A SUBVECTOR-BASED ERROR CONCEALMENT ALGORITHM FOR SPEECH RECOGNITION OVER MOBILE NETWORKS

Zheng-Hua Tan, Paul Dalsgaard and Børge Lindberg
{zt, pd, bli}@kom.auc.dk
SMC–Speech and Multimedia Communication, Department of Communication Technology[1], Aalborg University, Denmark

ABSTRACT

Conventional error concealment (EC) algorithms for distributed speech recognition (DSR) share a common characteristic namely the fact of conducting EC at the vector (or frame) level. This strategy, however, fails to effectively exploit the error-free fraction left within erroneous vectors where a substantial number of subvectors often are error-free. This paper proposes a novel EC approach for DSR encoded by split vector quantization (SVQ) where the detected erroneous vectors are submitted to a further analysis at the subvector level. Specifically, a data consistency test is applied to each erroneous vector to identify inconsistent subvectors. Only inconsistent subvectors are replaced by their nearest neighbouring consistent subvectors whereas consistent subvectors are kept untouched. Experimental results demonstrate that the proposed algorithm in terms of recognition accuracy is superior to conventional EC methods having almost the same complexity and resource requirement.

1. INTRODUCTION

Transmitting data across mobile networks adds a number of challenges to state-of-the-art speech technologies, for example bandwidth limitation and transmission errors. Inspired by the rapid growth of mobile communications, a standard for DSR has been published by ETSI with the aim of dealing with the degradations of speech recognition over mobile channels, caused by both low bit rate speech coding and transmission errors [1,2].

However, it is found that a poor channel, for example a 4 dB carrier-to-interference (C/I) ratio, still severely reduces the accuracy of speech recognition over mobile networks with the implementation of ETSI-DSR [2,3].

To mitigate transmission errors one of the following EC algorithms is generally employed in DSR: splicing, substitution, repetition or interpolation [1-8]. The ETSI-DSR standard employs a repetition scheme where the concealment is split into two parts by replacing the first half of a series of erroneous frames with a copy of the last correct frame before the error and replacing the second half with a copy of the first correct frame following the error [1-3]. The commonly used interpolation technique applies a polynomial interpolation as an estimate of the erroneous frames [4]. An interpolation scheme applying a trigonometric weight to the error-free frames received before and after the erroneous frames has been reported in [5]. In a splicing, erroneous frames are simply dropped [6]. In [7] a number of substitution schemes have been described. Partial splicing presented in [8] is a feature-domain concealment technique which substitutes lost/erroneous frames partly by a repetition of neighbouring frames and partly by a splicing. Under certain assumptions the partial splicing is equivalent to a modified Viterbi decoding algorithm. It is pointed out that in all the above feature-domain methods erroneous vectors are simply disregarded and substituted.

In a different way, both [9] and [10] integrate the reliability of the channel-decoded feature into the recognition process where the Viterbi decoding algorithms are modified such that contributions made by observation probabilities associated with vectors estimated from erroneous or lost vectors are decreased. To implement these methods, the reliability of the channel-decoded feature is required and the recognisers are changed accordingly which thereby can be classified as model-domain EC schemes. A more recent model-domain technique applies missing feature theory to error-robust speech recognition where lost and erroneous vectors generate constant contributions to the Viterbi decoding and therefore these vectors are neutralised [11].

All the algorithms referenced above share a common characteristic namely the fact of conducting error concealment at the vector level. A vector is considered the unit to be detected followed by a substitution or a reduced likelihood contribution if erroneous. A vector and a speech frame are equivalent in this paper.

However it is highly likely that not all elements in an erroneous vector are corrupted by error.

To utilize the (remaining) error-free information embedded in erroneous vectors, this paper proposes a subvector-level EC algorithm where each subvector in an SVQ is considered as an alternative unit for error detection and mitigation. The proposed algorithm is a server-side feature-domain EC technique where there is neither requirement for modification in the recogniser nor requirement for extra bandwidth.

