# Robust Speech Recognition over Mobile Networks Using Combined Weighted Viterbi Decoding and Subvector Based Error Concealment

*Zheng-Hua Tan, Paul Dalsgaard and Børge Lindberg*

Speech and Multimedia Communication (SMC), Department of Communication Technology
Aalborg University, Denmark
{zt, pd, bli}@kom.aau.dk

## ABSTRACT

Robustness against transmission errors is one of the primary barriers to the widespread application of automatic speech recognition (ASR) in mobile communications. We have previously proposed a subvector based error concealment (EC) method that conducts error detection and mitigation in the feature-domain at the subvector level. This paper presents a weighted Viterbi decoding (WVD) algorithm that works in the model domain for counteracting unreliable features generated by the subvector based EC. The reliability of each feature is estimated during the process of subvector based EC and is used by the WVD for modifying the observation probability of the feature. Recognition experiments are conducted on the Aurora 2 database corrupted by GSM error pattern EP3. Combining the WVD and the subvector EC achieves 70% and 24% performance improvement as compared to the ETSI-DSR standard and the subvector based EC, respectively.

**Index Terms**: distributed speech recognition, error concealment, split vector quantization, weighted Viterbi

## 1. INTRODUCTION

With the increasing number of mobile devices and the development of ubiquitous networking, distributed speech recognition (DSR) is advantageous in terms of low computational requirements and power consumption for devices at the client side and effortless update of the core part of the recogniser at the server side. Nevertheless, error-prone channels in mobile communications are the key problem in making DSR applications robust [1], [2] and severe degradations in recognition performance have been demonstrated for such channels (e.g. reported in [3] for GSM error pattern EP3).

Approaches to handle this problem consist of client-based error recovery, and server-based EC which is further classified into feature-reconstruction and ASR-decoder EC methods. Client-based error recovery requires the client to exploit the characteristics of channels and signals. The deployment of client-based techniques is always a trade-off between the achieved performance and the required resources. One disadvantage of client-based techniques is their weak compatibility due to the required modifications in the client. In contrast, server-based EC methods do not require modifications at the DSR client-side, thus guaranteeing compatibility with the existing ETSI-DSR standards. This paper focuses on server-based EC methods only.

Conventional server-based feature reconstruction methods such as repetition [4] and interpolation [5] disregard erroneous feature vectors and reconstruct them on the basis of received error-free vectors. As opposed to this, the recently proposed subvector EC [6] has shown substantially improved performance by conducting EC at the subvector level. It is however observed that the features reconstructed are potentially unreliable since the error detection at the subvector level uses a threshold based data consistency test rather than the more reliable cyclic redundancy check (CRC). On the other hand the data consistency test by nature generates a reliability measure for each subvector which can then be exploited by weighted Viterbi decoding [7] in the model domain.

WVD generally introduces exponential weighting factors into the calculation of the observation probability to decrease or neutralise contributions made by features or feature vectors with low reliability. Weighting factors may be computed either from the bit reliability information given by the network channel decoder that applies a soft-decision or by using an estimated value for a hard-decision channel decoder [8], [9]. The first method requires a known bit probability which is often not the case [10] whereas the second method removes the requirement of a known bit probability and is applicable to a wider range of channels. The WVD method introduced in this paper falls in the second category and is implemented as a follow-up to the subvector based EC.

## 2. SUBVECTOR BASED EC

Distinct from conventional vector level EC algorithms, the subvector based EC consider each subvector in a feature vector as a supplementary basis for error detection and mitigation. This is realised by exploiting the temporal correlation present in the speech features to identify inconsistent subvectors within erroneous vectors and replacing each inconsistent subvector with its nearest neighbouring consistent subvector.

Let us first introduce the ETSI-DSR standard [4]. In the standard, the front-end produces a 14-element vector consisting of log energy $(\log E)$ 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 $\{c_i \text{ and } c_{i+1}, i = 1, 3, ..., 11\}$ or $\{c_0 \text{ and } \log E\}$) 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 CRC creating a frame-pair. Frame-pairs are further concatenated to form a bit-stream 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 vector repetition scheme is applied to replace erroneous vectors.

