# CHANNEL ERROR PROTECTION SCHEME FOR DISTRIBUTED SPEECH RECOGNITION

*Zheng-Hua Tan and Paul Dalsgaard*
Center for PersonKommunikation, Aalborg University, Denmark
{zt, pd}@cpk.auc.dk

## ABSTRACT

This paper describes ongoing research preparing for the widespread deployment of spoken language processing in networks encompassing wired and wireless transmission channels.

The paper gives a brief overview of the standardized bit-error protection scheme aimed at minimising channel transmission errors and used within the distributed speech recognition (DSR) paradigm. Within the ETSI-DSR standard, two quantised mel-spectral frames – each of 10 ms duration - are grouped together and protected with a 4-bit Cyclic Redundancy Checking (CRC) forming a frame-pair. However, this causes the entire frame-pair erroneous if a one-bit error only occurs in the frame-pair packet. Over an error-prone transmission channel this format will cause severe problems.

To overcome this, the paper presents a one-frame architecture in which a 4-bit CRC is calculated to protect each frame independently. This scheme results in that the overall probability of one frame in error is lower, or that an error occurring in one frame does not affect another frame. A number of simple recognition experiments have been conducted 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 $2 \cdot 10^{-2}$. Experimental results show that the one-frame protection scheme is more robust to channel errors although a slight increase in the error-protection overhead is needed.

## 1. INTRODUCTION

With the rapid growth of the IP and mobile technologies, future solutions to communications emphasize a variety of access and IP-based core networks, which will enable end-users to communicate anywhere, anytime and on a variety of devices.

As the size of wireless devices shrinks, the use of traditional keyboard or keypad input and screen output becomes increasingly inconvenient. Therefore, efficient, flexible, and user-friendly approaches for human-device communication tuned for new communications scenarios are essential. Motivated by technology advances in the field of spoken language processing voice operated interfaces for a variety of services are becoming more and more prevalent as one way of circumventing the miniaturization problems.

However, a transmission channel adds a number of 'constraints' to currently used 'network free' methodologies, e.g. i) low bit-rate speech coding due to the limited bandwidth, ii) the effect of unreliable transmission channels due to multi-path fading propagation, iii) loss of frames due to transmission errors or network congestion, vi) long delay etc.

Seen from the network provider's as well as the end-user's point of view it is of paramount importance to initiate and conduct research leading to techniques that provide and maintain optimal Quality of Service (QoS) under dynamically varying transmission channel properties.

During the last five years one important research topic within speech and spoken language processing has focused on the problem of DSR in the context of speech-mediated communication in mobile, wireless and IP networks.

The results obtained so far are now being applied towards a next step development in which the concept of QoS additionally must embrace the functionality of spoken language systems into QoS-measures. QoS has hitherto been solely focused on network concepts to ensure that the network provider is able to 'deliver' access to transmission services having sufficient quality. But, in an end-user centred sense, QoS must be aimed at also offering access to applications that are judged of sufficient quality by the end-user.

Research on QoS for interactive spoken language systems – including now the end-user - is a paradigm shift. In the context of such 'Perceived' QoS it is expected that an application including its access via a speech driven interface will be designed to exploit the specific knowledge of the transmission network's QoS. Such application will be set-up to dynamically decide which specific server modules shall be used for the speech and spoken language processing and which structures to be used with a modified dialogue control. For instance, a low QoS transmission network would indicate the deployment of a more robust speech recognition algorithm and a simpler grammar structure used in the spoken language-understanding module. Integration of these combined knowledge sources aims at an overall optimising of the 'Perceived' QoS.

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 [1, 2]. A DSR system handles these problems by eliminating the speech channel and instead using an error protected data channel to send a parameterised representation - suitable for speech recognition - of the speech. One key point of Aurora is that the transmission channel is claimed not to affect the recognition system performance and channel invariability is achieved. The Aurora document [3] 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.

This paper investigates the channel error protection scheme aiming at providing a more robust scheme against transmission errors.

## 2. DSR AND THE ETSI STANDARD

Adopting the client-server architecture, 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.

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 4-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.

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.

## 3. FRAME-BASED CRC SCHEME

Both the DSR for circuit-switched data and for packet–switched data adopt the frame-pair format in which one 4-bit CRC is used to detect transmission errors in each frame-pair.