2. THE ETSI-DSR STANDARD

The ETSI-DSR standard defines the feature-estimation front-end processing together with an encoding scheme for speech input to be transmitted over the mobile network to the server-based speech recognition system [1]. The front-end adopts a standard mel-cepstral technique, which produces a 14-element vector consisting of log energy (logE) in addition to 13 cepstral coefficients ranging from c_0 to c_{12} – computed every 10 ms.

To reduce the bit rate of the encoded stream, each feature vector is compressed using an SVQ. The SVQ technique groups two features (either c_i and c_{i+1}, i=1,3...11 or c_0 and logE) into a

[1] Formerly CPK – Center for PersonKommunikation, which is now fully integrated into the Department of Communication Technology

0-7803-8484-9/04/$20.00 ©2004 IEEE

feature-pair subvector. Each subvector is quantized using its own SVQ codebook, in total resulting in seven codebooks and seven subvectors in one vector. The size of each codebook is 64 (6 bits) for c_i and c_{i+1} whereas 256 (8 bits) for c_0 and logE, resulting in a total of 44 bits for each vector.

Before being transmitted two quantized frames (vectors) are grouped together as a frame-pair. A 4-bit CRC is calculated for each frame-pair and appended, resulting in 92 bits for each frame-pair. Twelve frame-pairs are combined to form a 1104-bit feature stream. Adding the overhead bits of the synchronization sequence and the header in total results in a 1152-bit multi-frame representing 240 ms of speech. The multi-frames are concatenated into a bitstream for transmission with an overall bit rate of 4 800 bits/s.

Over error-prone channels the received bitstream may have been contaminated by errors. To determine if a frame-pair is received with errors two methods are applied, namely CRC and data consistency test. The data consistency test determines whether or not the decoded features for each of the two speech vectors in a frame-pair have a minimal continuity.

A frame-pair is labelled as erroneous when its CRC is detected as incorrect. It is moreover classified as erroneous if the previous frame-pair does not have the minimal continuity. The following frame-pairs are identified as erroneous until one frame-pair has a correct CRC and meet the consistency requirement.

In the error concealment process of the ETSI-DSR a repetition EC is applied to replace those erroneous vectors.

3. ANALYSIS OF THE EFFECT OF ERRORS

The problem of employing the above strategy is that two entire vectors in a frame-pair will be in error and substituted even though only a single bit error occurs in the 92-bit frame-pair.

This is a common characteristic of vector-level EC algorithms no matter whether splicing, substitution, repetition or interpolation is applied.

This is evidenced by the data shown in Table 1 comparing vector and subvector error rates calculated across a number of bit-error-rates (BER) according to the following formula

$$ErrorRate = 1 - (1 - BER)^{bits} \tag{1}$$

where *bits* is the number of bits in the vectors or subvectors.

Table 1: % Error rates of vectors and subvectors vs %BER

%BER	% Error Rate of Vectors	% Error Rate of Subvectors	
		[c_i, c_{i+1}], i=1,3...11	[c_0, logE]
0.1	8.8	0.6	0.8
0.5	36.9	3.0	3.9
1.0	60.3	5.9	7.7
1.5	75.1	8.7	11.4
2.0	84.4	11.4	14.9

From Table 1 it is noticed that the error rates of subvectors are significantly lower than error rates of vectors for the same value of BER and it may therefore be advantageous to exploit error-free subvector information still remaining in erroneous vectors rather than simply replacing them. The following sections focus on the detection, extraction and exploitation of error-free subvectors.

4. SUBVECTOR-BASED ERROR CONCEALMENT

Since there is no CRC-like channel coding applied (or error checking bits allocated) at the subvector-level, the error detection at this level makes use of the data consistency test. The test is appropriate and feasible due to the temporal correlation that is present in the speech feature stream and that originates from both the overlapping in the estimation procedure of the front-end processing and from the speech production process constrained by the vocal tract inertia.

