

OOV-DETECTION AND CHANNEL ERROR PROTECTION FOR DISTRIBUTED SPEECH RECOGNITION OVER WIRELESS NETWORKS

Zheng-Hua Tan, Paul Dalsgaard, Børge Lindberg
{zt, pd, bli}@cpk.auc.dk

Center for PersonKommunikation (CPK), Aalborg University, Denmark

ABSTRACT

This paper presents research on two aspects of distributed speech recognition (DSR) in the presence of channel transmission errors in wireless network environments.

The first is on experiments with a frame-based channel error protection scheme, where in previous research we reported results from experiments using randomly distributed bit-errors. This paper presents results from experiments using three additional, more realistic error distributions: burst-like packet loss, GSM error patterns and UMTS statistics.

The second is on exploiting the knowledge about channel transmission errors for the purpose of optimising the Out-of-Vocabulary (OOV) detection. Transmission errors influence the acoustic likelihood, and therefore affect the optimal threshold setting for discrimination between In-Vocabulary (IV) words and OOV words. An OOV-detection method is proposed in which the estimated Frame-Error-Rate (FER) is used to adjust the discrimination threshold. Results from experiments are reported over a range of transmission errors.

1. INTRODUCTION

In a client-server architecture the modules of a DSR system are split between the terminal (client) and the server. The front-end pre-processor is located in the terminal to which the remote back-end server is 'connected' via the transmission network. Non-perfect network transmission definitely induces a number of constraints to currently used processing methodologies conceptually similar to the influence of environmental noise to the speech signal. Without special compensation techniques, the performance of speech recognition degrades seriously when used in error-prone transmission environments.

In previous research [1] we proposed to use a frame-based channel error protection scheme instead of the frame-pair based scheme standardised by the ETSI-DSR Group [2,3]. The recognition experiments in [1,4] showed a significant increase in recognition accuracy for Additive White Gaussian Noise (AWGN) channels simulated over a range of Bit-Error-Rates (BER) (from 0 to 2%).

In section 2 we present results from a set of recognition experiments in which three additional and more realistic error distributions are used: burst-like channel errors to simulate a Rayleigh Fading channel, GSM error patterns [5] EP1, EP2 and EP3 and UMTS statistics.

A channel error protection scheme (detection and mitigation) can only partly alleviate the impairment on speech

features due to transmission errors. However, the modification of speech features will influence the performance of the recognition back-end. As a consequence transmission errors affect the acoustic likelihood and therefore also the threshold setting for OOV detection. However, if the Cyclic Redundancy Checking (CRC) information in channel error protection is exploited to estimate the current FER, then an FER-dependent threshold can be employed to optimise the OOV detection.

Section 3 presents details of this FER-based OOV detection method. Section 4 presents the summary and discussions.

2. CHANNEL ERROR PROTECTION

Within the ETSI-DSR standard, two quantised mel-cepstral frames are grouped together and protected with a 4-bit CRC forming a frame-pair [2,3]. This causes the entire frame-pair erroneous even if only a single bit error occurs in the frame-pair packet. No major degradation is observed for strong and medium GSM signal strengths using the frame-pair error protection scheme. However, for a poor channel, e.g. 4 dB carrier-to-interference (C/I), the recognition performance degrades from 10.0% to 16.2% for different tasks in comparison to the case of transmission without errors [5].

To overcome this, a frame-based error protection scheme was deployed in [1] to protect each frame independently causing the overall probability of one frame in error to be lower (at the cost of only a marginal increase in bit-rate, from 4,800 bits/s to 5,000 bits/s), see Figure 1.

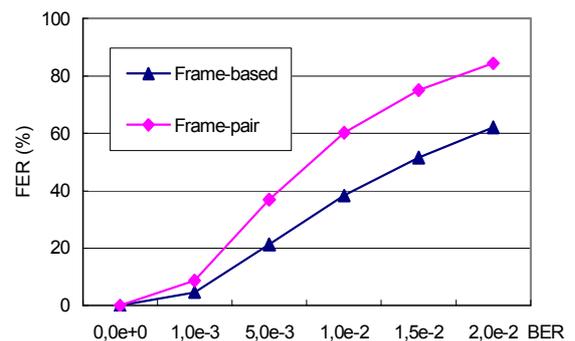


Figure 1. %FER vs BER for two different channel error protection schemes

To evaluate the frame-based error protection scheme, a number of recognition experiments have been conducted. Two different recognition tasks have been investigated: Danish digits

recognition (low perplexity) and city names recognition (medium perplexity).

The recogniser applied in the experiments is the SpeechDat/COST 249 reference recogniser [6]. 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).

The DA-FDB 4000 database is used for training 32 Gaussian mixture triphone models. Test data - isolated digits and city names - are from the same database.

