

A Comparative Study of Feature-Domain Error Concealment Techniques for Distributed Speech Recognition

Zheng-Hua Tan, Børge Lindberg and Paul Dalsgaard

SMC–Speech and Multimedia Communication, Department of Communication Technology,
Aalborg University, Denmark

{zt, bli, pd}@kom.aau.dk

Abstract

This paper presents a comparative study of different error concealment (EC) techniques in the context of distributed speech recognition (DSR) that exploits repetition, interpolation or subvector concealment to counteract transmission errors.

A number of experiments are conducted and the results demonstrate that repetition is as good as, or even better than, linear interpolation whereas the subvector concealment shows the best performance in terms of recognition accuracy. Further experiments and analyses are conducted with the purpose of uncovering the reasons for the different characteristics of the EC techniques: speech features are inspected, time normalised distances as well as hidden Markov model (HMM) state durations are compared for different EC techniques.

1. Introduction

Transmission across wireless networks may cause errors that severely degrade the accuracy of automatic speech recognition (ASR). The degradation introduced by transmission errors can, however, be partly alleviated by introducing various EC techniques.

Two classes of EC techniques exist within DSR where the client is always at the sender side and the server at the receiver side, namely client based EC and server based EC. Although client based techniques e.g. retransmission and forward error control (FEC) can result in recovering a large amount of transmission errors, generally the disadvantages are additional delay, increased bandwidth and higher computational overhead [1]. In server based EC techniques the redundancy in the transmitted signal is exploited. This may be used on its own or in combination with a client based technique. When used in combination, the purpose is to handle the remaining errors, which a pure client based technique fails to recover. In this paper only server based EC techniques are considered.

Server based EC can be conducted either in the feature-domain or in the model-domain. Feature-domain EC generally employs one of the following techniques: splicing, substitution (with silence, noise or source-data), repetition or interpolation [2]-[5]. Partial splicing and subvector concealment are recently proposed in [6] and [7], respectively.

In the model-domain, the effect of transmission errors can be mitigated by integration based EC techniques. Both [8] and [9] integrate the reliability of the channel-decoded features into the recognition process where the Viterbi decoding algorithms are modified such that contributions made by observation probabilities associated with features estimated from erroneous features are decreased. In [10], a theory of missing features has

been applied to error-robust ASR where erroneous features generate constant contributions to the Viterbi decoding with the aim of neutralising these features. In general, model-domain EC techniques have a requirement of modification in the recogniser.

This paper focuses on feature-domain EC techniques only, gives a survey of these and presents a number of experiments and analyses of their individual merits. The feature-domain EC techniques focused on in this paper are repetition, interpolation and subvector concealment.

2. Feature-domain EC techniques

The general purpose of feature-domain EC techniques is to generate substitutions for the erroneous features as close to the original ones as possible with the goal of improving the overall recognition accuracy of the system.

2.1. Insertion based techniques

In each of the insertion based EC techniques an erroneous frame is substituted by inserting a 'fill-in' frame [1]. Each 'fill-in' frame may equate silence, noise, an estimated value (for example a mean value over training data), or a repetition of a neighbouring frame.

In [2] two 'fill-in' principles, namely zeros (silence) and the mean-value frame over all training data, and a splicing are applied. The mean-value substitution is reported to outperform the zero-substitution and splicing.

In splicing a number of consecutive erroneous frames are dropped. An obvious side effect of employing splicing is a decrease in the Viterbi decoding time caused by a shorter feature stream [3].

The ETSI-DSR standard [4] applies a repetition that replaces the first half of a series of erroneous frames with a copy of the last correct frame before the error and the second half with a copy of the first correct frame following the error.

Partial splicing presented in [6] substitutes erroneous frames partly by a repetition of neighbouring frames and partly by a splicing. It can be shown that under certain assumptions the partial splicing is equivalent to a modified Viterbi decoding algorithm.

2.2. Interpolation based technique

Interpolation exploits the temporal correlation being present in the speech feature stream, originating from both the overlapping in the feature estimation procedure itself and from the speech production process.

The most commonly used interpolation technique is applying a linear interpolation as an estimate of the erroneous frames [5]. In [2] interpolation has achieved better results than splicing, silence substitution and mean-value insertion.

