A Comparison of Packet Loss Compensation Methods and Interleaving for Speech Recognition in Burst-Like Packet Loss

Alastair James¹, Ángel Gómez² and Ben Milner¹

¹School of Computing Sciences, University of East Anglia, UK
²Department of Electronics and Computer Technology, University of Granada, Spain

Abstract

This work compares the performance of three compensation methods for speech recognition in the presence of packet loss. Two methods, cubic interpolation and a novel maximum a posteriori (MAP) estimation, aim to restore the feature vector stream in the event of packet loss, while the third technique applies compensation in the decoding stage of recognition through missing feature theory. To improve performance in burst-like packet loss, interleaving is introduced to disperse bursts of loss. Experiments on the ETSI Aurora connected digit task show best performance to be given by a combination of missing feature theory and cubic interpolation. This raises performance from 50.3% to 69.8% at a packet loss rate of 50% and average burst length of 20 packets. Including interleaving further increases performance to over 76%.

1 Introduction

The move towards mobile and handheld devices for speech communication has lead to distributed speech recognition (DSR) systems being developed. The Aurora DSR standard proposed by the European Telecommunication Standards Institute (ETSI) gives improved speech recognition accuracy by replacing the low bit-rate speech codec on the terminal device with the static MFCC feature extraction component of the speech recogniser [1]. Including noise compensation on the terminal device gives good performance on noise contaminated speech. Figure 1 shows a typical DSR architecture along with the proposals outlined in this work.

Figure 1: DSR architecture with packet loss compensation and interleaving.

The networks across which DSR systems transmit packetised speech data often do not guarantee reliable delivery. When packet loss occurs, or too many bits are corrupted so that bit level forward error correction cannot correct the frame, then portions of the feature vector stream become lost. Techniques to improve recognition performance on these unreliable channels essentially fall into three groups. The first set of techniques aim to protect the feature vectors through forward error correction. Such methods include cyclic redundancy checks (CRC) and Reed-Solomon coding [1][2]. Other techniques restore the feature vector stream in the event of lost packets through estimation of missing vectors. Simple methods use repetition of previously received vectors or linear and non-linear interpolation [3][4]. The final set of techniques apply packet loss compensation either partly or fully at the decoding stage through modification of the observation probabilities within hidden Markov models (HMM). Missing feature theory has been shown effective for this and also weighted Viterbi recognition which adjusts the contribution of estimated feature vectors according to how accurate they are likely to be [5][6]. These techniques are all effective for short duration bursts of packet loss but degrade as burst lengths increase. To reduce burst lengths, interleaving the feature vectors prior to transmission has been shown effective at giving substantial gains in recognition accuracy, although at the expense of an increase in delay [4].

The aim of this work is to compare a range of packet loss compensation techniques both with and without interleaving. Packet loss compensation methods are discussed in section 2 and a novel method of estimating lost packets is also introduced based on maximum a posteriori (MAP) estimation. Section 3 describes the interleaving process. Experimental results which compare the compensation techniques are presented in section 4 as well as a study of the optimal interleaving depth. Finally a conclusion is made in section 5.

2 Compensation against Lost Vectors

This section describes three techniques to compensate for lost feature vectors in the event of packet loss. Two methods, interpolation and MAP estimation, attempt to restore the feature vector stream by estimating lost vectors, while the third method, missing feature theory, compensates for lost vectors at the decoding stage of recognition.

2.1 Interpolation

Several interpolation schemes have been considered for estimating vectors lost due to packet loss [3][4]. Of these non-linear interpolation using cubic Hermite polynomials has been found to give best performance [4], where the nth lost vector in a burst of length β, starting at vector h+l, is given as

\[
\hat{x}_n = a_0 + a_1 t + a_2 t^2 + a_3 t^3, \quad 1 \leq n \leq \beta
\]

where \( t=n/(\beta+1) \) and the multivariate coefficients, \( \{a_0, ..., a_3\} \), can be computed from the two vectors preceding and following the burst of loss, \( \hat{x}_{before} \) and \( \hat{x}_{after} \) and their first derivatives, \( \hat{x}_{before}' \) and \( \hat{x}_{after}' \). Expressing the interpolation function in terms of Hermite basis functions gives the estimate of the nth feature vector within the burst as