2.1 AWGN channel

In previous work, we reported on two recognition tasks, namely Danish digits and city names for the AWGN channel. Figure 2 and Figure 3 show the results for the digits and the city names, respectively [1,4].

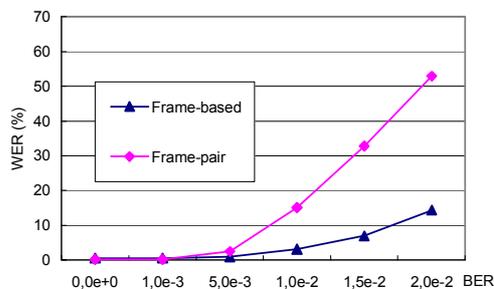


Figure 2. %WER vs. AWGN channel BER for Danish digits

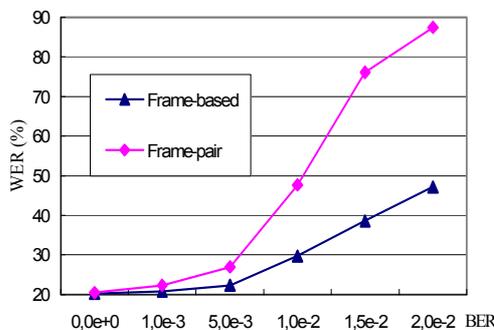


Figure 3. %WER vs. AWGN channel BER for city names

2.2 Burst-like packet loss

Burst-like errors occur in Rayleigh Fading channels. The errors are simulated using a three-state Markov Model as in [7].

A packet loss of 10% with an average loss-frame length of 8 is simulated. For the Danish digits task, the WER decreases from 8.5% to 7.1% - an improvement of 17%. The WER for the city names task decreases from 34.2% to 30.8% - an improvement of 10%.

2.3 GSM error patterns

Error patterns are commonly used for testing speech codecs and DSR error protection schemes. GSM transmission over a 9,600 bps data channel is simulated by adding error patterns to the DSR data stream.

For the frame-pair tests the error patterns are used according to [5]; for the frame-based CRC tests in a similar way. The three error patterns are: EP1, EP2 and EP3 corresponding to C/I ratios of 10 dB, 7 dB and 4 dB, respectively. The results of testing on the Danish digits task are shown in Figure 4 and the results for the city names task are shown in Figure 5.

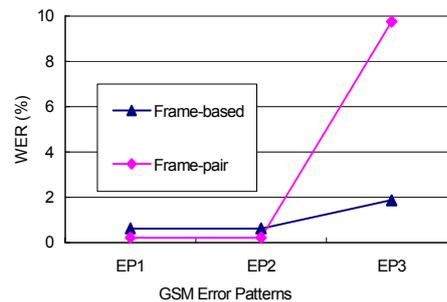


Figure 4 %WER vs. GSM error patterns for Danish digits

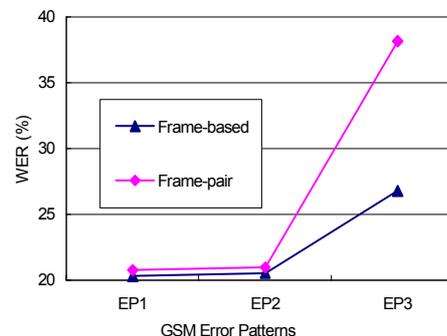


Figure 5 %WER vs. GSM error patterns for city names

It is observed that for EP3, the WER for the Danish digits task decreases from 9.8% to 1.9% - an improvement of 81%. The WER for the city names task decreases from 38.2% to 26.8% - an improvement of 30%.

2.4 UMTS statistics

UMTS statistics are provided from a system-level network simulator, which is able to simulate a large variety of scenarios and user deployments in order to extract realistic performance statistics regarding packet error rate or blocking probability. In this experiment, the statistics data encompasses 21,645 frames concatenated from 125 users (scenarios).

For UMTS statistics the WER of Danish digits task decreases from 3.3% to 2.5% - an improvement of 25%. The WER

of the city names task decreases from 26.6% to 23.9% - an improvement of 10%.

3. OOV DETECTION IN DSR SYSTEMS

OOV detection is a statistical hypothesis testing problem in which a decision algorithm accepts or rejects the hypothesis [8,9]. Given a speech signal observation O , the algorithm tests the null hypothesis H_0 against the alternative hypothesis H_1 . H_0 represents one of the IV words and H_1 represents OOV words modelled by one filler model. A likelihood ratio $LR(O)$ based on the null and alternative hypotheses is then used to detect OOV words. The test rejects the H_0 hypothesis if

$$LR(O) = \frac{p(O|H_0)}{p(O|H_1)} < T$$

where T is the threshold of the test. $p(O|H_0)$ and $p(O|H_1)$ are the probability density functions of the H_0 and the H_1 hypotheses, respectively.