2.3. Subvector concealment

It is observed that the conventional EC techniques discussed above conduct concealment at the full vector – in this paper equivalent to frame - level only: A vector is the unit selected for error detection, and if erroneous then followed by a full substitution. This is the common characteristic of vector level EC algorithms no matter whether splicing, substitution, repetition or interpolation is applied. The vector level EC strategy, however, fails to exploit the error free fractions left within erroneous vectors.

Prior to further discussion, let us first introduce the ETSI-DSR standard [4]. In the standard, the front-end produces a 14-element vector consisting of log energy (logE) and 13 mel-frequency cepstral coefficients (MFCC) ranging from c_0 to c_{12} . Each feature vector is compressed using split vector quantization (SVQ). The SVQ algorithm groups two features (either $\{c_i$ and c_{i+1} , $i=1, 3\dots 11\}$ or $\{c_0$ and $\log E\}$) into a feature-pair subvector resulting in seven subvectors in one vector. Each subvector is quantized using its own SVQ codebook.

Two quantized vectors are grouped together and protected by a cyclic redundancy check (CRC) creating a frame-pair. Frame-pairs further form a bitstream for transmission.

At the server side two calculations determine whether or not a frame-pair is received with errors, namely a CRC test and a data consistency test. In the EC processing, a repetition scheme is applied to replace erroneous vectors.

It is, however, highly likely that not all subvectors in an erroneous vector are corrupted by errors. It is noticed that the error rates of the subvectors are significantly lower than full vectors for the same bit error rate (BER) values [7]. The exploitation of the potential error-free information embedded in each erroneous vector – rather than simply substituting them – leads to a subvector-level EC scheme in which each subvector is selected as the basis for supplementary error detection and mitigation.

Since there is no CRC coding applied at the subvector level, error detection at this level can only make use of a data consistency test.

Given that n denotes the frame number and V the feature vector, each vector is formatted as

$$\begin{aligned} V^n &= [c_1^n, c_2^n \dots c_{12}^n, c_0^n, \log E^n]^T \\ &= [[c_1^n, c_2^n] \dots [c_{11}^n, c_{12}^n], [c_0^n, \log E^n]]^T \\ &= [[S_0^n]^T, [S_1^n]^T \dots [S_6^n]^T]^T \end{aligned} \quad (1)$$

where S_j^n ($j=0, 1 \dots 6$) denotes the j 'th subvector in frame n .

The consistency test is conducted across consecutive frame-pair vectors $[V^n, V^{n+1}]$ such that each subvector S_j^n from V^n is compared with its corresponding subvector S_j^{n+1} from V^{n+1} . If any of the two decoded features in a feature-pair subvector does not possess a minimal continuity, the subvector is classified as inconsistent. Specifically both subvectors S_j^n and S_j^{n+1} in a frame-pair are classified as inconsistent if

$$(d(S_j^{n+1}(0) - S_j^n(0)) > T_j(0)) \text{OR} (d(S_j^{n+1}(1) - S_j^n(1)) > T_j(1)) \quad (2)$$

where $d(x,y) = |x-y|$ and $S_j^n(0)$ and $S_j^{n+1}(0)$ and $S_j^n(1)$ and $S_j^{n+1}(1)$ are the first and second element, respectively, in the feature-pair subvectors S_j^n and S_j^{n+1} as given in (1); otherwise, they are

classified as consistent. The thresholds $T_j(0)$ and $T_j(1)$ are constants given on the basis of measuring the statistics of error free speech features.

The data consistency test generates a consistency matrix that discriminates between consistent and inconsistent subvectors. Only inconsistent subvectors are replaced by their nearest neighbouring consistent subvectors whereas the consistent subvectors are kept unchanged. Details are presented in [7].

3. Recognition experiments

To investigate the behaviour of the different EC techniques, recognition experiments are conducted for two recognition tasks, namely: Danish digits recognition and city names recognition.

The recogniser applied in the experiments is the SpeechDat/COST 249 reference recogniser [11]. A fully automatic, language-independent training procedure is used for building a phonetic recogniser based on the HTK toolkit and the SpeechDat (II) compatible database DA-FDB 4000. This database covers speech from 4000 Danish speakers collected over the fixed network (FDB). A part of the DA-FDB 4000 database is used for the training of 32 Gaussian mixture triphone models. The independent test data - isolated digits and city names - are from the same database.

