Soft Features for Improved Distributed Speech Recognition over Wireless Networks

Reinhold Haeb-Umbach, Valentin Ion

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

Abstract

A major drawback of distributed versus terminal-based speech recognition is the fact that transmission errors can lead to degraded recognition performance. In this paper we employ "soft features" to mitigate the effect of bit errors on wireless transmission links: At the receiver a posteriori probabilities of the transmitted feature vectors are computed by combining bit reliability information provided by the channel decoder and a priori knowledge about residual redundancy in the feature vectors. While the first-order moment of the a posteriori probability function is the MMSE estimate, the second-order moment is a measure of the uncertainty in the reconstructed features. We conducted realistic simulations of GSM transmission and achieved significant improvements in word accuracy compared to the error mitigation strategy described in the ETSI standard.

1. Introduction

In distributed speech recognition (DSR) the terminal extracts and codes the feature vectors (acoustic front-end), the transmission is performed over a data channel and the recognition engine resides on the server side (back-end). ETSI has standardized this approach in ETSI ES 201 108 and ETSI ES 202 050. Performing speech recognition in the network as opposed to the terminal has the advantage that the computational burden is shifted from the mobile terminal to a high performance platform in the network. Further, maintenance and provision of application dependent data is much simpler in the network than in a terminal.

Compared to completely terminal-based solutions, a drawback of DSR is, however, that transmission errors may degrade the performance of the recognizer [1]. The standard includes a basic error mitigation algorithm (CRC, consistency check, frame repetition), that has been shown to be effective on bursty channels if the channel is not too bad.

Better algorithms can be derived from joint source-channel coding/decoding, a powerful tool well-known in the mobile communications literature, see e.g. [2]. Weerackody et al. [3] proposed unequal error protection and a soft feature error concealment strategy. However, they assumed an idealized setup with perfectly known bit reliability. Further, speech parameter values (by *parameter* we mean a component of the speech feature vector) were considered either correct or false with no intemediate stages. Bernard and Alwan devised a coding scheme which allows error detection capabilities with soft-decision decoding, and the Viterbi decoder in the recognizer was modified to deal with unreliable features [1].

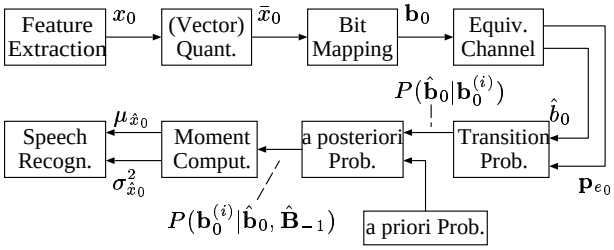

Figure 1: *Block diagram of soft feature distributed speech recognition system.*

Peinado et al. [4] applied the concept of softbit speech decoding introduced by Fingscheidt and Vary [5] for DSR and achieved good recognition performance for AWGN and bursty channels. They assumed the bit reliability information to be known, e.g. from a given or assumed SNR-value. We build upon the same concept here, however, we employ realistic bit reliability information computed in the channel decoder. We carried out simulations of a transmission over a complete GSM link, including channel (de)coding, (de)interleaving, GMSK (de)modulation and employing realistic channel models including multipath propagation, fast fading, and cochannel interference. Fig. 1 illustrates the processing stages in the proposed channel-noise robust DSR system. While for speech decoding one is only interested in the MMSE estimates of the parameters [5], the second-order moment of the a posteriori probability carries additional important information in the case of a subsequent speech recognizer: It is a measure of the confidence about the reconstructed features and can be utilized in uncertainty decoding [6].

2. Concept of Soft Feature Speech Recognition

2.1. Source Coding

The feature extraction delivers speech feature vectors, which are to be transmitted over a distorted channel. A component of the feature vector is called *parameter* in the following. Let

Table 1: *Code word lengths and entropies for the parameters of the ETSI advanced DSR front-end.*

Param.	c_1, c_2	c_3, c_4	c_5, c_6	c_7, c_8	c_9, c_{10}	c_{11}, c_{12}	$c_0, \log E$
M	6	6	6	6	6	5	8
$H(\mathbf{b}_n)$	5.84	5.80	5.79	5.76	5.80	4.78	7.72
$H(\mathbf{b}_n\|\mathbf{b}_{n-1})$	3.12	3.68	4.14	4.39	4.52	3.74	4.07

$x_n \in \mathbb{R}$ be such a parameter which is coded by M bits prior to transmission. n is a relative frame index, where $n = 0$ shall denote the present frame, $n = -1$ the previous frame, etc. The (vector) quantized parameter $Q(x_0) = \bar{x}_0$ with $\bar{x}_0 \in$ QT (QT: quantization table) is represented by the bit combination $\mathbf{b}_0 = (b_0(0), \ldots, b_0(m), \ldots, b_0(M-1))$. The bits are assumed to be bipolar: $b(m) \in \{-1, 1\}$. 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.