Let us then introduce the subvector based EC. Given that *t* denotes the frame number and *V* the feature vector, the vector is formatted as

$$\begin{aligned} \boldsymbol{V}^t &= [c_1{}^t, c_2{}^t \ldots c_{12}{}^t, c_0{}^t, \log E^t]^T \\ &= [[c_1{}^t, c_2{}^t] \ldots [c_{11}{}^t, c_{12}{}^t], [c_0{}^t, \log E^t]]^T \\ &= [[\mathrm{S}_0{}^t]^T, [\mathrm{S}_1{}^t]^T \ldots [\mathrm{S}_6{}^t]^T]^T \qquad (1) \end{aligned}$$

where $\boldsymbol{S}_j{}^t$ (*j*=0,1…6) denotes the *j*'th subvector in frame *t*. Two consecutive frames in a frame-pair are represented by [$\boldsymbol{V}^t, \boldsymbol{V}^{t+1}$]. The consistency test is conducted within the frame-pair such that each subvector $\boldsymbol{S}_j{}^t$ from $\boldsymbol{V}^t$ is compared with its corresponding subvector $\boldsymbol{S}_j{}^{t+1}$ from $\boldsymbol{V}^{t+1}$ to evaluate if any of the two subvectors is likely to be erroneous. 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 subvectors $\boldsymbol{S}_j{}^t$ and $\boldsymbol{S}_j{}^{t+1}$ in a frame-pair are classified as inconsistent if

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

where $d(x,y) = |x-y|$ and $S_j{}^t(0)$ and $S_j{}^{t+1}(0)$ and $S_j{}^t(1)$ and $S_j{}^{t+1}(1)$ are the first and second element, respectively, in the feature-pair subvectors $\boldsymbol{S}_j{}^t$ and $\boldsymbol{S}_j{}^{t+1}$ as given in Eq. (1); otherwise, they are consistent. Thresholds $T_j(0)$ and $T_j(1)$ are constants based on measuring the statistics of error-free speech features.

Assume there are 2*N* frames (*N* frame-pairs) in error to be mitigated. Using the notation *A* for the last error-free frame and *B* for the following error-free frame, the ETSI-DSR standard buffered vectors are [$\boldsymbol{V}^A, \boldsymbol{V}^{A+1}, \boldsymbol{V}^{A+2} \ldots \boldsymbol{V}^{A+2N-1}, \boldsymbol{V}^{A+2N}, \boldsymbol{V}^B$], as illustrated in Fig. 1 at the subvector level.

$$\begin{array}{ccccccc} \boldsymbol{V}^{A} & \boldsymbol{V}^{A+1} & \boldsymbol{V}^{A+2} & . & \boldsymbol{V}^{A+2N-1} & \boldsymbol{V}^{A+2N} & \boldsymbol{V}^{B} \\ \left[\begin{matrix} \boldsymbol{S}_0{}^A \\ \boldsymbol{S}_1{}^A \\ \boldsymbol{S}_2{}^A \\ \boldsymbol{S}_3{}^A \\ \boldsymbol{S}_4{}^A \\ \boldsymbol{S}_5{}^A \\ \boldsymbol{S}_6{}^A \end{matrix}\right. & \begin{matrix} \boldsymbol{S}_0{}^{A+1} \\ \boldsymbol{S}_1{}^{A+1} \\ \boldsymbol{S}_2{}^{A+1} \\ \boldsymbol{S}_3{}^{A+1} \\ \boldsymbol{S}_4{}^{A+1} \\ \boldsymbol{S}_5{}^{A+1} \\ \boldsymbol{S}_6{}^{A+1} \end{matrix} & \begin{matrix} \boldsymbol{S}_0{}^{A+2} \\ \boldsymbol{S}_1{}^{A+2} \\ \boldsymbol{S}_2{}^{A+2} \\ \boldsymbol{S}_3{}^{A+2} \\ \boldsymbol{S}_4{}^{A+2} \\ \boldsymbol{S}_5{}^{A+2} \\ \boldsymbol{S}_6{}^{A+2} \end{matrix} & \begin{matrix} . \\ . \\ . \\ . \\ . \\ . \\ . \end{matrix} & \begin{matrix} \boldsymbol{S}_0{}^{A+2N-1} \\ \boldsymbol{S}_1{}^{A+2N-1} \\ \boldsymbol{S}_2{}^{A+2N-1} \\ \boldsymbol{S}_3{}^{A+2N-1} \\ \boldsymbol{S}_4{}^{A+2N-1} \\ \boldsymbol{S}_5{}^{A+2N-1} \\ \boldsymbol{S}_6{}^{A+2N-1} \end{matrix} & \begin{matrix} \boldsymbol{S}_0{}^{A+2N} \\ \boldsymbol{S}_1{}^{A+2N} \\ \boldsymbol{S}_2{}^{A+2N} \\ \boldsymbol{S}_3{}^{A+2N} \\ \boldsymbol{S}_4{}^{A+2N} \\ \boldsymbol{S}_5{}^{A+2N} \\ \boldsymbol{S}_6{}^{A+2N} \end{matrix} & \left.\begin{matrix} \boldsymbol{S}_0{}^B \\ \boldsymbol{S}_1{}^B \\ \boldsymbol{S}_2{}^B \\ \boldsymbol{S}_3{}^B \\ \boldsymbol{S}_4{}^B \\ \boldsymbol{S}_5{}^B \\ \boldsymbol{S}_6{}^B \end{matrix}\right] \end{array}$$