Given that n denotes the frame number and $\boldsymbol{V}$ denotes the vector, the features in a vector are formatted as

$$\begin{aligned}\boldsymbol{V}^n &= [c_1{}^n, c_2{}^n ... c_{12}{}^n, c_0{}^n, \log E^n]^T \\ &= [[c_1{}^n, c_2{}^n]...[c_{11}{}^n, c_{12}{}^n],[c_0{}^n, \log E^n]]^T \\ &= [[\boldsymbol{S}_0{}^n]^T, [\boldsymbol{S}_1{}^n]^T ... [\boldsymbol{S}_6{}^n]^T]^T \end{aligned} \tag{2}$$

where $\boldsymbol{S}_j^n$ (j=0,1…6) denotes the j'th subvector in frame n.

Since two frames in a frame-pair are consecutive, a frame-pair is represented by [$\boldsymbol{V}^n$,$\boldsymbol{V}^{n+1}$]. The consistency test is conducted within the frame-pair such that each subvector $\boldsymbol{S}_j^n$ from vector $\boldsymbol{V}^n$ is compared with its corresponding subvector S_j^{n+1} from vector $\boldsymbol{V}^{n+1}$ in the same frame-pair to evaluate if either of the two subvectors is likely to be erroneous. If any of the two decoded features in a feature-pair subvector does not possess the minimal continuity, the subvector is classified as inconsistent. Specifically the two subvectors S_j^n and S_j^{n+1} in a frame-pair are classified as inconsistent if

$$(d(S_j{}^{n+1}(0) - S_j{}^n(0)) > T_j(0))\, OR\, (d(S_j{}^{n+1}(1) - S_j{}^n(1)) > T_j(1)) \tag{3}$$

where $d(x,y)=|x-y|$ and $S_j^n(0)$ and $S_j^n(1)$ are the first and second element in the feature-pair subvector $\boldsymbol{S}_j^n$ as given in (2) respectively; otherwise, they are consistent. Thresholds $T_j(0)$ and $T_j(1)$ are constants based on measuring the statistics of error-free speech features, directly taken from the ETSI-DSR standard, and then used for the experiments on Danish digits and city names as given in Section 5. Thresholds are thus neither task nor language-dependent.

In ETSI-DSR, the data consistency test is applied only to supplement the CRC in detecting erroneous vectors at the vector-level. In the proposed subvector-based EC, however, the data consistency test is applied for discriminating between consistent and inconsistent subvectors. Only inconsistent subvectors are replaced by their nearest neighbouring consistent subvectors.

Given that $\boldsymbol{V}^n$ represents the cepstral coefficient vector of the n'th erroneous frame and that 2N frames (N frame-pairs) in error have to be mitigated. Using the notation A for the last error-free frame and B for the following error-free frame, the resulting buffered vectors are [$\boldsymbol{V}^A$,$\boldsymbol{V}^1$,$\boldsymbol{V}^2$...$\boldsymbol{V}^{2N-1}\boldsymbol{V}^{2N}\boldsymbol{V}^B$] (the same as the ETSI-DSR standard buffering), which illustrated at the subvector level is as follows:

