A Unified Probabilistic Approach to Error Concealment for Distributed Speech Recognition

Valentin Ion, Reinhold Haeb-Umbach

University of Paderborn
Dept. of Communications Engineering
33098 Paderborn, Germany
{ion,haeb}@nt.uni-paderborn.de

Abstract

The transmission errors in a wireless or packet oriented network may dramatically decrease the performance of a distributed speech recognition (DSR) system. Error concealment has been shown to be an effective way to mantain an acceptable word error rate when dealing with error prone communication channels. In this paper we propose an extension of our previously introduced soft features approach for the case that the soft-output of the channel decoder is not available at the server side of the DSR system. We found a simple method to estimate bit reliability information which still gives good speech recognition results. It is shown that some other error concealment schemes turn out to be special cases of the method proposed here.

1. Introduction

The distributed speech recognition technique enables a mobile device, usually with limited processing and memory resources, to access sophisticated speech recognition services without the need to run or install complex speech recognition software. DSR works by splitting the process of speech recognition between the client (e.g. mobile device) and the network server. A so-called front-end runs on the client device, which extracts, compresses, protects against transmission errors and sends the speech features over a communication link to the back-end. In the back-end the features are block decoded, uncompressed and fed into the recognition engine. Note that additional channel coding, not part of DSR standard, is part of the communication link, e.g. convolutional or turbo coding in GSM/UMTS.

However, it has been shown that the accuracy of speech recognition is severely degraded in the case of a poor channel. Several error concealment strategies performing better than the original error mitigation scheme of ETSI ES 202 050 ([1]) have been proposed. Some introduce a small amount of redundancy in the feature bit stream which can be used at the back-end to better interpolate missing features [2]. In [3] the subvector structure of the compressed feature vector, where each subvector codes two parameters (i.e. feature vector components) was exploited.

In our previous work [4], which built upon the concept of softbit speech decoding introduced by Fingscheidt and Vary [5], we used the residual redundancy in the source bit stream and bit reliability information of the channel decoder to compute the a posteriori probability of the transmitted bit pattern. Further we employed uncertainty decoding and achieved a high degree of robustness of the speech recognition to transmission errors.

The assumption, that bit reliability information computed by the communication network channel decoder is available for reconstruction of corrupted features in the back-end, is arguable in practice. The channel decoder might not compute or at least not output this information, or, the transmission of the bit reliability information from the channel decoder to the DSR back-end requires bandwidth which one might not be willing to spend.

In this paper we present a method, how this bit reliability information can be estimated at the DSR back-end, and demonstrate that this results in only small loss of recognition accuracy. Further, the ETSI error mitigation scheme [1] and the scheme proposed in [3] are shown to be special cases of the scheme proposed here, resulting from introducing simplifying assumptions.

In the followings, we first review the soft feature speech recognition concept in Section 2, present our algorithm to estimate bit reliability in Section 3, Section 4 contains experimental results and we finish by drawing some conclusions.

2. Review of soft feature speech recognition

The feature vectors are computed on the client side and coded with a split vector quantizer resulting in the bit pattern to be sent over a distorted channel. Let x_n be an element of the multidimensional cepstral feature vector with n being a relative frame index, where $n = 0$ shall denote the present frame, $n = -1$ the previous frame, etc. In the following x_n is often called a "parameter". Two parameters are coded into a bit sequence of length M: $\mathbf{b}_n = (b_n(0), \ldots, b_n(M-1))$. Each bit combination $\mathbf{b}$ is assigned to a quantization table index i, $i \in \{0, 1, \ldots, 2^M - 1\}$ and we write for simplicity $\mathbf{b}_0 = \mathbf{b}_0^{(i)}$ to denote that $\mathbf{b}_0$ represents the i-th quantization table index.