*Figure 1*: ETSI-DSR buffering matrix.

In the buffering matrix columns $\boldsymbol{V}^A$ and $\boldsymbol{V}^B$ are the error-free vectors with 2*N* erroneous vectors received in between. The 2*N* vectors $\boldsymbol{V}^{A+1} \ldots \boldsymbol{V}^{A+2N}$ are all identified as erroneous by the frame error detection methods. In the subvector based EC, these erroneous vectors are now further submitted to a subvector consistency test which generates a consistency matrix C of dimensions $7 \times (2N+2)$ defined as follows:

$$c_{ij} = \begin{cases} 1, & j = 1 \, or \; j = 2N+2 \\ 0, & 2 \le j \le 2N+1 \text{ and } \mathrm{S}_{t-1}^{A+j-1} \text{ inconsistent from (2)} \\ 1, & 2 \le j \le 2N+1 \text{ and } \mathrm{S}_{t-1}^{A+j-1} \text{ consistent from (2)} \end{cases} \qquad (3)$$

On the basis of this consistency matrix, the EC is implemented in such a way that all inconsistent subvectors are replaced by their nearest neighbouring - in time - consistent subvectors whereas the consistent subvectors are kept unchanged.

# 3. WVD BASED ON SUBVECTOR EC

Subvector EC handles subvectors within erroneous vectors in two different ways. The first retains all consistent subvectors, and the second substitutes inconsistent subvectors with their nearest neighbouring consistent subvectors. This strategy beneficially exploits the remaining error-free information embedded in each erroneous vector, but it is noted that neither the retained consistent subvectors are necessarily correct (or reliable) nor the nearest neighbouring substitution generates the same features as their original. Therefore, these potentially unreliable features should not be given the same weight as error-free (reliable) features in the ASR decoder. This is realised by using WVD.

## 3.1 Weighted Viterbi Decoding

Vector based WVD modifies the observation probability of each feature vector in the Viterbi decoding by using the reliability of each vector as an exponential weighting factor [14]. The WVD uses the following formula to update the likelihood score accordingly

$$\delta_t(j) = \underset{i}{Max}\left[\delta_{t-1}(i) a_{ij}\right]\left[b_j(\boldsymbol{V}^t)\right]^{\gamma(t)} \qquad (4)$$