$$\begin{array}{c}
\begin{array}{ccccccc} \boldsymbol{V}^{A} & \boldsymbol{V}^{1} & \boldsymbol{V}^{2} & . & \boldsymbol{V}^{2N-1} & \boldsymbol{V}^{2N} & \boldsymbol{V}^{B} \end{array} \\
\begin{bmatrix}
\boldsymbol{S}_0^{A} & \boldsymbol{S}_0^{1} & \boldsymbol{S}_0^{2} & . & \boldsymbol{S}_0^{2N-1} & \boldsymbol{S}_0^{2N} & \boldsymbol{S}_0^{B} \\
\boldsymbol{S}_1^{A} & \boldsymbol{S}_1^{1} & \boldsymbol{S}_1^{2} & . & \boldsymbol{S}_1^{2N-1} & \boldsymbol{S}_1^{2N} & \boldsymbol{S}_1^{B} \\
\boldsymbol{S}_2^{A} & \boldsymbol{S}_2^{1} & \boldsymbol{S}_2^{2} & . & \boldsymbol{S}_2^{2N-1} & \boldsymbol{S}_2^{2N} & \boldsymbol{S}_2^{B} \\
\boldsymbol{S}_3^{A} & \boldsymbol{S}_3^{1} & \boldsymbol{S}_3^{2} & . & \boldsymbol{S}_3^{2N-1} & \boldsymbol{S}_3^{2N} & \boldsymbol{S}_3^{B} \\
\boldsymbol{S}_4^{A} & \boldsymbol{S}_4^{1} & \boldsymbol{S}_4^{2} & . & \boldsymbol{S}_4^{2N-1} & \boldsymbol{S}_4^{2N} & \boldsymbol{S}_4^{B} \\
\boldsymbol{S}_5^{A} & \boldsymbol{S}_5^{1} & \boldsymbol{S}_5^{2} & . & \boldsymbol{S}_5^{2N-1} & \boldsymbol{S}_5^{2N} & \boldsymbol{S}_5^{B} \\
\boldsymbol{S}_6^{A} & \boldsymbol{S}_6^{1} & \boldsymbol{S}_6^{2} & . & \boldsymbol{S}_6^{2N-1} & \boldsymbol{S}_6^{2N} & \boldsymbol{S}_6^{B}
\end{bmatrix}
\end{array}$$

The first and last columns in the matrix are the error-free vectors before and following the erroneous vectors, respectively. Notation ‘1’ of $\boldsymbol{V}^{A}$ and $\boldsymbol{V}^{B}$ indicate that these subvectors are error-free. The columns in between are the erroneous vectors to be submitted to consistency test. The test generates the following consistency matrix, where notation ‘X’ is either ‘1’ or ‘0’ representing consistency or inconsistency, respectively.

$$\begin{array}{cc}
 & \begin{array}{ccccccc} \boldsymbol{V}^{A} & \boldsymbol{V}^{1} & \boldsymbol{V}^{2} & . & \boldsymbol{V}^{2N-1} & \boldsymbol{V}^{2N} & \boldsymbol{V}^{B} \end{array} \\
\begin{array}{c} \boldsymbol{S}_0 \\ \boldsymbol{S}_1 \\ \boldsymbol{S}_2 \\ \boldsymbol{S}_3 \\ \boldsymbol{S}_4 \\ \boldsymbol{S}_5 \\ \boldsymbol{S}_6 \end{array} &
\begin{bmatrix}
1 & X & X & . & X & X & 1 \\
1 & X & X & . & X & X & 1 \\
1 & X & X & . & X & X & 1 \\
1 & X & X & . & X & X & 1 \\
1 & X & X & . & X & X & 1 \\
1 & X & X & . & X & X & 1 \\
1 & X & X & . & X & X & 1
\end{bmatrix}
\end{array}$$

As an example, the consistency matrix below gives data from a consistency test applied on transmission data corrupted by the GSM EP3 (error pattern 3), corresponding to 4dB C/I.

$$\begin{array}{cc}
 & \begin{array}{cccccccccc} \boldsymbol{V}^{A} & \boldsymbol{V}^{1} & \boldsymbol{V}^{2} & \boldsymbol{V}^{3} & \boldsymbol{V}^{4} & \boldsymbol{V}^{5} & \boldsymbol{V}^{6} & \boldsymbol{V}^{7} & \boldsymbol{V}^{8} & \boldsymbol{V}^{B} \end{array} \\
\begin{array}{c} \boldsymbol{S}_0 \\ \boldsymbol{S}_1 \\ \boldsymbol{S}_2 \\ \boldsymbol{S}_3 \\ \boldsymbol{S}_4 \\ \boldsymbol{S}_5 \\ \boldsymbol{S}_6 \end{array} &
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1
\end{bmatrix}
\end{array}$$