At the server side (see Figure 1), the received bit sequence $\hat{\mathbf{b}}_0$ may not be identical to the sent sequence $\mathbf{b}_0$ due to transmission errors. The soft output channel decoder delivers the instantaneous bit error probability $\mathbf{p}_{e_0} = (p_{e_0}(0), \ldots, p_{e_0}(m), \ldots, p_{e_0}(M-1))$ for each bit of the decoded $\hat{\mathbf{b}}_0$. This is called bit reliability information and it is used to compute the conditional bit probability of a transmitted bit

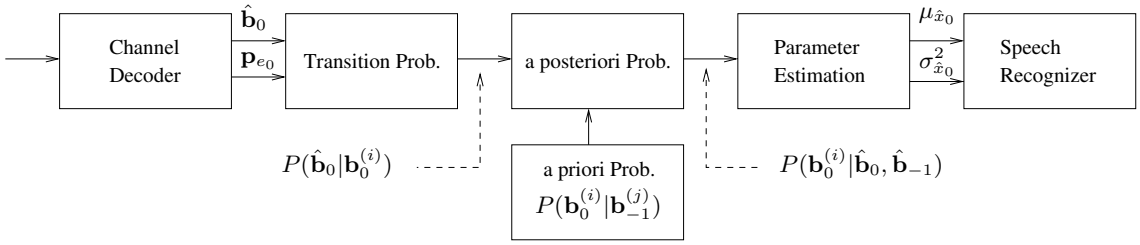

Figure 1: *Block diagram of soft feature distributed speech recognition system, server side.*

$b_0^{(i)}(m)$ to the known received bit $\hat{b}_0(m)$:

$$P(\hat{b}_0(m)|b_0^{(i)}(m)) = \begin{cases} 1 - p_{e_0}(m) & \text{if } \hat{b}_0(m) = b_0^{(i)}(m) \\ p_{e_0}(m) & \text{if } \hat{b}_0(m) \neq b_0^{(i)}(m) \end{cases} \quad (1)$$

Assuming a memoryless channel, the transition probability can be expressed as:

$$P(\hat{\mathbf{b}}_0|\mathbf{b}_0^{(i)}) = \prod_{m=0}^{M-1} P(\hat{b}_0(m)|b_0^{(i)}(m)). \quad (2)$$

The next knowledge source to be exploited is the residual redundancy in the bit stream of the source coder. We used *first-order* (AK1) a priori knowledge $P(\mathbf{b}_n|\mathbf{b}_{n-1})$, which captures the correlation between successive frames. It can be represented for each parameter as a matrix $P(\mathbf{b}_0^{(i)}|\mathbf{b}_{-1}^{(j)})$ with $i, j = 0, 1 \ldots 2^M - 1$. The matrix can be estimated once and for all on a training set. The a posteriori probabilities are computed from the transition probabilities (2) and the a priori probabilities. In the case of AK1 the a posteriori probabilities can be computed recursively [5]:

$$P(\mathbf{b}_0^{(i)}|\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_{-1}) = \frac{1}{C} \cdot P(\hat{\mathbf{b}}_0|\mathbf{b}_0^{(i)}) \cdot \sum_{j=0}^{2^M-1} P(\mathbf{b}_0^{(i)}|\mathbf{b}_{-1}^{(j)}) \cdot P(\mathbf{b}_{-1}^{(j)}|\hat{\mathbf{b}}_{-1}, \hat{\mathbf{b}}_{-2}) \quad (3)$$

where C is a normalizing constant.

Having the a posteriori probabilities, the MMSE estimate $\mu_{\hat{x}_0}$ [5], i.e. the first-order moment of the a posteriori probability mass function, can be computed. The second-order moment $\sigma^2_{\hat{x}_0}$ carries additional important information for uncertainty decoding [6].

3. Bit reliability estimation

Although the use of bit reliability information of the channel decoder delivers excellent speech recognition accuracy down to extremely bad channel conditions, it can be argued that, in a typical communication architecture, this information might not be available. A possible scenario is that the GSM base station (or an equivalent platform) performs the soft-output channel decoding and sends the decoded bits over another, e.g. wired, network to the DSR back-end. Transmission of bit reliability information would increase the required bandwidth and makes this approach unattractive.