\[
\hat{x}_n = \hat{x}_{before} \left( 1 - 3t^2 + 2t^3 \right) + \hat{x}_{before}' \left( 3t^2 - 2t^3 \right) + \hat{x}_{after} \left( t - 2t^2 + t^3 \right) + \hat{x}_{after}' \left( t^3 - t^2 \right), \quad 1 \leq n \leq \beta
\]
where the derivates are approximated by \( x_i^{\text{before}} = \beta (x_i - x_\text{before}) \) and \( x_i^{\text{after}} = \beta (x_i - x_\text{after}) \). This requires two vectors either side of the burst from which to compute the derivitives. If two vectors are not available the derivative is set to zero. In practice it was found that rapid fluctuations of the feature vector stream resulted in overestimation of derivative components causing the interpolation to overshoot. Improved performance was achieved by reducing large derivative estimates by applying logarithmic compression to the vector differences, 
\[
f(x) = \text{sgn}(x) \log(1 + |x|)
\]
which gives \( x_i^{\text{before}} \) and \( x_i^{\text{after}} \) as,
\[
x_i^{\text{before}} = \beta \times f(x_i - x_\text{before})
\]
\[
x_i^{\text{after}} = \beta \times f(x_i - x_\text{after})
\]

2.2 MAP estimation

A better approach for estimating the value of lost vectors is through statistical methods which use prior information about the nature of the signal [7]. In maximum a-posteriori (MAP) estimation a sequence of lost vectors, \( X_o \), can be calculated in order to maximize their likelihood conditioned on the values of the correctly received vectors, \( X_r \) and the overall distribution of the feature vector stream \( P(X_r|\mu, \Sigma) \). Assuming this distribution is Gaussian, MAP estimation can be simplified to a linear regression [8] given by,
\[
\hat{X}_m = \mu_m + \Sigma_m \Sigma_c^{-1}(X_m - \mu_c)
\]
where \( \mu_c \) and \( \mu_m \) are the mean vectors of \( X_r \) and \( X_o \) respectively, \( \Sigma_c \) is the auto-covariance matrix of \( X_r \) and \( \Sigma_m \) is the cross-covariance matrix between \( X_m \) and \( X_r \). The assumption of wide-sense stationarity leads to the assumption that the mean and covariance of the MFCC features are independent of their current position in time. Given the \( k_1 \) coefficient of the MFCC vector at time instant \( t_1, S(t_1, k_1) \), and the \( k_2 \) coefficient of the MFCC vector at time \( t_2, S(t_2, k_2) \), this assumption allows the means and covariance to be given as,
\[
\mu(t_1, k) = E(S(t_1, k)) = \mu(k)
\]
\[
\mu(t_2, k) = E(S(t_2, k)) = \mu(k)
\]
\[
E[(S(t_1, k) - \mu(k))(S(t_2, k) - \mu(k))] = c(t_1 - t_2, k_1, k_2) = c(t, k, k)
\]
where \( E[f] \) is the expectation operator, \( \mu(k) \) is the expected value of each coefficient in the MFCC vector and \( c(t, k, k) \) defines the covariance between any coefficient and any other coefficient \( r \) time instants later in the MFCC vector sequence.

Lost vectors are reconstructed by arranging the received MFCCs into \( X_r \), whilst the lost features are arranged into \( X_o \). Since both the mean of all the coefficients in the MFCC vector sequence and the covariance between any two components (equation 7) are known, an estimate of the lost vectors, \( \hat{X}_o \) can be made. In practice not all the received features in the utterance are used in \( X_r \) as this imposes too much of a computational overhead. Instead \( \hat{X}_o \) is limited to a few vectors in the region of the loss.

Isolated lost vectors can be efficiently reconstructed through MAP estimation, however bursts of lost vectors require an iterative strategy where each vector is individually reconstructed and their estimations are reused in the reconstruction of the next vector. Otherwise a large amount of \( c(t, k_1, k_2) \) values would be required or the most inner lost vectors of a burst could not be reconstructed since they are too far in time from the observed (received) features. Different strategies have been considered for the order of vector estimation within a burst with the most effective being iterative reconstruction from the outer to the inner vectors of a burst.

