Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Speech Communication 47 (2005) 220–242

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic speech recognition over error-prone wireless networks ☆

Zheng-Hua Tan *, Paul Dalsgaard, Børge Lindberg

Centre for TeleInFrastructure (CTIF), Speech and Multimedia Communication (SMC), Aalborg University, Niels Jernes Vej 12, DK-9220 Aalborg, Denmark

Received 19 December 2004; received in revised form 5 May 2005; accepted 11 May 2005

Abstract

The past decade has witnessed a growing interest in deploying automatic speech recognition (ASR) in communication networks. The networks such as wireless networks present a number of challenges due to e.g. bandwidth constraints and transmission errors. The introduction of distributed speech recognition (DSR) largely eliminates the bandwidth limitations and the presence of transmission errors becomes the key robustness issue. This paper reviews the techniques that have been developed for ASR robustness against transmission errors.

In the paper, a model of network degradations and robustness techniques is presented. These techniques are classified into three categories: error detection, error recovery and error concealment (EC). A one-frame error detection scheme is described and compared with a frame-pair scheme. As opposed to vector level techniques a technique for error detection and EC at the sub-vector level is presented. A number of error recovery techniques such as forward error correction and interleaving are discussed in addition to a review of both feature-reconstruction and ASR-decoder based EC techniques. To enable the comparison of some of these techniques, evaluation has been conduced on the basis of the same speech database and channel. Special attention is given to the unique characteristics of DSR as compared to streaming audio e.g. voice-over-IP. Additionally, a technique for adapting ASR to the varying quality of networks is presented. The frame-error-rate is here used to adjust the discrimination threshold with the goal of optimising out-of-vocabulary detection.

This paper concludes with a discussion of applicability of different techniques based on the channel characteristics and the system requirements.

Keywords: Distributed speech recognition; Channel error robustness; Out-of-vocabulary detection

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

* Corresponding author. Tel.: +45 9635 8686; fax: +45 9815 1583.
E-mail addresses: zt@kom.aau.dk (Z.-H. Tan), pd@kom.aau.dk (P. Dalsgaard), bli@kom.aau.dk (B. Lindberg).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.05.007

1. Introduction

To facilitate the access to services over wireless networks there is often a demand to include automatic speech recognition (ASR) as a key component in the user interface (Cox et al., 2000; Lee and Lee, 2001). However, terminals of today are often hand-held devices with limited battery life, computing power and memory size which all taken together constitute a challenge for the implementation of any complex ASR system into these devices. For ASR systems associated with large databases, security and consistency are additional challenges to be taken into account for terminal-based implementation (Rose et al., 2003). In addition, the 'always-on' facility of communication networks offers improved opportunities for operating ASR modules using a distributed architecture in which only the front-end processing requires specific porting and implementation into the hand-held devices. Therefore, the development of network-based ASR emerges as a recent trend.

To enable low bit-rate data transmission in distributed architectures, speech coding is applied to conduct data compression. Lossy speech codecs may, however, significantly degrade the performance of ASR (Haavisto, 1999). A way to eliminate this degradation is to introduce distributed speech recognition (DSR) (Pearce, 2000, 2004) and DSR has been an important research focus within ASR during the last decade. 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 are transmitted between the two parts via heterogeneous networks. Comparative studies have shown superior performance of DSR to codec-based ASR (Kelleher et al., 2002; Kiss, 2000).

Although DSR eliminates the degradations originating from the speech compression algorithms, the transmission of data across networks still brings in a number of problems to speech recognition technology, in particular transmission errors. This paper analyses the characteristics of transmission error degradations and attempts to model the degradations and the corresponding error-robustness techniques. In order to reduce transmission error degradation, client-driven recovery and server-based concealment techniques are applied within DSR systems (where the client is always at the sender side and the server at the receiver side) in addition to error detection techniques. Client-based techniques e.g. retransmission, interleaving and forward error correction (FEC) may result in recovering a large amount of transmission errors (e.g. Hsu and Lee, 2004; James and Milner, 2004). However, generally these methods have a number of disadvantages such as additional delay, increased bandwidth and computational overhead (Perkins et al., 1998). Server-based error concealment (EC) techniques exploit the redundancy in the transmitted signal and this may be used independent of, or in combination with, client-based techniques. When used in combination, the purpose is to handle the remaining errors that a purely client-based technique fails to recover.

Server-based EC for DSR may either be conducted through feature-reconstruction or modification of the ASR-decoder. Feature-reconstruction EC generally employs one of the following techniques: splicing, substitution (with silence, noise or source-data), repetition or interpolation (Boulis et al., 2002; Milner and Semnani, 2000). Tan et al. (2003b, 2004a) recently proposed a partial splicing and a sub-vector concealment technique that both demonstrate better performance. A group of statistical-based techniques have been developed which all exploit the statistical information about speech for feature-reconstruction (Gomez et al., 2004; James et al., 2004). In wireless communications, a number of studies exploit the reliability information of the received bits to improve feature-reconstruction (Haeb-Umbach and Ion, 2004; Peinado et al., 2003).

In the ASR-decoding stage, the effect of transmission errors can be mitigated by integration-based EC techniques. Both (Bernard and Alwan, 2002; Weerackody et al., 2002) integrate the reliability of the channel-decoded features into the recognition process where the Viterbi decoding algorithm is modified such that contributions made by observation probabilities associated with features estimated from erroneous features are decreased. Endo et al. (2003) apply a theory of missing features to error-robust ASR where erroneous features generate constant contributions to the Viterbi decoding with the goal of neutralising the effect from these features.

In addition to these error-robustness methods, ASR systems may benefit from adaptation to the varying quality of network in order to optimise its performance. This has been demonstrated in an experiment that introduces a frame-error-rate (FER) based out-of-vocabulary (OOV) detection method (Tan et al., 2003a).

This paper gives a survey of the robustness issues related to network degradations and presents a number of analyses and experiments with a focus on transmission error robustness.

The paper is organized as follows. Section 2 briefly reviews network-based speech recognition and the ETSI–DSR standards. Section 3 analyses the characteristics of transmission errors, presents a model and categorizes error-robustness techniques. Section 4 presents a number of error detection methods with an emphasis on error detection at different levels. A number of client-based error recovery techniques are investigated in Sections 5 and 6 presents server-based EC techniques. Experimental evaluations are presented in Section 7. Section 8 presents details of the FER-dependent OOV detection method. Concluding remarks are given in Section 9.

2. Speech recognition over networks

The deployment of ASR in networks requires specific attention due to a number of factors, such as the more complicated architecture, the limited resources in the terminals, the bandwidth constraints and the transmission errors. These network-linked issues have focussed the research in network-based ASR on front-end processing for remote speech recognition, on source coding and on channel coding and EC.

2.1. Remote ASR

The integration of ASR into network environments may be implemented as either a terminal- or network-based architecture. Because of their different advantages both architectures are expected to be used in connection with the deployment of future services. An overview and comparison of these architectures is presented in (Viikki, 2001). Since numerous devices already exist and the types and number are still increasing with an incredible rate, the limited resources in the devices and the demand for services with user-friendly interfaces are motivating factors for introducing remote (network-based) ASR. This paper limits its attention to the network-based solutions only. To enable remote speech recognition, data representing speech may be transmitted from the input device to the server as either coded speech or as ASR features, resulting in two types of network-based ASR: server only ASR and DSR.

In the server only ASR approach, the client compresses input speech via conventional speech coders and transmits the coded speech to the server (Kiss, 2000). One way of decoding is that the server re-synthesises the speech, conducts feature-extraction and subsequently performs recognition. The influence of speech coding algorithms on ASR performance, e.g. voice-over-IP (VoIP), Global System for Mobile Communications (GSM) and Universal Mobile Telecommunication System (UMTS) codecs, has been extensively investigated in the literature (e.g. 3GPP TR 26.943, 2004; Besacier et al., 2001; Fingscheidt et al., 2002; Kelleher et al., 2002; Lilly and Paliwal, 1996; Mayorga et al., 2003; Pearce, 2004). As the quality of the re-synthesised speech is highly dependent on the speech coder, a low bit-rate speech coder may cause significant degradations on recognition performance (Haavisto, 1998). Although the deployment of acoustic models trained in matched conditions for each individual coding scheme results in substantial improvement, degradation in ASR performance is still observed at bit-rates below 16 kbps (Euler and Zinke, 1994). Another way is to estimate the feature set directly from the bit-stream of the speech coder without re-synthesizing the coded speech (Huerta, 2000; Huerta and Stern, 1998; Kim and Cox, 2000, 2001; Pelaez-Moreno et al., 2001).

The degradations observed in the above studies have led to the introduction of DSR with the goal of avoiding the degradations from lossy speech compression. In DSR, speech features suitable for recognition are calculated and quantized in the client and transmitted to the server, where they are decoded, followed by appropriate EC and then processed by the recogniser as shown in Fig. 1. To provide reliable communication between the client and the server, the deployment of selected channel coding/decoding schemes together with appropriate EC schemes is needed. This architecture provides a good trade-off between bit-rate and recognition performance (Bernard and Alwan, 2002).

DSR standards have been produced within ETSI in the STQ Aurora DSR working group—often known as Aurora. The first standard for the well-known cepstral features was published in 2000 with the aim of handling the degradations of ASR over mobile channels caused by both lossy speech coding and transmission errors (ETSI ES 201 108, 2000). The goal was also to provide front-end standards to enable interoperability over mobile network (Pearce, 2004). It has been experimentally justified that DSR outperforms adaptive multi-rate (AMR) codecs according to both 'Aurora tests' (Kelleher et al., 2002) and a number of extensive industrial tests organised by 3GPP (3GPP TS 26.235 V6.1.0, 2004). When transmission errors are introduced speech codecs produce even more degradation since those coding algorithms generally have inter-frame dependency (Pearce, 2004). The strength of DSR is that the DSR frames are generally independent and thus more robust to error-prone channels.