where $\delta_t(j)$ is the likelihood of the most likely state sequence at time *t* that ends in state *j* and has generated the observation (feature vectors) from $\boldsymbol{V}^1$ to $\boldsymbol{V}^t$, $a_{ij}$ is the transition probability from state *i* to state *j*, $b_j(\boldsymbol{V}^t)$ is the probability of emitting observation $\boldsymbol{V}^t$ when state *j* is entered. The weighting factor $\gamma(t)$ is a normalised reliability coefficient – of value between 0 and 1 – that adjusts the contribution of each vector to the overall likelihood score. Choosing the value of $\gamma(t)$ close to one causes the output probability for the particular vector to contribute almost fully to the likelihood score. In contrast, choosing a value of $\gamma(t)$ close to zero causes the output probability to be equal to one and identically contribute to all models, and therefore neutralises the vector contribution. A vector based WVD is applied in [11] where a time varying weighting factor is used to handle the fact that the longer a burst is the less effective is the vector repetition technique.

In combining WVD with the subvector EC, each feature is given its own weighting factor. Consider an observation vector

$\boldsymbol{V}^t = [v^t(1), v^t(2), ..., v^t(K)]^T$ where the component $v^t(k)$ is either one of the MFCC coefficients $c_k^t, k = 1, 2, ..., 12$ or $\log E^t$ for $k$=13. In this work $c_0$ is not used for recognition. The mapping between $v^t(k)$ and $\boldsymbol{S}_j^{\,t}$ is defined by Eq. (1). For example, $v^t(1) = c_1^t = S_0^t(0)$ and $v^t(2) = c_2^t = S_0^t(1)$. In assuming a diagonal covariance matrix, the overall observation probability is the product of the probabilities of emitting each individual feature. A feature based WVD thus computes the likelihood score as follows:

$$\delta_t(j) = \underset{i}{Max}\left[\delta_{t-1}(i)a_{ij}\right]\prod_{k=1}^{K}\left[b_j(v^t(k))\right]^{\gamma_k(t)} \tag{5}$$

where $b_j(v^t(k))$ is the observation probability of observing feature $v^t(k)$ when entering state $j$, and $\gamma_k(t)$ is the reliability measure for feature $v^t(k)$ as given in the next subsection.

### 3.2 Reliability measure

The reliability of each feature $v^t(k)$ is calculated during the subvector EC processing. When the two corresponding subvectors $\boldsymbol{S}_j^{\,t}$ and $\boldsymbol{S}_j^{\,t+1}$ in a frame-pair pass the consistency test as given in Eq. (2) the reliability of each feature in the subvectors is calculated on the basis of the difference $d(v^t(k), v^{t+1}(k))$ between two corresponding features; otherwise, the reliability is dependent on both the reliability of the substituting features and the temporal distance between the substituted features and the substituting features. Specifically, the weightings are assigned according to the following formula:

$$\gamma_k(t) = \begin{cases} \alpha^{d(v^t(k), v^{t+1}(k))/T_k}, & S_j^t \text{ consistent from (2)} \\ \gamma_k(t+p)\cdot\beta^{|p|}, & v^t(k) \text{ substituted by } v^{t+p}(k) \end{cases} \tag{6}$$

where $\alpha$ and $\beta$ are two constant parameters, $p$ is the temporal distance between the two features, and $T_k$ is the threshold for subvector consistency test as used in Eq. (2). For error-free vectors the weighting factors are equal to 1.

## 4. EXPERIMENTS

The combination of the feature based WVD and the subvector EC is evaluated on the basis of the Aurora 2 Test Set A database [12]. The database is the TI digit database artificially distorted by adding noise and using a simulated channel distortion. Whole-word models are created for all digits with the HTK recogniser [13]. Each of the digit models has 16 hidden Markov model (HMM) states with three Gaussian mixtures per state. The silence model has only three states with six HMM Gaussian mixtures per state. The one-state short pause model is tied to the second state of the silence model. Training on clean speech is used in the experiments. The test data are the clean data from Test Set A.

The characteristics of the transmission channel are given by the widely used GSM error pattern EP3 [1], [2]. GSM error patterns are commonly used for testing speech codecs and DSR error protection schemes due to the realistic error distributions including both random errors and burst-like errors. Table 1 shows the WER performance for three EC methods - consisting of the repetition (ETSI-DSR baseline) [4], the vector based WVD [11], and the subvector based EC (SEC) [6] – as compared to error-free transmission.

