Adaptive Multi-Frame-Rate Scheme for Distributed Speech Recognition Based on a Half Frame-Rate Front-End

Zheng-Hua Tan, Paul Dalsgaard and Børge Lindberg
Centre for TeleInFrastructure (CTIF), Speech and Multimedia Communication (SMC)
Aalborg University, Denmark
{zt, pd, bli}@kom.aau.dk

***Abstract*—In this paper a half frame-rate (HFR) front-end is investigated for distributed speech recognition (DSR). The work is inspired from the need for low bit-rate and is justified by the redundancies known to exist in full frame-rate (FFR) features. At the client-side in the DSR architecture, implementation of the HFR is carried out by using double frame shifting as compared to the FFR resulting in the achievement of half the bit rate. At the server-side, each HFR feature vector is repeated once to construct the FFR features and no changes are therefore required in the recognition back-end. It is experimentally justified that the performance achieved by HFR is comparable to FFR and that repetition of each HFR feature vector is critical for the HFR front-end to maintain the performance. Motivated by the effectiveness of HFR, a number of additional FFR-based DSR schemes are further presented. Finally, this paper introduces an adaptive multi-frame-rate scheme in which the DSR system adapts to the characteristics of the transmission channel by switching between HFR and the FFR-based schemes. This multi-frame-rate scheme is found to be superior to the basic FFR.**

Keywords—distributed speech recognition; low bit-rate; multi-frame-rate; transmission error robustness

I. INTRODUCTION

Aimed at optimal performance of automatic speech recognition (ASR) over mobile networks, an important research topic within ASR has been to focus on the issue of DSR. In the client-server DSR system architecture, the ASR processing is split into the client based front-end feature extraction and the server based back-end recognition, where data transmission between the two parts may take place via heterogeneous networks. However, the transmission of data across networks presents a number of challenges to, for example bandwidth limitations and transmission errors [1].

Research in the context of DSR is mainly concerned with three aspects namely front-end processing, source coding and channel coding [2]. Since the mel-frequency cepstral coefficient (MFCC) features are extensively used and have proved to be successful for ASR, MFCCs are used for most DSR front-ends. The goal of source coding is to compress information aiming at low bit-rate. One common class of source coding schemes for DSR applies split vector quantization (VQ) for the coding of ASR features in addition to the recently introduced transform coding such as discrete cosine transform that pursues very low bit-rate [3].

Channel coding techniques attempt to protect information from transmission errors. [4] introduces linear block codes and a soft decision decoding and Red-Solomon coding is applied in [5]. [6] introduces interleaving to handle burst-like packet losses. All these techniques can recover a large amount of transmission errors, however, at such cost as additional delay, increased bandwidth and higher computational overhead. Motivated by the temporal correlation present in the speech features [7] introduced a data consistency test to identify inconsistent sub-vectors within erroneous vectors, resulting in a sub-vector error concealment (EC) scheme conducting EC at the sub-vector level instead of at the full vector level.

The temporal correlation is further exploited in this work with the aim of achieving both low bit-rate and high robustness against transmission errors. In particular, a HFR processing technique is investigated in detail followed by the introduction of a number of FFR-based DSR schemes each originating from the HFR principle. The HFR feature extraction is carried out based on choosing the double length of the normally applied frame shift. Before the server ASR decoding, each HFR feature vector is repeated once resulting in the construction of an approximation to FFR feature vector. Transmission of the HFR features therefore requires only half the bandwidth as compared to transmitting FFR features.

The HFR contains on the one hand less redundant information but it may on the other hand be more sensitive to transmission errors. Additionally, using HFR may cause the ASR performance to degrade for complex recognition tasks. With the goal of counteracting these effects an adaptive multi-frame-rate scheme is established where the DSR system is designed to adapt to the transmission channel by selecting either HFR or FFR.

The effectiveness of the HFR is the factor motivating the introduction of a number of additional FFR-based DSR schemes including multiple description coding (MDC) and a specific version of interleaving.

II. HALF FRAME-RATE FRONT-END

The temporal correlation between speech features from consecutive speech frames is caused partly by the vocal tract inertia partly by the overlapping in the feature extraction

This work has been supported by the FACE project, the CNTK consortium project and the EU 6th framework project MAGNET.

procedure. With the limited bandwidth requirements of DSR the goal of this section is to investigate ways to reduce the bit-rate in the feature extraction stage.