2.2. Source coding

The goal of source coding is to compress information for transmission over bandwidth-limited channels. One common class of coding schemes for DSR applies vector quantization (VQ) to ASR features. Split VQ together with scalar quantization used to compress Mel-frequency cepstral coefficients (MFCCs) were evaluated by Digalakis et al. (1999). In the split VQ, each cepstral vector was partitioned into sub-vectors and each sub-vector was independently quantized by using its own codebook. It was concluded that split VQ has lower storage and computational requirements as compared to full VQ and that split VQ performs significantly better than scalar quantization at any bit-rate. By using split VQ and a bit-allocation method, it was found that 2000 bps is sufficient for the transmission of the 13 cepstral coefficients to achieve ASR performance corresponding to unquantized coefficients. The ETSI–DSR standards use a particular form of split VQ.

Another class of source coding is transform coding in which features are transformed to remove the correlation in the features and where quantization is applied in the transformed domain. Transform coding usually gives better performance than quantizing in the original domain. Milner and Shao (2003) study both Karhunen–Loeve and discrete cosine transform (DCT) for the compression of MFCC features. Zhu and Alwan (2001) use a 2D-DCT to exploit inter-frame (temporal) correlations between speech features. Specifically, feature vectors from 12 frames are grouped together to form one block of features, which is then transformed by 2D-DCT. Since the DCT compacts energy into the low-order components, by setting the lowest energy components in each block to zero and quantizing the nonzero components only, a low bit-rate is achieved. Recognition performance is maintained even at 634 bps though at the expense of a block-sized delay.

Paliwal and So (2004) exploited the multi-frame Gaussian mixture model-based block quantizer for

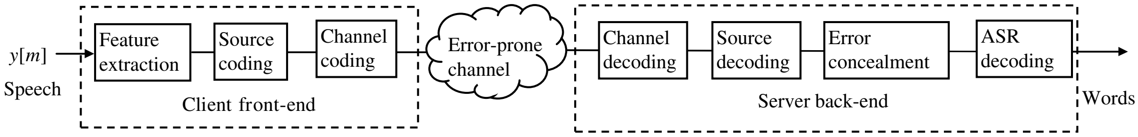

Fig. 1. Block diagram of DSR system.

the coding of MFCC features. The strengths of the block quantizer are computational simplicity and bit-rate scalability.

The inter-frame coding schemes discussed above all exploit the correlation across consecutive MFCC features. This results in the fact that the error in one frame has considerable impact on the quality of the following frames. Due to the removal of the inherent redundancy that exists in speech features, a low bit-rate source coding method is highly sensitive to transmission errors. To some extent, there is a trade-off between the error-resistance and the low bit-rate (achieved by the removal of redundancy)—sometimes called the "no free lunch theorem" (Ho, 1999): coding efficiency multiplied by robustness is constant.

Considerable work has been conducted on investigating how to allocate the bits among e.g. sub-vectors given an overall bit-rate. Digalakis et al. (1999) allocate bits among sub-vectors by using the word-error-rate (WER) as a metric. An iterative bit-allocation method is deployed in (Hsu and Lee, 2004) but driven by syllable error rates. Srinivasamurthy et al. (2004) propose a bit-allocation algorithm based on a mutual information measure, which is superior to the conventional mean square error (MSE) metric. The justification for using the mutual information measure is that the goal in DSR is to ensure minimal degradation in classification performance rather than minimal MSE.

2.3. Channel coding and error concealment

While source coding aims at compressing information, channel coding techniques attempt to protect (detect and/or correct) information from distortions (Bossert, 2000). Channel coding is defined as an error-control technique used for reliable data delivery across error-prone channels by means of adding redundancy to the data (Sklar and Harris, 2004). In the context of DSR, considerable research has been conducted aimed at exploring the potential of channel coding and EC techniques (e.g. Bernard, 2002; Boulis et al., 2002; James and Milner, 2004; Milner, 2001; Peinado et al., 2003; Potamianos and Weerackody, 2001; Tan et al., 2004a). In addition, some work investigates the joint design of source and channel coding (Riskin et al., 2001; Weerackody et al., 2001).

2.4. The ETSI–DSR standards

The ETSI–DSR basic front-end defines the feature-extraction processing together with an encoding scheme (ETSI ES 201 108). The front-end processing produces a 14-element vector consisting of log energy ($\log E$) in addition to 13 MFCCs ranging from c_0 to c_{12}—computed every 10 ms. Each feature vector is compressed using split VQ. The split VQ algorithm groups two features (either $\{c_i$ and c_{i+1}, $i = 1, 3, \ldots, 11\}$ or $\{c_0$ and $\log E\}$) into a feature-pair sub-vector resulting in seven sub-vectors in one vector. Each sub-vector is quantized using its own split VQ codebook. The size of each codebook is 64 (6 bits) for the feature-pair $\{c_i$ and $c_{i+1}\}$ and 256 (8 bits) for $\{c_0$ and $\log E\}$, resulting in a total of 44 bits for each vector.

Two quantized frames—in this paper equivalent to a vector—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. Multi-frames are concatenated into a 4800 bps bit-stream for transmission. The decoding algorithm at the server conducts two calculations to determine whether or not a frame-pair is received with errors, namely a CRC test and a data consistency test. The ETSI standard uses a repetition scheme in its EC processing to replace erroneous vectors.

The ETSI–DSR standard serves as a baseline for presenting various techniques and a platform for comparing a number of experiments in this paper with the baseline results.

A noisy acoustical environment is in the very nature of mobile services. Environmental noise and speaker variation are key issues to be taken into consideration for the success of ASR applications (Rose, 2004). Speech enhancement techniques—used to counteract the detrimental effects of acoustic noise—are mainly applied in the time and frequency domain. However, as features sent to the server-based recognition are commonly cepstral coefficients, the speech enhancement techniques must be applied at the client side. This motivated an update by ETSI to the basic front-end to include noise-robustness techniques, leading to the publication of the advanced front-end (ETSI ES 202 050, 2002).

On the basis of requests for server-side speech reconstruction and for enabling improved tonal language recognition, an additional extended version of the basic front-end was later issued and including fundamental frequency information (ETSI ES 202 211, 2003; Ramabadran et al., 2004; Sorin et al., 2004). The attempt of speech reconstruction in DSR implies a convergence of speech coding and DSR feature-extraction though the difference still exists, namely that the optimisation criterion for speech coding is perceptual quality while it is ASR performance for DSR feature-extraction. The combination of the advanced front-end and the server-side speech reconstruction has resulted in the extended advanced front-end ETSI ES 202 212 (2003). Presently, the DSR extended advanced front-end is selected by the 3rd Generation Partnership Project (3GPP) as the codec for speech enabled services (3GPP TS 26.235). Standards have also been agreed in the Internet Engineering Task Force (IETF) to define the Real-time Transport Protocol (RTP) payload formats for these DSR codecs (Xie and Pearce, 2004).

3. Modelling transmission error degradation

Network-linked degradations are mainly caused by the occurrence of transmission errors. This section analyses the characteristics of transmission errors, presents a model of error degradation and categorizes error-robustness techniques.

3.1. Transmission channel types and simulation

Two types of network connections exist: circuit-switched and packet-switched data channels over which the DSR client and server are interlinked. Transmission errors in connection with packet-switched networks occur in the form of packets that are lost (also known as erasure errors), or packets that are delayed and therefore in a real-time application discarded (Perkins et al., 1998). Packet loss and delay are mainly caused by congestion at routers. Bit errors seldom happen in packet-switched networks. In contrast, bit errors occur more frequently during transmission over circuit-switched mobile networks, resulting in bit errors in the speech data streams. When a speech frame is detected as erroneous and is dropped by the receiver, the situation is equivalent to a packet loss.

Considerable efforts have been spent on investigating and simulating various network degradation phenomena, mainly transmission errors. In general DSR system performance is evaluated either in channel simulations or in real transmission. Channels are simulated in three ways: by adding random errors, by adding burst errors or by using simulated networks. Random errors are intuitively tested in simulations by randomly adding errors to the speech bit-stream or by randomly dropping packet with a given probability. An AWGN (Additive White Gaussian Noise) channel generates random bit errors while a Rayleigh fading channel produces bit errors that occur in bursts. The well-known two-state Gilbert model is widely used to generate error bursts (Gilbert, 1960; Kanal and Sastry, 1978). Two states are used to model error-free transmissions (good) and erroneous transmissions (bad), respectively. Milner and James (2004) suggested using a three-state Markov model. The basic idea of the three-state model is to add the extra state associated with the bad state allowing for some short error-free periods to occur during intervals of error bursts. Tests showed that the three-state model traces real-network packet losses better.

In many investigations the widely used GSM circuit switched channel error patterns (EPs) are applied (Pearce, 2004). The EPs are realistic in the sense that they include a merging of both random and burst errors. The three EPs are EP1, EP2 and EP3 corresponding to a carrier-to-interference (C/I) ratio of 10 dB, 7 dB and 4 dB and corresponding to average bit error rate of 0.0049%, 0.18% and 3.55%, respectively.

The literature references a number of network simulators. Hsu and Lee (2004) use a General Packet Radio Service (GPRS) simulator where a large number of complicated transmission phenomena have been considered such as the propagation model and multi-path fading. Tan et al. (2003a) applied UMTS statistics. The UMTS statistics are provided from a system-level network simulator which is able to simulate a large variety of scenarios and user deployments and is able to extract realistic performance statistics regarding packet error rate or blocking probability.

Besides evaluations in channel simulations, real-world network transmissions have also been tested. Mayorga et al. (2003) have investigated the effect of both packet loss simulated by the Gilbert model and packet loss in real transmission. A strong correlation between WER and packet loss rate has been observed in simulated conditions. In real conditions, however, the same correlation observed in simulated conditions did not occur.

3.2. Modelling transmission error degradation

The traditional model of the degradation of speech signals is depicted in Fig. 2 (Acero, 1993; Huang et al., 2001). The speech signal is corrupted by both linear distortion and additive noise. To compensate for these distortions three categories of noise-robustness techniques are introduced. These are noise-resistant features, speech enhancement and speech model compensation for noise as shown in Fig. 2 (Gong, 1995).

The link between feature-extraction and the ASR-decoding module in the context of DSR is broken by one or more network connections that each introduces additional transmission errors into the data streams. The architectural model of degradations involving both acoustic noise and transmission errors is illustrated in Fig. 3. The figure demonstrates how errors may be introduced into the speech stream and shows which type of error-robustness techniques can be applied.