The transmission of the bit combination is described by an equivalent channel model with input $\mathbf{b}_0$ and output $\hat{\mathbf{b}}_0$, which may comprise the channel model itself, the channel encoder/decoder, modulation/demodulation and equalization. Due to transmission errors $\mathbf{b}_0$ and $\hat{\mathbf{b}}_0$ may not be identical. In a conventional decoding scheme an inverse bit mapping scheme is applied to $\hat{\mathbf{b}}_0$ to obtain the decoded parameter $\hat{x}_0$, which is then fed into the back-end recognition engine.

2.2. Channel Decoding with Soft Output

For soft feature speech recognition, a channel decoder is required which outputs the detected bit sequence $\hat{\mathbf{b}}_0$, and in addition reliability information in terms of estimated bit error probabilities: $\mathbf{p}_{e_0} = (p_{e_0}(0), \ldots, p_{e_0}(m), \ldots, p_{e_0}(M-1))$. The combination of the received bit and its reliability has been termed *softbit* in the literature [5]. We employed the MAP-decoder after Bahl et al.[1] [7], however, the less complex soft-output Viterbi algorithm (SOVA) can do as well [8]. The soft output consists of the so-called log-likelihood ratio

$$L_0(m) = \log \frac{P(b_0(m) = +1|\mathbf{q})}{P(b_0(m) = -1|\mathbf{q})} , \tag{1}$$

with $\mathbf{q}$ being the (real or complex valued) input sequence to the channel decoder. The instantaneous bit error probability of the decoded hard bit $\hat{b}_0(m) = \mathrm{sign}\{L_0(m)\}$ is given by

$$p_{e_0}(m) = \frac{1}{1 + \exp|L_0(m)|} . \tag{2}$$

The conditional bit probability of a transmitted bit $b_0^{(i)}(m)$ to the known received bit $\hat{b}_0(m)$ is then given by:

$$P(\hat{b}_0(m)|b_0^{(i)}(m)) = \begin{cases} 1 - p_{e_0}(m) & \text{if } \hat{b}_0(m) = b_0^{(i)}(m) \\ p_{e_0}(m) & \text{if } \hat{b}_0(m) \neq b_0^{(i)}(m) \end{cases} \tag{3}$$

[1] The algorithm is known in the speech recognition community under the name *Forward-Backward Algorithm*.

If we consider the equivalent channel to be memoryless, the *parameter* transition probability (which could also be termed "observation probability" [4]) becomes

$$P(\hat{\mathbf{b}}_0|\mathbf{b}_0^{(i)}) = \prod_{m=0}^{M-1} P(\hat{b}_0(m)|b_0^{(i)}(m)). \tag{4}$$

2.3. A Priori Probabilities

The next knowledge source to be exploited is the residual redundancy in the bit stream of the source coder. In specifying this a priori knowledge, models of different complexity may be chosen. In this paper we have chosen the two most simple: *zeroth-order* a priori knowledge "AK0", $P(\mathbf{b}_n)$, and *first-order* a priori knowledge "AK1" $P(\mathbf{b}_n|\mathbf{b}_{n-1})$.

Unlike speech coding, the goal of feature extraction for speech recognition is not in the first place to obtain a very low bit rate by removing all redundancy in the speech signal. Therefore it is likely that there is considerable redundancy left in the coded feature vectors, which can be exploited at the decoding side. The redundancy is defined as the difference between the average code word length, here M, and the entropy $H(\mathbf{b}_n)$, in case of AK0, and the conditional entropy $H(\mathbf{b}_n|\mathbf{b}_{n-1})$ in case of AK1, respectively.

Table 1 shows the entropies for the parameters of the ETSI advanced DSR front-end, measured on the clean training data set of Aurora 2 [9].