In the ASR front-end processing speech features are commonly computed with a 25 ms frame length and a 10 ms frame shift, causing a 15 ms overlap between consecutive frames. In this work the HFR is implemented by using a 20 ms frame shift and thereby resulting in a 5 ms overlap. Prior to server-side recognition, each HFR speech feature vector is repeated once to reconstruct the FFR vector and thus let the back-end recogniser unchanged. This is similar to the HFR reported in [8] where, however, the server interpolates the features by a factor of two. The present work uses repetition instead of interpolation in reconstructing the FFR features due to experiments reported in [2] showing that repetition performs better than interpolation in recovering missing features resulting from transmission errors.

In addition to providing a low bit-rate feature stream for DSR, the HFR front-end has the advantage of only requiring half the computational cost in its feature extraction process, which may be a significant merit for resource-limited hand-held devices. It is worth mentioning here that source coding in contrast achieves low bit-rate but at the cost of introducing additional computations.

III. Adaptive Multi-Frame-Rate Scheme

There is always a trade-off between the requirement of low delay and low bandwidth against the performance degradation caused by both coding compression and transmission errors. The HFR coding obviously has the benefit of low bit-rate but is on the other hand likely to be more sensitive to transmission errors, which as a consequence motivates an adaptive multi-frame-rate scheme that is able to switch between HFR and FFR processing. Before presenting this adaptive scheme, the ETSI-DSR standard is introduced as a baseline system.

A. TheFFR-based ETSI-DSR Standard

ETSI published the first DSR standard with the aim of handling the degradations of ASR over mobile networks caused by both lossy speech coding and channel errors [1].

The standard defines the feature extraction front-end together with a coding scheme [9]. The FFR front-end produces a 14-element vector consisting of log energy (logE) in addition to 13 MFCCs c_0 to c_{12} computed every 10 ms. Each feature vector is compressed using split VQ into 44 bits. Two quantized frames are grouped together and protected by a 4-bit cyclic redundancy check (CRC) creating a 92-bit frame-pair. Twelve frame-pairs are combined and appended with overhead bits resulting in an 1152-bit multi-frame, as shown in Fig. 1 where the numbering of frames in each multi-frame starts from one and the odd-numbered frames are coloured while the even-numbered frames are white. Multi-frames are concatenated into a 4 800 bps bit-stream for transmission.

At the server, two calculations determine whether a frame-pair is received with errors, namely a CRC test and a data consistency test. The CRC test determines if a frame-pair is received with errors. The data consistency test determines whether or not the decoded features for each of the two speech vectors in a frame-pair have the minimal continuity. In ETSI-DSR EC processing, a repetition scheme is applied to replace erroneous vectors.

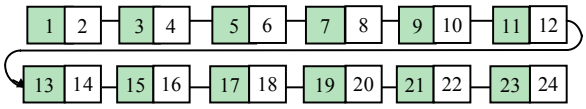

Figure 1. Frame-pair FFR coding as used in the ETSI-DSR standard.

B. FFR based One-Frame Coding

The use of frame-pair formatting in the ETSI-DSR standard causes the entire frame-pair to be designated erroneous even if only a single bit error occurs in the frame-pair. To overcome this, [10] proposed an alternative one-frame scheme to protect each frame independently and thus causing the overall probability of one frame in error to be lower and showed improved recognition performance (at the cost of only a marginal increase in bit-rate, from 4 800 bps to 5 000 bps). The FFR multi-frame architecture of the one-frame coding is illustrated in Fig. 2(a).

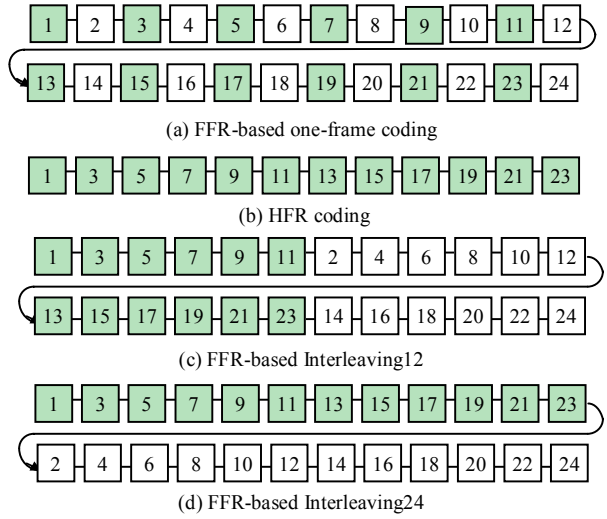

Figure 2. HFR and three different FFR coding schemes.