Note that the general goal of feature-extraction remains to be the estimation of features that are robust to acoustic noise since noise-robustness often is the dominating factor for the degradations in ASR performance (Rose, 2004; Sukkar et al., 2002). The degradation originating from

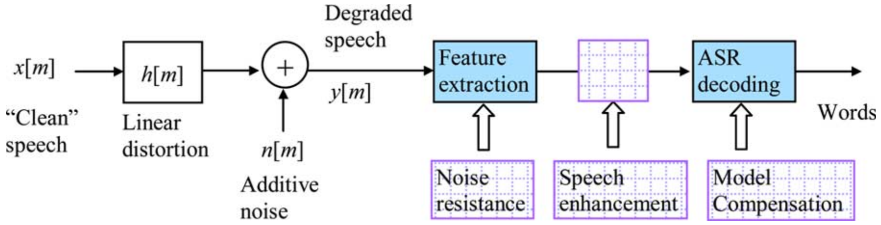

Fig. 2. Architectural model of degradation and robustness techniques.

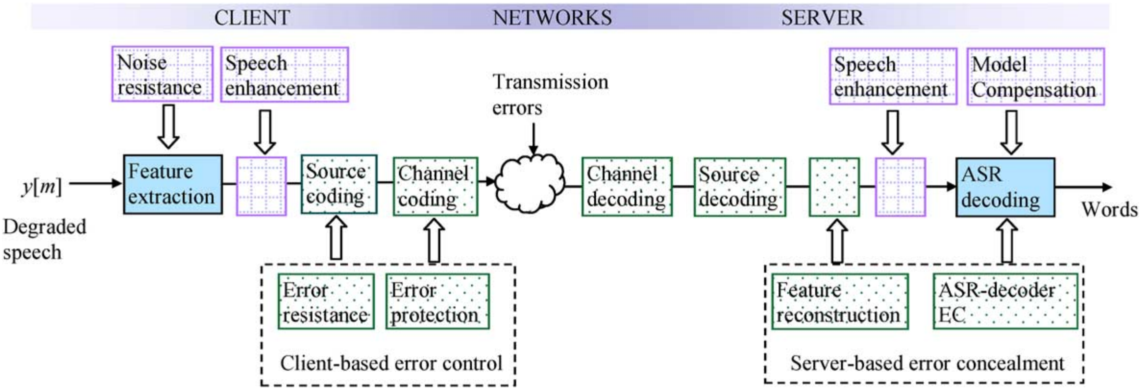

Fig. 3. Architectural model of network degradations and robustness techniques against transmission errors.

transmission errors is handled by a number of techniques such as error control and concealment. As the MFCC features are extensively used and have proved to be successful for ASR (Davis and Mermelstein, 1980), MFCCs are used for most DSR front-ends. One exception of this is the adoption of perceptual linear prediction (PLP) by Bernard and Alwan (2001). The advantage of applying PLP is low bit-rate since PLP coefficients can be quantized at 400 bps (Gunawan and Hasegawa-Johnson, 2001).

Source coding and decoding modules are added to compress speech features to meet bandwidth constraints as shown in Fig. 3. Data transferred over networks are subject to various error sources that cause changes in the stream of speech data either at the bit or the packet level, depending on the individual transmission channel. Channel coding/decoding is deployed with the aim of enabling error detection and recovery and thus providing reliable transmission across networks. Deployment of server-based EC techniques is a necessity in order to reduce the impact of transmission errors still remaining in the speech features.

3.3. Categorization of error-robustness techniques

As compared to acoustic noise, transmission errors have distinctive characteristics and influence speech signals and features differently. Firstly, transmission errors occur at discrete-time frame values whereas the acoustic noise influences the speech signal (and the derived speech features) as a running process. Secondly, transmission errors in general affect the speech data in the cepstral domain whereas the environmental noise affects in the time–frequency domain. Thirdly, transmission errors are introduced into the speech signal after the feature-extraction process which allows for applying methods for client-based error control prior to transmission. These characteristics are the foundations for applying different compensation techniques for handling transmission errors and environmental noise.

Generally, error-robustness techniques in DSR have been developed in three manifestations. Firstly, DSR features are protected by traditional error control and recovery techniques such as FEC, interleaving and joint source and channel coding with the aim of lossless recovery. These techniques require the participation of the client, termed as client-based recovery. Secondly, DSR shares a class of robustness techniques with audio (Perkins et al., 1998) and video (Wang and Zhu, 1998) transmission over networks where feature-reconstruction EC is applied to generate an estimation of the original signal. Thirdly, since the final receiver of DSR features is the ASR-decoder, transmission errors can be further mitigated in the ASR-decoding process through the modification of the decoder, either by marginalisation according to missing data theory (Cooke et al., 2001) or by weighted Viterbi decoding (Yoma et al., 1998). Transmission over networks may cause speech features completely lost or partially corrupted, which makes missing data techniques well suited for DSR to combat transmission errors (Potamianos and

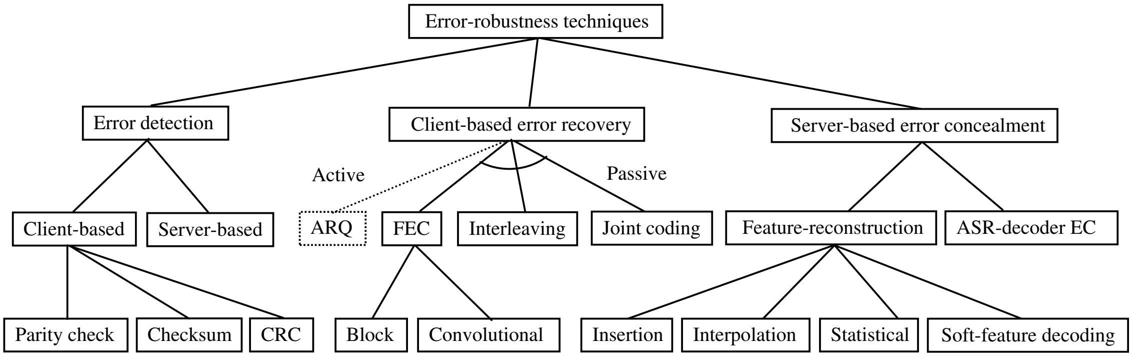

Fig. 4. The taxonomy of error-robustness techniques.

Weerackody, 2001). The taxonomy of these techniques is shown in Fig. 4.

DSR applications have strict end-to-end delay constraints and are generally requested to operate using a RTP protocol (Schulzrinne et al., 2003). Automatic repeat request (ARQ) is an error control mechanism in which a retransmission is requested by the server when an error is detected, resulting in a long delay. ARQ can be applied at various network protocol layers e.g. the transport layer or the application layer. ARQ at transport layer is generally deployed by network operators and it is in most cases not possible to turn this feature off. ARQ at the application layer, however, is not recommended for DSR and passive recovery techniques e.g. interleaving become the preferred client-based methods for correcting the errors. However, ASR can tolerate a certain amount of distortion in the speech features so that server-based EC techniques such as feature-reconstruction and ASR-decoder EC are applicable for concealing any remaining errors. Before any EC techniques can be implemented, error detection methods should first be applied to reveal whether and where a transmission error has been introduced (Wang and Zhu, 1998). Detection can be realised on the basis of either adding redundancy at the client-side or by exploiting the redundancy inherent in the signal itself purely at the server side. These techniques are thoroughly reviewed in the following sections.

4. Error detection

There are two types of error detection methods accomplished, either exploiting added redundancy from the channel coding or exploiting the redundancy of speech features.

In channel coding, redundant bits are added to the data being transmitted and these are subsequently used by the decoder to determine whether or not errors have been conveyed into the data. Two classes of such techniques are used: parity check and block check. Parity check is a simple character-based error detection method that is seldom used today for reliable communications. Two of the most commonly used block check methods are checksum and CRC. In the block check methodology, data are segmented into blocks and an additional check block is appended to each data block at the client. At the server side, the check block information is used to identify whether or not there is an error in the data block. When an error is detected, EC is conducted—as opposed to a retransmission.

For circuit switched networks, the ETSI–DSR standards (Pearce, 2004) apply CRC as the major error detection scheme together with an additional scheme in which a data consistency test exploits the characteristics of speech features. For packet switched networks, the CRC is still kept as part of the payload for two reasons. Firstly, this enables the interoperability with the circuit switched networks as a circuit switched network could potentially be connected to a network gateway that encodes the ETSI–DSR features into the RTP payload. If the features are sent in the RTP payload over a packet switched network, the RTP header information is then used for error detection. Secondly, in the Internet Protocol version 4 (IPv4), IP packets are dropped if there is any error detected in the payload in contrast to the Internet Protocol version 6 (IPv6) in which it may be possible to transport RTP payloads that contain errors and thus making the CRC internal to the ETSI–DSR payload useful for error detection.

In general, error detection by adding header information and/or FEC codes at the client side is more reliable than error detection exploiting redundancy in the signal at the server although at the cost of additional bandwidth (Wang and Zhu, 1998).

Tan et al. (2004c) raise the problem of the size (measured in bits) of a data block for error detection and concealment. Error rates corresponding to the size of a data block are calculated as a function of bit error rates (BER) of random errors according to the following formula

$$\text{Error Rate} = 1 - (1 - \text{BER})^{\text{bits}} \qquad (1)$$

where bits is the number of bits in the data block.

The formula shows that given a BER value, the smaller the number of bits is, the lower is the error rate of the data block. This has motivated the introduction of two methods: one-frame based error protection (Tan and Dalsgaard, 2002) and sub-vector based error detection and concealment (Tan et al., 2004a), which are channel coding based and feature characteristic based method, respectively.

4.1. Frame-pair versus one-frame

Within the ETSI–DSR standards, two quantized frames are grouped together and protected with a 4-bit CRC block together forming a 92-bit frame-pair. This method results in the entire frame-pair being labelled erroneous even if only a single bit error occurs in the frame-pair packet. To overcome this, a one-frame based error protection scheme was deployed to protect each frame by its own 4-bit CRC block which together generates a 48-bit one-frame (Tan and Dalsgaard, 2002; Tan et al., 2003a). The one-frame scheme causes the overall probability of one frame in error to be lower, as shown in Fig. 5 (at the cost of only a marginal increase in bit-rate, from 4800 bps to 5000 bps). The data in Fig. 5(a) are achieved by applying Eq. (1) with a range of BER values of bit errors with Gaussian distribution and the results in Fig. 5(b) come from the calculation on the basis of the GSM EPs where errors occur in bursts.