The disadvantage of the Aurora frame-pair format is that both feature frames will be in error if only one-bit error occurs in the frame-pair packet of 92 bits. Table 1 illustrates the elements of a multi-frame packet example from a bit-stream simulating 1% bit error rate (BER) and taken from the experiments. There are 24 frames in each multi-frame packet, shown in the first row of the table. The second row gives the errors occurred; "X" marked frames are erroneous frames. This results in an actual frame error rate (FER) of about 45.6% (11 out of 24). Using the Aurora frame-pair format, for example, frame 0 and frame 1 are in a frame-pair and protected by one 4-bit CRC. Therefore, the frame-pair consisting of frame 0 and 1 will be detected as an erroneous frame-pair due to the error in frame 0. Since neither frame 6 nor 7 have errors detected, the frame-pair consisting of frame 6 and 7 will be detected as an error-free frame-pair. Across the entire frame-pair packet, the result is that only two correct frame-pairs (6/7 and 18/19) are detected. This is shown in the third row, from which it is also observed that the detected FER increases to 83.3% (20 out of 24) due to the use of the Aurora frame-pair format.

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. Therefore, the frames numbered 6, 7, 18 and 19 are used to substitute 20 erroneous frames shown in the fourth row. The substitutions at the ends depend on the previous or following frame-pair packet.

From Table 1, it is clear that Aurora frame-pair protection scheme increases the FER. A different frame-based CRC protection scheme is presented in the following and tested in a number of experiments. In the frame-based scheme a 4-bit CRC is calculated for each frame independently and is appended. Therefore, error in one frame is not affecting a neighbouring frame.

| Frame number | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Errors (X) | X | | X | | X | | | | | X | | X | X | | X | | | X | | | X | | X | X |
| Aurora error correction | X | | X | | X | | 6 | 7 | X | | X | | X | | X | | X | | 18 | 19 | X | | X | |
| | **6** | | | | | | | **7** | | | | | | **18** | | | | | | **19** | | | | |
| Frame-based-CRC | X | 1 | X | 3 | X | 5 | 6 | 7 | 8 | X | 10 | X | X | 13 | X | 15 | 16 | X | 18 | 19 | X | 21 | X | X |
| | **1** | | **3** | | **5** | | **6** | **7** | **8** | | **10** | | **13** | | **15** | | **16** | | **18** | **19** | **21** | | | |

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

In the fifth row, each erroneous frame is detected by its own CRC and the error-free frames are still detected as error-free. This maintains the FER of 45.6%. Using the same substitution scheme as for the Aurora frame-pairs, the frame-based method obtains the results shown in the sixth row.

Row 6 in Table 1 shows that the frame-based CRC error correction scheme maintains 13 effective frames to be used to interpolate 24 frames under the same error condition.

The results in Table 1 also show that for the Aurora frame-pair error correction scheme, a series of frames may, e.g. be repeatedly substituted using the same frame feature vector across 7 frames, whereas the worst case for the frame-based CRC is 4 only.

Moreover, none of the CRC schemes are able to detect all errors. A data consistency test is thus 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. Applying the 4-bit frame-based CRC, in principle, will allow detection of more errors.

In the frame-based CRC scheme, 4 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.

## 4. EXPERIMENTAL RESULTS

In this section, we describe the experiments performed to evaluate the frame-based CRC error protection scheme. The recognition task is on 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.

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 [5]. It relies on the HTK toolkit and the SpeechDat (II) compatible database DA-FDB 4000. This database covers 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, as Aurora2 uses, A-law sampled.

The DA-FDB 4000 database is used for training 32 Gaussian mixture triphone models. Test data - isolated digits - are also from the database.

Within the Aurora framework erroneous frames are substituted either by the previous error-free frame or by the following. An alternative estimation scheme for erroneous frames is tried in these experiments. A polynomial interpolation is used to estimate the erroneous frames, e.g. given in reference [6]. Interpolation exploits the temporal correlation present in the speech feature stream, which is due to both the overlapping estimation procedure of the front-end processing and the speech production process constrained by the vocal tract. In the experiment - called Aurora-Int - a first degree Lagrange polynomial interpolation is used.

Given $c_i^n$ stands for the $i$th cepstral coefficient of the $n$th erroneous frame. There are $2N$ frames ($N$ frame-pairs) in error to be replaced. The previous error-free frame is frame $A$ and the following frame $B$. Then $c_i^n$ can be estimated using $c_i^B$ and $c_i^A$.

$$c_i^n = c_i^A + \frac{n}{2N+1}.(c_i^B - c_i^A),\ n = 1, 2, \cdots, 2N$$

The acoustic features for the recognition consist of the conventional set of a 39-dimensional MFCC vector, including the zero’th cepstral coefficient C0 and the first and second order deltas.

To simulate channel transmission errors various amounts of bit errors, ranging from 0% to 2%, are randomly added to the bit-stream, which is the concatenation of multi-frames. A closer analysis shows that 2% BER is equivalent to the relatively high value of 60% FER.