For each column in Table 1, 2^M probabilities in case of AK0, and $2^M \times 2^M$ in case of AK1, have to be estimated in advance from the training data and must be available to the soft feature decoder. In the case of AK1 this amounts to $5 \cdot 2^{12} + 2^{10} + 2^{16} = 87,040$ values.

2.4. Parameter Estimation

Finally the a posteriori probabilities can be computed from the transition probabilities (4) and the a priori probabilities. In the case of 0-th order a priori knowledge (AK0) we have

$$P(\mathbf{b}_0^{(i)}|\hat{\mathbf{b}}_0) = \frac{P(\hat{\mathbf{b}}_0|\mathbf{b}_0^{(i)})P(\mathbf{b}_0^{(i)})}{\sum_{l=0}^{2^M-1} P(\hat{\mathbf{b}}_0|\mathbf{b}_0^{(l)})P(\mathbf{b}_0^{(l)})} \tag{5}$$

In the case of first-order a priori knowledge (AK1) the a posteriori probabilities can be computed recursively [5]: Let $\hat{\mathbf{B}}_{-1} = \hat{\mathbf{b}}_{-1}, \hat{\mathbf{b}}_{-2}, \cdots$ be a short-hand notation for the complete history of the received bit combinations until the time instant $n = -1$.

Then we obtain

$$P(\mathbf{b}_0^{(i)}|\hat{\mathbf{b}}_0,\hat{\mathbf{B}}_{-1}) = \frac{1}{C}\cdot P(\hat{\mathbf{b}}_0|\mathbf{b}_0^{(i)})\cdot \sum_{j=0}^{2^M-1} P(\mathbf{b}_0^{(i)}|\mathbf{b}_{-1}^{(j)})\cdot P(\mathbf{b}_{-1}^{(j)}|\hat{\mathbf{b}}_{-1},\hat{\mathbf{B}}_{-2}) \quad (6)$$

where the normalizing constant is given by

$$C = \sum_{l=0}^{2^M-1} P(\hat{\mathbf{b}}_0|\mathbf{b}_0^{(l)})\cdot \sum_{j=0}^{2^M-1} P(\mathbf{b}_0^{(l)}|\mathbf{b}_{-1}^{(j)})\cdot P(\mathbf{b}_{-1}^{(j)}|\hat{\mathbf{b}}_{-1},\hat{\mathbf{B}}_{-2}). \quad (7)$$

Note, that the term $P(\mathbf{b}_{-1}^{(j)}|\hat{\mathbf{b}}_{-1},\hat{\mathbf{B}}_{-2})$ is nothing else than the resulting a posteriori probability of the previous time frame.

Now that the a posteriori probabilities are known, estimates of various kinds can be computed. For real-valued parameters, a good choice is the minimum mean squared error (MMSE) estimate, i.e. the conditional mean of x_0 given $\{\hat{\mathbf{b}}_0,\hat{\mathbf{B}}_{-1}\}$, in case of AK1. Since there is a one-to-one mapping from x_0 to $\mathbf{b}_0$, we have

$$\mu_{\hat{x}_0} = E[x_0|\hat{\mathbf{b}}_0,\hat{\mathbf{B}}_{-1}] = \sum_{i=0}^{2^M-1} x_0^{(i)}\cdot P(\mathbf{b}_0^{(i)}|\hat{\mathbf{b}}_0,\hat{\mathbf{B}}_{-1}). \quad (8)$$

For binary parameters (e.g. the VAD flag) the MAP-estimate is more appropriate since it guarantees that the estimate is again binary valued.

Likewise, the variance can be computed

$$\sigma^2_{\hat{x}_0} = E[x_0^2|\hat{\mathbf{b}}_0,\hat{\mathbf{B}}_{-1}] - \mu^2_{\hat{x}_0}. \quad (9)$$

It is reasonable to use the Gaussian assumption to characterize the uncertainty in the speech features:

$$p(\hat{x}_0) = \mathcal{N}(\hat{x}_0;\mu_{\hat{x}_0},\sigma^2_{\hat{x}_0}). \quad (10)$$

Instead of the "plug-in" classification rule, now an "optimal rule" can be used in the speech recognizer, which takes into account the uncertainty about the reconstructed feature vectors [6]:

$$\hat{W} = \underset{W}{\operatorname{argmax}} \left\{ \int p(\hat{X}|W)p(\hat{X})d\hat{X}\cdot P(W)\right\}, \quad (11)$$

where $P(W)$ is the prior probability of the word sequence, and $\hat{X}$ denotes the sequence of reconstructed feature vectors. We approximated the integration by increasing the variance of each Gaussian in the HMMs by the amount equal to $\sigma^2_{\hat{x}_0}$.

