To overcome this, a one-frame architecture in which a four-bit CRC is calculated to protect each frame independently is presented instead [4]. This scheme results in that the overall probability of one frame in error is significantly lower, see Figure 2, at the cost only of a slight increase in the overall bit rate.
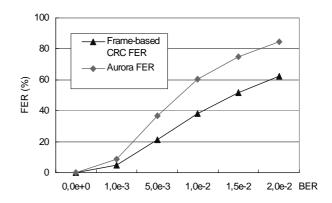

Figure 2: The theoretic FER of one- and two frame CRC

4. FRAME-BASED CRC-SCHEME

Both the DSR for circuit-switched data and for packet–switched data adopt the frame-pair format in which one four-bit CRC is used to detect transmission errors in each frame-pair.

When errors are detected, a substitution is needed for the frames received with errors. The last error-free frame before the erroneous frame-pair/s and the first correct frame following the erroneous frame-pair are used to substitute those received with errors. If there are N consecutive erroneous frame-pairs (corresponding to 2N frames), then the first N frames are replaced by a copy of the last correct frame before the error and the last N frames are replaced by a copy of the first error-free frame received following the error.

A data consistency test is applied to determine whether the frames in an Aurora frame-pair have a minimal continuity to search for erroneous frames missed by the CRC detection.

Frame #	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	
Errors X	X		X		X					X		X	X		X			X			X		X	X
Aurora Error Correction	X		X		X			X		X		X		X		X					X		X	
			6							7						18						19		
Frame-based CRC	X		X		X					X		X	X		X			X			X		X	X
		1		3		5		6	7		8		10		13		15		16	18	19		21	

Table 2: A multi-frame packet example with BER 1%

Applying the four-bit frame-based CRC, in principle, will allow detection of more errors.

In the frame-based CRC scheme, four bits are appended to each 44-bit frame vector resulting in a one-frame packet of 48 bits. Twenty-four of these one-frame packets are concatenated into an 1152-bit multi-frame packet stream. After the feature stream is combined with the overhead of the synchronization sequence and the header, a 1200-bit multi-frame is formed which results in an overall data rate of 5.000 bits/s.

Table 2 illustrates a multi-frame packet example with a BER of 1%. It is seen that in this case, the frame-based method maintains the actual FER.

4. EXPERIMENTAL SETUP

A number of recognition experiments have been conducted to verify the introduction of the one-frame CRC protection scheme. For each experiment, simulated transmission channel bit-error rates (BER) range from 0 (no transmission channel involved) to 2%. To simulate channel transmission errors various amounts of bit errors, ranging from 0% to 2%, are randomly added to the bitstream. A closer analysis shows that 2% BER is equivalent to the relatively high value of 60% FER.

The recogniser applied in the experiments is the SpeechDat reference recogniser established within the COST 249 Action, which is using a fully automatic, language-independent training procedure for building a phonetic recogniser [8]. It relies on the HTK toolkit and a SpeechDat (II) compliant database.

The database used in the experiments is the DA-FDB 4000, which contains speech from 4000 speakers collected over the fixed network (FDB) for the Danish language. The speech files are stored as sequences of 8-bit 8 kHz and A-law sampled.

In [4] results were reported for the Danish digits. The vocabulary consists of isolated words: nul, en, et, to, tre, fire, fem, seks, syv, otte, ni. The digit '1' pronounced either as 'en' or 'et' occurs twice as often as the remaining digits.

In this paper further experiments on a more difficult task are described, using "city names" recognition. The "city names" list was constructed from the 284 largest towns/cities in Denmark added with 103 names of international airport cities. Approximately 30 additional names were found from names of domestic airports and airports in Greenland whereas the remaining names were found from the spontaneous city names retrieved out of the first 1000 calls. This results in a set of 500 city names.

5. RESULTS AND DISCUSSIONS

Six different channel conditions (labelled O, A, B, C, D, E) are defined in terms of their bit error rates as listed in Table 3.

Recognition results are presented in word error rate (WER) against bit error rates for the two different error protection schemes (Aurora stands for frame-pair protection scheme within the ETSI-DSR standard Aurora and the corresponding experiments are used as the baseline; Frame-based CRC stands for the proposed scheme), see Table 3 and Figure 3.