MAP estimation is often computationally expensive due to the inversion of large covariance matrices. However, not all observed coefficients involved in the estimation are equally relevant. Therefore, the size of the auto-covariance matrix can be reduced by reconstructing each coefficient of the feature vector separately from the most relevant observed coefficients. This selection can be made by considering their relative covariances in relation to the coefficient under estimation,
\[
r(\tau, k, k) = \frac{c(\tau, k, k)}{\sqrt{c(0, k, k)c(k, k)}}
\]
Only those features having a high relative covariance, with respect to the coefficient under estimation, are used in the MAP estimation. In particular the DCT operation applied during feature extraction means that coefficients are considerably more correlated along time than quantefy and this fact is supported by their observed relative covariances. Therefore, each coefficient of the feature vector, \( S(t,k) \), can be reconstructed from received features using the same MFCC coefficient, i.e., \( S(t-1, k), S(t+1, k), ... S(t+\tau, k) \). This gives a considerable reduction in the size of the observation covariance matrix and leads to a substantial reduction in computation [8].

2.3 Missing feature theory

The methods outlined above attempt to reconstruct the feature vector stream based on those feature vectors correctly received. An alternative approach is to compensate for lost vectors at the decoding stage through the technique of missing feature theory [5]. The observation probability, \( b_j(x_j) \), associated with the \( j \)th feature vector, \( x_j \), in state \( j \) of the HMM is modified according to whether particular coefficients of the vector were received or not,
\[
b_j(x_j) = \prod_{i=1}^{K} \left( \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left( -\frac{(x_{ij} - \mu_i)^2}{2\sigma_i^2} \right) \right)^{P_{k,i}}
\]
where \( K \) is the dimensionality of the feature vector, \( \mu_i \) and \( \sigma_i \) are the mean and variance of the \( k \)th coefficient of the feature vector in state \( j \). The variable \( P_{k,i} \) is set to 1 if the \( k \)th coefficient of the \( i \)th feature vector is suitable for inclusion in the observation calculation, or set to 0 if it is unsuitable. The feature vector in this work comprises 12 static MFCC coefficients and a log energy term which are augmented by their velocity and acceleration derivatives to give a \( K=39 \) dimensional feature vector. Temporal derivative components are calculated using a ‘sliding-window’ technique [1], where the window extends \( w_{v+3} \) frames either side of the current element for the velocity and \( w_{a+2} \) frames either side for the acceleration. As shown in [9], the loss of a burst of \( \beta \) static vectors has the effect of corrupting \( 2w_{v+\beta} \) velocity and \( 2(w_{a+\beta}+\beta) \) acceleration features. Therefore, for every static vector removed from the calculation, several velocity and acceleration features will need to be removed by setting the corresponding \( P_{k,i} \) values to 0, as shown in figure 2a. Note that for a particular \( i \), when \( P_{k,i} \) is zero for all \( k \), no information can be obtained from the
observation and the decoded state lattice depends wholly on the HMM state transition probability matrix for this frame.

![Figure 2: Methods of considering temporal derivatives](image)

In severe packet loss, where bursts of loss are positioned close together, the regions of removed temporal derivative components can overlap, causing the higher-order temporal derivatives to be removed from the calculation for an increased number of frames. In order to prevent this, an alternative method is proposed, whereby the missing static feature vectors are reconstructed (using a method such as cubic interpolation) purely for the purpose of calculating the temporal derivatives. The observed values are only set to 0 for those feature vectors that are not received, as shown in figure 2b.

3 Interleaving

The packet loss compensation methods described in the previous section are effective for short duration bursts of loss but deteriorate at longer burst lengths. An effective method to reduce burst lengths in the received feature vector stream is to employ an interleaver on the terminal device. For a given sequence of feature vectors, \( X = \{x_0, x_1, x_2, \ldots, x_k\} \), the interleaving operation can be expressed as a permutation producing a re-ordered sequence, \( X' \), given as,

\[
X' = \{x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(N)}\}
\]

(10)

The interleaving function, \( \pi(i) \), gives the index of the vector to be output at the \( i^{th} \) time instance. Feature vectors are returned to their original order on the receiver side through de-interleaving which is given by the inverse function of \( \pi \), i.e.,

\[
\pi^{-1}(j) \quad \text{where} \quad \pi(\pi^{-1}(j)) = j
\]

(11)

In the presence of burst-like packet loss the interleaving process is able to distribute long bursts of loss into a series of shorter duration losses.