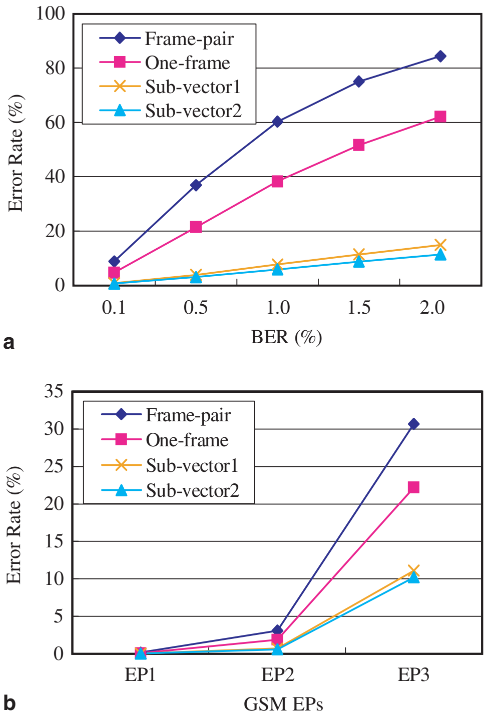

Fig. 5. Percentage error rates of frame-pair, one-frame (vector) and sub-vectors versus different channels. (a) Error rates versus random BER values. (b) Error rates versus GSM EPs.

4.2. Vector versus sub-vector

As one feature vector consists of seven sub-vectors, the error rates of the low bit sub-vectors are significantly lower than both frame-pair and one-frame as shown in Fig. 5. As compared to 92-bit for the frame-pair and 48-bit for the one-frame, sub-vector1 (corresponding to $[c_i, c_{i+1}]$, $i = 1,3,\ldots,11$) and sub-vector2 (corresponding to $[c_0, \log E]$) are represented by 6-bit and 8-bit, respectively. However, since there is no channel coding based error detection applied at the sub-vector level, error detection at this level can only make use of feature characteristics, e.g. by a data consistency test on each pair of the sub-vectors. This is realistic due to the temporal correlation between speech features in consecutive frames caused partly by the vocal tract inertia and partly by the overlapping in the feature-extraction procedure.

Given that n denotes the frame number and $\boldsymbol{V}$ the feature vector, each vector is formatted as

$$\begin{aligned} V^n &= [c_1^n, c_2^n, \ldots, c_{12}^n, c_0^n, \log E^n]^T \\ &= [[c_1^n, c_2^n], \ldots, [c_{11}^n, c_{12}^n], [c_0^n, \log E^n]]^T \\ &= [[\mathrm{S}_0^n]^T, [\mathrm{S}_1^n]^T, \ldots, [\mathrm{S}_6^n]^T]^T \end{aligned} \tag{2}$$

where $\boldsymbol{S}_j^n$ $(j = 0, 1, \ldots, 6)$ denotes the jth sub-vector in frame n (ETSI ES 201 108).

The consistency test is conducted across consecutive frame-pair vectors $[V^n, V^{n+1}]$ such that each sub-vector $\boldsymbol{S}_j^n$ from V^n is compared with its corresponding sub-vector $\boldsymbol{S}_j^{n+1}$ from V^{n+1}. If any of the two decoded features in a feature-pair sub-vector does not possess a minimal continuity criterion, the sub-vector is classified as inconsistent. Specifically both sub-vectors $\boldsymbol{S}_j^n$ and $\boldsymbol{S}_j^{n+1}$ in a frame-pair are classified as inconsistent if

$$\begin{aligned} &(d(S_j^{n+1}(0) - S_j^n(0)) > T_j(0)) \quad \text{or} \\ &(d(S_j^{n+1}(1) - S_j^n(1)) > T_j(1)) \end{aligned} \tag{3}$$

where $d(x,y) = |x - y|$ and $S_j^n(0)$ and $S_j^{n+1}(0)$ and $S_j^n(1)$ and $S_j^{n+1}(1)$ are the first and second element, respectively, in the feature-pair sub-vectors $\boldsymbol{S}_j^n$ and $\boldsymbol{S}_j^{n+1}$ as given in (2); otherwise, they are classified as consistent. The thresholds $T_j(0)$ and $T_j(1)$ are constants given on the basis of measuring the statistics of error-free speech features.

This test generates a consistency matrix that discriminates between consistent and inconsistent sub-vectors. Inconsistent sub-vectors are replaced by their nearest neighbouring consistent sub-vectors whereas the consistent sub-vectors are kept unchanged (Tan et al., 2004a).

5. Error recovery—client-based techniques

Aimed at lossless repair, error recovery techniques require the participation of the client including both source and channel coding. Channel coding such as FEC plays an important role in error recovery. On one hand, the process of both FEC and EC relies on redundant information. FEC techniques add redundancy to the speech signal using a channel code whereas EC techniques exploit the redundancy in the signal itself. On the other hand, source coding removes the redundancy from the speech signal to obtain a high compression rate, resulting in hindrance to the recovery and concealment of errors. One solution to this is to deliberately keep some redundancy in the coded signal to enable better error recovery and concealment (Wang and Zhu, 1998), termed as error-resistant source coding. Another solution is to jointly design source and channel coder. In a broader sense, layer coding (LC) and multiple description coding (MDC) fall into both solutions.

Interleaving is a method that attempts to rearrange burst errors into a set of random errors and by this enabling FEC and EC to become more effective.

5.1. Forward error correction

In using the FEC techniques redundant information is transmitted along with the original data to allow the server to detect and correct errors in the data without any reference to the client (Carle and Biersack, 1997). Two classes of techniques exist: block encoding and convolutional encoding (Bossert, 2000; Sklar and Harris, 2004). Block coding encodes a block of k information bits into a block of n coded bits for transmission and thus the codes are referred to as (n,k) codes. The $(n - k)$ redundant bits determine the error correction capability of the code. Commonly used block codes are Hamming codes, Golay codes, BCH codes and Reed–Solomon codes. Convolutional coding considers the entire stream of data as one single codeword. As a result, encoded data is dependent on not only the current bits but also the previous bits.

A convolutional code is used in combination with unequal error protection (UEP) to protect MFCC features in (Potamianos and Weerackody, 2001). However, block codes are preferred for DSR systems as opposed to convolutional codes due to independency between blocks, smaller delay and lower complexity. The ETSI–DSR standard applies a Golay code to protect the most important information in the data stream e.g. the header information. Bernard and Alwan (2002) use linear block codes mainly for error detection with a limited capacity to conduct bit error correction. Boulis et al. (2002) apply Reed–Solomon codes to achieve graceful degradation of ASR performance over packet-erasure networks. Hsu and Lee (2004) introduce BCH coding into DSR.

5.2. Multiple description coding and layered coding

Unlike most conventional coders, both MDC and LC encode a source into two or more sub-streams that can be delivered on separate channels in order to exploit channel diversity. MDC encodes the signal source into sub-streams (also called descriptions) of equal importance in the sense that each description can independently reproduce the original signal into some basic quality (Goyal, 2001). The quality incrementally increases when more descriptions are received. In contrast, LC generates one base layer stream and several enhancement layer streams. The base layer stream is the most important and can provide a basic level of quality. The enhancement layer streams can refine the quality of the signal reconstructed from the base layer stream but is useless on its own. According to (Wang et al., 2002), MDC is more effective for applications with strict delay constraints. On the other hand, LC is a good choice when retransmission of the base layer or UEP over different channels is feasible.

Kim and Kleijn (2004) observed that MDC generally outperforms Reed–Solomon based FEC. Zhong et al. (2002) compare the popular G.729 standard against a number of MDC schemes for recognizing VoIP and justify the superior performance of the MDC schemes. Aimed at scalable DSR, Srinivasamurthy et al. (2001) present a layered scheme encompassing two layers: the base layer encodes speech using a coarse DPCM (differential pulse code modulation) loop while the enhancement layer encodes the quantization error introduced by the coarse DPCM loop.

5.3. Joint source and channel coding

According to the well-known source–channel separation theorem proposed by Shannon (1948), the source and channel coder can be designed separately without loss of optimality. The assumptions are the property of stationary source and channels and the unlimited complexity and processing delay of the source and channel coder. Since the above assumptions are not true for most real-world applications, there is a need for joint design of source and channel coding that exploits the characteristics of source or source coder for providing better error protection.

In (Weerackody et al., 2001), UEP is applied to speech data by partitioning the data bit stream into classes of different error sensitivity. Riskin et al. (2001) introduce an unequal loss protection (ULP) algorithm to assign unequal amounts of FEC to different sub-vectors to minimise WER.

5.4. Interleaving

FEC and EC techniques have good efficiency in counteracting errors randomly distributed in the data stream but fail to manage burst errors. Interleaving techniques have therefore been broadly applied in communication systems as an optional addition to error correction codes to counteract the effect of burst errors at the cost of delay. On the basis of interleaving, channel coding is confronted with only a set of random errors that are converted from burst errors.

Specifically, interleaving is a method that rearranges the ordering of a sequence of code symbols in order to spread burst errors over multiple codewords for efficient error recovery and concealment (Ramsey, 1970). At the server, the counterpart de-interleaving restores the reordered sequence to its original order. A common way to implement interleaving is to divide symbol sequences into blocks corresponding to a two-dimensional array, and to read symbols in by rows and out by columns. Extensive work has been done by James and Milner (2004) to deploy interleaving in DSR.

5.5. Discussion of client-based techniques

Although the principles of the client-based techniques reviewed in this section are different from each other, they share a common attribute namely the participation of the client with the aim of exploiting the characteristics of channels and signals. The deployment of client-based techniques is always a trade-off between the achieved performance and the required resources. For example, FEC trades bandwidth for redundancy, MDC trades multiple channels for uncorrelated transmission errors among descriptions, and interleaving trades delay for random distribution of errors. Therefore, the employment of client-based techniques is highly dependent on networks and applications. One disadvantage of client-based techniques is their weak compatibility. Further discussions and comparisons are presented in Section 7.