There is therefore a need to estimate this bit reliability information $p_e(m)$ of the decoded bit $\hat{b}(m)$ at the recognition server site. It has been experimentally observed that the recognition accuracy is not very sensitive to estimation errors of p_e. The idea here is to estimate a single value of p_e over an interval, e.g. a frame or subvector, as the ratio of the number of (assumed) bit errors $\hat{N}_e$ to the total number of bits N within the interval:

$$\hat{p}_e(m) = \frac{\hat{N}_e}{N} \quad (4)$$

An estimate $\hat{N}_e$ for the number of bit errors can be obtained from the consistency test used in the ETSI front-end for distributed speech recognition [1]. Basically this data consistency test checks the continuity of a parameter in the frame pair. When the difference between two consecutive values of a parameter exceeds a predetermined threshold it will be decided that the frame pair is not reliable. Certainly not every bit error can be revealed by this test, only those bit errors that lead to exceeding the thresholds are detected, so it is very likely to under estimate $\hat{p}_e$. Nevertheless, the experiments have shown that better results are obtained, compared to the case when only the CRC (cyclic redundancy codes) test is employed. This is not surprising since the CRC can detect one or more errors with a high level of confidence but unfortunately gives no information about the number of errors.

Based on the data consistency test, we tried out two procedures in order to obtain an estimate of p_e. In the first one, referred as FR (frame oriented), the errors were counted over the whole frame-pair (86 bits) by applying the consistency test to each parameter. For each consistency test failure the error counter is increased by one. Note that the same value of $\hat{p}_e$ is shared by all 86 bits of the frame-pair.

In the second procedure, labeled SV (subvector based), the consistency test is triggered by a CRC failure, and the errors are counted on a subvector basis. Note that a subvector encodes two parameters of the cepstral feature vector. That is the $\hat{p}_e$ for all bits of the subvector is estimated as the number of bit errors occurring in that subvector divided by the number of bits allocated to that subvector over the frame-pair, which is $2M$. We assumed that a consistency test failure reveals one bit error.

Interestingly, other known error mitigation methods are obtained as special case of this approach. If we set $\hat{p}_e = \frac{1}{2}$ in the case of FR and if the matrix of first-order a priori knowledge $P(\mathbf{b}_0^{(i)}|\mathbf{b}_{-1}^{(j)})$ is set to an identity matrix, which essentially means that bits at the same position of the cepstral parameter vector in two successive frames are fully corelated, we obtain the error mitigation scheme described in the ETSI standard, up to the following difference: here, these settings result in a repetition of the last good frame for the whole duration of an error

burst, while, according to the standard, missing frames are repeated from both ends of the burst towards the middle. Our observation, however, is that the interpolation from both ends results in only minor improvement of the word accuracy of a subsequent speech recognizer (see results in Section 4.2: comparison of curves labeled ETSI-MIT and FR-DIAG in Figures 2 and 3).

Using the same settings in the case of SV, the subvector based error mitigation algorithm presented in [3] is obtained, up to the same difference in the interpolation as mentioned above.

The scheme proposed here, however, is more flexible: First it uses a more accurate estimate for p_e and not the binary values $p_e = 0$ (frame/subvector correct) or $p_e = \frac{1}{2}$ (frame/subvector corrupt). Second, the matrix of correlations among bits in successive frames, which can be estimated on clean speech training data, is utilized for a more sophisticated reconstruction of bits deemed corrupt than a pure frame repetition can do.

4. Performance evaluation

4.1. Experimental setup

Two channel models approximating COST 207 profiles have been used to simulate various conditions, from very low to medium link quality of a GSM data transmission at 4.8kbit/s full rate and half rate. The first models a "rural area" with 6 distinct propagation paths, delay spread 0.097 μs, Rician and Rayleigh fading whereas the second models a "typical urban" area characterized by 12 propagation paths, delay spread 1.03 μs and Rayleigh fading. The terminal was assumed to be moving at 50 km/h. The simulation of the GSM physical layer included (de)interleaving, (de)modulation and channel decoding with the Bahl algorithm. However, we assumed perfect clock and carrier phase synchronization.

Speech recognition experiments were conducted on the clean test set of AURORA 2 database (4004 utterances from 52 male and 52 female distributed over 4 subsets) employing the front-end for feature extraction standardized by ETSI [1] at various carrier-to-interference ratio on both channels. The ETSI front-end computes a 14-dimension feature vector per speech frame which consists of 13 Mel Frequency Cepstrum Coefficients $(c_0, ..., c_{12})$ and log-Energy $(logE)$. The vector components are grouped into pairs, each pair being quantized utilizing a split vector quantizer (SVQ), for each resulting a codebook index. These indexes are represented by $M = 6$ bits for (c_1, c_2), (c_3, c_4), ..., (c_9, c_{10}), $M = 5$ bits for (c_{11}, c_{12}) and $M = 8$ bits for $(c_0, logE)$, forming altogether a 44 bits sequence (one bit allocated for VAD). Two such sequences are protected by a 4-bit CRC and grouped into a 92-bit data frame which has to be transmitted to the remote server.