3. Experimental Results

We conducted experiments on the AURORA 2 database employing the advanced front-end feature extraction algorithm standardized by ETSI. First, we simulated uncoded binary transmission over an additive white Gaussian noise (AWGN) channel. This served as a kind of reference setup to validate the approach. We conducted experiments both for clean and (microphone) noisy test data. Since we observed the same trends both for clean test data and for test data corrupted by microphone noise, we are going to present in the following only the results obtained with clean test data.

In the figures to follow we show word accuracies for the following scenarios:

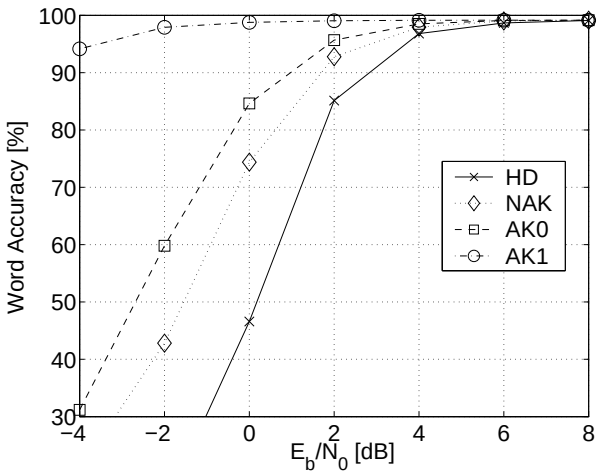

Figure 2: *Word accuracy for uncoded transmission over AWGN-channel. Clean test data.*

- A baseline setup without use of bit reliability information or a priori knowledge. For the GSM simulations we used the ETSI error mitigation strategy (curve labeled "ETSI-MIT"). For AWGN no error mitigation was employed, since the ETSI-technique resulted in degraded performance [4] (curve labeled "HD": *hard decision*).
- Use of bit reliability information but assuming uniform a priori probablities (curves labeled "NAK": *no apriori knowledge*).
- Use of bit reliability information and a priori knowledge: zero-order, eq. (5) ("AK0"), first-order, eq. (6) ("AK1"), and AK1 exploiting the uncertainity in speech recognizer, eq. (10) ("AK1-VAR"), respectively.

Fig. 2 shows the average word accuracies achieved on test sets A,B and C as a function of the bit energy to noise spectral density E_b/N_0 on an AWGN channel. For these experiments we assumed E_b/N_0 to be known at the receiver. The significant improvements obtained are similar with the results in [4].

Next we employed the GSM library of the SPW ("Signal Processing Worksystem") software suite to simulate a complete transmission over a GSM data channel. What was denoted as "Equivalent Channel" in Fig. 1 comprised the specified GSM data links TCH/F4.8 (full-rate) or TCH/H4.8 (half-rate) offering 4.8 kbit/s data rate in compliance with the ETSI standard:

- Channel coding with a rate $r = 1/3$ (Fig. 3) and $r = 1/2$ (Fig. 4) convolutional code
- Interleaving at the transmitter and deinterleaving at the receiver side
- Channel models approximating the COST 207 profiles: A "rural area" channel, modelled by 6 distinct propagation paths (delay spread 0.097 μs) with Rician and Rayleigh fading (Fig. 3), and a "typical urban" channel, modelled by 12 propagation paths (delay spread 1.03 μs) and Rayleigh fading only (Fig. 4). The mobile terminal was assumed to be moving at 50 km/h. Further, cochannel interference was simulated at various C/I (carrier-to-interference) ratios.