6. Error concealment—server-based techniques

Lossless error recovery is required in data transmission as even a single bit error may cause the entire data block to be discarded. In contrast, a certain amount of distortions in the speech features can be tolerated by the ASR-decoder. This fact makes EC a feasible method to complement client-based error recovery techniques to mitigate the effect of remaining transmission errors without the request for a retransmission. EC generally deploys the strong temporal correlation residing in speech features and utilises the statistical information about speech.

An EC scheme often relies on a set of error-free features received before and/or after erroneous or lost features: either one- or two-sided EC schemes. In real-time speech and audio streaming, one-sided schemes are often used since discontinuities in a re-synthesised signal may perceptually annoy the user, especially during burst intervals. In DSR applications, however, the discontinuities do not disturb the ASR engine although it may increase the end-to-end delay. Therefore, two-sided EC techniques are favoured in DSR because of its significantly superior performance.

The aim of EC is in general to create a substitution for a lost or erroneous packet as close to the original as possible. This type of concealment technique is termed as feature-reconstruction EC. In applications like voice transmission, the source and the sink of the communication channel are the voice and the human ears, respectively. In the context of DSR, however, the source is the speech features and the sink the ASR-decoder. Therefore, EC may be conducted during the recognition decoding process as well, which is unique for DSR. Specifically, the ASR-decoder may be modified to handle degradations introduced by transmission errors, termed as ASR-decoder EC.

6.1. Insertion-based techniques

Insertion-based EC techniques refer to a class of simple techniques that reconstruct lost packets without taking the signal characteristics into consideration (Perkins et al., 1998). An erroneous frame is substituted by inserting silence, noise, an estimated value (for example a mean value over training data), or a repetition of a neighbouring frame.

In applying splicing a number of consecutive erroneous frames are simply dropped. A side effect of employing splicing is a decrease in the Viterbi decoding time caused by the shorter feature stream (Kim and Cox, 2001). Boulis et al. (2002) reported that the mean-value substitution (estimated over all training data) outperforms splicing and silence substitution.

The ETSI–DSR standard applies a repetition where the first half of a series of erroneous frames is replaced with a copy of the last correct frame before the error and the second half with a copy of the first correct frame following the error.

In the partial splicing scheme presented in (Tan et al., 2003b) erroneous frames are partly substituted by a repetition of neighbouring frames and partly by a splicing. It can be shown that partial splicing under certain assumptions is equivalent to a weighted Viterbi decoding algorithm.

On the basis of error detection at the sub-vector level as presented in Section 4.2, the sub-vector EC (Tan et al., 2004) is considered as a repetition EC at the sub-vector level.

6.2. Interpolation-based techniques

Interpolation accounts for the changing characteristics of the signal and particularly exploits the temporal correlation that is present in the speech feature stream to aid the reconstruction of speech features.

The most commonly used interpolation technique is applying a polynomial interpolation as an estimate of the erroneous frames (Milner and Semnani, 2000). For DSR applications, repetition has been experimentally justified to perform better than linear interpolation (Pearce, 2004; Peinado et al., 2003; Tan et al., 2003b). Tan et al. (2004b) further conduct a comparative study with the aim of revealing the causes of this, which justifies a difference existing between traditional signal-reconstruction and feature-reconstruction for ASR. James and Milner (2004) propose to use cubic interpolation instead of linear interpolation which shows better performance than both repetition and linear interpolation.

6.3. Statistical-based techniques

Neither insertion- nor interpolation-based techniques use a priori information about the speech features though the mean-value substitution replaces lost packets by the mean estimated over all training data. A number of recently developed techniques deploy the statistical knowledge about speech source by introducing such knowledge into the concealment process.

Statistical-based techniques use a priori knowledge of the full data to estimate the missing data (Ramakrishana, 2000). In particular, the maximum a posteriori (MAP) estimation is a technique that estimates the missing data with the aim of maximising their likelihood conditioned on the observed data and the distribution of the full data. On the basis of the MAP technique, James et al. (2004) reconstruct lost speech vectors by employing both the correctly received vectors and the statistical information such as mean and variance calculated from a set of training utterances (assuming a Gaussian distribution). The method outperforms cubic interpolation in particular when the packet loss rate is high. MAP estimation, however, is in general computationally expensive due to its need of inversing large covariance matrices.

In (Gomez et al., 2003), a data-source model mitigation technique is presented for DSR over lossy packet channels. The technique models the data-source through transition probabilities from a sequence of quantized sub-vectors to another sequence. For example, in the first order data-source model, a set of comprehensive tables containing every combination of two split VQ indices are built for forward estimation and for backward estimation, respectively. Tables for forward estimation are constructed by searching each index in the training database and averaging the sequences of indices following it while tables for backward estimation by searching each index in the training database and averaging the sequences of indices previous to it. Sequences in lost packets are reconstructed by means of the trained data-source model and the received sequences preceeding and following the lost packets. Performance improvement has been observed particularly for high packet loss rate as compared to repetition EC. Similarly, Lee et al. (2004) proposed an *N*-gram model approach for packet loss concealment in a VoIP application. This work showed that trigram predictive models consistently outperform the repetition-based method in terms of distortion. Both of the above referenced techniques have low computational cost but high memory requirements.

The MAP estimation has a high computational complexity whilst the memory requirements of the data-source technique are high. Gomez et al. (2004) combine the data-source model technique and the MAP estimation technique to offer a trade-off between memory and computational resources requirement.

6.4. Soft-feature decoding based techniques

In wireless communications, a number of studies exploit the information about the reliability of the received bits. Specific channels and channel coding/decoding algorithms are often specified so that soft-decision of channel decoding is applicable (Bernard, 2002; Haeb-Umbach and Ion, 2004; Peinado et al., 2003; Potamianos and Weerackody, 2001). The reliability information is used either for feature-reconstruction or in combination with weighted Viterbi decoding that takes this information into account during the ASR-decoding process.

A minimum mean square error (MMSE) estimation and hidden Markov model (HMM) based EC are proposed in (Peinado et al., 2001, 2003, 2005). Applied for EC in speech coding (Fingscheidt and Vary, 2001), MMSE estimation models the speech source as a Markov process to exploit the correlation between consecutive frames. Peinado et al. (2003) first deploy MMSE based on soft-feature decoding. Secondly significant improvement is further obtained by considering previously received vectors for the estimation of the current feature vector, resulting in the HMM model based EC.

6.5. ASR-decoder based techniques

Speech features are, after transmission over networks, subject to being missing or unreliable. In addition to reconstruction of these features, the ASR-decoder can provide complementary means to handle the detriment in speech features by integrating the reliability of the channel-decoded features into the recognition process. Two well-known noise-robustness techniques match this purpose: namely marginalisation used in missing data theory (Cooke et al., 2001) and soft-decoding (weighted Viterbi decoding when HMM is applied) (Yoma et al., 1998). As compared with ASR in noisy environments, where identifying the reliabilities of spectral features is difficult, the advantages in this context are that missing features or the reliability of each feature is known from the channel decoding (Potamianos and Weerackody, 2001) and that the information is available in the cepstral domain.

In (Endo et al., 2003), missing data theory is applied such that erroneous features generate constant contributions to the Viterbi decoding with the aim of neutralising these features. James et al. (2004a) show that missing data technique is superior to a number of feature-reconstruction methods. Although both marginalisation and splicing do not utilise unreliable features, marginalisation reserves the time information—since HMM state transitions are possible in the intervals of unreliable features—so that much better performance is obtained (Bernard, 2002).

In weighted Viterbi decoding, exponential weighting factors are introduced into the calculation of the likelihood based on the probability of the speech observations such that contributions made by observation probabilities are decreased or neutralised if the features are estimated from erroneous frames.

The weighting factor may be computed from either the reliability measure of the received bits—when available—or from an estimation value when the hard-decision channel coding is applied. The first method requires the evaluation of the reliability of the decoding feature from soft-decision channel coding i.e. assuming a known bit probability (Bernard and Alwan, 2002; Weerackody et al., 2002). The second method is applicable to a wider range of channels including channels characterised by packet loss so that the range of channel conditions are extended from wireless to IP-based (Bernard and Alwan, 2002). Cardenal-Lopez et al. (2004) compare constant weighting factor with time varying weighting factor to cope with the fact that the longer the burst is the less effective is the repetition technique.

6.6. Discussion of server-based techniques

The following may be concluded on the five classes of server-based EC techniques reviewed in this section. One of the advantages of server-based EC is that there is no requirement for modifying the client-side of DSR, signifying the compatibility with the existing ETSI–DSR standards. Insertion- and interpolation-based techniques are traditional techniques widely used in many applications such as audio and video transmission. Among them, repetition EC has shown good performance with low complexity. The statistical-based techniques take advantage of a priori knowledge of speech features and show slightly better performance than repetition, however, at the expense of either high computational cost or high memory requirement. Soft-feature decoding based techniques achieve highest performance but generally at a high computational cost. Finally ASR-decoder based techniques are unique for DSR and can be applied in combination with other EC. Detailed performance comparisons and discussions are presented in the next section.

7. Performance evaluation

In the literature, various techniques have been developed but evaluated on the basis of a number of different speech databases and different channel simulations. This makes the comparison of these techniques difficult. In this evaluation, the same database and the same channel condition are used in order to effectively compare the performance of some of the techniques outlined above.

7.1. Experimental settings

The Aurora 2 database (Pearce and Hirsch, 2000) has been selected for this purpose. The 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 using the HTK recogniser. Each of the digit whole word models has 16 HMM states with three Gaussian mixtures per state. The silence model has three HMM states with six Gaussian mixtures per state. A one-state short pause model is tied to the second state of the silence model. The word models used in the experiments are trained on clean speech, no acoustic noise is added to the test data and transmission errors are the only cause of the decrease in the server-side recognition performance.

The three GSM EPs are commonly used for the error-robustness evaluation of speech coding algorithms and DSR schemes. As EP1 and EP2 generally do not cause noticeable performance degradation (Tan et al., 2003a), EP3 is specifically chosen for this evaluation. It should be noticed that the experiments conducted here are using this error pattern for circuit switched GSM channels. In case of packet switched channels such as the GSM GPRS (General Packet Radio Service), the lower levels of the protocol stack allow for retransmission and transmission errors will occur as packet loss and at a lower rate.

