Partial splicing packet loss concealment for distributed speech recognition

Zheng-Hua Tan, P. Dalsgaard and B. Lindberg

A technique for mitigating the effect of packet loss in the context of distributed speech recognition is presented. The proposed packet loss concealment (PLC) technique substitutes packet loss partly by a repetition of neighbouring packets and partly by a splicing in which a number of packets are dropped. Experimental results demonstrate that the proposed PLC technique outperforms existing techniques.

Introduction: Recently an important research topic within speech processing has been to focus on the issue of distributed speech recognition (DSR). In a client–server architecture, a DSR system breaks speech recognition processing down in front-end feature estimation conducted in the client and back-end recognition in the server, where data transmission between the two parts may be via heterogeneous networks. Transmission across wireless networks and IP networks may cause varying types of transmission errors that severely degrade the performance of speech recognition. To counteract the drop in recognition accuracy caused by transmission errors, a number of packet loss concealment (PLC) techniques for DSR have been introduced using one of the following general PLC techniques: splicing, repetition or interpolation [1, 2].

In [1, 3] the repetition technique is used and it is found that for speech streams corrupted by the GSM error pattern EP3 (4 dB carrier-to-interference ratio), the word error rate (WER) for Danish digit recognition increases from 0.2% to 9.7% compared to error-free transmission. This indicates that further research on even more efficient PLC techniques remains a challenging topic.

In this Letter we propose an alternative PLC technique to the above three techniques – called *partial splicing*. This technique combines repetition and splicing, as a packet loss is substituted partly by a repetition of the neighbouring packets and partly by a splicing. Since the PLC technique operates in the feature-processing domain in the server its employment requires neither extra bandwidth of the network nor extra computations in the client and there is no requirement for modification in the speech recogniser decoding algorithm. It is however shown that the concealment technique under certain assumptions is equivalent to a modified Viterbi decoding algorithm.

Partial splicing: The motivation for introducing the partial splicing technique comes from the success of the commonly used weighted/modified Viterbi algorithm for coping with impulsive noise, where the influence on recognition accuracy from a segment of noise corrupted speech is taken into account in the Viterbi decoding by modifying its contribution to the overall likelihood score. The modified Viterbi algorithm in [4] uses the following formula (notation as typically used in Viterbi algorithm and identical to [4]) to update the likelihood score

$$\delta_t(j) = \max_i[\delta_{t-1}(i) \times a_{ij}] \times [b_j(x_t)]^{R_N(t)} \quad (1)$$

$R_N(t)$ is a normalised reliability coefficient – of value between 0 and 1 – for each speech frame that adjusts the contribution of each frame to the overall likelihood score. If $R_N(t)$ is close to 1, the output probability for the particular frame contributes almost fully to the likelihood score. In contrast, if $R_N(t)$ is close to 0, the output probability approaches an identical (equal to 1) contribution for all models for the unreliable frame, and therefore neutralises the frame.

In DSR, a packet (speech frame) is either error-free or erroneous (and subsequently dropped by the server if applying splicing) implying $R_N(t) = 1$ or $R_N(t) = 0$, respectively. Applying substitution of lost packets has been shown to better maintain the recognition accuracy than splicing [2]. However, all substitution packets are different from the original error-free packets and this may be considered equivalent to representing speech frames corrupted by noise. Defining a constant value of $R_N(t) = \alpha$ for all substitution packets results in an output probability $[b_j(x_t)]^\alpha$ at any time t. Assuming that two consecutive lost packets occur at time t and $t+1$, and also assuming that the decoding stays at the same state j and that a repetition substitution is used (thus $x_t = x_{t+1}$), then the contribution to the output probability of the two packets is

$$[b_j(x_t)]^\alpha \cdot [b_j(x_{t+1})]^\alpha = [b_j(x_t) \cdot b_j(x_{t+1})]^\alpha = [b_j(x_t)]^{2\alpha}$$

Setting $\alpha = 1/2$, the contribution of the two frames is $b_j(x_t)$, equal to the contribution of one frame. To implement an action that has a similar effect on speech decoding, partial splicing is proposed to adjust the contribution of the lost packets.

Given that C^n represents the cepstral coefficient vector of the nth erroneous frame and that there are N erroneous frames to be replaced, in the following notation A is used to denote the last correct frame before the first erroneous frame, and B the first correct frame following the last erroneous frame.

In partial splicing, N_s of the N packets will be spliced where N_s is

$$N_s = floor\left(\frac{N-1}{2}\right) \quad (2)$$

The remaining $N_r = N - N_s$ packets are substituted by 'fill-in' packets that are chosen to represent repetitions of neighbouring error-free packets. Thereby, the first half of N_r packets are represented by copies of frame A, the second half represented by copies of frame B, i.e.

$$\begin{cases} C^n = C^A, & n = 1, 2, \ldots, floor\left(\frac{N_r+1}{2}\right) \\ C^n = C^B, & n = floor\left(\frac{N_r+3}{2}\right), \ldots, N_r \end{cases} \quad (3)$$

To evaluate the partial splicing technique, existing PLC techniques are briefly described in the following. In the interpolation technique a first-order Lagrange polynomial interpolation is applied to establish an estimate of the erroneous frames [5], as given in linear equation (4).

$$C^n = C^A + \frac{n}{N+1} \cdot (C^B - C^A), \quad n = 1, 2, \ldots, N \quad (4)$$

The PLC technique applied in the ETSI-DSR standard [1] is using repetition where the consecutive erroneous frames are separated into two parts: the first half represented by copies of frame A, the second half represented by copies of frame B.