The experimental setting for testing the repetition technique is as defined in the ETSI-DSR standard [4]. However, modifications are introduced to enable the testing of the interpolation and subvector concealment. The threshold values given in the ETSI-DSR standard for the data consistency test are used for subvector concealment when conducting the subvector consistency test as given in (2).

The realistic GSM error patterns (EP) are used as they include a merging of both random errors and burst-like errors. These EPs are commonly used for testing speech codecs and DSR EC techniques. The three EPs are EP1, EP2 and EP3 corresponding to carrier-to-interference (C/I) ratios of 10 dB, 7dB and 4dB, respectively.

3.1. Recognition results

The baseline word error rate (WER) (no transmission errors) for Danish digits and city names are 0.2% and 20.7%, respectively.

Figure 1 (a) and (b) provide the experimental results from the Danish digits and city names, respectively.

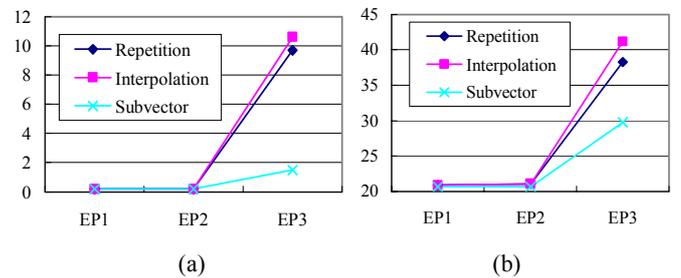


Figure 1: The %WER for three EC techniques tested on three GSM EPs for (a) Danish digits and (b) city names

The results show that repetition is slightly better than interpolation whereas the subvector concealment gives the best results. More comparisons and recognition results are available in [6] and [7].

4. Comparative study

In this section a comparative study is conducted with the aim of revealing possible causes for the different WERs observed for the different EC techniques. The study involves comparison of the MFCC features, the dynamic programming (DP) distances as well as the HMM state durations.

In all experiments used for the study transmission errors of a random BER value of 2% is used.

4.1. MFCC features

The original error-free MFCC features are directly compared with the features corrupted with errors but concealed either by repetition (rMFCC), by interpolation (iMFCC) or by subvector concealment (sMFCC). The test utterance is word “et”. The coefficient c_0 is taken as an example due to its ability of emphasising transitions between vowels and consonants.

The effects of the three techniques on the MFCC, the Δ and the $\Delta\text{-}\Delta$ coefficients are exemplified in Figure 2-5, respectively.

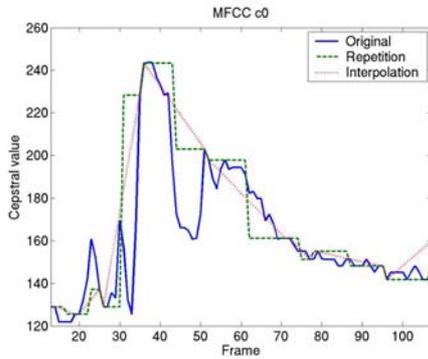


Figure 2: MFCC, rMFCC and iMFCC c_0 for word “et”

From Figure 2 it is observed that the rMFCC-curve traces the MFCC-curve better than the iMFCC-curve indicating a better reconstruction of erroneous frame values by employing a repetition technique. The sMFCC-curve, however, traces the MFCC-curve best, as shown in Figure 3.

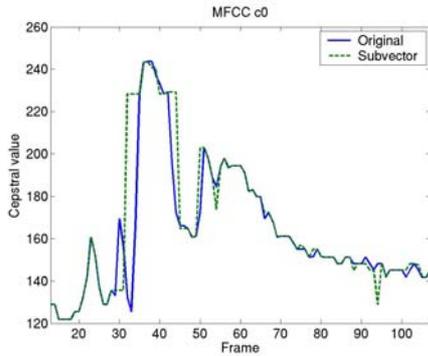


Figure 3: MFCC and sMFCC c_0 for word “et”

From Figures 4 and 5 it is seen that the $\Delta\text{-rMFCC}$ and $\Delta\text{-}\Delta\text{-rMFCC}$ features trace the corresponding error-free features better than the $\Delta\text{-iMFCC}$ and $\Delta\text{-}\Delta\text{-iMFCC}$ features. This may be explained as follows. As frames are reconstructed by an interpolation technique into a straight line of iMFCC features shown in Figure 2, this results in a constant value segment in the

$\Delta\text{-iMFCC}$ and consequently in a zero value segment in the corresponding $\Delta\text{-}\Delta\text{-iMFCC}$ thus causing less information available for the Viterbi decoding. In contrast, when applying repetition, a fast change is introduced in the middle of erroneous frames.