7.2. Experimental results and discussions

A number of techniques are tested ranging from client-based to server-based techniques. The tested client-based techniques include Reed–Solomon based FEC, interleaving, MDC and the one-frame scheme presented in Section 4.1, which all are used in combination with repetition EC. The implemented Reed–Solomon code is RS(32, 16) with 8-bit symbols where 16 information symbols are encoded into 32 coded symbols, indicating a capability of correcting 8 symbol errors or 16 symbol erasures in the code word. Two interleaving schemes are applied: Interleaving12 in which a sequence of 12 vectors is grouped into one block and Interleaving24 where a sequence of 24 vectors is grouped. Interleaving is implemented simply by reading odd-numbered features first and even-numbered features second from the blocks. As a result, Interleaving12 has 5 vectors or 50 ms (with a 10 ms frame shift) maximum delay and Interleaving24 has 110 ms maximum delay. In applying MDC, two descriptions are generated namely the odd-numbered and the even-numbered. Each description is encoded into 2600 bps comprising 2200 bps speech data, 200 bps head information and 200 bps CRC information. The two description encodings are transmitted over two uncorrelated channels which both are simulated by EP3.

The evaluated feature-reconstruction techniques encompass repetition (Aurora baseline), linear interpolation, splicing and sub-vector EC. The scheme without CRC error detection (No CRC) is also evaluated. In this case, transmission errors remain in the speech features and are passed through to the ASR-decoder. The result from error-free transmission is shown as well. The results for statistical-based techniques are cited from (Peinado et al., 2003) in which H-FBMMSE and H-MAP represent forward–backward MMSE with hard decisions and MAP with hard decisions, respectively. The detailed WER results for Test Set A are shown in Fig. 6.

A more detailed comparison is presented in Table 1 in terms of WER, bandwidth requirement and computational complexity. It is observed that the performance in applying MDC approaches the error-free channel, but at the additional requirement of multiple channels. H-FBMMSE and H-MAP both provide very low WER values but at the cost of very high computational complexity, indicating that they may not be applicable for real-time applications without the reduction of computation. The interleaving schemes achieve good performance, however at the expense of adding delay. As compared to the above techniques, sub-vector EC gives lower performance but it neither introduce extra complexity nor resource requirement. The one-frame scheme shows superior performance to the Aurora frame-pair scheme with the introduction of a marginal increase in bandwidth. RS(32, 16) gives a performance close

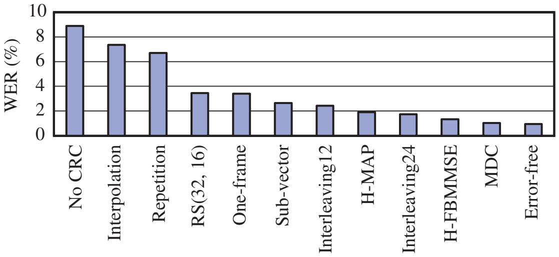

Fig. 6. WER (%) across the error-robustness techniques for EP3 for Test Set A.

Table 1
Performance comparison of some error-robustness techniques for EP3 for Test Set A

	WER (%)	Bit-rate (bps)	Complexity	Compatibility with ETSI–DSR standards
Splicing	24.00	4800	Low	Yes
No CRC	8.88	4600	Low	No
Linear interpolation	7.35	4800	Low	Yes
Repetition (Aurora)	6.70	4800	Low	Yes
Marginalisation	–	4800	Low	Yes
Weighted Viterbi	–	4800	Low	Yes
RS(32,16)	3.45	9600	High	No
One-frame	3.41	5000	Low	No
Sub-vector	2.65	4800	Low	Yes
Interleaving12	2.43	4800	Low	No
H-MAP	1.91	4800	High	Yes
Interleaving24	1.74	4800	Low	No
H-FBMMSE	1.34	4800	High	Yes
MDC	1.04	5200	Low	No
Error-free	0.95	4800	–	–

to the one-frame scheme but requiring more bandwidth and higher computation. It is noticed that Reed–Solomon based FEC aims at correcting errors whilst the one-frame scheme just increases the capability of detecting errors. This justifies that for DSR applications, channel coding should focus on error detection rather than error correction as also observed in (Bernard and Alwan, 2002).

With regard to the ASR-decoder EC, Endo et al. (2003) show that marginalisation offers superior performance to linear interpolation. Bernard (2002) demonstrates that repetition is superior to marginalisation for random packet loss conditions while marginalisation may outperform repetition when the average burst lengths are large. Repetition in combination with weighted Viterbi gives better performance than both repetition alone and marginalisation for all conditions. In accordance with these references only, marginalisation and a weighted Viterbi technique are included and ranked in Table 1 for performance comparisons.

The experiments show that linear interpolation gives lower ASR performance than repetition as discussed in Section 6.2. In the case of No CRC, no compensation (i.e. EC) is conducted so that erroneous features are fed directly to the ASR-decoder, resulting in reduced ASR performance. Splicing as described in Section 6.1—which is equivalent to no compensation when packet losses occur—gives the lowest performance and therefore is not an applicable technique. It should be noted that the client-based techniques including Reed–Solomon, one-frame, No CRC, interleaving and MDC are not compatible with the existing ETSI–DSR standards while all other techniques compared in Table 1 are server-based and therefore can be used at the server side without requiring modifications of the standards.

It is generally verified by the experiments that ASR performance can be improved by introducing a number of error recovery and concealment techniques. Depending on the techniques that are applied, for transmission over severe error-prone channels—as demonstrated by applying EP3—the degradation in recognition performance is still a matter of concern as compared to the baseline performance with no transmission errors. The following section focuses on the design of a potential method by which it is possible to modify the processing of the recogniser to further raise the overall performance of the ASR system.

8. Recogniser adaptation to transmission quality

The techniques as reviewed in the preceeding sections are designed with the goal of maintaining maximal ASR performance. This section presents research on adaptation of the server-side recogniser to the highly varying quality of the network in order to further optimise the overall performance.

One example of such optimisation is the introduction of frame-error-rate (FER) based out-of-vocabulary (OOV) detection (Tan et al., 2003a). The method is based on the observation that transmission errors influence the acoustic likelihood and thus affect the optimal threshold setting for discrimination between in-vocabulary words and OOV words.

In this section experiments are based on the basic ETSI–DSR standard. The selected database used for both training and testing is the Danish SpeechDat 2 database DA-FDB 4000 which covers speech from 4000 Danish speakers collected over the fixed network. A part of the database is used for the training of tri-phone models and one filler model, each having three HMM states, and each state having a mixture of 32 Gaussians. The independent test data consist of isolated Danish digits (11 words including two different pronunciations of the Danish digit '1') used as in-vocabulary words and city names (449 words) used as OOV words. The recogniser applied is the HTK-based SpeechDat/COST249 reference recogniser (Lindberg et al., 2000). The baseline WER (no transmission errors) for the Danish digits is 0.2%.

8.1. Effect of errors on likelihood ratio distribution

OOV detection is a statistical hypothesis testing problem in which a decision algorithm accepts or rejects the hypothesis (e.g. Rahim et al., 1997; Lleida and Rose, 2000). Given a speech signal observation sequence $\boldsymbol{O}$, the algorithm tests the null hypothesis H_0 against the alternative hypothesis H_1. H_0 represents one of the in-vocabulary words and H_1 represents OOV words modelled by one filler model. A likelihood ratio LR($\boldsymbol{O}$) based on the null and alternative hypotheses is used to detect OOV words. The test rejects the H_0 hypothesis if

$$\mathrm{LR}(\boldsymbol{O}) = \frac{p(\boldsymbol{O}|H_0)}{p(\boldsymbol{O}|H_1)} < T \tag{4}$$

where T is the threshold of the test. $p(\boldsymbol{O}|H_0)$ and $p(\boldsymbol{O}|H_1)$ are the probabilities of the H_0 and the H_1 hypotheses, respectively.

Transmission errors may, however adversely affect the likelihood ratios of both the in-vocabulary words and OOV words. Fig. 7 shows the probability density functions (PDFs) of the log-likelihood ratios of the in-vocabulary words and OOV words from the experiments for error-free channel and channel with 2% BER value.

The figure shows that the occurrence of transmission errors changes the PDFs of the log-likelihood ratios in two ways. Firstly, the standard deviations of the distributions are increased for increasing BER values. This has the effect of weakening the discrimination between in-vocabulary and OOV words. Secondly, the shifting in the mean value of the distributions affects the optimal threshold setting for OOV detection. A fixed threshold setting—as normally used in the context of error-free transmission—may therefore fail to maintain the balance of the false rejection and false acceptance rates.

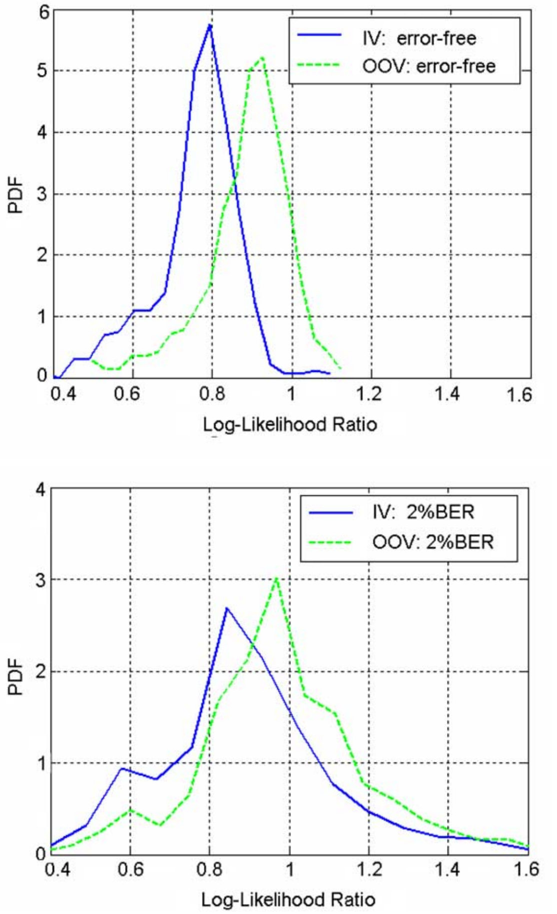

Fig. 7. PDFs of the log-likelihood ratios for in-vocabulary and OOV words.

8.2. FER-dependent threshold for OOV detection

A potential way of maintaining the balance is to adjust the threshold in accordance with the FER that is a measure of the instantaneous transmission error rate. The FER is calculated on the basis of the CRC information in the data stream. This results in a FER-dependent threshold that optimises the OOV detection.