At the server side, the ETSI standard provides an error mitigation scheme (ETSI-MIT) summarized as follows: For each data frame the CRC is checked, signaling a good or bad frame. Because of pairwise grouping the bad feature-frames form always error bursts of length 2B. Once a burst is detected, the first B frames are replaced by a copy of the last good frame before the burst and the last B ones by a copy of first good frame after the burst.

To validate the expression for estimating p_e, we conducted reference simulations where N_e was perfectly known by calculating the Hamming distance between frames or subvectors at transmitter and receiver. This shall give the upper limit of the performance that can be attained using any estimation method.

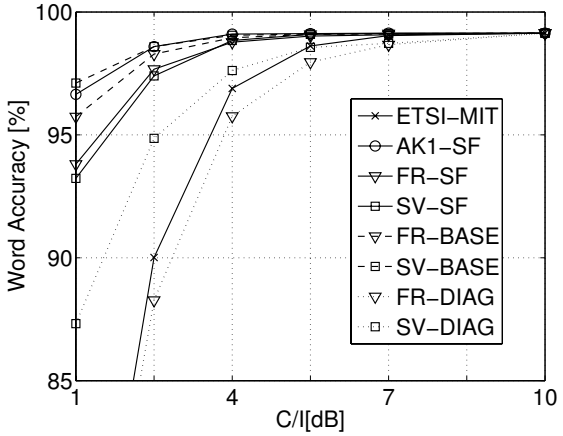

Figure 2: *Word accuracy for transmission over GSM TCH/F4.8 data link, "rural area" channel model.*

Table 1: *Bit error rates (BER) for "rural area"*

C/I [dB]	1	2.5	4	5.5	7	10
BER [%]	8.4	3.7	1.4	0.4	0.1	0.004

For those error mitigation schemes for which the variance of the a posteriori probability of the parameter can be computed, the speech recognizer was extended to perform uncertainty decoding: the variance of each Gaussian of the observation probabilities was increased by an amount equal to $\sigma^2_{\hat{x}_0}$ [4].

The following error mitigation schemes were evaluated for both described GSM scenarios:

- The ETSI error mitigation (labeled ETSI-MIT).
- Soft feature DSR scheme using the soft-output of channel decoder computed by means of the MAP-decoder after Bahl et al. [7], *first-order* a priori knowledge and uncertainty decoding (labeled AK1-SF) [4].
- The proposed error concealment, with p_e estimated over the whole frame and uncertainty decoding in the speech recognizer (labeled FR-SF).
- The proposed error concealment, with p_e estimated on a subvector basis and uncertainty decoding in the speech recognizer (labeled SV-SF).
- The reference simulations with N_e and therefore p_e assumed to be perfectly known for each frame or subvector (labeled FR-BASE and SV-BASE respectively).
- The simulations assuming $p_e = \frac{1}{2}$ and the identity matrix as a priori knowledge for both frame and subvector based approaches (labeled FR-DIAG and SV-DIAG respectively).

4.2. Experimental results

Figures 2 and 3 show the achieved word accuracies as a function of carrier-to-interference ratio. Note the different scaling

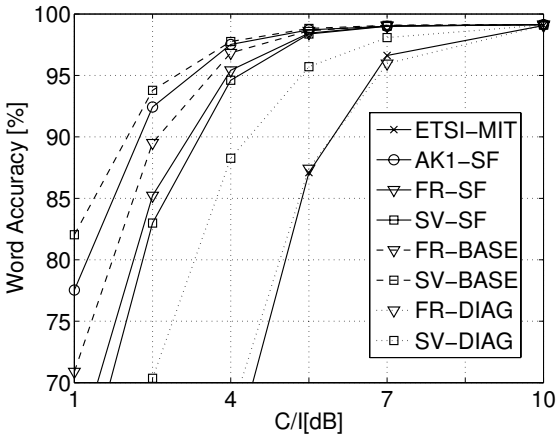

Figure 3: *Word accuracy for transmission over GSM TCH/H4.8 data link, "typical urban" channel model.*

Table 2: *Bit error rates (BER) for "typical urban"*

C/I [dB]	1	2.5	4	5.5	7	10
BER [%]	23	15	8.1	3.6	1.3	0.1

in Figures 2 and 3! For better comparison with simulations employing error patterns rather than a full link level simulation as is done here, Tables 1 and 2 list the measured bit error rate (BER) of the decoded bit stream at each C/I value.