Experiments and discussions: Two different recognition tasks have been investigated to analyse the influence of applying a number of PLC techniques in speech recognition. The first task is recognition of the Danish digits (low perplexity); the second is city names (medium perplexity). The recogniser applied in the experiments is the Speech-Dat/COST 249 reference recogniser. A part of the DA-FDB 4000 database is used for training 32 Gaussian mixture triphone models.

The experimental setting is as defined in [1]. Three different error distributions have been used namely 1) additive white Gaussian noise (AWGN) channels simulated by random bit error rates (BER), 2) burst-like packet loss and 3) the more realistic GSM error patterns. The baseline word error rates (WER) (no packets lost) on the two tasks are 0.2 and 20.7%, respectively.

AWGN channel: Table 1 shows the results for the Danish digits and the city names tasks for AWGN channels. It is observed that partial splicing achieves the lowest WER for all values of BER across the two tasks – except for 0.1% BER for the city names.

Table 1: %WER across PLC techniques for varying values of BER for Danish digits and city names

Tasks	Danish digits					City names				
BER (%)	0.1	0.5	1	1.5	2	0.1	0.5	1	1.5	2
Partial-splicing	**0.2**	**2.5**	**10.4**	**20.5**	**47.1**	21.8	**24.9**	**44.1**	**69.0**	**83.3**
Repetition	0.2	2.5	15.1	33.4	53.0	22.5	26.9	47.7	76.2	87.5
Interpolation	0.2	3.1	15.8	39.0	61.0	**21.4**	30.3	54.1	84.0	92.2
Splicing	0.4	6.6	31.1	56.6	74.7	22.3	41.0	80.8	96.4	98.4

Burst-like packet loss: The burst-like packet loss was simulated by a three-state Markov model according to [5]. The packet losses (PL) used in this experiment range from 10 to 50% with an average length of eight. Burst-like losses are more difficult to counteract since they may cause the information of whole phonemes to be lost, and consequently no way of regaining this information. However, the results in Table 2 show that PLC by partial splicing performs slightly better in terms of WER than for the other techniques.

Table 2: %WER across PLC techniques for varying values of burst-like PL for Danish digits and city names

Tasks	Danish digits					City names				
PL (%)	10	20	30	40	50	10	20	30	40	50
Partial-splicing	**7.0**	**18.5**	**31.1**	**39.8**	**54.6**	**31.2**	**42.5**	**57.5**	**75.1**	**82.4**
Repetition	8.5	19.3	32.8	42.1	58.3	34.3	45.4	62.8	77.7	84.4
Interpolation	8.9	21.2	35.5	43.8	59.1	36.3	46.8	63.9	79.3	88.0
Splicing	7.1	19.9	34.0	43.8	59.1	35.0	49.7	66.4	80.0	90.2

GSM error patterns: GSM error patterns are commonly used for testing speech codecs and DSR error protection schemes as they are more realistic error distributions including both random errors and burst-like errors. The three error patterns are EP1, EP2 and EP3, corresponding to C/I ratios of 10, 7 and 4 dB, respectively. The error patterns are used as specified in [1]. Table 3 provides the experimental results. Again the partial splicing PLC technique gives the best results – except for EP1 where splicing is slightly better for the city names task.

Table 3: %WER across PLC techniques for GSM error patterns for Danish digits and city names

Tasks	Danish digits			City names		
Patterns	EP1	EP2	EP3	EP1	EP2	EP3
Partial-splicing	**0.2**	**0.2**	**7.3**	20.9	**20.9**	**35.0**
Repetition	0.2	0.2	9.7	20.9	21.1	38.3
Interpolation	0.2	0.2	10.6	20.9	21.1	41.2
Splicing	0.2	0.4	13.1	**20.7**	22.5	48.8

Conclusion: A computationally simple and effective PLC technique has been presented and its power verified by experiments. Compared to existing PLC techniques the proposed technique significantly decreases WERs. Further to this improvement a speed-up of the recognition decoding is obtained due to the drop of packets.

1 August 2003
Electronics Letters Online No: 20031026
DOI: 10.1049/el:20031026

Zheng-Hua Tan, P. Dalsgaard and B. Lindberg (*Center for Person-Kommunikation, Department of Communication Technology, Aalborg University, 9220 Aalborg, Denmark*)

E-mail: zt@kom.auc.dk

References

1 PEARCE, D.: '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, CA, USA, May 2000
2 BOULIS, C., *et al.*: 'Graceful degradation of speech recognition performance over packet-erasure networks', *IEEE Trans. Speech Audio Process.*, 2002, **10**, (8)
3 TAN, Z.-H., DALSGAARD, P., and LINDBERG, B.: 'OOV-detection and channel error protection for distributed speech recognition over wireless network'. ICASSP-2003, Hong Kong, China, April 2003
4 CHO, H.Y., KIM, L.Y., and OH, Y.H.: 'Segmental reliability weighting for robust recognition of partly corrupted speech', *Electron. Lett.*, 2002, **38**, (12)
5 MILNER, B.: 'Robust speech recognition in burst-like packet loss'. ICASSP-01SP-01, Salt Lake City, UT, USA, May 2001