The threshold is modelled as a fourth-order polynomial function of the FER. The FER values are calculated from the BER values according to Eq. (1). To estimate the coefficients of the polynomial, five experiments (with BER values ranging from 0.1% to 2%) were conducted using a development database consisting of 282 digit utterances and 249 city names utterances. The thresholds for each of these experiments are chosen with the specifically chosen optimisation target of maintaining the false rejection rate approximately constant across a range of BER values.

The test data for the experiments described below are the remaining 200 digits and 200 city names utterances from the same database. During test, the FER is estimated by using the CRC error detection and then used for adjusting the threshold of the OOV detection based on the fourth-order polynomial function.

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

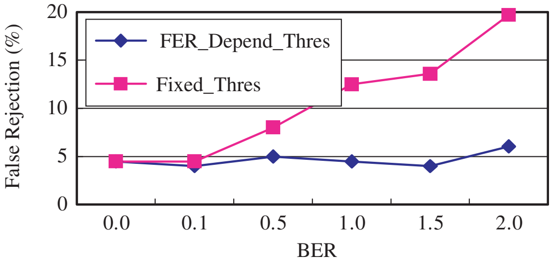

Fig. 8. False rejection rate versus BER values.

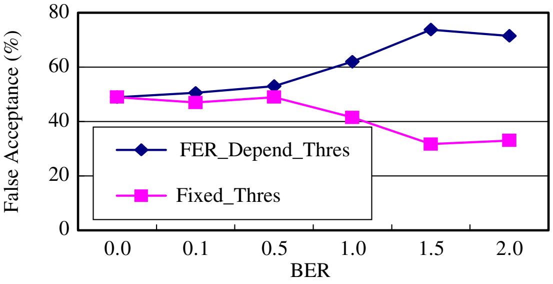

Fig. 9. False acceptance rate versus BER values.

By maintaining an almost constant false rejection rate, the false acceptance rate increases as shown in Fig. 9. In general, however threshold setting in general 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-dependent OOV detection.

The experimental results shown above justify that it is feasible to adapt the behaviour of the back-end recogniser and that it is possible to further improve the overall ASR performance according to the varying quality of the network in question.

9. Conclusion

In this paper, the developments and trends of incorporating ASR technology into wireless networks have been reviewed. Emphasis has been placed on robustness techniques against transmission errors enforced by error-prone communication channels. Three classes of techniques have been presented in detail namely error detection, client-based error recovery and server-based error concealment. Error detection can be accomplished either by adding redundancy at the client-side or by exploiting the redundancy inherent in the signal itself purely at the server side. It has been pointed out that it is important to identify the proper size of a data block for error detection and the following concealment.

For transmission over severely error-prone channels, deployment of client-based error recovery techniques is of importance in order to achieve high ASR performance. The deployment of client-based techniques will distinguish DSR systems from one another, e.g. by not being compatible with the existing ETSI–DSR standards. Based on the work conducted in this paper the following comments apply:

- Although FEC protection is essential to e.g. the head information in a DSR stream, FEC such as Reed–Solomon code is not effective for protecting speech features as the overhead introduced by FEC is superfluous for error-free channels and not useful for channels with burst errors.

- Since the ASR-decoder can tolerate a certain amount of distortions in the speech features, especially if these are caused by independent errors, it is a significant advantage to convert burst errors into independent (random) errors. This makes the deployment of interleaving techniques attractive, although at the expense of inherent delay.
- Another technique being able to circumvent burst errors is MDC. When multiple channels are available, MDC is highly recommended due to its excellent performance.

As pointed out in this review, it is not possible to recover from all transmission errors on the basis of deploying client-based techniques only. Server-based EC techniques are therefore employed for handling the remaining errors. On the basis of the experiences learned from the experiments, the following overall comments are presented for server-based EC techniques:

- For circuit-switched channels where errors occur at the bit level, error-free information is potentially available within the erroneous frames (vectors) for the EC process, and it is therefore strongly recommended to exploit the remaining error-free information within the erroneous vectors. The sub-vector base EC scheme and the MMSE estimation are examples of such techniques where the experiments have shown superior performance to conventional techniques.
- Statistical-based techniques that utilise a priori information about the speech signal show improved performance although at the expense of large memory requirement and/or high computational complexity.
- In addition to feature-reconstruction techniques, ASR-decoder based EC techniques are unique for DSR and provide good performance.
- The applicability of server-based EC is dependent on the trade-off between the achieved performance and the computational complexity. Some of these techniques may not be practical for real-time applications because of their inherent processing delay and computational cost. For future research, combinations of these techniques are relevant to pursue.

In general, each technique has its own strengths and weaknesses, so the selection of techniques to be deployed is dependent on the expected channel characteristics and the system requirements.

In addition to robustness techniques, the introduction of adaptation schemes may be worthwhile to exploit due to the dynamic nature of network transmission. Adaptation may be implemented in two ways. Firstly, the schemes chosen for source coding/decoding, channel coding/decoding and EC may be adaptive in order to optimise the trade-off between the required network and computational resources and the achieved performance. Secondly, the recogniser and spoken language modules can be made adjustable to the quality of network to obtain optimal overall performance. Preliminary experiments have verified this adaptation concept for OOV detection task.

Acknowledgments

The authors appreciate the insightful comments and suggestions from the reviewers, which have significantly improved the presentation of the paper.

References

Acero, A., 1993. Acoustical and Environmental Robustness in Automatic Speech Recognition. Kluwer Academic Publishers, Boston.

Bernard, A., 2002. Source and channel coding for speech transmission and remote speech recognition, PhD dissertation, University of California, Los Angeles, 2002.

Bernard, A., Alwan, A., 2001. Source and channel coding for remote speech recognition over error-prone channel. In: Proc. ICASSP01, USA, May 2001.

Bernard, A., Alwan, A., 2002. Low-bitrate distributed speech recognition for packet-based and wireless communication. IEEE Trans. Speech Audio Process. 10 (8), 570–579.

Besacier, L., Bergamini, C., Vaufreydaz, D., Castelli, E., 2001. The effect of speech and audio compression on speech recognition performance. In: Proc. IEEE Multimedia Signal Processing Workshop, 2001.

Boulis, C., Ostendorf, M., Riskin, E.A., Otterson, S., 2002. Gracefully degradation of speech recognition performance over packet-erasure networks. IEEE Trans. Speech Audio Process. 10 (8), 580–590.

Bossert, M., 2000. Channel Coding for Telecommunications. John Wiley & Sons.

Cardenal-Lopez, A., Docio-Fernandez, L., Garcia-Mateo, C., 2004. Soft decoding strategies for distributed speech recognition over IP networks. In: Proc. ICASSP04.

Carle, G., Biersack, E.W., 1997. Survey of error recovery techniques for IP-based audio–visual multicast applications. IEEE Network 11 (6), 24–36.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun. 34, 267–285.

Cox, R.V., Kamm, C.A., Rabiner, L.R., et al., 2000. Speech and language processing for next-millennium communications services. Proc. IEEE 88 (8), 1314–1337.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics Speech Signal Process. 28 (4), 357–366.

Digalakis, V., Neumeyer, L., Perakakis, M., 1999. Quantization of cepstral parameters for speech recognition over the World Wide Web. IEEE J. Select. Areas Commun. 17, 82–90.

Endo, T., Kuroiwa, S., Nakamura, S., 2003. Missing feature theory applied to robust speech recognition over IP networks. In: Proc. Eurospeech03, Geneva, Switzerland.

ETSI Standard ES 201 108, 2000. Distributed speech recognition; front-end feature extraction algorithm; compression algorithms.

ETSI Standard ES 202 050, 2002. Distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm.

ETSI Standard ES 202 211, 2003. Distributed speech recognition; extended front-end feature extraction algorithm; compression algorithm, back-end speech reconstruction algorithm.

ETSI Standard ES 202 212, 2003. Distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithm, back-end speech reconstruction algorithm.

Euler, S., Zinke, J., 1994. The influence of speech coding algorithms on automatic speech recognition. In: Proc. ICASSP94.

Fingscheidt, T., Vary, P., 2001. Softbit speech decoding: a new approach to error concealment. IEEE Trans. Speech Audio Process. 9 (3), 1–11.

Fingscheidt, T., Aalburg, S., Stan, S., Beaugeant, C., 2002. Network-based vs. distributed speech recognition in adaptive multi-rate wireless systems. In: Proc. ICSLP02.

Gilbert, E.N., 1960. Capacity of a burst-noise channel. Bell Syst. Tech. J. 39, 1253–1266.

Gomez, A.M., Peinado, A.M., Sanchez, V., Rubio, A.J., 2003. A source model mitigation technique for distributed speech recognition over lossy packet channels. In: Proc. Eurospeech03.

Gomez, A.M., Peinado, A.M., Sanchez, V., Miner, B.P., Rubio, A.J., 2004. Statistical-based reconstruction methods for speech recognition in IP networks. In: Proc. Robust2004.

Gong, Y., 1995. Speech recognition in noisy environments: a survey. Speech Commun. 16, 261–291.

Goyal, V.K., 2001. Multiple description coding: compression meets the network. IEEE Signal Process. Mag. 18 (5), 74–93.

3GPP TR 26.943 V6.0.0, 2004. Recognition performance evaluations of codecs for speech enabled services (SES) (Release 6).

3GPP TS 26.235 V6.1.0, 2004. Packet switched conversational multimedia applications; default codecs (Release 6).

Gunawan, W., Hasegawa-Johnson, 2001. PLP coefficients can be quantized at 400 bps. In: Proc. ICASSP.

Haavisto, P., 1998. Audio–visual signal processing for mobile communications. In: Proc. European Signal Processing Conference, Island of Rhodes, Greece.

Haavisto, P., 1999. Speech recognition for mobile communications. In: Proc. Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland.

Haeb-Umbach, R., Ion, V., 2004. Soft features for improved distributed speech recognition over wireless networks. In: Proc. ICSLP04.

Ho, Y.-C., 1999. The no free lunch theorem and the human-machine interface. IEEE Control Syst. 1999 (June), 8–10.

Huerta, J., 2000. Robust Speech Recognition in GSM Codec Environments, PhD thesis, CMU.