C. HFR Coding

In the HFR coding - shown in Fig. 2(b) - each multi-frame encompasses twelve rather than 24 frames. Each frame is protected by a 4-bit CRC resulting in a 48-bit frame. Twelve frames are joined and appended with overhead bits resulting in a 624-bit multi-frame. Multi-frames are concatenated into a 2 600 bps bit-stream (as opposed to 4 800 bps for the ETSI-DSR FFR). At the server, each CRC is used as the only error detection method and no data consistency test is conducted.

D. FFR-based Multiple Description Coding

The HFR feature vectors are simply the odd-numbered feature vectors in the corresponding FFR front-end. Together with the even-numbered feature vectors, two descriptions of the speech signal are created and each of them can be transmitted independently, resulting in a MDC coding scheme. A general characteristics of an MDC encoder is that a source is encoded into two or more sub-streams (descriptions) that each can be delivered on separate channels with the aim of exploiting channel diversity and thus improving robustness against transmission errors [11].

E. FFR-Based Interleaving

In both the ETSI-DSR and the one-frame coding, each even-numbered feature is transmitted immediately followed by its corresponding odd-numbered feature. Alternatively, a certain number of odd-numbered features can be concatenated and transmitted first and their corresponding even-numbered features transmitted later, resulting in a special version of interleaving that is capable of counteracting a large amount of burst-like transmission errors. Figures 2(c) and 2(d) present two interleaving schemes: Interleaving12 in which a sequence of 12 vectors is grouped into one block and Interleaving24 where a sequence of 24 vectors is grouped.

The difference between conventional interleaving and the special interleaving is that the latter may offer less overall delay. Immediately following reception of data for decoding, the CRC test determines whether the transmission has caused errors. If there are no errors, the odd-numbered feature vectors can be repeated without awaiting the even-numbered feature vectors causing no delay.

F. Adaptive Multi-Frame-Rate Scheme

The adaptive multi-frame-rate scheme is implemented in a way that a DSR system is able to switch between the HFR scheme and one of the FFR-based schemes presented above depending on the channel characteristics. The principle is illustrated in Fig. 3. At the client side, the speech signal is processed either by the HFR front-end or by the FFR front-end. If the FFR front-end is chosen, one of the schemes namely one-frame, MDC, interleaving12 and interleaving24 is selected for channel encoding. At the server side, the matching decoding process is conducted.

Switching between the HFR coding and the FFR coding results in a multi-frame-rate DSR codec, which is functionally similar to the adaptive multi-rate (AMR) speech codec.

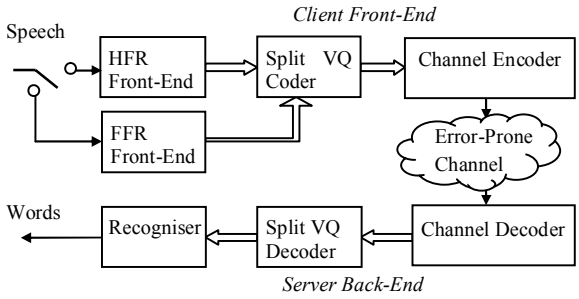

Figure 3. DSR system using adaptive multi-frame-rate scheme.

IV. Experiments and Discussions

Two speech databases are used for investigating the performance of the proposed methods namely the Danish SpeechDat 2 database DA-FDB 4000 and the Aurora 2 database.

The DA-FDB 4000 database covers speech from 4000 Danish speakers collected over the fixed network. A part of the database is used for the training of 32 Gaussian mixture tri-phone models. The independent test data - isolated digits (low perplexity) and city names (medium perplexity) - are from this database as well. The recogniser applied is the HTK-based SpeechDat/COST 249 reference recogniser [12].