*Table 1*: %WER for the GSM EP3 for Aurora 2 Test Set A

| Method | ETSI-DSR | WVD | SEC | Error-free |
|---|---|---|---|---|
| WER | 6.70 | 4.78 | 2.65 | 0.95 |

### 4.1 The reliability measure parameters α and β

This subsection investigates the relationship between performance and the parameters α and β used in calculating the reliability. Experiments show that $\alpha = 0.45$ and $\beta = 0.4$ give the highest performance with a WER of approximately 2%. This is significantly better than the ETSI-DSR standard result that is 6.7%. The effects on WER of varying the values of the two parameters are presented in Table 2 and 3. The results show that a setting of the two parameters $\alpha$ and $\beta$ around their optimum values only has a minor influence on the resulting WER.

*Table 2*: %WER across different $\alpha$ settings with a fixed $\beta = 0.4$ for the GSM EP3 for Aurora 2 Test Set A

| $\alpha$ | 0.35 | 0.40 | 0.43 | **0.45** | 0.47 | 0.50 | 0.55 |
|---|---|---|---|---|---|---|---|
| WER | 2.08 | 2.07 | **2.05** | **2.05** | 2.06 | 2.06 | 2.10 |

*Table 3*: %WER across different $\beta$ settings with a fixed $\alpha = 0.45$ for the GSM EP3 for Aurora 2 Test Set A

| $\beta$ | 0.30 | 0.35 | 0.38 | **0.40** | 0.42 | 0.45 | 0.50 |
|---|---|---|---|---|---|---|---|
| WER | 2.08 | 2.08 | **2.05** | **2.05** | 2.07 | 2.09 | 2.11 |

### 4.2 The consistency test thresholds

The effect of the settings of the threshold values (i.e. $T_j(0)$ and $T_j(1)$ in Eq. (2) which are the same as $T_k$ in Eq. (6)) on performance is investigated in this subsection. It is noted that two sets of thresholds are applied in this work. The first set is used in the vector consistency test as an additional test to the CRC checking according to the ETSI-DSR standard, and the second set in the subvector consistency test. In the experiments conducted above, the two sets of thresholds are given the same values as provided by the ETSI-DSR standard.

In the experiments in this subsection, the first set of thresholds is kept as given in the ETSI-DSR standard, whereas the second set of thresholds is varied across a range. This is done by multiplying the ETSI-DSR standard values with a scaling factor $\lambda$ as given in Table 4, which shows the WER across the $\lambda$ range.

*Table 4*: %WER across different threshold settings for the GSM EP3 for Aurora 2 Test Set A

| $\lambda$ | -1.0 | 0.1 | 0.6 | 0.8 | **1.0** | 1.2 | 2.0 |
|---|---|---|---|---|---|---|---|
| WER | 4.73 | 2.85 | 2.09 | **2.01** | 2.05 | 2.18 | 2.81 |

The results show that for a value $\lambda = -1$ the performance is still better than the ETSI-DSR standard (6.7%), but close to the WER obtained by applying vector based WVD (4.78%) as shown in Table 1. The effect of using a negative threshold value is that all subvectors are detected as inconsistent and replaced by their nearest neighbouring error-free substitutions, thus equivalent to the ETSI-DSR vector repetition scheme. The performance improvement over the ETSI-DSR is caused by the WVD. The lowest WER is achieved for $\lambda = 0.8$. It is shown that for $\lambda = 0.1$ where only almost identical subvectors are classified as consistent, the proposed method gives better performance than the vector based WVD. A possible explanation for this improvement is that keeping the identical features unchanged may be better than using substitutions. With a choice of $\lambda = 2$ - where only a small number of subvectors are detected as inconsistent – an improvement as compared to the ETSI-DSR repetition is still observed. This may indicate that in the context of ASR, information including some errors is better than no information.

The results show that varying the scaling factor $\lambda$ around the default setting only has a minor influence on the resulting WER.