Huerta, J., Stern, R., 1998. Speech recognition from GSM codec parameters. In: Proc. ICSLP98.

Hsu, W.-H., Lee, L.-S., 2004. Efficient and robust distributed speech recognition (DSR) over wireless fading channels: 2D-DCT compression, iterative bit allocation, short BCH code and interleaving. In: Proc. ICASSP04.

Huang, X.D., Acero, A., Hon, H.-W., 2001. Spoken Language Processing. Prentice Hall, New Jersey, USA.

James, A.B., Milner, B.P., 2004. An analysis of interleavers for robust speech recognition in burst-like packet loss. In: Proc. ICASSP04, Montreal, Quebec, Canada.

James, A.B., Gomez, A., Milner, B.P., 2004. A comparison of packet loss compensation methods and interleaving for speech recognition in burst-like packet loss. In: Proc. ICSLP04.

Kanal, L.N., Sastry, A.R.K., 1978. Models for channels with memory and their applications to error control. Proc. IEEE 66 (7), 724–744.

Kelleher, H., Pearce, D., Ealey, D., Mauuary, L., 2002. Speech recognition performance comparison between DSR and AMR transcoded speech. In: Proc. ICSLP02, Denver, USA.

Kim, H.K., Cox, R.V., 2000. Bitstream-based feature extraction for wireless speech recognition. In: Proc. ICASSP00, Turkey.

Kim, H.K., Cox, R.V., 2001. A bitstream-based front-end for wireless speech recognition on IS-136 communications system. IEEE Trans. Speech Audio Process. 9, 558–568.

Kim, M.Y., Kleijn, E.B., 2004. Comparison of transmitter-based packet-loss recovery techniques for voice transmission. In: Proc. ICSLP04.

Kiss, I., 2000. A comparison of distributed and network speech recognition for mobile communication systems. In: Proc. ICSLP00, Beijing, China.

Lee, M., Zitouni, I., Zhou, Q., 2004. On a N-gram model approach for packet loss concealment. In: Proc. ICSLP04.

Lee, L.-S., Lee, Y., 2001. Voice access of global information for broad-band wireless: technologies of today and challenges of tomorrow. Proc. IEEE 89 (1), 41–57.

Lilly, B.T., Paliwal, K.K., 1996. Effect of speech coders on speech recognition performance. In: Proc. ICSLP96.

Lindberg, B., Johansen, F.T., Warakagoda, N., et al., 2000. A noise robust multilingual reference recogniser based on SpeechDat(II). In: Proc. ICSLP00.

Lleida, E., Rose, R.C., 2000. Utterance verification in continuous speech recognition: decoding and training procedures. IEEE Trans. Speech Audio Process. 8 (2), 126–139.

Mayorga, P., Besacier, L., Lamy, R., Serignat, J.-F., 2003. Audio packet loss over IP and speech recognition. In: Proc. ASRU03, Virgin Islands, USA.

Milner, B., Semnani, S., 2000. Robust speech recognition over IP networks. In: Proc. ICASSP00, Turkey.

Milner, B., 2001. Robust speech recognition in burst-like packet loss. In: Proc. ICASSP01, USA.

Milner, B.P., James, A.B., 2004. An analysis of packet loss models for distributed speech recognition. In: Proc. ICSLP04.

Milner, B.P., Shao, X., 2003. Low bit-rate feature vector compression using transform coding and non-uniform bit allocation. In: Proc. ICASSP03.

Paliwal, K.K., So, S., 2004. Scalable distributed speech recognition using multi-frame GMM-based block quantization. In: Proc. ICSLP04.

Pearce, D., 2000. Enabling new speech driven services for mobile devices: an overview of the ETSI standards activities for distributed speech recognition front-ends. In Proc. AVIOS00: The Speech Applications Conference, San Jose, USA.

Pearce, D., 2004. Robustness to transmission channel—the DSR approach. In: Proc. Robustness 2004, Norwich, UK.

Pearce, D., Hirsch, H., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. ICSLP00, Beijing, China.

Peinado, A.M., Sanchez, V., Segura, J.C., Perez-Cordoba, J.L., 2001. MMSE-based channel error mitigation for distributed speech recognition. In: Proc. Eurospeech01.

Peinado, A., Sanchez, V., Perez-Cordoba, J., de la Torre, A., 2003. HMM-based channel error mitigation and its application to distributed speech recognition. Speech Commun. 41, 549–561.

Peinado, A., Sanchez, V., Perez-Cordoba, J., Rubio, A., 2005. Efficient MMSE-based channel error mitigation techniques—application to distributed speech recognition over wireless channels. IEEE Trans. Wireless Commun. 4 (1), 14–19.

Pelaez-Moreno, C., Gallardo-Antolin, A., Diaz-de-Maria, F., 2001. Recognizing voice over IP: a robust front-end for speech recognition on the World Wide Web. IEEE Trans. Multimedia (June), 2001.

Perkins, C., Hodson, O., Hardman, V., 1998. A survey of packet loss recovery techniques for streaming audio. IEEE Network 12, 40–48.

Potamianos, A., Weerackody, V., 2001. Soft-feature decoding for speech recognition over wireless channels. In: Proc. ICASSP01, USA.

Ramabadran, T., Sorin, A., McLaughlin, M., Chazan, D., Pearce, D., Hoory, R., 2004. The ETSI extended distributed speech recognition (DSR) standards: server-side speech reconstruction. In: Proc. ICASSP04, Montreal, Quebec, Canada.

Ramakrishana, B.R., 2000. Reconstruction of incomplete spectrograms for robust speech recognition, PhD dissertation, Carnegie Mellon University.

Rahim, M.G., Lee, C.-H., Juang, B.-H., 1997. Discriminative utterance verification for connected digits recognition. IEEE Trans. Speech Audio Process. 5 (3), 266–277.

Ramsey, J.L., 1970. Realization of optimum interleavers. IEEE Trans. Inform. Theory 16 (3), 338–345.

Riskin, E.V., Boulis, C., Otterson, S., Ostendorf, M., 2001. Graceful degradation of speech recognition performance over lossy packet networks. In: Proc. Eurospeech01.

Rose, R.C., 2004. Environmental robustness in automatic speech recognition. In: Proc. Robust2004.

Rose, R.C., Arizmendi, I., Parthasarathy S., 2003. An efficient framework for robust mobile speech recognition services. In: Proc. ICASSP03, Hong Kong, China.

Schulzrinne, H., Casner, S., Frederick, R., Jacobson, V., 2003. RTP: A Transport Protocol for Real-Time Applications. IETF Audio/Video Transport WG, RFC3550.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423, 623–656.

Sklar, B., Harris, F.J., 2004. The ABCs of linear block codes. IEEE Signal Process. Mag. 2004 (July), 14–35.

Sorin, A., Ramababran, T., Chazan, D., et al., 2004. The ETSI extended distributed speech recognition (DSR) standards: client side processing and tonal language recognition evaluation. In: Proc. ICASSP04.

Srinivasamurthy, N., Ortega, A., Narayanan, S., 2001. Efficient scalable speech compression for scalable speech recognition. In: Proc. Eurospeech01, Aalborg, Denmark.

Srinivasamurthy, N., Ortega A., Narayanan, S., 2004. Enhanced standard compliant distributed speech recognition (Aurora encoder) using rate allocation. In: Proc. ICASSP04.

Sukkar, R.A., Chengalvarayan, R., Jacob, J.J., 2002. Unified speech recognition for landline and wireless environments. In: Proc. ICASSP02.

Tan, Z.-H., Dalsgaard, P., 2002. Channel error protection scheme for distributed speech recognition. In: Proc. ICSLP02, Denver, USA.

Tan, Z.-H., Dalsgaard, P., Lindberg, B., 2003a. OOV-detection and channel error protection for distributed speech recognition over wireless networks. In: Proc. ICASSP03, Hong Kong, China.

Tan, Z.-H., Dalsgaard, P., Lindberg, B., 2003b. Partial splicing packet loss concealment for distributed speech recognition. IEE Electron. Lett. 39 (22), 1619–1620.

Tan, Z.-H., Dalsgaard, P., Lindberg, B., 2004a. A subvector-based error concealment algorithm for speech recognition over mobile networks. In: Proc. ICASSP04, Montreal, Quebec, Canada.

Tan, Z.-H., Lindberg, B., Dalsgaard, P., 2004b. A comparative study of feature-domain error concealment techniques for distributed speech recognition. In: Proc. Robust2004, Norwich, UK.

Tan, Z.-H., Dalsgaard, P., Lindberg, B., 2004c. On the integration of speech recognition into personal networks. In: Proc. ICSLP04, Jeju Island, Korea.

Viikki, O., 2001. ASR in portable wireless devices. In: Proc. ASRU01, Madonna di Campiglio, Italy.

Wang, Y., Zhu, Q.-F., 1998. Error control and concealment for video communication: a review. Proc. IEEE 86 (5), 974–997.

Wang, Y., Panwar, S., Lin, S., Mao, S., 2002. Wireless video transport using path diversity: multiple description vs. layered coding. In: Proc. IEEE ICIP 2002.

Weerackody, V., Reichl, W., Potamianos, A., 2001. Speech recognition for wireless applications. In: Proc. IEEE Int. Conf. Commun. 2001.

Weerackody, V., Reichl, W., Potamianos, A., 2002. An error-protected speech recognition system for wireless zcommunications. IEEE Trans. Wireless Commun. 1 (2), 282–291.

Xie, Q., Pearce, D., 2004. RTP Payload Formats for ETSI ES 202 050, ES 202 211, and ES 202 212 Distributed Speech Recognition Encoding, June 2004, Available from: <http://www.ietf.org/internet-drafts/draft-ietfavt-rtp-dsr-codecs-03.txt>.

Yoma, N.B., McInnes, F.R., Jack, M.A., 1998. Weighted Viterbi algorithm and state duration modelling for speech recognition in noise. In: Proc. ICASSP98.

Zhong, X., Arrowood, J., Moreno, A., Clements, M., 2002. Multiple description coding for recognizing voice over IP. In: Proc. IEEE Digital Signal Processing Workshop.

Zhu, Q., Alwan, A., 2001. An efficient and scalable 2D DCT-based feature coding scheme for remote speech recognition. In: Proc. ICASSP01.