Comparison of Decoder-based Transmission Error Compensation Techniques for Distributed Speech Recognition

Valentin Ion, Reinhold Haeb-Umbach

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

Abstract

In this study we evaluate transmission error compensation techniques for distributed speech recognition systems based on modification of the speech decoder. The candidates are marginalization, weighted Viterbi and our recently proposed soft-feature uncertainty decoding. For the latter, it is shown how the Bayesian speech recognition approach must be reformulated for recognition at the server side. The resulting predictive classifier is able to take account of the transmission errors by changing the contribution of the affected speech features to the acoustic score. The comparison of the experimental results has proven the superiority of our approach.

1. Introduction

Distributed speech recognition (DSR) is a client-server implementation of automatic speech recognition where the client, often a mobile device, extracts the feature vectors from the acoustic signal and sends them to the remote server which performs the speech recognition. Due to specific delay and bandwidth constraints, the link is likely to loose information, be it by altering bits or by actually dropping packets. The performance of the subsequent recognition degrades when such losses occur, expecially if they appear in bursts.

Several strategies have been proposed to conceal the transmission errors. Some of them, known as point estimate methods, attempt to correct or reconstruct the erroneous data using either forward error correction or the redundancy of the data [1, 2]. More sophisticated strategies assign a reliability measure to each feature vector and utilize this measure in a modified speech recognizer by changing the contribution of the associated feature to the acoustic score [3]. The reliability is inferred from the instantaneous bit error probability, delivered by the modern soft-bit decoding techniques like Turbo or soft-output Viterbi decoding and some a priori knowledge about the speech signal [4].

In the next section of this paper we describe three decoder-based error compensation methods: marginalization, weighted Viterbi (WV) and uncertainty decoding (UD). More accent is on the last one, also referred to by soft-feature decoding [4]. We compare these methods in terms of recognition accuracy obtained by performing AURORA 2 and Wall Street Journal (WSJ0) recognition tasks. The experimental setup is described in Section 3 and the word error rates obtained are given in Section 4. Finally some conclusions are drawn in Section 5.

2. Decoder-based compensation

To complement the point estimate methods which plug in into the speech decoder an estimation of the transmitted feature vector, the decoder-based methods associate a measure of

reliability to that estimation and modify the classification rule so that the reliability is considered.

2.1. Marginalization

Marginalization reformulates the classification to perform recognition based on the reliable features alone [5]. In the case of a packet oriented transmission there is a straight forward association of packet loss with unreliable data. However, on circuit switched channels bit errors might occur, which makes difficult to decide whether a feature is reliable or not, even if the error probability of each bit is available. In [3], a threshold of error probability for the feature was experimentally determined and that feature was marginalized when the threshold was exceeded.

In a Hidden Markov Model (HMM) speech decoder marginalization is performed by setting the state observation probability of the unreliable feature to one. In this way that feature does not change the acoustic score.

2.2. Weighted Viterbi

The one-bit reliability measure computed for marginalization can be extended to a continuous confidence measure γ, taking values between 0 and 1. The WV-decoding takes into account the confidence in the feature vector by raising the probability of observing the feature in the given state to the power of γ. In the logarithmic domain the operation becomes a weighting.

Obviously, for the successfully received feature vectors $\gamma = 1$ and no changes of observation probability occur. For a lost feature vector, the maximum uncertainty is expressed by setting $\gamma = 0$, and thus the observation probability becomes one, independent of the state. With binary weighting, the WV is equivalent to marginalization. Although a straight forward interpretation of the binary weighting is given, generally, raising the observation probability to some value γ lacks a probabilistic interpretation. Moreover, the methods proposed in [3, 6, 7] to determine the weighting coefficient are rather empirical.