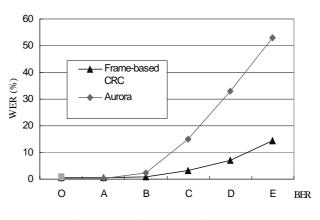| Channel conditions	O	A	B	C	D	E
BER	0	0.001	0.005	0.010	0.015	0.020
Aurora (WER)	0.2	0.2	2.5	15.1	32.8	52.9
Frame-based CRC (WER)	0.6	0.6	1.0	3.1	7.0	14.4

Table 3: Digit WER (in %) against BER

Figure 3: Digit WER against BER

The results show that an improved performance is obtained by the frame-based CRC. For the 2 % BER channel condition, the frame-based CRC protection scheme still achieves a WER of about 14.4% and it indicates a strong robustness against transmission errors. At the same conditions, however, the WER of Aurora rises to 52.9%.

City name recognition is a more difficult recognition task because of perplexity and phonetic similarities among the vocabulary items. Thus even without a transmission channel involved, the achieved performance is a WER of 20.5%. The results of this task against various BERs are shown in Figure 4 and Table 4. The experimental results again strongly support that the one-frame protection scheme is more robust to channel errors although a slight increase in the error-protection overhead is needed due to the more CRC bits needed.

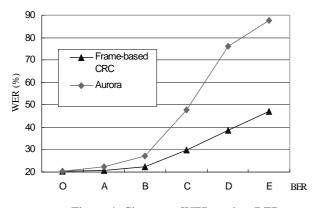| Channel conditions	O	A	B	C	D	E
BER	0	0.001	0.005	0.010	0.015	0.020
Aurora (WER)	20.5	22.3	27.0	47.7	76.1	87.5
Frame-based CRC (WER)	20.3	20.8	22.3	29.7	38.6	47.2

Table 4: City name WER (in %) against BER

Figure 4: City name WER against BER

6. CONCLUSIONS AND FUTURE WORK

This paper presents a number of experiments on a channel error protection scheme for DSR. The method uses a frame-based CRC for error protection. With a slight increase in the overall bit rate, the robustness against errors increases significantly for all the applications.

The recognition results observed for both sets of experiments consistently indicate an improved robustness of the frame-based CRC protection scheme, compared to the standard Aurora scheme.

CRC is an often-used approach for detecting the potential transmission errors. In further work, the embedded CRC in DSR may be applied to estimate the immediate performance of the involved networks. Such knowledge, for example, frame-error-rate (FER) may be utilized to determine the threshold of out-of-vocabulary (OOV) detection – an issue that is highly relevant for spoken input for mobile devices. Furthermore, adaptive grammars in variable network environments are an obvious subject to investigate with the aim of analysing its effect on perceived quality-of-service (QoS).

REFERENCES

[1] L. –S Lee and Y. Lee, "Voice access of global information for broad-band wireless: technologies of today and challenges of tomorrow," Proceedings of the IEEE, Vol. 89, No. 1, Jan. 2001, pp. 41 –57.

[2] R. V. Cox, C. A. Kamm, L. R. Rabiner et al., "Speech and Language Processing for Next-Millennium Communications Services," Proceedings of the IEEE, Vol. 88, No. 8, Aug. 2000, pp.1314-1337.

[3] E. A. Riskin, C. Boulis, S. Otterson and M. Ostendorf, "Graceful Degradation of Speech Recognition Performance Over Lossy Packet Networks," in Proc. Eurospeech-2001, September 2001.

[4] Z. –H. Tan and P. Dalsgaard, "Channel Error Protection Scheme For Distributed Speech Recognition," Proc. ICSLP-2002, September 2002.

[5] "ETSI ES 201 108 v1.1.2 Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithm", February 2000.

[6] D. Pearce, "Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI standards activities for Distributed Speech Recognition Front-ends". AVIOS 2000: The Speech Applications Conference, San Jose (USA), May 2000.

[7] Aurora document no. AU/266/00 "Recognition with WI007 Compression and Transmission over GSM Channel", Ericsson, December 2000.

[8] B. Lindberg, F.T. Johansen, N. Warakagoda, et al, "A Noise Robust Multilingual Reference Recogniser Based on SpeechDat (II)," in Proc. ICSLP-2000, October 2000.