Transmission errors may, however adversely affect the distribution of the likelihood of both the IV models and the filler model.

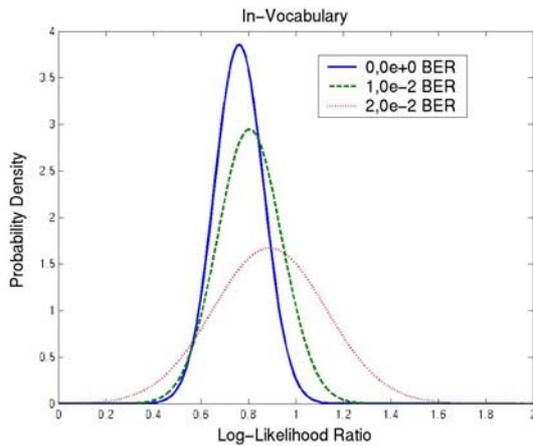


Figure 6. Probability densities of the log-likelihood ratios for the IV words for three different BER values

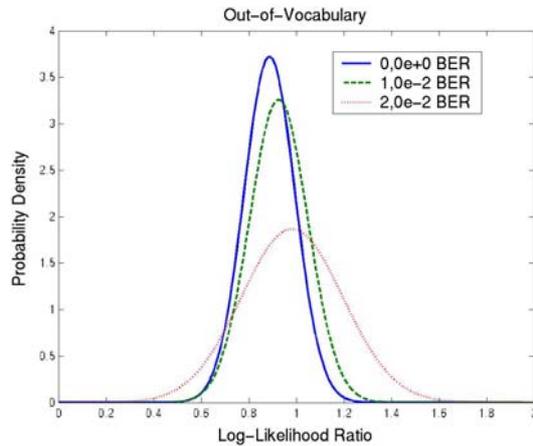


Figure 7. Probability densities of the log-likelihood ratios for OOV words for three different BER values

Figure 6 and Figure 7 show the best Gaussian fit to the log-likelihood ratios of IV words and OOV words from the experiments for channels with 0%, 1% and 2% BER values, respectively. In this paper, the IV words are the Danish digits and the OOV words are the city names.

These figures evidence that the transmission errors change the probability density of the log-likelihood ratio in two ways. One effect of transmission errors is that the standard deviation of the distributions are increased for increasing BER values. This weakens the discrimination between IV and OOV words. Another effect is the shift of the mean of the distributions which affects the optimal threshold setting for OOV detection. A fixed threshold method may therefore fail to maintain the balance of the false rejection and false acceptance rates.

One way to aim at maintaining the balance is to adjust the threshold according to the transmission errors. A FER-based OOV detection method is therefore suggested where the CRC in the error protection scheme is exploited to estimate the current FER – representing the transmission errors – and use this estimate to determine the threshold for OOV detection.

3.1 FER-based OOV detection

A FER-dependent threshold for OOV detection is deployed where the threshold is modelled as a fourth-order polynomial function of the FER. To calculate the coefficients of the polynomial, five experiments (with BER values ranging from 0.1% to 2%) were conducted using a training database consisting of 282 digits utterances and 249 city names utterances. The FER values are calculated from the BER values according to Figure 1.

The thresholds for each of these experiments are chosen with the specific optimisation target of maintaining the false rejection rate approximately constant across a range of BER values.

The filler model is a 32 Gaussian mixture five-state HMM model trained on the basis of a large amount of speech data with no transmission channel involved.

Test data for the experiments described below are the remaining 200 digits and 200 city names utterances from the same database. The CRC is utilised to estimate the FER, which is then used for adjusting the threshold of the OOV detection based on the fourth-order polynomial function.

Figure 8 shows that the OOV detection algorithm using FER-dependent threshold maintains the false rejection rate of IV words within the range from 4% to 6% whereas the false rejection rate using a fixed threshold is varying in the range from 4.5% to 20%. The experiments were targeted at a false rejection rate of 5%.

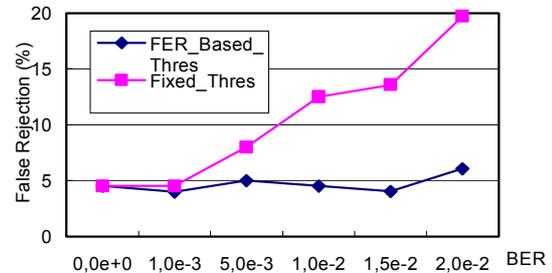


Figure 8. False rejection rate vs. AWGN channel BER values

The results in Figure 9 show that the overall recognition rate is improved using the FER-dependent threshold approach as compared to a fixed threshold.