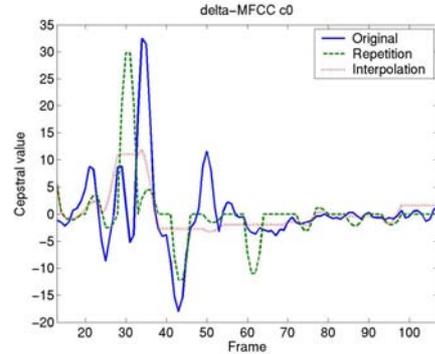


Figure 4: $\Delta\text{-MFCC}$, $\Delta\text{-rMFCC}$ and $\Delta\text{-iMFCC}$ c_0 for word “et”

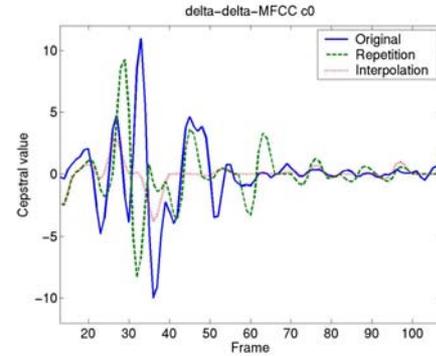


Figure 5: $\Delta\text{-}\Delta\text{-MFCC}$, $\Delta\text{-}\Delta\text{-rMFCC}$ and $\Delta\text{-}\Delta\text{-iMFCC}$ c_0 for word “et”

In Figure 2, 4 and 5, the MFCC and the rMFCC feature curves seem to display similar shapes even though there are some displacements along the time axis as compared to the iMFCC feature. However, the DP embedded in the Viterbi algorithm makes this displacement relatively irrelevant, which is evident from the discussion of DP distances in Section 4.2.

The above observations are found in other cepstral coefficients and on other utterances as well.

In general, it seems that the rapid changes often appearing in MFCC coefficients do not justify the introduction of linear interpolation, especially in segments spanning over phoneme boundaries (in Figure 2, frame 48 approximately corresponds to a vowel/consonant boundary).

4.2. DP distances

Theoretically, interpolation must result in a smaller Euclidean distance between MFCC and iMFCC than repetition when averaged over a full signal. That is why there has been a common/general expectation that interpolation must perform better than repetition in terms of recognition accuracy.

However, average distance is not analogue with the Viterbi based matching. Therefore, instead of Euclidean distance, time-normalized DP distances were computed by the symmetric dynamic time warping (DTW) between error-free features and features derived by repetition, interpolation and subvector concealment according to [12].

The Euclidean and DP distances between c_0 of MFCC and MFCC generated by different EC techniques for word “et” are shown in Figure 6.

The results show that the rMFCC feature has smaller DP distances to the original MFCC than the iMFCC feature though larger Euclidean distances. The distances between MFCC and sMFCC are always the smallest.

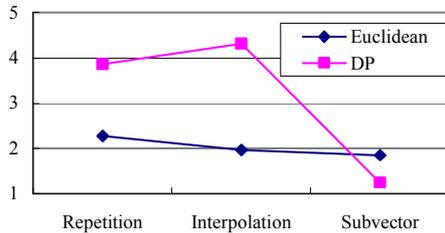


Figure 6: The Euclidean and DP distances between c_0 of MFCC and MFCC generated by different EC techniques for word “et”.

The experiments are extended to a number of utterances. Results show that, over 328 testing utterances, 295 iMFCC features have smaller Euclidean distances to the original MFCC than rMFCC features whereas only 33 rMFCC features have smaller Euclidean distances than iMFCC features. However, features having smaller DP distances to the original MFCC for iMFCC and for rMFCC are 146 and 182, respectively, indicating that repetition performs better in terms of DP distance. sMFCC features always have the smallest for both distances.