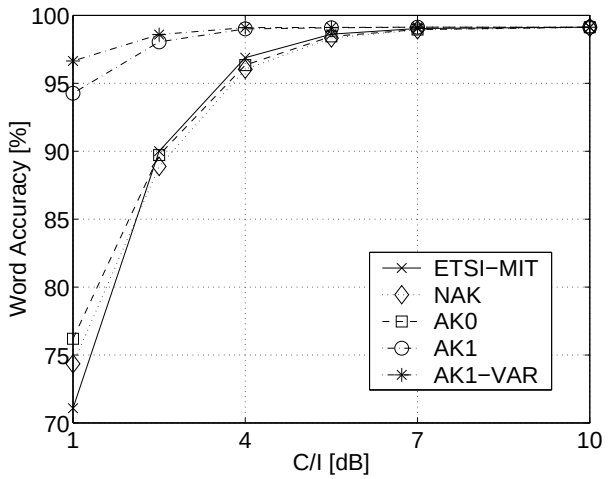

Figure 3: *Word accuracy for transmission over GSM TCH/F4.8 data link, "rural area" channel model.*

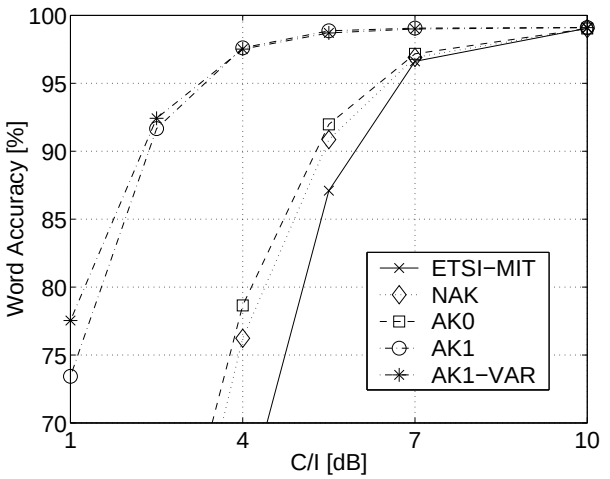

Figure 4: *Word accuracy for transmission over GSM TCH/H4.8 data link, "typical urban" channel model.*

- Channel decoding with the Bahl algorithm which delivers the L-values according to eq. (1). This decoder is by far the computationally most intensive part of the receiver. However, note that such a decoder will likely be used in future UMTS terminals for the decoding of turbo coded data.

Figs. 3 and 4 show word accuracies as a function of the C/I ratio for the rural and the urban channel models, respectively. In the case of the rural channel model and full-rate mode ($r = 1/3$) the word accuracies are hardly degraded for C/I-values as low as 4 dB in the case of soft feature speech recognition with AK1 a priori probabilities. Uncertainty decoding ("AK1-VAR") yields another noticeable improvement. Since the typical urban channel model exhibits more severe multipath and since the half-rate mode ($r = 1/2$) was employed, the results are worse in this scenario. Still, the improvement by AK1 is significant. For the results with low C/I-values it should be noted that in practice performance at C/I-values close to 0 dB will be degraded due to synchronization problems of the receiver, which are not reflected in our simulations.

4. Conclusions

We presented a soft feature speech recognition back-end for distributed speech recognition, which results from a consequent use of Bayesian estimation. Simulations of a realistic GSM transmission showed significant improvements over a conventional scheme which employs CRC and frame repetition for error concealment. The decoder also delivers feature vector uncertainty information which is used in the speech recognizer.

5. Acknowledgements

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

6. References

[1] A. Bernard and A. Alwan, "Low-bitrate distributed speech recognition for packet-based and wireless communications," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 570–579, 2002.

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

[3] V. Weerackody, W. Reichl, and A. Potamianos, "An error protected speech recognition system for wireless communications," *IEEE Trans. on Wireless Communications*, vol. 1, pp. 282–291, 2002.

[4] A. Peinado, J. Sánchez, V. Pérez-Córdoba, and A. de la Torre, "Hmm-based channel error mitigation and its application to distributed speech recognition," *Speech Communication*, vol. 41, pp. 549–561, 2003.

[5] T. Fingscheidt and P. Vary, "Softbit speech decoding: A new approach to error concealment," *IEEE Transactions on Speech and 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*, Denver, Co, Sept. 2002, pp. 2449–2452.

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

[8] J. Hagenauer and P. Höher, "A viterbi algorithm with soft-decision outputs and its applications," in *Proc. IEEE Global Communications Conference*, Dallas, Tx, Nov. 1989, pp. 2505–2509.

[9] H. Hirsch and 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.