The block interleaver of degree \( d \) operates by re-arranging the transmission order of a \( d \times d \) block of input vectors. Two block interleavers, \( \delta_{\text{block}} \) and \( s_{\text{block}} \) [4] are considered optimal in terms of maximising their spread for a given degree, and are given,

\[
\delta_{\text{block}} = (d-1-j)d + i \quad \text{where} \quad 0 \leq i,j \leq d-1
\]

(12)

\[
s_{\text{block}} = jd + (d-1-i) \quad \text{where} \quad 0 \leq i,j \leq d-1
\]

(13)

It is interesting to observe that \( \pi_1 \) and \( \pi_2 \) form an invertible pair as \( \pi_1 = \pi_2^{-1} \) and \( \pi_2 = \pi_1^{-1} \). The interleaving operation is equivalent to a rotation of the block of feature vectors either 90° clockwise or 90° anti-clockwise as shown in figure 3.

![Figure 3: Rotation of block 90° anti-clockwise.](image)

The degree of the interleaver determines both the spread and delay of the interleaver. For the block interleaver the delay, \( \delta_{\text{block}} \), and spread, \( s_{\text{block}} \), are given as,

\[
\delta_{\text{block}} = d^2 - d \quad \text{and} \quad s_{\text{block}} = d
\]

(14)

This shows that increasing the degree of the interleaver increases its ability to disperse bursts of loss, but at the expense of increasing delay.

4 Experimental Results

The experiments in this section first compare the effectiveness of the three lost vector compensation methods. The effect of interleaving is then considered together with an examination into the effect of increasing the interleaving depth.

The recognition task for all experiments is the Aurora connected digit database [1]. Digits are modelled using 16-state, 3-mode HMMs, trained from the set of clean digits. The test set comprises 4004 noise-free digits strings (13,159 digits in total) which gives baseline accuracy of 99% with 95% confidence error bands of +/- 0.38% at 95% accuracy. As per the ETSI standard, two vectors are carried by each packet.

4.1 Lost vector compensation

The effectiveness of the lost vector compensation methods are evaluated on four different channels which were simulated by a 3-state Markov chain [4]. Table 1 shows the conditions of the four channels which vary in terms of the packet loss rate, \( \alpha \), and average burst length, \( \beta \).

<table>
<thead>
<tr>
<th>Packet loss rate, ( \alpha )</th>
<th>Av. burst length, ( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel A 10%</td>
<td>4 packets</td>
</tr>
<tr>
<td>Channel B 10%</td>
<td>20 packets</td>
</tr>
<tr>
<td>Channel C 50%</td>
<td>4 packets</td>
</tr>
<tr>
<td>Channel D 50%</td>
<td>20 packets</td>
</tr>
</tbody>
</table>

Table 1: Simulated channel conditions

Table 2 shows recognition performance of the compensation methods for the channel conditions A to D. For MAP estimation and missing feature theory, the two variants of each technique, as discussed in sections 2.2 and 2.3, are evaluated. For the MAP estimation methods, \( \tau = 3 \) time instants before and after the lost were used.

Considerable improvements are attained by applying the methods of compensation considered in this work. MAP methods give higher accuracy than cubic interpolation. It can be seen that missing feature theory methods generally out-perform the reconstruction methods, particularly when the average burst length is large (channels B and D). Also, when the average burst length is short (channels A and C) there is a substantial difference between the performances of the two missing feature methods. This is because for these channels, the bursts of lost vectors are positioned closer together, causing the triangular regions removed by the first missing feature method to overlap. Restraining MAP estimation to the
The bandlimited case gives slightly higher accuracy than original MAP estimation but at a considerable reduction in complexity.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>No compensation</td>
<td>92.2</td>
<td>89.5</td>
<td>50.6</td>
<td>50.3</td>
</tr>
<tr>
<td>Cubic interpolation</td>
<td>96.9</td>
<td>91.3</td>
<td>86.1</td>
<td>59.3</td>
</tr>
<tr>
<td>MAP</td>
<td>96.9</td>
<td>92.0</td>
<td>86.5</td>
<td>61.5</td>
</tr>
<tr>
<td>MAP – bandlimited</td>
<td>97.0</td>
<td>92.0</td>
<td>86.6</td>
<td>61.9</td>
</tr>
<tr>
<td>Missing – triangle</td>
<td>96.8</td>
<td>93.1</td>
<td>85.8</td>
<td>67.3</td>
</tr>
<tr>
<td>Missing – interpolate</td>
<td>97.5</td>
<td>93.4</td>
<td>90.0</td>
<td>69.8</td>
</tr>
</tbody>
</table>