4.3. HMM state durations

For the purpose of studying the decoding process itself, a set of speech recognition experiments are conducted in which the Viterbi decoding keeps track of the HMM state alignment. Each state-duration (number of frames) is tracked and counted during recognition. The duration corresponding to the part of an utterance recognised as speech is summed up and divided by the total number of states of the speech part in the testing.

The average state-durations over eleven test utterances (one for each digit including two variants for one digit) for error-free features, for features calculated by repetition, by interpolation and by subvector concealment are 5.253, 4.023, 3.736 and 5.345 frames, respectively. From this, two facts are observed.

First, it shows that interpolation gives the smallest average state-duration indicating that features calculated by interpolation result in faster transition from one HMM state to the following state whereas features reconstructed by repetition result in a Viterbi search in which each state-duration is longer. This may be explained from the fact that interpolation, in contrast to repetition, potentially generates artefact-features that do not exist in the training data, and therefore ‘mislead’ the search in the decoding process.

Second, the average state-duration for features calculated by subvector concealment is very close to the one for error-free features. In addition, close analyses of detailed experimental data also show that both start and end frame of the recognised speech part are close to each other for error-free features and for features calculated by subvector concealment. This justifies that subvector concealment provides a good reconstruction of erroneous features.

5. Conclusions

In this paper, three different EC techniques have been compared. It has been experimentally verified that the simple repetition technique - measured by its influence on WER - in general is as good as or even better than linear interpolation. Subvector concealment is the best performing technique of the three.

The MFCC features have been directly compared and it is observed that the repetition generated features, and their derivatives, trace the original features better than those obtained by interpolation. Again the subvector concealment generated features exhibit superior tracing ability.

Further experiments are the comparison of Euclidean and DP distances. It is observed that interpolation generally gives smaller Euclidean distance but larger DP distance as compared to repetition. The subvector concealment achieves the smallest for both distances.

Finally, from measurements of HMM state durations, it is observed that interpolation results in faster state transitions in the decoding compared to repetition. This is explained from the fact that interpolation potentially introduces artefact features that do not exist in the training data. The subvector concealment gives almost the same average state-duration as observed for error-free features.

6. References

- [1] Perkins, C., Hodson, O., and Hardman V., “A Survey of Packet Loss Recovery Techniques for Streaming Audio”, *IEEE Network*, September/October 1998.
- [2] Boulis, C., Ostendorf, M., Riskin, E.A. and Otterson, S., “Graceful Degradation of Speech Recognition Performance over Packet-Erasure Networks”, *IEEE Trans. On Speech and Audio Processing*, November 2002.
- [3] 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.
- [4] 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
- [5] Milner, B. and Semnani, S., “Robust speech recognition over IP networks”, *ICASSP-00*, Turkey, May 2000.
- [6] Tan, Z. -H., Dalsgaard, P. and Lindberg, B., “Partial Splicing Packet Loss Concealment for Distributed Speech Recognition,” *IEE Electronics Letters*, vol.39, no.22, pp. 1619-1620, October 2003.
- [7] Tan, Z. -H., Dalsgaard, P. and Lindberg, B., “A Subvector-Based Error Concealment Algorithm for Speech Recognition over Mobile Networks,” *ICASSP-04*, Montreal, Quebec, Canada, May 2004.
- [8] 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.
- [9] Potamianos, A. and Weerackody, V., “Soft-feature Decoding for Speech Recognition over Wireless Channels,” *ICASSP-01*, USA, May 2001.
- [10] Endo, T., Kuroiwa, S., and Nakamura, S., “Missing Feature Theory Applied to Robust Speech Recognition over IP Networks”, *Eurospeech-03*, Geneva, Switzerland, September 2003.
- [11] Lindberg, B., Johansen, F. T., Warakagoda, N., et al, “A Noise Robust Multilingual Reference Recogniser Based on SpeechDat(II),” in Proc. ICSLP-2000, October 2000.
- [12] Sakoe H., and Chiba S., “Dynamic Programming Algorithm Optimization for Spoken Word Recognition”, *IEEE Trans. On Acoustics, Speech, and Signal Processing*, February 1978.