2.3. Soft-feature uncertainty decoding

In this section we summarize the mathematical support of our approach for speech recognition at the server side. Let $\mathbf{X}$ be the sequence of feature vectors generated by the DSR frontend. This sequence is sent in digital form over the communication channel. At the server side, the sequence $\mathbf{Y}$ is retrieved, however it might significantly differ from the original $\mathbf{X}$ due to channel errors. The classical speech recognition problem has to be reformulated as finding the word sequence $\hat{\mathbf{W}}$ which maximizes the conditional probability of word given the received sequence $\mathbf{Y}$, instead of $\mathbf{X}$, which is not available. Using the Bayes theorem and further statistical prediction theory we obtain the new decision rule as:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} \int P(\mathbf{X}|\mathbf{W}) \frac{P(\mathbf{X}|\mathbf{Y})}{P(\mathbf{X})} d\mathbf{X} \cdot P(\mathbf{W}) \tag{1}$$

where the integration is done over the space of the feature vector sequence. Note that a similar result was found by [8], however they neglected the term $P(\mathbf{X})$.

The posterior $P(\mathbf{X}|\mathbf{Y})$ depends on the channel properties during transmission and it is computed using the bit reliability from the channel decoder and some signal statistics. To exploit the inter-frame redundancy, the feature vectors are assumed to be generated by a Markov chain whose transition probabilities were determined in advance using a training database. Practically, the HMM speech recognizer is extended to use the new decision rule by modifying the output probability computation. Instead of evaluating the probability density function at a particular point, the integration over the whole feature space has to be done. We assume that the a priori distribution of feature vectors $P(\mathbf{X})$ and the a posterior $P(\mathbf{X}|\mathbf{Y})$ are Gaussian to reduce the computation of the integral to a simple adaptation of the original

acoustic models $P(\mathbf{X}|\mathbf{W})$.

3. Experimental Setup

The selected compensation techniques were tested on small and large vocabulary tasks. The small vocabulary recognition task was the clean test set of AURORA 2, which consists of 4004 utterances from 52 male and 52 female speakers. The large vocabulary task was the WSJ0 5k (4986 words) Nov'92 evaluation test, closed vocabulary bigram language model, comprising 330 utterances of 4 male and 4 female speakers, summing up to 40 min of speech. For feature extraction we employed the DSR front-end standardized by [1]. The standard also provides an error mitigation scheme based on nearest frame repetition (NFR) which was used as a baseline in our experiments. The word error rates (WER) in an error free scenario were 0.86% for AURORA 2 and 8.99% for WSJ0. The GSM data channel was chosen to be representative of transmission with bit errors. We performed a realistic simulation of the GSM physical layer including channel coding/decoding, interleaving/deinterleaving, modulation/demodulation using the CoWare SPW ("Signal Processing Worksystem") software suite.

The channel coding was TCH/F4.8 with convolutional coding of rate r=1/3 and Bahl decoder providing the bit reliability information at the other end. We have chosen a channel model approximating a "typical urban" profile specified by COST 207 to simulate various conditions by varying the carrier-to-interference ratio (C/I).

In the packet switched scenario, an IP network exhibiting packet loss was simulated using a Gilbert-Elliot model characterized by mean loss probability and conditional loss probability [9]. The investigated conditions, extensively used in the literature, are summarized in Table 1. The payload was one frame-pair per packet.

Table 1: *The conditional loss probability (clp) and mean loss probability (mlp).*

Condition	C1	C2	C3	C4
clp	0.147	0.33	0.5	0.6
mlp	0.006	0.09	0.286	0.385

4. Experimental Results

For each task, the recognition accuracy was evaluated at various channel conditions and using the presented error concealment techniques. At the GSM simulation, the results obtained by marginalization were bad and they were omitted. This happened because even the errors on less significant bits of a feature vector component alter its reliability.

For WV the confidence of each feature vector component was computed as in [3], however using the bit reliability from the channel decoder. The WERs are presented in Figures 1 and 2. The best performance was obtained by UD followed by WV and NFR. Even for the worst channel conditions (C/I less than 4dB), UD performed excellent.

In the Figures 3 and 4 the results of the IP scenario are depicted. The curve labeled M was obtained by marginalization of the lost features. If the loss bursts are relatively short (C1, C2) the NFR is slightly better than M, fact easier to observe in the small vocabulary task. Clearly, repetition is a basic way of exploiting redundancy which is not effective if loss length increases. For the curve labeled WV-α, the lost features were obtained by repetition and their confidence was computed based on the relative position (n) in the burst. Thus, it equaled one at extremities of burst and decreased exponentially $\gamma_n = \alpha^n$ towards the middle. The optimal value of α was experimentally determined and was about 0.7, see also [7].