On the basis of this consistency matrix, the error concealment is implemented in such a way that all inconsistent subvectors (marked with zero) are replaced by their nearest neighbouring consistent subvectors whereas the consistent subvectors (marked with one) are kept unchanged. For example, subvectors $\boldsymbol{S}_3$ in frame-pair [$\boldsymbol{V}^1$,$\boldsymbol{V}^2$] are inconsistent whereas subvectors $\boldsymbol{S}_3$ in frame-pair [$\boldsymbol{V}^3$,$\boldsymbol{V}^4$] are consistent. Accordingly, subvectors $\boldsymbol{S}_3$ in $\boldsymbol{V}^1$ and $\boldsymbol{V}^2$ will be replaced by $\boldsymbol{S}_3$ in $\boldsymbol{V}^A$ and $\boldsymbol{V}^3$, respectively. For subvectors $\boldsymbol{S}_6$ in vectors $\boldsymbol{V}^1$ and $\boldsymbol{V}^2$ the same substitution is conducted. The remaining subvectors in $\boldsymbol{V}^1$ and $\boldsymbol{V}^2$ are untouched.

5. PERFORMANCE EVALUATION

Three different error distributions have been used to evaluate the performance of the proposed EC algorithm, namely: 1) additive white Gaussian noise (AWGN) channels simulated by random bit error rates (BER), 2) burst-like bit errors simulated by Gilbert’s model and 3) the more realistic GSM error patterns. The recognition tasks involve the Danish digits (low perplexity) and city names (medium perplexity). The recogniser applied in the evaluation is the SpeechDat/COST 249 reference recogniser [13]. A part of the DA-FDB 4000 database is used for training 32 Gaussian mixture triphone models. Except from the EC-algorithm, the experimental settings are as described in [1,3]. The baseline word error rate (WER) (no transmission errors) for Danish digits and city names are 0.2% and 20.7%, respectively. It is shown in [8] that repetition is superior to linear interpolation and to splicing in terms of recognition accuracy. Therefore, in the evaluation of the proposed algorithm, repetition used by ETSI-DSR standard is chosen as a representative for other conventional algorithms.

5.1. AWGN Channels

The results for the Danish digits and the city names tasks for AWGN channels are illustrated in Table 2.

Table 2: %WER across the EC techniques for varying values of the BER for Danish digits and city names

BER (%)	Danish digits		City names	
	ETSI-DSR	Subvector-based	ETSI-DSR	Subvector-based
0.1	0.2	0.2	22.5	21.2
0.5	2.5	1.0	26.9	22.5
1.0	15.1	2.3	47.7	25.8
1.5	33.4	4.6	76.2	33.4
2.0	53.0	13.9	87.5	45.2

It is seen that the subvector-level EC offers better results across all BER values for both tasks.

5.2. Rayleigh Fading Channels

Errors occur not only due to noise but also due to a variety of transmission impairments so that in real communication environments errors occur in clusters separated by long error-free gaps [12], so called burst-like errors. To characterize such channels with memory, Gilbert proposed a well-known two-state Markov model composed of a “good state” G and a “burst state” B as shown in Figure 1 [12].

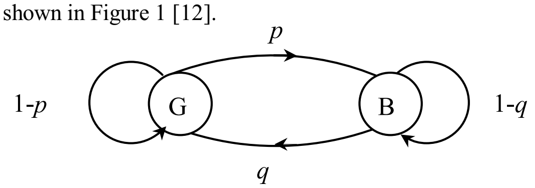

Figure 1: Gilbert’s model for bit error simulation

In this evaluation, Rayleigh fading channels are simulated by the Gilbert’s model. The parameter setting is p=0.001, h=0.1 while q is varying according to the different BER values and calculated by the following equation.

$$q = \frac{p}{BER} \times (h - BER) \quad (4)$$

where h is the bit error rate within the state B.

In applying this model, the simulated frame-error-rate (FER) can be calculated as

$$FER = \frac{p}{p+q} \times (1 - (1-h)^{Fbits}) \quad (5)$$

where *Fbits* represents the number of bits in a frame (for the frame-pair, 92-bit). As an example, given BER=1.0%, it can be worked out that q=0.009 and FER=10.0%. It is observed that this FER-value is much lower than the corresponding FER calculated for the random bit error situation, where it is 60.3%. This is due to the burst effect (non spreading of bit-errors in the speech stream).

Table 3 shows the results for the Danish digits and the city names tasks for Rayleigh fading channels.