The Aurora 2 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. Each of the digit models has 16 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. In this evaluation, clean speech training is used.

A. The HFR Front-End Testing - No Transmssion Errors

The performance of the HFR front-end is evaluated on the two databases. No transmission errors are involved in this evaluation.

1) Danish digits and city names: The results for the Danish digits and city names tasks for the HFR or FFR front-ends are shown in Table I. In the experiment, tri-phone models are trained using ETSI-DSR FFR features without VQ. The features for the test data are all after VQ processing. It is seen that the HFR front-end achieves results that are close to the FFR front-end. However, using the HFR feature without repeating each feature (HFR-NoRepeat) to construct the FFR gives substantially lower recognition accuracy.

TABLE I. Recognition Accuracy (%) across the Front-Ends for Danish Digits and City Names Using FFR Models without VQ

	Danish Digits	City Names
FFR	99.79	79.29
HFR	99.59	79.29
HFR-NoRepeat	96.68	61.25

2) Aurora 2: A number of more comprehensive experiments have been conducted with data from the Aurora 2 database. The test data are the clean data from Test Set A. Table II demonstrates the recognition accuracy across the same three FFR/HFR feature extraction techniques as used in the above experiment. The models used in this evaluation are trained on FFR features without VQ whereas the test data are quantized by VQ. The HFR front-end demonstrates comparable average performance to the FFR (relative drop of 0.07%) although the models are trained using the ETSI-DSR FFR features. HFR-NoRepeat gives significantly lower recognition accuracy.

TABLE II. Recognition Accuracy (%) across the Front-Ends for Test Set A Using FFR Models without VQ

	Clean1	Clean2	Clean3	Clean4	Average
FFR	98.86	99.00	99.08	99.23	99.05
HFR	98.93	98.97	98.99	99.04	98.98
HFR-NoRepeat	70.13	71.92	71.85	70.56	71.12

Table III gives the results for tests in which matching models are used, i.e. the feature processing is the same for both training and test data and VQ is applied. From the results it is observed that the HFR front-end gives results close to the FFR front-end. The performance of HFR-NoRepeat is still substantially lower although both training features and test features are matched.

TABLE III. RECOGNITION ACCURACY (%) ACROSS THE FRONT-ENDS FOR TEST SET A USING MATCHED MODELS AFTER VQ

	Clean1	Clean2	Clean3	Clean4	Average
FFR	98.96	99.03	98.90	99.11	99.00
HFR	98.89	99.03	98.90	99.11	98.98
HFR-NoRepeat	88.92	89.72	89.56	89.29	89.37

The results in Table III justify that the performance achieved by HFR is close to the FFR results; however, the HFR-NoRepeat results show that repetition of each HFR feature vector is critical even though using matching models.

B. Robustness against Transmission Errors

The evaluation is extended by testing the robustness of each of the coding schemes against transmission errors. Since the testing of the adaptive multi-frame-rate scheme requires a complex network simulator and the results will be highly dependent on the settings of the simulator, evaluation in this work is limited to the testing on the basis of individual schemes presented in Section III. Three GSM error patterns (EP) are often used as they include a merging of both random errors and burst-like errors. EP3 only is chosen for this evaluation since EP1 and EP2 do not cause noticeable performance degradation [7]. For testing MDC, the two description encodings are transmitted over two uncorrelated channels both simulated by EP3.

The same FFR models without VQ are applied in all the experiments. Table IV shows the recognition accuracy of a number of schemes for GSM EP3, including the sub-vector concealment [8]. The FFR (ETSI-DSR) represents the ETSI-DSR standard served as a baseline for this evaluation.