### 4.3 Overall performance comparison

On the basis of the above experiments on parameters settings, the combination of the feature based WVD and subvector EC (with a setting of $\alpha = 0.45$ $\beta = 0.4$ and $\lambda = 0.8$) is compared with the ETSI-DSR standard, the vector based WVD, and the subvector EC (SEC). The results are shown in Fig. 2.

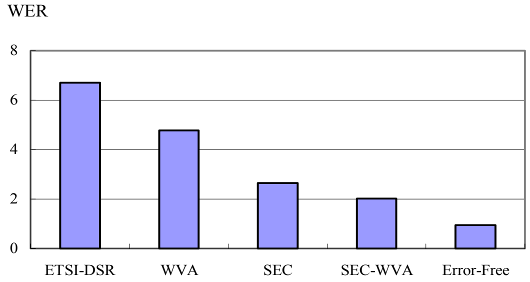

*Figure 2*: %WER across different methods for the GSM EP3 for Aurora 2 Test Set A.

The relative improvement over the ETSI-DSR standard, vector based WVD, and SEC are 70.0%, 58.0% and 24.2%, respectively.

## 5. CONCLUSIONS

The paper presents a WVD that conducts ASR decoding on the basis of the subvector based EC. During the process of subvector feature-reconstruction, a reliability measure for each feature is calculated and used for the WVD. Encouraging recognition performance has been obtained by combining the WVD and the subvector based EC. Due to the nature of server-based methods, there is no requirement of client-side modifications given full compatibility with the current ETSI-DSR standards. The method does not lead to an increase in computational cost or in bandwidth requirement.

## 6. ACKNOWLEDGEMENTS

This work has been supported by the FACE project, the CNTK project and the EU 6th framework project MAGNET. Furthermore, the authors would like to thank Haitian Xu for making available his WVD modification to HTK.

## 6. REFERENCES

[1] Tan, Z.-H., Dalsgaard, P. and Lindberg, B., "Automatic Speech Recognition over Error-Prone Wireless Networks," *Speech Communication*, 47(1-2), 220-242, Sept.- Oct., 2005.

[2] Pearce, D., “Robustness to transmission channel – the DSR approach”, ISCA ITRW Robustness2004, Norwich, UK, Aug. 2004.

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

[4] ETSI Standard ES 202 212. Distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithm, back-end speech reconstruction algorithm. Nov. 2003.

[5] Milner, B., “Robust speech recognition in burst-like packet loss”, ICASSP2001, USA, May 2001.

[6] Tan, Z.-H., Dalsgaard, P. and Lindberg, B., “A subvector-based error concealment algorithm for speech recognition over mobile networks,” ICASSP2004, Montreal, Quebec, Canada, May 2004.

[7] Yoma, N.B., McInnes, F.R., Jack, M.A., “Weighted Viterbi algorithm and state duration modelling for speech recognition in noise”, ICASSP98.

[8] Bernard, A., Alwan, A., “Low-bitrate distributed speech recognition for packet-based and wireless communication”, *IEEE Trans. Speech Audio Processing*, 10 (8), 570-579.

[9] Weerackody, V., Reichl, W., Potamianos, A., “An error-protected speech recognition system for wireless communications”, *IEEE Trans. Wireless Communications*, 1 (2), 282 – 291.

[10] Ion, V. and Haeb-Umbach, “A unified probabilistic approach to error concealment for distributed speech recognition”, Interspeech2005, Lisbon, Portugal, Sept. 2005.

[11] Cardenal-Lopez, A., Docio-Fernandez, L. and Garcia-Mateo, C., “Soft decoding strategies for distributed speech recognition over IP networks”, ICASSP2004, Montreal, Quebec, Canada, May 2004.

[12] Hirsch, H.G. and Pearce, D., “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions”, ISCA ITRW ASR00, Paris, France, Sept., 2000.

[13] Young, S. J. et al., “HTK: Hidden Markov Model Toolkit V3.2.1, Reference Manual”, Cambridge Univ. Speech Group, Mar. 2004.

[14] Yoma, N.B., McInnes, F.R., Jack, M.A., “Weighted Viterbi algorithm and state duration modelling for speech recognition in noise,” in Proc. ICASSP, May 1998.