As expected, the results obtained using bit reliability information computed in the channel decoder are better than those based on estimating p_e.

The curves FR-BASE and SV-BASE reveal the optimally achievable performance if the number of bit errors per frame or subvector is perfectly known. Interestingly, FR-BASE turned out to be worse than soft feature DSR. While the bit error rate per frame is perfectly known in FR-BASE, a single fixed value of p_e is used for all 86 bits of a frame, whereas soft feature DSR uses different values $p_e(m)$ for every bit m of a frame. The smaller the interval for which p_e is fixed the better the performance. Therefore SV-BASE is superior to FR-BASE and turns out to be also slightly superior to soft feature DSR.

If the number of bit errors in an interval is not perfectly known and has to be estimated, a shorter estimation interval leads to less reliable estimates. Therefore SV-SF turns out to be slightly worse than FR-SF. Note that both lead to reduced word accuracy compared to soft feature DSR. Still, however, they conceal errors significantly better than both, the ETSI error mitigation scheme and the subvector-based scheme proposed in [3].

5. Conclusions

In this paper we proposed a method to estimate the bit reliability information for soft feature DSR and showed that it delivers competitive results. It was also shown that the ETSI error mitigation scheme and the related subvector-based scheme are simplified versions of our approach.

While the above considerations assumed circuit-switched GSM data communication, the proposed method can also be adapted to packet oriented transmission as is used in IP networks. Here, typically a complete packet is lost. If a packet transports a frame, then packet loss is equivalent to $\hat{p}_e = \frac{1}{2}$ in the FR scheme and the reconstruction of the lost frame can be done exactly as described here.

The proposed method can therefore serve as an unified framework for error mitigation in DSR, dealing with both bit errors and packet loss, be it in a wireless or wired network.

Further work on quantifying the effect of lost and reconstructed packets on the word accuracy of the speech recognition is under way.

6. Acknowledgement

This work was supported by Deutsche Forschungsgemeinschaft under contract number HA 3455/2-1.

7. References

[1] ETSI ES 202 050 v1.1.1 "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms", Nov. 2003

[2] Antonio M. Peinado, Angel M. Gomez, Victoria Sanchez, Jose L. Perez-Cordoba, Antonio J. Rubio, "Packet loss concealment based on VQ replicas and MMSE estimation applied to distributed speech recognition", in *Proc. ICASSP 2005*, Philadelphia, March 2005.

[3] Z. Tan, P. Dalsgaard, and Borge Lindberg, "A subvector-based error concealment algorithm for speech recognition over mobile networks", in *Proc. ICASSP 2004*, Montreal, CA., May 2004.

[4] R. Haeb-Umbach, V. Ion, "Soft features for improved distributed speech recognition over wireless networks", in *Proc. ICSLP 2004*, Jeju, Korea, Oct. 2004.

[5] T. Fingscheidt, P. Vary, "Softbit speech decoding: A new approach to error concealment", *IEEE Trans. on Speech an Audio Processing*, vol. 9, no. 3, pp. 1-11, 2001.

[6] L. Deng, J. Droppo, and A. Acero, "Exploiting variances in robust feature extraction based on a parametric model of speech distortion", in *Proc. ICSLP 2002*, Denver, Co., Sep. 2002.

[7] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate", *IEEE Trans. on Information Theory*, vol. 20, pp. 284-287, March 1974.

[8] J. Hagenauer, "Source-controlled channel decoding", in *IEEE Transactions on Communications*, vol. 43, no. 9, pp. 2449–2457, 1995.

[9] J. Hagenauer, P. Höher, "A viterbi algorithm with soft-decision outputs and its applications", in *Proc. IEEE Global Communications Conference*, Dallas, Tx, Nov. 1989.

[10] H.G. Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions", in *ISCA ITRW Workshop ASR2000*, Paris, France, Sep. 2000.