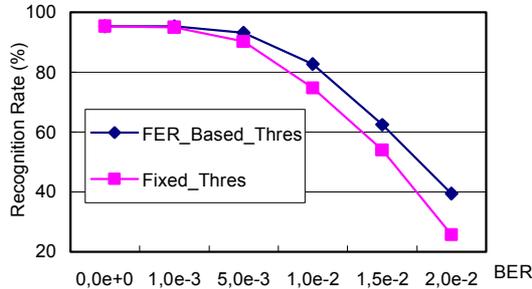


Figure 9. Recognition rate vs. AWGN channel BER values

In maintaining an almost constant false rejection rate, the false acceptance rate increases as shown in Figure 10. However, in general threshold setting is a trade-off between false rejection and false acceptance and therefore design criteria (such as equal error rate requirements) could be the basis for the FER-based OOV detection.

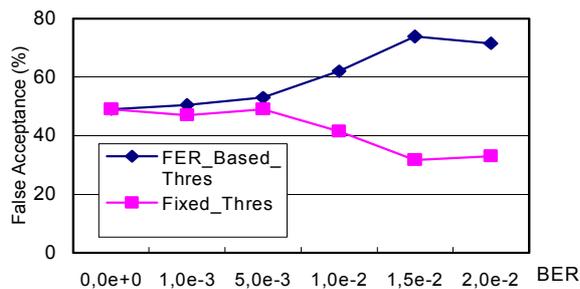


Figure 10. False acceptance rate vs. AWGN channel BER values

4. DISCUSSION

In this paper the frame-based error protection scheme has been tested on three different realistic transmission error distributions: burst-like packet loss, GSM error patterns and UMTS statistics.

The results verify and generalise the conclusions from previous research on more artificial error distributions: compared to the frame-pair scheme, the frame-based error protection scheme is able to better maintain the recognition rates. The cost of using the frame-based scheme is only marginal, as the bit-rate increases from 4,800 bits/s to 5,000 bits/s for which there is plenty of bandwidth available in GSM and higher bandwidth wireless channels. The results are consistent across two different recognition tasks – a low perplexity digits task and a medium perplexity city names task.

Further research is reported from experiments focussing on the back-end recogniser where knowledge about current channel transmission errors is exploited adaptively to optimise the OOV detection. Since the likelihood ratio pdf's are changed due to transmission errors, an FER-dependent threshold is proposed as an OOV-detection method. This method proved successful in maintaining a constant false rejection rate across a range of error rates.

As the frame-pair protection scheme is the current standard, it was chosen to use this as a basis for the OOV-detection experiments. However, if the frame-based scheme is introduced instead, a further improvement will be achieved.

The principle of exploiting information about the channel (e.g. fading channels, overloading services, congested networks and degraded acoustic environments) can be exploited to adapt the dialogue and the applied grammars and vocabularies. This will enable a graceful modification of the behaviour of a given dialogue application according to the current quality of the channel.

5. ACKNOWLEDGEMENT

We would like to thank David Pearce (from Motorola Labs) for providing us with the GSM EP patterns and Laurent Schumacher from CPK, Aalborg University for the UMTS statistics.

6. REFERENCES

- [1] Z.-H. Tan and P. Dalsgaard, "Channel Error Protection Scheme for Distributed Speech Recognition," *ICSLP-2002*, Denver, USA, September 2002.
- [2] D. Pearce, "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
- [3] "ETSI ES 201 108 v1.1.2 Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithm", February 2000.
- [4] Z.-H. Tan, B. Lindberg and P. Dalsgaard, "Experiments on A Channel Error Protection Scheme for Distributed Speech Recognition," *NORSIG-2002*, Norway, October 2002.
- [5] Aurora document no. AU/266/00 "Recognition with WI007 Compression and Transmission over GSM Channel", Ericsson, December 2000.
- [6] B. Lindberg, F.T. Johansen, N. Warakagoda, et al, "A Noise Robust Multilingual Reference Recogniser Based on SpeechDat(II)," in Proc. *ICSLP-2000*, October 2000.
- [7] B. Milner, "Robust Speech Recognition in Burst-Like Packet Loss," in Proc. *ICASSP-01*, USA, May 2001.
- [8] M.G. Rahim, C.-H. Lee and B.-H. Juang, "Discriminative Utterance Verification for Connected Digits Recognition," *IEEE Transaction on Speech and Audio Processing*, vol. 5, no. 3, pp. 266-277, May 1997.
- [9] E. Lleida and R.C. Rose, "Utterance Verification in Continuous Speech Recognition: Decoding and Training Procedures," *IEEE Transaction on Speech and Audio Processing*, vol. 8, no. 2, pp. 126-139, March 2000.