Table 2: Recognition accuracy with no interleaving

4.2 Interleaving

The experiments in this section now apply interleaving to the feature vector stream prior to transmission. The experimental configuration is identical to that in the previous section and table 3 shows recognition accuracy attained by the various forms of the three compensation methods and a block interleaver (as described in section 3) of d = 4.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>No compensation</td>
<td>94.1</td>
<td>90.0</td>
<td>54.0</td>
<td>52.4</td>
</tr>
<tr>
<td>Cubic interpolation</td>
<td>98.4</td>
<td>93.3</td>
<td>93.0</td>
<td>67.9</td>
</tr>
<tr>
<td>MAP</td>
<td>98.3</td>
<td>93.7</td>
<td>93.1</td>
<td>70.1</td>
</tr>
<tr>
<td>MAP – bandlimited</td>
<td>98.3</td>
<td>93.8</td>
<td>93.1</td>
<td>70.7</td>
</tr>
<tr>
<td>Missing – triangle</td>
<td>97.5</td>
<td>93.9</td>
<td>88.5</td>
<td>68.8</td>
</tr>
<tr>
<td>Missing – interpolate</td>
<td>98.4</td>
<td>94.6</td>
<td>94.5</td>
<td>76.2</td>
</tr>
</tbody>
</table>

Table 3: Recognition accuracy with block interleaving

Comparing the interleaved results with the no interleaving results of table 2 shows an increase in recognition accuracy for all compensation methods, the magnitude of which is greater if the average burst length of the channel was large before interleaving. Note that, for the first missing feature method, for channels A and C (where the average burst length was short before interleaving), a less pronounced increase in performance is observed. This is because the decrease in burst length results in a higher occurrence of overlap.

A shop interleaver’s ability to disperse packet loss is related to its degree [4]. Figures 4a and 4b show the effect of increasing interleaving depth in the range d=1 (corresponding to no interleaving) to d=8 on channels A and B described above. A similar pattern of results is observed as from table 3, in that increasing the interleaving depth results in an increase in word accuracy for all methods, however, this increase is less pronounced for the first missing feature method for the same reasons as out-lined above. In figure 4a a leveling off of accuracy is observed where the depth of the interleaver becomes sufficiently large to fully distribute the bursts of packet loss of channel A (d=6). However, this is not repeated in figure 4b as an interleaving degree of nearly d=40 would be required in order to distribute the longer burst lengths. As indicated by equation 14 this would introduce prohibitively long delays to the system.

5 Conclusions

This work has shown that packet loss can have a severe effect on the accuracy of DSR systems. The methods outlined here give substantial improvements in these conditions. Results suggest that it is more beneficial to compensate for lost vectors in the decoding stage of the recogniser, rather than attempting to reconstruct the feature vector stream beforehand. This is especially true in the presence of large bursts of losses, as the accuracy of such methods falls off as burst length increases.

Of the reconstruction methods, it has been shown that MAP estimation performs significantly better than cubic interpolation. In addition to this, performing MAP estimation using only those values in the same quefrency band results in a slight increase in recognition accuracy whilst greatly decreasing the complexity of the estimation.

With missing feature methods, it is important to consider how to treat corrupted temporal derivates within the feature vector stream. One approach is to remove all temporal derivates affected by the loss of static elements from the feature vector stream. However, substantially improved performance was achieved by calculating the temporal derivatives using a reconstructed version of the static feature vector stream prior to recognition.

Interleaving has given substantial increases in recognition accuracy by dispersing bursts of packet loss. However the degree of the interleaver, and hence its delay, must be carefully considered in the design of such a system.

6 Acknowledgements

The authors gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) (grant GR/R88243/01) and the Spanish CICYT Project TIC-2001-3323 in funding this work.

7 References