Table 3: %WER across the EC techniques for varying values of the burst-like BER for Danish digits and city names

BER (%)	q	Danish digits		City names	
		ETSI-DSR	Subvector-based	ETSI-DSR	Subvector-based
0.1	0.099	0.2	0.2	21.4	20.5
0.5	0.019	0.8	0.4	20.3	22.0
1.0	0.009	1.9	0.6	24.5	21.4
1.5	0.0057	4.4	0.8	29.8	23.4
2.0	0.004	7.3	1.9	37.2	26.1

It is observed that the subvector-level EC achieves better results. As compared with Table 2, it is seen that burst-like bit errors degrade the performance much less than random bit errors for the same value of BER, which can be justified by their different FER values as shown above.

5.3. GSM Error Patterns

The results for the Danish digits and the city names tasks for GSM error patterns are demonstrated in Table 4.

Table 4: %WER across the EC techniques for the GSM error patterns for Danish digits and city names

EP	Danish digits		City names	
	ETSI-DSR	Subvector-based	ETSI-DSR	Subvector-based
1	0.2	0.2	20.9	20.7
2	0.2	0.2	21.1	20.7
3	9.7	1.5	38.3	29.8

The subvector-level EC gives better results. It is noted that the improvement for EP3 is significant.

6. CONCLUSION

This paper presents a subvector-level error concealment technique where subvectors in an SVQ are considered as an alternative basis for error concealment rather than the full vector. Experimental results show that the proposed algorithm - tested on a set of recognition experiments - is superior to commonly used EC methods. An advantage of the proposed method is that it has neither requirements for modification in the recogniser nor requirements for extra bandwidth.

Further work will consider adaptive thresholds in the consistency test for the voiced regions and for the transient regions.

7. ACKNOWLEDGEMENTS

This work is conducted in the context of the projects FACE "Future Adaptive Communication Environment" and CNTK (Centre for Network and Service Convergence). For more information, please refer to www.cpk.auc.dk/FACE.

8. REFERENCES

[1] ETSI ES 201 108 v1.1.2 2000, "Distributed speech recognition; front-end feature extraction algorithm; compression algorithms," April 2000.

[2] 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, USA, May 2000.

[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] Milner, B. and Semnani, S., “Robust speech recognition over IP networks”, *ICASSP*-00, Turkey, May 2000.

[5] Bawab Z.A., Locher, I., Xue, J. and Alwan, A., “Speech Recognition over Bluetooth Wireless Channels”, *Eurospeech*-03, Geneva, Switzerland, September 2003.

[6] Kim, H. K., and Cox, R. V., “A Bitstream-Based Front-End for Wireless Speech Recognition on IS-136 Communications System”, *IEEE Trans. On Speech and Audio Processing*, July 2001.

[7] C. Boulis, M. Ostendorf, E. A. Riskin and S. Otterson, "Gracefully degradation of speech recognition performance over packet-erasure networks," IEEE Trans. On Speech and Audio Processing, November 2002.

[8] Tan, Z.-H., Dalsgaard, P., and Lindberg, B., “Partial Splicing Packet Loss Concealment for Distributed Speech Recognition”, *IEE Electronics Letters*, to be published.

[9] Bernard, A. and Alwan, A, “Low-Bitrate Distributed Speech Recognition for Packet-Based and Wireless Communication”, *IEEE Trans. On Speech and Audio Processing*, November 2002.

[10] Potamianos, A and Weerackody, V., “Soft-Feature Decoding for Speech Recognition over Wireless Channels”, *ICASSP*-01, USA, May 2001.

[11] Endo, T., Kuroiwa, S., and Nakamura, S., “Missing Feature Theory Applied to Robust Speech Recognition over IP Networks”, *Eurospeech*-03, Geneva, Switzerland, Sep. 2003.

[12] Kanal, L.N. and Sastry, A.R.K., “Models for Channels with Memory and Their Applications to Error Control”, Proceedings of the IEEE, vol. 66, no. 7, July 1978.

[13] 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.