Six different channels (labelled O, A, B, C, D, E) are defined in terms of their bit error rates which is shown in Figure 1 and Table 2. The Aurora experiment is used as the baseline.

The results show that an improved performance is obtained by the frame-based CRC. For the $2 \cdot 10^{-2}$ BER channel condition, frame-based CRC protection scheme still achieves the recognition accuracy of about 85,6% and it shows a strong robustness against transmission errors. Aurora-Int obtains the worst performance. This indicates that the error correction by Lagrange polynomial interpolation does not outperform the simple repeating substitution. Reference [7] shows that C0 exhibits strong temporal correlation even at lags of 20 frames. But the temporal correlation of higher-order MFCCs falls off rapidly with increasing frame lag. It can be deduced that by using interpolation in the mel-cepstral domain it is hard to achieve high performance. However, interpolation on log filter bank features can be more effective since the correlation is much stronger in the filter bank domain.

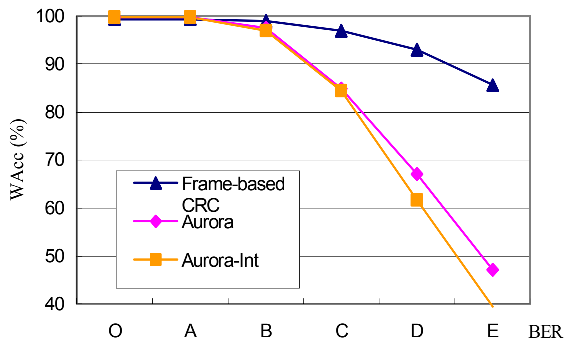

Figure 1: Digit recognition accuracy against bit errors

| Channel conditions (BER) | O | A | B | C | D | E |
|---|---|---|---|---|---|---|
| | 0 | $10^{-3}$ | $5.10^{-3}$ | $10^{-2}$ | $1,5.10^{-2}$ | $2.10^{-2}$ |
| Aurora | 99,8 | 99,8 | 97,5 | 84,9 | 67,2 | 47,1 |
| Aurora-Int | 99,8 | 99,8 | 96,9 | 84,4 | 61,6 | 39,5 |
| Frame-based CRC | 99,4 | 99,4 | 99,0 | 96,9 | 93,0 | 85,6 |

Table 2: Digit recognition accuracy against bit errors

## 5. SUMMARY AND DISCUSSION

This paper has demonstrated a robust error protection scheme for distributed speech recognition. The method uses a frame-based CRC for error detection. With a slight increase in the overall bit rate, the robustness against errors increases significantly. Since the DSR for packet-switched channel adopts the same front-end codec and frame-pair format, the proposed scheme also fits into DSR over IP.

In further work, decreasing the overall bit rate down to 4.800 bits/s will be investigated for maintaining the same robustness as in frame-based CRC. Empirical tests [8] have shown that no significant performance degradation occurred in the conducted experiment by replacing the last (here the 12th) cepstral coefficient with its fixed pre-calculated mean value. This means that it is not necessary to allocate bits to the last coefficient. Reference [9] presents a source and channel coding system that operates at 500 bits/s and provides good digit recognition performance over a wide range of channel conditions. On the other hand, the DSR bit-stream is suggested to be transmitted through a 9.600 bits/s GSM data channel. Thus, 5.000 bits/s is still acceptable.

We also plan to apply it in more complex application tasks and use a newly established UMTS emulator that is being deployed in the CPK cross-group initiative, the FACE project.

## 6. ACKNOWLEDGEMENT

This work is partly funded by the support given to CPK from the Danish Technical Research Council. The work is done within the CPK research project FACE “Future Adaptive Communication Environment” that is an activity encompassing experts from the three research groups SMC (Speech and Multi-media Communication), CSys (Cellular Systems) and WING (Wireless Networks).

## 7. REFERENCES

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

[2] 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.

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

[4] “IETF AVT WG Internet-Draft RTP Payload Format for Distributed Speech Recognition”, November 2001.

[5] 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.

[6] B. Milner and S. Semnani, “Robust speech recognition over IP networks”, in Proc. ICASSP-00

[7] B. Milner, “Robust Speech Recognition in Burst-Like Packet Loss,” in Proc. ICASSP-01, USA, May 2001.

[8] V. Weerackody, W. Reichl and A. Potamianos, “Speech Recognition for Wireless Applications”. IEEE International Conference on Communications, 2001.

[9] A. Bernard and A. Alwan, “Source and Channel Coding for Remote Speech Recognition over Error-prone Channel,” in Proc. ICASSP-01, USA, May 2001.