TABLE IV. RECOGNITION ACCURACY (%) ACROSS THE CODING SCHEMES FOR EP3 FOR TEST SET A USING FFR MODELS WITHOUT VQ

	Clean1	Clean2	Clean3	Clean4	Average
FFR (ETSI-DSR)	92.82	92.26	94.39	93.74	93.30
HFR	95.33	95.37	95.62	95.43	95.44
One-frame	96.19	96.43	97.32	96.42	96.59
Sub-vector	97.08	97.01	97.73	97.59	97.35
Interleaving12	97.21	97.58	97.88	97.59	97.57
Interleaving24	98.28	98.16	98.27	98.33	98.26
MDC	98.77	98.91	99.05	99.11	98.96
Error-free	98.86	99.00	99.08	99.23	99.05

The results show improved performance for the HFR front-end as compared to the ETSI-DSR FFR standard even though the HFR has only approximately half the bandwidth requirement. Improvements that are more significant are observed for the interleaving schemes. It is worth noticing that the special interleaving schemes introduce a delay in the decoding stage only when there are transmission errors, as discussed in Section 3.E. It is also observed that the performance of the MDC scheme approaches that of the error-free channel. Since all individual schemes introduced in this work are superior to the ETSI-DSR in terms of robustness against transmission errors, the deployment of the adaptive multi-frame-rate scheme should exhibit an overall performance superior to that of the ETSI-DSR.

V. CONCLUSIONS

This paper investigates a HFR front-end and an adaptive multi-frame-rate scheme for DSR. In the HFR front-end, feature extraction is carried out with a frame rate being only half of the conventional FFR. Prior to recognition at the server-side, each of the HFR features is repeated once to construct FFR features. Experimental results justify that the recognition accuracy in applying a HFR front-end is close that of the FFR front-end.

The exploitation of the HFR front-end is further extended into an adaptive multi-frame-rate scheme allowing the DSR system to switch between the HFR and the FFR-based schemes consisting of one-frame, MDC, Interleaving12 or Interleaving24 scheme. The one-frame scheme uses a FFR front-end with CRC protection for each frame instead of frame-pair. When only one transmission channel is available, the interleaving schemes maintain the highest recognition performance even for severe error-prone channels and they further have the advantage of not introducing delay when there are no transmission errors. The MDC scheme gives a performance close to that of the error-free channel with the requirement of the availability of independent multiple channels.

The adaptive multi-frame-rate method is also suitable for packet-switched networks where packet losses are the dominant causes for the degradation of ASR performance. Future work will consider applying sub-vector concealment into the adaptive multi-frame-rate scheme and generally investigating variable frame-rate feature extraction for DSR.

REFERENCES

[1] D. Pearce, "Robustness to transmission channel – the DSR approach," Robust-2004, Norwich, UK, August 2004.

[2] Z.-H. Tan, P. Dalsgaard and B. Lindberg, "Automatic speech recognition over error-prone wireless networks," *Speech Communication*, in press, 2005.

[3] W.-H Hsu and L.-S Lee, "Efficient and robust distributed speech recognition (DSR) over wireless fading channels: 2D-DCT compression, iterative bit allocation, short BCH code and interleaving," Proc. ICASSP04.

[4] A. Bernard, A. and A. Alwan, "Low-bitrate distributed speech recognition for packet-based and wireless communication", *IEEE Trans. SAP*, November 2002.

[5] C. Boulis, M. Ostendorf, E.A. Riskin and S. Otterson, "Gracefully degradation of speech recognition performance over packet-erasure networks," *IEEE Trans. SAP*, vol. 10, pp. 580-590, November 2002.

[6] A.B. James, and B.P. Milner, "An analysis of interleavers for robust speech recognition in burst-like packet loss," Proc. ICASSP 2004.

[7] Z. -H. Tan, P. Dalsgaard and B. Lindberg, "A subvector-based error concealment algorithm for speech recognition over mobile networks," Proc. ICASSP 2004.

[8] A. Bernard and A. Alwan, "Joint channel decoding – Viterbi recognition for wireless applications," Proc. Eurospeech, Aalborg, Denmark, 2001.

[9] Distributed speech recognition; front-end feature extraction algorithm; compression algorithms, ETSI ES 201 108 v1.1.2 2000.

[10] Z. -H. Tan, P. Dalsgaard and B. Lindberg, "OOV-detection and channel error protection for distributed speech recognition over wireless networks," Proc. ICASSP, Hong Kong, China, April 2003.

[11] Goyal, V.K., "Multiple Description Coding: Compression Meets the Network", *IEEE Signal Processing Magazine*, 18 (5), 2001.

[12] B. Lindberg et al, "A Noise Robust Multilingual Reference Recogniser Based on SpeechDat(II)," in Proc. ICSLP-2000, October 2000.