5. Conclusions

In this work it was experimentally proven that taking into account the confidence in the re-

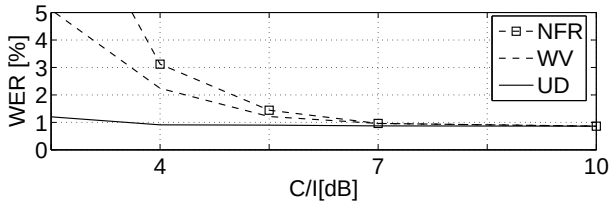

Figure 1: *WER for AURORA 2 on GSM*

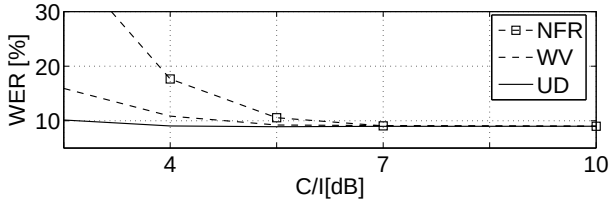

Figure 2: *WER for WSJ0 on GSM*

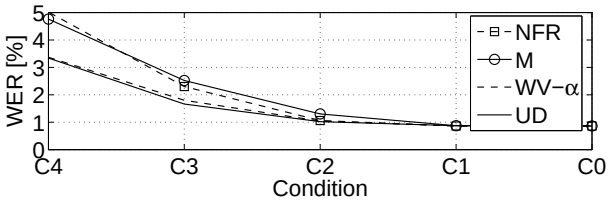

Figure 3: *WER for AURORA 2 on IP network*

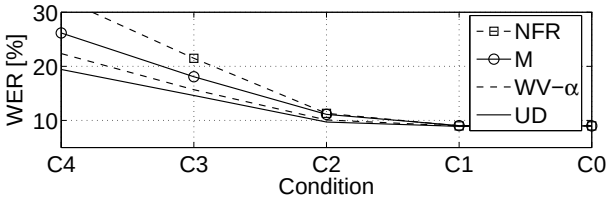

Figure 4: *WER for WSJ0 on IP network*

constructed feature vectors at the recognition stage results in superior performance compared to classical schemes using point estimates.

Based on these experiments it can be concluded that for packet loss channels UD is not significantly better than WV-α. The pitfall is that the optimal setting of α depends on the recognition task. For DSR over circuit switched channels, the bit reliability of the channel decoder and the residual redundancy of the signal are effectively exploited by UD. This results in virtually no degradation of WER, even for bad channel conditions.

6. Acknowledgement

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

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," *Tech. rep. ETSI*, Oct 2002.

[2] A.M. Peinado, V. Sanchez, J.L. Perez-Cordoba, and A. de la Torre, "HMM-based channel error mitigation and its application to distributed speech recognition," *Speech Communication*, vol. 41, no. 6, pp. 549–561, Nov. 2003.

[3] A. Potamianos and V. Weerackody, "Soft-feature decoding for speech recognition over wireless channels," in *Proc. of ICASSP, Salt Lake City, Utah*, 2001.

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

[5] T. Endo, S. Kuroiwa, and S. Nakamura, "Missing feature theory applied to robust speech recognition over IP networks," in *Proc. of EUROSPEECH, Geneva, Switzerland*, 2003.

[6] A. Bernard and A. Alwan, "Joint channel decoding - viterbi recognition for wireless applications," in *Proc. of EUROSPEECH, Aalborg, Denmark*, 2001.

[7] Cardenal-López A., L. Docío-Fernández, and C. García-Mateo, "Soft Decoding Strategies for Distributed Speech Recognition over IP Networks," in *Proc. of ICASSP, Montreal*, 2004.

[8] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 412–421, 2005.

[9] C. Boulis, M. Ostendorf, E:A. Riskin, and S. Otterson, "Gracefully degradation of speech recognition performance over packet-erasure networks," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 8, pp. 580–590, Nov. 2002.